Fast Two-Timescale Stochastic EM Algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

Using the Expectation-Maximization (EM) algorithm is the most popular choice for current latent data model learning tasks. For today's modern and complex models, variants of the EM have been initially introduced by [20], using incremental updates to scale to large datasets, and by [24, 8], using Monte-Carlo (MC) approximations to bypass the impossible 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 double stages of stochastic updates to tackle an essential large and nonconvex optimization task for latent data models. We motivate the choice of a double dynamics by invoking the variance reduction virtue of each stage of the method on both sources of noise: the incremental update and the MC approximation. We establish finite-time and global convergence bounds for nonconvex objective functions. Numerical applications are also presented in this article to illustrate our findings.

1 Introduction

Learning latent data models is critical for modern machine learning problems, see [18] for references. We formulate the training of such model as an empirical risk minimization problem:

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

We denote the observations by $\{y_i\}_{i=1}^n$, $\Theta \subset \mathbb{R}^d$ is the convex parameters space. We consider a smooth convex regularization noted $\mathrm{r} : \Theta \to \mathbb{R}$ and $g(y;\boldsymbol{\theta})$ is the (incomplete) likelihood of each observation. The objective function $\overline{\mathsf{L}}(\boldsymbol{\theta})$ is possibly *nonconvex* and is assumed to be lower bounded.

In the latent variable model, $g(y_i;\boldsymbol{\theta})$, 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 make the assumption of a complete model belonging to the curved exponential family:

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

where $\psi(\boldsymbol{\theta})$, $h(z_i, y_i)$ are scalar functions, $\phi(\boldsymbol{\theta}) \in \mathbb{R}^k$ is a vector function, and $S(z_i, y_i) \in \mathbb{R}^k$ is the complete data sufficient statistics. Full batch EM [9] is the method of reference for that kind of task and is a two steps procedure. The E-step amounts to computing the conditional expectation of the complete data sufficient statistics,

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

The M-step is given by

$$\mathsf{M}\text{-step: } \hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\overline{\mathbf{s}}(\boldsymbol{\theta})) := \underset{\vartheta \in \Theta}{\arg\min} \left\{ \mathrm{r}(\vartheta) + \psi(\vartheta) - \left\langle \overline{\mathbf{s}}(\boldsymbol{\theta}) \,|\, \phi(\vartheta)\right\rangle \right\}, \tag{4}$$

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 a full pass over the dataset at each iteration and (b) the complexity of modern models makes the expectation in (3) intractable. So far, both challenges have been addressed separately, to the best of our knowledge, as detailed the sequel.

Prior Work Inspired by stochastic optimization procedures, [20] and [5] developed respectively an incremental and an online variant of the E-step in models where the expectation is computable then extensively used and studied in [21, 15, 4]. Some improvements of that methods have been provided and analyzed, globally and in finite-time, in [12] 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 first method was the Monte-Carlo EM (MCEM) introduced in the seminal paper [24] where a MC approximation for this expectation is computed. A variant of that method is the Stochastic Approximation of the EM (SAEM) in [8] leveraging the power of Robbins-Monro type of update [23] to ensure pointwise convergence of the vector of estimated parameters rather using a decreasing stepsize than increasing the number of MC samples. The MCEM and the SAEM have been successfully applied in mixed effects models [17, 10, 3] or to do inference for joint modeling of time to event data coming from clinical trials in [7], among other applications. Recently, an incremental variant of the SAEM was proposed in [14] showing positive empirical results but its analysis is limited to asymptotic consideration. Gradient-based methods have been developed and analyzed in [25] but they remain out of the scope of this paper as they tackle the high-dimensionality issue.

Contributions This paper *introduces* and *analyzes* a new class of methods which purpose is to update two proxies for target expected quantities in a two-timescale manner. Those approximated quantities are then used to optimize (1) for modern examples and settings using EM Maximization step. The main contributions of the paper are:

- We propose a two-timescale method based on Stochastic Approximation (SA), to alleviate the problem of MC computation, and on Incremental updates, to scale to large datasets. We describe in details the edges of each level of our method based on variance reduction arguments. The derivation of 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 suboptimality condition in the nonconvex literature, *i.e.,* the second order moment of the gradient of the objective function.

In Section 2 we formalize both incremental and Monte-Carlo variants of the EM. Then, we introduce our two-timescale class of EM algoirithms for which we derive several global statistical guarantees in Section 3 for possibly *nonconvex* functions. Section 4 is devoted to numerical illustrations.

2 Two-Timescale Stochastic EM Algorithms

We recall and formalize in this section the different methods found in the literature that aim to solving the large-scale problem and the intractable expectation. 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 nonlinear 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 term. For all $i \in [\![1, n]\!]$, draw for $m \in [\![1, M]\!]$, samples $z_{i,m} \sim p(z_i|y_i;\theta)$ and compute the MC integration $\tilde{\mathbf{s}}$ of the deterministic quantity $\overline{\mathbf{s}}(\boldsymbol{\theta})$:

$$\textsf{MC-step}: \ \tilde{\mathbf{s}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}, y_i) \ , \tag{5}$$

and then update the parameter $\hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\tilde{\mathbf{s}})$. 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 following quantity

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

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

$$\textsf{SA-step}: \ \hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \ , \tag{7}$$

where $\{\gamma_k\}_{k>1} \in (0,1)$ is a sequence of decreasing step sizes 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 [8]. In the simulation step (6), since the loss function between the observed data y_i and the latent variable z_i can be nonconvex, sampling from the posterior distribution $p(z_i|y_i;\theta)$, under the current model θ, requires using an inference algorithm. [13] proved almost sure convergence of the sequence of parameters obtained by this algorithm coupled with an MCMC procedure during the simulation step. In simple scenarios, the samples $\{z_{i,m}\}_{m=0}^{M-1}$ are conditionally independent and identically distributed with distribution $p(z_i,\theta)$. Nevertheless, in most cases, sampling exactly from this distribution is not an option and the Monte Carlo batch is sampled by Monte Carlo Markov Chains (MCMC) algorithm. It is inappropriate to start with small values for step size γ_k and large values for the number of simulations M_k. Rather, it is recommended that one decrease γ_k, as in $\gamma_k = 1/k^\alpha$, where $\alpha \in (0,1)$, and keep a constant and small number of MC samples M_k which shows a great advantage over the MC-step (5), which requires large M_k to converge. This Robbins-Monro type of update represents the *first level* of our algorithm, needed to temper the variance and noise implied by MC 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 Bi-Level Stochastic EM Methods

Efficient strategies to scale to large datasets include incremental and variance reduced methods. We will explicit a general update that covers those latter variants and that represents the *second level* of our algorithm, namely the incremental update of the approached statistics $\hat{S}^{(k)}$ inside the SA-Step.

