
Fast Two-Time-Scale Noisy EM Algorithms

Anonymous Author(s)

Affiliation

Address

email

Abstract

Training latent data models using the EM algorithm is the most popular choice for current learning tasks. For today's modern and complex tasks, variants of the EM have been initially introduced by [Neal and Hinton, 1998], using incremental updates to scale to large dataset, and by [Wei and Tanner, 1990, Delyon et al., 1999], using Monte-Carlo (MC) approximations to bypass the impossible conditional expectation of the latent data for most nonconvex models. In this paper, we propose to combine both techniques in a single class of methods called Two-Time-Scale EM Methods. We motivate the choice of a double dynamics by invoking the variance reduction virtue of each stage of the method on both noises: the incremental update and the MC approximation. We establish finite-time and independent of the initialization convergence bounds for nonconvex objective function. 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 [McLachlan and Krishnan, 2007] for references. We formulate the training of such model as the following empirical risk minimization problem:

$$\min_{\theta \in \Theta} \bar{L}(\theta) := r(\theta) + L(\theta) \quad \text{with} \quad L(\theta) = \frac{1}{n} \sum_{i=1}^n L_i(\theta) := \frac{1}{n} \sum_{i=1}^n \{ -\log g(y_i; \theta) \}, \quad (1)$$

We denote the observations by $\{y_i\}_{i=1}^n$, $\Theta \subset \mathbb{R}^d$ is the convex parameters space. We consider a regularized model where $r : \Theta \rightarrow \mathbb{R}$ is a smooth convex regularization function and for $\theta \in \Theta$, $g(y; \theta)$ is the (incomplete) likelihood of each individual observation. The objective function $\bar{L}(\theta)$ is possibly *nonconvex* and is assumed to be lower bounded $\bar{L}(\theta) > -\infty$ for all $\theta \in \Theta$.

In the latent variable model, $g(y_i; \theta)$, is the marginal of the complete data likelihood defined as $f(z_i, y_i; \theta)$, i.e. $g(y_i; \theta) = \int_{\mathcal{Z}} f(z_i, y_i; \theta) \mu(dz_i)$, where $\{z_i\}_{i=1}^n$ are the (unobserved) latent variables. In this paper, we make the assumption of a complete model belonging to the curved exponential family, i.e.,

$$f(z_i, y_i; \theta) = h(z_i, y_i) \exp(\langle S(z_i, y_i) | \phi(\theta) \rangle - \psi(\theta)), \quad (2)$$

where $\psi(\theta)$, $h(z_i, y_i)$ are scalar functions, $\phi(\theta) \in \mathbb{R}^k$ is a vector function, and $S(z_i, y_i) \in \mathbb{R}^k$ is the complete data sufficient statistics.

Full batch EM [Dempster et al., 1977] 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,

$$\bar{s}(\theta) = \frac{1}{n} \sum_{i=1}^n \bar{s}_i(\theta) \quad \text{where} \quad \bar{s}_i(\theta) = \int_{\mathcal{Z}} S(z_i, y_i) p(z_i | y_i; \theta) \mu(dz_i). \quad (3)$$

30 The M-step is given by

$$\text{M-step: } \hat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}(\bar{\mathbf{s}}(\boldsymbol{\theta})) := \arg \min_{\boldsymbol{\vartheta} \in \Theta} \{r(\boldsymbol{\vartheta}) + \psi(\boldsymbol{\vartheta}) - \langle \bar{\mathbf{s}}(\boldsymbol{\theta}) | \phi(\boldsymbol{\vartheta}) \rangle\}, \quad (4)$$

31 Two caveats of this method are the following: (a) with the explosion of data, the first step of the EM
 32 is computationally inefficient as it requires a full pass over the dataset at each iteration and (b) the
 33 complexity of modern models makes the expectation intractable. So far, both challenges have been
 34 addressed separately, to the best of our knowledge, and we give an overview of current solutions in
 35 the sequel.

36 **Prior Work** Inspired by stochastic optimization procedures, [Neal and Hinton, 1998] and [Cappé
 37 and Moulines, 2009] developed respectively an incremental and an online variant of the E-step in
 38 models where the expectation is computable then extensively used and studied in [Nguyen et al.,
 39 2020, Liang and Klein, 2009, Cappé, 2011]. Some improvements of that methods have been pro-
 40 vided and analyzed, globally and in finite-time, in [Karimi et al., 2019] where variance reduction
 41 techniques taken from the optimization literature have been efficiently applied to scale the EM algo-
 42 rithm to large datasets.

43 Regarding the computation of the expectation under the posterior distribution, the first method was
 44 the Monte-Carlo EM (MCEM) introduced in the seminal paper [Wei and Tanner, 1990] where a MC
 45 approximation of this expectation is computed. A variant of that method is the Stochastic Approx-
 46 imation of the EM (SAEM) in [Delyon et al., 1999] leveraging the power of Robbins-Monro type of
 47 update [Robbins and Monro, 1951] to ensure pointwise convergence of the vector of estimated pa-
 48 rameters rather using a decreasing stepsize than increasing the number of MC samples. The MCEM
 49 and the SAEM have been successfully applied in mixed effects models [McCulloch, 1997, Hughes,
 50 1999, Baey et al., 2016] or to do inference for joint modelling of time to event data coming from
 51 clinical trials in [Chakraborty and Das, 2010], among other applications.

52 Recently, an incremental variant of the SAEM was proposed in [Kuhn et al., 2019] showing positive
 53 empirical results but its analysis is limited to asymptotic consideration. Gradient-based methods
 54 have been developed and analyzed in [Zhu et al., 2017] but they remain out of the scope of this
 55 paper as they tackle the high-dimensionality issue.

56 **Contributions** This paper *introduces* and *analyzes* a new class of methods which purpose is to
 57 combine both solutions proposed in the past years in a two-time-scale manner in order to optimize
 58 (1) for current modern examples and settings. The main contributions of the paper are:

- 59 • We propose a two-time-scale method based on Stochastic Approximation (SA), to alleviate
 60 the problem of MC computation, and on Incremental updates, to scale to large datasets.
 61 We describe in details the edges of each level of our method based on variance reduc-
 62 tion arguments. The derivation of such class of algorithms has two advantages. First, it
 63 combines two powerful ideas, commonly used separately, to tackle large scale and highly
 64 nonlinear learning tasks. Then, it gives a simple formulation as a *scaled-gradient method*,
 65 as introduced in [Karimi et al., 2019], which makes the global analysis accessible.
- 66 • We also establish global (independent of the initialization) and finite-time (true at each
 67 iteration) upper bounds on a classical suboptimality condition in the nonconvex literature,
 68 *i.e.*, the second order moment of the gradient of the objective function.

69 In Section 2 we give rigorous mathematical definitions of the various updates used for both incre-
 70 mental and Monte-Carlo EMs and we introduce the main class of new algorithms, based on two
 71 different dynamics, we are proposing to analyze and compare to baselines algorithms. Section 3
 72 presents the main theoretical guarantees of this newly introduced two-time-scale class of algorithms.
 73 Results are given both in finite-time and in the nonconvex setting. Finally, we illustrate the advan-
 74 tages of our method in Section 4 on two numerical experiments.

75 2 Two-Time-Scale Stochastic EM Algorithms

76 We recall and formalize in this section the different methods found in the literature that aim to solv-
 77 ing the large scale problem and the intractable expectation. We then provide the general framework
 78 of our method to efficiently tackle the optimization problem (1).

79 2.1 Monte Carlo Integration and Stochastic Approximation

80 As mentioned in the introduction, for complex and possibly nonlinear models, the expectation under
 81 the posterior distribution defined in (3) is not tractable. In that case, the first solution involves
 82 computing a Monte Carlo integration of that latter term. For all $i \in \llbracket 1, n \rrbracket$, draw for $m \in \llbracket 1, M \rrbracket$,
 83 samples $z_{i,m} \sim p(z_i|y_i; \theta)$ and compute the MC integration \tilde{s} of the deterministic quantity $\bar{s}(\theta)$:

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

84 and then update the parameter $\hat{\theta} = \bar{\theta}(\tilde{s})$. This algorithm bypasses the intractable expectation issue
 85 but is rather computationally expensive in order to reach point wise convergence (M needs to be
 86 large). An alternative to that stochastic algorithm is to use a Robbins-Monro (RM) type of update.
 87 We denote, at iteration k , the following quantity

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

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

$$\text{SA-step : } \hat{s}^{(k+1)} = \hat{s}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{s}^{(k)}) \quad (7)$$

89 where $\{\gamma_k\}_{k \geq 1} \in (0, 1)$ is a sequence of decreasing step sizes to ensure asymptotic convergence.
 90 This is called the Stochastic Approximation of the EM (SAEM) and has been shown theoretically to
 91 converge to a maximum of the likelihood of the observations under very general conditions [Delyon
 92 et al., 1999]. In the simulation step (6), since the relation between the observed data y_i and the
 93 latent variable z_i can be non linear, sampling from the posterior distribution $p(z_i|y_i; \theta)$, under the
 94 current model θ , could require using an inference algorithm. [?] proved almost sure convergence
 95 of the sequence of parameters obtained by this algorithm coupled with an MCMC procedure during
 96 the simulation step. In simple scenarios, the samples $\{z_{i,m}\}_{m=0}^{M-1}$ are conditionally independent and
 97 identically distributed with distribution $p(z_i, \theta)$. Nevertheless, in most cases, sampling exactly from
 98 this distribution is not an option and the Monte Carlo batch is sampled by Monte Carlo Markov
 99 Chains (MCMC) algorithm. In the SA-step, the sequence of decreasing positive integers $\{\gamma_k\}_{k \geq 1}$
 100 controls the convergence of the algorithm. In practice, γ_k is set equal to 1 during the first few
 101 iterations to let the algorithm explore the parameter space without memory and converge quickly
 102 to a neighbourhood of the target estimate. The Stochastic Approximation is performed during the
 103 remaining iterations where $\gamma_k = 1/k^\alpha$, where $\alpha \in (0, 1)$, ensuring the almost sure convergence of
 104 the estimate. It is inappropriate to start with small values for step size γ_k and large values for the
 105 number of simulations M_k . Rather, it is recommended that one decrease γ_k and keep a constant
 106 and small number of MC samples M_k which shows a great advantage over the MC-step (5), which
 107 requires large M_k to converge.

108 This Robbins-Monro type of update represents the *first level* of our algorithm, needed to temper
 109 the variance and noise implied by MC integration. In the next section, we derive variants of this
 110 algorithm to adapt to the sheer size of data of today's applications and formalize the *second level* of
 111 our class of Two-Time-Scale EM methods.

112 2.2 Incremental and Bi-Level Inexact EM Methods

113 Strategies to scale to large datasets include classical incremental and variance reduced variants. We
 114 will explicit a general update that will cover those variants and that represents the *second level* of our
 115 algorithm, namely the incremental update of the noisy statistics $\hat{S}^{(k)}$ inside the RM type of update.

$$\text{Incremental-step : } \tilde{S}^{(k+1)} = \tilde{S}^{(k)} + \rho_{k+1}(\mathcal{S}^{(k+1)} - \tilde{S}^{(k)}), \quad (8)$$

116 Note $\{\rho_k\}_{k \geq 1} \in (0, 1)$ is a sequence of step sizes, $\mathcal{S}^{(k)}$ is a proxy for $\tilde{S}^{(k)}$, If the stepsize is equal
 117 to one and the proxy $\mathcal{S}^{(k)} = \hat{S}^{(k)}$, i.e., computed in a full batch manner as in (6), then we recover
 118 the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\mathcal{S}^{(k)} = \tilde{S}^{(k)}$, then we recover the Monte Carlo
 119 EM algorithm.

