
Fast Two-Time-Scale Noisy EM Algorithms

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Training latent data models using the EM algorithm is the most common choice
2 for current learning tasks. Variants of the EM to scale to large datasets and by-
3 pass the impossible conditional expectation of the latent data for most nonlinear
4 models have been initially introduced respectively by [Neal and Hinton, 1998],
5 using incremental updates, and [Wei and Tanner, 1990, Delyon et al., 1999], using
6 Monte-Carlo (MC) approximations. In this paper, we propose to combine those
7 both techniques in a single class of methods called Two-Time-Scale EM Methods.
8 We motivate the choice of a double dynamics by invoking the variance reduction
9 virtue of each stage of the method on both noise: the incremental update and the
10 MC approximation. We establish finite-time convergence bounds for nonconvex
11 objective function and independent of the initialization. Numerical applications
12 are also presented in this article to illustrate our findings.

13 1 Introduction

14 Learning latent data models is critical for modern machine learning problems, see [McLachlan and
15 Krishnan, 2007] for references. We formulate the training of such model as the following empirical
16 risk minimization problem:

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

17 We denote the observations by $\{y_i\}_{i=1}^n$, $\Theta \subset \mathbb{R}^d$ is the convex parameters space. We consider a
18 regularized model where $r : \Theta \rightarrow \mathbb{R}$ is a smooth convex regularization function and for $\theta \in \Theta$,
19 $g(y; \theta)$ is the (incomplete) likelihood of each individual observation. The objective function $\bar{L}(\theta)$ is
20 possibly *nonconvex* and is assumed to be lower bounded $\bar{L}(\theta) > -\infty$ for all $\theta \in \Theta$.

21 In the latent variable model, $g(y_i; \theta)$, is the marginal of the complete data likelihood defined as
22 $f(z_i, y_i; \theta)$, i.e. $g(y_i; \theta) = \int_{\mathcal{Z}} f(z_i, y_i; \theta) \mu(dz_i)$, where $\{z_i\}_{i=1}^n$ are the (unobserved) latent vari-
23 ables. In this paper, we make the assumption of a complete model belonging to the curved expo-
24 nential family, i.e.,

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

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

27 Full batch EM [Dempster et al., 1977] is the method of reference for that kind of task and is a two
28 steps procedure. The E-step amounts to computing the conditional expectation of the complete data
29 sufficient statistics,

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

30 The M-step is given by

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

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

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

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

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

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

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

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

75 2 Two-Time-Scale Stochastic EM Algorithms

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

79 2.1 Monte Carlo Integration and Stochastic Approximation

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

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

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

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

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

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

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

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

112 2.2 Incremental and Bi-Level Inexact EM Methods

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

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

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

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

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

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

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

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

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

132 2.3 Two-Time-Scale Noisy EM methods

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

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

Algorithm 1 Two-Time-Scale Noisy EM methods.

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

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

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

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

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

138 3 Global and Finite Time Analysis of the Scheme

139 First, we consider the following minimization problem on the statistics space:

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

140 It has been shown that this minimization problem is equivalent to the optimization problem (1), see
 141 [Karimi et al., 2019, Lemma2]

142 **H1.** Θ is an open set of \mathbb{R}^d and the sets Z, S are measurable open sets such that:

$$S \supset \left\{ n^{-1} \sum_{i=1}^n u_i, u_i \in \text{conv}(\bar{s}_i(\theta)) \right\} \quad (16)$$

143 where $\bar{s}_i(\theta)$ is defined in (3).

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

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

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

149 Similar to [Karimi et al., 2019], we denote by $H_L^\theta(s, \theta)$ the Hessian (w.r.t to θ for a given value of
 150 s) of the function $\theta \mapsto L(s, \theta) = r(\theta) + \psi(\theta) - \langle s | \phi(\theta) \rangle$, and define

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

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

153 We now formulate the main difference with the work done in [Karimi et al., 2019]. The class of
 154 algorithms we develop in this paper are two time-scale where the first stage corresponds to the
 155 variance reduction trick used in [Karimi et al., 2019] in order to accelerate incremental methods and
 156 kill the variance induced by the index sampling. The second stage is the Robbins-Monro type of
 157 update that aims to kill the variance induced by the MC approximations

158 Indeed the expectations (3) are never available and requires Monte Carlo approximation. Thus, at
 159 iteration $k + 1$, we introduce the errors when approximating the quantity $\bar{s}_i(\hat{\theta}(\hat{s}^{(k-1)}))$. For all
 160 $i \in \llbracket 1, n \rrbracket$, $r > 0$ and $\vartheta \in \Theta$, define:

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

161 For instance, we consider that the MC approximation is unbiased if for all $i \in \llbracket 1, n \rrbracket$ and $m \in$
 162 $\llbracket 1, M \rrbracket$, the samples $z_{i,m} \sim p(z_i|y_i; \theta)$ are i.i.d. under the posterior distribution, i.e., $\mathbb{E}[\eta_i^{(r)} | \mathcal{F}_r] = 0$
 163 where \mathcal{F}_r is the filtration up to iteration r .

