FAST INCREMENTAL EXPECTATION-MAXIMIZATION ALGORITHM: $\sqrt{N}$ ITERATIONS FOR AN ϵ-STATIONARY POINT ?

Pierre Gach, Gersende Fort, Eric Moulines

FAST INCREMENTAL EXPECTATION-MAXIMIZATION ALGORITHM: $\sqrt{N}$ ITERATIONS FOR AN ϵ-STATIONARY POINT ?

P. Gach[1], *G. Fort*[1], *E. Moulines*[2],

[1] IMT, Université de Toulouse & CNRS, F-31062 Toulouse, France.
[2] CMAP, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France.

ABSTRACT

Fast Incremental Expectation Maximization (FIEM) is an iterative algorithm, based on the Expectation Maximization (EM) algorithm, which was introduced to design EM for the large scale learning framework by avoiding the full data set to be processed at each iteration. In this paper, we first recast this algorithm in the *Stochastic Approximation (SA) within EM* framework. Then, we provide non asymptotic convergence rates as a function of the batch size n and of the maximal number of iterations $K_{\max}$ fixed by the user. This allows a complexity analysis: in order to reach an ϵ-approximate solution, how does $K_{\max}$ depend upon n and ϵ?

Index Terms— Statistical Learning, Large Scale Learning, Non convex optimization, Iterative Expectation Maximization algorithm, Accelerated Stochastic Approximation, Control Variate.

1. INTRODUCTION

EM [1, 2] is a very popular computational tool, designed to solve non convex minimization problems on $\mathbb{R}^d$ when the objective function is not explicit but defined as $F(\theta) = -\log \int_{\mathsf{Z}} G(z;\theta)\mathrm{d}\mu(z)$ for a positive function G. EM is a Majorize-Minimization algorithm which, based on the current value of the point θ_{c}, defines a majorizing function $\theta \mapsto Q(\theta, \theta_{\mathrm{c}})$ through a Kullback-Leibler argument; then, the new point is chosen as the/a minimum of Q. The computation of a function at each iteration can be greedy and even intractable; in many applications, Q has a special form: there exist (known and explicit) functions ψ, ϕ, s such that $Q(\cdot, \theta_{\mathrm{c}}) = \psi(\cdot) - \langle \bar{s}(\theta_{\mathrm{c}}), \phi(\cdot)\rangle$ and $\bar{s}(\tau)$ is the expectation of the function s with respect to (w.r.t.) the probability distribution $G(\cdot;\tau)\exp(-F(\tau))\mathrm{d}\mu$. In these cases, the definition of Q consists in the computation of the vector $\bar{s}(\theta_{\mathrm{c}})$.
It may happen that the expectation $\bar{s}(\theta_{\mathrm{c}})$ is not explicit (see e.g. [3, section 6]); a natural idea is to substitute $\bar{s}$ for an approximation, possibly random. Many stochastic EM versions were proposed and studied: among them, let us cite Monte Carlo EM [4, 5] where $\bar{s}$ is approximated by a Monte Carlo sum; and SA EM ([6, 7]) where $\bar{s}$ is approximated by a SA scheme [8]. With the Big Data era, EM applied to statistical learning evolved into online versions and large scale versions: the objective function is a loss function associated to a set of observations (also called *examples*); in online versions, the data are not stored and are processed online which means that the objective function is time-varying (see e.g. [9, 10]); in large scale versions, a batch of observations is given but it is too large to be processed at each iteration of EM.
This paper is devoted to the convergence analysis of FIEM [11], an EM-based algorithm designed for large scale learning in a non-convex setting: FIEM is also a *SA within EM* algorithm, with a SA scheme for the approximation of $\bar{s}(\theta_{\mathrm{c}})$ which combines *(i)* a random selection of a single (or a few) observation(s) in the large batch, and *(ii)* a variance reduction technique inspired from SAGA [12] (see also sEM-VR [13] which uses SVRG [14]). In Section 2, some *SA within EM* algorithms are described; FIEM is explicitly recasted as such an algorithm. Our contribution is essentially the content of Section 3: we provide non asymptotic convergence rates as a function of the batch size n and of the maximal number of iterations $K_{\max}$ fixed by the user. In this non convex optimization setting, errors are relative to L^1-convergence when stopping FIEM at a random time K, prior to $K_{\max}$ and chosen independently of the FIEM sources of randomness. We recover previous rates by [11] in the case K is a uniform distribution, and improve them from $n^{2/3}K_{\max}^{-1}$ to $n^{1/3}K_{\max}^{-2/3}$; our analysis also includes a definition of the step sizes in the SA scheme, with an explicit dependence upon n and $K_{\max}$. A corollary of these bounds is a complexity analysis: to reach an ϵ-approximate solution, we show that either $K_{\max} = O(n^{2/3}\epsilon^{-1})$ and the step size is constant scaling as $O(n^{-2/3})$; or $K_{\max} = O(n^{1/2}\epsilon^{-3/2})$ and the step size is constant scaling as $O(n^{-1/2})$. The last contribution establishes that the same convergence rates are obtained with any random stopping rule K (bounded by $K_{\max}$), up to the choice of an adequate time-varying step size. In Section 4, through a toy example, we explore an extension of FIEM.

