
Fast Two-Timescale Stochastic EM Algorithms

Anonymous Author(s)

Affiliation

Address

email

Abstract

The Expectation-Maximization (EM) algorithm is a popular choice for learning latent variable models. Variants of the EM have been initially introduced by [25], using incremental updates to scale to large datasets, and by [30, 12], using Monte Carlo (MC) approximations to bypass the intractable conditional expectation of the latent data for most nonconvex models. In this paper, we propose a general class of methods called Two-Timescale EM Methods based on a two-stage approach of stochastic updates to tackle an essential nonconvex optimization task for latent variable models. We motivate the choice of a double dynamic by invoking the variance reduction virtue of each stage of the method on both sources of noise: the index sampling for the incremental update and the MC approximation. We establish finite-time and global convergence bounds for nonconvex objective functions. Numerical applications are also presented to illustrate our findings.

1 Introduction

Learning latent variable models is critical for modern machine learning problems, see (e.g.,) [23] for references. We formulate the training of such model as an empirical risk minimization problem:

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

We denote the observations by $\{y_i\}_{i=1}^n$, $\Theta \subset \mathbb{R}^d$ is the convex parameters set. We consider a smooth convex regularization noted $r : \Theta \rightarrow \mathbb{R}$ and $g(y; \theta)$ is the (incomplete) likelihood of each observation. The objective function $\bar{L}(\theta)$ is possibly *nonconvex* and is assumed to be lower bounded. 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 latent variables. In this paper, we assume that the complete model belongs to the curved exponential family [14]:

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

where $\psi(\theta)$, $h(z_i, y_i)$ are scalar functions, $\phi(\theta) \in \mathbb{R}^k$ is a vector function, and $\{S(z_i, y_i) \in \mathbb{R}^k\}_{i=1}^n$ is the vector of sufficient statistics of the complete model. Full batch EM [13, 31] is the method of reference for that type of task and is a two steps procedure. The **E-step** amounts to computing the conditional expectation of the complete data sufficient statistics,

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

and the **M-step** is given by

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

Two caveats of this method are the following: (a) with the explosion of data, the first step of the EM is computationally inefficient as it requires, at each iteration, a full pass over the dataset; and (b) the complexity of modern models makes the expectation in (3) intractable. So far, and to the best of our knowledge, both challenges have been addressed separately, as detailed in the sequel.

Prior Work: Inspired by stochastic optimization procedures, [25] and [8] develop respectively an incremental and an online variant of the E-step in models where the expectation is computable, and were then extensively used and studied in [27, 20, 7]. Some improvements of those methods have been provided and analyzed, globally and in finite-time, in [18] where variance reduction techniques taken from the optimization literature have been efficiently applied to scale the EM algorithm to large datasets. Regarding the computation of the expectation under the posterior distribution, the Monte Carlo EM (MCEM) has been introduced in the seminal paper [30] where an MC approximation for this expectation is computed. A variant of that algorithm is the Stochastic Approximation of the EM (SAEM) in [12] leveraging the power of Robbins-Monro update [29] to ensure pointwise convergence of the vector of estimated parameters using a decreasing stepsize rather than increasing the number of MC samples. The MCEM and the SAEM have been successfully applied in mixed effects models [22, 15, 4] or to do inference for joint modeling of time to event data coming from clinical trials in [10], unsupervised clustering in [26], variational inference of graphical models in [5] among other applications. Recently, an incremental variant of the SAEM was proposed in [19] showing positive empirical results but its analysis is limited to asymptotic consideration.

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

- We propose a two-timescale method based on (i) Stochastic Approximation (SA), to alleviate the problem of computing MC approximations, and on (ii) Incremental updates, to scale to large datasets. We describe in details the edges of each level of our method based on variance reduction arguments. Such class of algorithms has two advantages. First, it naturally leverages variance reduction and Robbins-Monro type of updates to tackle large-scale and highly nonlinear learning tasks. Then, it gives a simple formulation as a *scaled-gradient method* which makes the global analysis and the implementation accessible.
- We also establish global (independent of the initialization) and finite-time (true at each iteration) upper bounds on a classical sub-optimality condition in the nonconvex literature, *i.e.*, the second order moment of the gradient of the objective function.

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

2 Two-Timescale Stochastic EM Algorithms

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

2.1 Monte Carlo Integration and Stochastic Approximation

As mentioned in the Introduction, for complex and possibly nonconvex models, the expectation under the posterior distribution defined in (3) is not tractable. In that case, the first solution involves computing a Monte Carlo integration of that latter. For all $i \in [n]$, where $[n] := \{1, \dots, n\}$, draw $\{z_{i,m} \sim p(z_i|y_i; \theta)\}_{m \in [M]}$ samples and compute the MC integration \tilde{s} of $\bar{s}(\theta)$ defined by (3):

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

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

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

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

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

78 where $\{\gamma_k\}_{k>1} \in (0, 1)$ is a sequence of decreasing stepsizes to ensure asymptotic convergence.
 79 This is called the Stochastic Approximation of the EM (SAEM) and has been shown to converge to
 80 a maximum likelihood of the observations under very general conditions [12]. In simple scenarios,
 81 the samples $\{z_{i,m}\}_{m=0}^{M-1}$ are conditionally independent and identically distributed with distribution
 82 $p(z_i, \theta)$. Nevertheless, in most cases, since the loss function between the observed data y_i and the
 83 latent variable z_i can be nonconvex, sampling exactly from this distribution is not an option and the
 84 MC batch is sampled by Markov Chain Monte Carlo (MCMC) algorithm [24, 6].

85 **Role of the stepsize γ_k :** The sequence of decreasing positive integers $\{\gamma_k\}_{k>1}$ controls the con-
 86 vergence of the algorithm. It is inefficient to start with small values for the stepsize γ_k and large
 87 values for the number of simulations M_k . Rather, it is recommended that one decreases γ_k , as in
 88 $\gamma_k = 1/k^\alpha$, with $\alpha \in (0, 1)$, and keeps a constant and small number M_k bypassing the computationally
 89 involved sampling step in (5). In practice, γ_k is set equal to 1 during the first few iterations to let
 90 the iterates explore the parameter space without memory and converge quickly to a neighborhood
 91 of the target estimate. The Stochastic Approximation is performed during the remaining iterations
 92 ensuring the almost sure convergence of the vector of estimates.

