
Fast Two-Time-Scale Noisy EM Algorithms

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 T.B.C

2 1 Introduction

3 We formulate the following empirical risk minimization as:

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

4 where $\{y_i\}_{i=1}^n$ are the observations, Θ is a convex subset of \mathbb{R}^d for the parameters, $R : \Theta \rightarrow \mathbb{R}$ is a
5 smooth convex regularization function and for each $\theta \in \Theta$, $g(y; \theta)$ is the (incomplete) likelihood of
6 each individual observation. The objective function $\bar{\mathcal{L}}(\theta)$ is possibly *non-convex* and is assumed to
7 be lower bounded $\bar{\mathcal{L}}(\theta) > -\infty$ for all $\theta \in \Theta$. In the latent variable model, $g(y_i; \theta)$, is the marginal
8 of the complete data likelihood defined as $f(z_i, y_i; \theta)$, i.e. $g(y_i; \theta) = \int_{\mathcal{Z}} f(z_i, y_i; \theta) \mu(dz_i)$, where
9 $\{z_i\}_{i=1}^n$ are the (unobserved) latent variables. We make the assumption of a complete model be-
10 longing to the curved exponential family, *i.e.*,

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

11 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
12 the complete data sufficient statistics.

13 **Prior Work** Cite Kuhn [Kuhn et al., 2019] (for ISAEM) and incremental EM like papers. As well
14 as Optim papers (Variance reduction, SAGA etc.)

15 2 Expectation Maximization Algorithm

16 Full batch EM is a two steps procedure. The **E-step** amounts to computing the conditional expecta-
17 tion of the complete data sufficient statistics,

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

18 The **M-step** is given by

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

19 3 Monte Carlo Integration and Stochastic Approximation

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

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

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

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

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

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

29 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:

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

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

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

36 4 Incremental and Bi-Level Inexact EM Methods

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

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

40 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
 41 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
 42 the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\mathcal{S}^{(k)} = \tilde{S}^{(k)}$, then we recover the Monte Carlo
 43 EM algorithm.

44 We now introduce three variants of the SAEM update depending on different definitions of the proxy
 45 $\mathcal{S}^{(k)}$ and the choice of the stepsize ρ_k . Let $i_k \in \llbracket 1, n \rrbracket$ be a random index drawn at iteration k and
 46 $\tau_i^k = \max\{k' : i_{k'} = i, k' < k\}$ be the iteration index where $i \in \llbracket 1, n \rrbracket$ is last drawn prior to
 47 iteration k . For iteration $k \geq 0$, the fiSAEM method draws *two* indices *independently* and uniformly
 48 as $i_k, j_k \in \llbracket 1, n \rrbracket$. In addition to τ_i^k which was defined w.r.t. i_k , we define $t_j^k = \{k' : j_{k'} = j, k' <$
 49 $k\}$ 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
 50 the initialization $\bar{\mathcal{S}}^{(0)} = \bar{s}^{(0)}$, we use a slightly different update rule from SAGA inspired by [Reddi

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

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

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

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

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

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 $\mathcal{S}^{(0)} = \tilde{S}^{(0)}$; 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.

Algorithm 1 Two-Time-Scale Noisy EM methods.

- 1: **Input:** initializations $\hat{\theta}^{(0)} \leftarrow 0$, $\hat{s}^{(0)} \leftarrow \hat{S}^{(0)}$, $K_{\max} \leftarrow \max.$ 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 the surrogate sufficient statistics $\mathcal{S}^{(k+1)}$ using (9) or (10) or (11) and using the MC-step (5) to compute the Monte Carlo approximations.
 - 6: Compute $\hat{S}^{(k+1)}$ via the Inexact-step (8).
 - 7: Compute $\hat{s}^{(k+1)}$ via the SA-step (7).
 - 8: Compute $\hat{\theta}^{(k+1)}$ via the M-step (4).
 - 9: **end for**
 - 10: **Return:** $\hat{\theta}^{(K)}$.
-

5 Finite Time Analysis

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

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

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 \mathcal{Z}, \mathcal{S} are measurable open sets such that:

$$\mathcal{S} \supset \left\{ n^{-1} \sum_{i=1}^n u_i, u_i \in \text{conv}(\overline{\mathbf{s}}_i(\theta)) \right\} \quad (15)$$

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

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

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

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

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

$$\mathbf{B}(\mathbf{s}) := J_{\phi}^{\theta}(\overline{\theta}(\mathbf{s})) \left(H_L^{\theta}(\mathbf{s}, \overline{\theta}(\mathbf{s})) \right)^{-1} J_{\phi}^{\theta}(\overline{\theta}(\mathbf{s}))^{\top}. \quad (16)$$

