Two-Timescale Stochastic EM Algorithms

Belhal Karimi and Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004
Email: {belhal.karimi, pingli98}@gmail.com

***Abstract*—The Expectation-Maximization (EM) algorithm is a popular choice for learning latent variable models. Variants of the EM have been initially introduced by [1], using incremental updates to scale to large datasets, and by [2], [3], using Monte Carlo (MC) approximations to bypass the intractable conditional expectation of the latent data for most nonconvex models. In this paper, we propose a general class of methods called Two-Timescale EM Methods based on a two-stage approach of stochastic updates to tackle an essential nonconvex optimization task for latent variable models. We motivate the choice of a double dynamic by invoking the variance reduction virtue of each stage of the method on both sources of noise: the index sampling for the incremental update and the MC approximation. We establish finite-time and global convergence bounds for nonconvex objective functions. Numerical applications on various models such as deformable template for *image analysis* or nonlinear mixed-effects models for *pharmacokinetics* are also presented to illustrate our findings.**

I. Introduction

Learning latent variable models is critical for modern machine learning problems, see (e.g.,) [4] for references. We formulate the training of such model as an empirical risk minimization problem:

$$\min_{\boldsymbol{\theta}\in\Theta} \ \overline{\mathsf{L}}(\boldsymbol{\theta}) := \mathsf{L}(\boldsymbol{\theta}) + \mathrm{r}(\boldsymbol{\theta}) \tag{1}$$

$$\text{with} \ \ \mathsf{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{L}_i(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \big\{ -\log g(y_i;\boldsymbol{\theta}) \big\} \ , \tag{2}$$

where $\{y_i\}_{i=1}^n$ are observations, $\Theta \subset \mathbb{R}^d$ is the parameters set and $\mathrm{r} : \Theta \to \mathbb{R}$ is a smooth regularizer. The objective $\overline{\mathsf{L}}(\boldsymbol{\theta})$ is possibly *nonconvex* and is assumed to be lower bounded. In the latent data model, the likelihood $g(y_i;\boldsymbol{\theta})$, is the marginal of the complete data likelihood defined as $f(z_i, y_i;\boldsymbol{\theta})$, $g(y_i;\boldsymbol{\theta}) = \int_{\mathsf{Z}} f(z_i, y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i)$, where $\{z_i\}_{i=1}^n$ are the latent variables. In this paper, we assume that the complete model belongs to the curved exponential family [5]:

$$f(z_i, y_i;\boldsymbol{\theta}) = h(z_i, y_i) \exp\left(\big\langle S(z_i, y_i) \,|\, \phi(\boldsymbol{\theta}) \big\rangle - \psi(\boldsymbol{\theta}) \right) \ , \tag{3}$$

where $\psi(\boldsymbol{\theta})$, $h(z_i, y_i)$ are scalar functions, $\phi(\boldsymbol{\theta}) \in \mathbb{R}^k$ is a vector function, and $\{S(z_i, y_i) \in \mathbb{R}^k\}_{i=1}^n$ is the vector of sufficient statistics. Batch EM [6], [7], the method of reference for (1), is comprised of two steps. The E-step computes the conditional expectation of the sufficient statistics of (3), noted $\overline{\mathrm{s}}(\boldsymbol{\theta})$:

$$\overline{\mathrm{s}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \overline{\mathrm{s}}_i(\boldsymbol{\theta}) \tag{4}$$

$$\text{where} \quad \overline{\mathrm{s}}_i(\boldsymbol{\theta}) = \int_{\mathsf{Z}} S(z_i, y_i) p(z_i|y_i;\boldsymbol{\theta}) \mu(\mathrm{d}z_i) \ , \tag{5}$$

and the M-step is given by

$$\hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\overline{\mathrm{s}}(\boldsymbol{\theta})) := \operatorname*{arg\,min}_{\vartheta\in\Theta} \ \big\{ \mathrm{r}(\vartheta) + \psi(\vartheta) - \big\langle \overline{\mathrm{s}}(\boldsymbol{\theta}) \,|\, \phi(\vartheta) \big\rangle \big\} \ . \tag{6}$$

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, at each iteration, a full pass over the dataset; and (b) the complexity of modern models makes the expectation in (4) intractable. So far, and to the best of our knowledge, both challenges have been addressed separately, as detailed in the sequel.