93 This Robbins-Monro type of update constitutes the *first level* of our algorithm, needed to temper the
 94 variance and noise introduced by the Monte Carlo integration. In the next section, we derive variants
 95 of this algorithm to adapt to the sheer size of data of today's applications and formalize the *second*
 96 *level* of our class of two-timescale EM methods.

97 2.2 Incremental and Two-Stage Stochastic EM Methods

98 Efficient strategies to scale to large datasets include incremental [25] and variance reduced [11, 17]
 99 methods. We will explicit a general update that covers those latter variants and that represents the
 100 *second level* of our algorithm, i.e., the incremental update of the noisy statistics $\tilde{S}^{(k+1)}$ in (7):

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

101 Note that $\{\rho_k\}_{k>1} \in (0, 1)$ is a sequence of stepsizes, $\mathcal{S}^{(k)}$ is a proxy for $\tilde{S}^{(k)}$. If the stepsize
 102 is equal to one and the proxy $\mathcal{S}^{(k)} = \tilde{S}^{(k)}$, i.e., computed in a full batch manner as in (6), then
 103 we recover the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\mathcal{S}^{(k)} = \tilde{S}^{(k)}$, then we recover the
 104 MCEM [30]. For all methods, we define a random index drawn at iteration k , noted $i_k \in [n]$, and
 105 $\tau_i^k = \max\{k' : i_{k'} = i, k' < k\}$ as the iteration index where $i \in [n]$ is last drawn prior to iteration k .

106 The proposed fitTEM method draws
 107 two indices *independently* and uni-
 108 formly as $i_k, j_k \in [n]$. Thus, we de-
 109 fine $t_j^k = \{k' : j_{k'} = j, k' < k\}$
 110 to be the iteration index where the
 111 sample $j \in [n]$ is last drawn as j_k
 112 prior to iteration k in addition to τ_i^k
 113 which was defined w.r.t. i_k . Recall
 114 $\tilde{S}_{i_k}^{(k)} = \frac{1}{M_k} \sum_{m=1}^{M_k} S(z_{i_k,m}^{(k)}, y_{i_k})$ and

Table 1 Proxies for the Incremental-step (8)

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

115 $z_{i_k,m}^{(k)} \sim p(z_{i_k} | y_{i_k}; \theta^{(k)})$. The stepsize is set to $\rho_{k+1} = 1$ for the iSAEM method and we initialize
 116 with $\mathcal{S}^{(0)} = \tilde{S}^{(0)}$; $\rho_{k+1} = \rho$ is constant for the vrTTEM and fitTEM methods. Note that we ini-
 117 tialize as follows $\bar{\mathcal{S}}^{(0)} = \tilde{S}^{(0)}$ for the fitTEM which can be seen as a slightly modified version of
 118 SAGA inspired by [28]. For vrTTEM we set an epoch size of m and define $\ell(k) := m \lfloor k/m \rfloor$ as the
 119 first iteration number in the epoch that iteration k is in.

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

Algorithm 1 Two-Timescale Stochastic EM methods.

- 1: **Input:** $\hat{\theta}^{(0)} \leftarrow 0, \hat{s}^{(0)} \leftarrow \tilde{S}^{(0)}, \{\gamma_k\}_{k>0}, \{\rho_k\}_{k>0}$ and $K_m \in \mathbb{N}^*$.
 - 2: **for** $k = 0, 1, 2, \dots, K_m - 1$ **do**
 - 3: Draw index $i_k \in [n]$ uniformly (and $j_k \in [n]$ for fitTEM).
 - 4: Compute $\tilde{S}_{i_k}^{(k)}$ using the MC-step (5), for the drawn indices.
 - 5: Compute the surrogate sufficient statistics $\mathcal{S}^{(k+1)}$ using Lines 1, 2 or 3 in Table 1.
 - 6: Compute $\tilde{S}^{(k+1)}$ and $\hat{s}^{(k+1)}$ using respectively (8) and (7):
$$\begin{aligned}\tilde{S}^{(k+1)} &= \tilde{S}^{(k)} + \rho_{k+1}(\mathcal{S}^{(k+1)} - \tilde{S}^{(k)}) \\ \hat{s}^{(k+1)} &= \hat{s}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{s}^{(k)})\end{aligned}\tag{9}$$
 - 7: Compute $\hat{\theta}^{(k+1)} = \bar{\theta}(\hat{s}^{(k+1)})$ via the M-step.
 - 8: **end for**
-

The update in (9) is said to have a two-timescale property as the stepsizes satisfy $\lim_{k \rightarrow \infty} \gamma_k / \rho_k < 1$ such that $\tilde{S}^{(k+1)}$ is updated at a faster time-scale, determined by ρ_{k+1} , than $\hat{s}^{(k+1)}$, determined by γ_{k+1} . The next section introduces the main results of this paper and establishes global and finite-time bounds for the three different updates of our scheme.

3 Finite Time Analysis of the Two-Timescale Scheme

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

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

that we propose to study in this article.

3.1 Assumptions and Intermediate Lemmas

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

A1. *The sets \mathcal{Z}, \mathcal{S} are compact. There exist constants C_S, C_Z such that:*

$$C_S := \max_{\mathbf{s}, \mathbf{s}' \in \mathcal{S}} \|\mathbf{s} - \mathbf{s}'\| < \infty, \quad C_Z := \max_{i \in [n]} \int_{\mathcal{Z}} |S(z, y_i)| \mu(dz) < \infty. \tag{11}$$

A2. *For any $i \in [n]$, $z \in \mathcal{Z}$, $\theta, \theta' \in \text{int}(\Theta)^2$, we have $|p(z|y_i; \theta) - p(z|y_i; \theta')| \leq L_p \|\theta - \theta'\|$ where $\text{int}(\Theta)$ denotes the interior of Θ .*

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

A3. *For any $\mathbf{s} \in \mathcal{S}$, the function $\theta \mapsto L(\mathbf{s}, \theta) := r(\theta) + \psi(\theta) - \langle \mathbf{s} | \phi(\theta) \rangle$ admits a unique global minimum $\bar{\theta}(\mathbf{s}) \in \text{int}(\Theta)$. In addition, $J_\phi^\theta(\bar{\theta}(\mathbf{s}))$ is full rank, L_p -Lipschitz and $\bar{\theta}(\mathbf{s})$ is L_t -Lipschitz.*

We denote by $H_L^\theta(\mathbf{s}, \theta)$ the Hessian (w.r.t to θ for a given value of \mathbf{s}) of the function $\theta \mapsto L(\mathbf{s}, \theta) = r(\theta) + \psi(\theta) - \langle \mathbf{s} | \phi(\theta) \rangle$, and define $B(\mathbf{s}) := J_\phi^\theta(\bar{\theta}(\mathbf{s})) \left(H_L^\theta(\mathbf{s}, \bar{\theta}(\mathbf{s})) \right)^{-1} J_\phi^\theta(\bar{\theta}(\mathbf{s}))^\top$.