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

Note $\{\rho_k\}_{k>1} \in (0,1)$ is a sequence of step sizes, $\boldsymbol{\mathcal{S}}^{(k)}$ is a proxy for $\tilde{S}^{(k)}$, If the stepsize is equal to one and the proxy $\boldsymbol{\mathcal{S}}^{(k)} = \hat{S}^{(k)}$, 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 $\boldsymbol{\mathcal{S}}^{(k)} = \tilde{S}^{(k)}$, then we recover the MCEM [24].

The following table provides the definitions of the proxy $\boldsymbol{\mathcal{S}}^{(k)}$ for three variants of (8). For all methods, we define a random index drawn at iteration k, noted $i_k \in [\![1,n]\!]$, and $\tau_i^k = \max\{k' : i_{k'} = i,\ k' < k\}$ as the iteration index where $i \in [\![1,n]\!]$ is last drawn prior to iteration k. The proposed fiTTEM method draws *two* indices *independently* and uniformly as $i_k, j_k \in [\![1,n]\!]$. Thus, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample $j \in [\![1,n]\!]$ is last drawn as j_k prior to iteration k in addition to τ_i^k which was defined *w.r.t.* i_k.

$$\textsf{iSAEM} \quad \boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + \tfrac{1}{n}\big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)}\big) \tag{9}$$

$$\textsf{vrTTEM} \quad \boldsymbol{\mathcal{S}}^{(k+1)} = \tilde{S}^{(\ell(k))} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}\big) \tag{10}$$

$$\textsf{fiTTEM} \quad \boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big), \quad \overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\big(\tilde{S}_{j_k}^{(k)} - \tilde{S}_{j_k}^{(t_{j_k}^k)}\big) \tag{11}$$

where $\tilde{S}_{i_k}^{(k)}$ is the MC approximation of the expectation $\overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})$:

$$\tilde{S}_{i_k}^{(k)} = \frac{1}{M_k}\sum_{m=1}^{M_k} S(z_{i_k,m}^{(k)}, y_{i_k}) \quad \text{with} \quad z_{i_k,m}^{(k)}, \sim p(z_{i_k}|y_{i_k};\theta^{(k)}) \ .$$

The stepsize is set to $\rho_{k+1} = 1$ for 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 $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \overline{\mathrm{s}}^{(0)}$ for the fiTTEM which can be seen as a slightly modified version of SAGA inspired by [22]. Moreover, for vrTTEM we set an epoch size of m and 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 $\hat{\boldsymbol{\theta}}^{(K)}$ where K is a randomly chosen termination point.

Algorithm 1 Two-Timescale Stochastic EM methods.

Input: initializations $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathrm{s}}^{(0)} \leftarrow \hat{S}^{(0)}$, $K_{\mathsf{max}} \leftarrow$ max. iteration number.

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

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\mathsf{max}}-1} \gamma_\ell} = \frac{\gamma_k}{\mathsf{P}_{\mathsf{max}}} \,. \tag{12}$$

for $k = 0, 1, 2, \dots, K$ **do**

  Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiTTEM).

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

  Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using (9) or (10) or (11).

  Compute $\hat{S}^{(k+1)}$ and $\hat{\mathrm{s}}^{(k+1)}$ using respectively (8) and (7):

$$\begin{aligned} \tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho_{k+1}\big(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k)}\big) \\ \hat{\mathrm{s}}^{(k+1)} &= \hat{\mathrm{s}}^{(k)} + \gamma_{k+1}\big(\tilde{S}^{(k+1)} - \hat{\mathrm{s}}^{(k)}\big) \end{aligned} \tag{13}$$

  Compute $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathrm{s}}^{(k+1)})$ via the M-step (4).

end for

Return: $\hat{\boldsymbol{\theta}}^{(K)}$.

The update in (13) is said to have two-timescale as the step sizes 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, than $\hat{\mathrm{s}}^{(k+1)}$, determined by γ_k. The next section presents the main results of this paper and establishes global and finite-time bounds for the three different updates of our two-timescale scheme.

3 Finite Time Analysis of the Two-Timescale Scheme

Following [5], it can be shown that stationary points of the objective function (1) corresponds to the stationary points of the following *nonconvex* Lyapunov function:

$$\min_{\mathrm{s}\in\mathsf{S}} V(\mathrm{s}) := \overline{\mathrm{L}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) = \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathrm{s})) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathrm{s})) \,, \tag{14}$$

that we propose to study in this article. Several critical assumptions required to derive convergence guarantees read as follows:

H1. *The sets* Z, S *are compact. There exists constants* $C_{\mathsf{S}}, C_{\mathsf{Z}}$ *such that:*

$$C_{\mathsf{S}} := \max_{\mathbf{s},\mathbf{s}'\in\mathsf{S}} \|\mathbf{s} - \mathbf{s}'\| < \infty, \quad C_{\mathsf{Z}} := \max_{i\in[\![1,n]\!]} \int_{\mathsf{Z}} |S(z, y_i)| \mu(\mathrm{d}z) < \infty. \tag{15}$$

H2. *The conditional distribution is smooth on* $\mathrm{int}(\Theta)$*. For any* $i \in [\![1, n]\!]$*,* $z \in \mathsf{Z}$*,* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathrm{int}(\Theta)^2$*, we have* $\left|p(z|y_i; \boldsymbol{\theta}) - p(z|y_i; \boldsymbol{\theta}')\right| \leq \mathrm{L}_p \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

We also recall from the introduction that we consider curved exponential family models. besides:

H3. *For any* $\boldsymbol{s} \in \mathsf{S}$*, the function* $\boldsymbol{\theta} \mapsto L(s, \boldsymbol{\theta}) := \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \mathrm{s} \,|\, \phi(\boldsymbol{\theta}) \right\rangle$ *admits a unique global minimum* $\overline{\boldsymbol{\theta}}(\mathrm{s}) \in \mathrm{int}(\Theta)$*. In addition,* $\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))$ *is full rank,* L_ϕ*-Lipschitz and* $\overline{\boldsymbol{\theta}}(\mathrm{s})$ *is* L_θ*-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 $\boldsymbol{\theta} \mapsto L(\mathrm{s}, \boldsymbol{\theta}) = \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \mathrm{s} \,|\, \phi(\boldsymbol{\theta}) \right\rangle$, and define

$$\mathrm{B}(\mathrm{s}) := \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))\Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathrm{s}, \overline{\boldsymbol{\theta}}(\mathrm{s}))\Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))^{\top}. \tag{16}$$

