Statistics and Computing manuscript No.
(will be inserted by the editor)

Fast Incremental Expectation Maximization for non-convex finite-sum optimization: non asymptotic convergence bounds

G. Fort · P. Gach · E. Moulines

Received: date / Accepted: date

Abstract Fast Incremental Expectation Maximization was introduced to design Expectation-Maximization (EM) for the large scale learning framework involving finite-sum and possibly non-convex optimization, in order to avoid the processing of the full data set at each iteration by considering an incremental assimilation. In this paper, we first recast this iterative algorithm and other incremental EM type algorithms in the *Stochastic Approximation within EM* framework. Then, we provide non asymptotic convergence bounds as a function of the number of examples n and of the maximal number of iterations $K_{\max}$ fixed by the user; as a corollary, we propose two strategies for reaching an ϵ-approximate stationary point: either with $K_{\max} = O(n^{2/3}\epsilon^{-1})$ or with $K_{\max} = O(\sqrt{n}\epsilon^{-3/2})$, both strategies relying on a random termination rule before $K_{\max}$ and on a constant step size in the Stochastic Approximation step. Our bounds are explicit and can be compared to previous results. We are the first to provide a complexity bound which scales in n as $\sqrt{n}$ and such a rate is the best one among the incremental EM methods proposed so far; it is at the cost of a larger dependence upon the tolerance ϵ thus making

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

G. Fort
(corresponding author) Institut de Mathématiques de Toulouse & CNRS, France
E-mail: gersende.fort@math.univ-toulouse.fr

P. Gach
Institut de Mathématiques de Toulouse & Université Toulouse 3, France
E-mail: pierre.gach@math.univ-toulouse.fr

E. Moulines
CMAP & Ecole Polytechnique, France
E-mail: eric.moulines@polytechnique.edu

this control relevant for small to medium accuracy with respect to the number of examples n. For the $n^{2/3}$-rate, our bounds show a numerical improvement thanks to a tighter definition of crucial quantities playing a role in the efficiency of the algorithm.

L'abstract doit être entre 150 et 250 mots; il en fait 248. Si modifs, vérifier la longueur par exemple sur le site
https://www.compteurdelettres.com/mots.html

Keywords Computational Statistical Learning · Large Scale Learning · Incremental Expectation Maximization algorithm · Momentum Stochastic Approximation · Non-convex optimization.

Mathematics Subject Classification (2010) MSC: 65C60 · 68Q32 · 65K10

1 Introduction

Expectation Maximization (EM) is a very popular tool introduced by Dempster et al. (1977) to solve non linear programming on $\Theta \subseteq \mathbb{R}^d$ when the function F to be minimized, possibly non convex, is not explicit and defined through an integral:

$$F(\theta) = -\frac{1}{n} \log \int_{\mathsf{Z}_n} G(z;\theta) \, \mathrm{d}\mu_n(z) \, , \tag{1}$$

for $n \in \mathbb{N} \setminus \{0\}$, a positive function G and a σ-finite positive measure μ_n on a measurable set $(\mathsf{Z}_n, \mathcal{Z}_n)$. EM is a Majorize-Minimization (MM) algorithm which, based on the current value of the iterate θ_{curr}, defines a majorizing function $\theta \mapsto \mathsf{Q}(\theta, \theta_{\mathrm{curr}})$ through a Kullback-Leibler argument: up to an additive constant,

$$\mathsf{Q}(\theta, \theta_{\mathrm{curr}}) = -\frac{1}{n} \int_{\mathsf{Z}_n} \log G(z,\theta) \; G(z,\theta_{\mathrm{curr}}) \exp(n \, F(\theta_{\mathrm{curr}})) \; \mu_n(\mathrm{d}z) \, ;$$

then, the new point is chosen as the/a minimum of $\mathsf{Q}(\cdot, \theta_{\mathrm{curr}})$. Each iteration of EM is divided into two steps: the definition of the surrogate function is called the E step (expectation step), and its optimization is the M step (optimization step). The computation of a function at each iteration can be greedy and even intractable; in many models, $\log G$ has a special form: there exist explicit functions $\phi : \Theta \to \mathbb{R}^q$, $S : \mathsf{Z} \to \mathbb{R}^q$ such that $n^{-1} \log G(z,\theta) = \langle S(z), \phi(\theta) \rangle$. In these cases, the function Q is defined by a vector $\bar{s}(\theta_{\mathrm{curr}})$, equal to the expectation of the function S with respect to (w.r.t.) the probability distribution $G(\cdot;\theta_{\mathrm{curr}}) \exp(n \, F(\theta_{\mathrm{curr}}))\mathrm{d}\mu_n$.

This paper is concerned with the optimization of a function F when $\mathsf{Z}_n = \mathsf{Z}^n$, $S(z) = n^{-1} \sum_{i=1}^n s_i(z_i)$, $\mathrm{d}\mu_n(z) = \otimes_{i=1}^n \mathrm{d}\mu(z_i)$ so that

$$F(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta), \qquad f_i(\theta) \stackrel{\mathrm{def}}{=} -\log \int_{\mathsf{Z}} \exp(\langle s_i(z), \phi(\theta) \rangle) \, \mathrm{d}\mu(z) \, ; \tag{2}$$

it addresses the finite-sum setting in the case n is large so that the computation of the full sum $\bar{s}(\theta_{\rm curr})$ has to be avoided.

When $\bar{s}(\theta_{\rm curr})$ is not explicit (see e.g. (McLachlan and Krishnan, 2008, 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 Wei and Tanner (1990); Fort and Moulines (2003) where $\bar{s}$ is approximated by a Monte Carlo sum; and SA EM (Celeux and Diebolt (1985); Delyon et al. (1999)) where $\bar{s}$ is approximated by a Stochastic Approximation (SA) scheme (see e.g. Benveniste et al. (1990); Borkar (2008) for a description of SA).

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, EM is designed to deal with a stream of data; the algorithms process the data without storing them - or store at most, a mini-batch per iteration; hence, the objective function is time-varying but in some asymptotic and under stationarity conditions on the data stream, the objective function can be compared to a "population version" (see e.g. Cappé and Moulines (2009); Le Corff and Fort (2013a); Balakrishnan et al. (2017); Nguyen et al. (2020)). In large scale versions, a fixed data set is given but it is too large to be fully processed at each iteration of EM: algorithms based on a partial scan of the data at each iteration have to be designed.

This paper deals with this large scale setting, in the situation (2), when F is possibly non-convex and with Lipschitz-continuous gradient. This framework is motivated by computational problems in large scale learning when the n available data are modeled as independent; the function f_i stands for a non-convex loss associated to the example $\#i$ and it can also include a penalty (or a regularization) term. For finite-sum optimization, the existing EM-based algorithms are incremental: at each iteration, a single example or a mini-batch is selected and the E step uses this new information in an iterative updating mechanism of the Q-quantity. The time to convergence of these incremental procedures is always a trade-off between a loss of information since only part of the data are used per iterations, and a quicker exploitation of the new information since the parameter is updated more often without waiting the full scan of the data set.

A pioneering work in this vein is the *incremental EM* by Neal and Hinton (1998): the data set is divided into B blocks and one block is visited per EM iteration chosen through a deterministic cycle or through a random selection. The Q-quantity of *incremental EM* is again a sum over n terms, but each E step consists in updating only one term (or a block) in this sum while the original EM would update the n terms. The role of the size of the blocks is studied for some specific applications (see e.g. Ng and McLachlan (2003) for an application to inference of Gaussian mixture models). Let us cite also the work by Nowlan (1991) which

proposes to pick at random one example per iteration in the large data set - but the update mechanism is quite difficult to recast in the EM framework; and the work by Thiesson et al. (2001) which proposes to regularly identify the significant examples and proceed with these examples only, for few iterations.

The *online EM* algorithm, proposed by Cappé and Moulines (2009), can be easily adapted to the framework of an incremental processing of a large data set even if originally, is was designed to process a stream of data. It is derived in the case the function G is of the form $\exp(\langle S(z), \phi(\theta)\rangle)$, which in the statistical context when F is the negative log-likelihood of the observations, means that the complete likelihood G is from the exponential family. In that case, as explained above, the E step of EM is equivalent to the construction of expected sufficient statistics $\bar{s}(\theta_{\mathrm{curr}})$; when this integration is intractable, Delyon et al. (1999) proved that the successive E steps can be replaced with a SA procedure targeting the roots of a *mean field* h. Exploiting the parallel between SA and gradient descent algorithms, *online EM* mimes what *Stochastic Gradient Descent* (see e.g. Bottou and Le Cun (2004)) is for finite-sum optimization. Going further in this parallel, Chen et al. (2018) and Karimi et al. (2019c) proposed resp. *Stochastic EM with Variance Reduction (sEM-vr)* and *Fast Incremental EM (FIEM)* as variance reduction techniques within *online EM* as an echo to *Stochastic Variance Reduced Gradient* (SVRG, Johnson and Zhang (2013)) and *Stochastic Averaged Gradient* (SAGA, Defazio et al. (2014)) introduced as variance reduction techniques within *Stochastic Gradient Descent*.

In this paper, we aim to study such incremental EM methods combined with a Stochastic Approximation approach. The first goal of this paper is to cast *online EM*, *incremental EM* and *FIEM* into a framework called hereafter *Stochastic Approximation within EM* approaches; see subsection 2.3. We show that the E step of FIEM can be seen as the combination of a SA update and of a control variate; we propose to optimize the balance between these two quantities yielding to the *optimized FIEM* algorithm; this new algorithm is numerically explored in section 4.

The second and main objective of this paper, is to derive non asymptotic convergence bounds for FIEM (see section 3). In the non-convex setting, finding a point $\hat{\theta}^{\epsilon}$ such that $F(\hat{\theta}^{\epsilon}) - \min F \leq \epsilon$ is NP-hard (see Murty and Kabadi (1987)); hence, in non-convex deterministic optimization of a smooth function F, convergence is often characterized by the quantity $\inf_{1\leq k\leq K_{\max}} \|\nabla F(\theta^k)\|$ along a path of length $K_{\max}$; in non-convex stochastic optimization, this criterion gets into $\inf_{1\leq k\leq K_{\max}} \mathbb{E}\left[\|\nabla F(\theta^k)\|^2\right]$ when the expectation is w.r.t. the randomness introduced to replace intractable quantities with oracles. Nevertheless, in many frameworks such as the finite-sum optimization one we are interested in, such a criterion can not be used to define a termination rule for the algorithm since ∇F is intractable. Adopting the idea of Ghadimi and Lan (2013) (see also Allen-Zhu and Hazan (2016), Reddi et al. (2016), Fang et al. (2018), Zhou et al. (2018) and

Karimi et al. (2019c)), we propose to fix a maximal length $K_{\max}$ and terminate a path of the algorithm at some random time K uniformly sampled in the range $\{0, \dots, K_{\max} - 1\}$ prior the run and independently of it; our convergence bounds control $\mathbb{E}\left[\|\nabla F(\theta^K)\|^2\right]$ and as a corollary, we discuss how to fix $K_{\max}$ as a function of the sample size n in order to reach an ϵ-approximate stationary point i.e.to find $\hat{\theta}^{K,\epsilon}$ such that $\mathbb{E}\left[\|\nabla F(\hat{\theta}^{K,\epsilon})\|^2\right] \leq \epsilon$. Such a property is sometimes called ϵ-accuracy in expectation (see e.g. (Reddi et al., 2016, Definition 1)).

The first convergence analyses of incremental EM methods concerned their asymptotic behavior such as almost-sure convergence of the functional along the path, almost-sure convergence of the iterates to a stationary point or rate of convergence when starting from a neighborhood of such a point: such results can be found e.g. in Gunawardana and Byrne (2005) for incremental EM with a deterministic scan of the blocks; in Karimi et al. (2019a) for incremental EM with a random selection of the examples at each iteration; in Srivastava et al. (2019) for a distributed version of the incremental EM; in Nguyen et al. (2020) for mini-batch online EM applied to inference in a mixture of exponential models; in Chen et al. (2018) for sEM-vr, under convexity assumptions. Some of the analyses also rely on asymptotic convergence of MM algorithms (Neal and Hinton (1998), Mairal (2015)).

For non asymptotic results, let us cite the contributions by Karimi et al. (2019c) which prove that *incremental EM*, which picks at random one example per iteration, reaches ϵ-accuracy by choosing $K_{\max} = O(n\epsilon^{-1})$: even if the algorithm is terminated at a random time K, this random time is chosen as a function of $K_{\max}$ which has to increase linearly with the size n of the data set. Karimi et al. (2019b) and Karimi et al. (2019c) provide the same analysis for *online EM* and *FIEM* showing that for both methods, ϵ-approximate stationarity is reached with $K_{\max} = O(n^{2/3}\epsilon^{-1})$ - here again, with one example picked at random per iteration. For these reasons, *online EM* and *FIEM* are preferable especially when n is large (see section 4 for a numerical illustration). Our major contribution in this paper is to show that for *FIEM*, the rate depends on the choice of some design parameters. By choosing a constant step size sequence in the SA step, depending upon n as $O(n^{-2/3})$, then ϵ-accuracy requires $K_{\max} = O(n^{2/3}\epsilon^{-1})$; we therefore retrieve the conclusions of Karimi et al. (2019c) but provide a value of the step size and a value of the convergence bounds which improve on Karimi et al. (2019c) - as numerically illustrated in section 4. We then prove that such an ϵ-accuracy is possible with $K_{\max} = O(\sqrt{n}\epsilon^{-3/2})$ and another strategy for the definition of the step size. To our best knowledge, this second result is new: it provides a weaker dependence on n but a larger dependence on the tolerance ϵ; the second approach is preferable for small to medium accuracy ϵ w.r.t. the size of the data set n. Our third contribution relative to non asymptotic convergence bounds is the converse approach: given a sampling distribution for the random termination time K, how

to choose the stepsize sequence in the SA step in order to reach ϵ-approximate stationarity with $K_{\max} = O(n^{2/3}\epsilon^{-1})$?