120 We now introduce three variants of the SAEM update depending on different definitions of the
 121 proxy $\mathcal{S}^{(k)}$ and the choice of the stepsize ρ_k . Let $i_k \in \llbracket 1, n \rrbracket$ be a random index drawn at iteration
 122 k and $\tau_i^k = \max\{k' : i_{k'} = i, k' < k\}$ be the iteration index where $i \in \llbracket 1, n \rrbracket$ is last drawn
 123 prior to iteration k . For iteration $k \geq 0$, the fiTTSEM method draws *two* indices *independently* and
 124 uniformly as $i_k, j_k \in \llbracket 1, n \rrbracket$. In addition to τ_i^k which was defined w.r.t. i_k , we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample $j \in \llbracket 1, n \rrbracket$ is last drawn as j_k prior to
 125 iteration k . With the initialization $\bar{\mathcal{S}}^{(0)} = \bar{s}^{(0)}$, we use a slightly different update rule from SAGA
 126 inspired by [Reddi et al., 2016]. Then, we obtain:

$$(iSAEM [Karimi, 2019, Kuhn et al., 2019]) \quad \mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} + \frac{1}{n} (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)}) \quad (9)$$

$$(vrTTSEM This paper) \quad \mathcal{S}^{(k+1)} = \tilde{S}^{(\ell(k))} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}) \quad (10)$$

$$(fiTTSEM This paper) \quad \mathcal{S}^{(k+1)} = \bar{\mathcal{S}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) \quad (11)$$

$$\bar{\mathcal{S}}^{(k+1)} = \bar{\mathcal{S}}^{(k)} + n^{-1} (\tilde{S}_{j_k}^{(k)} - \tilde{S}_{j_k}^{(t_{j_k}^k)}). \quad (12)$$

128 The stepsize is set to $\rho_{k+1} = 1$ for the iSAEM method; $\rho_{k+1} = \gamma$ is constant for the vrTTSEM and
 129 fiTTSEM methods. Moreover, for iSAEM we initialize with $\mathcal{S}^{(0)} = \tilde{S}^{(0)}$; for vrTTSEM we set an
 130 epoch size of m and define $\ell(k) := m \lfloor k/m \rfloor$ as the first iteration number in the epoch that iteration
 131 k is in.

132 2.3 Two-Time-Scale Noisy EM methods

133 We now introduce the general method derived using the two variance reduction techniques described
 134 above. Algorithm 1 leverages both levels (7) and (8) in order to output a vector of fitted parameters
 135 $\hat{\theta}^{(K)}$ where K is some randomly chosen termination point.

136 The update in (14) is said to have two timescales as the step sizes satisfy $\lim_{k \rightarrow \infty} \gamma_k / \rho_k < 1$ such that
 137 $\tilde{S}^{(k+1)}$ is updated at a faster timescale than $\hat{s}^{(k+1)}$.

Algorithm 1 Two-Time-Scale Noisy EM methods.

- 1: **Input:** initializations $\hat{\theta}^{(0)} \leftarrow 0, \hat{s}^{(0)} \leftarrow \tilde{S}^{(0)}, K_{\max} \leftarrow \text{max. iteration number}$.
- 2: Set the terminating iteration number, $K \in \{0, \dots, K_{\max} - 1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\max}-1} \gamma_\ell}. \quad (13)$$

- 3: **for** $k = 0, 1, 2, \dots, K$ **do**
- 4: Draw index $i_k \in \llbracket 1, n \rrbracket$ uniformly (and $j_k \in \llbracket 1, n \rrbracket$ for fiTTSEM).
- 5: Compute $\hat{S}_{i_k}^{(k)}$ using the MC-step (5), for the drawn indices.
- 6: Compute the surrogate sufficient statistics $\mathcal{S}^{(k+1)}$ using (9) or (10) or (11).
- 7: Compute $\hat{S}^{(k+1)}$ and $\hat{s}^{(k+1)}$ using respectively (8) and (7):

$$\begin{aligned} \tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho_{k+1} (\mathcal{S}^{(k+1)} - \tilde{S}^{(k)}) \\ \hat{s}^{(k+1)} &= \hat{s}^{(k)} + \gamma_{k+1} (\tilde{S}^{(k+1)} - \hat{s}^{(k)}) \end{aligned} \quad (14)$$

- 8: Compute $\hat{\theta}^{(k+1)}$ via the M-step (4).
 - 9: **end for**
 - 10: **Return:** $\hat{\theta}^{(K)}$.
-

138 The next section presents the main results of this paper and establishes global and finite-time bounds
 139 for the three different updates of our two-time-scale scheme..

3 Finite Time Analysis of the Two-Time-Scale Scheme

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

$$\min_{\mathbf{s} \in \mathcal{S}} V(\mathbf{s}) := \bar{L}(\bar{\boldsymbol{\theta}}(\mathbf{s})) = r(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(\bar{\boldsymbol{\theta}}(\mathbf{s})) \quad (15)$$

We thus propose to study the latter minimization problem in the sequel.

An important assumption in order to derive convergence guarantees reads as follows:

H1. *The sets \mathcal{Z}, \mathcal{S} are compact. There exists constants C_S, C_Z such that:*

$$C_S := \max_{\mathbf{s}, \mathbf{s}' \in \mathcal{S}} \|\mathbf{s} - \mathbf{s}'\| < \infty, \quad C_Z := \max_{i \in \llbracket 1, n \rrbracket} \int_{\mathcal{Z}} |S(z, y_i)| \mu(dz) < \infty. \quad (16)$$

H2. *The conditional distribution is smooth on $\text{int}(\Theta)$. For any $i \in \llbracket 1, n \rrbracket$, $z \in \mathcal{Z}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \text{int}(\Theta)^2$, we have $|p(z|y_i; \boldsymbol{\theta}) - p(z|y_i; \boldsymbol{\theta}')| \leq L_p \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.*

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

H3. *For any $\mathbf{s} \in \mathcal{S}$, the function $\boldsymbol{\theta} \mapsto L(\mathbf{s}, \boldsymbol{\theta}) := r(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \langle \mathbf{s} | \phi(\boldsymbol{\theta}) \rangle$ admits a unique global minimum $\bar{\boldsymbol{\theta}}(\mathbf{s}) \in \text{int}(\Theta)$. In addition, $J_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))$ is full rank, L_{ϕ} -Lipschitz and $\bar{\boldsymbol{\theta}}(\mathbf{s})$ is L_{θ} -Lipschitz.*

We denote by $H_L^{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\theta})$ the Hessian (w.r.t to $\boldsymbol{\theta}$ for a given value of \mathbf{s}) of the function $\boldsymbol{\theta} \mapsto L(\mathbf{s}, \boldsymbol{\theta}) = r(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \langle \mathbf{s} | \phi(\boldsymbol{\theta}) \rangle$, and define

$$\mathbf{B}(\mathbf{s}) := J_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \left(H_L^{\boldsymbol{\theta}}(\mathbf{s}, \bar{\boldsymbol{\theta}}(\mathbf{s})) \right)^{-1} J_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^{\top}. \quad (17)$$

H4. *It holds that $v_{\max} := \sup_{\mathbf{s} \in \mathcal{S}} \|\mathbf{B}(\mathbf{s})\| < \infty$ and $0 < v_{\min} := \inf_{\mathbf{s} \in \mathcal{S}} \lambda_{\min}(\mathbf{B}(\mathbf{s}))$. There exists a constant L_B such that for all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^2$, we have $\|\mathbf{B}(\mathbf{s}) - \mathbf{B}(\mathbf{s}')\| \leq L_B \|\mathbf{s} - \mathbf{s}'\|$.*

We now formulate the main difference with the work done in [Karimi et al., 2019]. The class of algorithms we develop in this paper are two time-scale where the first stage corresponds to the variance reduction trick used in [Karimi et al., 2019] 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

Indeed the expectations (3) are never available and requires Monte Carlo approximation. Thus, at iteration $k + 1$, we introduce the errors when approximating the quantity $\bar{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k-1)}))$. For all $i \in \llbracket 1, n \rrbracket$, $r > 0$ and $\vartheta \in \Theta$, define:

$$\eta_i^{(r)} := \tilde{S}_i^{(r)} - \bar{\mathbf{s}}_i(\vartheta^{(r)}) \quad (18)$$

For instance, we consider that the MC approximation is unbiased if for all $i \in \llbracket 1, n \rrbracket$ and $m \in \llbracket 1, M \rrbracket$, the samples $z_{i,m} \sim p(z_i|y_i; \boldsymbol{\theta})$ are i.i.d. under the posterior distribution, i.e., $\mathbb{E}[\eta_i^{(r)} | \mathcal{F}_r] = 0$ 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 $k > 0$, $i \in \llbracket 1, n \rrbracket$ and $\vartheta \in \Theta$:*

$$\mathbb{E} \left[\left\| \eta_i^{(r)} \right\|^2 \right] \leq \frac{C_{\eta}}{M_r} \quad \text{and} \quad \mathbb{E} \left[\left\| \mathbb{E}[\eta_i^{(r)} | \mathcal{F}_r] \right\|^2 \right] \leq \frac{C}{M_r} \quad (19)$$

In that setting, we can prove two important results on the Lyapunov function. The first one suggests smoothness:

Lemma 1. [Karimi et al., 2019] *Assume H1-H4. For all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ and $i \in \llbracket 1, n \rrbracket$, we have*

$$\|\bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq L_S \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq L_V \|\mathbf{s} - \mathbf{s}'\|, \quad (20)$$

where $L_S := C_Z L_p L_{\theta}$ and $L_V := v_{\max}(1 + L_S) + L_B C_S$.

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 $\mathbf{s} \in \mathcal{S}$,

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \quad (21)$$

See proofs of this Lemma in Appendix A.

3.1 Global Convergence of Incremental Noisy 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 first intermediate result is the computation of the quantity $\hat{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}$, which corresponds to the drift term of (7) and reads as follows:

Lemma 3. The update (9) is equivalent to the following update on the resulting statistics

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1} (\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \quad \text{where} \quad \tilde{\mathbf{S}}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(\tau_i^{k+1})} \quad (22)$$

Also:

$$\mathbb{E} [\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E} [\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \left(1 - \frac{1}{n}\right) \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right] + \frac{1}{n} \mathbb{E} [\eta_{i_k}^{(k+1)}] \quad (23)$$

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

See proofs of this Lemma in Appendix B.

The following main result for the iSAEM algorithm is derived under a control of the Monte Carlo fluctuations as described by assumption H 5. Typically, the controls exhibited below are of interest when the number of MC samples M_k increase with the iteration index f .

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 $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = 1$ for any $k > 0$. We also set $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$, $\beta = \frac{c_1 \bar{L}}{n}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.

$$v_{\max}^{-2} \sum_{k=0}^{K_{\max}} \tilde{\alpha}_k \mathbb{E} \left[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2 \right] \leq \mathbb{E} [V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K)})] + \sum_{k=0}^{K_{\max}-1} \tilde{\Gamma}_k \mathbb{E} \left[\|\eta_{i_k}^{(k)}\|^2 \right] \quad (24)$$

See proof in Appendix C.