H4. *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 are two-timescale where the first stage corresponds to the variance reduction trick used in [12] in order to accelerate incremental methods and reduce the variance induced by the index sampling. The second stage is the Robbins-Monro type of update that aims to reduce the variance induced by the MC approximations As the expectations (3) are never available, we introduce the errors when approximating the quantity $\overline{\mathrm{s}}_i(\hat{\boldsymbol{\theta}}(\hat{\mathrm{s}}^{(k-1)}))$ at iteration $k+1$:

$$\eta_i^{(r)} := \tilde{S}_i^{(r)} - \overline{\mathrm{s}}_i(\vartheta^{(r)}) \quad \text{for all} \quad i \in [\![1, n]\!],\, r > 0 \quad \text{and} \quad \vartheta \in \Theta \, . \tag{17}$$

For instance, we consider that the MC approximation is unbiased if for all $i \in [\![1, n]\!]$ and $m \in [\![1, M]\!]$, the samples $z_{i,m} \sim p(z_i|y_i;\theta)$ are i.i.d. under the posterior distribution, *i.e.*, $\mathbb{E}[\eta_i^{(r)}|\mathcal{F}_r] = 0$ where $\mathcal{F}_r$ is the filtration up to iteration r. The following results are derived under the assumption of control of the fluctuations implied by the approximation stated as follows:

H5. *There exist a positive sequence of MC batch size* $\{M_r\}_{r>0}$ *and constants* (C, C_η) *such that for all* $k > 0$, $i \in [\![1, n]\!]$ *and* $\vartheta \in \Theta$:

$$\mathbb{E}\left[\left\|\eta_i^{(r)}\right\|^2\right] \leq \frac{C_\eta}{M_r} \quad \textit{and} \quad \mathbb{E}\left[\left\|\mathbb{E}[\eta_i^{(r)}|\mathcal{F}_r]\right\|^2\right] \leq \frac{C}{M_r} \, . \tag{18}$$

We can prove two important results on the Lyapunov function. The first one suggests smoothness:

Lemma 1. *[12] Assume H1-H4. For all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}$ *and* $i \in [\![1, n]\!]$, *we have*

$$\|\overline{\mathrm{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathrm{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \, \|\mathbf{s} - \mathbf{s}'\|, \;\; \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \mathrm{L}_V \, \|\mathbf{s} - \mathbf{s}'\| \, , \tag{19}$$

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

and the second one suggests a growth condition on the gradient of V depending on the mean field of the algorithm:

Lemma 2. *Assume H3,H4. For all* $\mathbf{s} \in \mathsf{S}$,

$$\upsilon_{\min}^{-1} \big\langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \big\rangle \geq \big\|\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\big\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2 \, , \tag{20}$$

Proof of this Lemma can be found in Appendix A.

3.1 Global Convergence of Incremental Stochastic EM Algorithms

We present in this section a finite-time analysis of the incremental variant of the Stochastic Approximation of the EM algorithm. We want to draw the attention of the readers that the word "global" here does not mean for a global optimum of the nonconvex function, but of the independence of our analysis on the initialization and the iteration k (finite time).

The following main result for the iSAEM algorithm, which proof can be found in Appendix B, is derived under a control of the Monte Carlo fluctuations as described by assumption H 5 and is built upon an intermediary Lemma, detailed in in Appendix B, characterizing the quantity of interest $\hat{S}^{(k+1)} - \hat{\mathrm{s}}^{(k)}$. Typically, the controls exhibited above are of interest when the number of MC samples M_k increase with k.

Theorem 1. *Assume H1-H5. Let* $K_{\max}$ *be a positive integer. Let* $\{\gamma_k, k \in \mathbb{N}\}$ *be a sequence of positive step sizes and consider the iSAEM sequence* $\big\{\hat{\mathrm{s}}^{(k)}, k \in \mathbb{N}\big\}$ *obtained with* $\rho_{k+1} = 1$ *for any* $k > 0$. *We also set* $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\gamma_{k+1} = \frac{1}{k^a \alpha c_1 \overline{L}}$ *where* $a \in (0, 1)$, $\beta = \frac{c_1 \overline{L}}{n}$. *Assume that* $\hat{\mathrm{s}}^{(k)} \in \mathcal{S}$ *for any* $k \leq K_{\max}$.

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

3.2 Global Convergence of Two-Timescale Stochastic EM Algorithms

We now proceed by giving our main result regarding the global convergence of the fiTTEM algorithm. Two important auxiliary Lemmas, which proofs are given in Appendix C.1, are need in order to derive our finite-time bound. The first one derives an identity for the quantity $\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2]$ using the vrTTEM update:

Lemma 3. *For any $k \geq 0$ and consider the vrTTEM update in (10) with $\rho_k = \rho$, it holds for all $k > 0$*

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

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 4. *For any $k \geq 0$ and consider the fiTTEM update in (11) with $\rho_k = \rho$, it holds for all $k > 0$*

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

Recalling that K is an independent discrete r.v. drawn from $\{1, \ldots, K_{\mathsf{max}}\}$ with distribution $\{\gamma_k/\mathsf{P}_{\mathsf{max}}, 0 \leq k \leq K_{\max} - 1\}$, as in (12), we have

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

We now state the main result regarding the vrTTEM method, see proof in Appendix D:

Theorem 2. *Assume H1-H5. Let $K_{\max}$ be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the vrTTEM sequence $\left\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\right\}$ obtained with $\rho_{k+1} = \rho$ for any $k > 0$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$. Setting $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\rho = \frac{\mu}{c_1 \overline{L} n^{2/3}}$, $m = \frac{n c_1^2}{2\mu^2 + \mu c_1^2}$, a constant $\mu \in (0,1)$, $\gamma_{k+1} = \frac{1}{k^a \overline{L}}$ where $a \in (0,1)$, we have the following bound:*

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

We now state the main result regarding the fiTTEM method.

Theorem 3. *Assume H1-H5. Let $K_{\max}$ be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the fiTTEM sequence $\left\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\right\}$ obtained with $\rho_{k+1} = \rho$ for any $k > 0$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$. By setting $\alpha = \max\{2, 1 + 2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \overline{L} n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$ and $\gamma_{k+1} = \frac{1}{k^a \alpha c_1 \overline{L}}$ where $a \in (0,1)$, we have the following bound:*

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

Proof of this Theorem can be found in Appendix E. Note that in those two bounds, the quantities $\tilde{\eta}^{(k+1)}$ and $\Xi^{(k+1)}$ depend only on the MC fluctuations $\mathbb{E}\left[\left\|\eta_{i_k}^{(k)}\right\|^2\right]$ and some constants.