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

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

76 Indeed the expectations (3) are never available and requires Monte Carlo approximation. Thus, at
77 iteration $k + 1$, we introduce the errors when approximating the quantity $\bar{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k-1)}))$. For all
78 $i \in \llbracket 1, n \rrbracket$, $r > 0$ and $\vartheta \in \Theta$, define:

$$\eta_{i,\vartheta}^{(r)} := \tilde{S}_i^{(r)} - \bar{\mathbf{s}}_i(\vartheta) \quad (17)$$

79 For instance, we consider that the MC approximation is unbiased if for all $i \in \llbracket 1, n \rrbracket$ and $m \in$
80 $\llbracket 1, M \rrbracket$, the samples $z_{i,m} \sim p(z_i | y_i; \theta)$ are i.i.d. under the posterior distribution, i.e., $\mathbb{E}[\eta_{i,\vartheta}^{(r)} | \mathcal{F}_r] = 0$
81 where \mathcal{F}_r is the filtration up to iteration r .

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

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

$$\mathbb{E} \left[\left\| \eta_{i,\vartheta}^{(r)} \right\|^2 \right] \leq \frac{C_\eta}{M_r} \quad \text{and} \quad \mathbb{E} \left[\left\| \mathbb{E}[\eta_{i,\vartheta}^{(r)} | \mathcal{F}_r] \right\|^2 \right] \leq \frac{C}{M_r} \quad (18)$$

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

$$\|\bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}}(\mathbf{s})) - \bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq L_s \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq L_V \|\mathbf{s} - \mathbf{s}'\|, \quad (19)$$

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

88 5.1 Global Convergence of Incremental Noisy EM Algorithms

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

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

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

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_{k+1} \left(n^{-1} \sum_{i=1}^n \hat{S}_i^{(\tau_i^k)} - \hat{\mathbf{s}}^{(k)} \right) \quad (20)$$

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

$$\mathbb{E} \left[\left\| \hat{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \leq \mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] + 2L_s^2 \left(1 - \frac{1}{n} \right)^2 \mathbb{E} \left[\left\| n^{-1} \sum_{i=1}^n \tilde{S}_i^{(\tau_i^{k+1})} - \bar{\mathbf{s}}^{(k)} \right\|^2 \right] + \frac{2C}{M_k} \quad (21)$$

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

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

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

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

103 **Proof** Under some regularity conditions of the Lyapunov function V , cf. Lemma 19, and the fol-
 104 lowing growth condition for all $\mathbf{s} \in \mathcal{S}$,

$$v_{\min}^{-1} \langle \nabla V(\mathbf{s}) | \mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \|\mathbf{s} - \bar{\mathbf{s}}(\bar{\boldsymbol{\theta}}(\mathbf{s}))\|^2 \geq v_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \quad (22)$$

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

$$V(\hat{\mathbf{s}}^{(k+1)}) \leq V(\hat{\mathbf{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \hat{S}^{(k+1)}\|^2 \quad (23)$$

106 Taking the expectation on both sides and using the growth condition (22), we obtain:

$$\begin{aligned} \mathbb{E}[V(\hat{\mathbf{s}}^{(k+1)})] &\leq \mathbb{E}[V(\hat{\mathbf{s}}^{(k)})] - \gamma_{k+1} v_{\min} \mathbb{E} \left[\left\| \bar{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] + \mathbb{E} \left[\frac{\gamma_{k+1}^2 L_V}{2} \|\hat{\mathbf{s}}^{(k)} - \hat{S}^{(k+1)}\|^2 \right] \\ &\quad - \gamma_{k+1} \mathbb{E} \left[\langle \bar{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] \end{aligned} \quad (24)$$

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

108 $\mathbb{E} \left[\langle \bar{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right]$ where:

$$\bar{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} = \bar{\mathbf{s}}^{(k)} - \left(\tilde{S}^{(k)} + \frac{1}{n} (\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)}) \right) \quad (25)$$

Lemma 3.

$$\mathbb{E} \left[\langle \bar{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} | \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \right] \leq \quad (26)$$

109

□

110 5.2 Global Convergence of Two-Time-Scale Noisy EM Algorithms

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

113 6 Numerical Examples

114 6.1 Gaussian Mixture Models

115 Graphs obtained and relevant

116 6.2 Deep Latent Variable Models using noisy EM

117 See if makes sense to use EM instead of Variational Inference

118 6.3 Deformable Template Model for Image Analysis

119 See Kuhn et.al. paper.

120 7 Conclusion

References

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