164 The following results are derived under the assumption of control of the fluctuations implied by the
 165 approximation stated as follows:

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

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

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

170 **Lemma 1.** [Karimi et al., 2019] Assume H2, H3, H4. For all $s, s' \in S$ and $i \in \llbracket 1, n \rrbracket$, we have

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

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

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

174 **Lemma 2.** Assume H3, H4. For all $s \in S$,

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

175 See proofs of this Lemma in Appendix A.

3.1 Global Convergence of Incremental Noisy EM Algorithms

Following the asymptotic analysis of update (9), we present a finite-time analysis of the incremental variant of the Stochastic Approximation of the EM algorithm.

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

Lemma 3. Assume H1. The update (9) is equivalent to the following update on the resulting statistics

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} + \gamma_{k+1} (\tilde{S}^{(k+1)} - \hat{s}^{(k)}) \quad (22)$$

Also:

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

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

See proofs of this Lemma in Appendix B.

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

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

Assume that $\hat{s}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.

TO COMPLETE WITH BOUND

See proof in Appendix C.

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

We now proceed by giving our main result regarding the global convergence of the fiSAEM algorithm.

Theorem 2. Let K_{\max} be a positive integer. Let $\{\gamma_k, k \in \mathbb{N}\}$ be a sequence of positive step sizes and consider the fiSAEM sequence $\{\hat{s}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = \rho$ for any $k > 0$.

Assume that $\hat{s}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.

TO COMPLETE WITH BOUND

See proof in Appendix D.

4 Numerical Examples

4.1 Gaussian Mixture Models

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:

$$\log f(z_i, y_i; \theta) = \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) [\log(\omega_m) - \mu_m^2/2] + \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) \mu_m y_i + \text{constant} . \quad (24)$$

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 on θ is given by

$$\Theta = \{\omega_m, m = 1, \dots, M-1 : \omega_m \geq 0, \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R}, m = 1, \dots, M\}. \quad (25)$$

Exact two time scale updates are given in Appendix E.1.

In the following experiments on synthetic data, we generate samples from a GMM model with $M = 2$ components with two mixtures with means $\mu_1 = -\mu_2 = 0.5$. We use $n = 10^4$ synthetic samples and run the bEM method until convergence (to double precision) to obtain the ML estimate μ^* averaged on 50 datasets. We compare the bEM, SAEM, iSAEM, vrSAEM and fiSAEM 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 of the Incremental-step for vrSAEM and the fiSAEM to a constant stepsize equal to $1/n^{2/3}$.

The number of MC samples is fixed to $M = 40$ chains. Figure 1 shows the convergence of the precision $|\mu - \mu^*|^2$ for the different methods against the epoch(s) elapsed (one epoch equals n iterations). We observe that the vrSAEM and fiSAEM methods outperform the other methods, supporting our analytical results.

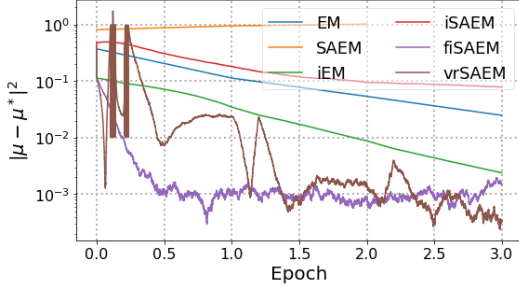


Figure 1: TO COMPLETE

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:

$$y_i(u) = I(x_u - \Phi_i(x_u, z_i)) + \varepsilon_i(u) \quad (26)$$

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

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

$$I_\xi = \mathbf{K}_p \beta, \quad \text{where} \quad (\mathbf{K}_p \beta)(x) = \sum_{k=1}^{k_p} \mathbf{K}_p(x, p_k) \beta_k \quad (27)$$

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

$$\Phi_i = \mathbf{K}_g z_i \quad \text{where} \quad (\mathbf{K}_g z_i)(x) = \sum_{k=1}^{k_g} \mathbf{K}_g(x, g_k) \left(z_i^{(1)}(k), z_i^{(2)}(k) \right) \quad (28)$$

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

$$\begin{aligned} S_1(z) &= \sum_{i=1}^n S_1(y_i, z_i) = \sum_{i=1}^n (\mathbf{K}_p^{z_i})^t y_i \\ S_2(z) &= \sum_{i=1}^n S_2(y_i, z_i) = \sum_{i=1}^n (\mathbf{K}_p^{z_i})^t (\mathbf{K}_p^{z_i}) \\ S_3(z) &= \sum_{i=1}^n S_3(y_i, z_i) = \sum_{i=1}^n z_i^t z_i \end{aligned} \quad (29)$$