3.2 Global Convergence of Two-Time-Scale Noisy EM Algorithms

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

Lemma 4. For any $k \geq 0$ and consider the vrTTSEM update in (10) with $\rho_k = \rho$, it holds for all $k > 0$

$$\begin{aligned} \mathbb{E} \left[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2 \right] &\leq 2\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 L_s^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2] \\ &\quad + 2(1 - \rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(\ell(k))} - \tilde{\mathbf{S}}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \end{aligned} \quad (25)$$

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

204 The second one derives an identity for the quantity $\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2]$ using the fTTSEM update:
 205 **Lemma 5.** For any $k \geq 0$ and consider the fTTSEM update in (11) with $\rho_k = \rho$, it holds for all
 206 $k > 0$

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \right\|^2 \right] &\leq 2\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 \frac{L_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &\quad + 2(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \end{aligned} \quad (26)$$

207 We now state the main result regarding the vrTTSEM method.

208 **Theorem 2.** Assume H1-H5. Let K_{\max} be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of
 209 positive step sizes and consider the vrTTSEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = \rho$ for
 210 any $k > 0$.

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

$$\begin{aligned} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \frac{2n^{2/3} \bar{L}}{\mu v_{\min}^2 v_{\max}^2} \mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})] \\ &\quad + \frac{2n^{2/3} \bar{L}}{\mu v_{\min}^2 v_{\max}^2} \sum_{k=0}^{K_{\max}-1} \left[\tilde{\eta}^{(k+1)} + \chi^{(k+1)} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right] \right] \end{aligned} \quad (27)$$

213 See proof in Appendix E. We now state the main result regarding the fTTSEM method.

214 **Theorem 3.** Assume H1-H5. Let K_{\max} be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of
 215 positive step sizes and consider the fTTSEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = \rho$ for
 216 any $k > 0$.

217 Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$. By setting $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$,
 218 $\beta = \frac{c_1 \bar{L}}{n}$, $\rho = \frac{1}{n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$, we have the following bound:

$$\begin{aligned} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \frac{\alpha \bar{L} n^{2/3}}{v_{\min} v_{\max}^2} [V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})] \\ &\quad + \frac{\alpha \bar{L} n^{2/3}}{v_{\min} v_{\max}^2} \sum_{k=0}^{K_{\max}-1} \left[\Xi^{(k+1)} + \Gamma_{k+1} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right] \right] \end{aligned} \quad (28)$$

219 See proof in Appendix F. Note that in those two bounds, the quantities $\tilde{\eta}^{(k+1)}$ and $\Xi^{(k+1)}$ are
 220 abstraction that depends only on the MC fluctuations $\mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right]$ and some constants.

221 **Remarks:** The following remarks are worth noting on the quantity $\mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right]$:

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

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

224 which strengthen the fact that this quantity characterizes the impact of the variance reduc-
 225 tion technique introduced in our two stages class of methods.

226 The following lemma, which proof can be found in Appendix D.2, can be derived to characterize
 227 this gap:

228 **Lemma 6.** Consider a decreasing stepsize $\gamma_k \in (0, 1)$ and a constant $\rho \in (0, 1)$, then the following
 229 inequality holds:

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\|^2] \leq \frac{\rho}{1-\rho} \sum_{\ell=0}^k (1-\gamma_\ell)^2 (\mathbf{S}^{(\ell)} - \tilde{\mathbf{S}}^{(\ell)}) \quad (29)$$

230 where $\mathbf{S}^{(k)}$ is defined either by (10) (vrTTSEM) or (11) (fiTTSEM).

231 In the next section, we illustrate the benefits of our two-time-scale class of methods on several
 232 numerical applications.

233 4 Numerical Examples

234 4.1 Gaussian Mixture Models

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

$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) [\log(\omega_m) - \mu_m^2/2] + \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) \mu_m y_i + \text{constant} . \quad (30)$$

241 where $\boldsymbol{\theta} := (\boldsymbol{\omega}, \boldsymbol{\mu})$ with $\boldsymbol{\omega} = \{\omega_m\}_{m=1}^{M-1}$ are the mixing weights with the convention $\omega_M =$
 242 $1 - \sum_{m=1}^{M-1} \omega_m$ and $\boldsymbol{\mu} = \{\mu_m\}_{m=1}^M$ are the means. We use the penalization $r(\boldsymbol{\theta}) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 -$
 243 $\log \text{Dir}(\boldsymbol{\omega}; M, \epsilon)$ where $\delta > 0$ and $\text{Dir}(\cdot; M, \epsilon)$ is the M dimensional symmetric Dirichlet distribu-
 244 tion with concentration parameter $\epsilon > 0$. The constraint set on $\boldsymbol{\theta}$ is given by

$$\Theta = \{\omega_m, m = 1, \dots, M-1 : \omega_m \geq 0, \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R}, m = 1, \dots, M\}. \quad (31)$$

245 Exact two time scale updates are given in Appendix G.1.

246 In the following experiments on synthetic data, we generate samples from a GMM model with
 247 $M = 2$ components with two mixtures with means $\mu_1 = -\mu_2 = 0.5$. We use $n = 10^5$
 248 synthetic samples and run the bEM method until convergence (to double precision) to obtain the
 249 ML estimate μ^* averaged on 50 datasets. We compare the bEM, iEM (incremental EM), SAEM,
 250 iSAEM, vrTTSEM and fiTTSEM methods in terms of their precision measured by $|\mu - \mu^*|^2$. We
 251 set the stepsize of the SA-step of all method as $\gamma_k = 1/k^\alpha$ with $\alpha = 0.5$, and the stepsizes
 252 of the Incremental-step for vrTTSEM and the fiTTSEM to a constant stepsize equal to $1/n^{2/3}$.
 253

254 The number of MC samples is fixed to $M = 10$
 255 chains. Figure 1 shows the convergence of the
 256 precision $|\mu - \mu^*|^2$ for the different methods
 257 against the epoch(s) elapsed (one epoch equals
 258 n iterations). We observe that the vrTTSEM
 259 and fiTTSEM methods outperform the other
 260 stochastic methods, supporting the benefits of
 261 our newly introduced scheme.

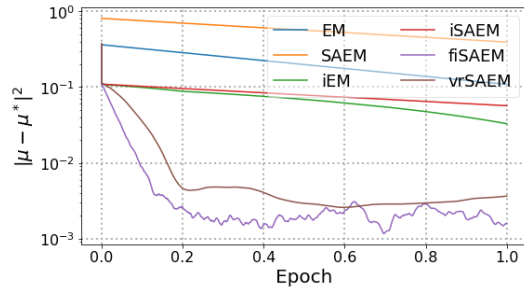


Figure 1: TO COMPLETE

262 4.2 Deformable Template Model for Image 263 Analysis

264 Let $(y_i, i \in \llbracket 1, n \rrbracket)$ be observed gray level images defined on a grid of pixels. Let $u \in \mathcal{U} \subset \mathbb{R}^2$
 265 denotes the pixel index on the image and $x_u \in \mathcal{D} \subset \mathbb{R}^2$ its location. The model used in this
 266 experiment suggests that each image y_i is a deformation of a template, noted $I : \mathcal{D} \rightarrow \mathbb{R}$, common
 267 to all images of the dataset:

$$y_i(u) = I(x_u - \Phi_i(x_u, z_i)) + \varepsilon_i(u) \quad (32)$$

where $\phi_i : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a deformation function, z_i some latent variable parametrizing this deformation and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an observation error.

The template model, given $(p_k, k \in \llbracket 1, k_p \rrbracket)$ landmarks on the template, a fixed known kernel \mathbf{K}_p and a vector of parameters $\beta \in \mathbb{R}^{k_p}$ is defined as follows:

$$I_\xi = \mathbf{K}_p \beta, \quad \text{where} \quad (\mathbf{K}_p \beta)(x) = \sum_{k=1}^{k_p} \mathbf{K}_p(x, p_k) \beta_k \quad (33)$$

Besides, we parameterize the deformation model given some landmarks $(g_k, k \in \llbracket 1, k_g \rrbracket)$ and a fixed kernel \mathbf{K}_g as:

$$\Phi_i = \mathbf{K}_g z_i \quad \text{where} \quad (\mathbf{K}_g z_i)(x) = \sum_{k=1}^{k_g} \mathbf{K}_g(x, g_k) \left(z_i^{(1)}(k), z_i^{(2)}(k) \right) \quad (34)$$

where we put a Gaussian prior on the latent variables, $z_i \sim \mathcal{N}(0, \Gamma)$ and $z_i \in (\mathbb{R}^{k_g})^2$. The vector of parameters we ought to estimate is thus $\theta = (\beta, \Gamma, \sigma)$. The complete model belongs to the curved exponential family, see [Allasonnière et al., 2007], which vector of sufficient statistics $S = (S_1(z), S_2(z), S_3(z))$ read:

$$\begin{aligned} S_1(z) &= \frac{1}{n} \sum_{i=1}^n S_1(y_i, z_i) = \frac{1}{n} \sum_{i=1}^n (\mathbf{K}_p^{z_i})^\top y_i \\ S_2(z) &= \frac{1}{n} \sum_{i=1}^n S_2(y_i, z_i) = \frac{1}{n} \sum_{i=1}^n (\mathbf{K}_p^{z_i})^\top (\mathbf{K}_p^{z_i}) \\ S_3(z) &= \frac{1}{n} \sum_{i=1}^n S_3(y_i, z_i) = \frac{1}{n} \sum_{i=1}^n z_i^t z_i \end{aligned} \quad (35)$$

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

$$\mathbf{K}_p^{z_i}(x_u, j) = \mathbf{K}_p^{z_i}(x_u - \phi_i(x_u, z_i), p_j) \quad (36)$$

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

$$\bar{\theta}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z) \hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n} \hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z) \beta(\hat{s}) - 2\beta(\hat{s}) \hat{s}_1(z) \end{pmatrix} \quad (37)$$

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 $(S_1(z), S_2(z), S_3(z))$ defined in (42).

Comparison using epochs credit

Comparison using number of training samples credit

4.3 PK Model with Absorption Lag Time

This numerical example was conducted in order to characterize the pharmacokinetics of orally administered drug to simulated patients, using a population pharmacokinetic approach. $M = 50$ synthetic datasets were generated for $n = 500$ patients with 10 observations (concentration measures) per patient.

The model: We consider a one-compartment pharmacokinetics (PK) model for oral administration with an absorption lag-time (T^{lag}), assuming first-order absorption and linear elimination processes. The final model includes ka is the absorption rate constant, V the volume of distribution, k the elimination rate constant. 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 such as V is function of it.