2. INCREMENTAL EM ALGORITHMS

This paper addresses explicit convergence rates for an algorithm designed to solve the optimization problem

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

when $\Theta \subseteq \mathbb{R}^d$ and F can not be explicitly evaluated (nor its gradient or derivatives of higher order when they exist). Two levels of intractability of $F(\theta)$ are considered. The first one is motivated by the large scale learning setting when the number n is so large that computations involving a sum over n terms are not allowed or have to be quite rare along the run of the optimization algorithm. The second one is motivated by the latent variable statistical framework, where for all i, the function $\mathcal{L}_i$ is not explicit and defined through an integral. A large class of computational learning problems is covered by this framework: it includes for example the situations when n is the number of examples, $\mathcal{L}_i$ is a loss function associated to example $\#i$ and R is a penalty term. In this paper, a specific expression for $\mathcal{L}_i$ is considered: we restrict our attention to the case when

$$\mathcal{L}_i(\theta) \stackrel{\mathrm{def}}{=} -\log \int_{\mathsf{Z}} h_i(z)\exp\left(\langle s_i(z), \phi(\theta)\rangle - \psi(\theta)\right)\mu(\mathrm{d}z)\ . \quad (2)$$

This work is partially supported by the French *Agence Nationale de la Recherche* (ANR), project under reference ANR-PRC-CE23 MASDOL; and by the *Fondation Simone et Cino Del Duca* through the project OpSiMorE.

$n^{-1}\sum_{i=1}^n \mathcal{L}_i$ can be seen as the normalized negative log-likelihood of n observations in a latent variable model: *(i)* conditionally to the missing variables, the observations are independent thus yielding to an additive expression of the log-likelihood; *(ii)* the model for the joint distribution of the observation y_i and its associated missing variable is a curved exponential family w.r.t. a dominating positive measure $\mathrm{d}\mu$ on a measurable set $(\mathsf{Z}, \mathcal{Z})$. In (2), the dependence upon the observation y_i is omitted (but appears implicitly through the dependence upon i of the functions h_i, s_i). Let us introduce rigorous conditions on the problem at hand. Denote by $[[n]] \stackrel{\text{def}}{=} \{1, \ldots, n\}$.

A1 *$\Theta \subseteq \mathbb{R}^d$ is an open set. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\mathsf{R} : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$, $\psi : \Theta \to \mathbb{R}$, $s_i : \mathsf{Z} \to \mathbb{R}^q$, $h_i : \mathsf{Z} \to \mathbb{R}_+$ for all $i \in [[n]]$ are measurable functions. Finally, for any $\theta \in \Theta$ and $i \in [[n]]$, $-\infty < \mathcal{L}_i(\theta) < \infty$.*

Under A1, for any $\theta \in \Theta$ and $i \in [[n]]$, the quantity $p_i(\cdot;\theta)\,\mathrm{d}\mu$

$$p_i(z;\theta) \stackrel{\text{def}}{=} h_i(z)\exp\left(\langle s_i(z), \phi(\theta)\rangle - \psi(\theta) + \mathcal{L}_i(\theta)\right) , \qquad (3)$$

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

A2 *The expectation $\bar{s}_i(\theta) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} s_i(z)\; p_i(z;\theta)\mu(\mathrm{d}z)$ exists for all $\theta \in \Theta$ and $i \in [[n]]$.*
For any $s \in \mathbb{R}^q$, the following set is a (non empty) singleton denoted by $\{\mathsf{T}(s)\}$:

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

Define

$$\bar{s}(\theta) \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n \bar{s}_i(\theta) \; . \qquad (4)$$

