2 Motivation

2.1 The problem

sec:motivation

The problem to be solved is

$$\text{Argmin}_{\theta\in\Theta} F(\theta), \qquad \text{where} \quad F(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + \mathsf{R}(\theta), \tag{2}$$

eq:problem

and

$$\mathcal{L}_i(\theta) \stackrel{\text{def}}{=} -\log \int_{\mathsf{Z}} h_i(z) \exp\left(\langle s_i(z), \phi(\theta)\rangle - \psi_i(\theta)\right)\ \mu(\mathrm{d}z). \tag{3}$$

eq:def:loss

hyp:Rand

H1. *1. $\Theta \subseteq \mathbb{R}^d$ is an open set.*

hyp:curvedexpo

2. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\mathsf{R} : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$, $\psi_i : \Theta \to \mathbb{R}$, $s_i : \mathsf{Z} \to \mathbb{R}^q$, $h_i : \mathsf{Z} \to \mathbb{R}_+$ for all $i \in \{1, \ldots, n\}$ are measurable functions. Finally, for any $\theta \in \Theta$, $-\infty < \mathcal{L}_i(\theta) < \infty$.

Under H1-item 2, for any $\theta \in \Theta$ and $i \in \{1, \ldots, n\}$, the quantity $p_i(z;\theta)\,\mu(\mathrm{d}z)$ where

$$p_i(z;\theta) \stackrel{\text{def}}{=} h_i(z) \exp\left(\langle s_i(z), \phi(\theta)\rangle - \psi_i(\theta) + \mathcal{L}_i(\theta)\right).$$

defines a probability distribution on $\mathcal{Z}$.

hyp:bars

H2. *For all $\theta \in \Theta$ and $i \in \{1, \ldots, n\}$, the expectation*

$$\bar{s}_i(\theta) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} s_i(z)\ p_i(z;\theta)\mu(\mathrm{d}z)$$

exists.

Let us define $\mathsf{Q}_i : \Theta \times \Theta \to \mathbb{R}$ by

$$\mathsf{Q}_i(\theta, \theta') \stackrel{\text{def}}{=} \psi_i(\theta) - \langle \bar{s}_i(\theta'), \phi(\theta)\rangle\,.$$

Then, the Jensen's inequality implies that for any $\theta, \theta' \in \Theta$,

$$\mathcal{L}_i(\theta) \le \mathsf{Q}_i(\theta, \theta') + \mathcal{L}_i(\theta') - \psi_i(\theta') + \langle \bar{s}_i(\theta'), \phi(\theta')\rangle\,.$$

Therefore, for any $\theta, \theta' \in \Theta$,

$$F(\theta) \le \overline{\psi}(\theta) - \langle \bar{s}(\theta'), \phi(\theta)\rangle + \mathsf{R}(\theta) + \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta') - \overline{\psi}(\theta') + \langle \bar{s}(\theta'), \phi(\theta')\rangle \right\} \tag{4}$$

eq:MMequation

where

$$\bar{s} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i, \qquad \overline{\psi} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \psi_i. \tag{5}$$

eq:def:bars

The RHS of (4) defines a family of majorizing functions of F on the whole set Θ; this family is indexed by $\theta' \in \Theta$ and for any θ', the majorizing function is equal to F at the point $\theta = \theta'$.

hyp:Tmap hypxTmap

H3. *1. Let $\mathcal{S} \subseteq \mathbb{R}^q$ be a measurable open set such that*

$$\mathcal{S} \supset \left\{ \frac{1}{n} \sum_{i=1}^{n} u_i,\ u_i \in Conv(\bar{s}_i(\Theta)) \right\}.$$

For any $s \in \mathcal{S}$

$$\mathrm{Argmin}_{\theta \in \Theta} \ \left(\overline{\psi}(\theta) - \langle s, \phi(\theta) \rangle + \mathsf{R}(\theta) \right),$$

is a (non empty) singleton denoted by $\{\mathsf{T}(s)\}$.