Prior Work: Inspired by stochastic optimization procedures, [1], [8] develop respectively an incremental and an online variant of the E-step in models where the expectation is computable, and were then extensively used and studied in [9]–[11]. Some improvements of those methods have been provided and analyzed, globally and in finite-time, in [12] where variance reduction techniques taken from the optimization literature have been efficiently applied to scale the EM algorithm to large datasets. Regarding the computation of the expectation under the posterior distribution, the Monte Carlo EM (MCEM) has been introduced in [2] where a Monte Carlo (MC) approximation for this expectation is computed. A variant of that algorithm is the Stochastic Approximation of the EM (SAEM) in [3] leveraging the power of Robbins-Monro update [13] to ensure pointwise convergence of the vector of estimated parameters using a decreasing stepsize rather than increasing the number of MC samples. The MCEM and the SAEM have been successfully applied in mixed effects models [14]–[16] or to do inference for joint modeling of time to event data coming from clinical trials in [17], unsupervised clustering in [18], variational inference of graphical models in [19] among other applications. An incremental variant of the SAEM was proposed in [20] showing positive empirical results but its analysis is limited to asymptotic consideration.

Contributions: This paper *introduces* and *analyzes* a new class of methods which purpose is to update two proxies

for the target expected quantities in a two-timescale manner. Those approximated quantities are then used to optimize the objective function (1) for modern examples and settings using the M-step of the EM algorithm. The main contributions of the paper are:

- We propose a two-timescale method based on (i) Stochastic Approximation (SA), to alleviate the problem of computing MC approximations, and on (ii) Incremental updates, to scale to large datasets. We describe in details the edges of each level of our method based on variance reduction arguments. Such class of algorithms has two advantages. First, it naturally leverages variance reduction and Robbins-Monro type of updates to tackle large-scale and highly nonlinear learning tasks. Then, it gives a simple formulation as a *scaled-gradient method* which makes the global analysis and the implementation accessible.
- We also establish global (independent of the initialization) and finite-time (true at each iteration) upper bounds on a classical sub-optimality condition in the nonconvex literature [21], [22], *i.e.,* the second order moment of the gradient of the objective function. We discuss the double dynamic of those bounds due to the two-timescale property of our algorithm update and we theoretically show the advantages of introducing variance reduction in a *Stochastic Approximation* [13] scheme.
- We stress on the originality of our theoretical findings including such MC sampling noise contrary to existing studies related to the EM where the expectations are computed exactly. Adding a layer of MC approximation and the stochastic approximation step to reduce its variance introduce some new technicalities and challenges that need careful considerations and constitues the originality of our paper on the algorithmic and theoretical plans.

In Section II we formalize both incremental and Monte Carlo variants of the EM. Then, we introduce our two-timescale class of EM algorithms for which we derive several global statistical guarantees in Section III for possibly *nonconvex* functions. Section IV is devoted to numerical illustrations. The supplementary material of this paper includes proofs of our theoretical results.

II. Two-Timescale Stochastic EM Algorithms

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

A. Monte Carlo Integration and Stochastic Approximation

As mentioned in the Introduction, for complex and possibly nonconvex models, the expectation under the posterior distribution defined in (4) is not tractable. In that case, the first solution involves computing a Monte Carlo integration of that latter. For all $i \in [n]$, where $[n] := \{1, \cdots, n\}$, draw $\{z_{i,m} \sim p(z_i|y_i;\theta)\}_{m=1}^M$ samples and compute the MC integration of $\tilde{S}$ of $\bar{\mathrm{s}}(\boldsymbol{\theta})$ defined by (4):

$$\text{MC-step}: \ \tilde{S} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}, y_i) \ . \tag{7}$$

Then update the parameter via the maximization function $\overline{\boldsymbol{\theta}}(\tilde{S})$. This algorithm bypasses the intractable expectation issue but is rather computationally expensive in order to reach point wise convergence (M needs to be large). An alternative to that stochastic algorithm is to use a Robbins-Monro (RM) type of update. We denote, at iteration k, the number of samples M_k and the following MC approximation by $\tilde{S}^{(k+1)}$:

$$\begin{aligned} &\tilde{S}^{(k+1)} := \frac{1}{n}\sum_{i=1}^{n}\tilde{S}_i^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_k}\sum_{m=1}^{M_k} S(z_{i,m}^{(k)}, y_i) \\ &\text{where} \quad z_{i,m}^{(k)} \sim p(z_i|y_i;\theta^{(k)}) \ . \end{aligned} \tag{8}$$

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

$$\text{SA-step}: \ \hat{\mathrm{s}}^{(k+1)} = \hat{\mathrm{s}}^{(k)} + \gamma_{k+1}(\tilde{S}^{(k+1)} - \hat{\mathrm{s}}^{(k)}) \ , \tag{9}$$

where $\{\gamma_k\}_{k>1} \in (0,1)$ is a sequence of decreasing stepsizes to ensure asymptotic convergence. The combination of (8) and (9) is called the Stochastic Approximation of the EM (SAEM) and has been shown to converge to a maximum likelihood of the observations under very general conditions [3]. In simple scenarios, the samples $\{z_{i,m}\}_{m=1}^M$ are conditionally independent and identically distributed with distribution $p(z_i,\theta)$. Nevertheless, in most cases, since the loss function between the observed data y_i and the latent variable z_i can be nonconvex, sampling exactly from this distribution is not an option and the MC batch is sampled by Markov Chain Monte Carlo (MCMC) algorithm [23], [24]. It has been proved in [25] that (9) converges almost surely when coupled with an MCMC procedure. This Robbins-Monro type of update constitutes the *first level* of our algorithm, needed to temper the variance and noise introduced by the Monte Carlo integration. In the next section, we derive variants of this algorithm to adapt to the sheer size of data of today's applications and formalize the *second level* of our class of two-timescale EM methods.