294 More precisely, the log-volume $\log(V)$ is a linear function of the log-weight $lw70 = \log(wt/70)$.
 295 The final reads:

$$f(t, ka, V, k) = \frac{D ka}{V(ka - k)} (e^{-ka(t-T^{\text{lag}})} - e^{-k(t-T^{\text{lag}})}) , \quad (38)$$

296 Here, T^{lag} , ka , V and k are PK parameters that can change from one individual to another.

297 Let $z_i = (T_i^{\text{lag}}, ka_i, V_i, k_i)$ be the vector of individual PK parameters for individual i . The model
 298 for the j -th measured concentration, noted y_{ij} , for individual i writes:

$$y_{ij} = f(t_{ij}, z_i) + \varepsilon_{ij} . \quad (39)$$

299 We assume in this example that the residual errors are independent and normally distributed with
 300 mean 0 and variance σ^2 . Lognormal distributions are used for the three PK parameters:

$$\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega_{T^{\text{lag}}}^2) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2) , \quad (40)$$

$$\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2) , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2) . \quad (41)$$

301 The complete model belongs to the curved exponential family, which vector of sufficient statistics
 302 $S = (S_1(z), S_2(z), S_3(z))$ read:

$$S_1(z) = \frac{1}{n} \sum_{i=1}^n z_i, \quad S_2(z) = \frac{1}{n} \sum_{i=1}^n z_i^\top z_i, \quad S_3(z) = \frac{1}{n} \sum_{i=1}^n (y_i - f(t_i, z_i))^2 \quad (42)$$

303 where we have noted y_i and t_i the vector of observations and time for each patient i .

304 **Monte Carlo study:** We conduct a Monte
 305 Carlo study to showcase the benefits of our
 306 scheme.

307 $M = 50$ datasets have been simulated using
 308 the following PK parameters values: $T_{\text{pop}}^{\text{lag}} =$
 309 1 , $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$,
 310 $\omega_{T^{\text{lag}}} = 0.4$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k =$
 311 0.3 and $\sigma^2 = 0.5$. We defined the mean
 312 square distance over the M replicates $E_k(\ell) =$

$$\frac{1}{M} \sum_{m=1}^M \left(\theta_k^{(m)}(\ell) - \theta^* \right)^2 .$$

314 Note that the MC simulation is performed using a Metropolis Hastings procedure since the posterior
 315 distribution under the model θ noted $p(z_i|y_i, \theta)$ is intractable due to the nonlinearity of the model
 316 (38). Figure 2 shows

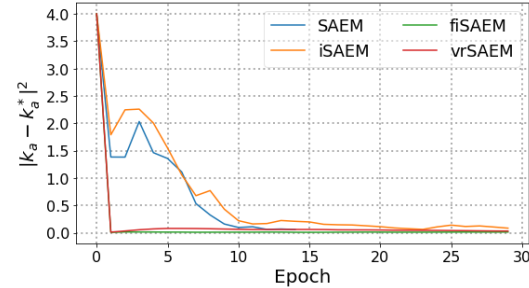


Figure 2: TO COMPLETE

317 5 Conclusion

References

- S. Allasonnière, 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.
- S. Allasonnière, 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.
- 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.
- O. Cappé. Online em algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749, 2011.
- 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.
- 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.
- 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>.
- 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.
- J. P. Hughes. Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55(2):625–629, 1999.
- B. Karimi. *Non-Convex Optimization for Latent Data Models: Algorithms, Analysis and Applications*. PhD thesis, 2019.
- 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.
- E. Kuhn, C. Matias, and T. Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.
- 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.
- C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.
- G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- 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.
- 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.

- 359 S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for nonconvex optimization.
360 *arXiv preprint arXiv:1603.06159*, 2016.
- 361 H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statis-*
362 *tics*, pages 400–407, 1951.
- 363 G. C. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man’s
364 data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704,
365 1990.
- 366 R. Zhu, L. Wang, C. Zhai, and Q. Gu. High-dimensional variance-reduced stochastic gradient
367 expectation-maximization algorithm. In *Proceedings of the 34th International Conference on*
368 *Machine Learning-Volume 70*, pages 4180–4188. JMLR. org, 2017.

A Proof of Lemma 2

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

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \quad (43)$$

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

$$\begin{aligned} \nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \left(\nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{L}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \left(\nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top (\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))), \end{aligned} \quad (44)$$

Consider the following vector map:

$$\mathbf{s} \rightarrow \nabla_{\boldsymbol{\theta}} L(\mathbf{s}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(\mathbf{s})} = \nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{s}. \quad (45)$$

Taking the gradient of the above map w.r.t. \mathbf{s} and using assumption H3, we show that:

$$\mathbf{0} = -\mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \underbrace{\left(\nabla_{\boldsymbol{\theta}}^2 (\psi(\boldsymbol{\theta}) + \mathbf{r}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) | \mathbf{s} \rangle) \right)|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(\mathbf{s})}}_{=\mathbf{H}_L^{\boldsymbol{\theta}}(\mathbf{s}; \boldsymbol{\theta})} \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s}). \quad (46)$$

The above yields

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathbf{B}(\mathbf{s})(\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))) \quad (47)$$

where we recall $\mathbf{B}(\mathbf{s}) = \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \left(\mathbf{H}_L^{\boldsymbol{\theta}}(\mathbf{s}; \bar{\boldsymbol{\theta}}(\mathbf{s})) \right)^{-1} \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top$. The proof of (43) follows directly from the assumption H4. \square

B Proof of Lemma 3

Lemma. Assume H??. The update (9) is equivalent to the following update on the resulting statistics

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1} (\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \quad (48)$$

Also:

$$\mathbb{E} [\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E} [\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \left(1 - \frac{1}{n}\right) \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right] + \frac{1}{n} \mathbb{E} [\eta_{i_k}^{(k+1)}] \quad (49)$$

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

Proof From update (9), we have:

$$\begin{aligned} \tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= \tilde{\mathbf{S}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \frac{1}{n} \left(\tilde{S}_{i_k}^{(k+1)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)} \right) \\ &= \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{\mathbf{S}}^{(k)} - \bar{\mathbf{s}}^{(k)} - \frac{1}{n} \left(\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \tilde{S}_{i_k}^{(k+1)} \right) \end{aligned} \quad (50)$$

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

$$\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{\mathbf{S}}^{(k)} - \bar{\mathbf{s}}^{(k)} - \frac{1}{n} \left(\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \bar{s}_{i_k}(\boldsymbol{\theta}^{(k)}) \right) + \frac{1}{n} \eta_{i_k}^{(k+1)} \quad (51)$$

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

$$\begin{aligned} \mathbb{E} [\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] &= \mathbb{E} [\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right] \\ &\quad - \frac{1}{n} \mathbb{E} \left[\mathbb{E} [\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \bar{s}_{i_k}(\boldsymbol{\theta}^{(k)}) | \mathcal{F}_k] \right] + \frac{1}{n} \mathbb{E} [\eta_{i_k}^{(k+1)}] \end{aligned} \quad (52)$$

The following equalities:

$$\mathbb{E} [\tilde{S}_i^{(\tau_i^k)} | \mathcal{F}_k] = \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} \quad \text{and} \quad \mathbb{E} [\bar{s}_{i_k}(\boldsymbol{\theta}^{(k)}) | \mathcal{F}_k] = \bar{\mathbf{s}}^{(k)} \quad (53)$$

concludes the proof of the Lemma. \square

C Proof of Theorem 1

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 $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = 1$ for any $k > 0$. We also set $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$, $\beta = \frac{c_1 \bar{L}}{n}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.

$$v_{\max}^{-2} \sum_{k=0}^{K_{\max}} \tilde{\alpha}_k \mathbb{E} \left[\left\| \nabla V(\hat{\mathbf{s}}^{(k)}) \right\|^2 \right] \leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K)}) \right] + \sum_{k=0}^{K_{\max}-1} \tilde{\Gamma}_k \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \quad (54)$$

Proof We begin our proof by giving this auxiliary Lemma setting an upper bound for the quantity $\mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right]$

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

$$\begin{aligned} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] &\leq 4\mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] + \frac{2L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \right\|^2 \right] \\ &\quad + 2\frac{C_\eta}{M_k} + 4\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \end{aligned} \quad (55)$$

Proof Applying the iSAEM update yields:

$$\begin{aligned} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] &= \mathbb{E} \left[\left\| \tilde{S}^{(k)} - \hat{\mathbf{s}}^{(k)} - \frac{1}{n} (\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \tilde{S}_{i_k}^{(k)}) \right\|^2 \right] \\ &\leq 4\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] + 4\mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \\ &\quad + \frac{2}{n^2} \mathbb{E} \left[\left\| \bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)} \right\|^2 \right] + 2\frac{C_\eta}{M_k} \end{aligned} \quad (56)$$

The last expectation can be further bounded by

$$\frac{2}{n^2} \mathbb{E} \left[\left\| \bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)} \right\|^2 \right] = \frac{2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \bar{\mathbf{s}}_i^{(k)} - \bar{\mathbf{s}}_i^{(t_i^k)} \right\|^2 \right] \stackrel{(a)}{\leq} \frac{2L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \right\|^2 \right], \quad (57)$$

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

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

$$V(\hat{\mathbf{s}}^{(k+1)}) \leq V(\hat{\mathbf{s}}^{(k)}) + \gamma_{k+1} \langle \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \left\| \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \quad (58)$$

Taking the expectation on both sides yields:

$$\mathbb{E} \left[V(\hat{\mathbf{s}}^{(k+1)}) \right] \leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(k)}) \right] + \gamma_{k+1} \mathbb{E} \left[\langle \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \quad (59)$$

402 Using Lemma 3, we obtain:

$$\begin{aligned}
& \mathbb{E} \left[\langle \tilde{S}^{(k+1)} - \hat{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] = \\
& \mathbb{E} \left[\langle \bar{s}^{(k)} - \hat{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] + \left(1 - \frac{1}{n}\right) \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \right\rangle \right] + \frac{1}{n} \mathbb{E} \left[\langle \eta_{i_k}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] \\
& \stackrel{(a)}{\leq} -v_{\min} \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + \left(1 - \frac{1}{n}\right) \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \right\rangle \right] + \frac{1}{n} \mathbb{E} \left[\langle \eta_{i_k}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] \\
& \stackrel{(b)}{\leq} -v_{\min} \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + \frac{1 - \frac{1}{n}}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] \\
& + \frac{\beta(n-1)+1}{2n} \mathbb{E} \left[\left\| \nabla V(\hat{s}^{(k)}) \right\|^2 \right] + \frac{1}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
& \stackrel{(a)}{\leq} \left(v_{\max}^2 \frac{\beta(n-1)+1}{2n} - v_{\min} \right) \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + \frac{1 - \frac{1}{n}}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] + \frac{1}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right]
\end{aligned} \tag{60}$$

403 where (a) is due to the growth condition (2) and (b) is due to Young's inequality (with $\beta \rightarrow 1$). Note

404 $a_k = \gamma_{k+1} \left(v_{\min} - v_{\max}^2 \frac{\beta(n-1)+1}{2n} \right)$ and

$$\begin{aligned}
a_k \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] & \leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] \\
& + \frac{\gamma_{k+1}(1 - \frac{1}{n})}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] + \frac{\gamma_{k+1}}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right]
\end{aligned} \tag{61}$$

405 We now give an upper bound of $\mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right]$ using Lemma 7 and plug it into (61):

$$\begin{aligned}
(a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] & \leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] \\
& + \gamma_{k+1} \left(\frac{1}{2\beta} \left(1 - \frac{1}{n}\right) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] \\
& + \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
& + \frac{\gamma_{k+1}^2 L_V L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{s}^{(k)} - \hat{s}^{(\tau_i^k)} \right\|^2 \right]
\end{aligned} \tag{62}$$

