
Fast Two-Timescale Stochastic EM Algorithms

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The Expectation-Maximization (EM) algorithm is a popular choice for learning
2 latent variable models. Variants of the EM have been initially introduced by [21],
3 using incremental updates to scale to large datasets, and by [25, 9], using Monte
4 Carlo (MC) approximations to bypass the intractable conditional expectation of
5 the latent data for most nonconvex models. In this paper, we propose a general
6 class of methods called Two-Timescale EM Methods based on a two-stage ap-
7 proach of stochastic updates to tackle an essential nonconvex optimization task
8 for latent variable models. We motivate the choice of a double dynamic by invoking
9 the variance reduction virtue of each stage of the method on both sources of
10 noise: the index sampling for the incremental update and the MC approximation.
11 We establish finite-time and global convergence bounds for nonconvex objective
12 functions. Numerical applications are also presented to illustrate our findings.

1 Introduction

14 Learning latent variable models is critical for modern machine learning problems, see (e.g.,) [19]
15 for references. We formulate the training of such model as an empirical risk minimization problem:

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

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

19 In the latent variable model, $g(y_i; \theta)$, is the marginal of the complete data likelihood defined as
20 $f(z_i, y_i; \theta)$, i.e., $g(y_i; \theta) = \int_{\mathbb{Z}} f(z_i, y_i; \theta) \mu(dz_i)$, where $\{z_i\}_{i=1}^n$ are the latent variables. In this
21 paper, we make the assumption of a complete model belonging to the curved exponential family [11]:

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

23 where $\psi(\theta)$, $h(z_i, y_i)$ are scalar functions, $\phi(\theta) \in \mathbb{R}^k$ is a vector function, and $\{S(z_i, y_i) \in \mathbb{R}^k\}_{i=1}^n$
24 is the vector of sufficient statistics of the complete model. Full batch EM [10] is the method of
25 reference for that type of task and is a two steps procedure. The **E-step** amounts to computing the
26 conditional expectation of the complete data sufficient statistics,

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

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

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

28 Two caveats of this method are the following: (a) with the explosion of data, the first step of the EM
29 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, [21] and [5] 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 [22, 16, 4]. Some improvements of those methods have been provided and analyzed, globally and in finite-time, in [14] 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 [25] where a MC approximation for this expectation is computed. A variant of that algorithm is the Stochastic Approximation of the EM (SAEM) in [9] leveraging the power of Robbins-Monro update [24] 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 [18, 12, 3] or to do inference for joint modeling of time to event data coming from clinical trials in [7], among other applications. Recently, an incremental variant of the SAEM was proposed in [15] showing positive empirical results but its analysis is limited to asymptotic consideration. Gradient-based methods have been developed and analyzed in [26] but they remain out of the scope of this paper as they tackle the high-dimensionality issue.

Contributions: This paper *introduces* and *analyzes* a new class of methods which purpose is to update two proxies for 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 main 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 term. For all $i \in \llbracket 1, n \rrbracket$, draw $\{z_{i,m} \sim p(z_i | y_i; \theta)\}_{m \in \llbracket 1, M \rrbracket}$ samples and compute the MC integration \tilde{s} of the quantity $\bar{s}(\theta)$ (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).

77 An alternative to that stochastic algorithm is to use a Robbins-Monro (RM) type of update. We
 78 denote, at iteration k , the following quantity

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

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

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

80 where $\{\gamma_k\}_{k>1} \in (0, 1)$ is a sequence of decreasing step sizes to ensure asymptotic convergence.
 81 This is called the Stochastic Approximation of the EM (SAEM) and has been shown to converge to
 82 a maximum likelihood of the observations under very general conditions [9]. In simple scenarios,
 83 the samples $\{z_{i,m}\}_{m=0}^{M-1}$ are conditionally independent and identically distributed with distribution
 84 $p(z_i, \theta)$. Nevertheless, in most cases, since the loss function between the observed data y_i and the
 85 latent variable z_i can be nonconvex, sampling exactly from this distribution is not an option and the
 86 MC batch is sampled by Markov Chain Monte Carlo (MCMC) algorithm.

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

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

99 2.2 Incremental and Two-Stage Stochastic EM Methods

100 Efficient strategies to scale to large datasets include incremental [21] and variance reduced [8] meth-
 101 ods. We will explicit a general update that covers those latter variants and that represents the *second*
 102 *level* of our algorithm, namely the incremental update of the noisy statistics $\tilde{S}^{(k+1)}$ in the **SA-Step**:

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

104 Note that $\{\rho_k\}_{k>1} \in (0, 1)$ is a sequence of step sizes, $\mathcal{S}^{(k)}$ is a proxy for $\tilde{S}^{(k)}$. If the stepsize
 105 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
 106 we recover the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\mathcal{S}^{(k)} = \tilde{S}^{(k)}$, then we recover the
 107 MCEM [25]. For all methods, we define a random index drawn at iteration k , noted $i_k \in \llbracket 1, n \rrbracket$,
 108 and $\tau_i^k = \max\{k' : i_{k'} = i, k' < k\}$ as the iteration index where $i \in \llbracket 1, n \rrbracket$ is last drawn prior
 109 to iteration k . The proposed fitTEM method draws *two* indices *independently* and uniformly as
 110 $i_k, j_k \in \llbracket 1, n \rrbracket$. Thus, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the
 111 sample $j \in \llbracket 1, n \rrbracket$ is last drawn as j_k prior to iteration k in addition to τ_i^k which was defined *w.r.t.* i_k .

$$\text{iSAEM} \quad \mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} + n^{-1} (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)}) \quad (9)$$

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

$$\text{fitTEM} \quad \mathcal{S}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}), \quad \overline{\mathcal{S}}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + n^{-1} (\tilde{S}_{j_k}^{(k)} - \tilde{S}_{j_k}^{(t_{j_k}^k)}) \quad (11)$$

112 where $\tilde{S}_{i_k}^{(k)} = \frac{1}{M_k} \sum_{m=1}^{M_k} S(z_{i_k,m}^{(k)}, y_{i_k})$ and $z_{i_k,m}^{(k)} \sim p(z_{i_k} | y_{i_k}; \theta^{(k)})$. The stepsize is set to $\rho_{k+1} =$
 113 1 for the iSAEM method and we initialize with $\mathcal{S}^{(0)} = \tilde{S}^{(0)}$; $\rho_{k+1} = \rho$ is constant for the vrTTEM

114 and fitTEM methods. Note that we initialize as follows $\bar{\mathcal{S}}^{(0)} = \tilde{\mathcal{S}}^{(0)}$ for the fitTEM which can be
 115 seen as a slightly modified version of SAGA inspired by [23]. For vrTEM we set an epoch size of
 116 m and define $\ell(k) := m \lfloor k/m \rfloor$ as the first iteration number in the epoch that iteration k is in.