We conclude this introduction by discussing the choice of an EM approach for solving the optimization problem at hand w.r.t. other strategies like Stochastic Gradient approaches. As already pointed out by several discussants of the original paper of Dempster et al. (1977), there are no reasons why EM should be systematically preferred to other strategies, and conversely. The Fisher relation

$$\nabla F(\theta) = -\frac{1}{n}\mathbb{E}_\theta \left[\partial_\theta \log G(Z,\theta)\right]$$

where $Z \sim G(\cdot,\theta)\exp(n\,F(\theta))\mathrm{d}\mu_n$ under $\mathbb{E}_\theta$, shows that in the case given by (2) when $\log G(z,\theta) = \sum_{i=1}^n \langle s_i(z_i), \phi(\theta)\rangle$, we have $\nabla F(\theta) = n^{-1}\sum_{i=1}^n \nabla f_i(\theta)$ with $\nabla f_i(\theta) = \left(\dot\phi(\theta)\right)^T \bar s_i(\theta)$. Therefore, both EM and Gradient Descent require the computation of the full sum $\bar s = n^{-1}\sum_{i=1}^n \bar s_i$. A comparison of EM and its interpretation as a quasi-Newton method and more generally as a descent method, is investigated e.g. in Springer and Urban (2014). Xu and Jordan (1996) is among the many contributions discussing the advantages and disadvantages of EM for a specific application (here, application to maximum likelihood learning of Gaussian mixtures); see also Fort et al. (2019) for a discussion on SA within EM and gradient descent algorithms for solving problems of the form (1), possibly in the non smooth case.

A review on incremental gradient based approaches for the minimization of a sum of convex functions, in the case of a deterministic cyclic order and a randomized selection of the functions as well, can be found in Bertsekas (2011). In the deterministic setting, (Nesterov, 2004, p.29) proves that for a gradient descent algorithm targeting a stationary point of a function F with Lipschitz-continuous first derivative, it holds: $\min_{1\le k\le K_{\max}}\|\nabla F(\theta^k)\| \le \epsilon$ with $K_{\max} = O(\epsilon^{-2})$; Cartis et al. (2010) shows that this bound is tight. Convergence bounds for accelerated (deterministic) gradient procedures in the non-convex case are obtained by Carmon et al. (2018): a point $\theta^\star$ satisfying $\|\nabla F(\theta^\star)\| \le \epsilon$ can be reached with $O(\epsilon^{-7/4}\log(1/\epsilon))$ iterations under the assumptions that F possesses Lipschitz-continuous first and second derivatives.

The Stochastic Gradient Descent (SGD) approach consists in picking at random one example $\#i$ in $\{1,\dots,n\}$ at each iteration, and substitute the full gradient $\nabla F(\theta_{\mathrm{curr}})$ for $\nabla f_i(\theta_{\mathrm{curr}})$. Non asymptotic convergence rates are derived by Ghadimi and Lan (2013) in the non-convex setting, for finding a root of a Lipschitz function when only stochastic oracles are available: it is proved that with a random termination rule K upper bounded by $K_{\max}$ and a constant step size scaling as $O(1/\sqrt{K_{\max}})$, ϵ-approximate stationarity is reached with $K_{\max} = O(\epsilon^{-2})$, a complexity which is higher than what is proved for the incremental EM algorithms mentioned above. A slight different analysis is provided in Ghadimi et al. (2016)

and Wang et al. (2017) resp. when the optimization is also constrained and when the objective function is not smooth: given a number N of calls to the stochastic oracle, the accuracy is $O(N^{-1/2})$. In the terminology used in this paper, a call to the stochastic oracle corresponds to pick at random one example in the data set; and this result states that ϵ-accuracy is reached after ϵ^{-2} calls.

Many recent works combine Stochastic Gradient Descent algorithm and variance reduction techniques to design algorithms for non-convex finite-sum optimization of a smooth objective function: it is proved that ϵ-approximate stationarity is reached with a number of calls to one of the gradients ∇f_i given by $O(n^{2/3}\epsilon^{-1})$ for SAGA adapted from *Stochastic Average Gradient* by Reddi et al. (2016), and for an algorithm adapted from *Stochastic Variance Reduced Gradient* by Allen-Zhu and Hazan (2016)); $O(n^{2/3}\epsilon^{-1} \wedge \epsilon^{-5/3})$ for *Stochastically Controlled Stochastic Gradient* proposed by Lei et al. (2017); $O(\sqrt{n}\epsilon^{-1})$ for *Stochastic Path-Integrated Differential Estimator* proposed by Fang et al. (2018), and $O(\sqrt{n}\epsilon^{-1} \wedge \epsilon^{-3/2})$ for *Stochastic Nested Variance Reduced Gradient descent* proposed by Zhou et al. (2018). ERIC: on met ce qui suit ? c'est pour éviter une demande du genre "et que donnent ces idées appliquées à SA-EM ? Works in progress by authors of this paper, consist in analyzing SA within EM approaches miming the ideas introduced by Fang et al. (2018) and Zhou et al. (2018).

2 Stochastic Approximation within EM algorithms for non-convex optimization

Notations. $\langle a, b\rangle$ denotes the standard Euclidean scalar product on $\mathbb{R}^\ell$, for $\ell \geq 1$; and $\|a\|$ the associated norm. For a matrix A, A^T is its transpose. For a smooth function ϕ, $\dot{\phi}$ denotes its gradient.

2.1 A non-convex finite-sum optimization problem

This paper deals with EM-based algorithms designed to solve the optimization problem

$$\text{Argmin}_{\theta\in\Theta} F(\theta), \qquad F(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + \mathsf{R}(\theta) \,, \tag{3}$$

where

$$\mathcal{L}_i(\theta) \stackrel{\text{def}}{=} -\log \int_{\mathsf{Z}} h_i(z) \, \exp\left(\langle s_i(z), \phi(\theta)\rangle\right) \, \mu(\mathrm{d}z) \,, \tag{4}$$

and

H1 *$\Theta \subseteq \mathbb{R}^d$ is a measurable convex subset. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\mathsf{R} : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathbb{R}^q$ and $h_i : \mathsf{Z} \to \mathbb{R}_+$, $s_i : \mathsf{Z} \to \mathbb{R}^q$ for $i \in \{1, \ldots, n\}$ are measurable functions. Finally, for any $\theta \in \Theta$ and $i \in \{1, \ldots, n\}$, $-\infty < \mathcal{L}_i(\theta) < \infty$.*

The Eq. (4) and the assumption H1 claim that $\exp(-\mathcal{L}_i)$ is positive and defined as the integral of a non-negative function which is separable with respect to the integration variable z and the variable θ. Under H1, for any $\theta \in \Theta$ and $i \in \{1, \ldots, n\}$, the quantity $p_i(z;\theta)\,\mu(\mathrm{d}z)$ where

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

defines a probability distribution on $\mathcal{Z}$. We will assume that

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 and is computationally tractable.

Define

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

The framework defined by (3) and (4) covers many computational learning problems such as empirical risk minimization with non-convex losses: R may include a regularization condition on the parameter θ, $n^{-1}\sum_{i=1}^{n}\mathcal{L}_i$ is the empirical loss and $\mathcal{L}_i$ is the loss function associated to example $\#i$.

Among applications concerned with the form (4) of the loss function, let us cite the normalized negative log-likelihood in a latent variable models (see e.g. Little and Rubin (2002)), when the complete data likelihood is from a multi dimensional exponential family, not necessarily canonical (see e.g. Brown (1986); Sundberg (2019) for properties of exponential families): the additive form of the global loss $n^{-1}\sum_{i=1}^{n}\mathcal{L}_i(\theta)$ is the consequence of an independence assumption on the n observations; in such models, the likelihood of the observation $\#i$ is of the form

$$y \mapsto \int_{\mathsf{Z}} h(y,z)\ \exp(\langle T(y,z), \phi(\theta)\rangle - \psi(\theta))\ \mathrm{d}\mu(z)$$

which corresponds to (4) by setting $h_i(z) \stackrel{\text{def}}{=} h(y,z)$ and $s_i(z) \stackrel{\text{def}}{=} T(y,z)$. The normalizing constant $\exp(-\psi(\theta))$ can be part of the term $\mathsf{R}(\theta)$ in (3).

2.2 A Majorize-Minimization approach based on EM

Given $\theta' \in \Theta$, define the function $\overline{F}(\cdot,\theta') : \Theta \to \mathbb{R}$ by

$$\overline{F}(\theta,\theta') \stackrel{\text{def}}{=} -\left\langle \bar{s}(\theta'), \phi(\theta)\right\rangle + \mathsf{R}(\theta) + \frac{1}{n}\sum_{i=1}^{n}\mathcal{C}_i(\theta')\ ,$$

$$\mathcal{C}_i(\theta') \stackrel{\text{def}}{=} \mathcal{L}_i(\theta') + \left\langle \bar{s}_i(\theta'), \phi(\theta')\right\rangle\ .$$

The following result shows that $\{\overline{F}(\cdot,\theta'), \theta' \in \Theta\}$ is a family of majorizing function of the objective function F from which a Majorize-Minimization approach for solving (3) can be derived under the following assumption:

H3 *For any* $s \in \mathbb{R}^q$, $\operatorname{Argmin}_{\theta\in\Theta}\ (-\langle s, \phi(\theta)\rangle + \mathsf{R}(\theta))$ *exists and is unique. It is denoted by* $\mathsf{T}(s)$.

When $\theta \mapsto -\langle s, \phi(\theta)\rangle + \mathsf{R}(\theta)$ is continuous and either Θ is compact or the function is coercive[1] on Θ, then $\mathsf{T}(s)$ exists.

Proposition 1 *Assume H1 and H2.*

1. *For any* $i \in \{1, \cdots, n\}$ *and* $\theta' \in \Theta$, $\mathcal{L}_i(\cdot) \le -\left\langle \bar{s}_i(\theta'), \phi(\cdot)\right\rangle + \mathcal{C}_i(\theta')$.
2. *For any* $\theta' \in \Theta$, $F \le \overline{F}(\cdot, \theta')$, *and* $\overline{F}(\theta', \theta') = F(\theta')$.
3. *Assume also H3. Given* $\theta^0 \in \Theta$, *the sequence defined by* $\theta^{k+1} \stackrel{\text{def}}{=} \mathsf{T} \circ \bar{s}(\theta^k)$ *for any* $k \ge 0$, *satisfies* $F(\theta^{k+1}) \le F(\theta^k)$.

The proof, while relying on standard tools, is provided in subsubsection 6.1.1.

The algorithm described by item 3 of Proposition 1, is the EM algorithm: upon noting that (3) is equivalent to the maximization of

$$\theta \mapsto \log \int_{\mathsf{Z}^n} \left(\prod_{i=1}^n h_i(z_i)\right) \exp\left(\left\langle \sum_{i=1}^n s_i(z_i), \phi(\theta)\right\rangle\right) \mu(\mathrm{d}z_1)\dots\mu(\mathrm{d}z_n) - n\mathsf{R}(\theta)\ ,$$

the E step of the EM algorithm would compute the auxiliary quantity

$$\begin{aligned}\mathsf{Q}(\theta, \tau^k) &\stackrel{\text{def}}{=} \int_{\mathsf{Z}^n} \left\langle \sum_{i=1}^n s_i(z_i), \phi(\theta)\right\rangle \prod_{i=1}^n p_i(z_i; \tau^k)\, \mu(\mathrm{d}z_i) - n\mathsf{R}(\theta) \\ &= n\left\langle \bar{s}(\tau^k), \phi(\theta)\right\rangle - n\mathsf{R}(\theta)\ ,\end{aligned}$$

given the current parameter τ^k; and then the M step would update the parameter by setting $\tau^{k+1} \in \operatorname{Argmax}_{\theta\in\Theta} \mathsf{Q}(\theta, \tau^k)$. It is easily seen that this mechanism is equal to $\tau^{k+1} = \mathsf{T} \circ \bar{s}(\tau^k)$.

The map T defined in H3 is therefore the maximization map of the M step in EM. We assume that T is explicit even if the optimization may be constrained; in some applications, the optimization is intractable and many strategies are proposed in the EM literature (see e.g. (McLachlan and Krishnan, 2008, Chapter 5) and references in Nettleton (1999); Takai (2012); Zhu et al. (2017)). It is not the purpose of this paper to cope with the optimization step.

The assumption that T is defined for any $s \in \mathbb{R}^q$ may be restrictive for some applications. When deriving theoretical analysis of EM-based algorithms, it is sometimes assumed that T is defined on a convex subset (and sometimes also compact) $\mathcal{S} \subseteq \mathbb{R}^q$ (see e.g. (Delyon et al., 1999, Assumption M5), (Fort and Moulines, 2003, Assumption M2), (Kuhn and Lavielle, 2004, Theorem 1), (Cappé and Moulines, 2009, Assumption 1), (Allassonnière et al., 2010, Theorem 1), (Le Corff and Fort,

[1] for any $A > 0$, there exist $\rho, B > 0$ such that if $\theta \in \Theta$ and either $\|\theta\| \ge B$ or $\mathrm{d}(\theta, \Theta^c) \le \rho$, then $|-\langle s, \phi(\theta)\rangle + \mathsf{R}(\theta)| \ge A$.

2013b, Section 4.1), (Karimi et al., 2019c, Assumption H4)). While in many applications, it is difficult to prove that the argument of T remains in $\mathcal{S}$ - and often is not even discussed - it is often observed that a smart implementation - such as a convenient initialization - may make the conditions to be satisfied numerically (see e.g. Donnet and Samson (2007), Cappé and Moulines (2009)). For the theoretical results derived hereafter, we assume H3 which is the easiest way to cover the EM-based algorithms studied here; it is out of the scope of this paper to address a more general case.

Starting from the current point θ^k, the iterative scheme $\theta^{k+1} = \mathsf{T} \circ \bar{s}(\theta^k)$ first computes a point in $\bar{s}(\Theta)$ through the expectation $\bar{s}$, and then apply the map T to obtain the new iterate θ^{k+1}. It can therefore be described in the $\bar{s}(\Theta)$-space, a space sometimes called the *expectation space*, being equivalently defined as follows: define $\{\bar{s}^k,\ k \in \mathbb{N}\}$ by $\bar{s}^0 \in \mathbb{R}^q$ and for any $k \geq 0$, and is

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

This approach in the expectation space comes up more naturally in the derivation of incremental algorithms designed for the large scale learning setting; it will be adopted throughout this paper.