406 Next, we observe that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{s}^{(k+1)} - \hat{s}^{(\tau_i^{k+1})} \right\|^2 \right] = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \mathbb{E} \left[\left\| \hat{s}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] + \frac{n-1}{n} \mathbb{E} \left[\left\| \hat{s}^{(k+1)} - \hat{s}^{(\tau_i^k)} \right\|^2 \right] \right) \tag{63}$$

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

$$\begin{aligned}
& \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\
&= \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)} \rangle\right] \\
&= \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2 - 2\gamma_{k+1}\langle \hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)} \rangle\right] \\
&\leq \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2 + \gamma_{k+1}\beta\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2\right]
\end{aligned} \tag{64}$$

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

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\tau_i^{k+1})}\|^2] \\
&\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n^2} \sum_{i=1}^n \mathbb{E}\left[(1 + \gamma_{k+1}\beta)\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2\right]
\end{aligned} \tag{65}$$

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

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\tau_i^{k+1})}\|^2] \\
&\leq (\gamma_{k+1}^2 + \frac{n-1}{n} \frac{\gamma_{k+1}}{\beta}) \mathbb{E}[\|\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{i=1}^n \mathbb{E}\left[\frac{1 - \frac{1}{n} + \gamma_{k+1}\beta}{n} \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2\right] \\
&\leq 4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + 2(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \mathbb{E}\left[\|\eta_{i_k}^{(k)}\|^2\right] \\
&\quad + 4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\
&\quad + \sum_{i=1}^n \mathbb{E}\left[\frac{1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})}{n} \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2\right]
\end{aligned} \tag{66}$$

410 Let us define

$$\Delta^{(k)} := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2] \tag{67}$$

411 From the above, we get

$$\begin{aligned}
\Delta^{(k+1)} &\leq (1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}))\Delta^{(k)} + 4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] \\
&\quad + 2(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \mathbb{E}\left[\|\eta_{i_k}^{(k)}\|^2\right] + 4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right]
\end{aligned} \tag{68}$$

412 Setting $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_{\mathbf{s}}, L_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$, $\beta = \frac{c_1 \bar{L}}{n}$,

413 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 6$, $\alpha \geq 8$, we observe that

$$1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \leq 1 - \frac{c_1(k\alpha - 1) - 4}{k\alpha n c_1} \leq 1 - \frac{2}{k\alpha n c_1} \tag{69}$$

414 which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} =$
 415 $\frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1}L_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

$$\begin{aligned} \Delta^{(k+1)} &\leq 4 \sum_{\ell=0}^k \prod_{j=\ell+1}^k \left(1 - \Lambda_{(j)}\right) (\gamma_{\ell+1}^2 + \frac{\gamma_{\ell+1}}{\beta}) \mathbb{E}[\|\bar{\mathbf{s}}^{(\ell)} - \hat{\mathbf{s}}^{(\ell)}\|^2] + 2 \sum_{\ell=0}^k \prod_{j=\ell+1}^k \left(1 - \Lambda_{(j)}\right) (\gamma_{\ell+1}^2 + \frac{\gamma_{\ell+1}}{\beta}) \mathbb{E}[\|\eta_{i_\ell}^{(\ell)}\|^2] \\ &\quad + 4 \sum_{\ell=0}^k \prod_{j=\ell+1}^k \left(1 - \Lambda_{(j)}\right) (\gamma_{\ell+1}^2 + \frac{\gamma_{\ell+1}}{\beta}) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^\ell)} - \bar{\mathbf{s}}^{(\ell)} \right\|^2 \right] \end{aligned} \quad (70)$$

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

$$\begin{aligned} &\sum_{k=0}^{K_{\max}-1} \Delta^{(k+1)} \\ &= 4 \sum_{k=0}^{K_{\max}-1} (\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \omega_{k,1} \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + 2 \sum_{k=0}^{K_{\max}-1} (\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \omega_{k,1} \mathbb{E}[\|\eta_{i_\ell}^{(k)}\|^2] \\ &\quad + \sum_{k=0}^{K_{\max}-1} 4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}) \omega_{k,1} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \\ &\leq \sum_{k=0}^{K_{\max}-1} \frac{4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})}{\Lambda_{(k+1)}} \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{k=0}^{K_{\max}-1} \frac{2(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})}{\Lambda_{(k+1)}} \mathbb{E}[\|\eta_{i_\ell}^{(k)}\|^2] \\ &\quad + \sum_{k=0}^{K_{\max}-1} \frac{4(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta})}{\Lambda_{(k+1)}} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \end{aligned} \quad (71)$$

417 We recall (62) where we have summed on both sides from $k = 0$ to $k = K_{\max} - 1$:

$$\begin{aligned} &\sum_{k=0}^{K_{\max}-1} (a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] \leq \mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K)})] \\ &\quad + \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \left(\frac{1}{2\beta} (1 - \frac{1}{n}) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \\ &\quad + \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \\ &\quad + \sum_{k=0}^{K_{\max}-1} \frac{\gamma_{k+1}^2 L_V L_s^2}{n^2} \Delta^{(k)} \end{aligned} \quad (72)$$

418 Plugging (71) into (72) results in:

$$\begin{aligned} &\sum_{k=0}^{K_{\max}-1} \tilde{\alpha}_k \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{k=0}^{K_{\max}-1} \tilde{\beta}_k \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \leq \mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K)})] \\ &\quad + \sum_{k=0}^{K_{\max}-1} \tilde{\Gamma}_k \mathbb{E}[\|\eta_{i_k}^{(k)}\|^2] \end{aligned} \quad (73)$$

419 where:

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

420 and

$$\begin{aligned}a_k &= \gamma_{k+1} \left(v_{\min} - v_{\max}^2 \frac{\beta(n-1) + 1}{2n} \right) \\ \Lambda_{(k+1)} &= \frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1} L_{\mathbf{s}}^2}{n^2} (\gamma_{k+1} + \frac{1}{\beta}) \\ c_1 &= v_{\min}^{-1}, \alpha = \max\{8, 1 + 6v_{\min}\}, \bar{L} = \max\{L_{\mathbf{s}}, L_V\}, \gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}, \beta = \frac{c_1 \bar{L}}{n}\end{aligned}$$

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

$$\sum_{k=0}^{K_{\max}} \tilde{\alpha}_k \mathbb{E} \left[\left\| \nabla V(\hat{\mathbf{s}}^{(k)}) \right\|^2 \right] \leq v_{\max}^2 \sum_{k=0}^{K_{\max}} \tilde{\alpha}_k \mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \quad (74)$$

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

424 D Proofs of Auxiliary Lemmas

425 D.1 Proof of Lemma 4 and Lemma 5

426 **Lemma.** For any $k \geq 0$ and consider the vrTTSEM update in (10) with $\rho_k = \rho$, it holds for all
427 $k > 0$

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \right\|^2 \right] &\leq 2\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 L_s^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2] \\ &\quad + 2(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{((k))} - \tilde{S}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \end{aligned} \quad (75)$$

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

429 **Proof** Beforehand, we provide a rewriting of the quantity $\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}$ that will be useful through-
430 out this proof:

$$\begin{aligned} \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}) = -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho\mathbf{S}^{(k+1)}) \\ &= -\gamma_{k+1} \left((1-\rho) \left[\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right] + \rho \left[\hat{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)} \right] \right) \end{aligned} \quad (76)$$

431 We observe, using the identity (76), that

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq 2\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)}\|^2] + 2(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{((k))} - \tilde{S}^{(k)}\|^2] \quad (77)$$

432 For the latter term, we obtain its upper bound as

$$\begin{aligned} \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)}\|^2] &= \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n (\bar{\mathbf{s}}_i^{(k)} - \tilde{S}_i^{\ell(k)}) - (\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{\ell(k)}) \right\|^2 \right] \\ &\stackrel{(a)}{\leq} \mathbb{E}[\|\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{\ell(k)}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \stackrel{(b)}{\leq} L_s^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{\ell(k)}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \end{aligned} \quad (78)$$

433 where (a) uses the variance inequality and (b) uses Lemma 1. Substituting into (77) proves the
434 lemma. \square

435 **Lemma.** For any $k \geq 0$ and consider the fiTTSEM update in (11) with $\rho_k = \rho$, it holds for all $k > 0$
436

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \right\|^2 \right] &\leq 2\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 \frac{L_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &\quad + 2(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{((k))} - \tilde{S}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \end{aligned} \quad (79)$$

437 **Proof** Beforehand, we provide a rewriting of the quantity $\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}$ that will be useful through-
438 out this proof:

$$\begin{aligned} \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}) \\ &= -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho\mathbf{S}^{(k+1)}) \\ &= -\gamma_{k+1} \left((1-\rho) \left[\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right] + \rho \left[\hat{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)} \right] \right) \\ &= -\gamma_{k+1} \left((1-\rho) \left[\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right] + \rho \left[\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} - (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) \right] \right) \end{aligned} \quad (80)$$

439 We observe, using the identity (80), that

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq 2\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)}\|^2] + 2(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{((k))} - \tilde{S}^{(k)}\|^2] \quad (81)$$

440 For the latter term, we obtain its upper bound as

$$\begin{aligned}\mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)}\|^2] &= \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n (\bar{\mathbf{s}}_i^{(k)} - \bar{\mathbf{S}}_i^{(k)}) - (\tilde{\mathbf{S}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(t_{i_k}^k)})\right\|^2\right] \\ &\stackrel{(a)}{\leq} \mathbb{E}[\|\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\ell(k))}\|^2] + \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2]\end{aligned}\quad (82)$$

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

$$\mathbb{E}[\|\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\bar{\mathbf{s}}_i^{(k)} - \bar{\mathbf{s}}_i^{(t_i^k)}\|^2] \stackrel{(a)}{\leq} \frac{L_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \quad (83)$$

442 Substituting into (81) proves the lemma. \square

443 D.2 Proof of Lemma 6

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

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\|^2] \leq \frac{\rho}{1-\rho} \sum_{\ell=0}^k (1-\gamma_\ell)^2 (\mathbf{S}^{(\ell)} - \tilde{\mathbf{S}}^{(\ell)}) \quad (84)$$

446 where $\mathbf{S}^{(k)}$ is defined either by (11) (fTTSEM) or (10) (vrTTSEM)

447 **Proof** We begin by writing the two-time-scale update:

$$\begin{aligned}\tilde{\mathbf{S}}^{(k+1)} &= \tilde{\mathbf{S}}^{(k)} + \rho(\mathbf{S}^{(k+1)} - \tilde{\mathbf{S}}^{(k)}) \\ \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)})\end{aligned}\quad (85)$$

448 where $\mathbf{S}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(t_i^k)} + (\tilde{\mathbf{S}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(t_{i_k}^k)})$ according to (11). Denote $\delta^{(k+1)} = \hat{\mathbf{s}}^{(k+1)} -$
449 $\tilde{\mathbf{S}}^{(k+1)}$. Then from (85), doing the subtraction of both equations yields:

$$\delta^{(k+1)} = (1 - \gamma_{k+1})\delta^{(k)} + \frac{\rho}{1-\rho} (1 - \gamma_{k+1})(\mathbf{S}^{(k+1)} - \tilde{\mathbf{S}}^{(k+1)}) \quad (86)$$

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

$$\delta^{(k+1)} \leq \frac{\rho}{1-\rho} \sum_{\ell=0}^k (1 - \gamma_{\ell+1})^2 (\mathbf{S}^{(\ell+1)} - \tilde{\mathbf{S}}^{(\ell+1)}) \quad (87)$$

451 \square

452 D.3 Additional Intermediary Result

453 **Lemma 8.** At iteration $k + 1$, the drift term of update (11), with $\rho_{k+1} = \rho$, is equivalent to the
454 following :

$$\begin{aligned}\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} &= \rho(\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}) + \rho\eta_{i_k}^{(k+1)} + \rho \left[(\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(t_{i_k}^k)}] \right] \\ &\quad + (1 - \rho) (\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)})\end{aligned}\quad (88)$$

455 where we recall that $\eta_{i_k}^{(k+1)}$, defined in (19), which is the gap between the MC approximation and
456 the expected statistics.

