Fast Two-Time-Scale Noisy EM Algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

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

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

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

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

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

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

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

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

The M-step is given by

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

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

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

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

Recently, an incremental variant of the SAEM was proposed in [Kuhn et al., 2019] showing positive empirical results but its analysis is limited to asymptotic consideration. Gradient-based methods have been developed and analyzed in [Zhu et al., 2017] 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 combine both solutions proposed in the past years in a two-time-scale manner in order to optimize (1) for current modern examples and settings. The main contributions of the paper are:

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

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

2 Two-Time-Scale Stochastic EM Algorithms

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

2.1 Monte Carlo Integration and Stochastic Approximation

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

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

and compute $\hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\hat{\mathrm{s}})$.

This algorithm bypasses the intractable expectation issue but is rather computationally expensive in order to reach point wise convergence (M needs to be large).

As a result, an alternative to that stochastic algorithm is to use a Robbins-Monro (RM) type of update. We denote

$$\tilde{S}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}^{(k)}, y_i) \tag{6}$$

where $z_{i,m}^{(k)} \sim p(z_i|y_i;\theta^{(k)})$. At iteration k, the sufficient statistics $\hat{\mathrm{s}}^{(k+1)}$ is approximated as follows:

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

where $\{\gamma_k\}_{k=1}^{\infty} \in [0, 1]$ is a sequence of decreasing step sizes to ensure asymptotic convergence. This is called the Stochastic Approximation of the EM (SAEM), see [Delyon et al., 1999] and allows a smooth convergence to the target parameter. It represents the *first level* of our algorithm (needed to temper the variance and noise implied by MC integration).

In the next section, we derive variants of this algorithm to adapt of the sheer size of data of today's applications.

2.2 Incremental and Bi-Level Inexact EM Methods

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

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

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

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

$$
\begin{aligned}
&\textit{(iSAEM [Karimi, 2019, Kuhn et al., 2019])} && \boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + \tfrac{1}{n}\big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)}\big) && (9)\\
&\textit{(vrSAEM This paper)} && \boldsymbol{\mathcal{S}}^{(k+1)} = \tilde{S}^{(\ell(k))} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}\big) && (10)\\
&\textit{(fiSAEM This paper)} && \boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big) && (11)\\
& && \overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\big(\tilde{S}_{j_k}^{(k)} - \tilde{S}_{j_k}^{(t_{j_k}^k)}\big). && (12)
\end{aligned}
$$

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

2.3 Two-Time-Scale Noisy EM methods

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

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

Algorithm 1 Two-Time-Scale Noisy EM methods.

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

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

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

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

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

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

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

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

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

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

end for

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

---

3 Global and Finite Time Analysis of the Scheme

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

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

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

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

$$\mathsf{S} \supset \left\{ n^{-1} \sum_{i=1}^{n} u_i, u_i \in \mathrm{conv}(\overline{\mathbf{s}}_i(\boldsymbol{\theta})) \right\} \tag{16}$$

where $\overline{\mathbf{s}}_i(\boldsymbol{\theta})$ is defined in (3).

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

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

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

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

$$\mathrm{B}(\mathrm{s}) := \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) \Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathrm{s}, \overline{\boldsymbol{\theta}}(\mathrm{s})) \Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))^\top. \tag{17}$$

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

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

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

$$\eta_i^{(r)} := \tilde{S}_i^{(r)} - \overline{\mathrm{s}}_i(\vartheta^{(r)}) \tag{18}$$

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

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

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

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

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

Lemma 1. *[Karimi et al., 2019] Assume H2, H3, H4. For all* $\mathrm{s}, \mathrm{s}' \in \mathsf{S}$ *and* $i \in [\![1, n]\!]$, *we have*

$$\|\overline{\mathrm{s}}_i(\overline{\boldsymbol{\theta}}(\mathrm{s})) - \overline{\mathrm{s}}_i(\overline{\boldsymbol{\theta}}(\mathrm{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \|\mathrm{s} - \mathrm{s}'\|, \;\; \|\nabla V(\mathrm{s}) - \nabla V(\mathrm{s}')\| \leq \mathrm{L}_V \|\mathrm{s} - \mathrm{s}'\|, \tag{20}$$

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

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

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

$$\upsilon_{\min}^{-1} \langle \nabla V(\mathrm{s}) \,|\, \mathrm{s} - \overline{\mathrm{s}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) \rangle \geq \left\| \mathrm{s} - \overline{\mathrm{s}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) \right\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathrm{s})\|^2, \tag{21}$$

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 $\hat{S}^{(k+1)} - \hat{\mathrm{s}}^{(k)}$, which corresponds to the dirft term of (7) and reads as follows:

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

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

Also:

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

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

See proofs of this Lemma in Appendix B.