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

The class of algorithms we develop in this paper is composed of two levels where the second stage corresponds to the variance reduction trick used in [18] in order to accelerate incremental methods and reduce the variance introduced by the index sampling. The first stage is the Robbins-Monro type of update that aims at reducing the Monte Carlo noise of the quantity $\bar{s}_i(\hat{\theta}(\hat{s}^{(k)}))$ at iteration k . We denote those latter MC fluctuations terms as follows:

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

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

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

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

Lemma 1. [18] Assume A1-A4. For all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ and $i \in [n]$, we have

$$\|\bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq L_s \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq L_V \|\mathbf{s} - \mathbf{s}'\|, \quad (13)$$

where $L_s := C_Z L_p L_t$ and $L_V := v_{\max}(1 + L_s) + L_b C_S$.

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

Lemma 2. Assume A3 and A4. For all $\mathbf{s} \in \mathcal{S}$,

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2. \quad (14)$$

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

3.2 Global Convergence of Incremental Stochastic EM Algorithms

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

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

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

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

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

Theorem 1. Assume A1-A5. Consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k \geq 0} \in \mathcal{S}$ obtained with $\rho_{k+1} = 1$ for any $k \leq K_m$ where K_m is a positive integer. Let $\{\gamma_k = 1/(k^a \alpha c_1 \bar{L})\}_{k \geq 0}$, where $a \in (0, 1)$, be a sequence of stepsizes, $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\beta = c_1 \bar{L}/n$. Then:

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

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

3.3 Global Convergence of Two-Timescale Stochastic EM Algorithms

Two important intermediate Lemmas are needed in order to establish finite-time bounds for the vrTTEM and the fitTEM methods. We first derive an identity for the drift term of the vrTTEM :

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

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

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

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

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

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

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

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

185 where $P_m = \sum_{\ell=0}^{K_m-1} \gamma_\ell$ and the expectation is over the stochasticity of the algorithm. Denote
186 $\Delta V = V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_m)})$. We now state the main result regarding the vrTTEM method:

187 **Theorem 2.** *Assume A1-A5. Consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq K_m$ where
188 K_m is a positive integer. Let $\{\gamma_{k+1} = 1/(k^a \bar{L})\}_{k>0}$, where $a \in (0, 1)$, be a sequence of stepsizes,
189 $\bar{L} = \max\{L_s, L_V\}$, $\rho = \mu/(c_1 \bar{L} n^{2/3})$, $m = nc_1^2/(2\mu^2 + \mu c_1^2)$ and a constant $\mu \in (0, 1)$. Then:*

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

190 Furthermore, the fiTTEM method has the following convergence rate:

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

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

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

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

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

203 which strengthen the fact that this quantity characterizes the impact of the variance reduction tech-
204 nique introduced in our class of methods. The following Lemma characterizes this gap:

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

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

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

4 Numerical Examples

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

4.1 Gaussian Mixture Models

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

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

where $\theta := (\omega, \mu)$ with $\omega = \{\omega_m\}_{m=1}^{M-1}$ are the mixing weights with the convention $\omega_M = 1 - \sum_{m=1}^{M-1} \omega_m$ and $\mu = \{\mu_m\}_{m=1}^M$ are the means. We use the penalization $r(\theta) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \log \text{Dir}(\omega; M, \epsilon)$ where $\delta > 0$ and $\text{Dir}(\cdot; M, \epsilon)$ is the M dimensional symmetric Dirichlet distribution with concentration parameter $\epsilon > 0$. The constraint set is given by $\Theta = \{\omega_m, m = 1, \dots, M-1 : \omega_m \geq 0, \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R}, m = 1, \dots, M\}$. In the following experiments on synthetic data, we generate 30 synthetic datasets of size $n = 10^5$ from a GMM model with $M = 2$ components with two mixtures with means $\mu_1 = -\mu_2 = 0.5$. We run the EM method until convergence (to double precision) to obtain the ML estimate μ^* averaged on 50 datasets. We compare the EM, iEM, SAEM, iSAEM, vrTTEM and fitTTEM methods in terms of their precision measured by $|\mu - \mu^*|^2$. We set the stepsize of the SA-step of all method as $\gamma_k = 1/k^\alpha$ with $\alpha = 0.5$, and the stepsize ρ_k for vrTTEM and the fitTTEM to a constant stepsize equal to $1/n^{2/3}$. The number of MC samples is fixed to $M = 10$ chains. Figure 1 shows the precision $|\mu - \mu^*|^2$ for the different methods against the epoch(s) elapsed (one epoch equals n iterations). vrTTEM and fitTTEM methods outperform the other stochastic methods, supporting the benefits of our scheme.

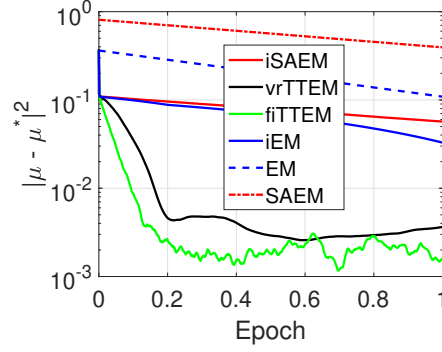


Figure 1: Precision $|\mu^{(k)} - \mu^*|^2$ per epoch

4.2 Deformable Template Model for Image Analysis

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

$$y_i(u) = I(x_u - \Phi_i(x_u, z_i)) + \varepsilon_i(u) \quad (15)$$

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

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

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

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

where we put a Gaussian prior on the latent variables, $z_i \sim \mathcal{N}(0, \Gamma)$ and $z_i \in (\mathbb{R}^{k_g})^2$. The vector of parameters we estimate is thus $\theta = (\beta, \Gamma, \sigma)$. The complete model (15) belongs to the curved exponential family, see [2], which vector of sufficient statistics for all $i \in [n]$ is defined by $S(y_i, z_i) = (\mathbf{K}_{p, z_i}^\top y_i, \mathbf{K}_{p, z_i}^\top \mathbf{K}_{p, z_i} z_i)$ where we denote $\mathbf{K}_{p, z_i} = \mathbf{K}_{p, z_i}(x_u - \phi_i(x_u, z_i), p_j)$. Then, the Two-Timescale M-step yields the following parameter

updates $\bar{\theta}(\hat{s}) = (\beta(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z), \Gamma(\hat{s}) = \hat{s}_3(z)/n, \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z)\beta(\hat{s}) - 2\beta(\hat{s})\hat{s}_1(z))$
 where $\hat{s} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via (9) in Algorithm 1.