117 **Two-Timescale Stochastic EM methods:** We now introduce the general method derived using the
 118 two variance reduction techniques described above. Algorithm 1 leverages both levels (7) and (8) in
 119 order to output a vector of fitted parameters $\hat{\theta}^{(K)}$ where K is a randomly chosen termination point.

Algorithm 1 Two-Timescale Stochastic EM methods.

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

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_m-1} \gamma_\ell} = \frac{\gamma_k}{P_m}. \quad (12)$$

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

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

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

120 The update in (13) is said to have two-timescale property as the step sizes satisfy $\lim_{k \rightarrow \infty} \gamma_k / \rho_k < 1$
 121 such that $\tilde{\mathcal{S}}^{(k+1)}$ is updated at a faster time-scale, determined by ρ_{k+1} , than $\hat{\mathbf{s}}^{(k+1)}$, determined by
 122 γ_{k+1} . The next section introduces the main results of this paper and establishes global and finite-
 123 time bounds for the three different updates of our scheme.

124 3 Finite Time Analysis of the Two-Timescale Scheme

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

$$\min_{\mathbf{s} \in \mathcal{S}} V(\mathbf{s}) := \bar{L}(\bar{\theta}(\mathbf{s})) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(\bar{\theta}(\mathbf{s})) + \mathbf{r}(\bar{\theta}(\mathbf{s})), \quad (14)$$

127 that we propose to study in this article.

128 3.1 Assumptions and Intermediate Lemmas

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

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

$$C_S := \max_{\mathbf{s}, \mathbf{s}' \in \mathcal{S}} \|\mathbf{s} - \mathbf{s}'\| < \infty, \quad C_Z := \max_{i \in \llbracket 1, n \rrbracket} \int_{\mathcal{Z}} |S(z, y_i)| \mu(dz) < \infty. \quad (15)$$

131 **H2.** *For any $i \in \llbracket 1, n \rrbracket, z \in \mathcal{Z}, \theta, \theta' \in \text{int}(\Theta)^2$, we have $|p(z|y_i; \theta) - p(z|y_i; \theta')| \leq L_p \|\theta - \theta'\|$*
 132 *where $\text{int}(\Theta)$ denotes the interior of Θ .*

133 We also recall from the introduction that we consider curved exponential family models with:

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

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

H4. 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 [14] 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 $\bar{s}_i(\hat{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(r)}))$ at iteration r . We denote those latter MC fluctuations terms as follows:

$$\eta_i^{(r)} := \tilde{S}_i^{(r)} - \bar{s}_i(\vartheta^{(r)}) \quad \text{for all } i \in \llbracket 1, n \rrbracket, r > 0 \quad \text{and} \quad \vartheta \in \Theta. \quad (16)$$

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

H5. There exist a positive sequence of MC batch size $\{M_r\}_{r>0}$ and constants (c, c_η) such that for all $k > 0$, $i \in \llbracket 1, n \rrbracket$ and $\vartheta \in \Theta$:

$$\mathbb{E}[\|\eta_i^{(r)}\|^2] \leq \frac{c_\eta}{M_r} \quad \text{and} \quad \mathbb{E}[\|\mathbb{E}[\eta_i^{(r)} | \mathcal{F}_r]\|^2] \leq \frac{c}{M_r}. \quad (17)$$

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

Lemma 1. [14] Assume H1-H4. For all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ and $i \in \llbracket 1, n \rrbracket$, we have

$$\|\bar{s}_i(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \bar{s}_i(\bar{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq L_S \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq L_V \|\mathbf{s} - \mathbf{s}'\|, \quad (18)$$

where $L_S := C_Z L_p L_\theta$ 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 H3, H4. For all $\mathbf{s} \in \mathcal{S}$,

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2. \quad (19)$$

3.2 Global Convergence of Incremental and Two-Timescale Stochastic EM Algorithms

We present in this section a finite-time and global (independent of the initialization) analysis of both the incremental and two-timescale variants of the Stochastic Approximation of the EM algorithm.

The following result for the iSAEM algorithm is derived under the control of the Monte Carlo fluctuations as described by Assumption H5 and is built upon an intermediary Lemma, detailed in the supplementary material, characterizing the quantity of interest $(\hat{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)})$. Typically, the controls exhibited above are of interest when the number of MC samples M_k increase with k .

Theorem 1. Assume H1-H5. Let K_m be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = 1$ for any $k > 0$. We also set $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_S, L_V\}$, $\gamma_{k+1} = \frac{1}{k^\alpha \alpha c_1 \bar{L}}$ where $a \in (0, 1)$, $\beta = \frac{c_1 \bar{L}}{n}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_m$, then it holds:

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

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 3. Consider the vrTTEM update in (10) with $\rho_k = \rho$, it holds for all $k > 0$

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

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

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

172 **Lemma 4.** Consider the fitTEM update in (11) with $\rho_k = \rho$. It holds for all $k > 0$ that

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

173 Recalling that K is an independent discrete r.v. drawn from $\{1, \dots, K_m\}$ with distribution
174 $\{\gamma_k/P_m, 0 \leq k \leq K_m - 1\}$, as in (12), then the convergence criterion used in our study reads
175

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

176 where the expectation is over the stochasticity of the algorithm.