While Theorem 1 suffers only from the MC noise induced by the latent data sampling step, 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 $V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})$, and a double dynamics burden exemplified by the term $\mathbb{E}[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2]$.

Indeed, the following remarks are worth noting on the quantity $\mathbb{E}[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2]$:

- This term is the price we pay for the two-timescale dynamics and corresponds to the gap between the two asynchronous updates (one is on $\hat{\boldsymbol{s}}^{(k)}$ and the other on $\tilde{S}^{(k)}$).
- It is trivial to see that if $\rho = 1$, *i.e.,* there is no variance reduction, then for any $k > 0$

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

  which strengthen the fact that this quantity characterizes the impact of the variance reduction technique introduced in our two stages class of methods.

The following lemma, which proof can be found in Appendix C.2, characterizes this gap:

Lemma 5. *Consider a decreasing stepsize $\gamma_k \in (0,1)$ and a constant $\rho \in (0,1)$, then the following inequality holds:*

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

where $\boldsymbol{\mathcal{S}}^{(k)}$ is defined either by (10) (vrTTEM) or (11) (fiTTEM).

In the next section, we illustrate the benefits of our two-timescale class of algorithms on several numerical applications.

4 Numerical Examples

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, and 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 Gaussian mixture of M 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:

$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i)\left[\log(\omega_m) - \mu_m^2/2\right] + \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i)\mu_m y_i + \text{constant} \ . \quad (21)$$

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 on Θ 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 30 synthetic datasets of size $n = 10^5$ from a GMM model with $M = 2$ components with two mixtures with means

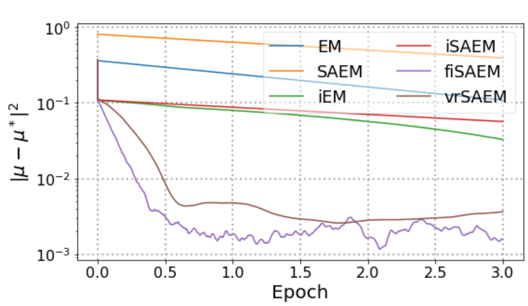

Figure 1: Precision $|\mu^{(k)} - \mu^*|^2$ per epoch

$\mu_1 = -\mu_2 = 0.5$. We run the bEM method until convergence (to double precision) to obtain the ML estimate $\mu^\star$ averaged on 50 datasets. We compare the bEM, 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 of all method as $\gamma_k = 1/k^\alpha$ with $\alpha = 0.5$, and the stepsizes of the Incremental-step for vrTTEM and the fiTTEM to a constant stepsize equal to $1/n^{2/3}$. The number of MC samples is fixed to $M = 10$ chains. Figure 1 shows the precision $|\mu - \mu^*|^2$ for the different methods against the epoch(s) elapsed (one epoch equals n iterations). Besides, vrTTEM and fiTTEM methods outperform the other stochastic methods, supporting the benefits of our scheme.

4.2 Deformable Template Model for Image Analysis

Let $(y_i, i \in [\![1, n]\!])$ be observed gray level images defined on a grid of pixels. Let $u \in \mathcal{U} \subset \mathbb{R}^2$ denotes 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:

$$y_i(u) = I\left(x_u - \Phi_i\left(x_u, z_i\right)\right) + \varepsilon_i(u) \tag{22}$$

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

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

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

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

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

Numerical Experiment: We apply model (22) and our algorithms to a collection of handwritten digits, called the US postal database [11], featuring $n = 1\,000$ (16×16)-pixel images for each class of digits from 0 to 9. The main difficulty with these data comes 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 (22) in order to account for both sources of variability: the intrinsic template to each class of digit and the small and local deformation in each observed image.

Figure 2: Training set of the USPS database (20 images for figit 5)

Figure 3 shows the resulting synthetic images for digit 5 through several epochs and for the batch method, the online SAEM, the incremental SAEM and the various TTS methods. We choose 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. $k_p = 15$ and $k_g = 6$ equidistributed landmarks points are chosen on the grid for the training. Those hyperparameters are inspired by a relevant study in [2]. The kernel covariance matrices are important hyperparameters in such study since they have a direct impact on the sharpness of the templates. Intuitively, if those variances are large, the kernels centered arounds the equidistributed landmark spread out on too many of its neighbors. Bad choices of such hyperparameters can lead to thicker shapes.

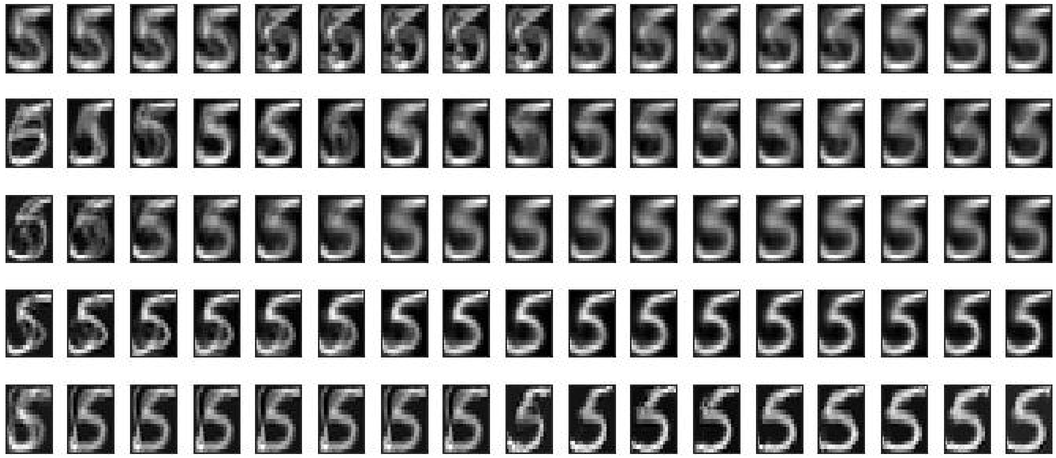

Figure 3: Estimation of the template: from top to bottom: batch, online, iSAEM ,vrTTEM and fiTTEM through 7 epochs. Note that Batch method templates are replicated in-between epochs for a fair comparison with incremental variants.

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 batch method. After a few epochs, the batch SAEM seem to estimate similar template as the incremental an online methods due to the high variance of the latter procedures. Our variance reduced and fast incremental come into play in the long run and sharpen the final template estimates contrasting between the background and the regions of interest in the image.

Third numerical application on PK Model with Absorption Lag Time is in the appendix for lack of space.

5 Conclusion

This paper introduces a new class of two-timescale EM methods for learning latent data 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 represent the *first* level of our class of methods and the scalability with the number of samples is performed through a variance reduced and incremental type of update, the *second* and last level of our newly introduced scheme. The various methods are interpreted as scaled gradient methods, in the space of the sufficient statistics, and our convergence results are *global*, in the sense of independent of the initial values, and *non-asymptotic*, true for any random termination number.