sec:EM

2.2 An EM algorithm

From (4), a natural idea for solving (2) is the use of a MM algorithm defined as follows: define the sequence $\{\tau^k,\ k \in \mathbb{N}\}$ by $\tau^0 \in \Theta$, and for any $k \geq 0$,

$$\tau^{k+1} \stackrel{\text{def}}{=} \mathsf{T}(\bar{s}(\tau^k)). \tag{6}$$

eq:exact:update:tau

Starting from the current point τ^k, the algorithm first compute a point in $\bar{s}(\Theta)$ through the expectation $\bar{s}$, and then apply the map T to obtain the new point τ^{k+1}. It can therefore be equivalently defined in the $\bar{s}(\Theta)$-space, as follows: define $\{\bar{s}^k,\ k \in \mathbb{N}\}$ by $\bar{s}^0 \in \mathcal{S}$ and for any $k \geq 0$,

$$\bar{s}^{k+1} \stackrel{\text{def}}{=} \bar{s}\left(\mathsf{T}(\bar{s}^k)\right); \tag{7}$$

eq:exact:update:bars

upon noting that we have $\tau^{k+1} = \mathsf{T}(\bar{s}^{k+1})$ (for the initialization, choose $\tau^0 \stackrel{\text{def}}{=} \mathsf{T}(\bar{s}^0)$). The following lemma shows that there exists a natural Lyapunov function for these algorithms; it also establishes that this MM algorithm is an EM algorithm.

Data: $K_{\max} \in \mathbb{N}$, $\bar{s}^0 \in \mathcal{S}$
Result: The EM sequence: $\bar{s}^k, k = 0, \ldots, K_{\max}$
for $k = 0, \ldots, K_{\max} - 1$ **do**
  $\bar{s}^{k+1} = \bar{s} \circ \mathsf{T}(\bar{s}^k)$

Algorithm 1: EM algorithm

lem:lyapunovEM

Lemma 1. *Assume H1-item 1,item 2, H2 and H3-item 1. Let $\{\bar{s}^k, k \geq 0\}$ be given by (7) and set $\tau^k \stackrel{\text{def}}{=} \mathsf{T}(\bar{s}^k)$. Then for any $k \geq 0$,*

$$F(\tau^{k+1}) \leq F(\tau^k), \qquad F \circ \mathsf{T}(\bar{s}^{k+1}) \leq F \circ \mathsf{T}(\bar{s}^k),$$

and the sequence $\{\tau^k, k \geq 0\}$ is an EM sequence.

Proof. By definition of the map T,

$$\overline{\psi}(\tau^{k+1}) - \left\langle \bar{s}^{k+1}, \phi(\tau^{k+1}) \right\rangle + \mathsf{R}(\tau^{k+1}) \leq \overline{\psi}(\tau^k) - \left\langle \bar{s}^{k+1}, \phi(\tau^k) \right\rangle + \mathsf{R}(\tau^k).$$

In addition, by (4), we have

$$F(\tau^{k+1}) \leq \overline{\psi}(\tau^{k+1}) - \left\langle \bar{s}^{k+1}, \phi(\tau^{k+1}) \right\rangle + \mathsf{R}(\tau^{k+1}) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\tau^k) - \overline{\psi}(\tau^k) + \left\langle \bar{s}^{k+1}, \phi(\tau^k) \right\rangle .$$

Combining these inequalities yields

$$F(\tau^{k+1}) \leq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\tau^k) + \mathsf{R}(\tau^k) = F(\tau^k).$$

This concludes the proof of the first statement. We now show this MM algorithm is equivalent to the EM algorithm. Given the current value τ^k, the E-step would compute at iteration k the quantity:

$$\begin{aligned}\widetilde{Q}_i(\tau, \tau^k) &\stackrel{\text{def}}{=} \int_{\mathsf{Z}} \log(h_i(z) \exp\left(\langle s_i(z), \phi(\tau)\rangle - \psi_i(\tau)\right) p_i(z;\tau^k)\mu(\mathrm{d}z) \\ &= \int_{\mathsf{Z}} \log(h_i(z))\ p_i(z;\tau^k)\mu(\mathrm{d}z) + \mathsf{Q}_i(\tau, \tau^k)\end{aligned}$$

and the M step defined the next value $\tau_{\text{EM}}^{k+1} \stackrel{\text{def}}{=} \text{Argmax}_{\tau\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \widetilde{Q}_i(\tau, \tau^k) - \mathsf{R}(\tau)$, that is