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

$$\mathbf{K}_p^{z_i}(x_u, j) = \mathbf{K}_p^{z_i}(x_u - \phi_i(x_u, z_i), p_j) \quad (30)$$

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

$$\bar{\theta}(\hat{s}) = \begin{pmatrix} \beta(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z) \\ \Gamma(\hat{s}) = \frac{1}{n}\hat{s}_3(z) \\ \sigma(\hat{s}) = \beta(\hat{s})^\top \hat{s}_2(z)\beta(\hat{s}) - 2\beta(\hat{s})\hat{s}_1(z) \end{pmatrix} \quad (31)$$

245 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
 246 MC approximation of the sufficient statistics $(S_1(z), S_2(z), S_3(z))$ defined in **(29)**.

247 **Comparison using epochs credit**

248 **Comparison using number of training samples credit**

249 **5 Conclusion**

References

- S. Allasonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.
- S. Allasonnière, E. Kuhn, A. Trouvé, et al. Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. *Bernoulli*, 16(3):641–678, 2010.
- C. Baey, S. Trevezas, and P.-H. Cournède. A non linear mixed effects model of plant growth and estimation via stochastic variants of the em algorithm. *Communications in Statistics-Theory and Methods*, 45(6):1643–1669, 2016.
- O. Cappé. Online em algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749, 2011.
- O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.
- A. Chakraborty and K. Das. Inferences for joint modelling of repeated ordinal scores and time to event data. *Computational and mathematical methods in medicine*, 11(3):281–295, 2010.
- B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL <https://doi.org/10.1214/aos/1018031103>.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
- J. P. Hughes. Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55(2):625–629, 1999.
- B. Karimi. *Non-Convex Optimization for Latent Data Models: Algorithms, Analysis and Applications*. PhD thesis, 2019.
- B. Karimi, H.-T. Wai, É. Moulines, and M. Lavielle. On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems*, pages 2833–2843, 2019.
- E. Kuhn, C. Matias, and T. Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.
- P. Liang and D. Klein. Online em for unsupervised models. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 611–619, 2009.
- C. E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.
- G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.
- H. D. Nguyen, F. Forbes, and G. J. McLachlan. Mini-batch learning of exponential family finite mixture models. *Statistics and Computing*, pages 1–18, 2020.

- 291 S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for nonconvex optimization.
292 *arXiv preprint arXiv:1603.06159*, 2016.
- 293 H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statis-*
294 *tics*, pages 400–407, 1951.
- 295 G. C. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man’s
296 data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704,
297 1990.
- 298 R. Zhu, L. Wang, C. Zhai, and Q. Gu. High-dimensional variance-reduced stochastic gradient
299 expectation-maximization algorithm. In *Proceedings of the 34th International Conference on*
300 *Machine Learning-Volume 70*, pages 4180–4188. JMLR. org, 2017.

A Proof of Lemma 2

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

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \quad (32)$$

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

$$\begin{aligned} \nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \left(\nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{L}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \left(\nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \right) \\ &= \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^\top \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top (\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))), \end{aligned} \quad (33)$$

Consider the following vector map:

$$\mathbf{s} \rightarrow \nabla_{\boldsymbol{\theta}} L(\mathbf{s}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(\mathbf{s})} = \nabla_{\boldsymbol{\theta}} \psi(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathbf{r}(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{s}. \quad (34)$$

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

$$\mathbf{0} = -\mathbf{J}_{\phi}^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) + \underbrace{\left(\nabla_{\boldsymbol{\theta}}^2 (\psi(\boldsymbol{\theta}) + \mathbf{r}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) | \mathbf{s} \rangle) \right)|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}(\mathbf{s})}}_{=\mathbf{H}_L^{\boldsymbol{\theta}}(\mathbf{s}; \boldsymbol{\theta})} \mathbf{J}_{\bar{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s}). \quad (35)$$

The above yields

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathbf{B}(\mathbf{s})(\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))) \quad (36)$$

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 (32) follows directly from the assumption H4. \square

B Proof of Lemma 3

Lemma. Assume H1. The update (9) is equivalent to the following update on the resulting statistics

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1} (\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \quad (37)$$

Also:

$$\mathbb{E} [\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] = \mathbb{E} [\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \left(1 - \frac{1}{n}\right) \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right] + \frac{1}{n} \mathbb{E} [\eta_{i_k}^{(k+1)}] \quad (38)$$