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

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

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

460 *Important Note:* Note that $\overline{\mathcal{S}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not equal to $\eta_{i_k}^{(k+1)}$, defined in (19), which is the gap
461 between the MC approximation and the expected statistics. Indeed $\tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not computed under the
462 same model as $\overline{\mathcal{S}}_{i_k}^{(k)}$. □

463 E Proof of Theorem 2

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

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

$$\begin{aligned} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \frac{2n^{2/3}\bar{L}}{\mu v_{\min}^2 v_{\max}^2} \mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})] \\ &\quad + \frac{2n^{2/3}\bar{L}}{\mu v_{\min}^2 v_{\max}^2} \sum_{k=0}^{K_{\max}-1} \left[\tilde{\eta}^{(k+1)} + \chi^{(k+1)} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right] \right] \end{aligned} \quad (89)$$

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

$$\begin{aligned} V(\hat{\mathbf{s}}^{(k+1)}) &\leq V(\hat{\mathbf{s}}^{(k)}) + \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{L_V}{2} \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (90)$$

470 Denote $\mathbf{H}_{k+1} := \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}$ the drift term of the fiTTSEM update in (7) and $\mathbf{h}_k = \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}$.
 471 Taking expectations on both sides show that

$$\begin{aligned} &\mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] \\ &\stackrel{(a)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1}(1-\rho) \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1}\rho \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|\mathbf{H}_{k+1}\|^2] \\ &\stackrel{(b)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1}\rho \mathbb{E}[\langle \mathbf{h}_k | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1}(1-\rho) \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad - \gamma_{k+1}\rho \mathbb{E}[\langle \eta_{i_k}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|\mathbf{H}_{k+1}\|^2] \\ &\stackrel{(c)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - (\gamma_{k+1}\rho v_{\min} + \gamma_{k+1}v_{\max}^2) \mathbb{E}[\|\mathbf{h}_k\|^2] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|\mathbf{H}_{k+1}\|^2] \\ &\quad - \gamma_{k+1}\rho \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] - \gamma_{k+1}(1-\rho) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \end{aligned} \quad (91)$$

472 where we have used (76) in (a) and $\mathbb{E}[\mathbf{S}^{(k+1)}] = \bar{\mathbf{s}}^{(k)} + \mathbb{E}[\eta_{i_k}^{(k+1)}]$ in (b), the growth condition in
 473 Lemma 2 and the Young's inequality with the constant equal to 1 in (c).

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

$$\begin{aligned} &\mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2] = \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} + \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2] \\ &= \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2 + \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))} | \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \rangle] \\ &= \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2 \|\mathbf{H}_{k+1}\|^2 \\ &\quad - 2\gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))} | \rho(\mathbf{h}_k - \eta_{i_k}^{(k+1)}) + (1-\rho)(\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}) \rangle] \\ &\leq \mathbb{E}[(1 + \gamma_{k+1}\beta) \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2 \|\mathbf{H}_{k+1}\|^2 + \frac{\gamma_{k+1}\rho}{\beta} \|\mathbf{h}_k\|^2 \\ &\quad + \frac{\gamma_{k+1}\rho}{\beta} \|\eta_{i_k}^{(k+1)}\|^2 + \frac{\gamma_{k+1}(1-\rho)}{\beta} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2], \end{aligned} \quad (92)$$

475 where we first used (76) and the last inequality is due to the Young's inequality.

476 Consider the following sequence

$$R_k := \mathbb{E}[V(\hat{\mathbf{s}}^{(k)}) + b_k \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2] \quad (93)$$

477 where $b_k := \bar{b}_{k \bmod m}$ is a periodic sequence where:

$$\bar{b}_i = \bar{b}_{i+1}(1 + \gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2 L_{\mathbf{s}}^2) + \gamma_{k+1}^2\rho^2 L_V L_{\mathbf{s}}^2, \quad i = 0, 1, \dots, m-1 \quad \text{with } \bar{b}_m = 0. \quad (94)$$

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

$$\bar{b}_i \leq \bar{b}_0 = \gamma_{k+1}^2\rho^2 L_V L_{\mathbf{s}}^2 \frac{(1 + \gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2 L_{\mathbf{s}}^2)^m - 1}{\gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2 L_{\mathbf{s}}^2}, \quad i = 1, 2, \dots, m. \quad (95)$$

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

$$\begin{aligned} R_{k+1} &\leq \mathbb{E}\left[V(\hat{\mathbf{s}}^{(k)}) - (\gamma_{k+1}\rho v_{\min} + \gamma_{k+1}v_{\max}^2) \|\mathbf{h}_k\|^2 + \frac{\gamma_{k+1}^2 L_V}{2} \|\mathbf{H}_{k+1}\|^2\right] \\ &\quad + \gamma_{k+1} \mathbb{E}\left[\rho \left\|\eta_{i_k}^{(k+1)}\right\|^2 - (1-\rho) \left\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right\|^2\right] \\ &\quad + b_{k+1} \mathbb{E}\left[(1 + \gamma_{k+1}\beta) \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2 + \gamma_{k+1}^2 \|\mathbf{H}_{k+1}\|^2 + \frac{\gamma_{k+1}\rho}{\beta} \|\mathbf{h}_k\|^2\right] \\ &\quad + b_{k+1} \mathbb{E}\left[\frac{\gamma_{k+1}\rho}{\beta} \left\|\eta_{i_k}^{(k+1)}\right\|^2 + \frac{\gamma_{k+1}(1-\rho)}{\beta} \left\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right\|^2\right] \end{aligned} \quad (96)$$

480 And using Lemma 4 we obtain:

$$\begin{aligned} R_{k+1} &\leq \mathbb{E}\left[V(\hat{\mathbf{s}}^{(k)}) - (\gamma_{k+1}\rho v_{\min} + \gamma_{k+1}v_{\max}^2 - \gamma_{k+1}^2\rho^2 L_V) \|\mathbf{h}_k\|^2 + \gamma_{k+1}^2\rho^2 L_V L_{\mathbf{s}}^2 \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2\right] \\ &\quad + b_{k+1} \mathbb{E}\left[(1 + \gamma_{k+1}\beta + 2\gamma_{k+1}^2\rho^2 L_{\mathbf{s}}^2) \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2 + \left(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2\right) \|\mathbf{h}_k\|^2\right] \\ &\quad + \gamma_{k+1} \mathbb{E}\left[(\rho + \rho^2\gamma_{k+1} L_V) \left\|\eta_{i_k}^{(k+1)}\right\|^2 - (1-\rho - (1-\rho)^2\gamma_{k+1} L_V) \left\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right\|^2\right] \\ &\quad + b_{k+1} \mathbb{E}\left[\left(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2\right) \left\|\eta_{i_k}^{(k+1)}\right\|^2 + \left(\frac{\gamma_{k+1}(1-\rho)}{\beta} + 2\gamma_{k+1}^2(1-\rho)^2\right) \|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\|^2\right] \end{aligned} \quad (97)$$

481 Rearranging the terms yields:

$$\begin{aligned} R_{k+1} &\leq \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1}(\rho v_{\min} + v_{\max}^2 - \gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)) \mathbb{E}[\|\mathbf{h}_k\|^2] \\ &\quad + \underbrace{\left(b_{k+1}(1 + \gamma\beta + 2\gamma^2\rho^2 L_{\mathbf{s}}^2) + \gamma^2\rho^2 L_V L_{\mathbf{s}}^2\right)}_{=b_k \text{ since } k+1 \leq \ell(k) + m} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\ell(k))}\|^2] + \tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)} \end{aligned} \quad (98)$$

482 where

$$\begin{aligned} \tilde{\eta}^{(k+1)} &= \left(\gamma_{k+1}(\rho + \rho^2\gamma_{k+1} L_V) + b_{k+1}(\frac{\gamma_{k+1}\rho}{\beta} + 2\gamma_{k+1}^2\rho^2)\right) \mathbb{E}\left[\left\|\eta_{i_k}^{(k+1)}\right\|^2\right] \\ \chi^{(k+1)} &= \left(b_{k+1}(\frac{\gamma_{k+1}(1-\rho)}{\beta} + 2\gamma_{k+1}^2(1-\rho)^2) - \gamma_{k+1}(1-\rho - (1-\rho)^2\gamma_{k+1} L_V)\right) \\ \tilde{\chi}^{(k+1)} &= \chi^{(k+1)} \mathbb{E}\left[\left\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right\|^2\right] \end{aligned} \quad (99)$$

483 This leads, using Lemma 2, that for any γ_{k+1} , ρ and β such that $\rho v_{\min} + v_{\max}^2 -$
484 $\gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2) > 0$,

$$\begin{aligned} v_{\max}^2 \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] \leq \frac{R_k - R_{k+1}}{\gamma_{k+1}(\rho v_{\min} + v_{\max}^2 - \gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2))} \\ &\quad + \frac{\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}}{\gamma_{k+1}(\rho v_{\min} + v_{\max}^2 - \gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2))} \end{aligned} \quad (100)$$

485 We first remark that

$$\begin{aligned} & \gamma_{k+1}(\rho v_{\min} + v_{\max}^2 - \gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2)) \\ & \geq \frac{\gamma_{k+1}\rho}{c_1}(1 - \gamma_{k+1}c_1\rho L_V - b_{k+1}(\frac{c_1}{\beta} + 2\gamma_{k+1}\rho c_1)) \end{aligned} \quad (101)$$

486 where $c_1 = v_{\min}^{-1}$. By setting $\bar{L} = \max\{L_s, L_V\}$, $\beta = \frac{c_1\bar{L}}{n^{1/3}}$, $\rho = \frac{\mu}{c_1\bar{L}n^{2/3}}$, $m = \frac{nc_1^2}{2\mu^2 + \mu c_1^2}$ and
 487 $\{\gamma_{k+1}\}$ any sequence of decreasing stepsizes in $(0, 1)$, it can be shown that there exists $\mu \in (0, 1)$,
 488 such that the following lower bound holds

$$\begin{aligned} & 1 - \gamma_{k+1}c_1\rho L_V - b_{k+1}(\frac{c_1}{\beta} + 2\gamma_{k+1}\rho c_1) \geq 1 - \frac{\mu}{n^{\frac{2}{3}}} - \bar{b}_0(\frac{n^{\frac{1}{3}}}{\bar{L}} + \frac{2\mu}{\bar{L}n^{\frac{2}{3}}}) \\ & \geq 1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{L_V\mu^2}{c_1^2n^{\frac{4}{3}}}\frac{(1 + \gamma\beta + 2\gamma^2L_s^2)^m - 1}{\gamma\beta + 2\gamma^2L_s^2}(\frac{n^{\frac{1}{3}}}{\bar{L}} + \frac{2\mu}{\bar{L}n^{\frac{2}{3}}}) \\ & \stackrel{(a)}{\geq} 1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{\mu}{c_1^2}(e - 1)(1 + \frac{2\mu}{n}) \geq 1 - \mu - \mu(1 + 2\mu)\frac{e - 1}{c_1^2} \stackrel{(b)}{\geq} \frac{1}{2} \end{aligned} \quad (102)$$