A3 *The functions ϕ, ψ and R are continuously differentiable on Θ. T is continuously differentiable on $\mathbb{R}^q$.*
For any $s \in \mathbb{R}^q$, $B(s) \stackrel{\text{def}}{=} \nabla(\phi \circ \mathsf{T})(s)$ is a symmetric $q \times q$ matrix and there exist $0 < v_{min} \le v_{max} < \infty$ such that for all $s \in \mathbb{R}^q$, the spectrum of $B(s)$ is in $[v_{min}, v_{max}]$.
For any $i \in [[n]]$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i. Set $L^2 \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$. The function $s \mapsto B^T(s)\,(\bar{s} \circ T(s) - s)$ is globally Lipschitz on $\mathbb{R}^q$ with constant $L_{\dot{V}}$.

2.1. An EM algorithm in the statistic-space

In this framework - even including a penalty term R in the objective function-, the EM algorithm is usually described as an iterative algorithm in the Θ-space: given a current value $\tau_k \in \Theta$, the next point is obtained by a combination of an expectation step which here, computes $\bar{s}(\tau^k)$, and a maximization step through the map T yielding to $\tau^{k+1} \stackrel{\text{def}}{=} \mathsf{T} \circ \bar{s}(\tau^k)$. Equivalently (see [7]), by using T which maps $\mathbb{R}^q$ to Θ, it can be described in the $\mathbb{R}^q$-space: given the current value $\bar{s}^k \in \bar{s}(\Theta)$, set $\bar{s}^{k+1} \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(\bar{s}^k)$. In this second point of view, which is adopted throughout this paper, the limiting points are characterized as the roots of the field h

$$\{s \in \bar{s}(\Theta) : h(s) = 0\}, \qquad h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s \; . \qquad (5)$$

Assumption A3 implies that the roots of h are the zeros of $\nabla(F \circ \mathsf{T})$ i.e. the critital points of the objective function. Unfortunately, in the large scale learning setting, EM can not be used since each iteration involves the computation of a sum over n terms through $\bar{s}$.

2.2. A SA-based algorithm

A natural idea to overcome this intractability is the use of an iterative SA scheme for finding the roots of h upon noting that

$$h(s) = \frac{1}{n}\sum_{i=1}^n \bar{s}_i \circ \mathsf{T}(s) - s = \mathbb{E}\left[\bar{s}_I \circ \mathsf{T}(s) - s\right] \; ,$$

where $I \sim \mathcal{U}([[n]])$ is a uniform random variable (r.v.) on $[[n]]$. This yields Algorithm 1, where $K_{\max}$ is the total number of iterations, $\widehat{S}^0$ is the initial value and $\{\gamma_k, k \ge 1\}$ are positive step sizes.

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathbb{R}^q$, $\gamma_k \in (0,\infty)$ for $k \in [[K_{\max}]]$
Result: The SA sequence: $\widehat{S}^k, k = 0, \ldots, K_{\max}$
for $k = 0, \ldots, K_{\max} - 1$ **do**
  $I_{k+1} \sim \mathcal{U}([[n]])$;
  $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}(\bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k)$

Algorithm 1: The Stochastic Approximation (SA) algorithm

2.3. An Incremental EM algorithm (i-EM)

Another idea, introduced by [15] and studied in [16], can be seen as a two-level SA schemes where an auxiliary level is introduced in order to mimic the computation of $n^{-1}\sum_{i=1}^n \bar{s}_i \circ \mathsf{T}(\widehat{S}^k)$ at each iteration of the algorithm; $\widehat{S}^k$ is the current point of the algorithm at iteration k. i-EM is described by Algorithm 2 with a slight adaptation (in the original algorithm, $\gamma_{k+1} = 1$). Line (8) defines the i-EM sequence; the update rule is based on $\widetilde{S}^{k+1}$, defined through Lines (4) to (7), and which satisfies for any $k \ge 0$

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

where $\widehat{S}^{<0,i} \stackrel{\text{def}}{=} \widehat{S}^0$ for all $i \in [[n]]$ and for $k \ge 0$, $\widehat{S}^{<k+1,i} = \widehat{S}^\ell$ where ℓ stands for the current index of the statistics $\widehat{S}^\bullet$ during the last "visit" to the observation $\#i$: $\ell = k$ if $I_{k+1} = i$; $\ell \in [[0, k-1]]$ if $I_{k+1} \ne i, \ldots, I_{\ell+2} \ne i$ and $I_{\ell+1} = i$; $\ell = 0$ otherwise. The auxiliary quantity $\widetilde{S}^{k+1}$ is an online estimate of $\bar{s} \circ \mathsf{T}(\widehat{S}^k)$, where at iteration $k+1$, only the contribution of the observation $\#I_{k+1}$ of the sum is updated. Note however that this algorithm, while avoiding a sum over n terms at each iteration, necessitates the storage of a vector $\mathsf{S}_{k,\cdot} \in \mathbb{R}^{qn}$ whose length is proportional to n.