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

Proof From update (9), we have:

$$\begin{aligned} \tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} &= \tilde{\mathbf{S}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \frac{1}{n} \left(\tilde{S}_{i_k}^{(k+1)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)} \right) \\ &= \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{\mathbf{S}}^{(k)} - \bar{\mathbf{s}}^{(k)} - \frac{1}{n} \left(\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \tilde{S}_{i_k}^{(k+1)} \right) \end{aligned} \quad (39)$$

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

$$\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} + \tilde{\mathbf{S}}^{(k)} - \bar{\mathbf{s}}^{(k)} - \frac{1}{n} \left(\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \bar{s}_{i_k}(\boldsymbol{\theta}^{(k)}) \right) + \frac{1}{n} \eta_{i_k}^{(k+1)} \quad (40)$$

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

$$\begin{aligned} \mathbb{E} [\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}] &= \mathbb{E} [\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}] + \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right] \\ &\quad - \frac{1}{n} \mathbb{E} \left[\mathbb{E} [\tilde{S}_{i_k}^{(\tau_{i_k}^k)} - \bar{s}_{i_k}(\boldsymbol{\theta}^{(k)}) | \mathcal{F}_k] \right] + \frac{1}{n} \mathbb{E} [\eta_{i_k}^{(k+1)}] \end{aligned} \quad (41)$$

The following equalities:

$$\mathbb{E} [\tilde{S}_i^{(\tau_i^k)} | \mathcal{F}_k] = \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} \quad \text{and} \quad \mathbb{E} [\bar{s}_{i_k}(\boldsymbol{\theta}^{(k)}) | \mathcal{F}_k] = \bar{\mathbf{s}}^{(k)} \quad (42)$$

concludes the proof of the Lemma. \square

320 C Proof of Theorem 1

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

323 Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.

324 **TO COMPLETE WITH BOUND**

325 **Proof** We begin our proof by giving this auxiliary Lemma setting an upper bound for the quantity

$$326 \mathbb{E} \left[\|\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \right]$$

327 **Lemma 4.** For any $k \geq 0$ and consider the iSAEM update in (9), it holds that

$$\begin{aligned} \mathbb{E} \left[\|\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \right] &\leq 4\mathbb{E} \left[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2 \right] + \frac{2L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 \right] \\ &\quad + 2\frac{C_\eta}{M_k} + 4\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] \end{aligned} \quad (43)$$

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

$$\begin{aligned} \mathbb{E} \left[\|\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \right] &= \mathbb{E} \left[\left\| \tilde{S}^{(k)} - \hat{\mathbf{s}}^{(k)} - \frac{1}{n} (\tilde{S}_{i_k}^{(\tau_i^k)} - \tilde{S}_{i_k}^{(k)}) \right\|^2 \right] \\ &\leq 4\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] + 4\mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \\ &\quad + \frac{2}{n^2} \mathbb{E} \left[\left\| \bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)} \right\|^2 \right] + 2\frac{C_\eta}{M_k} \end{aligned} \quad (44)$$

329 The last expectation can be further bounded by

$$\frac{2}{n^2} \mathbb{E} \left[\left\| \bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)} \right\|^2 \right] = \frac{2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \bar{\mathbf{s}}_i^{(k)} - \bar{\mathbf{s}}_i^{(t_i^k)} \right\|^2 \right] \stackrel{(a)}{\leq} \frac{2L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 \right], \quad (45)$$

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

331 □

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

$$V(\hat{\mathbf{s}}^{(k+1)}) \leq V(\hat{\mathbf{s}}^{(k)}) + \gamma_{k+1} \langle \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \quad (46)$$

333 Taking the expectation on both sides yields:

$$\mathbb{E} \left[V(\hat{\mathbf{s}}^{(k+1)}) \right] \leq \mathbb{E} \left[V(\hat{\mathbf{s}}^{(k)}) \right] + \gamma_{k+1} \mathbb{E} \left[\langle \tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E} \left[\|\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 \right] \quad (47)$$

334 Using Lemma 3, we obtain:

$$\begin{aligned}
& \mathbb{E} \left[\langle \tilde{S}^{(k+1)} - \hat{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] = \\
& \mathbb{E} \left[\langle \bar{s}^{(k)} - \hat{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] + \left(1 - \frac{1}{n} \right) \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \right\rangle \right] + \frac{1}{n} \mathbb{E} \left[\langle \eta_{i_k}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] \\
& \stackrel{(a)}{\leq} -v_{\min} \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + \left(1 - \frac{1}{n} \right) \mathbb{E} \left[\left\langle \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \right\rangle \right] + \frac{1}{n} \mathbb{E} \left[\langle \eta_{i_k}^{(k)} \mid \nabla V(\hat{s}^{(k)}) \rangle \right] \\
& \stackrel{(b)}{\leq} -v_{\min} \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + \frac{1 - \frac{1}{n}}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] \\
& + \frac{\beta(n-1)+1}{2n} \mathbb{E} \left[\left\| \nabla V(\hat{s}^{(k)}) \right\|^2 \right] + \frac{1}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
& \stackrel{(a)}{\leq} \left(v_{\max}^2 \frac{\beta(n-1)+1}{2n} - v_{\min} \right) \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + \frac{1 - \frac{1}{n}}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] + \frac{1}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right]
\end{aligned} \tag{48}$$