References

[1] S. Allassonnière, Y. Amit, and A. 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.

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

[3] C. Baey, S. Trevezas, and P.-H. 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.

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

[5] O. Cappé and E. 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.

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

[7] A. Chakraborty and K. 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.

[8] B. Delyon, M. Lavielle, and E. 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.

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

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

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

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

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

[14] E. Kuhn, C. Matias, and T. Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.

[15] P. Liang and D. 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.

[16] F. Maire, E. Moulines, and S. Lefebvre. Online em for functional data, 2016. URL http://arxiv.org/abs/1604.00570. cite arxiv:1604.00570v1.pdf.

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

[18] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

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

[20] R. M. Neal and G. 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.

[21] H. D. Nguyen, F. Forbes, and G. J. McLachlan. Mini-batch learning of exponential family finite mixture models. *Statistics and Computing*, pages 1–18, 2020.

[22] S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*, 2016.

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

[24] G. C. Wei and M. 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.

[25] R. Zhu, L. Wang, C. Zhai, and Q. Gu. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4180–4188. JMLR. org, 2017.

A Proof of Lemma 2

Lemma. *Assume H3,H4. For all* $\mathbf{s} \in \mathsf{S}$,

$$\upsilon_{\min}^{-1}\left\langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{s}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\right\rangle \geq \left\|\mathbf{s} - \overline{s}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\right\|^2 \geq \upsilon_{\max}^{-2}\|\nabla V(\mathbf{s})\|^2, \tag{23}$$

Proof Using H3 and the fact that we can exchange integration with differentiation and the Fisher's identity, we obtain

$$\begin{aligned}
\nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^{\top}\left(\nabla_{\boldsymbol{\theta}}\, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathsf{L}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\right) \\
&= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^{\top}\left(\nabla_{\boldsymbol{\theta}}\psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}}\, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^{\top}\overline{s}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\right) \\
&= \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^{\top}\, \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^{\top}\left(\mathbf{s} - \overline{s}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\right),
\end{aligned} \tag{24}$$

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 H3, we show that:

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

The above yields

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

where we recall $\mathrm{B}(\mathbf{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\left(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s};\overline{\boldsymbol{\theta}}(\mathbf{s}))\right)^{-1}\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^{\top}$. The proof of (23) follows directly from the assumption H4. $\square$

B 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 $\mathbb{E}\left[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\right]$:

Lemma 6. *Assume H1. The update (9) 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:

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

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

Proof From update (9), 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:

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

The following equalities:

$$\mathbb{E}\left[\tilde{S}_i^{(\tau_i^k)}|\mathcal{F}_k\right] = \frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)} \quad \text{and} \quad \mathbb{E}\left[\overline{\mathbf{s}}_{i_k}(\boldsymbol{\theta}^{(k)})|\mathcal{F}_k\right] = \overline{\mathbf{s}}^{(k)}$$

concludes the proof of the Lemma. □

And the following auxiliary Lemma setting an upper bound for the quantity $\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right]$

Lemma 7. *For any $k \geq 0$ and consider the iSAEM update in (9), it holds that*

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

Proof Applying the iSAEM update yields:

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

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 H1-H5. Let $K_{\max}$ be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the iSAEM sequence $\left\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\right\}$ obtained with $\rho_{k+1} = 1$ for any $k > 0$. We also set $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\gamma_{k+1} = \frac{1}{k^a \alpha c_1 \overline{L}}$ where $a \in (0,1)$, $\beta = \frac{c_1 \overline{L}}{n}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.*

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

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 sidesyields:

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

Using Lemma 6, we obtain:

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

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

$$\begin{aligned}
a_k\mathbb{E}\left[\left\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right\|^2\right] \leq &\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] + \frac{\gamma_{k+1}^2 \, \mathrm{L}_V}{2}\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right] \\
&+ \frac{\gamma_{k+1}(1 - \frac{1}{n})}{2\beta}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] + \frac{\gamma_{k+1}}{2n}\mathbb{E}\left[\left\|\eta_{i_k}^{(k)}\right\|^2\right]
\end{aligned} \tag{25}$$

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 (25):

$$\begin{aligned}
\left(a_k - 2\gamma_{k+1}^2 \, \mathrm{L}_V\right)\mathbb{E}\left[\left\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right\|^2\right] \leq &\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] \\
&+ \gamma_{k+1}\left(\frac{1}{2\beta}(1 - \frac{1}{n}) + 2\gamma_{k+1} \, \mathrm{L}_V\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+ \gamma_{k+1}\left(\gamma_{k+1} \, \mathrm{L}_V + \frac{1}{2n}\right)\mathbb{E}\left[\left\|\eta_{i_k}^{(k)}\right\|^2\right] \\
&+ \frac{\gamma_{k+1}^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2}{n^3}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(\tau_i^k)}\|^2]
\end{aligned} \tag{26}$$

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 the 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

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

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

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

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:

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

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

$$
\begin{aligned}
&\sum_{k=0}^{K_{\max}-1} \Delta^{(k+1)} \\
&= 4\sum_{k=0}^{K_{\max}-1} (\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})\omega_{k,1}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + 2\sum_{k=0}^{K_{\max}-1} (\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})\omega_{k,1}\mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\
&+ \sum_{k=0}^{K_{\max}-1} 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)\omega_{k,1}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
&\leq \sum_{k=0}^{K_{\max}-1} \frac{4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \sum_{k=0}^{K_{\max}-1} \frac{2\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\
&+ \sum_{k=0}^{K_{\max}-1} \frac{4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right]
\end{aligned}
\tag{27}
$$

We recall (26) where we have summed on both sides from $k=0$ to $k=K_{\max}-1$:

$$
\begin{aligned}
\sum_{k=0}^{K_{\max}-1} \left(a_k - 2\gamma_{k+1}^2 \mathrm{L}_V\right)\mathbb{E}\left[\left\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right\|^2\right] \leq &\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(K)})\right] \\
&+ \sum_{k=0}^{K_{\max}-1} \gamma_{k+1}\left(\frac{1}{2\beta}(1-\frac{1}{n}) + 2\gamma_{k+1}\mathrm{L}_V\right)\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(\tau_i^k)} - \overline{\mathbf{s}}^{(k)}\right\|^2\right] \\
&+ \sum_{k=0}^{K_{\max}-1} \gamma_{k+1}\left(\gamma_{k+1}\mathrm{L}_V + \frac{1}{2n}\right)\mathbb{E}\left[\left\|\eta_{i_k}^{(k)}\right\|^2\right] \\
&+ \sum_{k=0}^{K_{\max}-1} \frac{\gamma_{k+1}^2 \mathrm{L}_V \mathrm{L}_{\mathbf{s}}^2}{n^2}\Delta^{(k)}
\end{aligned}
\tag{28}
$$