B. Incremental and Two-Stage Stochastic EM Methods

Efficient strategies to scale to large datasets include incremental [1] and variance reduced [26], [27] methods. We will explicit a general update that covers those latter variants and that represents the *second level* of our algorithm, *i.e.,* the incremental update of the noisy statistics $\tilde{S}^{(k+1)}$ in (8). Instead of computing its full batch noted $\tilde{S}^{(k+1)}$ as in (8), the MC approximation is incrementally evaluated through the quantity $S_{\text{tts}}^{(k+1)}$ as:

$$\text{Inc-step}: \ S_{\text{tts}}^{(k+1)} = S_{\text{tts}}^{(k)} + \rho_{k+1}\big(\boldsymbol{\mathcal{S}}^{(k+1)} - S_{\text{tts}}^{(k)}\big) \ . \tag{10}$$

Note that $\{\rho_k\}_{k>1} \in (0,1)$ is a sequence of stepsizes, $\boldsymbol{\mathcal{S}}^{(k)}$ is a proxy for $\tilde{S}^{(k)}$ defined in (8). If the stepsize is equal to 1 and $\boldsymbol{\mathcal{S}}^{(k)} = \tilde{S}^{(k)}$, i.e., computed in a full batch manner as

in (8), 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 MCEM.

Remarks on Table 1: For all methods, we define a random index drawn at iteration k, noted $i_k \in [n]$, and $\tau_i^k = \max\{k' : i_{k'} = i,\ k' < k\}$ as the iteration index where $i \in [n]$ is last drawn prior to iteration k.

Table 1 Proxies for the Incremental-step (10)

1:	iSAEM	$\boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + n^{-1}\big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\tau_{i_k}^k)}\big)$
2:	vrTTEM	$\boldsymbol{\mathcal{S}}^{(k+1)} = S_{\text{tts}}^{(\ell(k))} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(\ell(k))}\big)$
3:	fiTTEM	$\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\tilde{S}_{i_k}^{(k)} - \tilde{S}_{i_k}^{(t_{i_k}^k)}\big)$
		$\overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\big(\tilde{S}_{j_k}^{(k)} - \tilde{S}_{j_k}^{(t_{j_k}^k)}\big)$

The proposed fiTTEM method draws *two* indices *independently* and uniformly as $i_k, j_k \in [n]$. Thus, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample $j \in [n]$ is last drawn as j_k prior to iteration k in addition to τ_i^k which was defined *w.r.t.* i_k.

Recall $\tilde{S}_{i_k}^{(k)} = \frac{1}{M_k}\sum_{m=1}^{M_k} S(z_{i_k,m}^{(k)}, y_{i_k})$ where $z_{i_k,m}^{(k)}$ are samples drawn from $p(z_{i_k}|y_{i_k};\theta^{(k)})$. The stepsize in (10) is set to $\rho_{k+1} = 1$ for the iSAEM method and we initialize with $\boldsymbol{\mathcal{S}}^{(0)} = \tilde{S}^{(0)}$; $\rho_{k+1} = \rho$ is constant for the vrTTEM and fiTTEM methods. Note that we initialize as follows $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \tilde{S}^{(0)}$ for the fiTTEM which can be seen as a slightly modified version of SAGA inspired by [28]. For vrTTEM we set an epoch size of m and we define $\ell(k) := m\lfloor k/m \rfloor$ as the first iteration number in the epoch that iteration k is in.

Two-Timescale Stochastic EM methods: We now introduce the general method derived using the two variance reduction techniques described above. Algorithm 1 leverages both levels (9) and (10) in order to output a vector of fitted parameters $\hat{\boldsymbol{\theta}}^{(\mathsf{K_m})}$ where $\mathsf{K_m}$ is the total number of iterations.

Algorithm 1 Two-Timescale Stochastic EM methods.