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathbb{R}^q$, $\gamma_k \in (0,\infty)$ for $k \in [[K_{\max}]]$
Result: The iEM sequence: $\widehat{S}^k, k = 0, \ldots, K_{\max}$
1 $\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}^0)$ for all $i \in [[n]]$;
2 $\widetilde{S}^0 = n^{-1}\sum_{i=1}^n \mathsf{S}_{0,i}$;
3 **for** $k = 0, \ldots, K_{\max} - 1$ **do**
4 $I_{k+1} \sim \mathcal{U}([[n]])$;
5 $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \ne I_{k+1}$;
6 $\mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;
7 $\widetilde{S}^{k+1} = \widetilde{S}^k + n^{-1}\left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}}\right)$;
8 $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}(\widetilde{S}^{k+1} - \widehat{S}^k)$

Algorithm 2: The incremental EM (i-EM) algorithm

2.4. A Fast Incremental EM algorithm (FIEM)

More recently, [11] introduced FIEM, another incremental EM algorithm; they showed it is faster than i-EM and SA. FIEM combines the two-levels SA idea of i-EM with an acceleration technique inspired from SAGA [12]. The algorithm is described by Algorithm 3. This algorithm can be seen as a mix of the SA update (see the term $T_{k+1} \stackrel{\text{def}}{=} \bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k$ in Line 9) and the (centered) control variate $V_{k+1} \stackrel{\text{def}}{=} \widetilde{S}^{k+1} - \mathsf{S}_{k+1,J_{k+1}}$, which is correlated to T_{k+1} through the random index J_{k+1}. The auxiliary quantity $\widetilde{S}^{k+1}$ is the same as the one introduced in i-EM (see section 2.3).

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathbb{R}^q$, $\gamma_k \in (0,\infty)$ for $k \in [[K_{\max}]]$
Result: The FIEM sequence: $\widehat{S}^k, k = 0, \ldots, K_{\max}$

1 $\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}^0)$ for all $i \in [[n]]$;
2 $\widetilde{S}^0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;
3 **for** $k = 0, \ldots, K_{\max} - 1$ **do**
4 $\quad I_{k+1} \sim \mathcal{U}([[n]])$;
5 $\quad \mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \neq I_{k+1}$;
6 $\quad \mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;
7 $\quad \widetilde{S}^{k+1} = \widetilde{S}^k + n^{-1}\left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}}\right)$;
8 $\quad J_{k+1} \sim \mathcal{U}([[n]])$;
9 $\quad \widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}(\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \widetilde{S}^{k+1} - \mathsf{S}_{k+1,J_{k+1}})$

Algorithm 3: The Fast Incremental EM (FIEM) algorithm

3. FIEM: NON ASYMPTOTIC CONVERGENCE RATES

The originality of our contribution is to provide new explicit non asymptotic error rates for FIEM. The proofs of the statements below can be found in [17] Since the problem (1) is most often a non convex one, convergence is considered here in terms of the rate at which the following quantities E_0 to E_2 decay to zero as a function of the size of the sample n, and a function of a total number of iterations $K_{\max}$ chosen by the user. As in [11] (see also [18]), we derive L^1-error rates along a FIEM sequence stopped at a random time K, chosen independently of the sequence. The quantities of interest are

$$\mathsf{E}_0 \stackrel{\text{def}}{=} \mathbb{E}\left[\|\nabla V(\widehat{S}^K)\|^2\right], \qquad V \stackrel{\text{def}}{=} F \circ \mathsf{T} ,$$
$$\mathsf{E}_1 \stackrel{\text{def}}{=} \mathbb{E}\left[\|\bar{s} \circ \mathsf{T}(\widehat{S}^K) - \widehat{S}^K\|^2\right] = \mathbb{E}\left[\|h(\widehat{S}^K)\|^2\right],$$
$$\mathsf{E}_2 \stackrel{\text{def}}{=} \mathbb{E}\left[\|\widetilde{S}^{K+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^K)\|^2\right] .$$

They respectively quantify, at the random stopping time K, the mean squared norm of the gradient of the objective function F (when seen as a function on $\mathbb{R}^q$ through the map T); the mean squared (kind of) distance to the set of the limiting points (see (5)); and the mean squared error when approximating $\bar{s} \circ \mathsf{T}(\widehat{S}^K)$ by $\widetilde{S}^{K+1}$.