177 Denote $\Delta V = V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_m)})$. We now state the main result regarding the vrTTEM method:

178 **Theorem 2.** Assume H1-H5. Let K_m be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of
179 positive step sizes and consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for
180 any $k \leq K_m$. Setting $\bar{L} = \max\{L_s, L_V\}$, $\rho = \mu/(c_1 \bar{L} n^{2/3})$, $m = nc_1^2/(2\mu^2 + \mu c_1^2)$, a constant
181 $\mu \in (0, 1)$, $\gamma_{k+1} = 1/(k^a \bar{L})$ where $a \in (0, 1)$, it holds:

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

182 Furthermore, the fitTEM method has the following convergence rate:

183 **Theorem 3.** Assume H1-H5. Let K_m be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of
184 positive step sizes and consider the fitTEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any
185 $k \leq K_m$. Setting $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\beta = 1/(\alpha n)$, $\rho = 1/(\alpha c_1 \bar{L} n^{2/3})$,
186 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$ and $\gamma_{k+1} = 1/(k^a \alpha c_1 \bar{L})$ where $a \in (0, 1)$, it holds:

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

187 Note that in those two bounds, the quantities $\tilde{\eta}^{(k+1)}$ and $\Xi^{(k+1)}$ depend only on the MC fluctuations
188 $\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2]$ and some constants. While Theorem 1 suffers only from the MC noise created by the la-
189 tent data sampling step, Theorem 2 and Theorem 3 exhibit in their convergence bounds *two different*
190 *phases*. The upper bounds display a *bias term* due to the initial conditions, i.e., the term ΔV , and a
191 *double dynamic burden* exemplified by the term $\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]$.

192 Indeed, the following remarks are worth doing on this quantity. (i) This term is the price we pay for
193 the two-timescale dynamic and corresponds to the gap between the two *asynchronous* updates (one
194 on $\hat{\mathbf{s}}^{(k)}$ and the other on $\tilde{S}^{(k)}$). (ii) It is readily understood that if $\rho = 1$, i.e., there is no variance
195 reduction, then for any $k > 0$

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

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

198 **Lemma 5.** Considering a decreasing stepsize $\gamma_k \in (0, 1)$ and a constant $\rho \in (0, 1)$, we have

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

199 where $\mathcal{S}^{(k)}$ is defined either by (10) (vrTTEM) or (11) (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 $z_i \in \llbracket M \rrbracket$ 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 fitTEM 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 stepsizes ρ_k for vrTTEM and the fitTEM to a constant stepsize equal to $1/n^{2/3}$. The number of MC samples is fixed to $M = 10$ chains. Figure 1 shows the precision $|\mu - \mu^*|^2$ for the different methods against the epoch(s) elapsed (one epoch equals n iterations). vrTTEM and fitTEM methods outperform the other stochastic methods, supporting the benefits of our scheme.

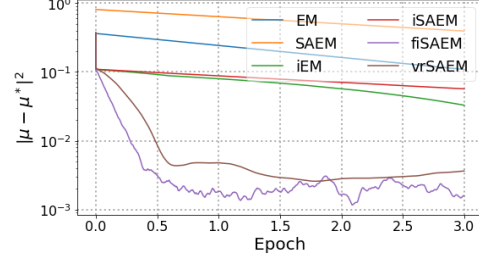


Figure 1: Precision $|\mu^{(k)} - \mu^*|^2$ per epoch

4.2 Deformable Template Model for Image Analysis

Let $(y_i, i \in \llbracket 1, n \rrbracket)$ 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:

$$y_i(u) = I(x_u - \Phi_i(x_u, z_i)) + \varepsilon_i(u) \quad (20)$$

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 $\theta = (\beta, \Gamma, \sigma)$.

240 **Numerical Experiment:** We apply model (20) and our algorithms 1 to a collection of handwritten
 241 digits, called the US postal database [13], featuring $n = 1\,000$ (16×16)-pixel images for each
 242 class of digits from 0 to 9. The main difficulty with these data comes from the geometric dispersion
 243 within each class of digit as shown Figure 2 for digit 5. We thus ought to use our deformable
 244 template model (20) in order to account for both sources of variability: the intrinsic template to each
 245 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)

246 Figure 3 shows the resulting synthetic images for digit 5 through several epochs, for the batch
 247 method, the online SAEM, the incremental SAEM and the various TTS methods. For all methods,
 248 the initialization of the template (21) is the mean of the gray level images. In our experiments, we
 249 have chosen Gaussian kernels for both, \mathbf{K}_p and \mathbf{K}_g , defined on \mathbb{R}^2 and centered on the landmark
 250 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
 251 $k_p = 15$ and $k_g = 6$ equidistributed landmarks points on the grid for the training procedure. Those
 252 hyperparameters are inspired by a relevant study in [2]. In particular, the choice of the geometric
 253 covariance, indexed by g , in such study is critical since it has a direct impact on the *sharpness* of
 254 the templates. As for the photometric hyperparameter, indexed by p , both the template and the
 255 geometry are impacted, in the sense that with a large photometric variance, the kernel centered on
 256 one landmark *spreads out* to many of its neighbors.

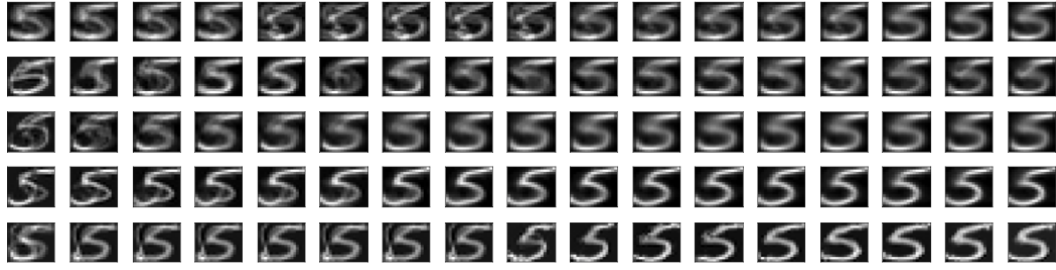


Figure 3: (USPS Digits) Estimation of the template. From top to bottom: batch, online, iSAEM, vrT-TEM and fitTEM through 7 epochs. Note that Batch method templates are replicated in-between epochs for a fair comparison with incremental variants.