Plugging (27) into (28) results in:

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

where:

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

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:

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

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. $\square$

C Proofs of Auxiliary Lemmas

C.1 Proof of Lemma 3 and Lemma 4

Lemma. *For any $k \geq 0$ and consider the vrTTEM update in (10) with $\rho_k = \rho$, it holds for all $k > 0$*

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

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:

$$\begin{aligned}\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) = -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho \boldsymbol{\mathcal{S}}^{(k+1)}) \\ &= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\right]\right)\end{aligned} \tag{29}$$

We observe, using the identity (29), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq 2\rho^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] + 2(1-\rho)^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - \tilde{S}^{(k)}\|^2] \tag{30}$$

For the latter term, we obtain its upper bound as

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

where (a) uses the variance inequality and (b) uses Lemma 1. Substituting into (30) proves the lemma. ☐

Lemma. *For any $k \geq 0$ and consider the fiTTEM update in (11) with $\rho_k = \rho$, it holds for all $k > 0$*

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

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:

$$\begin{aligned}\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}) \\ &= -\gamma_{k+1}(\hat{\boldsymbol{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho \boldsymbol{\mathcal{S}}^{(k+1)}) \\ &= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\right]\right) \\ &= -\gamma_{k+1}\left((1-\rho)\left[\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right] + \rho\left[\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)} - \left(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\right)\right]\right)\end{aligned} \tag{31}$$

We observe, using the identity (31), that

$$\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq 2\rho^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] + 2(1-\rho)^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{((k))} - \tilde{S}^{(k)}\|^2] \tag{32}$$

For the latter term, we obtain its upper bound as

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

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 (32) proves the lemma. □

C.2 Proof of Lemma 5

Lemma. *Consider a decreasing stepsize $\gamma_k \in (0,1)$ and a constant ρ, then the following inequality holds:*

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

where $\boldsymbol{\mathcal{S}}^{(k)}$ is defined either by (11) (fiTTEM) or (10) (vrTTEM)

Proof We begin by writing the two-timescale update:

$$\begin{aligned}\tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho\big(\boldsymbol{\mathcal{S}}^{(k+1)} - \tilde{S}^{(k)}\big)\\ \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)})\end{aligned} \tag{33}$$

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 (11). Denote $\delta^{(k+1)} = \hat{\boldsymbol{s}}^{(k+1)} - \tilde{S}^{(k+1)}$. Then from (33), 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

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

□

C.3 Additional Intermediary Result

Lemma 8. *At iteration $k+1$,the drift term of update (11), 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 (18), 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 (18), 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 $\overline{\mathbf{s}}_{i_k}^{(k)}$. $\square$

D Proof of Theorem 2

Theorem. *Assume H1-H5. Let $K_{\max}$ be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the vrTTEM sequence $\{\hat{\mathrm{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = \rho$ for any $k > 0$.*

Assume that $\hat{\mathrm{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$. By setting $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\rho = \frac{\mu}{c_1 \overline{L} n^{2/3}}$, $m = \frac{nc_1^2}{2\mu^2 + \mu c_1^2}$ and a constant $\mu \in (0,1)$ and $\gamma_{k+1} = \frac{1}{k^a \overline{L}}$ where $a \in (0,1)$, we have the following bound:

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

Proof Using the smoothness of V and update (10), we obtain:

$$\begin{aligned} V(\hat{\boldsymbol{s}}^{(k+1)}) &\leq V(\hat{\boldsymbol{s}}^{(k)}) + \left\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle + \frac{\mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \\ &\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1} \left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \tag{34}$$

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

$$\begin{aligned} &\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \\ &\overset{(a)}{\leq} \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle\right] - \gamma_{k+1}\rho \mathbb{E}\left[\left\langle \hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle\right] \\ &+ \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\ &\overset{(b)}{\leq} \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1}\rho \mathbb{E}\left[\left\langle \mathsf{h}_k \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle\right] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle\right] \\ &- \gamma_{k+1}\rho \mathbb{E}\left[\left\langle \eta_{i_k}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle\right] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\ &\overset{(c)}{\leq} \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \left(\gamma_{k+1}\rho \upsilon_{\min} + \gamma_{k+1}\upsilon_{\max}^2\right) \mathbb{E}\left[\|\mathsf{h}_k\|^2\right] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \mathbb{E}[\|\mathsf{H}_{k+1}\|^2] \\ &- \gamma_{k+1}\rho \mathbb{E}\left[\left\|\eta_{i_k}^{(k+1)}\right\|^2\right] - \gamma_{k+1}(1-\rho)\mathbb{E}\left[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2\right] \end{aligned} \tag{35}$$

where we have used (29) 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 the 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

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

where we first used (29) and the last inequality is due to the 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

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

And using Lemma 3 we obtain:

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

Rearranging the terms yields:

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

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

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

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

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

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:

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

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

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

This concludes our proof.

$\square$

E Proof of Theorem 3

Theorem. *Assume H1-H5. Let $K_{\max}$ be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the fiTTEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = \rho$ for any $k > 0$.*

Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$. By setting $\alpha = \max\{2, 1 + 2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, $\beta = \frac{c_1 \overline{L}}{n}$, $\rho = \frac{1}{n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$ and $\gamma_{k+1} = \frac{1}{k^a \alpha c_1 \overline{L}}$ where $a \in (0, 1)$, we have the following bound:

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

Proof Using the smoothness of V and update (11), we obtain:

$$
\begin{aligned}
V(\hat{\boldsymbol{s}}^{(k+1)}) &\leq V(\hat{\boldsymbol{s}}^{(k)}) + \left\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle + \frac{\mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \\
&\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1} \left\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \right\rangle + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2} \|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2
\end{aligned} \tag{36}
$$

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:

$$
\mathbb{E}\left[\left(\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\right) - \mathbb{E}[\overline{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}]\right] = 0 \tag{37}
$$

we have:

