Fast Two-Time-Scale Noisy EM Algorithms

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

T.B.C

1 Introduction

We formulate the following empirical risk minimization as:

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

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

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

2 Expectation Maximization Algorithm

Full batch EM 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 $\textsf{M-step}$ is given by

$$\textsf{M-step: } \hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\overline{\mathbf{s}}(\boldsymbol{\theta})) := \arg\min_{\vartheta\in\Theta} \big\{ \mathrm{R}(\vartheta) + \psi(\vartheta) - \big\langle \overline{\mathbf{s}}(\boldsymbol{\theta}) \,|\, \phi(\vartheta)\big\rangle \big\}, \tag{4}$$

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

3 Monte Carlo Integration and Stochastic Approximation

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 $\overline{\mathrm{s}}(\boldsymbol{\theta})$:

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

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:

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

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.

4 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.

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

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.

Algorithm 1 Two-Time-Scale Noisy EM methods.

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

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

3: **for** $k = 0, 1, 2, \ldots, K$ **do**
4: Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiSAEM).
5: Compute $\hat{S}_i^{(k)}$ using the MC-step (5), for the drawn indices.
6: Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using (9) or (10) or (11).
7: Compute $\hat{S}^{(k+1)}$ via the Inexact-step (8).
8: Compute $\hat{\mathrm{s}}^{(k+1)}$ via the SA-step (7).
9: Compute $\hat{\boldsymbol{\theta}}^{(k+1)}$ via the M-step (4).
10: **end for**
11: **Return**: $\hat{\boldsymbol{\theta}}^{(K)}$.

5 Finite Time Analysis

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

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

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 Z, S are measurable open sets such that:*

$$
\mathsf{S} \supset \left\{ n^{-1} \sum_{i=1}^{n} u_i, u_i \in \operatorname{conv}(\overline{\mathrm{s}}_i(\boldsymbol{\theta})) \right\} \tag{15}
$$

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

H2. *The conditional distribution is smooth on $\operatorname{int}(\Theta)$. For any $i \in [\![1, n]\!]$, $z \in \mathsf{Z}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \operatorname{int}(\Theta)^2$, 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 $s \in \mathsf{S}$, the function $\boldsymbol{\theta} \mapsto L(s, \boldsymbol{\theta}) := \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \mathrm{s} \,|\, \phi(\boldsymbol{\theta}) \right\rangle$ admits a unique global minimum $\overline{\boldsymbol{\theta}}(\mathrm{s}) \in \operatorname{int}(\Theta)$. 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 $\boldsymbol{\theta} \mapsto L(\mathrm{s}, \boldsymbol{\theta}) = \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \mathrm{s} \,|\, \phi(\boldsymbol{\theta}) \right\rangle$, and define

$$
\mathrm{B}(\mathrm{s}) := \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s})) \Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathrm{s}, \overline{\boldsymbol{\theta}}(\mathrm{s})) \Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))^{\top}. \tag{16}
$$

H4. *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* $\| \mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\| \leq \mathrm{L}_B \, \|\mathbf{s} - \mathbf{s}'\|$.

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 $\overline{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k-1)}))$. For all $i \in [\![1, n]\!]$, $r > 0$ and $\vartheta \in \Theta$, define:

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

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.*, $\mathbb{E}[\eta_{i,\vartheta}^{(r)}|\mathcal{F}_r] = 0$ 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* $\{M_k\}_{k>0}$ *and constants* (C, C_η) *such that for all* $k > 0$, $i \in [\![1, n]\!]$ *and* $\vartheta \in \Theta$:

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

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

$$\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \, \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \mathrm{L}_V \, \|\mathbf{s} - \mathbf{s}'\|, \tag{19}$$

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

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

Lemma 2. *Assume H1. The update* (9) *is equivalent to the following update on the resulting statistics*

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

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

$$\mathbb{E}\left[\left\|\tilde{S}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\right\|^2\right] \leq \mathbb{E}\left[\left\|\overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)}\right\|^2\right] + 2\,\mathrm{L}_{\mathbf{s}}^2 \left(1 - \frac{1}{n}\right)^2 n^{-1} \sum_{i=1}^{n} \mathbb{E}\left[\left\|\hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)}\right\|^2\right] + \frac{2C}{M_k} \tag{21}$$

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

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.