Before deriving incremental EM-based methods, we conclude this section by a discussion on the limiting points of the iterative method (6). Sufficient conditions for the characterization of the limit points of any instance $\{\bar{s}^k, k \geq 0\}$ as the critical points of $F \circ \mathsf{T}$, for the convergence of the functional along the sequence $\{F \circ \mathsf{T}(\bar{s}^k), k \geq 0\}$, or for the convergence of the iterates $\{\bar{s}^k, k \geq 0\}$ towards the critical points of $F \circ \mathsf{T}$ exist in the literature (see e.g. Wu (1983); Lange (1995); Delyon et al. (1999) in the EM context and Zangwill (1967); Csiszár and Tusnády (1984); Gunawardana and Byrne (2005); Parisi et al. (2019) for general iterative Majorize-Minimization algorithms). Proposition 2 characterizes the fixed points of $\mathsf{T} \circ \bar{s}$ and of $\bar{s} \circ \mathsf{T}$ under a set of conditions which will be adopted for the convergence analysis in Section 3.

H4 *1. The functions ϕ and R are continuously differentiable on Θ^v where $\Theta^v \stackrel{\text{def}}{=} \Theta$ if Θ is open, or Θ^v is a neighborhood of Θ otherwise. T is continuously differentiable on $\mathbb{R}^q$.*

2. The function F is continuously differentiable on Θ^v and for any $\theta \in \Theta$,

$$\dot{F}(\theta) = -\left(\dot{\phi}(\theta)\right)^T \bar{s}(\theta) + \dot{\mathsf{R}}(\theta) \ .$$

3. For any $s \in \mathbb{R}^q$, $B(s) \stackrel{\text{def}}{=} (\phi \stackrel{\cdot}{\circ} \mathsf{T})(s)$ is a symmetric $q \times q$ matrix with positive minimal eigenvalue.

Under H1 to H4-item 1 and the assumption that Θ and $\phi(\Theta)$ are open subsets of resp. $\mathbb{R}^d$ and $\mathbb{R}^q$, then Lemma 1 in subsubsection 6.1.3 shows that H4-item 2 holds and the functions $\mathcal{L}_i$ are continuously differentiable on Θ for all $i \in \{1, \ldots, n\}$.

Under H1, H3 and the assumptions that *(i)* T is continuously differentiable on $\mathbb{R}^q$ and *(ii)* for any $s \in \mathbb{R}^q$, $\tau \mapsto \mathsf{Q}(s,\tau) \stackrel{\text{def}}{=} -\langle s, \phi(\tau)\rangle + \mathsf{R}(\tau)$ is twice continuously differentiable on Θ^v (defined in H4-item 1), then for any $s \in \mathbb{R}^q$, $\partial^2_\tau \mathsf{Q}(s, \mathsf{T}(s))$ is positive-definite and

$$B(s) = \left(\dot{\mathsf{T}}(s)\right)^T \ \partial^2_\tau \mathsf{Q}(s, \mathsf{T}(s)) \ \left(\dot{\mathsf{T}}(s)\right) \ ;$$

see Lemma 2 in subsubsection 6.1.3. Therefore, $B(s)$ is a symmetric matrix and if $\text{rank}(\dot{\mathsf{T}}(s)) = q = q \wedge d$, its minimal eigenvalue is positive.

Proposition 2 *Assume H1, H2 and H3. Define the measurable functions* $V : \mathbb{R}^q \to \mathbb{R}$ *and* $h : \mathbb{R}^q \to \mathbb{R}^q$ *by*

$$V(s) \stackrel{\text{def}}{=} F \circ \mathsf{T}(s) \ , \qquad h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s \ .$$

1. *If* $s^\star$ *is a fixed point of* $\bar{s} \circ \mathsf{T}$*, then* $\theta^\star \stackrel{\text{def}}{=} \mathsf{T}(s^\star)$ *is a fixed point of* $\mathsf{T} \circ \bar{s}$*. Conversely, if* $\theta^\star$ *is a fixed point of* $\mathsf{T} \circ \bar{s}$ *then* $s^\star \stackrel{\text{def}}{=} \bar{s}(\theta^\star)$ *is a fixed point of* $\bar{s} \circ \mathsf{T}$*.*
2. *Assume in addition H4. Then for all* $s \in \mathbb{R}^q$*,* $\dot{V}(s) = -B(s)\, h(s)$*; and the zeros of* h *are the critical points of* V*.*

The proof is in subsubsection 6.1.2. As a conclusion, the EM algorithm summarized in algorithm 1, is designed to converge to the zeros of

$$s \mapsto h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s \ , \tag{7}$$

which, for some models, are the critical points of $F \circ \mathsf{T}$.

However, the computational cost per iteration of EM is proportional to the number n of examples, since it requires the computation of $\bar{s}$ i.e. a sum over n terms. It is therefore intractable in the large scale learning framework. We review in subsection 2.3 few alternatives based on an incremental approach, and proposed in the literature to overcome this intractability.

Data: $K_{\max} \in \mathbb{N}$, $\bar{s}^0 \in \mathbb{R}^q$
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: The EM algorithm in the expectation space

2.3 Stochastic Approximation within EM approaches

It was proposed in Delyon et al. (1999) to overcome the intractability of the expectation $\bar{s}$ by substituting the induction $\bar{s}^{k+1} = \bar{s} \circ \mathsf{T}(\bar{s}^k)$ with the definition of a random sequence $\{\widehat{S}^k, k \geq 0\}$ satisfying

$$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} \left(s^{k+1} - \widehat{S}^k \right) , \tag{8}$$

where $\{\gamma_k, k \geq 1\}$ is a $[0,1]$-valued deterministic positive sequence of *step sizes* (also called *learning rates*) chosen by the user and s^{k+1} is a Monte Carlo approximation of $\bar{s} \circ \mathsf{T}(\widehat{S}^k)$; the definition of s^{k+1} is such that the updating rule (8) is a Stochastic Approximation (SA) algorithm designed to target the zeros of the mean field $h(s)$ (see (7)); see e.g. Benveniste et al. (1990); Borkar (2008) for a general review on SA. Many stochastic perturbations of EM can be described by (8): let us cite for example the Stochastic EM Celeux and Diebolt (1985), or the Monte Carlo EM (introduced by Wei and Tanner (1990) and studied by Fort and Moulines (2003)) which corresponds to $\gamma_{k+1} = 1$.

We review below some stochastic perturbations of EM, recently introduced to overcome the large scale learning intractability. The originality of this review is to recast these algorithms in a SA framework. For this purpose, the key observation is the equality

$$h(s) = \mathbb{E}\left[\bar{s}_I \circ \mathsf{T}(s) - s + V\right] \tag{9}$$

where I is a uniform random variable on $\{1, \ldots, n\}$ and V is a centered variable. Such an expression gives insights for the definition of SA schemes, including the combination with a variance reduction techniques through an adequate choice of V (see e.g. (Glasserman, 2004, Section 4.1.) for an introduction to control variates). Since I is finitely sampled, the mean field h is from the finite-sum family of functions.

2.3.1 Online EM

A first natural idea is given by algorithm 2: at iteration $(k+1)$, s^{k+1} is obtained by sampling at random an unique example (say, example $\#I_{k+1}$) among the available n, and by computing the expectation $\bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$. Each iteration only requires the computation of one expectation $\bar{s}_i$. This algorithm, hereafter called `SA`, corresponds to a SA scheme: $\widehat{S}^{k+1} - \widehat{S}^k = \gamma_{k+1} H_{k+1}$ with $H_{k+1} \stackrel{\text{def}}{=} \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k$ satisfying $\mathbb{E}\left[H_{k+1} | \mathcal{F}_k\right] = h(\widehat{S}^k)$; the filtration $\mathcal{F}_k$ is defined by $\mathcal{F}_k \stackrel{\text{def}}{=} \sigma(\widehat{S}^0, I_1, \cdots, I_k)$. It is a natural extension of the `online EM` by Cappé and Moulines (2009) which is designed to process data streams i.e. situations when data appears one at a time, are used once in order to incorporate the information brought by a new observation, and then are no longer available.

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathbb{R}^q$, $\gamma_k \in (0, \infty)$ for $k = 1, \ldots, K_{\max}$
Result: The SA sequence: $\widehat{S}^k, k = 0, \ldots, K_{\max}$
1 **for** $k = 0, \ldots, K_{\max} - 1$ **do**
2 Sample I_{k+1} uniformly on $\{1, \ldots, n\}$;
3 $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} \left(\bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k \right)$.

Algorithm 2: The SA algorithm

Different variants were proposed: instead of sampling a single observation among a batch of size n (or incorporating a single new observation in a data stream), a mini-batch of examples can be used: line 3 would get into

$$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} \left(N^{-1} \sum_{i \in \mathcal{I}_{k+1}} \bar{s}_i \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k \right)$$

where $\mathcal{I}_{k+1}$ is a set of integers of cardinal N, sampled uniformly and with replacement in $\{1, \ldots, n\}$. Convergence of the iterates in the long-time behavior ($K_{\max} \to \infty$) for `online EM` is addressed in Cappé and Moulines (2009); similar convergence results in the mini-batch case for the ML estimation of exponential family mixture models were recently established by Nguyen et al. (2020). Karimi et al. (2019a) and Kuhn et al. (2019) also proposed an asymptotic convergence result, in the mini-batch case for the ML estimation in latent variable models from the exponential family, combined with a Monte Carlo approximation of the expectations $\bar{s}_i$. Finally, non asymptotic error rates are derived in Karimi et al. (2019b).

Note that an algorithm close to `SA` is proposed in Nowlan (1991): it corresponds to an update of the statistic of the form $\widehat{S}^{k+1} = \gamma \widehat{S}^k + \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$. A convenient choice of γ seems to favor an exponential forgetting of out-of-date statistics; and convergence to the same limiting value of EM is observed - when convergence is observed which is not guaranteed.

2.3.2 The incremental EM algorithm

The `Incremental EM (iEM)` algorithm is described by algorithm 3. Lines 4 to 7 are a recursive computation of

$$\widetilde{S}^{k+1} = n^{-1} \sum_{i=1}^{n} \mathsf{S}_{k+1,i} \ , \tag{10}$$

where for $k \geq 0$,

$$\mathsf{S}_{k+1,i} \stackrel{\text{def}}{=} \begin{cases} \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k) & \text{if } i = I_{k+1} \ , \\ \mathsf{S}_{k,i} & \text{otherwise} \ . \end{cases} \tag{11}$$

Data: $K_{\max} \in \mathbb{N}$, $\widehat{S}^0 \in \mathbb{R}^q$, $\gamma_k \in (0,\infty)$ for $k = 1, \ldots, K_{\max}$
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 $i = 1, \ldots, n$;
2 $\widetilde{S}^0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;
3 **for** $k = 0, \ldots, K_{\max} - 1$ **do**
4 $\quad I_{k+1} \sim \mathcal{U}(\{1, \ldots, n\})$;
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 $\quad \widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} (\widetilde{S}^{k+1} - \widehat{S}^k)$

Algorithm 3: The iEM algorithm

It avoids the implementation of a sum over n terms. The sequence $\{\widehat{S}^k, k \geq 0\}$ is not a SA sequence but the bivariate sequence $\{(\widehat{S}^k, \mathsf{S}_{k,\cdot}), k \geq 0\}$ is: we have $\mathsf{S}_{k+1,\cdot} - \mathsf{S}_{k,\cdot} = n^{-1} H_{k+1}^{(1)}$ and $\widehat{S}^{k+1} - \widehat{S}^k = \gamma_{k+1} H_{k+1}^{(2)}$ where

$$\mathbb{E}\left[H_{k+1}^{(1)} | \mathcal{F}_k \right] = \begin{bmatrix} \bar{s}_1 \circ \mathsf{T}(\widehat{S}^k) \\ \cdots \\ \bar{s}_n \circ \mathsf{T}(\widehat{S}^k) \end{bmatrix} - \mathsf{S}_{k,\cdot} \ ,$$

$$\mathbb{E}\left[H_{k+1}^{(2)} | \mathcal{F}_k \right] = h(\widehat{S}^k) + (n^{-1} - 1) \left(\bar{s} \circ \mathsf{T}(\widehat{S}^k) - \widetilde{S}^k \right) \ ;$$

here again, $\mathcal{F}_k \stackrel{\text{def}}{=} \sigma(\widehat{S}^0, I_1, \ldots, I_k)$. If there exists $(s^\star, \mathsf{S}_{\star,\cdot})$ such that $\lim_k (\widehat{S}^k, \mathsf{S}_{k,\cdot}) = (s^\star, \mathsf{S}_{\star,\cdot})$, then it may be seen (with no rigorous proof) that it satisfies $n^{-1} \sum_{i=1}^n \mathsf{S}_{\star,i} = \bar{s} \circ \mathsf{T}(s^\star) = s^\star$. This observation and the following equality obtained from lines 5 to 8 of algorithm 3

$$\widehat{S}^{k+1} = \widehat{S}^k + \frac{\gamma_{k+1}}{n} \left\{ \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \widetilde{S}^k - \mathsf{S}_{k,I_{k+1}} + (n-1) \left(\widetilde{S}^k - \widehat{S}^k \right) \right\} \ ,$$

show that *(i)* the update mechanism for $\{\widehat{S}^k, k \geq 0\}$ is of the form (9) with a random variable V correlated to I whose conditional expectation is $(n-1) \left(\widetilde{S}^k - \widehat{S}^k \right)$; *(ii)* if convergence holds, a convergence of $\{\widehat{S}^k, k \geq 0\}$ to a fixed point of h is expected and the conditional expectation of V vanishes to zero.

algorithm 3 generalizes the original incremental EM proposed by Neal and Hinton (1998), which corresponds to the case $\gamma_{k+1} = 1$ and to a deterministic visit to the successive examples. With $\gamma_{k+1} = 1$ for any $k \geq 0$, we have $\widehat{S}^k = \widetilde{S}^k = n^{-1} \sum_{i=1}^n \mathsf{S}_{k,i}$. algorithm 3 can be adapted in order to use a mini-batch of examples per iteration: the data set is divided into B blocks prior running `iEM`: per iteration, the examples of only a block are processed for the update of $\widehat{S}^k$ (see line 6) and along iterations, either the blocks are visited in turn or they are chosen randomly through a mechanism possibly depending on the fluctuations of the current iterate. The efficiency of `iEM` is therefore a trade-off between the size

of the block which is related to the computational cost of a full scan of the data, and the fewer number of total scans required for convergence since `iEM` exploits information more quickly. Ng and McLachlan (2003) provide a numerical analysis of the role of B when `iEM` is applied to fitting a normal mixture model with fixed number of components. Gunawardana and Byrne (2005) provides sufficient conditions for the convergence to stationary points of F in the case the data set is processed through B blocks visited according to a deterministic cycling.