1: **Input:** $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathbf{s}}^{(0)} \leftarrow \tilde{S}^{(0)}$, $\{\gamma_k\}_{k>0}$, $\{\rho_k\}_{k>0}$ and $\mathsf{K_m} \in \mathbb{N}^*$.
2: **for** $k = 0, 1, 2, \ldots, \mathsf{K_m} - 1$ **do**
3: Draw index $i_k \in [n]$ uniformly (and $j_k \in [n]$ for fiTTEM).
4: Compute $\tilde{S}_{i_k}^{(k)}$ using the MC-step (7), for the drawn indices.
5: Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using Lines 1, 2 or 3 in Table 1.
6: Compute $S_{\text{tts}}^{(k+1)}$ and $\hat{\mathbf{s}}^{(k+1)}$ using respectively (10) and (9):

$$\begin{aligned} S_{\text{tts}}^{(k+1)} &= S_{\text{tts}}^{(k)} + \rho_{k+1}\big(\boldsymbol{\mathcal{S}}^{(k+1)} - S_{\text{tts}}^{(k)}\big) \\ \hat{\mathbf{s}}^{(k+1)} &= \hat{\mathbf{s}}^{(k)} + \gamma_{k+1}(S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}) \end{aligned} \tag{11}$$

7: Update $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$ via the M-step.
8: **end for**

The update in (11) is said to have a two-timescale property as the stepsizes satisfy $\lim_{k\to\infty}\gamma_k/\rho_k < 1$ such that $\tilde{S}^{(k+1)}$ is updated at a faster time-scale, determined by ρ_{k+1}, than $\hat{\mathbf{s}}^{(k+1)}$, determined by γ_{k+1}. The next section introduces the main results of this paper and establishes global and finite-time bounds for the three different updates of our scheme.

III. Finite Time Analysis of the Two-Timescale Scheme

Following [8], it can be shown that stationary points of the objective function (1) corresponds to the stationary points of the following *nonconvex* Lyapunov function:

$$\min_{\mathbf{s}\in\mathsf{S}}\ V(\mathbf{s}) := \overline{\mathrm{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \mathrm{r}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\,, \tag{12}$$

that we propose to study in this article.

A. Assumptions and Intermediate Lemmas

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

A1: The sets Z, S are compact. There exist constants $C_{\mathsf{S}}, C_{\mathsf{Z}}$ such that:

$$\begin{aligned} C_{\mathsf{S}} &:= \max_{\mathbf{s},\mathbf{s}'\in\mathsf{S}} \|\mathbf{s} - \mathbf{s}'\| < \infty, \\ C_{\mathsf{Z}} &:= \max_{i\in[n]} \int_{\mathsf{Z}} |S(z, y_i)|\mu(\mathrm{d}z) < \infty. \end{aligned} \tag{13}$$

A2: For any $i \in [n]$, $z \in \mathsf{Z}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathrm{int}(\Theta)^2$, we have $\big|p(z|y_i;\boldsymbol{\theta}) - p(z|y_i;\boldsymbol{\theta}')\big| \le \mathrm{L}_p \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ where $\mathrm{int}(\Theta)$ denotes the interior of Θ.

We also recall that we consider curved exponential family models assuming the following:

A3: For any $\boldsymbol{s} \in \mathsf{S}$, the function $\boldsymbol{\theta} \mapsto L(s, \boldsymbol{\theta}) := \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \big\langle \mathbf{s} \,|\, \phi(\boldsymbol{\theta})\big\rangle$ admits a unique global minimum $\overline{\boldsymbol{\theta}}(\mathbf{s}) \in \mathrm{int}(\Theta)$.

In addition, $\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))$, the Jacobian of the function ϕ at $\boldsymbol{\theta}$, is full rank, L_p-Lipschitz and $\overline{\boldsymbol{\theta}}(\mathbf{s})$ is L_t-Lipschitz.

We denote by $\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\theta})$ the Hessian (w.r.t to $\boldsymbol{\theta}$ for a given value of $\mathbf{s}$) of the function $\boldsymbol{\theta} \mapsto L(\mathbf{s}, \boldsymbol{\theta}) = \mathrm{r}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \big\langle \mathbf{s} \,|\, \phi(\boldsymbol{\theta})\big\rangle$, and define $\mathrm{B}(\mathbf{s}) := \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}, \overline{\boldsymbol{\theta}}(\mathbf{s}))\Big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top$.

A4: It holds that $\upsilon_{\max} := \sup_{\mathbf{s}\in\mathsf{S}} \|\mathrm{B}(\mathbf{s})\| < \infty$ and $0 < \upsilon_{\min} := \inf_{\mathbf{s}\in\mathsf{S}} \lambda_{\min}(\mathrm{B}(\mathbf{s}))$. There exists a constant L_b such that for all $\mathbf{s}, \mathbf{s}' \in \mathsf{S}^2$, we have $\|\mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\| \le \mathrm{L}_b \|\mathbf{s} - \mathbf{s}'\|$.