Proposition 1 provides a control of E_i's upper bound in the case K is sampled uniformly on $\{0, \ldots, K_{\max} - 1\}$: as in [11], the control is proportional to $n^{2/3}/K_{\max}$ and the dependence upon n of the step size is $O(n^{-2/3})$; the constant in the control, and the exact value of the step size are improved w.r.t. [11] (see section 4). Proposition 2 shows that by using another strategy for the step size, while still being constant over iterations, the control of the errors evolves as $n^{1/3}/K_{\max}^{2/3}$. To our best knowledge, this is a new result in the literature. Set $\Delta V \stackrel{\text{def}}{=} \mathbb{E}\left[V(\widehat{S}^0) - V(\widehat{S}^K_{\max})\right]$ and for $n \geq 2$, $C \in (0,1)$,

$$f_n(C) \stackrel{\text{def}}{=} \frac{L_{\dot{V}}}{2L}\left(\frac{1}{n^{2/3}} + \frac{1}{1 - n^{-1/3}}\left(\frac{1}{n} + \frac{1}{1-C}\right)\right) .$$

Proposition 1 *Assume A1 to A3 and choose $\mu \in (0,1)$. Let K be a $\{0, \ldots, K_{\max} - 1\}$-valued uniform r.v. Run FIEM with a constant step size $\gamma_\ell = \sqrt{C} n^{-2/3} L^{-1}$ where $C \in (0,1)$ is the unique solution of*

$$\sqrt{C} f_n(C) = \mu v_{\min} . \tag{6}$$

Then, for any $n \geq 2$ and $K_{\max} \geq 1$

$$v_{\max}^{-2}\mathsf{E}_0 \leq \mathsf{E}_1 + \frac{L_{\dot{V}}}{2L}\frac{\sqrt{C}}{v_{\min} n^{2/3}}\mathsf{E}_2 \leq \frac{n^{2/3}}{K_{\max}}\frac{L\,\Delta V}{\sqrt{C}(1-\mu)v_{\min}} .$$

The constant C in (6) is upper bounded by the unique point C^+ in $(0,1)$ solving $v_{\min} L(1-x) - \sqrt{x} L_{\dot{V}} = 0$; thus showing that $L_{\dot{V}}(1-C^+)^{-1}/(2L) \leq f_n(C) \leq \sup_n f_n(C^+) < \infty$. Hence, the constant C can be lower bounded and upper bounded (away from 0 and 1) by a constant depending only upon $v_{\min}, L$ and $L_{\dot{V}}$.

Proposition 2 *Assume A1 to A3 and choose $\mu \in (0,1)$. Let K be a $\{0, \ldots, K_{\max} - 1\}$-valued uniform r.v. Run FIEM with a constant step size $\gamma_\ell = \sqrt{C} n^{-1/3} K_{\max}^{-1/3} L^{-1}$ where $C > 0$ satisfies*

$$\sqrt{C}\frac{L_{\dot{V}}}{2L}\left(1 + C\left(1 + \frac{1}{1-\lambda}\right)\right) \leq \mu v_{\min} , \tag{7}$$

for some $\lambda \in (0,1)$. Then, for any $n, K_{\max} \geq 1$ such that $n^{1/3} K_{\max}^{-2/3} \leq \lambda/C$,

$$v_{\max}^{-2}\mathsf{E}_0 \leq \mathsf{E}_1 + \frac{L_{\dot{V}}}{2L}\frac{\sqrt{C}}{v_{\min} n^{1/3} K_{\max}^{1/3}}\mathsf{E}_2 \leq \frac{n^{1/3}}{K_{\max}^{2/3}}\frac{L\,\Delta V}{\sqrt{C}(1-\mu)v_{\min}} .$$

As a corollary of Proposition 2, it may be shown that there exists a constant $M \in (1, +\infty)$ depending upon $v_{\min}, L, L_{\dot{V}}, \mu$ such that for any $\varepsilon \in (0,1)$, we have

$$\frac{n^{1/3}}{K_{\max}^{2/3}}\frac{L}{\sqrt{C}(1-\mu)v_{\min}} \leq \varepsilon ,$$

by setting $K_{\max} = M\sqrt{n}\varepsilon^{-3/2}$; which in turn implies that $\gamma_\ell \propto 1/\sqrt{n}$. From Proposition 1, the complexity is $K_{\max} = O(n^{2/3}\varepsilon^{-1})$. Proposition 2 provides a new rate which improves on the known result $n^{2/3}$, but at a cost on the dependence upon the precision ε. These two propositions are complementary, one providing a better strategy than the other one depending on how $\sqrt{\varepsilon}$ compares with $n^{-1/6}$.