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

where $\xi^{(k+1)} = \mathbb{E}\left[\left\|\mathbb{E}[\eta_{i_k}^{(k+1)} | \mathcal{F}_k]\right\|^2\right]$. **Bounding** $\mathbb{E}\left[\|\mathsf{H}_{k+1}\|^2\right]$ Using Lemma 4, we obtain:

$$
\begin{aligned}
&\gamma_{k+1}(\upsilon_{\min}\rho + \upsilon_{\max}^2 - \gamma_{k+1}\rho^2 \mathrm{L}_V)\mathbb{E}\left[\|\mathsf{h}_k\|^2\right] \\
&\leq \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)}) - V(\hat{\boldsymbol{s}}^{(k+1)})\right] + \tilde{\xi}^{(k+1)} + \left((1-\rho)^2\gamma_{k+1}^2 \mathrm{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2}\right)\mathbb{E}[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2] \\
&\frac{\gamma_{k+1}^2 \mathrm{L}_V \rho^2 \mathrm{L}_{\mathbf{s}}^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]
\end{aligned} \tag{38}
$$

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

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]\right) \tag{39}$$

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

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

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

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

where the last inequality is due to the Young's inequality. Plugging this into (39) yields:

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

Subsequently, we have

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

We now use Lemma 4 on $\Big\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\Big\|^2 = \gamma_{k+1}^2\Big\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\Big\|^2$ and obtain:

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

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

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

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

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

We recall (38) where we have summed on both sides from $k = 0$ to $k = K_{\max} - 1$:

$$\begin{aligned}& \mathbb{E}\big[V(\hat{\mathbf{s}}^{(K_{\mathsf{max}})}) - V(\hat{\mathbf{s}}^{(0)})\big] \\ & \leq \sum_{k=0}^{K_{\mathsf{max}}-1}\left\{\gamma_{k+1}(-(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}\rho^2\,\mathrm{L}_V)\mathbb{E}\left[\|\mathsf{h}_k\|^2\right] + \gamma^2\,\mathrm{L}_V\,\rho^2\,\mathrm{L}_{\mathbf{s}}^2\,\Delta^{(k)}\right\} \\ & + \sum_{k=0}^{K_{\mathsf{max}}-1}\left\{\tilde{\xi}^{(k+1)} + \left((1-\rho)^2\gamma_{k+1}^2\,\mathrm{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2}\right)\mathbb{E}[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2]\right\} \\ & \leq \sum_{k=0}^{K_{\mathsf{max}}-1}\left\{\left[-\gamma_{k+1}(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}^2\rho^2\,\mathrm{L}_V + \frac{\rho^2\gamma_{k+1}^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right]\mathbb{E}\left[\|\mathsf{h}_k\|^2\right]\right\} \\ & + \sum_{k=0}^{K_{\mathsf{max}}-1}\Xi^{(k+1)} + \sum_{k=0}^{K_{\mathsf{max}}-1}\Gamma_{k+1}\mathbb{E}\left[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2\right]\end{aligned} \tag{40}$$

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

$$\begin{aligned} &- \gamma_{k+1}(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}^2\rho^2 \, \mathrm{L}_V + \frac{\rho^2\gamma_{k+1}^2 \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \\ &= \gamma_{k+1}\left[-(\upsilon_{\min}\rho + \upsilon_{\max}^2) + \gamma_{k+1}\rho^2 \, \mathrm{L}_V + \frac{\rho^2\gamma_{k+1} \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}}\right] \end{aligned} \tag{41}$$

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,

$$\begin{aligned} &\gamma_{k+1}\rho^2 \, \mathrm{L}_V + \frac{\rho^2\gamma_{k+1} \, \mathrm{L}_V \, \mathrm{L}_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2\rho^2 \, \mathrm{L}_{\mathbf{s}}^2} \\ &\leq \frac{1}{k\alpha^2 c_1^2 \overline{L} n^{4/3}} + \frac{\overline{L}(k\alpha^2 c_1^2 n^{4/3})^{-1}\left(\frac{2}{k^2\alpha^2 c_1^2 \overline{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \overline{L}^2 n^{1/3}}\right)}{\frac{1}{n} - \frac{1}{k\alpha n} - \frac{1}{k^2\alpha^2 c_1^2 n^{4/3}}} \\ &= \frac{1}{k\alpha^2 c_1^2 \overline{L} n^{4/3}} + \frac{\overline{L}\left(\frac{2}{k^2\alpha^2 c_1^2 \overline{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \overline{L}^2 n^{1/3}}\right)}{(k\alpha c_1 n^{1/3})(k\alpha - 1)c_1 - 1} \\ &\overset{(a)}{\leq} \frac{1}{k\alpha^2 c_1^2 \overline{L} n^{4/3}} + \frac{\frac{1}{k\alpha c_1^2 \overline{L} n^{1/3}}\left(\frac{2}{k\alpha n} + 1\right)}{2(\alpha c_1 n^{1/3}) - 1} \\ &\leq \frac{1}{k^2\alpha c_1^2 \overline{L} n^{4/3}} + \frac{1}{4k\alpha^2 c_1^3 \overline{L} n^{2/3}} \\ &\leq \frac{3/4}{\alpha c_1^2 \overline{L} n^{2/3}} \end{aligned} \tag{42}$$

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 (42) on (40) yields:

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

proving the final bound on the gradient of the Lyapunov function:

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

F Practical Implementations of Two-Timescale EM Methods

F.1 Application on GMM

F.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

$$\begin{aligned}
&s_{i,m}^{(1)} = \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\} - \left\{ \log(1 - \textstyle\sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2} \right\} , \\
&s_{i,m}^{(2)} = \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m \ , \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M \ ,
\end{aligned} \tag{43}$$

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

$$z_{i,m} \sim \mathbb{P}\left(z_i = m | y_i; \boldsymbol{\theta}\right) \tag{44}$$

where $m \in \llbracket 1, M \rrbracket$, $i \in \llbracket 1, n \rrbracket$ 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

$$\tilde{S}_{i_k}^{(k)} = \big(\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}) y_{i_k}, ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1}) y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\tilde{\mathbf{s}}_{i_k}^{(3)}(\boldsymbol{\theta}^{(k)})} \big)^\top . \tag{45}$$

Recall that we have used the following regularizer:

$$\mathrm{r}(\boldsymbol{\theta}) = \tfrac{\delta}{2} \textstyle\sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log\left(1 - \sum_{m=1}^{M-1} \omega_m\right) , \tag{46}$$

It can be shown that the regularized M-step in (4) evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1} \big(s_1^{(1)} + \epsilon, \dots, s_{M-1}^{(1)} + \epsilon\big)^\top \\ \big((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \dots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)}\big)^\top \\ \big(1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta\big)^{-1} \big(s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}\big) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{47}$$

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

F.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 H1 For instance, the GMM example satisfies (15) as the sufficient statistics are composed of indicator functions and observations as defined Section F.1 Equation (43).