$$\begin{aligned}\tau_{\text{EM}}^{k+1} &= \text{Argmax}_{\tau\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \mathsf{Q}_i(\tau, \tau^k) - \mathsf{R}(\tau) \\ &= \text{Argmax}_{\tau\in\Theta} \left\langle \bar{s}(\tau^k), \phi(\tau)\right\rangle - \overline{\psi}(\tau) - \mathsf{R}(\tau) = \mathsf{T}(\bar{s}^{k+1});\end{aligned}$$

thus showing that $\tau^{k+1} = \tau_{\text{EM}}^{k+1}$. □

The updating rule (7) shows that if the algorithm converges to $s^\star$, then $s^\star$ is a root of

$$s \mapsto h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s. \tag{8}$$

eq:meanfield

The computational cost of the algorithm is proportional to n per iteration (since it requires the computation of $\bar{s}$, a sum over n terms); it is therefore untractable in the large scale learning framework. To overcome this drawback, and upon noting that

$$h(s) = \mathbb{E}\left[\bar{s}_I \circ \mathsf{T}(s) - s\right],$$

where I is a uniform random variable on $\{1, \ldots, n\}$, a natural idea is to replace $\bar{s} \circ \mathsf{T}(s^k)$ by a stochastic approximation involving the computation of one (or let us say, a fixed small number) expectation $\bar{s}_i \circ \mathsf{T}(s^k)$ at each iteration. Among possible strategies, we introduce in Section 2.3 and Section 2.4 two different algorithms where the deterministic algorithm producing $\bar{s}^{k+1} = \bar{s} \circ \mathsf{T}(\bar{s}^k)$ is replaced with a stochastic algorithm producing a sequence $\{\widehat{S}^k, k \geq 0\}$ satisfying

$$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} H(\widehat{S}^k, U_{k+1}) \tag{9}$$

eq:SAscheme

for two different choices of the field H and of the random sequence $\{U^k,\ k \in \mathbb{N}\}$. $\{\gamma^k,\ k \in \mathbb{N}\}$ is a deterministic positive stepsize sequence chosen by the user, and the random variable U_{k+1} is usually chosen such that

$$\mathbb{E}\left[H(\widehat{S}^k, U_{k+1})|\widehat{S}^0, U_1, \ldots, U_k\right] = h(\widehat{S}^k),$$

(see Section 2.4) but not necessarily (see Section 2.3).

sec:i-EM

2.3 The incremental EM algorithm

Incremental EM (iEM) defines a sequence $\{\widehat{S}^k,\ k \in \mathbb{N}\}$ as described in algorithm 2.

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathcal{S}$, $\gamma_k \in (0, \infty)$ for $k = 1, \ldots, K_{\max}$
Result: The iEM sequence: $\widehat{S}^k, k = 0, \ldots, K_{\max}$
$\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}^0)$ for all $i = 1, \ldots, n$;
$\widetilde{S}^0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;
for $k = 0, \ldots, K_{\max} - 1$ **do**
  $I_{k+1} \sim \mathcal{U}(\{1, \ldots, n\})$;
  $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \neq I_{k+1}$;
  $\mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;
  $\widetilde{S}^{k+1} = \widetilde{S}^k + n^{-1}\left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}}\right)$;
  $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}(\widetilde{S}^{k+1} - \widehat{S}^k)$

Algorithm 2: The incremental EM (iEM) algorithm algo:iEM

Upon noting that for any $k \geq 0$,

$$\frac{1}{n}\left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}}\right) = \frac{1}{n}\sum_{i=1}^n \mathsf{S}_{k+1,i} - \frac{1}{n}\sum_{i=1}^n \mathsf{S}_{k,i},$$

a trivial induction shows that $\widetilde{S}^k = \frac{1}{n}\sum_{i=1}^n \mathsf{S}_{k,i}$ for any $k \geq 0$. Note however that the above algorithmic description allows the computation of this sum of n terms (at each iteration k) through a call to a single $\bar{s}_i$ at each iteration. $\widetilde{S}^{k+1}$ is an approximation of $\bar{s} \circ \mathsf{T}(\widehat{S}^k)$; more precisely, we have for any $k \geq 0$,