489 where the simplification in (a) is due to

$$\frac{\mu}{n} \leq \gamma\beta + 2\gamma^2L_s^2 \leq \frac{\mu}{n} + \frac{2\mu^2}{c_1^2n^{\frac{4}{3}}} \leq \frac{\mu c_1^2 + 2\mu^2}{c_1^2} \frac{1}{n} \text{ and } (1 + \gamma\beta + 2\gamma^2L_s^2)^m \leq e - 1. \quad (103)$$

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

491 Finally, these results yield:

$$v_{\max}^2 \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{s}^{(k)})\|^2] \leq \frac{2(R_0 - R_{K_{\max}})}{v_{\min}\rho} + 2 \sum_{k=0}^{K_{\max}-1} \frac{\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}}{v_{\min}\rho} \quad (104)$$

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

$$\sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{s}^{(k)})\|^2] \leq \frac{2n^{2/3}\bar{L}}{\mu v_{\min}^2 v_{\max}^2} \mathbb{E}[V(\hat{s}^{(0)}) - V(\hat{s}^{(K_{\max})})] + \frac{2n^{2/3}\bar{L}}{\mu v_{\min}^2 v_{\max}^2} \sum_{k=0}^{K_{\max}-1} [\tilde{\eta}^{(k+1)} + \tilde{\chi}^{(k+1)}] \quad (105)$$

494 This concludes our proof.

495 □

496 F Proof of Theorem 3

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

500 Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$. By setting $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$,
 501 $\beta = \frac{c_1 \bar{L}}{n}$, $\rho = \frac{1}{n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$, we have the following bound:

$$\begin{aligned} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \frac{\alpha \bar{L} n^{2/3}}{v_{\min} v_{\max}^2} [V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})] \\ &\quad + \frac{\alpha \bar{L} n^{2/3}}{v_{\min} v_{\max}^2} \sum_{k=0}^{K_{\max}-1} \left[\Xi^{(k+1)} + \Gamma_{k+1} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right] \right] \end{aligned} \quad (106)$$

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

$$\begin{aligned} V(\hat{\mathbf{s}}^{(k+1)}) &\leq V(\hat{\mathbf{s}}^{(k)}) + \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{L_V}{2} \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (107)$$

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

$$\mathbb{E} \left[(\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(k)}) - \mathbb{E}[\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(k)}] \right] = 0 \quad (108)$$

505 we have:

$$\begin{aligned} &\mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] \\ &\leq \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1} \rho \mathbb{E}[\langle \mathbf{h}_k | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1} \mathbb{E} \left[\langle \rho \mathbb{E}[\eta_{i_k}^{(k+1)} | \mathcal{F}_k] + (1 - \rho) \mathbb{E}[\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}] | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2} \|\mathbf{H}_{k+1}\|^2 \\ &\stackrel{(a)}{\leq} -v_{\min} \gamma_{k+1} \rho \mathbb{E}[\|\mathbf{h}_k\|^2] - \gamma_{k+1} \mathbb{E} \left[\left\| \nabla V(\hat{\mathbf{s}}^{(k)}) \right\|^2 \right] - \frac{\gamma_{k+1} \rho^2}{2} \xi^{(k+1)} - \frac{\gamma_{k+1} (1 - \rho)^2}{2} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2} \|\mathbf{H}_{k+1}\|^2 \\ &\stackrel{(b)}{\leq} -(v_{\min} \gamma_{k+1} \rho + \gamma_{k+1} v_{\max}^2) \mathbb{E}[\|\mathbf{h}_k\|^2] - \frac{\gamma_{k+1} \rho^2}{2} \xi^{(k+1)} - \frac{\gamma_{k+1} (1 - \rho)^2}{2} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2} \|\mathbf{H}_{k+1}\|^2 \end{aligned} \quad (109)$$

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

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

$$\begin{aligned} &\gamma_{k+1} (v_{\min} \rho + v_{\max}^2 - \gamma_{k+1} \rho^2 L_V) \mathbb{E}[\|\mathbf{h}_k\|^2] \\ &\leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(k)}) - V(\hat{\mathbf{s}}^{(k+1)}) \right] + \tilde{\xi}^{(k+1)} + \left((1 - \rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1} (1 - \rho)^2}{2} \right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &\quad + \frac{\gamma_{k+1}^2 L_V \rho^2 L_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \end{aligned} \quad (110)$$

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

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n} \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \right) \quad (111)$$

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

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &= \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle] \end{aligned} \quad (112)$$

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

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &= \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle] \\ &\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + (1 + \gamma_{k+1}\beta)\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathbf{h}_k\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \\ &\quad + \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]] \end{aligned} \quad (113)$$

513 where the last inequality is due to the Young's inequality. Plugging this into (111) yields:

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &= \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle] \\ &\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + (1 + \gamma_{k+1}\beta)\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathbf{h}_k\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \\ &\quad + \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]] \end{aligned} \quad (114)$$

514 Subsequently, we have

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] \\ &\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n^2} \sum_{i=1}^n \mathbb{E}[(1 + \gamma_{k+1}\beta)\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\|\mathbf{h}_k\|^2] \\ &\quad + \frac{\gamma_{k+1}\rho^2}{\beta}\mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] + \frac{\gamma_{k+1}(1-\rho)^2}{\beta}\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]] \end{aligned} \quad (115)$$

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

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] \\
& \leq \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1} \rho^2}{\beta}\right) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{i=1}^n \left(\frac{\gamma_{k+1}^2 \rho^2 L_s^2}{n} + \frac{(n-1)(1+\gamma_{k+1}\beta)}{n^2}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\
& \quad + \gamma_{k+1}(1-\rho)^2 \left(2\gamma_{k+1} + \frac{1}{\beta}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \left(2\gamma_{k+1}^2 + \frac{\gamma_{k+1} \rho^2}{\beta}\right) \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] \\
& \leq \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1} \rho^2}{\beta}\right) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{i=1}^n \left(\frac{1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 L_s^2}{n}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\
& \quad + \gamma_{k+1}(1-\rho)^2 \left(2\gamma_{k+1} + \frac{1}{\beta}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \left(2\gamma_{k+1}^2 + \frac{\gamma_{k+1} \rho^2}{\beta}\right) \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2]
\end{aligned} \tag{116}$$

516 Let us define

$$\Delta^{(k)} := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \tag{117}$$

517 From the above, we get

$$\begin{aligned}
\Delta^{(k+1)} & \leq \left(1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 L_s^2\right) \Delta^{(k)} + \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1} \rho^2}{\beta}\right) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] \\
& \quad + \gamma_{k+1}(1-\rho)^2 \left(2\gamma_{k+1} + \frac{1}{\beta}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \gamma_{k+1} \left(2\gamma_{k+1} + \frac{\rho^2}{\beta}\right) \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2]
\end{aligned} \tag{118}$$

518 Setting $c_1 = v_{\min}^{-1}$, $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$, $\beta = \frac{c_1 \bar{L}}{n}$,
519 $\rho = \frac{1}{n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$, we observe that

$$1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 L_s^2 \leq 1 - \frac{1}{n} + \frac{1}{\alpha k n} + \frac{1}{\alpha^2 c_1^2 k^2 n^{4/3}} \leq 1 - \frac{c_1(k\alpha - 1) - 1}{k\alpha n c_1} \leq 1 - \frac{1}{k\alpha n c_1} \tag{119}$$

520 which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 L_s^2 \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} = \frac{1}{n} -$
521 $\gamma_{k+1}\beta - \gamma_{k+1}^2 \rho^2 L_s^2$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

$$\begin{aligned}
\Delta^{(k+1)} & \leq \sum_{\ell=0}^k \omega_{k,\ell} \left(2\gamma_{\ell+1}^2 \rho^2 + \frac{\gamma_{\ell+1} \rho^2}{\beta}\right) \mathbb{E}[\|\bar{\mathbf{s}}^{(\ell)} - \hat{\mathbf{s}}^{(\ell)}\|^2] \\
& \quad + \sum_{\ell=0}^k \omega_{k,\ell} \gamma_{\ell+1} (1-\rho)^2 \left(2\gamma_{\ell+1} + \frac{1}{\beta}\right) \mathbb{E}[\|\tilde{S}^{(\ell)} - \hat{\mathbf{s}}^{(\ell)}\|^2] + \sum_{\ell=0}^k \omega_{k,\ell} \gamma_{\ell+1} \tilde{\epsilon}^{(\ell+1)}
\end{aligned} \tag{120}$$

522 where $\omega_{k,\ell} = \prod_{j=\ell+1}^k (1 - \Lambda_{(j)})$ and $\tilde{\epsilon}^{(\ell+1)} = \left(2\gamma_{\ell+1} + \frac{\rho^2}{\beta}\right) \mathbb{E}[\|\eta_{i_k}^{(\ell+1)}\|^2]$.

523 Summing on both sides over $k = 0$ to $k = K_{\max} - 1$ yields:

$$\begin{aligned}
\sum_{k=0}^{K_{\max}-1} \Delta^{(k+1)} & \leq \sum_{k=0}^{K_{\max}-1} \frac{2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1} \rho^2}{\beta}}{\Lambda_{(k+1)}} \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] \\
& \quad + \sum_{k=0}^{K_{\max}-1} \frac{\gamma_{k+1}(1-\rho)^2 \left(2\gamma_{k+1} + \frac{1}{\beta}\right)}{\Lambda_{(k+1)}} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + \sum_{k=0}^{K_{\max}-1} \frac{\gamma_{k+1}}{\Lambda_{(k+1)}} \tilde{\epsilon}^{(k+1)}
\end{aligned} \tag{121}$$

524 We recall (110) where we have summed on both sides from $k = 0$ to $k = K_{\max} - 1$:

$$\begin{aligned}
& \mathbb{E}[V(\hat{\mathbf{s}}^{(K_{\max})}) - V(\hat{\mathbf{s}}^{(0)})] \\
& \leq \sum_{k=0}^{K_{\max}-1} \left\{ \gamma_{k+1}(-(\nu_{\min}\rho + \nu_{\max}^2) + \gamma_{k+1}\rho^2 L_V) \mathbb{E}[\|\mathbf{h}_k\|^2] + \gamma^2 L_V \rho^2 L_{\mathbf{s}}^2 \Delta^{(k)} \right\} \\
& + \sum_{k=0}^{K_{\max}-1} \left\{ \tilde{\xi}^{(k+1)} + \left((1-\rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1}(1-\rho)^2}{2} \right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \right\} \\
& \leq \sum_{k=0}^{K_{\max}-1} \left\{ -\gamma_{k+1}(\nu_{\min}\rho + \nu_{\max}^2) + \gamma_{k+1}^2 \rho^2 L_V + \frac{\rho^2 \gamma_{k+1}^2 L_V L_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\Lambda_{(k+1)}} \right\} \mathbb{E}[\|\mathbf{h}_k\|^2] \\
& + \sum_{k=0}^{K_{\max}-1} \Xi^{(k+1)} + \sum_{k=0}^{K_{\max}-1} \Gamma_{k+1} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]
\end{aligned} \tag{122}$$

where

$$\Xi^{(k+1)} = \tilde{\xi}^{(k+1)} + \frac{\gamma_{k+1}^3 L_V \rho^2 L_{\mathbf{s}}^2}{\Lambda_{(k+1)}} \epsilon^{(k+1)}$$

and

$$\Gamma_{k+1} = \left((1-\rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1}(1-\rho)^2}{2} \right) + \frac{\gamma_{k+1}^3 L_V \rho^2 L_{\mathbf{s}}^2 (1-\rho)^2 \left(2\gamma_{k+1} + \frac{1}{\beta} \right)}{\Lambda_{(k+1)}}$$