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

264 5 Conclusion

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

References

- [1] S. Allasonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.
- [2] S. Allasonnière, E. Kuhn, A. Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.
- [3] C. Baey, S. Trevezas, and P.-H. Cournède. A non linear mixed effects model of plant growth and estimation via stochastic variants of the em algorithm. *Communications in Statistics-Theory and Methods*, 45(6):1643–1669, 2016.
- [4] O. Cappé. Online EM algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749, 2011.
- [5] O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.
- [6] B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484, 1995.
- [7] A. Chakraborty and K. Das. Inferences for joint modelling of repeated ordinal scores and time to event data. *Computational and mathematical methods in medicine*, 11(3):281–295, 2010.
- [8] J. Chen, J. Zhu, Y. W. Teh, and T. Zhang. Stochastic expectation maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988, 2018.
- [9] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL <https://doi.org/10.1214/aos/1018031103>.
- [10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- [11] B. 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.
- [12] J. P. Hughes. Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55(2):625–629, 1999.
- [13] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.
- [14] B. Karimi, H.-T. Wai, É. Moulines, and M. Lavielle. On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems*, pages 2833–2843, 2019.
- [15] E. Kuhn, C. Matias, and T. Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.
- [16] P. Liang and D. Klein. Online em for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619, 2009.
- [17] F. Maire, E. Moulines, and S. Lefebvre. Online em for functional data, 2016. URL <http://arxiv.org/abs/1604.00570>. cite arxiv:1604.00570v1.pdf.

- 316 [18] C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.
- 317
- 318 [19] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- 319
- 320 [20] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- 321
- 322 [21] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.
- 323
- 324 [22] H. D. Nguyen, F. Forbes, and G. J. McLachlan. Mini-batch learning of exponential family finite mixture models. *Statistics and Computing*, pages 1–18, 2020.
- 325
- 326 [23] S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*, 2016.
- 327
- 328 [24] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- 329
- 330 [25] G. C. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man’s data augmentation algorithms. *Journal of the American statistical Association*, 85(411): 699–704, 1990.
- 331
- 332
- 333 [26] R. Zhu, L. Wang, C. Zhai, and Q. Gu. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4180–4188. JMLR. org, 2017.
- 334
- 335

336 A Proof of Lemma 2

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

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

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

$$\begin{aligned} \nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^{\top} \left(\nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{L}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^{\top} \left(\nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^{\top} \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^{\top} \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^{\top} (\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))), \end{aligned} \quad (22)$$

340 Consider the following vector map:

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

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

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

342 The above yields

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

343 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 (21) follows directly
344 from the assumption H4. \square

345 B Proof of Theorem 1

346 Beforehand, We present two intermediary Lemmas important for the analysis of the incremental
347 update of the iSAEM algorithm. The first one gives a characterization of the quantity $\mathbb{E}[\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}]$:
348

349 **Lemma 6.** Assume H1. The update (9) is equivalent to the following update on the resulting statis-
350 tics

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

351 Also:

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

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

353 **Proof** From update (9), we have:

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

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

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

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

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

356 The following equalities:

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

357 concludes the proof of the Lemma. \square

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

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

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

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

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

361 The last expectation can be further bounded by

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

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

363

364 **Theorem.** Assume H1-H5. Let K_m be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive
365 step sizes and consider the iSAEM sequence $\{\hat{s}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = 1$ for any $k > 0$.
366 We also set $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\gamma_{k+1} = \frac{1}{k^a \alpha c_1 \bar{L}}$ where
367 $a \in (0, 1)$, $\beta = \frac{c_1 \bar{L}}{n}$. Assume that $\hat{s}^{(k)} \in \mathcal{S}$ for any $k \leq K_m$, then it holds:

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

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

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

369 Taking the expectation on both sides yields:

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

370 Using Lemma 6, we obtain:

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

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

372 $a_k = \gamma_{k+1} \left(v_{\min} - v_{\max}^2 \frac{\beta(n-1)+1}{2n} \right)$ and

$$\begin{aligned}
a_k \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] &\leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] \\
&\quad + \frac{\gamma_{k+1}(1 - \frac{1}{n})}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] + \frac{\gamma_{k+1}}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \quad (23)
\end{aligned}$$

373 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 (23):

$$\begin{aligned}
(a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] &\leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] \\
&\quad + \gamma_{k+1} \left(\frac{1}{2\beta} (1 - \frac{1}{n}) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] \\
&\quad + \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
&\quad + \frac{\gamma_{k+1}^2 L_V L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{s}^{(k)} - \hat{s}^{(\tau_i^k)} \right\|^2 \right] \quad (24)
\end{aligned}$$

374 Next, we observe that

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

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

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

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

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

377 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

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

378 Let us define

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

379 From the above, we get

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

380 Setting $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6v_{\min}\}$, $\bar{L} = \max\{\mathbf{L}_s, \mathbf{L}_V\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$, $\beta = \frac{c_1 \bar{L}}{n}$,

381 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 6$, $\alpha \geq 8$, we observe that

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

382 which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}\mathbf{L}_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} =$

383 $\frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1}\mathbf{L}_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

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

384 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:

$$\begin{aligned}
& \sum_{k=0}^{K_m-1} \Delta^{(k+1)} \\
&= 4 \sum_{k=0}^{K_m-1} \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta} \right) \omega_{k,1} \mathbb{E} [\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + 2 \sum_{k=0}^{K_m-1} \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta} \right) \omega_{k,1} \mathbb{E} \left[\left\| \eta_{i_\ell}^{(k)} \right\|^2 \right] \\
&+ \sum_{k=0}^{K_m-1} 4 \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta} \right) \omega_{k,1} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \\
&\leq \sum_{k=0}^{K_m-1} \frac{4 \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta} \right)}{\Lambda_{(k+1)}} \mathbb{E} [\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{k=0}^{K_m-1} \frac{2 \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta} \right)}{\Lambda_{(k+1)}} \mathbb{E} \left[\left\| \eta_{i_\ell}^{(k)} \right\|^2 \right] \\
&+ \sum_{k=0}^{K_m-1} \frac{4 \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta} \right)}{\Lambda_{(k+1)}} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right]
\end{aligned} \tag{25}$$

385 We recall (24) where we have summed on both sides from $k = 0$ to $k = K_m - 1$:

$$\begin{aligned}
& \sum_{k=0}^{K_m-1} (a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K)}) \right] \\
&+ \sum_{k=0}^{K_m-1} \gamma_{k+1} \left(\frac{1}{2\beta} \left(1 - \frac{1}{n} \right) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \\
&+ \sum_{k=0}^{K_m-1} \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
&+ \sum_{k=0}^{K_m-1} \frac{\gamma_{k+1}^2 L_V L_s^2}{n^2} \Delta^{(k)}
\end{aligned} \tag{26}$$