We conclude this section by another point of view: given a probability distribution $p_0, \ldots, p_{K_{\max}-1}$ for the random stopping time K, how to choose the step sizes γ_k in order to reach the same controls (in n and $K_{\max}$) as in the above propositions ? we restrict here to the "mirror" of Proposition 1. For $C \in (0,1)$ and $n \geq 2$, define the function $F_{n,C}$

$$F_{n,C} : \quad x \mapsto \frac{1}{Ln^{2/3}} x\left(v_{\min} - x f_n(C)\right) .$$

$F_{n,C}$ is positive, increasing and continuous on $(0, v_{\min}/(2f_n(C))]$.

Proposition 3 *Assume A1 to A3. Let K be a $\{0,\ldots,K_{\max}-1\}$-valued r.v. with distribution $p_0,\ldots,p_{K_{\max}-1}$, $\inf_k p_k > 0$. Let $C \in (0,1)$ solving*

$$\sqrt{C} f_n(C) = \frac{1}{2} v_{\min} \ . \tag{8}$$

For any $n \geq 2$ and $K_{\max} \geq 1$, we have

$$v_{\max}^{-2} \mathsf{E}_0 \leq \mathsf{E}_1 + \frac{L_{\dot{V}}}{L} \frac{\min_k \upsilon_k}{\sqrt{C} v_{\min} n^{2/3}} \mathsf{E}_2 \leq n^{2/3} \max_k p_k \frac{2L\Delta V}{\sqrt{C} v_{\min}} \ ,$$

where $\upsilon_k \stackrel{\text{def}}{=} g_k^2 \max_\ell p_\ell / p_k$ and FIEM is run with

$$\gamma_{k+1} \stackrel{\text{def}}{=} \frac{g_k}{n^{2/3} L} \ , \ g_k \stackrel{\text{def}}{=} F_{n,C}^{-1}\left(\frac{p_k}{\max_\ell p_\ell} \frac{v_{\min}\sqrt{C}}{2L} \frac{1}{n^{2/3}} \right) \ .$$

Since $\sum_k p_k = 1$, we have $\max_k p_k \geq K_{\max}^{-1}$ thus showing that among the distributions $\{p_k, k = 0,\ldots,K_{\max}-1\}$, $\max_k p_k$ is minimal with the uniform distribution. In that case, Proposition 1 applied with $\mu = 1/2$ and Proposition 3 provide exactly the same control: *(i)* the control evolves as $n^{2/3}/K_{\max}$; *(ii)* the constant C solving (6) in the case $\mu = 1/2$ is the same as the one solving (8); *(iii)* since $F_{n,C}^{-1}(v_{\min}^2 \sqrt{C} n^{-2/3}/(2L)) = \sqrt{C}$, then $g_k^2 = C$ and $\upsilon_k = C$ so that the controls of E_2 are the same in both propositions; *(iv)* the step sizes are equal since $g_k = \sqrt{C}$.
The choice of the constant C is also crucial on a numerical point of view, since it defines the step size γ_ℓ: a large one may cause instability and a small one makes the convergence longer (see section 4). We provided here simple conditions for finding C but there are more intricate conditions than (6), (7) (8) yielding to larger constants C. For example, Proposition 1 holds for C satisfying: there exists $\lambda \in (0,1)$ s.t. $n^{-1/3} < \lambda/C$ and

$$\sqrt{C} \frac{L_{\dot{V}}}{2L} \left(\frac{1}{n^{2/3}} + \frac{C}{\lambda - C n^{-1/3}} \left(\frac{1}{n} + \frac{1}{1-\lambda} \right) \right) = \mu v_{\min} \ . \tag{9}$$