Numerical Experiment: We apply model (15) and our Algorithm 1 to a collection of handwritten digits, called the US postal database [16], featuring $n = 1000$ (16×16)-pixel images for each class of digits from 0 to 9. The main difficulty with these data comes from the geometric dispersion within each class of digit as shown Figure 2 for digit 5. We thus ought to use our deformable template model (15) in order to account for both sources of variability: the intrinsic template to each class of digit and the small and local deformation in each observed image.



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

Figure 3 shows the resulting synthetic images for digit 5 through several epochs, for the batch method, the online SAEM, the incremental SAEM and the various TTS methods. For all methods, the initialization of the template (16) is the mean of the gray level images. In our experiments, we have chosen Gaussian kernels for both, \mathbf{K}_p and \mathbf{K}_g , defined on \mathbb{R}^2 and centered on the landmark points $\{p_k\}_{k=1}^{k_p}$ and $\{g_k\}_{k=1}^{k_g}$ with standard respective standard deviations of 0.12 and 0.3. We set $k_p = 15$ and $k_g = 6$ equidistributed landmarks points on the grid for the training procedure. Those hyperparameters are inspired by relevant studies [1, 3]. In particular, the choice of the geometric covariance, indexed by g , in such study is critical since it has a direct impact on the *sharpness* of the templates. As for the photometric hyperparameter, indexed by p , both the template and the geometry are impacted, in the sense that with a large photometric variance, the kernel centered on one landmark *spreads out* to many of its neighbors.

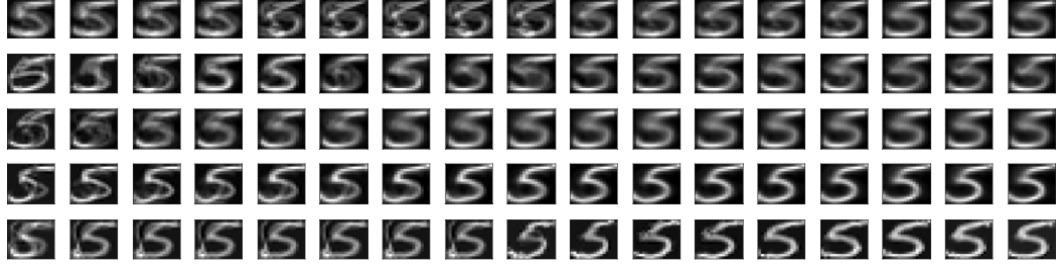


Figure 3: (USPS Digits) Estimation of the template. From top to bottom: batch, online, iSAEM, vrT-TEM and fitTEM through 7 epochs. Note that Batch method templates are replicated in-between epochs for a fair comparison with incremental variants.

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

5 Conclusion

This paper introduces a new class of two-timescale EM methods for learning latent variable models. In particular, the models dealt with in this paper belong to the curved exponential family and are possibly nonconvex. The nonconvexity of the problem is tackled using a Robbins-Monro type of update, which represents the *first level* of our class of methods. The scalability with the number of samples is performed through a variance reduced and incremental update, the *second* and last level of our newly introduced scheme. The various algorithms are interpreted as scaled gradient methods, in the space of the sufficient statistics, and our convergence results are *global*, in the sense of independence of the initial values, and *non-asymptotic*, *i.e.*, true for any random termination number. Numerical examples illustrate the benefits of our scheme on synthetic and real tasks.

References

- [1] Stéphanie Allasonnière and Estelle Kuhn. Stochastic algorithm for parameter estimation for dense deformable template mixture model. *arXiv preprint arXiv:0802.1521*, 2008.
- [2] Stéphanie Allasonnière, Yali Amit, and Alain Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.
- [3] Stéphanie Allasonnière, Estelle Kuhn, Alain Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.
- [4] Charlotte Baey, Samis Trevezas, and Paul-Henry Cournède. A non linear mixed effects model of plant growth and estimation via stochastic variants of the em algorithm. *Communications in Statistics-Theory and Methods*, 45(6):1643–1669, 2016.
- [5] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112(518):859–877, JUN 2017. ISSN 0162-1459. doi: {10.1080/01621459.2017.1285773}.
- [6] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.
- [7] Olivier Cappé. Online EM algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749, 2011.
- [8] Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 593–613, 2009.
- [9] Bradley P Carlin and Siddhartha Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484, 1995.
- [10] Arindom Chakraborty and Kalyan Das. Inferences for joint modelling of repeated ordinal scores and time to event data. *Computational and mathematical methods in medicine*, 11(3): 281–295, 2010.
- [11] Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. Stochastic expectation maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988, 2018.
- [12] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL <https://doi.org/10.1214/aos/1018031103>.
- [13] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- [14] Bradley Efron et al. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 3(6):1189–1242, 1975.
- [15] James P Hughes. Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55(2):625–629, 1999.
- [16] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