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

336 $a_k = \gamma_{k+1} \left(v_{\min} - v_{\max}^2 \frac{\beta(n-1)+1}{2n} \right)$ and

$$\begin{aligned}
a_k \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] & \leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] + \frac{\gamma_{k+1}^2 L_V}{2} \mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] \\
& + \frac{\gamma_{k+1} (1 - \frac{1}{n})}{2\beta} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] + \frac{\gamma_{k+1}}{2n} \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right]
\end{aligned} \tag{49}$$

337 We now give an upper bound of $\mathbb{E} \left[\left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right]$ using Lemma 4 and plug it into (49):

$$\begin{aligned}
(a_k - 2\gamma_{k+1}^2 L_V) \mathbb{E} \left[\left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] & \leq \mathbb{E} \left[V(\hat{s}^{(k)}) - V(\hat{s}^{(k+1)}) \right] \\
& + \gamma_{k+1} \left(\frac{1}{2\beta} (1 - \frac{1}{n}) + 2\gamma_{k+1} L_V \right) \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{s}^{(k)} \right\|^2 \right] \\
& + \gamma_{k+1} \left(\gamma_{k+1} L_V + \frac{1}{2n} \right) \mathbb{E} \left[\left\| \eta_{i_k}^{(k)} \right\|^2 \right] \\
& + \frac{\gamma_{k+1}^2 L_V L_s^2}{n^3} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{s}^{(k)} - \hat{s}^{(t_i^k)} \right\|^2 \right]
\end{aligned} \tag{50}$$

338 Next, we observe that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| \hat{s}^{(k+1)} - \hat{s}^{(t_i^{k+1})} \right\|^2 \right] = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \mathbb{E} \left[\left\| \hat{s}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] + \frac{n-1}{n} \mathbb{E} \left[\left\| \hat{s}^{(k+1)} - \hat{s}^{(t_i^k)} \right\|^2 \right] \right) \tag{51}$$

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

$$\begin{aligned}
& \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\
&= \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle\right] \\
&= \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 - 2\gamma_{k+1}\langle \hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} \mid \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle\right] \\
&\leq \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2 + \gamma_{k+1}\beta\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2\right]
\end{aligned} \tag{52}$$

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

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] \\
&\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n^2} \sum_{i=1}^n \mathbb{E}\left[\left(1 + \gamma_{k+1}\beta\right)\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2\right]
\end{aligned} \tag{53}$$

341 Observe that $\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = -\gamma_{k+1}(\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)})$. Applying Lemma 4 yields

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] \\
&\leq \left(\gamma_{k+1}^2 + \frac{n-1}{n} \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}[\|\tilde{\mathbf{S}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{i=1}^n \mathbb{E}\left[\frac{1 - \frac{1}{n} + \gamma_{k+1}\beta}{n} \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2\right] \\
&\leq 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + 2\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}\left[\|\eta_{i_k}^{(k)}\|^2\right] \\
&\quad + 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\
&\quad + \sum_{i=1}^n \mathbb{E}\left[\frac{1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})}{n} \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2\right]
\end{aligned} \tag{54}$$

342 Let us define

$$\Delta^{(k)} := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \tag{55}$$

343 From the above, we get

$$\begin{aligned}
\Delta^{(k+1)} &\leq \left(1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})\right) \Delta^{(k)} + 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] \\
&\quad + 2\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}\left[\|\eta_{i_k}^{(k)}\|^2\right] + 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{S}}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right]
\end{aligned} \tag{56}$$

344 Setting $c_1 = v_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6c_1, \bar{L}\}$, $\gamma_{k+1} = \frac{1}{k\alpha c_1 \bar{L}}$, $\beta = \frac{c_1 \bar{L}}{n}$, $c_1(k\alpha - 1) \geq$

345 $c_1(\alpha - 1) \geq 6$, $\alpha \geq 8$, we observe that

$$1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_{\mathbf{s}}^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \leq 1 - \frac{c_1(k\alpha - 1) - 4}{k\alpha n c_1} \leq 1 - \frac{2}{k\alpha n c_1} \tag{57}$$