4. NUMERICAL INVESTIGATION

Consider a toy example in order to *(i)* illustrate the role of the step size on the efficiency of FIEM and to compare different definitions; *(ii)* illustrate the interest of the variance reduction technique by comparing SA, FIEM and a third strategy called below *FIEM-coeff*. Both SA and FIEM update $\widehat{S}^k$ by a scheme of the form $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} H_{k+1}$ where $H_{k+1} = T_{k+1} + \lambda_{k+1} V_{k+1}$ and $\lambda_{k+1} = 1$ for FIEM; and $H_{k+1} = T_{k+1}$ for SA (see section 2.4 for a definition of T_{k+1} and V_{k+1}). The use of V_{k+1} can be seen as a control variate approach [19]: if such, the optimal coefficient $\lambda_{k+1}^\star$, defined as the quantity minimizing $\mathbb{E}\left[\|H_{k+1}\|^2 | \widehat{S}^k \right]$, depends on the correlation of V_{k+1} and T_{k+1}, which is not always equal to one. Below, under the name *FIEM-coeff*, we explore the benefit of the strategy $\lambda_{k+1} = \lambda_{k+1}^\star$ - which in a realistic example, has a non negligible computational cost and will necessitate to design an approximation.

F is a penalized negative log-likelihood function: the $n = 1e3$ observations are modeled as independent; each observation Y_i, conditionally to a latent variable Z_i, is a $\mathbb{R}^{15}$-valued Normal distribution with mean AZ_i and covariance matrix I; Z_i is a $\mathbb{R}^{10}$-valued Normal distribution with mean $X\theta_{\text{true}}$ and covariance matrix I; A and X are known; $\theta_{\text{true}} \in \mathbb{R}^{20}$ is unknown. The penalty term is $\mathsf{R}(\theta) = 0.1\|\theta\|^2$. In this toy example, F possesses an unique minimum which has an explicit expression in terms of A, X and $n^{-1}\sum_{i=1}^n Y_i$; the constants $v_{\min}, v_{\max}, L, L_{\dot{V}}$ are also explicit. We have $s_i(z) = X^T z$, $\phi(\theta) = \theta$ and $\psi(\theta) = \theta^T X^T X \theta/2$; $p_i(z;\theta)$ is a Normal density with explicit parameters; $\mathsf{T}(s) = (\lambda \mathrm{I} + X^T X)^{-1} s$.

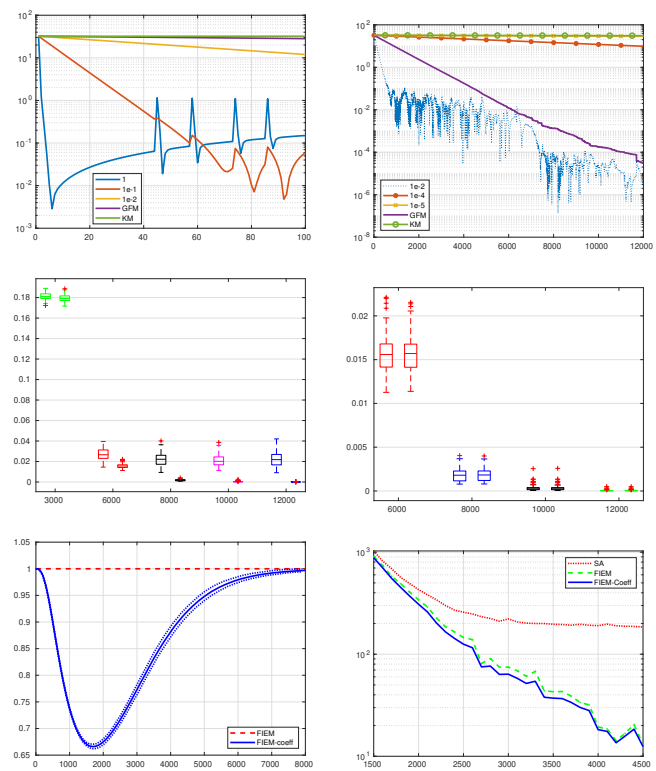

Fig. 1. [top] Role of the step size; [middle] SA, FIEM and FIEM-coeff; [bottom] FIEM and FIEM-coeff