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

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

Assume that $\hat{\mathbf{s}}^{(k)} \in \mathcal{S}$ 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 $\left\{\hat{\mathbf{s}}^{(k)}, k \in \mathbb{N}\right\}$ obtained with $\rho_{k+1} = \rho$ for any $k > 0$.*

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

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 $z_i \in [\![M]\!]$ be the latent labels of each component, the complete log-likelihood is defined as:

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

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

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

Exact two time scale updates are given in Appendix **??**.

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 $\mu^\star$ averaged on 50 datasets. We compare the bEM, SAEM, iSAEM, vrSAEM and fiSAEM methods in terms of their precision measured by $|\mu - \mu^\star|^2$. We set the stepsize of the SA-step of all method as $\gamma_k = 1/k^\alpha$ with $\alpha = 0.5$, and the stepsizes of the Incremental-step for 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.

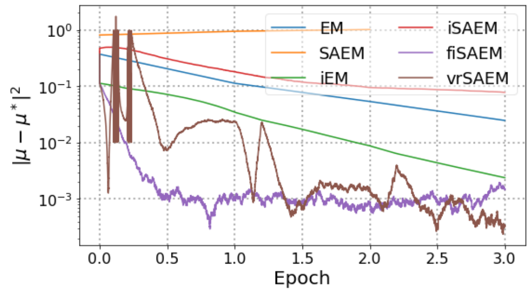

Figure 1: TO COMPLETE

4.2 Deformable Template Model for Image Analysis

We now run our different methods using an example taken from [Allassonnière et al., 2010]. Let $(y_i, i \in [\![1, n]\!])$ be observed images. Let $u \in \mathcal{U} \subset \mathbb{R}^2$ denote the pixel index on the image and $x_u \in \mathcal{D} \subset \mathbb{R}^2$ its location.

The model used in this experiment suggests that each image y_i is a deformation of a template, noted $I : \mathcal{D} \to \mathbb{R}$, common to all images of the dataset:

$$y_i(u) = I\left(x_u - \Phi_i\left(x_u, z_i\right)\right) + \varepsilon_i(u) \tag{26}$$

where $\phi_i : \mathbb{R}^2 \to \mathbb{R}^2$ 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 $(p_k, k \in [\![1, k_p]\!])$ landmarks on the template, a fixed known kernel $\mathbf{K_p}$ and a vector of parameters $\beta \in \mathbb{R}^{k_p}$ is defined as follows:

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

Besides, we parameterize the deformation model given some landmarks $(g_k, k \in [\![1, k_g]\!])$ and a fixed kernel $\mathbf{K_g}$ as:

$$\Phi_i(x, z_i) = \left(\mathbf{K_g} z_i\right)(x) = \sum_{k=1}^{k_s} \mathbf{K_g}\left(x, g_k\right)\left(z_i^{(1)}(k), z_i^{(2)}(k)\right) \tag{28}$$

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

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

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

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

Finally, the maximization step yields the following parameter updates:

$$\bar{\boldsymbol{\theta}}(S) = \begin{pmatrix} \beta(S) = S_2^{-1}(z) S_1(z) \\ \Gamma(S) = \frac{1}{n} S_3(z) \\ \sigma(S) = \beta(S)^\top S_2(z)\beta(S) - 2\beta(SS_1(z) \end{pmatrix} \tag{31}$$

5 Conclusion

References

S. Allassonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):3–29, 2007.

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

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.

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

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

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.

