

---

# 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.

## 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:

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

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,

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

30 The M-step is given by

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

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 compute  $\hat{\theta} = \bar{\theta}(\hat{s})$ .

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

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

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

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

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

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

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

## 96 2.2 Incremental and Bi-Level Inexact EM Methods

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

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

100 Note  $\{\rho_k\}_{k=1}^\infty \in [0, 1]$  is a sequence of step sizes,  $\mathcal{S}^{(k)}$  is a proxy for  $\tilde{S}^{(k)}$ , If the stepsize is equal  
 101 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  
 102 the SAEM algorithm. Also if  $\rho_k = 1$ ,  $\gamma_k = 1$  and  $\mathcal{S}^{(k)} = \tilde{S}^{(k)}$ , then we recover the Monte Carlo  
 103 EM algorithm.

104 We now introduce three variants of the SAEM update depending on different definitions of the proxy  
 105  $\mathcal{S}^{(k)}$  and the choice of the stepsize  $\rho_k$ . Let  $i_k \in \llbracket 1, n \rrbracket$  be a random index drawn at iteration  $k$  and  
 106  $\tau_i^k = \max\{k' : i_{k'} = i, k' < k\}$  be the iteration index where  $i \in \llbracket 1, n \rrbracket$  is last drawn prior to  
 107 iteration  $k$ . For iteration  $k \geq 0$ , the fiSAEM method draws *two* indices *independently* and uniformly  
 108 as  $i_k, j_k \in \llbracket 1, n \rrbracket$ . In addition to  $\tau_i^k$  which was defined w.r.t.  $i_k$ , we define  $t_j^k = \{k' : j_{k'} = j, k' <$   
 109  $k\}$  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  
 110 the initialization  $\bar{\mathcal{S}}^{(0)} = \bar{s}^{(0)}$ , we use a slightly different update rule from SAGA inspired by [Reddi

111 et al., 2016]. Then, we obtain:

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

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

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

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

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

### 116 2.3 Two-Time-Scale Noisy EM methods

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

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

---

#### Algorithm 1 Two-Time-Scale Noisy EM methods.

---

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

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

- 3: **for**  $k = 0, 1, 2, \dots, K$  **do**
- 4: Draw index  $i_k \in \llbracket 1, n \rrbracket$  uniformly (and  $j_k \in \llbracket 1, n \rrbracket$  for 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  $\tilde{S}^{(k+1)}$  and  $\hat{s}^{(k+1)}$  using respectively (8) and (7):

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

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

## 122 3 Global and Finite Time Analysis of the Scheme

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

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

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

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

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

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

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

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

131 **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  
 132 minimum  $\bar{\theta}(s) \in \text{int}(\Theta)$ . In addition,  $J_\phi^\theta(\bar{\theta}(s))$  is full rank and  $\bar{\theta}(s)$  is  $L_\theta$ -Lipschitz.

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

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

135 **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  
 136 a constant  $L_B$  such that for all  $s, s' \in S^2$ , we have  $\|B(s) - B(s')\| \leq L_B \|s - s'\|$ .

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

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

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

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

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

150 **H5.** There exist a positive sequence of MC batch size  $\{M_k\}_{k>0}$  and constants  $(C, C_\eta)$  such that for  
 151 all  $k > 0$ ,  $i \in \llbracket 1, n \rrbracket$  and  $\vartheta \in \Theta$ :

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

152 **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

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

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

### 154 3.1 Global Convergence of Incremental Noisy EM Algorithms

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

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

159 **Lemma 2.** Assume H1. The update (9) is equivalent to the following update on the resulting statis-  
 160 tics

$$\hat{s}^{(k+1)} = \hat{s}^{(k)} + \gamma_{k+1} \left( n^{-1} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^k)} - \hat{s}^{(k)} \right) \quad (21)$$