- 329 [17] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive vari-
330 ance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.
- 331 [18] Belhal Karimi, Hoi-To Wai, Éric Moulines, and Marc Lavielle. On the global convergence
332 of (fast) incremental expectation maximization methods. In *Advances in Neural Information
333 Processing Systems*, pages 2833–2843, 2019.
- 334 [19] Estelle Kuhn, Catherine Matias, and Tabea Rebafka. Properties of the stochastic approximation
335 em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.
- 336 [20] Percy Liang and Dan Klein. Online em for unsupervised models. In *Proceedings of human
337 language technologies: The 2009 annual conference of the North American chapter of the
338 association for computational linguistics*, pages 611–619, 2009.
- 339 [21] Florian Maire, Eric Moulines, and Sidonie Lefebvre. Online em for functional data, 2016.
340 URL <http://arxiv.org/abs/1604.00570>. cite arxiv:1604.00570v1.pdf.
- 341 [22] Charles E McCulloch. Maximum likelihood algorithms for generalized linear mixed models.
342 *Journal of the American statistical Association*, 92(437):162–170, 1997.
- 343 [23] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume
344 382. John Wiley & Sons, 2007.
- 345 [24] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science
346 & Business Media, 2012.
- 347 [25] Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental,
348 sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.
- 349 [26] SK Ng and GJ McLachlan. On the choice of the number of blocks with the incremental EM
350 algorithm for the fitting of normal mixtures. *Statistics and Computing*, 13(1):45–55, FEB
351 2003. ISSN 0960-3174. doi: {10.1023/A:1021987710829}.
- 352 [27] Hien D Nguyen, Florence Forbes, and Geoffrey J McLachlan. Mini-batch learning of expo-
353 nential family finite mixture models. *Statistics and Computing*, pages 1–18, 2020.
- 354 [28] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Fast incremental method for
355 nonconvex optimization. *arXiv preprint arXiv:1603.06159*, 2016.
- 356 [29] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of math-
357 ematical statistics*, pages 400–407, 1951.
- 358 [30] Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and
359 the poor man’s data augmentation algorithms. *Journal of the American statistical Association*,
360 85(411):699–704, 1990.
- 361 [31] CF Jeff Wu et al. On the convergence properties of the EM algorithm. *The Annals of statistics*,
362 11(1):95–103, 1983.

363 A Proof of Lemma 2

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

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

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

$$\begin{aligned} \nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \left(\nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{L}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \left(\nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top (\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))), \end{aligned} \quad (17)$$

367 Consider the following vector map:

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

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

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

369 The above yields

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

370 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 (16) follows directly
371 from the assumption A4. \square

372 B Proof of Theorem 1

373 Beforehand, We present two intermediary Lemmas important for the analysis of the incremental
374 update of the iSAEM algorithm. The first one gives a characterization of the quantity $\mathbb{E}[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}]$:
375

376 **Lemma.** Assume A1. The update (1) is equivalent to the following update on the resulting statistics
377

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

378 Also:

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

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

380 **Proof** From update (1), we have:

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

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

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

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

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

383 The following equalities:

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

384 concludes the proof of the Lemma. \square

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

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

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

387 **Proof** Applying the iSAEM update yields:

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

388 The last expectation can be further bounded by

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

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

390 \square

391 **Theorem.** Assume A1-A5. Consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ obtained with $\rho_{k+1} = 1$
392 for any $k \leq K_m$ where K_m is a positive integer. Let $\{\gamma_k = 1/(k^a \alpha c_1 \bar{L})\}_{k>0}$, where $a \in (0, 1)$, be a
393 sequence of stepsizes, $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\beta = c_1 \bar{L}/n$. Then:

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

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

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

395 Taking the expectation on both sides yields:

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

396 Using Lemma 3, we obtain:

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

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

398 $a_k = \gamma_{k+1} \left(v_{\min} - v_{\max}^2 \frac{\beta(n-1)+1}{2n} \right)$ and

$$\begin{aligned}
a_k \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] &\leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] \\
&\quad + \frac{\gamma_{k+1}(1 - \frac{1}{n})}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] + \frac{\gamma_{k+1}}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \quad (18)
\end{aligned}$$

399 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 (18):

$$\begin{aligned}
& (a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] \\
&\leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] \\
&\quad + \gamma_{k+1} \left(\frac{1}{2\beta} \left(1 - \frac{1}{n}\right) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] \\
&\quad + \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
&\quad + \frac{\gamma_{k+1}^2 L_V L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{s}^{(k)} - \hat{s}^{(\tau_i^k)} \right\|^2 \right] \quad (19)
\end{aligned}$$

400 Next, we observe that

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

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

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

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

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

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

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

404 Let us define

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

405 From the above, we get

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

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

407 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 6$, $\alpha \geq 8$, we observe that

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

408 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 $\Lambda_{(k+1)} =$

409 $\frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1}L_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

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

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

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

411 We recall (19) where we have summed on both sides from $k = 0$ to $k = K_m - 1$:

$$\begin{aligned}
& \sum_{k=0}^{K_m-1} (a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \\
&\leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K)}) \right] \\
&+ \sum_{k=0}^{K_m-1} \gamma_{k+1} \left(\frac{1}{2\beta} \left(1 - \frac{1}{n} \right) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \\
&+ \sum_{k=0}^{K_m-1} \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
&+ \sum_{k=0}^{K_m-1} \frac{\gamma_{k+1}^2 L_V L_s^2}{n^2} \Delta^{(k)}
\end{aligned} \tag{21}$$

412 Plugging (20) into (21) results in:

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

413 where

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

414 and

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

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

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

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

418 C Proofs of Auxiliary Lemmas

419 C.1 Proof of Lemma 4 and Lemma 5

420 **Lemma.** For any $k \geq 0$ and consider the vrTTEM update in (2) with $\rho_k = \rho$, it holds for all $k > 0$

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

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

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

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

424 We observe, using the identity (22), that

$$\mathbb{E}[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \right\|^2] \leq 2\rho^2 \mathbb{E}[\left\| \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} \right\|^2] + 2\rho^2 \mathbb{E}[\left\| \bar{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)} \right\|^2] + 2(1-\rho)^2 \mathbb{E}[\left\| \hat{\mathbf{s}}^{(\ell(k))} - \tilde{S}^{(k)} \right\|^2] \tag{23}$$

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

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

426 where (a) uses the variance inequality and (b) uses Lemma 1. Substituting into (23) proves the
427 lemma. \square

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

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

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:

$$\begin{aligned}
 \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}) \\
 &= -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - (1 - \rho)\tilde{\mathbf{S}}^{(k)} - \rho\mathbf{S}^{(k+1)}) \\
 &= -\gamma_{k+1}\left((1 - \rho)\left[\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right] + \rho\left[\hat{\mathbf{s}}^{(k)} - \mathbf{S}^{(k+1)}\right]\right) \\
 &= -\gamma_{k+1}\left((1 - \rho)\left[\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right] + \rho\left[\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{S}}^{(k)} - (\tilde{\mathbf{S}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(t_{i_k}^k)})\right]\right)
 \end{aligned} \tag{24}$$

431 We observe, using the identity (24), that

$$\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2] \leq 2\rho^2\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{S}}^{(k)}\|^2] + 2\rho^2\mathbb{E}[\|\bar{\mathbf{S}}^{(k)} - \mathbf{S}^{(k+1)}\|^2] + 2(1 - \rho)^2\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\|^2] \tag{25}$$