386 Plugging (25) into (26) results in:

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

387 where

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

388 and

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

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

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

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

392 C Proofs of Auxiliary Lemmas

393 C.1 Proof of Lemma 3 and Lemma 4

394 **Lemma.** For any $k \geq 0$ and consider the vrTTEM update in (10) with $\rho_k = \rho$, it holds for all $k > 0$
395

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

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

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

$$\begin{aligned}
\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}) = -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - (1-\rho)\tilde{S}^{(k)} - \rho\mathcal{S}^{(k+1)}) \\
&= -\gamma_{k+1} \left((1-\rho) \left[\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} \right] + \rho \left[\hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)} \right] \right)
\end{aligned} \tag{27}$$

399 We observe, using the identity (27), that

$$\mathbb{E}[\left\| \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \right\|^2] \leq 2\rho^2 \mathbb{E}[\left\| \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} \right\|^2] + 2\rho^2 \mathbb{E}[\left\| \bar{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)} \right\|^2] + 2(1-\rho)^2 \mathbb{E}[\left\| \hat{\mathbf{s}}^{(\ell(k))} - \tilde{S}^{(k)} \right\|^2] \tag{28}$$

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

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

401 where (a) uses the variance inequality and (b) uses Lemma 1. Substituting into (28) proves the
402 lemma. \square

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

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

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

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

406 We observe, using the identity (29), that

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

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

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

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

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

409 Substituting into (30) proves the lemma. \square

410 C.2 Proof of Lemma 5

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

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

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

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

$$\begin{aligned}
 \tilde{\mathbf{S}}^{(k+1)} &= \tilde{\mathbf{S}}^{(k)} + \rho(\mathbf{S}^{(k+1)} - \tilde{\mathbf{S}}^{(k)}) \\
 \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)})
 \end{aligned} \tag{31}$$

415 where $\mathbf{S}^{(k+1)} = \frac{1}{n}\sum_{i=1}^n\tilde{\mathbf{S}}_i^{(t_i^k)} + (\tilde{\mathbf{S}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(t_{i_k}^k)})$ according to (11). Denote $\delta^{(k+1)} = \hat{\mathbf{s}}^{(k+1)} -$
 416 $\tilde{\mathbf{S}}^{(k+1)}$. Then from (31), doing the subtraction of both equations yields:

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

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

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

418 \square

419 **C.3 Additional Intermediary Result**

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

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

422 *where we recall that $\eta_{i_k}^{(k+1)}$, defined in (17), which is the gap between the MC approximation and*
 423 *the expected statistics.*

424 **Proof** Using the fitTEM update $\tilde{S}^{(k+1)} = (1 - \rho)\tilde{S}^{(k)} + \rho\mathcal{S}^{(k+1)}$ where $\mathcal{S}^{(k+1)} = \bar{\mathcal{S}}^{(k)} + (\tilde{S}_{i_k}^{(k)} -$
 425 $\tilde{S}_{i_k}^{(t_{i_k}^k)})$ leads to the following decomposition:

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

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

427 *Important Note:* Note that $\bar{\mathbf{s}}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not equal to $\eta_{i_k}^{(k+1)}$, defined in (17), which is the gap
 428 *between the MC approximation and the expected statistics. Indeed $\tilde{S}_{i_k}^{(t_{i_k}^k)}$ is not computed under the*
 429 *same model as $\bar{\mathbf{s}}_{i_k}^{(k)}$. \square*

430 D Proof of Theorem 2

431 **Theorem.** Assume H1-H5. Let K_m be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive
 432 step sizes and consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\}$. Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_m$.
 433 Setting $\bar{L} = \max\{L_s, L_V\}$, $\rho = \frac{\mu}{c_1 L n^{2/3}}$, $m = \frac{nc_1^2}{2\mu^2 + \mu c_1^2}$, a constant $\mu \in (0, 1)$, $\gamma_{k+1} = \frac{1}{k^a \bar{L}}$ where
 434 $a \in (0, 1)$, it holds:

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

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

$$\begin{aligned} V(\hat{\mathbf{s}}^{(k+1)}) &\leq V(\hat{\mathbf{s}}^{(k)}) + \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{L_V}{2} \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (32)$$

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

$$\begin{aligned} &\mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] \\ &\stackrel{(a)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1}(1 - \rho) \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1} \rho \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|\mathbf{H}_{k+1}\|^2] \\ &\stackrel{(b)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1} \rho \mathbb{E}[\langle \mathbf{h}_k | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] - \gamma_{k+1}(1 - \rho) \mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad - \gamma_{k+1} \rho \mathbb{E}[\langle \eta_{i_k}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|\mathbf{H}_{k+1}\|^2] \\ &\stackrel{(c)}{\leq} \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - (\gamma_{k+1} \rho v_{\min} + \gamma_{k+1} v_{\max}^2) \mathbb{E}[\|\mathbf{h}_k\|^2] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E}[\|\mathbf{H}_{k+1}\|^2] \\ &\quad - \gamma_{k+1} \rho \mathbb{E}[\|\eta_{i_k}^{(k+1)}\|^2] - \gamma_{k+1}(1 - \rho) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \end{aligned} \quad (33)$$

438 where we have used (27) 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
 439 Lemma 2 and the Young's inequality with the constant equal to 1 in (c).

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

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

441 where we first used (27) and the last inequality is due to the Young's inequality.