When $\gamma_{k+1} = 1$ for any $k \geq 0$ and the examples are chosen randomly at each iteration, the sequence $\{\widehat{S}^k, k \geq 0\}$ is the same as the one given by a Majorize-Minorization algorithm based on the inequality, at iteration $(k+1)$

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

where $\theta^{k,i} \stackrel{\text{def}}{=} \mathsf{T}(\widehat{S}^k)$ if $i = I_{k+1}$ and $\theta^{k,i} \stackrel{\text{def}}{=} \theta^{k-1,i}$ otherwise (see subsection 2.2 and Proposition 1 for the definition of $\mathcal{C}_i$ and the properties of these surrogate functions). This point of view and its link with *Minimization by Incremental Surrogate Optimization* (MISO, introduced by Mairal (2015)) is observed by Karimi et al. (2019c). Sufficient conditions for the asymptotic convergence of the functionals and of the iterates, in the non-convex case, can be deduced from the convergence analysis of MISO (Mairal, 2015, Proposition 2.5); it is also addressed in Karimi et al. (2019a). Karimi et al. (2019c) provide non asymptotic convergence rates for `iEM` (case $\gamma_{k+1} = 1$). Asymptotic convergence analysis of `iEM` (with $\gamma_{k+1} = 1$) is also addressed by Srivastava et al. (2019) in the case of a asynchronous distributed implementation of the algorithm.

In algorithm 3, the computational cost of each iteration is the draw of one example $\#I_{k+1}$, the computation of the expectation $\bar{s}_{I_{k+1}}(\widehat{S}^k)$ and the update (and storage) of an auxiliary quantity $\mathsf{S}_{k+1,\cdot} \in \mathbb{R}^{qn}$; the initialization step also requires the computation of a sum over the n examples. Observe that the $K_{\max}$ integers I_k can be sampled before running the algorithm, so the space cost for the storage of $\mathsf{S}_{k,\cdot}$ can be reduced to $q(n \wedge K_{\max})$.

2.3.3 The Fast Incremental EM algorithm

The `Fast Incremental EM (FIEM)`, introduced by Karimi et al. (2019c), is defined by algorithm 4. Each iteration selects two examples independently, say $\#I_{k+1}$ and $\#J_{k+1}$, and computes the expectations $\bar{s}_{I_{k+1}}(\widehat{S}^k)$ and $\bar{s}_{J_{k+1}}(\widehat{S}^k)$; as in `iEM`, `FIEM` computes the sum $\widetilde{S}^{k+1} = n^{-1} \sum_{i=1}^{n} \mathsf{S}_{k+1,i}$ (see (11)) in a recursive way, avoiding a sum over n terms at each iteration (see line 4 to line 7 of algorithm 4). Then, this auxiliary quantity is used in the update mechanism of the sequence (see line 9). Here again, the sequence $\{\widehat{S}^k, k \geq 0\}$ is not a SA sequence, but $\{(\widehat{S}^k, \mathsf{S}_{k,\cdot}), k \geq 0\}$

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

Algorithm 4: The Fast Incremental EM (FIEM) algorithm

is. We have $\mathsf{S}_{k+1,\cdot} - \mathsf{S}_{k,\cdot} = n^{-1} H_{k+1}^{(1)}$ and $\widehat{S}^{k+1} - \widehat{S}^k = \gamma_{k+1} H_{k+1}^{(2)}$ where

$$\mathbb{E}\left[H_{k+1}^{(1)} | \mathcal{F}_k \right] = \begin{bmatrix} \bar{s}_1 \circ \mathsf{T}(\widehat{S}^k) \\ \cdots \\ \bar{s}_n \circ \mathsf{T}(\widehat{S}^k) \end{bmatrix} - \mathsf{S}_{k,\cdot} \; ,$$

$$\mathbb{E}\left[H_{k+1}^{(2)} | \mathcal{F}_{k+1/2} \right] = h(\widehat{S}^k) \; ;$$

we set $\mathcal{F}_k \overset{\text{def}}{=} \sigma(\widehat{S}^0, I_1, J_1, \ldots, J_{k-1}, I_k, J_k)$ and $\mathcal{F}_{k+1/2} \overset{\text{def}}{=} \sigma(\mathcal{F}_k \cup \{I_{k+1}\})$. It is easily seen that if there exists $(s^\star, \mathsf{S}_{\star,\cdot})$ such that $\lim_k (\widehat{S}^k, \mathsf{S}_{k,\cdot}) = (s^\star, \mathsf{S}_{\star,\cdot})$, then it satisfies $n^{-1} \sum_{i=1}^n \mathsf{S}_{\star,i} = \bar{s} \circ \mathsf{T}(s^\star) = s^\star$. This observation shows that *(i)* the update mechanism for $\{\widehat{S}^k, k \geq 0\}$ is of the form (9) with a random variable V, conditionally centered, and correlated to I; *(ii)* if convergence holds, a convergence of $\{\widehat{S}^k, k \geq 0\}$ to a fixed point of h is expected.

The introduction of such a variable V can be seen as a variance reduction technique, inherited from the Stochastic Averaged Gradient (SAGA, by Defazio et al. (2014)) which was proposed to improve convergence properties of incremental stochastic gradient methods. A similar idea is developed in Chen et al. (2018), a paper which adapts Stochastic Variance Reduced Gradient (SVRG, Johnson and Zhang (2013)) to incremental EM algorithms: their motivation is to improve on `online EM` (see subsubsection 2.3.1) which surpasses EM in the burn-in phase but is penalized by the large variance when approximating the E step in the convergence phase. The SVRG approach was also plugged in a generalized EM by Zhu et al. (2017), but when performing the M step: they propose to replace an exact gradient by a stochastic gradient in order to avoid the processing of the full data set in the M step, and they combine this idea with a variance reduction method.

To our best knowledge, the convergence analyses of `FIEM` are only given in Karimi et al. (2019c): non asymptotic convergence rates for `FIEM` are derived. The

novel theoretical contribution of our paper, detailed in section 3, is to complement and improve on these results. Since `FIEM` is among the family of the *SA within EM* methods, its asymptotic behavior could be studied by miming the proofs in Delyon et al. (1999) - nevertheless, this analysis of `FIEM` is not published as far as we know.

On the computational side, each iteration of `FIEM` requires two samplings from $\{1,\dots,n\}$ and two computations of an expectation of the form $\bar{s}_i(\theta)$; as for `iEM`, there is a space complexity through the storage of the auxiliary quantity $\mathsf{S}_{k,\cdot}$ - its size being proportional to $q(2K_{\max} \wedge n)$ (in some specific situations, see the comment in (Schmidt et al., 2017, Section 4.1), the size can be reduced). The initialization step also requires the computation of a sum over the n examples.

2.3.4 An optimized FIEM algorithm, opt-FIEM

From (9), line 9 of algorithm 4 and the control variate technique, we explore here the idea to modify the original `FIEM` as follows (compare to line 9 in algorithm 4)

$$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}\left(\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \lambda_{k+1}\left(\widetilde{S}^{k+1} - \mathsf{S}_{k+1,J_{k+1}}\right)\right) \tag{12}$$

where $\lambda_{k+1} \in \mathbb{R}$ is chosen in order to minimize the conditional fluctuation

$$\gamma_{k+1}^{-2}\ \mathbb{E}\left[\|\widehat{S}^{k+1} - \widehat{S}^k\|^2 | \mathcal{F}_{k+1/2}\right]\ .$$

Upon noting that $\mathbb{E}\left[\widehat{S}^{k+1} - \widehat{S}^k | \mathcal{F}_{k+1/2}\right] = \gamma_{k+1} h(\widehat{S}^k)$, it is easily seen that equivalently, λ_{k+1} is chosen as the minimum of the conditional variance

$$\mathbb{E}\left[\|\gamma_{k+1}^{-1}\left(\widehat{S}^{k+1} - \widehat{S}^k\right) - h(\widehat{S}^k)\|^2 | \mathcal{F}_{k+1/2}\right]\ .$$

We will refer to this technique as the optimized FIEM (`opt-FIEM`) below. Observe that `SA` corresponds to the choice $\lambda_{k+1} = 0$ for any $k \geq 0$ (see algorithm 2); and `FIEM` corresponds to the choice $\lambda_{k+1} = 1$ for any $k \geq 0$ (see algorithm 4).

Upon noting that, given two random variables U, V such that $\mathbb{E}[\|V\|^2] > 0$, the function $\lambda \mapsto \mathbb{E}\left[\|U + \lambda V\|^2\right]$ reaches its minimum at a unique point given by $\lambda_\star \overset{\text{def}}{=} -\mathbb{E}\left[U^T V\right] / \mathbb{E}\left[\|V\|^2\right]$, the optimal choice for λ_{k+1} is given by (remember that conditionally to $\mathcal{F}_{k+1/2}$, $\widetilde{S}^{k+1} - \mathsf{S}_{k+1,J_{k+1}}$ is centered),

$$\lambda_{k+1}^\star \overset{\text{def}}{=} -\frac{\text{Tr Cov}\left(\bar{s}_J \circ \mathsf{T}(\widehat{S}^k), \widetilde{S}^{k+1} - \mathsf{S}_{k+1,J} | \mathcal{F}_{k+1/2}\right)}{\text{Tr Var}\left(\widetilde{S}^{k+1} - \mathsf{S}_{k+1,J} | \mathcal{F}_{k+1/2}\right)} \tag{13}$$

where J is a uniform random variable on $\{1,\dots,n\}$, independent of $\mathcal{F}_{k+1/2}$, Tr denotes the trace of a matrix, and Cov, Var are resp. the covariance and variance

matrices. With this optimal value, it holds from (12)

$$\begin{aligned}\gamma_{k+1}^{-2}\,\mathbb{E}\left[\|\widehat{S}^{k+1}-\widehat{S}^{k}\|^2|\mathcal{F}_{k+1/2}\right]\\ =\operatorname{Tr}\operatorname{Var}\left(\bar{s}_J\circ\mathsf{T}(\widehat{S}^k)-\widehat{S}^k|\mathcal{F}_{k+1/2}\right)\cdots\\ \times\left(1-\operatorname{Corr}^2\left(\bar{s}_J\circ\mathsf{T}(\widehat{S}^k),\widetilde{S}^{k+1}-\mathsf{S}_{k+1,J}|\mathcal{F}_{k+1/2}\right)\right)\,,\end{aligned}\tag{14}$$

where $\operatorname{Corr}(U,V)\stackrel{\text{def}}{=}\operatorname{TrCov}(U,V)/\{\operatorname{TrVar}(U)\ \operatorname{TrVar}(V)\}^{1/2}$. If the `opt-FIEM` algorithm $\{(\widehat{S}^k,\mathsf{S}_{k,\cdot}),k\geq 0\}$ converges to $(s^\star,\mathsf{S}_{\star,\cdot})$, we have $n^{-1}\sum_{i=1}^n\mathsf{S}_{\star,i}=s^\star=\bar{s}\circ\mathsf{T}(s^\star)$ and $\mathsf{S}_{\star,i}=\bar{s}_i\circ\mathsf{T}(s^\star)$ (see subsubsection 2.3.2 and subsubsection 2.3.3 for a similar discussion) thus showing that asymptotically, $\lambda_k^\star\approx 1$ (which implies that the correlation is 1 in (14)). This value is the value proposed in the original `FIEM`: therefore, asymptotically `opt-FIEM` and `FIEM` should be equivalent and `opt-FIEM` should have a better behavior in the transient phase. We will compare numerically `SA`, `FIEM` and `opt-FIEM` in section 4.

Upon noting that

$$\begin{aligned}\lambda_{k+1}^\star&=-\frac{n^{-1}\sum_{j=1}^n\left\langle\bar{s}_j\circ\mathsf{T}(\widehat{S}^k),\widetilde{S}^{k+1}-\mathsf{S}_{k+1,j}\right\rangle}{n^{-1}\sum_{j=1}^n\|\widetilde{S}^{k+1}-\mathsf{S}_{k+1,j}\|^2}\,,\\&=-\frac{n^{-1}\sum_{j=1}^n\left\langle\bar{s}_j\circ\mathsf{T}(\widehat{S}^k),\widetilde{S}^{k+1}-\mathsf{S}_{k+1,j}\right\rangle}{n^{-1}\sum_{j=1}^n\|\mathsf{S}_{k+1,j}\|^2-\|\widetilde{S}^{k+1}\|^2}\,.\end{aligned}$$

the computational cost of $\lambda_{k+1}^\star$ is proportional to n: it is therefore an intractable quantity in the large scale learning setting considered in this paper. A numerical approximation has to be designed: for example, a Monte Carlo approximation of the numerator; and a recursive approximation (along the iterations k) of the denominator, miming the same idea as the recursive computation of the sum $\widetilde{S}^k=n^{-1}\sum_{i=1}^n\mathsf{S}_{k,i}$ in `online-EM` and `FIEM`.

3 Non asymptotic convergence bounds

3.1 A general result

In this non-convex setting, convergence is characterized by the behavior of the gradient of the objective function along the path, or in the EM setting, by a "distance" of the path to the set of the roots of h. Under our assumptions, both quantities are related as stated in Proposition 3. The convergence bounds are obtained by strengthening H4 with the following assumptions

H5 *1. There exist $0<v_{\min}\leq v_{max}<\infty$ such that for all $s\in\mathbb{R}^q$, the spectrum of $B(s)$ is in $[v_{\min},v_{\max}]$; $B(s)$ is defined in H4.*

2. For any $i\in\{1,\ldots,n\}$, $\bar{s}_i\circ\mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i.

3. *The function* $s \mapsto \dot{V}(s) = -B(s)h(s)$ *is globally Lipschitz on* $\mathbb{R}^q$ *with constant* $L_{\dot{V}}$.