161 where  $\tau_i^k = \max\{k' : i_{k'} = i, k' < k\}$ . Also:

$$\mathbb{E} \left[ \left\| \tilde{S}^{(k+1)} - \hat{s}^{(k)} \right\|^2 \right] \leq \mathbb{E} \left[ \left\| \bar{s}^{(k)} - \hat{s}^{(k)} \right\|^2 \right] + 2L_s^2 \left( 1 - \frac{1}{n} \right)^2 n^{-1} \sum_{i=1}^n \mathbb{E} \left[ \left\| \hat{s}^{(k)} - \hat{s}^{(\tau_i^k)} \right\|^2 \right] + \frac{2C}{M_k} \quad (22)$$

162 where  $\bar{s}^{(k)}$  is defined by (3).

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

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

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

### 169 3.2 Global Convergence of Two-Time-Scale Noisy EM Algorithms

170 We now proceed by giving our main result regarding the global convergence of the fiSAEM algo-  
 171 rithm.

172 See proof in Appendix A.

173 **TO COMPLETE**

## 174 4 Numerical Examples

### 175 4.1 Gaussian Mixture Models

176 Given  $n$  observations  $\{y_i\}_{i=1}^n$ , we want to fit a Gaussian Mixture Model (GMM) whose distribution  
 177 is modeled as a Gaussian mixture of  $M$  components, each with a unit variance. Let  $z_i \in \llbracket M \rrbracket$  be  
 178 the latent labels of each component, the complete log-likelihood is defined as:

$$\log f(z_i, y_i; \theta) = \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) [\log(\omega_m) - \mu_m^2/2] + \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) \mu_m y_i + \text{constant} . \quad (23)$$

179 where  $\theta := (\omega, \mu)$  with  $\omega = \{\omega_m\}_{m=1}^{M-1}$  are the mixing weights with the convention  $\omega_M =$   
 180  $1 - \sum_{m=1}^{M-1} \omega_m$  and  $\mu = \{\mu_m\}_{m=1}^M$  are the means. We use the penalization  $r(\theta) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 -$   
 181  $\log \text{Dir}(\omega; M, \epsilon)$  where  $\delta > 0$  and  $\text{Dir}(\cdot; M, \epsilon)$  is the  $M$  dimensional symmetric Dirichlet distribu-  
 182 tion with concentration parameter  $\epsilon > 0$ . The constraint set on  $\theta$  is given by

$$\Theta = \{\omega_m, m = 1, \dots, M-1 : \omega_m \geq 0, \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R}, m = 1, \dots, M\}. \quad (24)$$

183 In the following experiments of synthetic data, we generate samples from a GMM model with  $M =$   
 184 2 components with two mixtures with means  $\mu_1 = -\mu_2 = 0.5$ .

185 We use  $n = 10^4$  synthetic samples and run the bEM method until convergence (to double preci-  
 186 sion) to obtain the ML estimate  $\mu^*$ . We compare the bEM, SAEM, iSAEM, vrSAEM and fiSAEM  
 187 methods in terms of their precision measured by  $|\mu - \mu^*|^2$ . The left plot of Figure ?? shows the  
 188 convergence of the precision  $|\mu - \mu^*|^2$  for the different methods against the epoch(s) elapsed (one  
 189 epoch equals  $n$  iterations). We observe that the vrSAEM and fiSAEM methods outperform the other  
 190 methods, supporting our analytical results.

### 191 4.2 Deformable Template Model for Image Analysis

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

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

$$y_i(u) = I(x_u - \Phi_i(x_u)) + \varepsilon_i(u) \quad (25)$$

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

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

$$I_\xi = \mathbf{K}_p \beta, \quad \text{where} \quad (\mathbf{K}_p \beta)(x) = \sum_{k=1}^{k_p} \mathbf{K}_p(x, p_k) \beta_k \quad (26)$$

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

$$\Phi_i(x) = (\mathbf{K}_{\mathbf{g}} z_i)(x) = \sum_{k=1}^{k_s} \mathbf{K}_{\mathbf{g}}(x, g_k) \left( z_i^{(1)}(k), z_i^{(2)}(k) \right) \quad (27)$$

202 where  $z_i \sim (0, \Gamma)$  and  $z_i \in (\mathbb{R}^{k_g})^2$ .