346 which shows that $1 - \frac{1}{n} + \gamma_{k+1}\beta + \frac{2\gamma_{k+1}L_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta}) \in (0, 1)$ for any $k > 0$. Denote $\Lambda_{(k+1)} =$
 347 $\frac{1}{n} - \gamma_{k+1}\beta - \frac{2\gamma_{k+1}L_s^2}{n^2}(\gamma_{k+1} + \frac{1}{\beta})$ and note that $\Delta^{(0)} = 0$, thus the telescoping sum yields:

$$\begin{aligned} \Delta^{(k+1)} \leq & 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^{k-\ell} \mathbb{E}[\|\bar{\mathbf{s}}^{(\ell)} - \hat{\mathbf{s}}^{(\ell)}\|^2] + 2\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^{k-\ell} \mathbb{E}\left[\left\|\eta_{i_\ell}^{(\ell)}\right\|^2\right] \\ & + 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^{k-\ell} \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^\ell)} - \bar{\mathbf{s}}^{(\ell)}\right\|^2\right] \end{aligned} \quad (58)$$

348 Summing on both sides over $k = 0$ to $k = K - 1$ yields:

$$\begin{aligned} & \sum_{k=0}^{K-1} \Delta^{(k+1)} \\ & \leq 4 \sum_{k=0}^{K-1} \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^{k-\ell} \mathbb{E}[\|\bar{\mathbf{s}}^{(\ell)} - \hat{\mathbf{s}}^{(\ell)}\|^2] + 2 \sum_{k=0}^{K-1} \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^{k-\ell} \mathbb{E}\left[\left\|\eta_{i_\ell}^{(\ell)}\right\|^2\right] \\ & + \sum_{k=0}^{K-1} 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^{k-\ell} \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^\ell)} - \bar{\mathbf{s}}^{(\ell)}\right\|^2\right] \\ & = 4 \sum_{k=0}^{K-1} \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^\ell \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + 2 \sum_{k=0}^{K-1} \left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^\ell \mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\ & + \sum_{k=0}^{K-1} 4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right) \sum_{\ell=0}^k (1 - \Lambda_{(\ell+1)})^\ell \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\ & \leq \sum_{k=0}^{K-1} \frac{4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}} \mathbb{E}[\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\|^2] + \sum_{k=0}^{K-1} \frac{2\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}} \mathbb{E}\left[\left\|\eta_{i_\ell}^{(k)}\right\|^2\right] \\ & + \sum_{k=0}^{K-1} \frac{4\left(\gamma_{k+1}^2 + \frac{\gamma_{k+1}}{\beta}\right)}{\Lambda_{(k+1)}} \mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \end{aligned} \quad (59)$$

349 **FINAL**

350 When, for any $k > 0$, $\alpha_k > 0$, we have by Lemma 2 that:

$$\sum_{k=0}^{K_{\max}} \alpha_k \mathbb{E}\left[\left\|\nabla V(\hat{\mathbf{s}}^{(k)})\right\|^2\right] \leq v_{\max}^2 \sum_{k=0}^{K_{\max}} \alpha_k \mathbb{E}\left[\left\|\bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\right\|^2\right] \quad (60)$$

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

353 D Proof of Theorem 2

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

356 Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ for any $k \leq K_{\max}$.

357 **TO COMPLETE WITH BOUND**

358 **Proof** We begin our proof by giving this auxiliary Lemma setting an upper bound for the quantity
 359 $\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2]$

360 **Lemma 5.** For any $k \geq 0$ and consider the fiSAEM update in (11) with $\rho_k = \rho$, it holds that

$$\begin{aligned} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] &\leq 4\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + \frac{4\rho^2 \mathbf{L}_s^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &\quad + 2\rho^2 \frac{C_\eta}{M_k} + 4(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \end{aligned} \quad (61)$$

361 **Proof** Applying the fiSAEM update yields:

$$\begin{aligned} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] &= \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} + \rho(\tilde{S}^{(k)} - \mathcal{S}^{(k+1)})\|^2] \\ &= \mathbb{E}[\|(1-\rho)(\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}) + \rho(\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}) + \rho[\bar{\mathbf{s}}^{(k)} - \mathcal{S}^{(k)}] - (\tilde{S}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)})\|^2] \end{aligned} \quad (62)$$

362 We observe that $\bar{\mathcal{S}}^{(k)} = \frac{1}{n} \sum_{i=1}^n \bar{\mathbf{s}}_i^{(t_i^k)}$ and $\mathbb{E}[\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)}] = \bar{\mathbf{s}}^{(k)} - \bar{\mathcal{S}}^{(k)} + \mathbb{E}[\eta_{i_k}^{(\tau_i^k)}]$. Thus

$$\begin{aligned} \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] &\leq 4(1-\rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2] + 4\rho^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] + 2\rho^2 \mathbb{E}\left[\left\|\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_i^k)}\right\|^2\right] + 2\rho^2 \frac{C_\eta}{M_k} \end{aligned} \quad (63)$$