525 We now analyse the following quantity

$$\begin{aligned}
& -\gamma_{k+1}(\nu_{\min}\rho + \nu_{\max}^2) + \gamma_{k+1}^2 \rho^2 L_V + \frac{\rho^2 \gamma_{k+1}^2 L_V L_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\Lambda_{(k+1)}} \\
& = -\gamma_{k+1} \left[(\nu_{\min}\rho + \nu_{\max}^2) + \gamma_{k+1} \rho^2 L_V + \frac{\rho^2 \gamma_{k+1} L_V L_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\Lambda_{(k+1)}} \right]
\end{aligned} \tag{123}$$

526 Furthermore, we recall that $c_1 = \nu_{\min}^{-1}$, $\alpha = \max\{2, 1+2\nu_{\min}\}$, $\bar{L} = \max\{L_{\mathbf{s}}, L_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$,

527 $\beta = \frac{c_1 \bar{L}}{n}$, $\rho = \frac{1}{n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$. Then,

$$\begin{aligned}
& \gamma_{k+1}^2 \rho^2 L_V + \frac{\rho^2 \gamma_{k+1}^2 L_V L_{\mathbf{s}}^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2 \rho^2 L_{\mathbf{s}}^2} \\
& \leq \frac{1}{k^2 \alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\bar{L} (k^2 \alpha^2 c_1^2 n^{4/3})^{-1} \left(\frac{2}{k^2 \alpha^2 c_1^2 \bar{L}^2 n^{4/3}} + \frac{1}{k \alpha c_1^2 \bar{L}^2 n^{1/3}} \right)}{\frac{1}{n} - \frac{1}{k \alpha n} - \frac{1}{k^2 \alpha^2 c_1^2 n^{4/3}}} \\
& = \frac{1}{k^2 \alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\bar{L} \left(\frac{2}{k^2 \alpha^2 c_1^2 \bar{L}^2 n^{4/3}} + \frac{1}{k \alpha c_1^2 \bar{L}^2 n^{1/3}} \right)}{(k \alpha c_1 n^{1/3})(k \alpha - 1) c_1 - 1} \\
& \stackrel{(a)}{\leq} \frac{1}{k^2 \alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\frac{1}{k \alpha c_1^2 \bar{L} n^{1/3}} \left(\frac{2}{k \alpha n} + 1 \right)}{2(\alpha c_1 n^{1/3}) - 1} \\
& \leq \frac{1}{k^2 \alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{1}{k \alpha^2 c_1^3 \bar{L} n^{2/3}} \\
& \leq \frac{1}{\alpha c_1 \bar{L} n^{2/3}}
\end{aligned} \tag{124}$$

528 where (a) is due to $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$ and $k \alpha c_1 n^{1/3} \geq 1$. Also, since $-\gamma_{k+1}(\nu_{\min}\rho +$
529 $\nu_{\max}^2) \leq -\gamma_{k+1}\rho \nu_{\min} = -$

530 Using the Lemma 2, we know that $v_{\max}^2 \|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2 \leq \|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2$ and using (124) on (122)
 531 yields:

$$\begin{aligned} v_{\max}^2 \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \frac{\alpha \bar{L} n^{2/3}}{v_{\min}} [V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})] \\ &\quad + \frac{\alpha \bar{L} n^{2/3}}{v_{\min}} \sum_{k=0}^{K_{\max}-1} \Xi^{(k+1)} + \sum_{k=0}^{K_{\max}-1} \Gamma_{k+1} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right] \end{aligned} \quad (125)$$

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

$$\begin{aligned} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2] &\leq \frac{\alpha \bar{L} n^{2/3}}{v_{\min} v_{\max}^2} [V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})] \\ &\quad + \frac{\alpha \bar{L} n^{2/3}}{v_{\min} v_{\max}^2} \sum_{k=0}^{K_{\max}-1} \Xi^{(k+1)} + \sum_{k=0}^{K_{\max}-1} \Gamma_{k+1} \mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right] \end{aligned} \quad (126)$$

533 **Bounding** $\mathbb{E} \left[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right\|^2 \right]$ Remark that this term is the price we pay for the two time scale
 534 dynamics and corresponds to the gap between the two asynchronous updates (one is on $\hat{\mathbf{s}}^{(k)}$ and the
 535 other on $\tilde{S}^{(k)}$).

536 **FIND AN UPPER BOUND TO THAT GAP**

537 □

G Practical Implementations of Two-Time-Scale EM Methods

G.1 Application on GMM

We first recognize that the constraint set for θ is given by

$$\Theta = \Delta^M \times \mathbb{R}^M. \quad (127)$$

Using the partition of the sufficient statistics as $S(y_i, z_i) = (S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i)^\top)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, the partition $\phi(\theta) = (\phi^{(1)}(\theta)^\top, \phi^{(2)}(\theta)^\top, \phi^{(3)}(\theta)^\top)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$ 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

$$\begin{aligned} s_{i,m}^{(1)} &= \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\theta) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\} - \left\{ \log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2} \right\}, \\ s_{i,m}^{(2)} &= \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\theta) = \mu_m, \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\theta) = \mu_M, \end{aligned} \quad (128)$$

and $\psi(\theta) = -\left\{ \log(1 - \sum_{m=1}^{M-1} \omega_m) - \frac{\mu_M^2}{2\sigma^2} \right\}$. 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 $\mathbb{E}_{\theta}[\mathbb{1}_{\{z_i=m\}} | y = y_i]$:

$$z_{i,m} \sim \mathbb{P}(z_i = m | y_i; \theta) \quad (129)$$

where $m \in \llbracket 1, M \rrbracket$, $i \in \llbracket 1, n \rrbracket$ and $\theta = (\mathbf{w}, \boldsymbol{\mu}) \in \Theta$.

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

$$\tilde{S}_{i_k}^{(k)} = \left(\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1})y_{i_k}, \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\tilde{s}_{i_k}^{(3)}(\theta^{(k)})} \right)^\top. \quad (130)$$

Recall that we have used the following regularizer:

$$\mathbf{r}(\theta) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log(1 - \sum_{m=1}^{M-1} \omega_m), \quad (131)$$

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

$$\bar{\theta}(\mathbf{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1} (s_1^{(1)} + \epsilon, \dots, s_{M-1}^{(1)} + \epsilon)^\top \\ ((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \dots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)})^\top \\ (1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta)^{-1} (s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}) \end{pmatrix} = \begin{pmatrix} \bar{\omega}(\mathbf{s}) \\ \bar{\boldsymbol{\mu}}(\mathbf{s}) \\ \bar{\mu}_M(\mathbf{s}) \end{pmatrix}. \quad (132)$$

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

G.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 (16) as the sufficient statistics are composed of indicator functions and observations as defined Section G.1 Equation (128).

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

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

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

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

564 Under H1 and H3, we have $\|\hat{s}^{(k)}\| < \infty$ since S is compact and $\hat{\theta}^{(k)} \in \text{int}(\Theta)$ for any $k \geq 0$ which
 565 thus ensure that the EM methods operate in a closed set throughout the optimization process.

566 G.3 Algorithms updates

567 In the sequel, recall that, for all $i \in \llbracket n \rrbracket$ and iteration k , the computed statistic $\tilde{S}_{i_k}^{(k)}$ is defined by
 568 (130). At iteration k , the several E-steps defined by (9) or (10) and (11) leads to the definition of the
 569 quantity $\hat{s}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{s}^{(0)} = n^{-1} \sum_{i=1}^n \bar{s}_i^{(0)}$,
 570 those E-steps break down as follows:

571 **Batch EM (EM):** for all $i \in \llbracket 1, n \rrbracket$, compute $\bar{s}_i^{(k)}$ and set

$$\hat{s}^{(k+1)} = n^{-1} \sum_{i=1}^n \bar{s}_i^{(k)}. \quad (133)$$

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

$$\tilde{\omega}_m(y_i; \theta) := \mathbb{E}_{\theta}[\mathbb{1}_{\{z_i=m\}} | y = y_i] = \frac{\omega_m \exp(-\frac{1}{2}(y_i - \mu_i)^2)}{\sum_{j=1}^M \omega_j \exp(-\frac{1}{2}(y_i - \mu_j)^2)}, \quad (134)$$

573 **Incremental EM (iEM):** draw an index i_k uniformly at random on $\llbracket n \rrbracket$, compute $\bar{s}_{i_k}^{(k)}$ and set

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} + \frac{1}{n} (\bar{s}_{i_k}^{(k)} - \bar{s}_{i_k}^{(\tau_i^k)}) = n^{-1} \sum_{i=1}^n \bar{s}_i^{(\tau_i^k)}. \quad (135)$$

574 **batch SAEM (SAEM):** draw an index i_k uniformly at random on $\llbracket n \rrbracket$, compute $\bar{s}_{i_k}^{(k)}$ and set

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} (1 - \gamma_{k+1}) + \gamma_{k+1} \tilde{S}^{(k)}. \quad (136)$$

575 where $\tilde{S} = \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(k)}$ with $\tilde{S}_i^{(k)}$ defined in (130).

576 **Incremental SAEM (iSAEM):** draw an index i_k uniformly at random on $\llbracket n \rrbracket$, compute $\bar{s}_{i_k}^{(k)}$ and set
 577

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} (1 - \gamma_{k+1}) + \gamma_{k+1} (\tilde{S}^{(k)} + \frac{1}{n} (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)})). \quad (137)$$

578 **Variance Reduced Two-Time-Scale EM (vrTTSEM):** draw an index i_k uniformly at random on
 579 $\llbracket n \rrbracket$, compute $\bar{s}_{i_k}^{(k)}$ and set

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} (1 - \gamma_{k+1}) + \gamma_{k+1} (\tilde{S}^{(k)} (1 - \rho) + \rho (\tilde{S}^{(\ell(k))} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}))). \quad (138)$$

580 **Fast Incremental Two-Time-Scale EM (fiTTSEM):** draw an index i_k uniformly at random on $\llbracket n \rrbracket$,
 581 compute $\bar{s}_{i_k}^{(k)}$ and set

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} (1 - \gamma_{k+1}) + \gamma_{k+1} (\tilde{S}^{(k)} (1 - \rho) + \rho (\bar{\mathcal{S}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}))). \quad (139)$$

582 Finally, the k -th update reads $\hat{\theta}^{(k+1)} = \bar{\theta}(\hat{s}^{(k+1)})$ where the function $s \rightarrow \bar{\theta}(s)$ is defined by (132).