442 Consider the following sequence

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

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

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

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

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

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

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

446 And using Lemma 3 we obtain:

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

447 Rearranging the terms yields:

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

448 where

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

449 This leads, using Lemma 2, that for any γ_{k+1} , ρ and β such that $\rho v_{\min} + v_{\max}^2 -$
450 $\gamma_{k+1}\rho^2 L_V - b_{k+1}(\frac{\rho}{\beta} + 2\gamma_{k+1}\rho^2) > 0$,

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

451 We first remark that

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

452 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
 453 $\{\gamma_{k+1}\}$ any sequence of decreasing stepsizes in $(0, 1)$, it can be shown that there exists $\mu \in (0, 1)$,
 454 such that the following lower bound holds

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

455 where the simplification in (a) is due to

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

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

457 Finally, these results yield:

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

458 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
 459 condition, we have

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

460 This concludes our proof.

461

□

462 E Proof of Theorem 3

463 **Theorem.** Assume *H1-H5*. Let K_m be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive
 464 step sizes and consider the *fitTEM* sequence $\{\hat{s}^{(k)}, k \in \mathbb{N}\}$. Assume that $\hat{s}^{(k)} \in \mathcal{S}$ for any $k \leq K_m$.
 465 Setting $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \bar{L} n^{2/3}}$, $c_1(k\alpha - 1) \geq$
 466 $c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$ and $\gamma_{k+1} = \frac{1}{k^a \alpha c_1 \bar{L}}$ where $a \in (0, 1)$, it holds:

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

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

$$\begin{aligned} V(\hat{s}^{(k+1)}) &\leq V(\hat{s}^{(k)}) + \langle \hat{s}^{(k+1)} - \hat{s}^{(k)} | \nabla V(\hat{s}^{(k)}) \rangle + \frac{L_V}{2} \|\hat{s}^{(k+1)} - \hat{s}^{(k)}\|^2 \\ &\leq V(\hat{s}^{(k)}) - \gamma_{k+1} \langle \hat{s}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{s}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{s}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (34)$$

468 Denote $H_{k+1} := \hat{s}^{(k)} - \tilde{S}^{(k+1)}$ the drift term of the *fitTEM* update in (7) and $h_k = \hat{s}^{(k)} - \bar{s}^{(k)}$.
 469 Using Lemma 8 and the additional following identity:

$$\mathbb{E} \left[(\bar{s}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}) - \mathbb{E}[\bar{s}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}] \right] = 0 \quad (35)$$

470 we have:

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

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

$$\begin{aligned} &\gamma_{k+1} (v_{\min} \rho + v_{\max}^2 - \gamma_{k+1} \rho^2 L_V) \mathbb{E}[\|h_k\|^2] \\ &\leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] + \tilde{\xi}^{(k+1)} + \left((1 - \rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1} (1 - \rho)^2}{2} \right) \mathbb{E}[\|\hat{s}^{(k)} - \tilde{S}^{(k)}\|^2] \\ &\quad + \frac{\gamma_{k+1}^2 L_V \rho^2 L_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{s}^{(k)} - \hat{s}^{(t_i^k)}\|^2] \end{aligned} \quad (36)$$

472 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

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{s}^{(k+1)} - \hat{s}^{(t_i^{k+1})}\|^2] = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \mathbb{E}[\|\hat{s}^{(k+1)} - \hat{s}^{(k)}\|^2] + \frac{n-1}{n} \mathbb{E}[\|\hat{s}^{(k+1)} - \hat{s}^{(t_i^k)}\|^2] \right) \quad (37)$$

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

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

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

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

477 where the last inequality is due to the Young's inequality. Plugging this into (37) yields:

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

478 Subsequently, we have

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

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

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

480 Let us define

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

481 From the above, we get

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

482 Setting $c_1 = v_{\min}^{-1}$, $\alpha = \max\{2, 1+2v_{\min}\}$, $\bar{L} = \max\{L_s, L_V\}$, $\gamma_{k+1} = \frac{1}{k}$, $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 L n^{2/3}}$,
483 $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$, we observe that

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

484 which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \gamma_{k+1}^2\rho^2 L_s^2 \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} = \frac{1}{n} -$
485 $\gamma_{k+1}\beta - \gamma_{k+1}^2\rho^2 L_s^2$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

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

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

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

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

488 We recall (36) where we have summed on both sides from $k = 0$ to $k = K_m - 1$:

$$\begin{aligned} &\mathbb{E}[V(\hat{\mathbf{s}}^{(K_m)}) - V(\hat{\mathbf{s}}^{(0)})] \\ &\leq \sum_{k=0}^{K_m-1} \left\{ \gamma_{k+1}(-(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}\rho^2 L_V) \mathbb{E}[\|\mathbf{h}_k\|^2] + \gamma^2 L_V \rho^2 L_s^2 \Delta^{(k)} \right\} \\ &\quad + \sum_{k=0}^{K_m-1} \left\{ \tilde{\xi}^{(k+1)} + \left((1-\rho)^2 \gamma_{k+1}^2 L_V - \frac{\gamma_{k+1}(1-\rho)^2}{2} \right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \right\} \\ &\leq \sum_{k=0}^{K_m-1} \left\{ \left[-\gamma_{k+1}(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}^2\rho^2 L_V + \frac{\rho^2 \gamma_{k+1}^2 L_V L_s^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \right] \mathbb{E}[\|\mathbf{h}_k\|^2] \right\} \\ &\quad + \sum_{k=0}^{K_m-1} \Xi^{(k+1)} + \sum_{k=0}^{K_m-1} \Gamma^{(k+1)} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \end{aligned} \tag{38}$$

where

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

and

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

489 We now analyse the following quantity

$$\begin{aligned}
& -\gamma_{k+1}(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}^2\rho^2 L_V + \frac{\rho^2\gamma_{k+1}^2 L_V L_S^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \\
& = \gamma_{k+1} \left[-(v_{\min}\rho + v_{\max}^2) + \gamma_{k+1}\rho^2 L_V + \frac{\rho^2\gamma_{k+1}^2 L_V L_S^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\Lambda_{(k+1)}} \right]
\end{aligned} \tag{39}$$

490 Furthermore, we recall that $c_1 = v_{\min}^{-1}$, $\alpha = \max\{2, 1 + 2v_{\min}\}$, $\bar{L} = \max\{L_S, L_V\}$, $\gamma_{k+1} = \frac{1}{k}$,
491 $\beta = \frac{1}{\alpha n}$, $\rho = \frac{1}{\alpha c_1 \bar{L} n^{2/3}}$, $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$. Then,