363 where we use the variance inequality. The last expectation can be further bounded by

$$\mathbb{E}[\|\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_i^k)}\|^2] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\bar{\mathbf{s}}_i^{(k)} - \bar{\mathbf{s}}_i^{(t_i^k)}\|^2] \stackrel{(a)}{\leq} \frac{\mathbf{L}_s}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2], \quad (64)$$

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

365 □

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

$$\begin{aligned} V(\hat{\mathbf{s}}^{(k+1)}) &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 \mathbf{L}_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \\ &\quad - \gamma_{k+1} \rho \langle \tilde{S}^{(k)} - \mathcal{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 \mathbf{L}_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \\ &\leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \rho \langle \hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \\ &\quad - \gamma_{k+1} (1-\rho) \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 \mathbf{L}_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \end{aligned} \quad (65)$$

367 Taking the expectaitons on both sides and noting that $\mathbb{E}[\mathcal{S}^{(k+1)}] = \mathbb{E}[\mathbb{E}[\mathcal{S}^{(k+1)}|\mathcal{F}_k]] = \bar{\mathbf{s}}^{(k)}$
 368 (independence of both indices i_k and j_k), we have:

$$\begin{aligned} \mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)}) - V(\hat{\mathbf{s}}^{(k)})] &\leq -\gamma_{k+1}\rho\mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] \\ &\quad - \gamma_{k+1}(1-\rho)\mathbb{E}[\langle \hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle] + \frac{\gamma_{k+1}^2 L_V}{2}\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2] \\ &\stackrel{(a)}{\leq} -(\gamma_{k+1}v_{\max}^2\frac{(1-\rho)}{2} + \gamma_{k+1}\rho)\mathbb{E}\left[\left\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] - \gamma_{k+1}\frac{1-\rho}{2}\mathbb{E}\left[\left\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right\|^2\right] \\ &\quad + \frac{\gamma_{k+1}^2 L_V}{2}\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2] \end{aligned} \quad (66)$$

369 where (a) used the growth condition (32) twice (on $\langle \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle$ and $\|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2$ and
 370 the triangle inequality.

371 **USE LEMMA 19 AND 20 of Gersende on $\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2]$**

372 **Bounding $\mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2]$** Using Lemma 5, we obtain:

$$\begin{aligned} \mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)}) - V(\hat{\mathbf{s}}^{(k)})] &\leq -\gamma_{k+1}\left(v_{\max}^2\frac{(1-\rho)}{2} + \rho - 2\rho^2\gamma_{k+1}L_V\right)\mathbb{E}\left[\left\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\right\|^2\right] \\ &\quad + \frac{2\gamma_{k+1}^2 L_V \rho^2 L_{\mathbf{s}}^2}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &\quad + \gamma_{k+1}(1-\rho)\left(2\gamma_{k+1}L_V(1-\rho) - \frac{1}{2}\right)\mathbb{E}\left[\left\|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k)}\right\|^2\right] \\ &\quad + \gamma_{k+1}^2 \rho^2 L_V \frac{C_{\eta}}{M_k} \end{aligned} \quad (67)$$

373 Next, we observe that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n} \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \right) \quad (68)$$

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

$$\begin{aligned} &\mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &= \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + 2\langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle\right] \\ &= \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 - 2\gamma_{k+1}\langle \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} | \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle\right] \\ &\leq \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2 + \gamma_{k+1}\beta\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2\right] \end{aligned} \quad (69)$$

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

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] \\ &\leq \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n^2} \sum_{i=1}^n \mathbb{E}\left[\left(1 + \gamma_{k+1}\beta\right)\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2\right] \end{aligned} \quad (70)$$

376 Observe that $\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} = \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)})$. Applying Lemma 5 yields:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^{k+1})}\|^2] &\leq \left(4\rho^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] \\
&\quad + \left(\frac{4\rho^2 L_s^2}{n} + \frac{(n-1)(1 + \gamma_{k+1}\beta)}{n^2}\right) \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\
&\quad + 2\rho^2 \frac{C_n}{M_k} + 4(1 - \rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]
\end{aligned} \tag{71}$$

377 Define

$$\Delta^{(k)} := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \tag{72}$$

378 and note that

$$\begin{aligned}
\Delta^{(k+1)} &\leq \left(1 - \frac{1}{n} + 4\rho^2 L_s^2 + \gamma_{k+1}\beta\right) \Delta^{(k)} + \left(4\rho^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2] \\
&\quad + 2\rho^2 \frac{C_n}{M_k} + 4(1 - \rho)^2 \mathbb{E}[\|\hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k)}\|^2]
\end{aligned} \tag{73}$$