$$\widetilde{S}^k = \frac{1}{n}\sum_{i=1}^n \bar{s}_i \circ \mathsf{T}(\widehat{S}^{<k,i})$$

where $\widehat{S}^{<0,i} \stackrel{\text{def}}{=} \widehat{S}^0$ for all $i \in \{1, \ldots, n\}$ and for $k \geq 0$,

$$\widehat{S}^{<k+1,i} = \widehat{S}^{\ell}, \begin{cases} \ell = k & \text{if } I_{k+1} = i \\ 1 \leq \ell \leq k-1 & \text{if } I_{k+1} \neq i, I_k \neq i, \ldots, I_{\ell+1} = i \\ \ell = 0 & \text{otherwise} \end{cases} \quad (10)$$

eq:memory:lastupdate

The above algorithm slightly extends the original incremental EM (see Neal and Hinton (1998)) by introducing a stepsize sequence. In Neal and Hinton (1998), we have $\gamma_{k+1} = 1$ for any $k \geq 0$ so that $\widehat{S}^k = \widetilde{S}^k = n^{-1} \sum_{i=1}^n \mathsf{S}_{k,i}$ for any $k \geq 0$.

This algorithm is defined as soon as $\mathsf{T}(\widehat{S}^k)$ exists that is $\widehat{S}^k \in \mathcal{S}$ at each iteration. Based on H3item 1, a sufficient condition is $\mathcal{S} = \mathbb{R}^q$ or $\gamma_k \in (0, 1]$ for any k (note that in the original incremental EM by Neal and Hinton (1998), $\widehat{S}^k \in \mathcal{S}$ for any k under the assumption H3item 1).

In the literature, several variants of the EM algorithm share a similar procedure, meaning that at each step, a stochastic approximation of an expectation is defined through

$$\widehat{S}^k = (1 - \gamma_k)\widehat{S}^{k-1} + \gamma_k H_{k+1}$$

where H_{k+1} is in the convex hull of $\cup_i s_i(\mathsf{Z})$; and then the updated parameter is obtained by $\mathsf{T}(\widehat{S}^k)$. There exist different ways to ensure $\widehat{S}^k \in \mathcal{S}$. In (Delyon et al., 1999, Section 4), $\mathcal{S}$ is assumed to contain the convex hull of $\cup_i s_i(\mathsf{Z})$ and the step sizes γ_k are in $(0, 1)$. In (Kuhn and Lavielle, 2004, Theorem 1), (Allassonnière et al., 2010, Theorem 1) and (Donnet and Samson, 2007, Theorem 6), it is assumed $\gamma_k \in (0, 1)$ and that the sequence $\widehat{S}^k$ remains in a compact subset of $\mathcal{S}$; the first two papers verify these assumptions in their applications. Furthermore, in (Cappé and Moulines, 2009, Assumption 1), $\gamma_k \in (0, 1)$, and $\mathcal{S}$ is assumed to be convex and to contain the whole sequence $\widehat{S}^k$; they show that these conditions hold in their application as soon as the algorithm is suitably initialized. Finally, in the algorithm proposed in (Le Corff and Fort, 2013, Section 4.1) (which corresponds to the case $\gamma_k = 1$), it is again assumed that $\mathcal{S}$ contains the convex hull of $\cup_i s_i(\mathsf{Z})$.

lem:iEM:equivalent

Lemma 2. *Assume H1item 1-item 2, H2 and H3item 1. Let $\{\gamma^k,\ k \in \mathbb{N}\}$ be a positive stepsize sequence such that for any k, $\widehat{S}^k \in \mathcal{S}$. The incremental EM algorithm is equivalent to the following algorithm : initialize*

$$\widehat{S}^0 \in \mathcal{S}, \qquad \mathsf{S}_0 \stackrel{\text{def}}{=} (\bar{s}_1 \circ \mathsf{T}(\widehat{S}^0), \ldots, \bar{s}_n \circ \mathsf{T}(\widehat{S}^0)),$$

and then repeat for $k \geq 0$:

$$\begin{aligned}
&\textit{Draw: } I_{k+1} \sim \mathcal{U}(\{1, \ldots, n\}), \\
&\textit{Update: } \mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}, i \neq I_{k+1}, \qquad \mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k), \\
&\textit{Update: } \widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} \left(\frac{1}{n} \sum_{i=1}^n \mathsf{S}_{k+1,i} - \widehat{S}^k \right).
\end{aligned}$$