$$\begin{aligned}
& \gamma_{k+1}\rho^2 L_V + \frac{\rho^2\gamma_{k+1}^2 L_V L_S^2 \left(2\gamma_{k+1}^2\rho^2 + \frac{\gamma_{k+1}\rho^2}{\beta}\right)}{\frac{1}{n} - \gamma_{k+1}\beta - \gamma_{k+1}^2\rho^2 L_S^2} \\
& \leq \frac{1}{k\alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\bar{L}(k\alpha^2 c_1^2 n^{4/3})^{-1} \left(\frac{2}{k^2 \alpha^2 c_1^2 \bar{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \bar{L}^2 n^{1/3}}\right)}{\frac{1}{n} - \frac{1}{k\alpha n} - \frac{1}{k^2 \alpha^2 c_1^2 n^{4/3}}} \\
& = \frac{1}{k\alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\bar{L} \left(\frac{2}{k^2 \alpha^2 c_1^2 \bar{L}^2 n^{4/3}} + \frac{1}{k\alpha c_1^2 \bar{L}^2 n^{1/3}}\right)}{(k\alpha c_1 n^{1/3})(k\alpha - 1)c_1 - 1} \\
& \stackrel{(a)}{\leq} \frac{1}{k\alpha^2 c_1^2 \bar{L} n^{4/3}} + \frac{\frac{1}{k\alpha c_1^2 \bar{L} n^{1/3}} \left(\frac{2}{k\alpha n} + 1\right)}{2(\alpha c_1 n^{1/3}) - 1} \\
& \leq \frac{1}{k^2 \alpha c_1^2 \bar{L} n^{4/3}} + \frac{1}{4k\alpha^2 c_1^3 \bar{L} n^{2/3}} \\
& \leq \frac{3/4}{\alpha c_1^2 \bar{L} n^{2/3}}
\end{aligned} \tag{40}$$

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

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

492 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 (40) on (38)
493 yields:

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

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

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

495 □

496 F Practical Implementations of Two-Timescale EM Methods

497 F.1 Application on GMM

498 F.1.1 Explicit Updates

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

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

500 Using the partition of the sufficient statistics as $S(y_i, z_i) =$
 501 $(S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i)^\top)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, the partition
 502 $\phi(\theta) = (\phi^{(1)}(\theta)^\top, \phi^{(2)}(\theta)^\top, \phi^{(3)}(\theta)^\top)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$ and the fact that
 503 $\mathbb{1}_{\{M\}}(z_i) = 1 - \sum_{m=1}^{M-1} \mathbb{1}_{\{m\}}(z_i)$, the complete data log-likelihood can be expressed as in
 504 (2) with

$$\begin{aligned} s_{i,m}^{(1)} &= \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\theta) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\} - \left\{ \log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2} \right\}, \\ s_{i,m}^{(2)} &= \mathbb{1}_{\{m\}}(z_i)y_i, \quad \phi_m^{(2)}(\theta) = \mu_m, \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\theta) = \mu_M, \end{aligned} \quad (41)$$

505 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 $m \in \llbracket 1, M \rrbracket$, $j \in \llbracket 1, 3 \rrbracket$,
 506 $s_m^{(j)} = n^{-1} \sum_{i=1}^n s_{i,m}^{(j)}$. Consider the following latent sample used to compute an approximation of
 507 the conditional expected value $\mathbb{E}_\theta[\mathbb{1}_{\{z_i=m\}}|y = y_i]$:

$$z_{i,m} \sim \mathbb{P}(z_i = m|y_i; \theta) \quad (42)$$

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

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

$$\tilde{S}_{i_k}^{(k)} = \left(\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1})y_{i_k}, \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\tilde{s}_{i_k}^{(3)}(\theta^{(k)})} \right)^\top. \quad (43)$$

511 Recall that we have used the following regularizer:

$$\mathbf{r}(\theta) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log(1 - \sum_{m=1}^{M-1} \omega_m), \quad (44)$$

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

$$\bar{\theta}(\mathbf{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1} (s_1^{(1)} + \epsilon, \dots, s_{M-1}^{(1)} + \epsilon)^\top \\ ((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \dots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)})^\top \\ (1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta)^{-1} (s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}) \end{pmatrix} = \begin{pmatrix} \bar{\omega}(\mathbf{s}) \\ \bar{\boldsymbol{\mu}}(\mathbf{s}) \\ \bar{\mu}_M(\mathbf{s}) \end{pmatrix}. \quad (45)$$

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

514 F.1.2 Model Assumptions (GMM example)

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

516 Many practical models can satisfy the compactness of the sets as in Assumption H1. For instance,
 517 the GMM example satisfies (15) as the sufficient statistics are composed of indicator functions and
 518 observations as defined Section F.1 Equation (41).

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

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

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

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

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

525 F.1.3 Algorithms updates

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

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

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

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

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

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

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

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

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

534 where $= \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(k)}$ with $\tilde{S}_i^{(k)}$ defined in (43).

535 **Incremental SAEM (iSAEM):** draw an index i_k uniformly at random on $\llbracket n \rrbracket$, compute $\bar{s}_{i_k}^{(k)}$ and set
 536

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

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

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

539 **Fast Incremental Two-Timescale EM (fiTTEM):** draw an index i_k uniformly at random on $\llbracket n \rrbracket$,
 540 compute $\bar{s}_{i_k}^{(k)}$ and set

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

541 Finally, the k -th update reads $\hat{\theta}^{(k+1)} = \bar{\theta}(\hat{s}^{(k+1)})$ where the function $s \rightarrow \bar{\theta}(s)$ is defined by (45).

542 F.2 Deformable Template Model for Image Analysis

543 F.2.1 Model and Updates

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

$$\begin{aligned} S_1(z) &= \frac{1}{n} \sum_{i=1}^n S_1(y_i, z_i) = \frac{1}{n} \sum_{i=1}^n (\mathbf{K}_p^{z_i})^\top y_i \\ S_2(z) &= \frac{1}{n} \sum_{i=1}^n S_2(y_i, z_i) = \frac{1}{n} \sum_{i=1}^n (\mathbf{K}_p^{z_i})^\top (\mathbf{K}_p^{z_i}) \\ S_3(z) &= \frac{1}{n} \sum_{i=1}^n S_3(y_i, z_i) = \frac{1}{n} \sum_{i=1}^n z_i^t z_i \end{aligned} \quad (46)$$

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

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

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

$$\bar{\theta}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z) \hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n} \hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z) \beta(\hat{s}) - 2\beta(\hat{s}) \hat{s}_1(z) \end{pmatrix} \quad (47)$$

548 where $\hat{s} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via the SA-step (7) and using the
549 MC approximation of the sufficient statistics $(S_1(z), S_2(z), S_3(z))$ defined in (46).