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

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

383

□

E Practical Implementations of Two-Time-Scale EM Methods

E.1 Gaussian mixture models

E.1.1 Model assumptions

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

$$\Theta = \Delta^M \times \mathbb{R}^M. \quad (74)$$

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

$$\begin{aligned} s_{i,m}^{(1)} &= \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\theta) = \left\{ \log(\omega_m) - \frac{\mu_m^2}{2} \right\} - \left\{ \log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2} \right\}, \\ s_{i,m}^{(2)} &= \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\theta) = \mu_m, \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\theta) = \mu_M, \end{aligned} \quad (75)$$

and $\psi(\theta) = -\left\{ \log(1 - \sum_{m=1}^{M-1} \omega_m) - \frac{\mu_M^2}{2\sigma^2} \right\}$. We also define for each $m \in \llbracket 1, M \rrbracket$, $j \in \llbracket 1, 3 \rrbracket$, $s_m^{(j)} = n^{-1} \sum_{i=1}^n s_{i,m}^{(j)}$. Consider the following latent sample used to compute an approximation of the conditional expected value $\mathbb{E}_\theta[\mathbb{1}_{\{z_i=m\}} | y = y_i]$:

$$z_{i,m} \sim \mathbb{P}(z_i = m | y_i; \theta) \quad (76)$$

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

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

$$\tilde{S}_{i_k}^{(k)} = \left(\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1})y_{i_k}, \dots, \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\tilde{s}_{i_k}^{(3)}(\theta^{(k)})} \right)^\top. \quad (77)$$

Recall that we have used the following regularizer:

$$\mathbf{r}(\theta) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log(1 - \sum_{m=1}^{M-1} \omega_m), \quad (78)$$

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

$$\bar{\theta}(\mathbf{s}) = \begin{pmatrix} (1 + \epsilon M)^{-1} (s_1^{(1)} + \epsilon, \dots, s_{M-1}^{(1)} + \epsilon)^\top \\ ((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \dots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)})^\top \\ (1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta)^{-1} (s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}) \end{pmatrix} = \begin{pmatrix} \bar{\omega}(\mathbf{s}) \\ \bar{\boldsymbol{\mu}}(\mathbf{s}) \\ \bar{\mu}_M(\mathbf{s}) \end{pmatrix}. \quad (79)$$

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

E.1.2 Algorithms updates

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 (77). At iteration k , the several E-steps defined by (9) or (10) and (11) leads to the definition of the quantity $\hat{\mathbf{s}}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{\mathbf{s}}^{(0)} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i^{(0)}$, those E-steps break down as follows:

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

$$\hat{\mathbf{s}}^{(k+1)} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i^{(k)}. \quad (80)$$

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

$$\tilde{\omega}_m(y_i; \boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y=y_i] = \frac{\omega_m \exp(-\frac{1}{2}(y_i - \mu_i)^2)}{\sum_{j=1}^M \omega_j \exp(-\frac{1}{2}(y_i - \mu_j)^2)}, \quad (81)$$

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

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} + \frac{1}{n} (\bar{s}_{i_k}^{(k)} - \bar{s}_{i_k}^{(\tau_i^k)}) = n^{-1} \sum_{i=1}^n \bar{s}_i^{(\tau_i^k)}. \quad (82)$$

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

$$\hat{s}^{(k+1)} = \hat{s}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1} \tilde{S}^{(k)}. \quad (83)$$

411 where $= \frac{1}{n} \sum_{i=1}^n \tilde{S}_i^{(k)}$ with $\tilde{S}_i^{(k)}$ defined in (77).

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

$$\hat{s}^{(k+1)} = \hat{s}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1} \left(\tilde{S}^{(k)} + \frac{1}{n} (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)}) \right). \quad (84)$$

414 **Variance Reduced Two-Time-Scale EM (vrSAEM):** draw an index i_k uniformly at random on
415 $\llbracket n \rrbracket$, compute $\bar{s}_{i_k}^{(k)}$ and set

$$\hat{s}^{(k+1)} = \hat{s}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1} (\tilde{S}^{(k)}(1 - \rho) + \rho(\tilde{S}^{(\ell(k))} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}))). \quad (85)$$

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

$$\hat{s}^{(k+1)} = \hat{s}^{(k)}(1 - \gamma_{k+1}) + \gamma_{k+1} (\tilde{S}^{(k)}(1 - \rho) + \rho(\bar{\mathcal{S}}^{(k)} + (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}))). \quad (86)$$

418 Finally, the k -th update reads $\hat{\boldsymbol{\theta}}^{(k+1)} = \bar{\boldsymbol{\theta}}(\hat{s}^{(k+1)})$ where the function $\boldsymbol{s} \rightarrow \bar{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined by (79).