In addition, if $\mathcal{F}_0 \stackrel{\text{def}}{=} \sigma(\widehat{S}^0)$, and for any $k \geq 1$, $\mathcal{F}_k \stackrel{\text{def}}{=} \sigma\left(\widehat{S}^0, I_1, \ldots, I_k\right)$, then

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \mathsf{S}_{k+1,i} - \widehat{S}^k \middle| \mathcal{F}_k \right] = h(\widehat{S}^k) + \left(1 - \frac{1}{n}\right) \left(\frac{1}{n} \sum_{i=1}^n \mathsf{S}_{k,i} - \bar{s} \circ \mathsf{T}(\widehat{S}^k) \right).$$

Note that the sequence $\{\widehat{S}^k,\ k \in \mathbb{N}\}$ is not a Stochastic Approximation algorithm, but the bivariate sequence $\{(\widehat{S}^k, \mathsf{S}_{k,\cdot}), k \in \mathbb{N}\}$ is. While being an equivalent algorithmic description, the implementation is not equivalent: in Lemma 2, a sum over n terms is required for each update of $\widehat{S}^{k+1}$ (i.e. the computational cost is equivalent to the cost of the EM algorithm) while the first description of the algorithm is based on a recursive computation of this sum through the quantity $\widetilde{S}^{k+1}$.

sec:Fi-EM

2.4 The Fast Increment EM algorithm

Fast Incremental EM (FIEM) defines a sequence $\{\widehat{S}^k,\ k \in \mathbb{N}\}$ based on the scheme (9) where

$$H(\widehat{S}, U) \stackrel{\text{def}}{=} \left(\bar{s}_J \circ \mathsf{T}(\widehat{S}) - \widehat{S}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{S}_i - \mathsf{S}_J\right), \quad U \stackrel{\text{def}}{=} (J, \mathsf{S}) \in \{1, \ldots, n\} \times \mathcal{S}^n.$$

This field can be seen as the sum of two terms: the natural field associated to the mean field h (8) when conditionally to $(\widehat{S}, \mathsf{S})$, J is sampled from the uniform distribution on the integers $\{1, \ldots, n\}$; and a random variable whose conditional expectation is zero. The fundamental property is that, conditionnally to $(\widehat{S}, \mathsf{S})$ these two terms are correlated through the use of the same random index J (see the variance reduction technique based on control variates, e.g. in (Glasserman, 2004, Section 4.1.).) We introduce a slight modification of the original algorithm, by using a sequence of coefficients $\{\lambda_k,\ k \in \mathbb{N}\}$ of real numbers. In the original algorithm (see Karimi et al. (2019b)), $\lambda_k = 1$.

FIEM is defined by algorithm 3. As in Section 2.3, it is easily seen that the

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathcal{S}$, $\gamma_k \in (0, \infty)$ for $k = 1, \ldots, K_{\max}$
Result: The iEM sequence: $\widehat{S}^k, k = 0, \ldots, K_{\max}$
$\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}^0)$ for all $i = 1, \ldots, n$;
$\widetilde{S}^0 = n^{-1} \sum_{i=1}^{n} \mathsf{S}_{0,i}$;
for $k = 0, \ldots, K_{\max} - 1$ **do**
  $I_{k+1} \sim \mathcal{U}(\{1, \ldots, n\})$;
  $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \neq I_{k+1}$;
  $\mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;
  $\widetilde{S}^{k+1} = \widetilde{S}^k + n^{-1}\left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}}\right)$;
  $J_{k+1} \sim \mathcal{U}(\{1, \ldots, n\})$;
  $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}(\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \lambda_{k+1}(\widetilde{S}^{k+1} - \mathsf{S}_{k+1,J_{k+1}}))$