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

A Proof of Lemma 2

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

$$\upsilon_{\min}^{-1}\langle \nabla V(\mathrm{s}) \,|\, \mathrm{s} - \overline{s}(\overline{\boldsymbol{\theta}}(\mathrm{s}))\rangle \geq \left\|\mathrm{s} - \overline{s}(\overline{\boldsymbol{\theta}}(\mathrm{s}))\right\|^2 \geq \upsilon_{\max}^{-2}\|\nabla V(\mathrm{s})\|^2, \tag{32}$$

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

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

Consider the following vector map:

$$\mathrm{s} \to \nabla_{\boldsymbol{\theta}} L(\mathrm{s}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\mathrm{s})} = \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathrm{s})) + \nabla_{\boldsymbol{\theta}}\, \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathrm{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))^\top \mathrm{s} \,. \tag{34}$$

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

$$\mathbf{0} = -\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) + \Big(\underbrace{\nabla_{\boldsymbol{\theta}}^2\big(\psi(\boldsymbol{\theta}) + \mathrm{r}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) \,|\, \mathrm{s}\rangle\big)}_{=\mathrm{H}_L^{\boldsymbol{\theta}}(\mathrm{s};\boldsymbol{\theta})}|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\mathrm{s})}\Big)\, \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathrm{s}) \,. \tag{35}$$

The above yields

$$\nabla_{\mathbf{s}} V(\mathrm{s}) = \mathrm{B}(\mathrm{s})(\mathrm{s} - \overline{s}(\overline{\boldsymbol{\theta}}(\mathrm{s}))) \tag{36}$$

where we recall $\mathrm{B}(\mathrm{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))\Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathrm{s}; \overline{\boldsymbol{\theta}}(\mathrm{s}))\Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))^\top$. The proof of (32) follows directly from the assumption H4. □

B Proof of Lemma 3

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

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

Also:

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

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

Proof From update (9), we have:

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

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

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

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

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

The following equalities:

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

concludes the proof of the Lemma. □

C Proof of Theorem 1

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

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

TO COMPLETE WITH BOUND

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

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

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

Proof Applying the iSAEM update yields:

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

The last expectation can be further bounded by

$$\frac{2}{n^2}\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{2}{n^3}\sum_{i=1}^{n}\mathbb{E}[\|\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{(t_i^k)}\|^2] \overset{(a)}{\leq} \frac{2\,\mathrm{L}_{\mathbf{s}}}{n^3}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2], \tag{45}$$

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

□

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

$$V(\hat{\boldsymbol{s}}^{(k+1)}) \leq V(\hat{\boldsymbol{s}}^{(k)}) + \gamma_{k+1}\langle\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \mid \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 \tag{46}$$

Taking the expectation on both sidesyields:

$$\mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k+1)})\right] \leq \mathbb{E}\left[V(\hat{\boldsymbol{s}}^{(k)})\right] + \gamma_{k+1}\mathbb{E}\left[\langle\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \mid \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle\right] + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\mathbb{E}\left[\|\tilde{S}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2\right] \tag{47}$$

Using Lemma 3, we obtain:

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

where (a) is due to the growth condition (2) and (b) is due to Young's inequality (with $\beta \to 1$). Note $a_k = \gamma_{k+1}\left(\upsilon_{\min} - \upsilon_{\max}^2\frac{\beta(n-1)+1}{2n}\right)$ and

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

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

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

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

$$
\sum_{k=0}^{K_{\max}} \alpha_k\mathbb{E}\left[\left\|\nabla V(\hat{\boldsymbol{s}}^{(k)})\right\|^2\right] \leq \upsilon_{\max}^2 \sum_{k=0}^{K_{\max}} \alpha_k\mathbb{E}\left[\left\|\overline{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k)}\right\|^2\right] \tag{51}
$$

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

D Proof of Theorem 2

Theorem. *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 $\{\hat{\mathrm{s}}^{(k)}, k \in \mathbb{N}\}$ obtained with $\rho_{k+1} = 1$ for any $k > 0$.*

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

TO COMPLETE WITH BOUND

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

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

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \leq & 4\rho^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}\|^2] + \frac{4\rho^2\,\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\ & + 2\rho^2\frac{C_\eta}{M_k} + 4(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]\end{aligned} \tag{52}$$

Proof Applying the fiSAEM update yields:

$$\begin{aligned}&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] = \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} + \rho(\tilde{S}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)})\|^2] \\ &= \mathbb{E}[\|(1-\rho)(\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}) + \rho(\hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}) + \rho\left[(\overline{\mathbf{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)}) - (\tilde{S}_{i_k}^{(k)}) - \tilde{S}_{i_k}^{(\tau_i^k)}\right]\|^2]\end{aligned} \tag{53}$$

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

$$\begin{aligned}&\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \\ &\leq 4(1-\rho)^2\mathbb{E}\big[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2\big] + 4\rho^2\mathbb{E}\big[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\mathbf{s}}^{(k)}\|^2\big] + 2\rho^2\mathbb{E}\Big[\Big\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\Big\|^2\Big] + 2\rho^2\frac{C_\eta}{M_k}\end{aligned} \tag{54}$$