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

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

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

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

434 Substituting into (25) proves the lemma. \square

435 C.2 Proof of Lemma 6

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

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

438 where $\mathbf{S}^{(k)}$ is defined either by (3) (fiTTEM) or (2) (vrTTEM)

439 **Proof** We begin by writing the two-timescale update:

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

440 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 (3). Denote $\delta^{(k+1)} = \hat{\mathbf{s}}^{(k+1)} - \tilde{\mathbf{S}}^{(k+1)}$.
 441 Then from (26), doing the subtraction of both equations yields:

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

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

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

443 \square

444 **C.3 Additional Intermediary Result**

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

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

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

449 **Proof** Using the fitTEM update $\tilde{S}^{(k+1)} = (1 - \rho)\tilde{S}^{(k)} + \rho\mathcal{S}^{(k+1)}$ where $\mathcal{S}^{(k+1)} = \bar{\mathcal{S}}^{(k)} + (\tilde{S}_{i_k}^{(k)} -$
 450 $\tilde{S}_{i_k}^{(t_k^k)})$ leads to the following decomposition:

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

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

452 *Important Note:* Note that $\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_k^k)}$ is not equal to $\eta_{i_k}^{(k+1)}$, defined in (12), which is the gap
 453 *between the MC approximation and the expected statistics. Indeed $\tilde{S}_{i_k}^{(t_k^k)}$ is not computed under the*
 454 *same model as $\bar{\mathbf{s}}_{i_k}^{(k)}$.* □

455 D Proof of Theorem 2

456 **Theorem.** Assume A1-A5. Consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq K_m$ where
 457 K_m is a positive integer. Let $\{\gamma_{k+1} = 1/(k^a \bar{L})\}_{k>0}$, where $a \in (0, 1)$, be a sequence of stepsizes,
 458 $\bar{L} = \max\{L_s, L_V\}$, $\rho = \mu/(c_1 \bar{L} n^{2/3})$, $m = nc_1^2/(2\mu^2 + \mu c_1^2)$ and a constant $\mu \in (0, 1)$. Then:

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

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

$$\begin{aligned} V(\hat{\mathbf{s}}^{(k+1)}) &\leq V(\hat{\mathbf{s}}^{(k)}) + \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{L_V}{2} \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (27)$$

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

$$\begin{aligned} &\mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] \\ &\stackrel{(a)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1}(1-\rho) \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1}\rho \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|H_{k+1}\|^2] \\ &\stackrel{(b)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1}\rho \mathbb{E}[\langle \mathbf{h}_k | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1}(1-\rho) \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad - \gamma_{k+1}\rho \mathbb{E}[\langle \eta_{i_k}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|H_{k+1}\|^2] \\ &\stackrel{(c)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - (\gamma_{k+1}\rho v_{\min} + \gamma_{k+1}v_{\max}^2) \mathbb{E}[\|\mathbf{h}_k\|^2] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|H_{k+1}\|^2] \\ &\quad - \gamma_{k+1}\rho \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] - \gamma_{k+1}(1-\rho) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \end{aligned} \quad (28)$$

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

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

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

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

466 Consider the following sequence

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

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

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

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

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

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

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

470 And using Lemma 4 we obtain:

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

471 Rearranging the terms yields:

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

472 where

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

473 This leads, using Lemma 2, that for any γ_{k+1} , ρ and β such that $\rho v_{\min} + v_{\max}^2 -$
474 $\gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2) > 0$,

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

475 We first remark that

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

476 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
 477 $\{\gamma_{k+1}\}$ any sequence of decreasing stepsizes in $(0, 1)$, it can be shown that there exists $\mu \in (0, 1)$,
 478 such that the following lower bound holds

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

479 where the simplification in (a) is due to

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

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

481 Finally, these results yield:

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

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

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

484 This concludes our proof.

485

□

486 E Proof of Theorem 3

487 **Theorem.** Assume A1-A5. Consider the fitTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq K_m$ where
 488 K_m be a positive integer. Let $\{\gamma_{k+1} = 1/(k^a \alpha c_1 \bar{L})\}_{k>0}$, where $a \in (0, 1)$, be a sequence of
 489 positive stepsizes, $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\beta = 1/(\alpha n)$, $\rho = 1/(\alpha c_1 \bar{L} n^{2/3})$
 490 and $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$. Then:

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

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

$$\begin{aligned} V(\hat{\mathbf{s}}^{(k+1)}) &\leq V(\hat{\mathbf{s}}^{(k)}) + \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{L_V}{2} \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (29)$$

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

$$\mathbb{E} \left[(\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}] \right] = 0 \quad (30)$$

494 we have:

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

495 where $\xi^{(k+1)} = \mathbb{E}[\|\mathbb{E}[\eta_{i_k}^{(k+1)} | \mathcal{F}_k]\|^2]$. **Bounding** $\mathbb{E}[\|\mathbf{H}_{k+1}\|^2]$ Using Lemma 5, we obtain:

$$\begin{aligned} &\gamma_{k+1} (v_{\min} \rho + v_{\max}^2 - \gamma_{k+1} \rho^2 L_V) \mathbb{E}[\|\mathbf{h}_k\|^2] \\ &\leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(k)}) - V(\hat{\mathbf{s}}^{(k+1)}) \right] + \tilde{\xi}^{(k+1)} + \left((1 - \rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1} (1 - \rho)^2}{2} \right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &\quad - \frac{\gamma_{k+1}^2 L_V \rho^2 L_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \end{aligned} \quad (31)$$

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

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

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

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

498 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
 499 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,
 500 for any $\beta > 0$, it holds

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

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

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

502 Subsequently, we have

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

503 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:

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

504 Let us define

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

505 From the above, we get

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

506 Setting $c_1 = v_{\min}^{-1}$, $\alpha = \max\{2, 1+2v_{\min}\}$, $\bar{L} = \max\{\mathbf{L}_s, \mathbf{L}_V\}$, $\gamma_{k+1} = \frac{1}{k}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \bar{L} n^{2/3}}$,
507 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$, we observe that

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

508 which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2 \rho^2 \mathbf{L}_s^2 \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} = \frac{1}{n} -$
509 $\gamma_{k+1}\beta - \gamma_{k+1}^2 \rho^2 \mathbf{L}_s^2$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