550 F.2.2 Numerical Applications

551 For the inference of the template, we use the Matlab code (online SAEM) used in [17] and implement
552 our own batch, incremental, Variance reduced and Fast Incremental variants. The hyperparameters
553 are kept the same and reads as follows $M = 400$, $\gamma_k = 1/k^{0.6}$ and $p = 16$. The number of
554 landmarks for the template is $k_p = 15$ points and for the deformation $k_g = 6$ points. Both have
555 Gaussian kernels with respectively standard deviation of 0.08 and 0.16. The standard deviation of
556 the measurement errors is set to 0.1.

557 For the simulation part, we use the Carlin and Chib MCMC procedure, see [6]. Refer to [17] for
558 more details.

559 G Additional Experiment: Pharmacokinetics (PK) Model with Absorption 560 Lag Time

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

566 **Model and Explicit Updates:** We consider a one-compartment PK model for oral administration
567 with an absorption lag-time (T^{lag}), assuming first-order absorption and linear elimination processes.
568 The final model includes the following variables: ka the absorption rate constant, V the volume of
569 distribution, k the elimination rate constant and T^{lag} the absorption lag-time. We also add several
570 covariates to our model such as D the dose of drug administered, t the time at which measures
571 are taken and the weight of the patient influencing the volume V . More precisely, the log-volume
572 $\log(V)$ 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
573 vector of individual PK parameters, different for each individual i . The final model reads:

$$y_{ij} = f(t_{ij}, z_i) + \varepsilon_{ij} \quad \text{where} \quad f(t_{ij}, z_i) = \frac{D ka_i}{V(ka_i - k_i)} (e^{-ka_i(t_{ij} - T_i^{\text{lag}})} - e^{-k_i(t_{ij} - T_i^{\text{lag}})}) \quad (48)$$

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:

$$\log(T_i^{\text{lag}}) \sim \mathcal{N}(\log(T_{\text{pop}}^{\text{lag}}), \omega_{T^{\text{lag}}}^2), \log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2), \quad (49)$$

$$\log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2), \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2). \quad (50)$$

We recall that the complete model (y, z) defined by (48) belongs to the curved exponential family, which vector of sufficient statistics $S = (S_1(z), S_2(z), S_3(z))$ read:

$$S_1(z) = \frac{1}{n} \sum_{i=1}^n z_i, \quad S_2(z) = \frac{1}{n} \sum_{i=1}^n z_i^\top z_i, \quad S_3(z) = \frac{1}{n} \sum_{i=1}^n (y_i - f(t_i, z_i))^2 \quad (51)$$

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 $\hat{S}^{(k+1)}$ and $\hat{s}^{(k+1)}$ are then updated according to the different methods. Finally the maximization step yields:

$$\bar{\theta}(s) = \begin{pmatrix} \hat{s}_1^{(k+1)} \\ \hat{s}_2^{(k+1)} - \hat{s}_1^{(k+1)} (\hat{s}_1^{(k+1)})^\top \\ \hat{s}_3^{(k+1)} \end{pmatrix} = \begin{pmatrix} \overline{z_{\text{pop}}}(\hat{s}^{(k+1)}) \\ \overline{\omega_z}(\hat{s}^{(k+1)}) \\ \overline{\sigma}(\hat{s}^{(k+1)}) \end{pmatrix}. \quad (52)$$

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 [20] 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:  $z_{i,M}$ 

```

Monte Carlo study: We conduct a Monte Carlo study to showcase the benefits of our scheme. $M = 50$ datasets have been simulated using the following PK parameters values: $T_{\text{pop}}^{\text{lag}} = 1$, $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$, $\omega_{T^{\text{lag}}} = 0.4$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. We define the mean square distance over the M replicates $E_k(\ell) = \frac{1}{M} \sum_{m=1}^M (\theta_k^{(m)}(\ell) - \theta^*)^2$ and plot it against the epochs (passes over the data) Figure 4. Note that the MC-step (5) is performed using a Metropolis Hastings procedure since the posterior distribution under the model θ noted $p(z_i|y_i, \theta)$ is intractable due to the nonlinearity of the model (48). Figure 4 shows clear advantage of variance reduced methods (vrTTEM and fitTEM) avoiding the twists and turns displayed by the incremental and the batch methods.

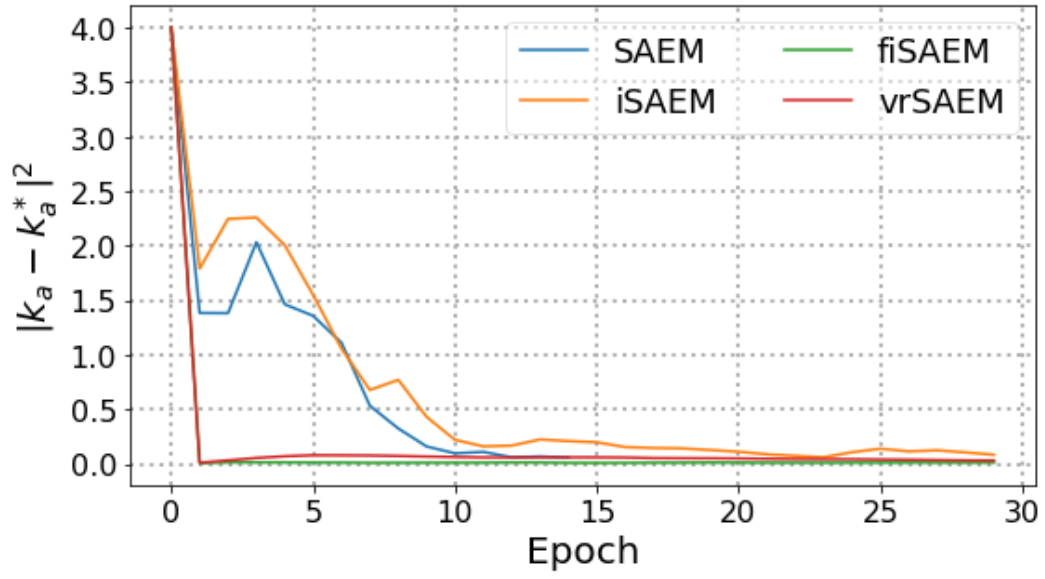


Figure 4: Precision $|ka^{(k)} - ka^*|^2$ per epoch