The class of algorithms we develop in this paper is composed of two levels where the second stage corresponds to the variance reduction trick used in [12] in order to accelerate incremental methods and reduce the variance introduced by the index sampling. The first stage is the Robbins-Monro update that aims at reducing the Monte Carlo noise of $\tilde{S}^{(k+1)}$ at iteration k:

$$\eta_i^{(k)} := \tilde{S}_i^{(k)} - \overline{\mathbf{s}}_i(\vartheta^{(k)}) \quad \text{for all} \quad i \in [n] \quad \text{and} \quad k > 0\,. \tag{14}$$

For instance, we consider that the MC approximation is unbiased if for all $i \in [n]$ and $m \in [M]$, the samples $z_{i,m} \sim p(z_i|y_i;\theta)$ are i.i.d. under the posterior distribution,

i.e., $\mathbb{E}[\eta_i^{(k)}|\mathcal{F}_k] = 0$ where $\mathcal{F}_k$ is the filtration up to iteration k. The following results are derived under the assumption that the fluctuations implied by the approximation are bounded:

A 5: For all $k > 0$, $i \in [n]$, it holds: $\mathbb{E}[\|\eta_i^{(k)}\|^2] < \infty$ and $\mathbb{E}[\|\mathbb{E}[\eta_i^{(k)}|\mathcal{F}_k]\|^2] < \infty$.

Note that typically, the controls exhibited above are vanishing when the number of MC samples M_k increases with k.

We present in the following sections a finite-time and global (independent of the initialization) analysis of both the incremental and two-timescale variants our method.

B. Global Convergence of Incremental Stochastic EM Algorithms

The following result for the iSAEM algorithm is derived under the control of the Monte Carlo fluctuations as described by Assumption A5 and is built upon an intermediary Lemma, found in the full version paper, characterizing the quantity of interest $(S_{\text{tts}}^{(k+1)} - \hat{\mathbf{s}}^{(k)})$:

Theorem 1: Assume A1-A5. Consider the iSAEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ obtained with $\rho_{k+1} = 1$ for any $k \leq \mathsf{K_m}$ where $\mathsf{K_m}$ is a positive integer. Let $\{\gamma_k = 1/(k^a \alpha c_1 \overline{L})\}_{k>0}$, where $a \in (0,1)$, be a sequence of stepsizes, $c_1 = \upsilon_{\min}^{-1}$, $\alpha = \max\{8, 1 + 6\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L_s}, \mathrm{L}_V\}$, $\beta = c_1\overline{L}/n$. Then:

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

Note that, in Theorem 1, the convergence bound is composed of an initialization term $V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(\mathsf{K_m})})$ and suffers from the Monte Carlo noise introduced by the posterior sampling step, see the second term on the RHS of the inequality. We observe, in the next section, that when variance reduction is applied ($\rho_k < 1$), a second phase of convergence will be included in our bounds.

C. Global Convergence of Two-Timescale Stochastic EM Algorithms

We now deal with the analysis of Algorithm 1 when variance reduction is applied *i.e.,* $\rho < 1$. Two important intermediate Lemmas are developed in the full version of this paper and lead to the finite-time bounds for the vrTTEM and the fiTTEM methods that we describe below. Let K be an independent discrete r.v. drawn from $\{1, \ldots, \mathsf{K_m}\}$ with distribution $\{\gamma_{k+1}/\mathsf{P_m}\}_{k=0}^{\mathsf{K_m}-1}$, then, for any $\mathsf{K_m} > 0$, the convergence criterion used in our study reads

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

where $\mathsf{P_m} = \sum_{\ell=0}^{\mathsf{K_m}-1} \gamma_\ell$ and the expectation is over the stochasticity of the algorithm. Denote $\Delta V := V(\hat{\boldsymbol{s}}^{(0)}) - V(\hat{\boldsymbol{s}}^{(\mathsf{K_m})})$ and $\|\Delta S\|^2 := \|\hat{\boldsymbol{s}}^{(k)} - S_{\text{tts}}^{(k)}\|^2$. We now state the main result regarding the vrTTEM method:

Theorem 2: Assume A1-A5. Consider the vrTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K_m}$ where $\mathsf{K_m}$ is a positive integer. Let $\{\gamma_{k+1} = 1/(k^a \overline{L})\}_{k>0}$, where $a \in (0,1)$, be a sequence of stepsizes, $\overline{L} = \max\{\mathrm{L_s}, \mathrm{L}_V\}$, $\rho = \mu/(c_1 \overline{L} n^{2/3})$, $m = nc_1^2/(2\mu^2 + \mu c_1^2)$ and a constant $\mu \in (0,1)$. Then:

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

Furthermore, the fiTTEM method has the following convergence rate:

Theorem 3: Assume A1-A5. Consider the fiTTEM sequence $\{\hat{\mathbf{s}}^{(k)}\}_{k>0} \in \mathcal{S}$ for any $k \leq \mathsf{K_m}$ where $\mathsf{K_m}$ be a positive integer. Let $\{\gamma_{k+1} = 1/(k^a \alpha c_1 \overline{L})\}_{k>0}$, where $a \in (0,1)$, be a sequence of positive stepsizes, $\alpha = \max\{2, 1 + 2\upsilon_{\min}\}$, $\overline{L} = \max\{\mathrm{L_s}, \mathrm{L}_V\}$, $\beta = 1/(\alpha n)$, $\rho = 1/(\alpha c_1 \overline{L} n^{2/3})$ and $c_1(k\alpha - 1) \geq c_1(\alpha - 1) \geq 2$, $\alpha \geq 2$. Then:

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

Note that in those two bounds, the quantities $\tilde{\eta}^{(k+1)}$ and $\Xi^{(k+1)}$ depend only on the Monte Carlo noises $\mathbb{E}[\|\eta_{i_k}^{(k)}\|^2]$, $\mathbb{E}[\|\mathbb{E}[\eta_i^{(r)}|\mathcal{F}_r]\|^2]$, bounded under Assumption A5, and some constants.

Remarks: Theorem 2 and Theorem 3 exhibit in their convergence bounds *two different phases*. The upper bounds display a *bias term* due to the initial conditions, *i.e.,* the term ΔV, and a *double dynamic* burden exemplified by the term $\mathbb{E}[\|\Delta S\|^2$. Indeed, the following remarks are worth doing on this quantity: (i) This term is the price we pay for the two-timescale dynamic and corresponds to the gap between the two *asynchronous* updates (one on $\hat{\boldsymbol{s}}^{(k)}$ and the other on $\tilde{S}^{(k)}$). (ii) It is readily understood that if $\rho = 1$, *i.e.,* there is no variance reduction, then for any $k > 0$

$$\mathbb{E}[\|\Delta S\|^2] = \mathbb{E}[\|\boldsymbol{\mathcal{S}}^{(k+1)} - S_{\text{tts}}^{(k+1)}\|^2] = 0 \, ,$$

with $\hat{\boldsymbol{s}}^{(0)} = \tilde{S}^{(0)} = 0$, which strengthen the fact that this quantity characterizes the impact of the variance reduction technique introduced in our class of methods.

The following Lemma characterizes this gap:

Lemma 1: Considering a decreasing stepsize $\gamma_k \in (0,1)$ and a constant $\rho \in (0,1)$, we have

$$\mathbb{E}[\|\Delta S\|^2] \leq \frac{\rho}{1-\rho} \sum_{\ell=0}^{k} (1-\gamma_\ell)^2 (\boldsymbol{\mathcal{S}}^{(\ell)} - S_{\text{tts}}^{(\ell)}) \, ,$$

where $\boldsymbol{\mathcal{S}}^{(\ell)}$ is defined either by Line 2 (vrTTEM) or Line 3 (fiTTEM).

IV. Numerical Examples

This section presents several numerical applications for our proposed class of Algorithms 1.

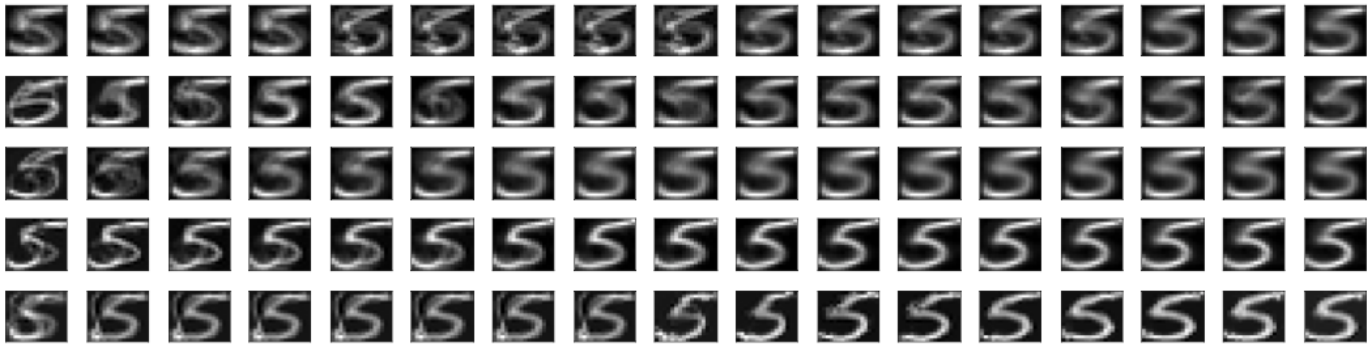

Fig. 1. (USPS Digits) Estimation of the template. From top to bottom: batch, online, iSAEM, vrTTEM and fiTTEM through 7 epochs. Note that Batch method templates are replicated in-between epochs for a fair comparison with incremental variants.