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

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

511 Summing on both sides over $k = 0$ to $k = K_m - 1$ yields:

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

512 We recall (31) where we have summed on both sides from $k = 0$ to $k = K_m - 1$:

$$\begin{aligned} &\mathbb{E}[V(\hat{\mathbf{s}}^{(K_m)}) - V(\hat{\mathbf{s}}^{(0)})] \\ &\leq \sum_{k=0}^{K_m-1} \left\{ \gamma_{k+1}(-(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}\rho^2 \mathbf{L}_V) \mathbb{E}[\|\mathbf{h}_k\|^2] + \gamma^2 \mathbf{L}_V \rho^2 \mathbf{L}_s^2 \Delta^{(k)} \right\} \\ &\quad + \sum_{k=0}^{K_m-1} \left\{ \tilde{\xi}^{(k+1)} + \left((1-\rho)^2 \gamma_{k+1}^2 \mathbf{L}_V - \frac{\gamma_{k+1}(1-\rho)^2}{2} \right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\|^2] \right\} \\ &\leq \sum_{k=0}^{K_m-1} \left\{ \left[-\gamma_{k+1}(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}^2 \rho^2 \mathbf{L}_V + \frac{\rho^2 \gamma_{k+1}^2 \mathbf{L}_V \mathbf{L}_s^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \right] \mathbb{E}[\|\mathbf{h}_k\|^2] \right\} \\ &\quad + \sum_{k=0}^{K_m-1} \Xi^{(k+1)} + \sum_{k=0}^{K_m-1} \Gamma^{(k+1)} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\|^2] \end{aligned} \tag{33}$$

where

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

and

$$\Gamma^{(k+1)} = \left((1-\rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1}(1-\rho)^2}{2} \right) + \frac{\gamma_{k+1}^3 L_V \rho^2 L_s^2 (1-\rho)^2 \left(2\gamma_{k+1} + \frac{1}{\beta} \right)}{\Lambda_{(k+1)}}$$

513 We now analyse the following quantity

$$\begin{aligned} & -\gamma_{k+1}(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}^2 \rho^2 L_V + \frac{\rho^2 \gamma_{k+1}^2 L_V L_s^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\Lambda_{(k+1)}} \\ & = \gamma_{k+1} \left[-(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1} \rho^2 L_V + \frac{\rho^2 \gamma_{k+1} L_V L_s^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\Lambda_{(k+1)}} \right] \end{aligned} \quad (34)$$

514 Furthermore, we recall that $c_1 = v_{\min}^{-1}$, $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\gamma_{k+1} = \frac{1}{k}$,
515 $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \bar{L} n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$. Then,

$$\begin{aligned} & \gamma_{k+1} \rho^2 L_V + \frac{\rho^2 \gamma_{k+1} L_V L_s^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2 \rho^2 L_s^2} \\ & \leq \frac{1}{k\alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\bar{L}(k\alpha^2 c_1^2 n^{4/3})^{-1} \left(\frac{2}{k^2 \alpha^2 c_1^2 \bar{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \bar{L}^2 n^{1/3}} \right)}{\frac{1}{n} - \frac{1}{k\alpha n} - \frac{1}{k^2 \alpha^2 c_1^2 n^{4/3}}} \\ & = \frac{1}{k\alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\bar{L} \left(\frac{2}{k^2 \alpha^2 c_1^2 \bar{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \bar{L}^2 n^{1/3}} \right)}{(k\alpha c_1 n^{1/3})(k\alpha - 1)c_1 - 1} \\ & \stackrel{(a)}{\leq} \frac{1}{k\alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\frac{1}{k\alpha c_1^2 \bar{L} n^{1/3}} \left(\frac{2}{k\alpha n} + 1 \right)}{2(\alpha c_1 n^{1/3}) - 1} \\ & \leq \frac{1}{k^2 \alpha c_1^2 \bar{L} n^{4/3}} + \frac{1}{4k\alpha^2 c_1^3 \bar{L} n^{2/3}} \\ & \leq \frac{3/4}{\alpha c_1^2 \bar{L} n^{2/3}} \end{aligned} \quad (35)$$

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

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

which yields that

$$\left[-(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1} \rho^2 L_V + \frac{\rho^2 \gamma_{k+1} L_V L_s^2 \left(2\gamma_{k+1}^2 \rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta} \right)}{\Lambda_{(k+1)}} \right] \leq -\frac{1/4}{\alpha c_1^2 \bar{L} n^{2/3}}$$

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

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

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

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

519

□

520 F Practical Implementations of Two-Timescale EM Methods

521 F.1 Application on GMM

522 F.1.1 Explicit Updates

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

$$\Theta = \Delta^M \times \mathbb{R}^M.$$

524 Using the partition of the sufficient statistics as $S(y_i, z_i) =$
 525 $(S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i)^\top)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, the partition
 526 $\phi(\theta) = (\phi^{(1)}(\theta)^\top, \phi^{(2)}(\theta)^\top, \phi^{(3)}(\theta)^\top)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$ and the fact that
 527 $\mathbb{1}_{\{M\}}(z_i) = 1 - \sum_{m=1}^{M-1} \mathbb{1}_{\{m\}}(z_i)$, the complete data log-likelihood can be expressed as in
 528 (2) with

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

529 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 $m \in \llbracket 1, M \rrbracket$, $j \in \llbracket 1, 3 \rrbracket$,
 530 $s_m^{(j)} = n^{-1} \sum_{i=1}^n s_{i,m}^{(j)}$. Consider the following latent sample used to compute an approximation of
 531 the conditional expected value $\mathbb{E}_\theta[\mathbb{1}_{\{z_i=m\}} | y = y_i]$:

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

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

533 In particular, given iteration $k+1$, the computation of the approximated quantity $\tilde{S}_{i_k}^{(k)}$ during
 534 Incremental-step updates, see (8) can be written as

$$\tilde{S}_{i_k}^{(k)} = \left(\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1})y_{i_k}, \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\tilde{s}_{i_k}^{(3)}(\theta^{(k)})} \right)^\top. \quad (38)$$

535 Recall that we have used the following regularizer:

$$\mathbf{r}(\theta) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log(1 - \sum_{m=1}^{M-1} \omega_m), \quad (39)$$

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

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

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

538 F.1.2 Model Assumptions (GMM example)

539 We use the GMM example to illustrate the required assumptions.

540 Many practical models can satisfy the compactness of the sets as in Assumption A1. For instance,
 541 the GMM example satisfies (11) as the sufficient statistics are composed of indicator functions and
 542 observations as defined Section F.1 Equation (36).