We consider here the control of the following quantities (as in Ghadimi and Lan (2013), Allen-Zhu and Hazan (2016), Reddi et al. (2016), Fang et al. (2018), Zhou et al. (2018) and Karimi et al. (2019c) to cite few contributions in the stochastic non-convex finite-sum optimization). Given a maximal number of iterations $K_{\max}$, and a random variable K taking values in $\{0, \ldots, K_{\max} - 1\}$, define

$$\begin{aligned}
\mathsf{E}_0 &\stackrel{\text{def}}{=} \frac{1}{v_{\min}^2} \mathbb{E}\left[\|\dot{V}(\widehat{S}^K)\|^2\right] , \\
\mathsf{E}_1 &\stackrel{\text{def}}{=} \mathbb{E}\left[\|h(\widehat{S}^K)\|^2\right] , \\
\mathsf{E}_2 &\stackrel{\text{def}}{=} \mathbb{E}\left[\|\widetilde{S}^{K+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^K)\|^2\right] ,
\end{aligned}$$

where K is a random termination rule, chosen independently of the path: upper bounds of these quantities provide insights on the behavior of FIEM when stopped at a random time. Except in subsection 3.4, below K is the uniform r.v. on $\{0, \ldots, K_{\max} - 1\}$.

The quantities E_0 and E_1 are classical in the literature: they stand for a way to measure resp. a distance to a stationary point of the objective $V = F \circ \mathsf{T}$, and a distance to the fixed points of EM. E_2 is specific to FIEM: it quantifies how far the control variate $\widetilde{S}^{k+1}$ is from the intractable mean $\bar{s} \circ \mathsf{T}(\widehat{S}^k)$ (see (10) for the definition of $\widetilde{S}^{k+1}$).

From Proposition 2-item 2, we trivially have (the proof is omitted)

Proposition 3 *Assume H1, H2, H3, H4 and H5-item 1. Then for any* $s \in \mathbb{R}^q$, $\left\langle h(s), \dot{V}(s) \right\rangle \leq -v_{\min} \|h(s)\|^2$ *and* $\mathsf{E}_0 \leq \mathsf{E}_1$.

Theorem 1 is a general result for the control of quantities of the form

$$\sum_{k=0}^{K_{\max}-1} \alpha_k \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] + \sum_{k=0}^{K_{\max}-1} \delta_k \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^k)\|^2\right]$$

where $\alpha_k \in \mathbb{R}$ and $\delta_k > 0$. In subsection 3.2 and subsection 3.3, we discuss how to choose the step sizes $\{\gamma_k, k \geq 1\}$ such that for any $k \in \{0, \ldots, K_{\max} - 1\}$, α_k is non-negative and such that $A_{K_{\max}} \stackrel{\text{def}}{=} \sum_{k=0}^{K_{\max}-1} \alpha_k$ is positive. We then deduce from Theorem 1 an upper bound for

$$\sum_{k=0}^{K_{\max}-1} \frac{\alpha_k}{A_{K_{\max}}} \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] + \sum_{k=0}^{K_{\max}-1} \frac{\delta_k}{A_{K_{\max}}} \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^k)\|^2\right] \tag{15}$$

such that the larger $A_{K_{\max}}$ is, the better the bound is. (15) is then used to obtain upper bounds on E_1 and E_2; which provides also an upper bound on E_0 by Proposition 3.

Theorem 1 *Assume H1, H2, H3, H4 and H5. Define* $L^2 \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$.

Let $K_{\max}$ *be a positive integer,* $\{\gamma_k,\ k\in\mathbb{N}\}$ *be a sequence of positive stepsizes and* $\widehat{S}^0 \in \mathbb{R}^q$. *Consider the FIEM sequence* $\{\widehat{S}^k,\ k\in\mathbb{N}\}$ *given by algorithm 4. Set* $\Delta V \stackrel{\text{def}}{=} \mathbb{E}\left[V(\widehat{S}^0)\right] - \mathbb{E}\left[V(\widehat{S}^{K_{\max}})\right]$.

For any positive numbers $\beta_1,\ldots,\beta_{K_{\max}-1}$, *we have*

$$\sum_{k=0}^{K_{\max}-1} \alpha_k\ \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] + \sum_{k=0}^{K_{\max}-1} \delta_k \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}^k)\|^2\right] \leq \Delta V \quad ,$$

where for any $k=0,\ldots,K_{\max}-1$,

$$\alpha_k \stackrel{\text{def}}{=} \gamma_{k+1} v_{min} - \gamma_{k+1}^2\left(1+\Lambda_k L^2\right)\frac{L_{\dot V}}{2}, \quad \delta_k \stackrel{\text{def}}{=} \gamma_{k+1}^2\left(1+\frac{\Lambda_k\beta_{k+1}L^2}{(1+\beta_{k+1})}\right)\frac{L_{\dot V}}{2}\ ,$$

with $\Lambda_{K_{\max}-1}=0$ *and for* $k=0,\ldots,K_{\max}-2$,

$$\Lambda_k \stackrel{\text{def}}{=} \left(1+\frac{1}{\beta_{k+1}}\right)\sum_{j=k+1}^{K_{\max}-1}\gamma_{j+1}^2\prod_{\ell=k+2}^{j}\left(1-\frac{1}{n}+\beta_\ell+\gamma_\ell^2 L^2\right)\ .$$

Proof The detailed proof is provided in Section 6.2; let us give here a sketch of proof. Define H_{k+1} such that $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}H_{k+1}$. V is regular enough so that

$$V(\widehat{S}^{k+1}) - V(\widehat{S}^k) - \gamma_{k+1}\left\langle H_{k+1}, \dot V(\widehat{S}^k)\right\rangle \leq \gamma_{k+1}^2\frac{L_{\dot V}}{2}\|H_{k+1}\|^2\ .$$

Then, the next step is to prove that

$$\begin{aligned}\mathbb{E}\left[V(\widehat{S}^{k+1})\right] - \mathbb{E}\left[V(\widehat{S}^k)\right] + \gamma_{k+1}\left(v_{\min} - \gamma_{k+1}\frac{L_{\dot V}}{2}\right)\mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right]\\ \leq \gamma_{k+1}^2\frac{L_{\dot V}}{2}\mathbb{E}\left[\|H_{k+1} - \mathbb{E}\left[H_{k+1}|\mathcal{F}_{k+1/2}\right]\|^2\right]\ ,\end{aligned}$$

which, by summation from $k=0$ to $k=K_{\max}-1$, yields

$$\begin{aligned}\sum_{k=0}^{K_{\max}-1}\gamma_{k+1}\left(v_{\min}-\gamma_{k+1}\frac{L_{\dot V}}{2}\right)\mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] \leq \mathbb{E}\left[V(\widehat{S}^0)\right] - \mathbb{E}\left[V(\widehat{S}^K_{\max})\right]\\ + \frac{L_{\dot V}}{2}\sum_{k=0}^{K_{\max}-1}\gamma_{k+1}^2\mathbb{E}\left[\|H_{k+1} - \mathbb{E}\left[H_{k+1}|\mathcal{F}_{k+1/2}\right]\|^2\right]\ .\end{aligned}$$

The most technical part is to prove that the last term on the RHS is upper bounded by

$$\begin{aligned}\frac{L_{\dot V}}{2}\sum_{k=0}^{K_{\max}-1}\gamma_{k+1}^2 L^2\ \Big\{\Lambda_k\mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right]\\ - \left(1+(1+\beta_{k+1}^{-1})^{-1}\Lambda_k\right)\ \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}^k)\|^2\right]\Big\}\ .\end{aligned}$$

This concludes the proof.

3.2 A uniform random stopping rule for a $n^{2/3}$-complexity

The main result of this section establishes that by choosing a constant stepsize sequence and a termination rule K sampled uniformly on $\{0, \cdots, K_{\max}-1\}$, an ϵ-approximate stationary point can be reached before $K_{\max} = O(n^{2/3}\epsilon^{-1}L_{\dot{V}}^{1/3}L^{2/3})$ iterations. In the Stochastic Gradient Descent literature, complexity is evaluated in terms of *Incremental First-order Oracle* introduced by Agarwal and Bottou (2015), that is, roughly speaking, number of calls to an oracle which returns a pair $(f_i(x), \nabla f_i(x))$. In our case, the equivalent cost is the number of computations of an expectation $\bar{s}_i(\theta)$ - see H2. $K_{\max}$ iterations of FIEM calls $2K_{\max}$ computations of such expectations. As a consequence, the complexity analyses consist in discussing how $K_{\max}$ has to be chosen as a function of n and ϵ.

For $\lambda \in (0,1)$, $C > 0$ and n such that $n^{-1/3} < \lambda/C$, define

$$f_n(C,\lambda) \stackrel{\text{def}}{=} \left(\frac{1}{n^{2/3}} + \frac{C}{\lambda - Cn^{-1/3}}\left(\frac{1}{n} + \frac{1}{1-\lambda}\right)\right) . \tag{16}$$

Proposition 4 (application of Theorem 1) *Let $\mu \in (0,1)$. Choose $\lambda \in (0,1)$ and $C \in (0,+\infty)$ such that*

$$\sqrt{C} f_n(C,\lambda) = 2\mu v_{\min}\frac{L}{L_{\dot{V}}} . \tag{17}$$

Let $\{\widehat{S}^k, k \in \mathbb{N}\}$ be the FIEM sequence given by algorithm 4 run with the constant step size

$$\gamma_\ell = \gamma_{\text{FGM}} \stackrel{\text{def}}{=} \frac{\sqrt{C}}{n^{2/3}L} = \frac{2\mu v_{\min}}{f_n(C,\lambda)\, n^{2/3} L_{\dot{V}}} . \tag{18}$$

For any $n > (C/\lambda)^3$ and $K_{\max} \geq 1$, we have

$$\mathsf{E}_1 + \frac{\mu}{(1-\mu) f_n(C,\lambda)\, n^{2/3}}\mathsf{E}_2 \leq \frac{n^{2/3}}{K_{\max}}\frac{L_{\dot{V}}\, f_n(C,\lambda)}{2\mu(1-\mu)v_{\min}^2}\Delta V , \tag{19}$$

where the errors E_i are defined with a random variable K sampled uniformly on $\{0, \cdots, K_{\max}-1\}$.

The proof of Proposition 4 is given in subsubsection 6.2.2. A first suggestion to solve the equation (17) is to choose $\lambda = C$ and $C \in (0,1)$ such that $\sqrt{C} f_n(C,C) = 2\mu v_{\min} L/L_{\dot{V}}$. This equation possesses an unique solution in $(0,1)$ which is upper bounded by C^+ given by

$$C^+ \stackrel{\text{def}}{=} \frac{\sqrt{1 + 16\mu v_{\min}^2 L^2 L_{\dot{V}}^{-2}} - 1}{4\mu v_{\min} L L_{\dot{V}}^{-1}}$$

The consequence is that, given $\varepsilon \in (0,1)$, by setting

$$M \stackrel{\text{def}}{=} \frac{L_{\dot{V}}}{\mu(1-\mu)v_{\min}^2} f_2(C^+, C^+) ,$$

we have

$$K_{\max} = M\ n^{2/3}\varepsilon^{-1} \Longrightarrow \mathsf{E}_1 + \frac{L_{\dot{V}}}{2L}\frac{\sqrt{C}}{v_{\min} n^{2/3}}\mathsf{E}_2 \leq \varepsilon\, \Delta V\ ;$$

see subsubsection 6.2.2 for a detailed proof of this comment.

Another suggestion is to exploit how (16) behaves when $n \to +\infty$; we prove in subsubsection 6.2.2 again that, there exists $N_\star$ depending only upon $L, L_{\dot{V}}, v_{\min}$ such that for any $n \geq N_\star$,

$$\mathsf{E}_1 + \frac{4}{3}\frac{C}{n^{2/3}}\left(\frac{L_{\dot{V}}}{L v_{\min}}\right)^{4/3}\mathsf{E}_2 \leq \frac{n^{2/3}}{K_{\max}}\frac{8}{3}\frac{L}{v_{\min}}\left(\frac{L_{\dot{V}}}{L v_{\min}}\right)^{1/3}\ \Delta V\ ,$$

by choosing $C \leftarrow 0.25\ (v_{\min} L/L_{\dot{V}})^{2/3}$ in the definition of the step size γ_{FGM}.

The conclusions of Proposition 4 confirm and improve previous results in the literature: (Karimi et al., 2019c, Theorem 2) proves that for FIEM run with the constant size

$$\gamma_{\mathrm{K}} \stackrel{\mathrm{def}}{=} \frac{v_{min}}{\max(6, 1+4v_{\min})\ \max(L_{\dot{V}}, L_1, \ldots, L_n)\ n^{2/3}}\ , \tag{20}$$

it holds

$$\mathsf{E}_1 \leq \frac{n^{2/3}}{K_{\max}}\frac{(\max(6, 1+4v_{\min}))^2\ \max(L_{\dot{V}}, L_1, \ldots, L_n)}{v_{\min}^2}\ \Delta V\ . \tag{21}$$

We improve on this result by first showing that the RHS in (19) controls a larger quantity than E_1. In addition, numerical explorations (see e.g. section 4) show that our step size γ_{FGM} is larger than the step size γ_{K} thus providing a more aggressive step size which may have a beneficial effect on the numerical implementation. It also shows that Proposition 4 provides a tighter control of convergence. In both contributions however, the step size depends on n as $O(1/n^{2/3})$ and, the explicit control increases at the rate $n^{2/3}$ and decreases at the rate $1/K_{\max}$. The rate of the step size is the same as what is observed for Stochastic Gradient Descent (see e.g. Allen-Zhu and Hazan (2016)).

3.3 A uniform random stopping rule for a $\sqrt{n}$-complexity

In subsubsection 6.2.3, we prove the following control obtained, here again, along a FIEM path run with a constant stepsize sequence and stopped at a random time K sampled uniformly on $\{0, \cdots, K_{\max}-1\}$: an ϵ-stationary point can be reached before $K_{\max} = O(\sqrt{n}\epsilon^{-3/2})$ iterations.