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

$$\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{(t_i^k)}\|^2] \overset{(a)}{\leq} \frac{\mathrm{L}_{\mathbf{s}}}{n}\sum_{i=1}^{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2], \tag{55}$$

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

□

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

$$\begin{aligned}V(\hat{\boldsymbol{s}}^{(k+1)}) &\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1}\big\langle\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \\ &\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1}\big\langle\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle \\ &- \gamma_{k+1}\rho\big\langle\tilde{S}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2 \\ &\leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1}\rho\big\langle\hat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle \\ &- \gamma_{k+1}(1-\rho)\big\langle\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\,|\,\nabla V(\hat{\boldsymbol{s}}^{(k)})\big\rangle + \frac{\gamma_{k+1}^2\,\mathrm{L}_V}{2}\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2\end{aligned} \tag{56}$$

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

$$
\begin{aligned}
\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)}) - V(\hat{\boldsymbol{s}}^{(k)})] \leq & -\gamma_{k+1}\rho\mathbb{E}[\langle \hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle] \\
& - \gamma_{k+1}(1-\rho)\mathbb{E}[\langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle] + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2] \\
& \overset{(a)}{\leq} -(\gamma_{k+1}\upsilon_{\max}^2\frac{(1-\rho)}{2} + \gamma_{k+1}\rho)\mathbb{E}\left[\left\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\right\|^2\right] - \gamma_{k+1}\frac{1-\rho}{2}\mathbb{E}[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2] \\
& + \frac{\gamma_{k+1}^2 \mathrm{L}_V}{2}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)}\|^2]
\end{aligned}
\tag{57}
$$

where (a) used the growth condition (32) twice (on $\langle \hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)})\rangle$ and $\left\|\nabla V(\hat{\boldsymbol{s}}^{(k)}\right\|^2$ and the triangle inequality.

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

$$
\begin{aligned}
\mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)}) - V(\hat{\boldsymbol{s}}^{(k)})] \leq & - \gamma_{k+1}\big(\upsilon_{\max}^2\frac{(1-\rho)}{2} + \rho - 2\rho^2\gamma_{k+1}\mathrm{L}_V\big)\mathbb{E}\left[\left\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\right\|^2\right] \\
& + \frac{2\gamma_{k+1}^2 \mathrm{L}_V \rho^2 \mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
& + \gamma_{k+1}(1-\rho)\big(2\gamma_{k+1}\mathrm{L}_V(1-\rho) - \frac{1}{2}\big)\mathbb{E}[\left\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\right\|^2] \\
& + \gamma_{k+1}^2\rho^2 \mathrm{L}_V \frac{C_\eta}{M_k}
\end{aligned}
\tag{58}
$$

Next, we observe that

$$
\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n}\sum_{i=1}^n \Big(\frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n}\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2]\Big) \tag{59}
$$

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

$$
\begin{aligned}
& \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
& = \mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\langle \hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\rangle\Big] \\
& = \mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 - 2\gamma_{k+1}\langle \hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\rangle\Big] \\
& \leq \mathbb{E}\Big[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2 + \|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 + \gamma_{k+1}\beta\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2\Big]
\end{aligned}
\tag{60}
$$

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

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] \\
& \leq \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n^2}\sum_{i=1}^n \mathbb{E}\Big[(1+\gamma_{k+1}\beta)\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta}\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2\Big]
\end{aligned}
\tag{61}
$$

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

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] \leq & \left(4\rho^2 + \frac{\gamma_{k+1}}{\beta}\right)\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \\
& + \left(\frac{4\rho^2 \mathrm{L}_{\mathbf{s}}^2}{n} + \frac{(n-1)(1+\gamma_{k+1}\beta)}{n^2}\right)\sum_{i=1}^n \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
& + 2\rho^2\frac{C_\eta}{M_k} + 4(1-\rho)^2\mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2]
\end{aligned}
\tag{62}
$$