## 203 **5 Conclusion**

## References

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



- 244 H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statis-*  
245 *tics*, pages 400–407, 1951.
- 246 G. C. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man’s  
247 data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704,  
248 1990.
- 249 R. Zhu, L. Wang, C. Zhai, and Q. Gu. High-dimensional variance-reduced stochastic gradient  
250 expectation-maximization algorithm. In *Proceedings of the 34th International Conference on*  
251 *Machine Learning-Volume 70*, pages 4180–4188. JMLR. org, 2017.

## 252 A Proof of Theorem 1

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

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

256 **Proof** Under some regularity conditions of the Lyapunov function  $V$ , cf. Lemma 20, and the fol-  
 257 lowing growth condition for all  $\mathbf{s} \in \mathcal{S}$ ,

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \quad (28)$$

258 proven in [Karimi et al., 2019, Lemma 3], we can write:

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

259 Taking the expectation on both sides and using the growth condition (28), we obtain:

$$\begin{aligned} \mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] &\leq \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1} v_{\min} \mathbb{E} \left[ \left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] + \mathbb{E} \left[ \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)}\|^2 \right] \\ &\quad - \gamma_{k+1} \mathbb{E} \left[ \langle \bar{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] \end{aligned} \quad (30)$$

260 We then establish an auxiliary Lemma yielding an upper-bound on the quantity

261  $\mathbb{E} \left[ \langle \bar{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right]$  where:

$$\bar{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} = \bar{\mathbf{s}}^{(k)} - \left( \tilde{\mathbf{S}}^{(k)} + \frac{1}{n} (\tilde{\mathbf{S}}_{i_k}^{(k)} - \tilde{\mathbf{S}}_{i_k}^{(\tau_{i_k}^k)}) \right) \quad (31)$$

262

**Lemma 3.**

$$\mathbb{E} \left[ \langle \bar{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] \leq \quad (32)$$

263 Using Lemma 2:

$$\begin{aligned} \mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] &\leq \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1} \left( v_{\min} - \frac{\gamma_{k+1} L_V}{2} \right) \mathbb{E} \left[ \left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \\ &\quad + \gamma_{k+1}^2 L_V L_s^2 \left( 1 - \frac{1}{n} \right)^2 n^{-1} \sum_{i=1}^n \mathbb{E} \left[ \left\| \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)} \right\|^2 \right] + \frac{\gamma_{k+1}^2 L_V C}{M_k} \\ &\quad - \gamma_{k+1} \mathbb{E} \left[ \langle \bar{\mathbf{s}}^{(k)} - \tilde{\mathbf{S}}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] \end{aligned} \quad (33)$$

264 Besides,

$$n^{-1} \sum_{i=1}^n \mathbb{E} \left[ \left\| \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\tau_i^{k+1})} \right\|^2 \right] = n^{-1} \sum_{i=1}^n \left( \frac{1}{n} \mathbb{E} [\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n} \mathbb{E} [\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2] \right) \quad (34)$$

265 yielding for any numbers  $\beta_k > 0$ ,

$$\begin{aligned} &\mathbb{E} [\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2] \\ &= \mathbb{E} \left[ \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + 2 \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} | \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle \right] \\ &= \mathbb{E} \left[ \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 - 2 \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)} | \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)} \rangle \right] \\ &\leq \mathbb{E} \left[ \|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta_{k+1}} \|\hat{\mathbf{s}}^{(k)} - \bar{\mathbf{s}}^{(k)}\|^2 + \gamma_{k+1} \beta_{k+1} \|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(t_i^k)}\|^2 \right] \end{aligned} \quad (35)$$

266

□