Assumptions A2 and A3 are standard for the curved exponential family models. For GMM, the following (strongly convex) regularization $r(\theta)$ ensures A3:

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

543 since it ensures $\theta^{(k)}$ is unique and lies in $\text{int}(\Delta^M) \times \mathbb{R}^M$. We remark that for A2, it is possible to
 544 define the Lipschitz constant L_p independently for each data y_i to yield a refined characterization.

545 Again, A4 is satisfied by practical models. For GMM, it can be verified by deriving the closed form
 546 expression for $B(s)$ and using A1.

547 Under A1 and A3, we have $\|\hat{s}^{(k)}\| < \infty$ since S is compact and $\hat{\theta}^{(k)} \in \text{int}(\Theta)$ for any $k \geq 0$ which
 548 thus ensure that the EM methods operate in a closed set throughout the optimization process.

549 F.1.3 Algorithms updates

550 In the sequel, recall that, for all $i \in [n]$ and iteration k , the computed statistic $\tilde{S}_{i_k}^{(k)}$ is defined by (38).
 551 At iteration k , the several E-steps defined by (1) or (2) and (3) leads to the definition of the quantity
 552 $\hat{s}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{s}^{(0)} = n^{-1} \sum_{i=1}^n \bar{s}_i^{(0)}$, those
 553 E-steps break down as follows:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

565 F.2 Deformable Template Model for Image Analysis

566 F.2.1 Model and Updates

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

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

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

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

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

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

571 where $\hat{\mathbf{s}} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via the **SA-step** (7) and using the
 572 MC approximation of the sufficient statistics $(S_1(z), S_2(z), S_3(z))$ defined in (41).

573 F.2.2 Numerical Applications

574 For the inference of the template, we use the Matlab code (online SAEM) used in [21] and implement
 575 our own batch, incremental, Variance reduced and Fast Incremental variants. The hyperparameters
 576 are kept the same and reads as follows $M = 400$, $\gamma_k = 1/k^{0.6}$ and $p = 16$. The number of
 577 landmarks for the template is $k_p = 15$ points and for the deformation $k_g = 6$ points. Both have
 578 Gaussian kernels with respectively standard deviation of 0.12 and 0.3. The standard deviation of the
 579 measurement errors is set to 0.1.

580 For the simulation part, we use the Carlin and Chib MCMC procedure, see [9]. Refer to [21] for
 581 more details.

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

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

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

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

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

Lognormal distributions are used for the four PK parameters:

$$\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega_{T^{\text{lag}}}^2), \log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2), \quad (44)$$

$$\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2), \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2). \quad (45)$$

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

$$S_1(z) = \frac{1}{n} \sum_{i=1}^n z_i, \quad S_2(z) = \frac{1}{n} \sum_{i=1}^n z_i^\top z_i, \quad S_3(z) = \frac{1}{n} \sum_{i=1}^n (y_i - f(t_i, z_i))^2 \quad (46)$$

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

$$\bar{\theta}(s) = \begin{pmatrix} \hat{s}_1^{(k+1)} \\ \hat{s}_2^{(k+1)} - \hat{s}_1^{(k+1)} (\hat{s}_1^{(k+1)})^\top \\ \hat{s}_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \overline{z_{\text{pop}}}(\hat{s}^{(k+1)}) \\ \overline{\omega_z}(\hat{s}^{(k+1)}) \\ \overline{\sigma}(\hat{s}^{(k+1)}) \end{pmatrix}. \quad (47)$$

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

Metropolis Hastings algorithm. During the simulation step of the MISSO method, the sampling from the target distribution $\pi(z_i, \theta) := p(z_i | y_i, \theta)$ is performed using a Metropolis Hastings (MH) algorithm [24] with proposal distribution $q(z_i, \delta)$ where $\theta = (z_{\text{pop}}, \omega_z)$ and δ is the vector of parameters of the proposal distribution. Commonly they parameterize a Gaussian proposal. The MH algorithm is summarized in 2.

Algorithm 2 MH algorithm

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

614 **Monte Carlo study:** We conduct a Monte Carlo study to showcase the benefits of our scheme. $M =$
615 50 datasets have been simulated using the following PK parameters values: $T_{\text{pop}}^{\text{lag}} = 1$, $ka_{\text{pop}} = 1$,
616 $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$, $\omega_{T^{\text{lag}}} = 0.4$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. We define
617 the mean square distance over the M replicates $E_k(\ell) = \frac{1}{M} \sum_{m=1}^M \left(\theta_k^{(m)}(\ell) - \theta^* \right)^2$ and plot it
618 against the epochs (passes over the data) Figure 4. Note that the MC-step (5) is performed using a
619 Metropolis Hastings procedure since the posterior distribution under the model θ noted $p(z_i | y_i, \theta)$
620 is intractable due to the nonlinearity of the model (43). Figure 4 shows clear advantage of variance
621 reduced methods (vrTTEM and fitTEM) avoiding the twists and turns displayed by the incremental
622 and the batch methods.

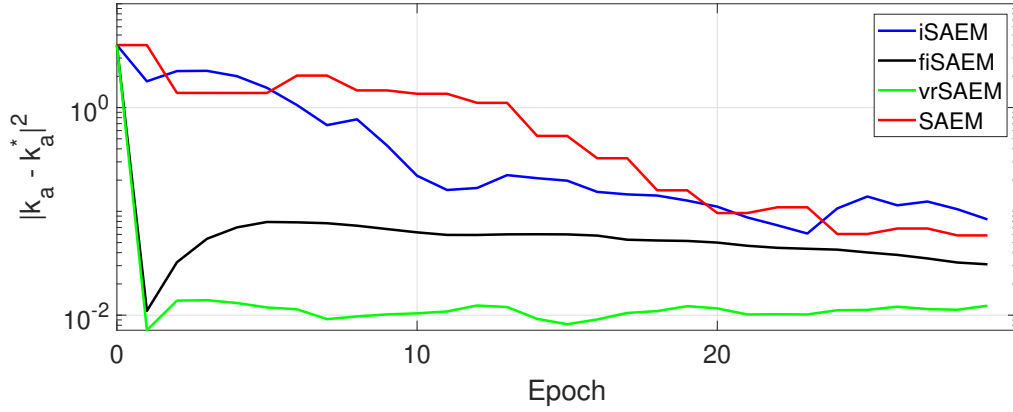


Figure 4: Precision $|ka^{(k)} - ka^*|^2$ per epoch