Define

$$\Delta^{(k)} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \tag{63}$$

and note that

$$\begin{aligned} \Delta^{(k+1)} \leq & \left(1 - \frac{1}{n} + 4\rho^2 \, \mathrm{L}_{\mathbf{s}}^2 + \gamma_{k+1}\beta\right) \Delta^{(k)} + \left(4\rho^2 + \frac{\gamma_{k+1}}{\beta}\right) \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \\ & + 2\rho^2 \frac{C_\eta}{M_k} + 4(1-\rho)^2 \mathbb{E}[\|\hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k)}\|^2] \end{aligned} \tag{64}$$

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

FIND AN UPPER BOUND TO THAT GAP

$\square$

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 $\boldsymbol{\theta}$ is given by

$$\Theta = \Delta^M \times \mathbb{R}^M. \tag{65}$$

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

$$\begin{aligned} &s_{i,m}^{(1)} = \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{\log(\omega_m) - \frac{\mu_m^2}{2}\right\} - \left\{\log(1 - \textstyle\sum_{j=1}^{M-1}\omega_j) - \frac{\mu_M^2}{2}\right\} , \\ &s_{i,m}^{(2)} = \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m \ , \quad s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M \ , \end{aligned} \tag{66}$$

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

$$z_{i,m} \sim \mathbb{P}\left(z_i = m | y_i; \boldsymbol{\theta}\right) \tag{67}$$

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

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

$$\tilde{S}_{i_k}^{(k)} = (\underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}), ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1})}_{:=\tilde{s}_{i_k}^{(1)}}, \underbrace{\mathbb{1}_{\{1\}}(z_{i_k,1}) y_{i_k}, ..., \mathbb{1}_{\{M-1\}}(z_{i_k,M-1}) y_{i_k}}_{:=\tilde{s}_{i_k}^{(2)}}, \underbrace{y_{i_k}}_{:=\bar{\mathbf{s}}_{i_k}^{(3)}(\boldsymbol{\theta}^{(k)})})^\top . \tag{68}$$

Recall that we have used the following regularizer:

$$\mathrm{r}(\boldsymbol{\theta}) = \tfrac{\delta}{2} \textstyle\sum_{m=1}^M \mu_m^2 - \epsilon \sum_{m=1}^M \log(\omega_m) - \epsilon \log\left(1 - \sum_{m=1}^{M-1} \omega_m\right) , \tag{69}$$

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

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1+\epsilon M)^{-1}\left(s_1^{(1)} + \epsilon, \dots, s_{M-1}^{(1)} + \epsilon\right)^\top \\ \left((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \dots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)}\right)^\top \\ \left(1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta\right)^{-1}\left(s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}\right) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{70}$$

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

E.1.2 Algorithms updates

In the sequel, recall that, for all $i \in [\![n]\!]$ and iteration k, the computed statistic $\tilde{S}_{i_k}^{(k)}$ is defined by (68). 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 $\hat{\mathbf{s}}^{(0)} = n^{-1} \sum_{i=1}^n \bar{\mathbf{s}}_i^{(0)}$, those E-steps break down as follows:

Batch EM (EM): for all $i \in [\![1, n]\!]$, compute $\bar{\mathbf{s}}_i^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = n^{-1} \sum_{i=1}^n \bar{\mathbf{s}}_i^{(k)} \ . \tag{71}$$

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

$$\widetilde{\omega}_m(y_i;\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y=y_i] = \frac{\omega_m \exp(-\frac{1}{2}(y_i-\mu_i)^2)}{\sum_{j=1}^M \omega_j \exp(-\frac{1}{2}(y_i-\mu_j)^2)} \ , \tag{72}$$

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

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \frac{1}{n}\big(\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\tau_i^k)}\big) = n^{-1}\sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(\tau_i^k)} \ . \tag{73}$$

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

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\tilde{S}^{(k)} \ . \tag{74}$$

where $= \frac{1}{n}\sum_{i=1}^n \tilde{S}_i^{(k)}$ with $\tilde{S}_i^{(k)}$ defined in (68).

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

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)} + \frac{1}{n}(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_i^k)})\big) \ . \tag{75}$$

Variance Reduced Two-Time-Scale EM (vrSAEM): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)}(1-\rho) + \rho(\tilde{S}^{(\ell(k))} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}\big))\big) \ . \tag{76}$$

Fast Incremental Two-Time-Scale EM (fiSAEM): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)}(1-\gamma_{k+1}) + \gamma_{k+1}\big(\tilde{S}^{(k)}(1-\rho) + \rho(\overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)\big) \ . \tag{77}$$

Finally, the k-th update reads $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$ where the function $\boldsymbol{s} \to \overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined by (70).