A. Gaussian Mixture Models

We begin by a simple and illustrative example. The authors acknowledge that the following model can be trained using deterministic EM-type of algorithms but propose to apply stochastic methods, including theirs, in order to compare their performances. Given n observations $\{y_i\}_{i=1}^n$, we want to fit a Gaussian Mixture Model (GMM) whose distribution is modeled as a mixture of M Gaussian components, each with a unit variance. 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 is given by $\Theta = \{\omega_m,\ m = 1, ..., M-1 : \omega_m \geq 0,\ \sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R},\ m = 1, ..., M\}$. We generate 50 synthetic datasets of size $n = 10^5$ from a GMM model with $M = 2$ components of means $\mu_1 = -\mu_2 = 0.5$. We run the EM method until convergence (to double precision) to obtain the ML estimate $\mu^\star$ averaged on 50 datasets. We compare the EM, iEM (incremental EM), SAEM, iSAEM, vrTTEM and fiTTEM methods in terms of their precision measured by $|\mu - \mu^\star|^2$. For all methods, $\gamma_k = 1/k^\alpha$ with $\alpha = 0.5$, and the stepsize $\rho_k \propto 1/n^{2/3}$ for the vrTTEM and the fiTTEM . The number of MC samples is fixed to $M = 10$. Figure 2 shows the precision $|\mu - \mu^*|^2$ for the different methods through the epoch(s) (one epoch equals n iterations). The vrTTEM and fiTTEM methods outperform the other stochastic methods, supporting the benefits of our scheme.

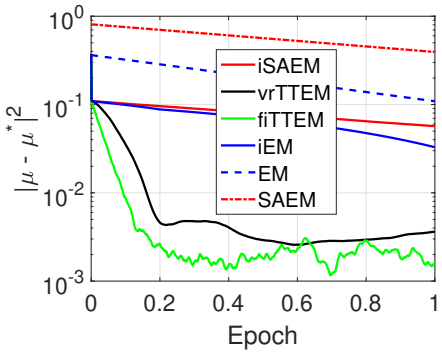

Fig. 2. Precision $|\mu^{(k)} - \mu^*|^2$ per epoch

B. Deformable Template Model for Image Analysis

Let $(y_i, i \in [n])$ be observed gray level images defined on a grid of pixels. 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:

$$y_i(u) = I\left(x_u - \Phi_i\left(x_u, z_i\right)\right) + \varepsilon_i(u) \tag{15}$$

where $\Phi_i : \mathbb{R}^2 \to \mathbb{R}^2$ is a deformation function, z_i some latent variable parameterizing this deformation and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an observation error. The template model, given $\{p_k\}_{k=1}^{k_p}$ landmarks on the template, a fixed known kernel $\mathbf{K_p}$ and a vector of parameters $\beta \in \mathbb{R}^{k_p}$ is defined as follows:

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

Given a set of landmarks $\{g_k\}_{k=1}^{k_g}$ and a fixed kernel $\mathbf{K_g}$, we parameterize the deformation Φ_i as $\Phi_i = \mathbf{K_g} z_i$ where $\left(\mathbf{K_g} z_i\right)(x) = \sum_{k=1}^{k_s} \mathbf{K_g}\left(x, g_k\right)\left(z_i^{(1)}(k), z_i^{(2)}(k)\right)$. We also 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 estimate is thus $\boldsymbol{\theta} = \left(\beta, \Gamma, \sigma\right)$. The complete model (15) belongs to the curved exponential family, see [29], which vector of sufficient statistics for all $i \in [n]$ is defined by $S(y_i, z_i) = (\mathbf{K}_{p,z_i}^\top y_i, \mathbf{K}_{p,z_i}^\top \mathbf{K}_{p,z_i}, z_i^t z_i)$ where we denote $\mathbf{K}_{p,z_i} = \mathbf{K}_{p,z_i}(x_u - \phi_i(x_u, z_i), p_j)$. Then, the two-timescale M-step (6) yields the following parameter updates $\boldsymbol{\beta}(\hat{s}) = \hat{s}_2^{-1}(z)\hat{s}_1(z)$,$\boldsymbol{\Gamma}(\hat{s}) = \hat{s}_3(z)/n$, $\boldsymbol{\sigma}(\hat{s}) = \boldsymbol{\beta}(\hat{s})^\top \hat{s}_2(z)\boldsymbol{\beta}(\hat{s}) - 2\boldsymbol{\beta}(\hat{s})\hat{s}_1(z)$ where $\hat{s} = (\hat{s}_1(z), \hat{s}_2(z), \hat{s}_3(z))$ is the vector of statistics obtained via update (11) in Algorithm 1.

Numerical Experiment: We apply model (15) and our Algorithm 1 to a collection of handwritten digits, called the US postal database [30], featuring $n = 1\,000$, (16×16)-pixel images for each class of digits from 0 to 9. The main challenge with this dataset stems from the geometric dispersion within each class of digit as shown Figure **??** for digit 5. We thus ought to use our deformable template model (15) in order to account for both sources of variability: the intrinsic template

to each class of digit and the small and local deformations in each observed image.