Assumptions H2 and H3 are standard for the curved exponential family models. For GMM, the following (strongly convex) regularization $\mathrm{r}(\boldsymbol{\theta})$ ensures H3:

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

since it ensures $\boldsymbol{\theta}^{(k)}$ is unique and lies in $\mathrm{int}(\Delta^M) \times \mathbb{R}^M$. We remark that for H2, it is possible to define the Lipschitz constant L_p independently for each data y_i to yield a refined characterization.

Again, H4 is satisfied by practical models. For GMM, it can be verified by deriving the closed form expression for $\mathrm{B}(\mathbf{s})$ and using H1.

Under H1 and H3, 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.

F.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 (45). At iteration k, the several E-steps defined by (9) or (10) and (11) 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 $i \in [\![1, n]\!]$, compute $\overline{\mathbf{s}}_i^{(k)}$ and set

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

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

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

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 (45).

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 (47).

F.2 Deformable Template Model for Image Analysis

F.2.1 Model and Updates

The complete model belongs to the curved exponential family, see [1], which vector of sufficient statistics $S = \big(S_1(z), S_2(z), S_3(z)\big)$ read:

$$
\begin{aligned}
S_1(z) &= \frac{1}{n}\sum_{i=1}^{n} S_1(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top y_i \\
S_2(z) &= \frac{1}{n}\sum_{i=1}^{n} S_2(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{K}_p^{z_i}\right)^\top \left(\mathbf{K}_p^{z_i}\right) \\
S_3(z) &= \frac{1}{n}\sum_{i=1}^{n} S_3(y_i, z_i) = \frac{1}{n}\sum_{i=1}^{n} z_i^t z_i
\end{aligned}
\tag{48}
$$

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

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

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

$$
\bar{\boldsymbol{\theta}}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n}\hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z)\beta(\hat{s}) - 2\beta(\hat{s})\hat{s}_1(z) \end{pmatrix}
\tag{49}
$$

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 (53).

F.2.2 Numerical Applications

For the inference of the template, we use the Matlab code (online SAEM) used in [16] 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 0.08 and 0.16. 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 [6]. Refer to [16] for more details.

G 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 pharmacokinetic 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 data 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:

$$
y_{ij} = f(t_{ij}, z_i) + \varepsilon_{ij} \quad \text{where} \quad f(t_{ij}, z_i) = \frac{D\, ka_i}{V(ka_i - k_i)}(\mathrm{e}^{-ka_i\,(t_{ij}-T_i^{\text{lag}})} - \mathrm{e}^{-k_i\,(t_{ij}-T_i^{\text{lag}})})\ , \tag{50}
$$

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:

$$\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega_{T^{\text{lag}}}^2)\ , \log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2)\ , \tag{51}$$

$$\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2)\ , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2)\ . \tag{52}$$

We recall that the complete model (y, z) defined by (50) belongs to the curved exponential family, which vector of sufficient statistics $S = \big(S_1(z), S_2(z), S_3(z)\big)$ read:

$$S_1(z) = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad S_2(z) = \frac{1}{n}\sum_{i=1}^{n} z_i^\top z_i, \quad S_3(z) = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(t_i, z_i))^2 \tag{53}$$

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:

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} \hat{\mathbf{s}}_1^{(k+1)} \\ \hat{\mathbf{s}}_2^{(k+1)} - \hat{\mathbf{s}}_1^{(k+1)} \left(\hat{\mathbf{s}}_1^{(k+1)}\right)^\top \\ \hat{\mathbf{s}}_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{z}_{\mathbf{pop}}}(\hat{\mathbf{s}}^{(k+1)}) \\ \overline{\boldsymbol{\omega}_{\boldsymbol{z}}}(\hat{\mathbf{s}}^{(k+1)}) \\ \overline{\boldsymbol{\sigma}}(\hat{\mathbf{s}}^{(k+1)}) \end{pmatrix} . \tag{54}$$

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 [19] 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

1: **Input:** initialization $z_{i,0} \sim q(z_i; \boldsymbol{\delta})$
2: **for** $m = 1, \cdots, M$ **do**
3:     Sample $z_{i,m} \sim q(z_i; \boldsymbol{\delta})$
4:     Sample $u \sim \mathcal{U}(\llbracket 0, 1 \rrbracket)$
5:     Calculate the ratio $r = \frac{\pi(z_{i,m};\boldsymbol{\theta})/q(z_{i,m});\boldsymbol{\delta})}{\pi(z_{i,m-1};\boldsymbol{\theta})/q(z_{i,m-1});\boldsymbol{\delta})}$
6:     **if** $u < r$ **then**
7:         Accept $z_{i,m}$
8:     **else**
9:         $z_{i,m} \leftarrow z_{i,m-1}$
10:     **end if**
11: **end for**
12: **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 (50). 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.

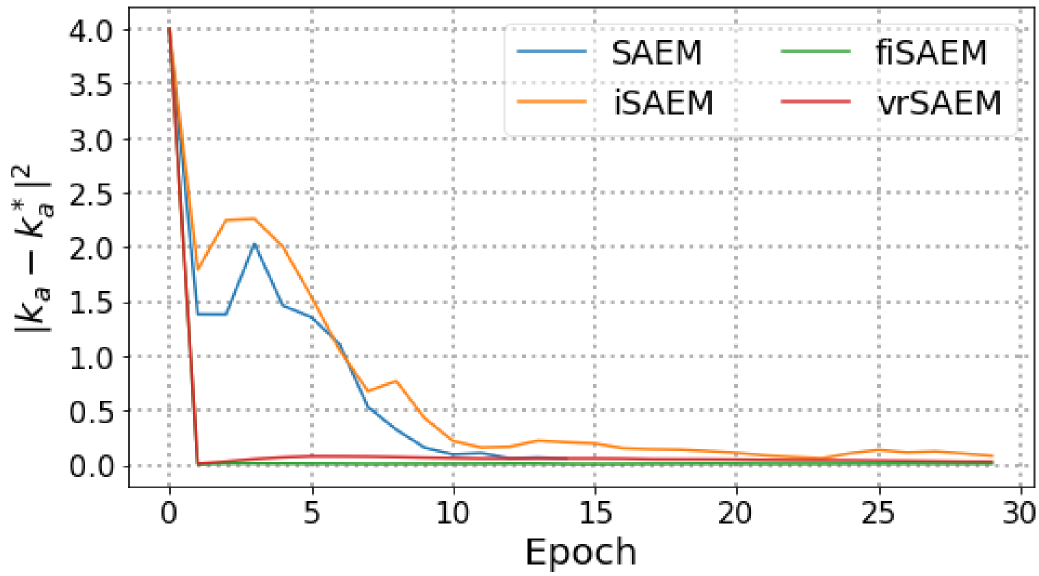

Figure 4: Precision $|ka^{(k)} - ka^*|^2$ per epoch