Algorithm 3: The Fast Incremental EM (FIEM) algorithm algo:FIEM

second and third instructions of the algorithm are a recursive computation of a

sum of n terms: it holds for any $k \geq 0$ (see (10) for the definition of $\widehat{S}^{<k,i}$)

$$\widetilde{S}^k = \frac{1}{n}\sum_{i=1}^{n} \mathsf{S}_{k,i} = \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i \circ \mathsf{T}(\widehat{S}^{<k,i}).$$

Here again, this algorithm is well defined as soon as $\widehat{S}^k \in \mathcal{S}$ for any $k \geq 0$, which ensures that $\mathsf{T}(\widehat{S}^k)$ exists. This is trivial when $\mathcal{S} = \mathbb{R}^q$; it holds true when $\lambda_k = 0$ and $\gamma_k \in (0, 1]$ under H3item 1. In the literature, there exist results which, similarly to FIEM, combine an update scheme of the form $\widehat{S}^{k+1} = (1-\gamma_{k+1})\widehat{S}^k + \gamma_{k+1} H_{k+1}$ when H_{k+1} is **not** in the convex hull of $\bigcup_{i=1}^n s_i(\mathsf{Z})$, and a condition that $\widehat{S}^k$ remains in a definition set $\mathcal{S} \subseteq \mathbb{R}^q$ of a transformation: nevertheless, they assume $\mathcal{S} = \mathbb{R}^q$ (see e.g. Johnson and Zhang (2013), Defazio et al. (2014) and Chen et al. (2018)). Therefore, to our best knowledge, the case when $\mathcal{S} \neq \mathbb{R}^q$ is an open question.

Algorithm 3 is equivalent to the following one, but as for iEM, the computational cost of the implementation is not equivalent.

lem:FIEM:MeanField

Lemma 3. *Assume H1item 1-item 2, H2 and H3item 1. Let $\{\gamma_k,\ k \in \mathbb{N}\}$ be a positive stepsize sequence and $\{\lambda_k,\ k \in \mathbb{N}\}$ be a real valued sequence, such that $\widehat{S}^k \in \mathcal{S}$ for any k.*

1. *Fast Incremental EM is equivalent to the following algorithm : initialize*

$$\widehat{S}^0 \in \mathcal{S}, \qquad \mathsf{S}_{0,i} \stackrel{\text{def}}{=} \bar{s}_i \circ \mathsf{T}(\widehat{S}^0),\ 1 \leq i \leq n$$

and repeat for $k \geq 0$: draw independently $I_{k+1}, J_{k+1} \sim \mathcal{U}(\{1, \ldots, n\})$ and set

$$\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}, i \neq I_{k+1}, \qquad \mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k),$$
$$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}\left(\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \lambda_{k+1}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathsf{S}_{k+1,i} - \mathsf{S}_{k+1,J_{k+1}}\right\}\right).$$

2. *Define the filtrations $\mathcal{F}_0 \stackrel{\text{def}}{=} \sigma(\widehat{S}^0)$, $\mathcal{F}_{1/2} \stackrel{\text{def}}{=} \sigma(\widehat{S}^0, I_1)$ and for $k \geq 1$,*

$$\mathcal{F}_k \stackrel{\text{def}}{=} \sigma\left(\widehat{S}^0, I_1, J_1, I_2, \ldots, I_k, J_k\right),$$
$$\mathcal{F}_{k+1/2} \stackrel{\text{def}}{=} \sigma\left(\widehat{S}^0, I_1, J_1, I_2, \ldots, I_k, J_k, I_{k+1}\right);$$

then $\widehat{S}^k \in \mathcal{F}_k$ and

$$\mathbb{E}\left[\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \lambda_{k+1}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathsf{S}_{k+1,i} - \mathsf{S}_{k+1,J_{k+1}}\right\} \middle| \mathcal{F}_{k+1/2}\right] = h(\widehat{S}^k),$$

where h is given by (8).

Lemma 3 outlines that the sequence $\{(\widehat{S}^k, \mathsf{S}_{k,\cdot}), k \in \mathbb{N}\}$ is a Stochastic Approximation scheme. It necessitates the storage of a quantity $\mathsf{S}_{k,\cdot}$ whose size is proportional to n; in some cases (see e.g. (Schmidt et al., 2017, Section 4.1)), it is exactly a vector of length n.