Figure 1 shows the resulting synthetic images for digit 5 through several epochs, for the batch method, the online SAEM, the incremental SAEM and the various two-timescale methods. For all methods, the initialization of the template (16) is the mean of the gray level images. In our experiments, we have chosen Gaussian kernels for both, $\mathbf{K_p}$ and $\mathbf{K_g}$, defined on $\mathbb{R}^2$ and centered on the landmark points$\{p_k\}_{k=1}^{k_p}$ and $\{g_k\}_{k=1}^{k_g}$ with standard respective standard deviations of 0.12 and 0.3. We set $k_p = 15$ and $k_g = 6$ equidistributed landmarks points on the grid for the training procedure. Those hyperparameters are inspired by relevant studies [31], [32]. In particular, the choice of the geometric covariance, indexed by g, in such study is critical since it has a direct impact on the *sharpness* of the templates. As for the photometric hyperparameter, indexed by p, both the template and the geometry are impacted, in the sense that with a large photometric variance, the kernel centered on one landmark *spreads out* to many of its neighbors.

As the iterations proceed, the templates become sharper. Figure 1 displays the virtue of the vrTTEM and fiTTEM methods that obtain a more *contrasted* and *accurate* template estimate. The incremental and online versions are better in the very first epochs compared to the batch method, given the high computational cost of the latter. After a few epochs, the batch SAEM estimates similar template as the incremental and online methods due to their high variance. Our variance reduced and fast incremental variants are effective in the long run and sharpen the template estimates contrasting between the background and the regions of interest in the image.

V. Conclusion

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

References

[1] 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*. Springer, 1998, pp. 355–368.

[2] 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*, vol. 85, no. 411, pp. 699–704, 1990.

[3] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the em algorithm," *Ann. Statist.*, vol. 27, no. 1, pp. 94–128, 03 1999. [Online]. Available: https://doi.org/10.1214/aos/1018031103

[4] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[5] B. Efron *et al.*, "Defining the curvature of a statistical problem (with applications to second order efficiency)," *The Annals of Statistics*, vol. 3, no. 6, pp. 1189–1242, 1975.

[6] 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)*, pp. 1–38, 1977.

[7] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[8] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[9] H. D. Nguyen, F. Forbes, and G. J. McLachlan, "Mini-batch learning of exponential family finite mixture models," *Statistics and Computing*, pp. 1–18, 2020.

[10] 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*, 2009, pp. 611–619.

[11] O. Cappé, "Online EM algorithm for hidden markov models," *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 728–749, 2011.

[12] 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*, 2019, pp. 2833–2843.

[13] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[14] C. E. McCulloch, "Maximum likelihood algorithms for generalized linear mixed models," *Journal of the American statistical Association*, vol. 92, no. 437, pp. 162–170, 1997.

[15] J. P. Hughes, "Mixed effects models with censored data with application to hiv rna levels," *Biometrics*, vol. 55, no. 2, pp. 625–629, 1999.

[16] 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*, vol. 45, no. 6, pp. 1643–1669, 2016.

[17] A. Chakraborty and K. Das, "Inferences for joint modelling of repeated ordinal scores and time to event data," *Computational and mathematical methods in medicine*, vol. 11, no. 3, pp. 281–295, 2010.

[18] S. Ng and G. McLachlan, "On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures," *Statistics and Computing*, vol. 13, no. 1, pp. 45–55, FEB 2003.

[19] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, JUN 2017.

[20] E. Kuhn, C. Matias, and T. Rebafka, "Properties of the stochastic approximation em algorithm with mini-batch sampling," *arXiv preprint arXiv:1907.09164*, 2019.

[21] P. Jain and P. Kar, "Non-convex optimization for machine learning," *arXiv preprint arXiv:1712.07897*, 2017.

[22] S. Ghadimi and G. Lan, "Stochastic first-and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[23] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[24] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of markov chain monte carlo*. CRC press, 2011.

[25] E. Kuhn and M. Lavielle, "Coupling a stochastic approximation version of em with an mcmc procedure," *ESAIM: Probability and Statistics*, vol. 8, pp. 115–131, 2004.

[26] J. Chen, J. Zhu, Y. W. Teh, and T. Zhang, "Stochastic expectation maximization with variance reduction," in *Advances in Neural Information Processing Systems*, 2018, pp. 7978–7988.

[27] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315–323.

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

[29] 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)*, vol. 69, no. 1, pp. 3–29, 2007.
[30] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.
[31] S. Allassonnière and E. Kuhn, "Stochastic algorithm for parameter estimation for dense deformable template mixture model," *arXiv preprint arXiv:0802.1521*, 2008.
[32] S. Allassonnière, E. Kuhn, A. Trouvé *et al.*, "Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study," *Bernoulli*, vol. 16, no. 3, pp. 641–678, 2010.