Define

$$\tilde{f}_n(C, \lambda) \stackrel{\mathrm{def}}{=} \frac{1}{(nK_{\max})^{1/3}} + C\left(\frac{1}{n} + \frac{1}{1-\lambda}\right)\ . \tag{22}$$

Proposition 5 (application of Theorem 1) *Let $\mu \in (0,1)$. Choose $\lambda \in (0,1)$ and $C > 0$ such that*

$$\sqrt{C}\tilde{f}_n(C,\lambda) = 2\mu v_{\min}\frac{L}{L_{\dot{V}}} \ . \tag{23}$$

Let $\{\widehat{S}^k, k \in \mathbb{N}\}$ be the FIEM sequence given by algorithm 4 run with the constant step size

$$\gamma_\ell = \tilde{\gamma}_{\mathrm{FGM}} \stackrel{\mathrm{def}}{=} \frac{\sqrt{C}}{n^{1/3}K_{\max}^{1/3}L} = \frac{2\mu v_{\min}}{L_{\dot{V}}\tilde{f}_n(C,\lambda)\, n^{1/3}K_{\max}^{1/3}} \ . \tag{24}$$

Then for any $n, K_{\max} \geq 1$ such that $n^{1/3}K_{\max}^{-2/3} \leq \lambda/C$, we have

$$\mathsf{E}_1 + \frac{\mu}{(1-\mu)\tilde{f}_n(C,\lambda)}\frac{1}{(nK_{\max})^{1/3}}\mathsf{E}_2 \leq \frac{n^{1/3}}{K_{\max}^{2/3}}\frac{L_{\dot{V}}\,\tilde{f}_n(C,\lambda)}{2\mu(1-\mu)v_{\min}^2}\,\Delta V \ ,$$

where the errors E_i are defined with a random variable K sampled uniformly on $\{0,\cdots,K_{\max}-1\}$.

The proof of Proposition 5 is given in subsubsection 6.2.3. From this upper bound, it can be shown (see subsubsection 6.2.3) that for any $\tau > 0$, there exists $M > 0$ depending upon $L, L_{\dot{V}}, v_{\min}, \mu$ and τ such that for any $\varepsilon > 0$,

$$K_{\max} \geq \left(\sqrt{n}\tau^{3/2}\right) \vee \left(M\sqrt{n}\varepsilon^{-3/2}\right) \Longrightarrow \frac{n^{1/3}}{K_{\max}^{2/3}}\frac{L_{\dot{V}}\,\tilde{f}_n(\lambda\tau,\lambda)}{2\mu(1-\mu)v_{\min}^2} \leq \varepsilon \ .$$

To our best knowledge, this is the first result in the literature which establishes a non asymptotic control for FIEM at such a rate: the upper bound is an increasing function of n at the rate $n^{1/3}$ and a decreasing function of $K_{\max}$ at the rate $K_{\max}^{-2/3}$.

As a corollary of Proposition 4 and Proposition 5, we have two upper bounds of the errors $\mathsf{E}_1, \mathsf{E}_2$: the first one is $O(n^{2/3}K_{\max}^{-1})$ and the second one is $O(n^{1/3}K_{\max}^{-2/3})$. Given a tolerance $\varepsilon > 0$, the first or second strategy will be chosen depending on how $\sqrt{n}\varepsilon^{-3/2}$ and $n^{2/3}\varepsilon^{-1}$ compare: for any $A > 0$, $\sqrt{n}\epsilon^{-3/2} < An^{2/3}\epsilon^{-1}$ iff $n^{-1/3} < \epsilon A^2$.

When $K_{\max} = A\sqrt{n}\epsilon^{-3/2}$, then $\tilde{\gamma}_{\mathrm{FGM}} = \sqrt{C}\sqrt{\epsilon}/(LA^{1/3}\sqrt{n})$. In the case $\sqrt{n}\epsilon^{-3/2} < \tilde{A}n^{2/3}\epsilon^{-1}$, we have $\tilde{\gamma}_{\mathrm{FGM}} > \sqrt{C}/(LA^{1/3}\tilde{A}n^{2/3})$ thus showing that the step size is lower bounded by $O(n^{-2/3})$ (see γ_{FGM} in Proposition 4). We have $\tilde{\gamma}_{\mathrm{FGM}} \propto 1/\sqrt{n}$ when $K_{\max} \propto \sqrt{n}$: the result of Proposition 5 is obtained with a slower step size (seen as a function of n) than what was required in Proposition 4.

We now discuss a choice for the pair (C,λ) which exploits how (22) behaves when $n \to +\infty$; we prove in subsubsection 6.2.3 that for any $\tau > 0$, there exists $N_\star$ depending only upon $L, L_{\dot{V}}, v_{\min}, \tau$ such that for any $N_\star \leq n \leq \tau^3 K_{\max}^2$,

$$\mathsf{E}_1 + \frac{2^{4/3}(1-\lambda_\star)^{-1/3}\mu^2}{\tilde{f}_n^2(\lambda_\star\tau,\lambda_\star)}\left(\frac{Lv_{\min}}{L_{\dot{V}}}\right)^{2/3}\frac{1}{(nK_{\max})^{1/3}}\mathsf{E}_2 \leq \frac{n^{1/3}}{K_{\max}^{2/3}}\frac{4}{3}\left(\frac{2L^2L_{\dot{V}}}{v_{\min}^4}\right)^{1/3}(1-\lambda_\star)^{-1/3}\,\Delta V \ ,$$

where $\lambda_\star$ is the unique solution of $(v_{\min}L)^2\,\tau^3(1-\lambda_\star)^2 = (2L_{\dot{V}})^2\lambda_\star^3$.

3.4 A non-uniform random stopping rule

Given a distribution $p_0, \ldots, p_{K_{\max}-1}$ for the r.v. K, we show how to fix the step sizes $\gamma_1, \ldots, \gamma_{K_{\max}}$ in order to deduce from Theorem 1 a control of the errors E_1 and E_2. For $\lambda \in (0,1)$, $C > 0$ and $n > (C/\lambda)^3$, define the function $F_{n,C,\lambda}$

$$F_{n,C,\lambda} : x \mapsto \frac{L_{\dot V}}{2L^2 n^{2/3}} x \left(v_{\min} \frac{2L}{L_{\dot V}} - x f_n(C,\lambda) \right) ,$$

where f_n is defined by (16). $F_{n,C,\lambda}$ is positive, increasing and continuous on $(0, v_{\min} L/(L_{\dot V} f_n(C,\lambda))]$.

Proposition 6 (application of Theorem 1) *Let K be a $\{0, \ldots, K_{\max}-1\}$-valued random variable with positive weights $p_0, \ldots, p_{K_{\max}-1}$. Choose $\lambda \in (0,1)$ and $C > 0$ such that*

$$\sqrt{C}\, f_n(C,\lambda) = v_{\min} \frac{L}{L_{\dot V}} . \tag{25}$$

For any $n > (C/\lambda)^3$ and $K_{\max} \geq 1$, we have

$$\mathsf{E}_1 + \frac{L_{\dot V}^2}{v_{\min}^2} n^{2/3} \max_k p_k \, f_n(C,\lambda) \sum_{k=0}^{K_{\max}-1} \gamma_{k+1}^2 \mathbb{E}\left[\| \widetilde{S}^{k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^k) \|^2 \right] \leq n^{2/3} \max_k p_k \, \frac{2 L_{\dot V} \, f_n(C,\lambda)}{v_{\min}^2} \, \Delta V ,$$

where the FIEM sequence $\{\widehat{S}^k, k \in \mathbb{N}\}$ is obtained with

$$\gamma_{k+1} = \frac{1}{n^{2/3} L} \, F_{n,C,\lambda}^{-1} \left(\frac{p_k}{\max_\ell p_\ell} \frac{v_{\min}^2}{2 L_{\dot V} f_n(C,\lambda)} \frac{1}{n^{2/3}} \right) .$$

The proof of Proposition 6 is in subsubsection 6.2.4.

Since $\sum_k p_k = 1$, we have $\max_k p_k \geq 1/K_{\max}$ thus showing that among the distributions $\{p_j, 0 \leq j \leq K_{\max}-1\}$, the term $\max_k p_k$ is minimal with the uniform distribution. In that case, the results of Proposition 6 can be compared to the results of Proposition 4: both RHS are increasing functions of n at the rate $n^{2/3}$; both are decreasing functions of $K_{\max}$ at the rate $1/K_{\max}$; the constants C, λ solving the equality in (17) in the case $\mu = 1/2$ are the same as the constants C, λ solving (25): as a consequence,

$$\frac{2 L_{\dot V} f_n(C,\lambda)}{v_{\min}^2} = \frac{L_{\dot V} f_n(C,\lambda)}{2\mu(1-\mu) v_{\min}^2} , \qquad \mu = 1/2.$$

Finally, when $k \mapsto p_k$ is constant, the step sizes given by Proposition 6 are constant as in Proposition 4; and they are equal since

$$F_{n,C,\lambda}^{-1} \left(\frac{v_{\min}^2 n^{-2/3}}{2 L_{\dot V} f_n(C,\lambda)} \right) = \sqrt{C} = \frac{v_{\min} L}{L_{\dot V} f_n(C,\lambda)} .$$

Hence Proposition 6 and Proposition 4 are the same when $p_k = 1/K_{\max}$ for any k.

The strategy $p_0 = \ldots = p_{K_{\max}-2} = \iota/(K_{\max}-1)$ and $p_{K_{\max}-1} = 1-\iota/(K_{\max}-1)$ for some $\iota << 1$, which consists in terminating the algorithm at $k = K_{\max}-1$ with high probability, provides a very bad upper bound since there is no decay when $K_{\max} \to \infty$.

As already commented in subsection 3.2, if we choose $C = \lambda$, then (25) gets into

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

There exists an unique solution $C^\star$, which is upper bounded by a quantity which only depends upon $L, L_{\dot{V}}, v_{\min}$; hence, so $f_n(C^\star, C^\star)$ is and the control of E_i given in Proposition 6 has the same behavior in $n, K_{\max}$ as $n^{2/3}\max_k p_k$.

If we choose $\lambda = 1/2$, the constant C satisfies (see subsubsection 6.2.4)

$$C \leq \left(\frac{v_{\min}L}{4L_{\dot{V}}}\right)^{2/3} ,$$

and the non asymptotic control given by Proposition 6 is available for $8n > (v_{\min}L/L_{\dot{V}})^2$.

4 A toy example

When introducing the novel algorithm `FIEM`, Karimi et al. (2019c) applied it to non trivial models. In this section, we consider a very simple optimization problem which does not require the incremental EM machinery to be solved. However, since the quantities of interest are explicit, it is used here to illustrate the role of the design parameters and to fairly compare the algorithms.

4.1 Description

n $\mathbb{R}^y$-valued observations are modeled as the realization of n vectors $Y_i \in \mathbb{R}^y$ whose distribution is described as follows: conditionally to $(Z_1, \ldots, Z_n)$, the r.v. are independent with distribution $Y_i \sim \mathcal{N}_y(AZ_i, \mathrm{I}_y)$ where $A \in \mathbb{R}^{y\times p}$ is a deterministic matrix and I_y denotes the $y \times y$ identity matrix; $(Z_1, \ldots, Z_n)$ are i.i.d. under the distribution $\mathcal{N}_p(X\theta, \mathrm{I}_p)$, where $\theta \in \Theta \stackrel{\text{def}}{=} \mathbb{R}^q$ and $X \in \mathbb{R}^{p\times q}$ is a deterministic matrix. Here, X and A are known, and θ is unknown; the objective is estimate θ, as a solution of a (possibly) penalized maximum likelihood estimator, with penalty term $\rho(\theta) \stackrel{\text{def}}{=} \upsilon\|\theta\|^2/2$ for some $\upsilon \geq 0$. If $\upsilon = 0$, it is assumed that the rank of X and AX is resp. $q = q \wedge y$ and $p = p \wedge y$. In this model, the r.v. $(Y_1, \ldots, Y_n)$ are i.i.d. with distribution $\mathcal{N}_y(AX\theta; \mathrm{I}_y + AA^T)$. The minimum of the function

$\theta \mapsto F(\theta) \stackrel{\text{def}}{=} -n^{-1} \log g(Y_{1:n}; \theta) + \rho(\theta)$, where $g(Y_{1:n}; \cdot)$ denotes the likelihood of the vector $(Y_1, \ldots, Y_n)$, is unique and is given by

$$\theta_\star \stackrel{\text{def}}{=} \left(\upsilon \mathrm{I}_q + X^T A^T \left(\mathrm{I}_y + A A^T \right)^{-1} A X \right)^{-1} X^T A^T \left(\mathrm{I}_y + A A^T \right)^{-1} \overline{Y}_n,$$

$$\overline{Y}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Nevertheless, using the above description of the distribution of Y_i, this optimization problem can be cast into the general framework described in Section 2.1. The loss function (see (4)) is the normalized negative log-likelihood of the distribution of Y_i and is of the form (4) with

$$\phi(\theta) \stackrel{\text{def}}{=} \theta, \quad \mathsf{R}(\theta) \stackrel{\text{def}}{=} \frac{1}{2} \theta^T (X^T X + \upsilon \mathrm{I}_q) \theta, \quad s_i(z) \stackrel{\text{def}}{=} X^T z.$$

Under the stated assumptions on X, the function $\theta \mapsto -\langle s, \phi(\theta) \rangle + R(\theta)$ is defined on $\mathbb{R}^q$ and for any $s \in \mathbb{R}^q$, it possesses an unique minimum given by

$$\mathsf{T}(s) \stackrel{\text{def}}{=} (\upsilon \mathrm{I}_q + X^T X)^{-1} s.$$

Define

$$\begin{aligned} \Pi_1 &\stackrel{\text{def}}{=} X^T (\mathrm{I}_p + A^T A)^{-1} A^T \in \mathbb{R}^{q \times y}, \\ \Pi_2 &\stackrel{\text{def}}{=} X^T (\mathrm{I}_p + A^T A)^{-1} X (\upsilon \mathrm{I}_q + X^T X)^{-1} \in \mathbb{R}^{q \times q}. \end{aligned}$$

The a posteriori distribution $p_i(\cdot, \theta) \mathrm{d}\mu$ of the latent variable Z_i given the observation Y_i is a Gaussian distribution

$$\mathcal{N}_p \left((\mathrm{I}_p + A^T A)^{-1} (A^T Y_i + X\theta), (\mathrm{I}_p + A^T A)^{-1} \right),$$

so that for all $i \in \{1, \ldots, n\}$,

$$\bar{s}_i(\theta) \stackrel{\text{def}}{=} X^T (\mathrm{I}_p + A^T A)^{-1} (A^T Y_i + X\theta) = \Pi_1 Y_i + X^T (\mathrm{I}_p + A^T A)^{-1} X \theta \in \mathbb{R}^q,$$

$$\bar{s}_i \circ \mathsf{T}(s) = \Pi_1 Y_i + \Pi_2 s.$$

Therefore, the assumptions H1, H2, H3 and H4-item 1,item 2 are satisfied. Since $\phi \circ \mathsf{T}(s) = \mathsf{T}(s)$ then $B(s) = (\upsilon \mathrm{I}_q + X^T X)^{-1}$ for any $s \in \mathbb{R}^q$, H4-item 3 and H5-item 1 hold with

$$\upsilon_{\min} \stackrel{\text{def}}{=} \frac{1}{\upsilon + \max\_\mathrm{eig}(X^T X)}, \qquad \upsilon_{\max} \stackrel{\text{def}}{=} \frac{1}{\upsilon + \min\_\mathrm{eig}(X^T X)};$$

here, max_eig and min_eig denotes resp. the maximum and the minimum of the eigenvalues. $\bar{s}_i \circ \mathsf{T}(s) = \Pi_1 Y_i + \Pi_2 s$ thus showing that H5-item 2 holds with the same constant $L_i = L$ for all i. Finally, $s \mapsto B^T(s) (\bar{s} \circ \mathsf{T}(s) - s)$ is globally Lipschitz with constant

$$L_{\dot{V}} \stackrel{\text{def}}{=} \max \left| \mathrm{eig} \left((\upsilon \mathrm{I}_q + X^T X)^{-1} (\Pi_2 - \mathrm{I}_q) \right) \right|;$$

here eig denotes the eigenvalues. This concludes the proof of H5-item 3.