Figure 1[top] displays the error $k \mapsto \|\theta_k - \theta_\star\|$ along FIEM paths (top,right) with a zoom on the first iterations (top,left). The paths are run with different step sizes: $\gamma \in \{1, 0.1, 0.01, \ldots\}$; $\gamma_{\text{GFM}} = \sqrt{C_{\text{GFM}}} n^{-2/3} L^{-1}$ where C_{GFM} solves (9), and γ_{KM} is given in [11]. Here, $\gamma_{\text{GFM}} \approx 1.4\,10^{-3}$ and $\gamma_{\text{KM}} \approx 3.4\,10^{-6}$. We observe that large step sizes may cause numerical instability; and here, γ_{GFM} is larger than γ_{KM} by a factor 300, thus yielding to a faster convergence rate. Figure 1[middle] displays the boxplot of $\|\theta_k - \theta_\star\|$ at iteration $k = \{3e3, 6e3, 8e3, 10e3, 12e3\}$; on the left subplot, SA and FIEM are compared (resp. left/right boxplot) and on the right subplot, FIEM and *FIEM-coeff* are compared. FIEM clearly improves on SA; and in the convergence phase ($k \geq 6e3$), FIEM and *FIEM-coeff* look similar. The boxplots are obtained with 100 independent realizations. Figure 1[bottom,left] displays $k \mapsto \lambda_k$, for FIEM ($\lambda_k = 1$) and for *FIEM-coeff* ($\lambda_k = \lambda_k^\star$). The plot is the mean value over 1e3 independent runs (the quantiles 0.25 and 0.75 are displayed in dotted line). It is shown that λ_k is smaller than 0.85 in the transient phase $k \in [5e2, 3.5e3]$. On Figure 1[bottom,right], a Monte Carlo estimation (over $1e3$ independent runs) of $\mathbb{E}\left[\|H_{k+1}\|^2\right]$ for SA, FIEM and *FIEM-coeff* is displayed for $k \in \{1.5e3, \ldots, 4.5e3\}$: *FIEM-coeff* can improve on FIEM up to 15% in the transient phase. In this example, $\lambda_{k+1}^\star$ is explicit; but the benefit will be investigated in future works, taking into account a numerical cost for its approximation. $\mathbb{E}\left[\|H_{k+1}\|^2\right]$ is a crucial quantity since the convergence rates derived in section 3 are obtained by upper bounding this quantity (see [17]).

5. REFERENCES

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc. B Met.*, vol. 39, no. 1, pp. 1–38, 1977.

[2] C.F.J. Wu, "On the Convergence Properties of the EM Algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[3] G.J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, Wiley series in probability and statistics. Wiley, 2008.

[4] G.C.G. Wei and M.A. Tanner, "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *J. Am. Stat. Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.

[5] G. Fort and E. Moulines, "Convergence of the Monte Carlo expectation maximization for curved exponential families," *Ann. Statist.*, vol. 31, no. 4, pp. 1220–1259, 2003.

[6] G. Celeux and J. Diebolt, "The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem," *Computational Statistics Quarterly*, vol. 2, pp. 73–82, 1985.

[7] 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, 1999.

[8] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer Verlag, 1990.

[9] O. Cappé and E. Moulines, "On-line Expectation Maximization algorithm for latent data models," *J. Roy. Stat. Soc. B Met.*, vol. 71, no. 3, pp. 593–613, 2009.

[10] S. Le Corff and G. Fort, "Online Expectation Maximization based algorithms for inference in Hidden Markov Models," *Electron. J. Statist.*, vol. 7, pp. 763–792, 2013.

[11] B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle, "On the Global Convergence of (Fast) Incremental Expectation Maximization Methods," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds., pp. 2837–2847. Curran Associates, Inc., 2019.

[12] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 1646–1654. Curran Associates, Inc., 2014.

[13] J. Chen, J. Zhu, Y.W. Teh, and T. Zhang, "Stochastic Expectation Maximization with Variance Reduction," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., pp. 7967–7977. Curran Associates, Inc., 2018.

[14] R. Johnson and T. Zhang, "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 315–323. Curran Associates, Inc., 2013.

[15] R. M. Neal and G. E. Hinton, *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*, pp. 355–368, Springer Netherlands, Dordrecht, 1998.

[16] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *J. Mach. Learn. Res.*, vol. 6, pp. 2049–2073, 2005.

[17] P. Gach, G. Fort, and E. Moulines, "(to be confirmed) Fast Incremental Expectation-Maximization algorithm: non asymptotic convergence rates for an ϵ-stationary point," Tech. Rep., https://perso.math.univ-toulouse.fr/gfort, 2020, A preliminary version of the paper is available upon request (write to the second author, Pr G. Fort) and the paper should be accessible from the webpage of G. Fort by April 1st, 2020.

[18] S. Ghadimi and G. Lan, "Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming," *SIAM J. Optimiz.*, vol. 23, no. 4, pp. 2341–2368, 2013.

[19] P. Glasserman, *Monte Carlo methods in financial engineering*, Springer, New York, 2004.