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

Proof Under some regularity conditions of the Lyapunov function V, cf. Lemma 19, and the following growth condition for all $\mathbf{s} \in \mathsf{S}$,

$$\upsilon_{\min}^{-1} \langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \rangle \geq \left\| \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \right\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \tag{22}$$

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

$$V(\hat{\boldsymbol{s}}^{(k+1)}) \leq V(\hat{\boldsymbol{s}}^{(k)}) - \gamma_{k+1} \langle \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle + \frac{\gamma_{k+1}^2 \,\mathrm{L}_V}{2} \| \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \|^2 \tag{23}$$

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

$$\begin{aligned} \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \leq \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1} \upsilon_{\min} \mathbb{E}\left[\left\| \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] + \mathbb{E}\left[\frac{\gamma_{k+1}^2 \,\mathrm{L}_V}{2} \| \hat{\boldsymbol{s}}^{(k)} - \tilde{S}^{(k+1)} \|^2 \right] \\ - \gamma_{k+1} \mathbb{E}\left[\langle \overline{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle \right] \end{aligned} \tag{24}$$

We then establish an auxiliary Lemma yielding an upper-bound on the quantity $\mathbb{E}\left[\langle \overline{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle \right]$ where:

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

Lemma 3.

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

Using Lemma 2:

$$\begin{aligned} \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k+1)})] \leq \mathbb{E}[V(\hat{\boldsymbol{s}}^{(k)})] - \gamma_{k+1} \left(\upsilon_{\min} - \frac{\gamma_{k+1} \,\mathrm{L}_V}{2} \right) \mathbb{E}\left[\left\| \overline{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(k)} \right\|^2 \right] \\ + \gamma_{k+1}^2 \,\mathrm{L}_V \,\mathrm{L}_{\mathbf{s}}^2 \left(1 - \frac{1}{n}\right)^2 n^{-1} \sum_{i=1}^n \mathbb{E}\left[\left\| \hat{\mathbf{s}}^{(k)} - \hat{\mathbf{s}}^{(\tau_i^k)} \right\|^2 \right] + \frac{\gamma_{k+1}^2 \,\mathrm{L}_V \, C}{M_k} \\ - \gamma_{k+1} \mathbb{E}\left[\langle \overline{\mathbf{s}}^{(k)} - \tilde{S}^{(k+1)} \,|\, \nabla V(\hat{\boldsymbol{s}}^{(k)}) \rangle \right] \end{aligned} \tag{27}$$

Besides,

$$n^{-1} \sum_{i=1}^n \mathbb{E}\left[\left\| \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\tau_i^{k+1})} \right\|^2 \right] = n^{-1} \sum_{i=1}^n \left(\frac{1}{n} \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2] + \frac{n-1}{n} \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(\tau_i^k)}\|^2] \right) \tag{28}$$

yielding for any numbers $\beta_k > 0$,

$$\begin{aligned} & \mathbb{E}[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\ & = \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2 \langle \hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)} \,|\, \hat{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)} \rangle \right] \\ & = \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 - 2\gamma_{k+1} \langle \hat{\mathbf{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \hat{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)} \rangle \right] \\ & \leq \mathbb{E}\left[\|\hat{\mathbf{s}}^{(k+1)} - \hat{\mathbf{s}}^{(k)}\|^2 + \|\hat{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma_{k+1}}{\beta_{k+1}} \|\hat{\mathbf{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 + \gamma_{k+1} \beta_{k+1} \|\hat{\mathbf{s}}^{(k)} - \hat{\boldsymbol{s}}^{(t_i^k)}\|^2 \right] \end{aligned} \tag{29}$$

$\square$

5.2 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.

6 Numerical Examples

6.1 Gaussian Mixture Models

Graphs obtained and relevant

6.2 Deep Latent Variable Models using noisy EM

See if makes sense to use EM instead of Variational Inference

6.3 Deformable Template Model for Image Analysis

See Kuhn et.al. paper.

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.