4.2 The algorithms

Given the current value $\widehat{S}^k$, one iteration of `EM`, `SA`, `FIEM` and `opt-FIEM` are given by algorithm 5 and algorithm 6.

All the algorithms (except EM) require $K_{\max}$ random draws from $\{1, \ldots, n\}$ per run of length $K_{\max}$ iterations; `FIEM` and `opt-FIEM` require $2 \times K_{\max}$ draws. For a fair comparison of the algorithms along one run, one vector of integers is sampled prior the runs and is common to all the algorithms. Such a protocol allows to compare the strategies by "freezing" the randomness due to the random choice of the examples, and to really explain the different behaviors only by the values of the design parameters (the step size, for example) or by the updating scheme specific to each algorithm.

All the paths, whatever the algorithms, are started at the same value $\widehat{S}^0$.

Data: $\widehat{S}^k \in \mathbb{R}^q$, $\mathit{\Pi}_1$, $\mathit{\Pi}_2$ and $\overline{Y}_n$
Result: $\widehat{S}^{k+1}_{\mathrm{EM}}$
$\widehat{S}^{k+1}_{\mathrm{EM}} = \mathit{\Pi}_1 \overline{Y}_n + \mathit{\Pi}_2 \widehat{S}^k$

Algorithm 5: Toy example: one iteration of `EM`.

Data: $\widehat{S}^k \in \mathbb{R}^q$, $\mathsf{S} \in \mathbb{R}^{qn}$, $\widetilde{S} \in \mathbb{R}^q$; a step size $\gamma_{k+1} \in (0, 1]$ and a coefficient λ_{k+1}; the matrices $\mathit{\Pi}_1$, $\mathit{\Pi}_2$; the examples $Y_1, \cdots, Y_n$
Result: $\widehat{S}^{k+1}_{\mathrm{FIEM}}$
Sample independently $I_{k+1}, J_{k+1} \sim \mathcal{U}(\{1, \cdots, n\})$;
Store $s = \mathsf{S}_{I_{k+1}}$;
Update $\mathsf{S}_{I_{k+1}} = \mathit{\Pi}_1 Y_{I_{k+1}} + \mathit{\Pi}_2 \widehat{S}^k$;
Update $\widetilde{S} = \widetilde{S} + n^{-1}(\mathsf{S}_{I_{k+1}} - s)$;
Update $\widehat{S}^{k+1}_{\mathrm{FIEM}} = \widehat{S}^k + \gamma_{k+1} \left(\mathit{\Pi}_1 Y_{J_{k+1}} + \mathit{\Pi}_2 \widehat{S}^k - \widehat{S}^k + \lambda_{k+1} \left\{ \widetilde{S} - \mathsf{S}_{J_{k+1}} \right\} \right)$

Algorithm 6: Toy example: one iteration of `SA` ($\lambda_{k+1} = 0$), `FIEM` ($\lambda_{k+1} = 1$) and `opt-FIEM`.

4.3 Numerical analysis

We choose $Y_i \in \mathbb{R}^{15}$, $Z_i \in \mathbb{R}^{10}$ and $\theta_{\mathrm{true}} \in \mathbb{R}^{20}$. The entries of the matrix A (resp. X) are obtained as a stationary Gaussian auto-regressive process: the first column is sampled from $\sqrt{1-\rho^2}\,\mathcal{N}_{15}(0; \mathrm{I})$ (resp. from $\sqrt{1-\tilde{\rho}^2}\,\mathcal{N}_{10}(0; \mathrm{I})$) with $\rho = 0.8$ (resp. $\tilde{\rho} = 0.9$). θ_{true} is sparse with 40% of the components set to zero; and the other ones are sampled uniformly in $[-5, 5]$.

The regularization parameter υ is set to 0.1.

`FIEM`*: the step sizes and the non asymptotic controls.* The first analysis is to compare the non asymptotic bounds and the constant step sizes provided by Proposition 4, Proposition 5 and (Karimi et al., 2019c, Theorem 2) (see also (20) and (21)): the bounds are of the form

$$\frac{n^{\mathsf{a}}}{K_{\max}^{\mathsf{b}}}\ \mathcal{B}\ \Delta V\ ;$$

the numerical results below correspond to $\Delta V = 1$ and are obtained with a data set of size $n = 1e6$. Figure 1 shows the value of the constant C solving (17) when λ is successively set to $\{0.25, 0.5, 0.75\}$ and as a function of $\mu \in (0.01, 0.9)$. Figure 2 shows the same analysis for the constant C solving (23). Figure 3 and Figure 4 display the quantity $\mathcal{B}$ as a function of μ and when (λ, C) is fixed to $\lambda \in \{0.25, 0.5, 0.75\}$ and C solves resp. (17) and (23). The role of λ looks quite negligible; the bound $\mathcal{B}$ seems to be optimal with $\mu \approx 0.25$. Note that the constants C and $\mathcal{B}$ given by Proposition 5 also depends on $K_{\max}$: the results displayed here correspond to $K_{\max} = n$ but we observed that the plots are the same with $K_{\max} = 1e2\, n$ and $K_{\max} = 1e3\, n$ (remember that $n = 1e6$).

Figure 5 displays the step sizes as a function of $\mu \in (0.01, 0.9)$, when $\lambda = 1/2$ and for different strategies of $K_{\max}$: $K_{\max} \in \{n, 1e2\, n, 1e3\, n\}$. Figure 6 displays the quantity $n^{\mathsf{a}} K_{\max}^{-\mathsf{b}} \mathcal{B}$. `Case 1` (resp. `Case 2`) corresponds to the definition given in Proposition 4 (resp. Proposition 5). For `Case 1` and `Karimi et al`, $(\mathsf{a}, \mathsf{b}) = (2/3, 1)$ and for `Case 2`, $(\mathsf{a}, \mathsf{b}) = (1/3, 2/3)$. The first conclusion is that our results improve on those by Karimi et al. (2019c): we provide a larger step size (improved by a factor up to 55, with the strategy `Case 1`, $\mu = 0.25$, $\lambda = 0.5$) and a tighter bound (reduced by a factor up to 235, with the strategy `Case 1`, $\mu = 0.25$, $\lambda = 0.5$). The second conclusion is about the comparison of Proposition 4 and Proposition 5: as already commented (see subsection 3.3), the first strategy is preferable when the tolerance level ϵ is small (w.r.t. $n^{-1/3}$).

Comparison of `SA`, `FIEM` *and* `opt-FIEM`. The algorithms are run with the same constant step size given by (18) when C solves (17) with $\mu = 0.25$ and $\lambda = 0.5$. The size of the data set is $n = 1e3$ and the maximal number of iterations is $K_{\max} = 20\, n$. Since the non asymptotic convergence bounds are essentially based on the control of $\gamma_{k+1}^{-2}\, \mathbb{E}\left[\|\widehat{S}^{k+1} - \widehat{S}^{k}\|^2\right]$ (see the sketch of proof of Theorem 1 in section 3), we first compare the algorithms through this criterion: the expectation is approximated by a Monte Carlo sum over $1e3$ independent runs. The second criterion for comparison is a distance of the iterates to the unique solution $\theta_\star$ via the expectation $\mathbb{E}\left[\|\theta^k - \theta_\star\|\right]$ and the standard deviation $\mathrm{std}\left(\|\theta^k - \theta_\star\|\right)$ again approximated by a Monte Carlo sum over the same $1e3$ independent runs.

Figure 7 displays the evolution of $k \mapsto \lambda_{k+1}^\star$, the optimal coefficient given by (13); in this toy example, it is computed explicitly. We have $\lambda_{k+1}^\star \approx 1$ for large

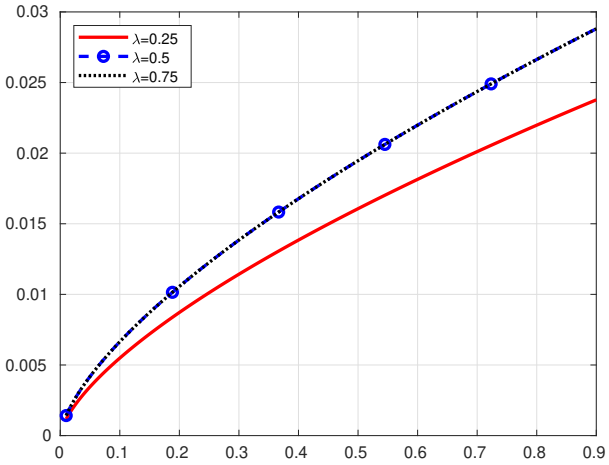

Fig. 1 For $\lambda \in \{0.25, 0.5, 0.75\}$ and $\mu \in (0.01, 0.9)$, evolution of the constant C solving (17)

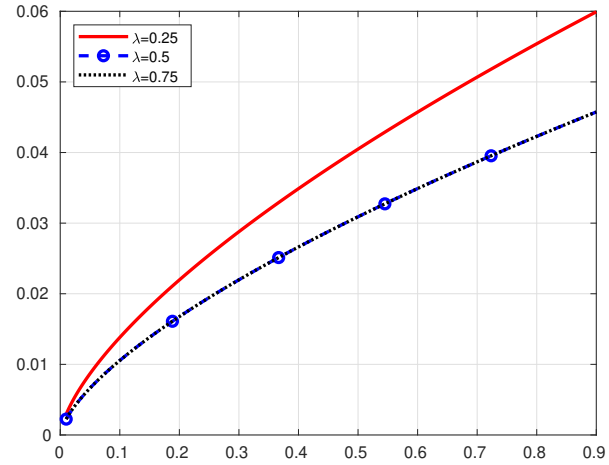

Fig. 2 For $\lambda \in \{0.25, 0.5, 0.75\}$ and $\mu \in (0.01, 0.9)$, evolution of the constant C solving (23)

iteration index k: `FIEM` and `opt-FIEM` are expected to be equivalent in the convergence phase. The ratio of the expectations $\mathbb{E}\left[\|\theta^k_{\text{opt\_FIEM}} - \theta_\star\|\right] / \mathbb{E}\left[\|\theta^k_{\text{alg}} - \theta_\star\|\right]$ and of the standard deviations $\text{std}(\|\theta^k_{\text{opt\_FIEM}} - \theta_\star\|)/\text{std}(\|\theta^k_{\text{alg}} - \theta_\star\|)$ are displayed on Figure 8 when alg is `FIEM` and `SA`. They are shown as a function of $k \in \{1e2, 5e2, 1e3, 1.5e3, \ldots, 6e3, 7e3, \ldots, 20e3\}$. The plot illustrates that if `opt-FIEM` and `FIEM` are equivalent in expectation, `opt-FIEM` surpasses `FIEM` in the transient phase by reducing the variance up to 22%. It also shows that `SA` has a really poor behavior w.r.t. `opt-FIEM` (and therefore also with `FIEM`) in the convergence

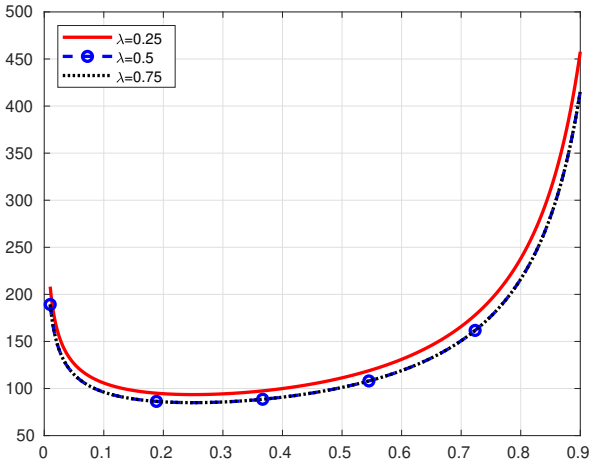

Fig. 3 For $\lambda \in \{0.25, 0.5, 0.75\}$ and $\mu \in (0.01, 0.9)$, evolution of the quantity $\mathcal{B}$ given by Proposition 4

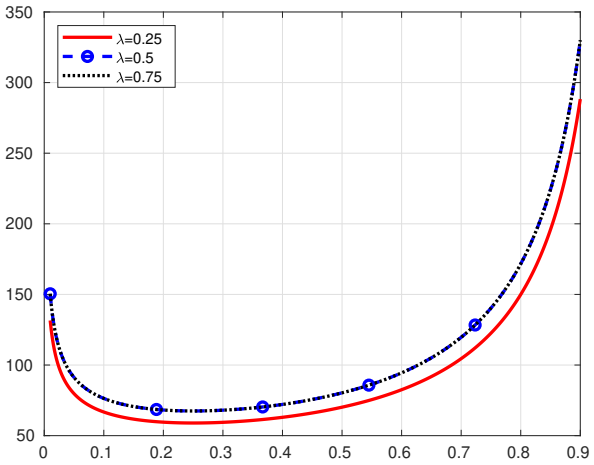

Fig. 4 For $\lambda \in \{0.25, 0.5, 0.75\}$ and $\mu \in (0.01, 0.9)$, evolution of the quantity $\mathcal{B}$ given by Proposition 5

phase, `SA` reduces the variability of `opt-FIEM` up to 18% in the transient phase, but `opt-FIEM` provides a drastic variability reduction in the first iterations. Since we advocate to stop `FIEM` at a random time K sampled in the range $\{0, \ldots, K_{\max} - 1\}$, `opt-FIEM` gives insights on how to improve the behavior of incremental EM algorithms in the transient phase. Figure 9 shows $k \mapsto \gamma_{k+1}^{-2} \mathbb{E}\left[\|\widehat{S}^{k+1} - \widehat{S}^{k}\|^2\right]$ for the three algorithms, in the transient phase $k \in [1.5e3, 5e3]$. The plot illustrates again

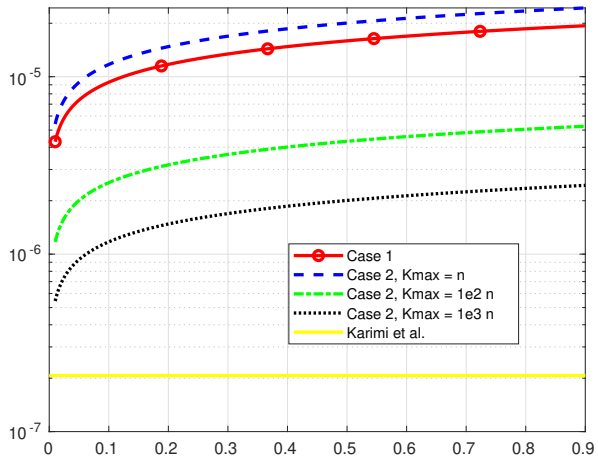

Fig. 5 Value of the constant step size given by Karimi et al., Proposition 4 (`Case 1`) and Proposition 5 (`Case 2`). The step size is shown as a function of $\mu \in (0.01, 0.9)$. In `Case 2`, different strategies for $K_{\max}$ are considered.

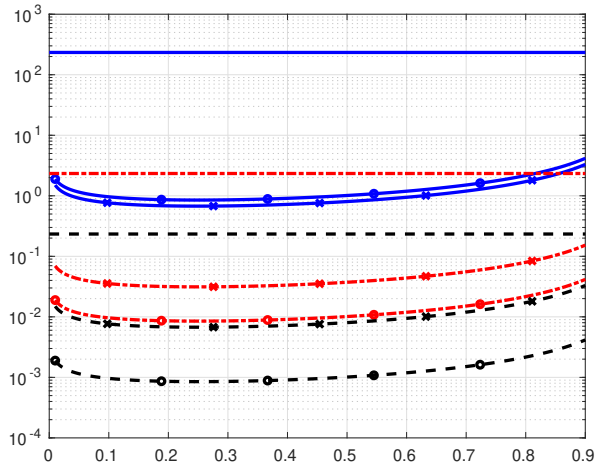

Fig. 6 Value of the control $n^{\mathsf{a}} K_{\max}^{-\mathsf{b}} \mathcal{B}$ given by Proposition 4 (`Case 1`, with a circle), Proposition 5 (`Case 2`, with a cross) and Karimi et al. (no markers). The control is displayed as a function of $\mu \in (0.01, 0.9)$ and for different values of $K_{\max}$: $K_{\max} = n$ (solid line), $K_{\max} = 1e2\, n$ (dash-dot line) and $K_{\max} = 1e3\, n$ (dashed line).

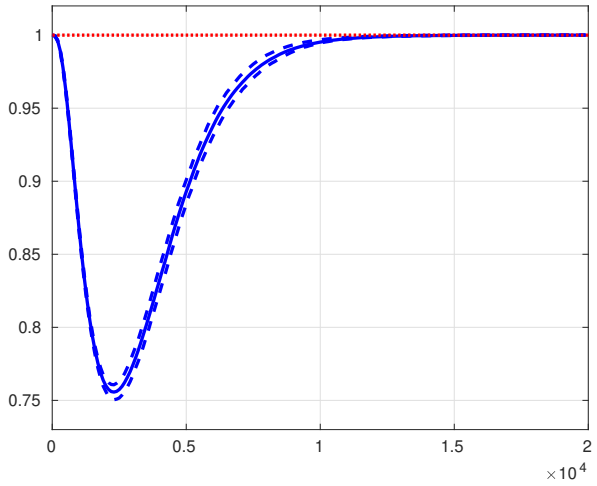

Fig. 7 The coefficient $\lambda_k^\star$ (see (13)) as a function of the number of iterations k; it is a random variable, and the solid line is the mean value (the dashed lines are resp. the quantiles 0.25 and 0.75) over $1e3$ independent paths.

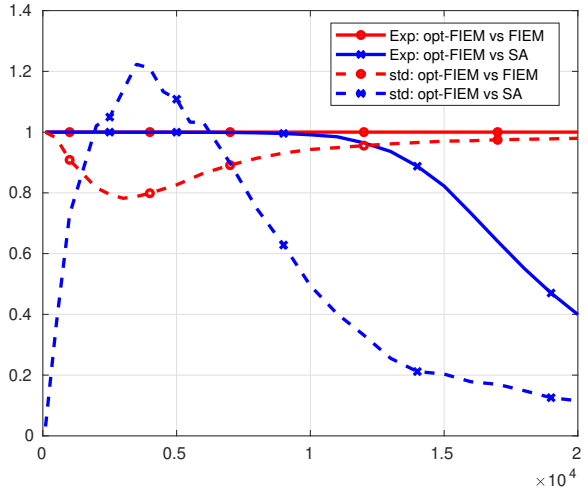

Fig. 8 For $k \in \{1e2, 5e2, 1e3, 1.5e3, \ldots, 6e3, 7e3, \ldots, 20e3\}$, ratio of the expectations (Exp) $\mathbb{E}\left[\|\theta^k_{\mathrm{opt-FIEM}} - \theta_\star\|\right] / \mathbb{E}\left[\|\theta^k_{\mathrm{Alg}} - \theta_\star\|\right]$ when Alg is FIEM (solid line with circle) and then `SA` (solid line with cross); and the standard deviations (std) $\mathrm{std}(\|\theta^k_{\mathrm{opt_FIEM}} - \theta_\star\|)/\mathrm{std}(\|\theta^k_{\mathrm{FIEM}} - \theta_\star\|)$ when Alg is `FIEM` (dashed line with circle) and then `SA` (dashed line with cross). The expectations and standard deviations are approximated by a Monte Carlo sum over $1e3$ independent runs.

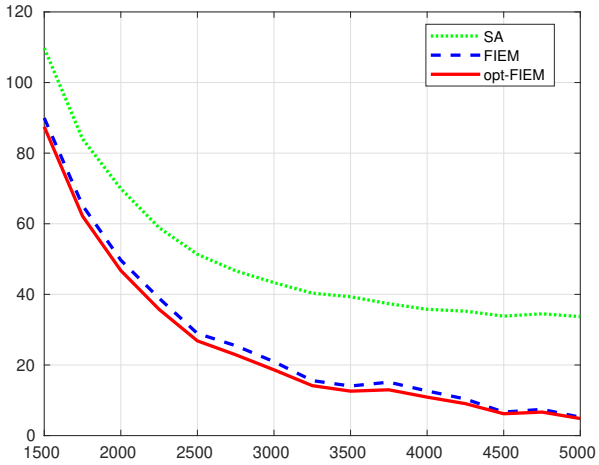

Fig. 9 Monte Carlo approximation (over 1e3 independent runs) of $k \mapsto \gamma_{k+1}^{-2}\mathbb{E}\left[\|\widehat{S}^{k+1} - \widehat{S}^k\|^2\right]$ for `SA`, `FIEM` and `opt-FIEM`.

that `opt-FIEM` improves on `FIEM` during this phase of the algorithm; and improves drastically on `SA`.

5 Mixture of Gaussian distributions

In this section, `FIEM` is applied to solve the Maximum Likelihood inference in a multidimensional mixture of Gaussian distributions (see Frühwirth-Schnatter et al. (2019) for a recent review on mixture models): given n $\mathbb{R}^d$-valued observations $y_1, \ldots, y_n$, find a point $\hat{\theta}_n^{\mathrm{ML}} \in \Theta$ satisfying $\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i(\hat{\theta}_n^{\mathrm{ML}}) \geq \frac{1}{n}\sum_{i=1}^n \mathcal{L}_i(\theta)$ for any $\theta \in \Theta$ where

$$\exp\left(\mathcal{L}_i(\theta)\right) \stackrel{\mathrm{def}}{=} \sum_{\ell=1}^{L} \alpha_\ell \, \mathcal{N}_d(\mu_\ell, \Sigma_\ell)[y_i] \, ,$$

and

$$\Theta \stackrel{\mathrm{def}}{=} \left\{ \alpha_\ell \geq 0, \sum_{\ell=1}^{L} \alpha_\ell = 1 \right\} \times \mathbb{R}^{dL} \times (\mathcal{M}_+^d)^L \, ;$$

$\mathcal{M}_+^d$ denotes the invertible $d \times d$ covariance matrices. In this example, we have $\mathcal{L}_i(\theta) = -\log \sum_{z=1}^L \exp(\langle s_i(z), \phi(\theta)\rangle)$ with $s_i = (s_i^{(1)}, s_i^{(2)}, s_i^{(3)})$ given by

$$\begin{aligned}
s_i^{(1)}(z) &\stackrel{\mathrm{def}}{=} (\mathbb{1}_{z=1}, \ldots, \mathbb{1}_{z=L}) \, , \\
s_i^{(2)}(z) &\stackrel{\mathrm{def}}{=} (y_i \mathbb{1}_{z=1}, \ldots, y_i \mathbb{1}_{z=L}) \, , \\
s_i^{(3)}(z) &\stackrel{\mathrm{def}}{=} \left(y_i y_i^T \mathbb{1}_{z=1}, \ldots, y_i y_i^T \mathbb{1}_{z=L}\right) \, .
\end{aligned}$$

The data set is the MNIST one [2]. The data are pre-processed as follows (we follow the same protocol as in Nguyen et al. (2020)): the training set contains $n = 6e4$ pictures of size 28×28; among these 784 pixels, 67 are non informative since they are constant over all the pictures, and they are removed thus yielding to n observations of length 717; each feature is centered and standardized (among the n observations) and a PCA is applied in order to summarize the features by the first $d = 20$ principal components.

The data set Décrire comment il est obtenu

The optimization problem to be solved is donner l'expression de F

$$XXX$$

By using the well known derivations of E step and M step in this model reference, `FIEM` is described by **??**. insérer la description algorithmique de FIEM pour ce modèle, avec un paramètre λ pour dire qu'on va explorer $\lambda = 0$, $\lambda = \hat{\lambda}_k$ et $\lambda = 1$.

This model indiquer quelles hypothèses sont vérifiées. Dire que "tant pis" pour celles qui ne le sont pas. Que l'objectif de cette section est d'explorer au-delà de la théorie du papier, notamment en proposant (a) une technique pour démarrer en SA, puis switcher en FIEM à un moment adapté; (b) une technique pour approcher $\hat{\lambda}_k$.

Description d'une technique d'estimation de λ_k

Description des analyses menées et des résultats obtenus

Des éléments de biblio Rahmani et al. (2020) ajuste un modèle de mélange de gaussiennes par un algo qui n'est pas EM: mais un gradient de type "alternating gradient descent" - grosso modo, c'est une technique d'estimation par méthodes des moments qui exploite les moments d'ordre 2 et 3 (d'où tenseur). Ils appliquent leur algo pour des données en batch, mais aussi des données en streaming. Ils affirment montrer empiriquement que leur algo fait mieux que l'EM. Ils prennent un mélange multi-dim, à matrice de covariance $\sigma_k^2 \mathrm{I}$.

Zhao et al. (2020) établit l'identifiabilité des modèles de mélange gaussien, dans le cas où les poids et les variances des gaussiennes sont connus. Seules les moyennes sont estimées, en multi-dimensionnel. Ils établissent d'abord un résultat sur la version population (la moyenne sur les n dans la formule d'update est remplacée par une espérance), puis reviennent à la version "échantillon" par des arguments de concentration. Ils prouvent que si sous la vraie loi (la loi "population"), les moyennes sont suffisamment éloignées selon un critère qui dépend de la dimension et du nombre de composantes dans le mélange notamment: alors l'algorithme va produire des points qui rentrent dans des voisinages de ces vraies moyennes. Ils explorent numériquement le cas où les poids et les covariances ne sont pas connus.

[2] available at http://yann.lecun.com/exdb/mnist/

5.1 For the supplementary material

5.1.1 The expression of $\mathcal{L}_i$

Set $\Gamma_\ell \stackrel{\text{def}}{=} \Sigma_\ell^{-1}$. We write, up to the constant $\sqrt{2\pi}^{-d}$,

$$\begin{aligned}\sum_{\ell=1}^{L} \alpha_\ell \mathcal{N}_d(\mu_\ell, \Sigma_\ell)[y] &= \sum_{z=1}^{L} \alpha_z \sqrt{\det(\Gamma_z)} \exp\left(-\frac{1}{2}(y-\mu_z)^T \Gamma_z (y-\mu_z)\right) \\ &= \sum_{z=1}^{L} \exp\left(\ln \alpha_z + \frac{1}{2}\ln\det(\Gamma_z) - \frac{1}{2}(y-\mu_z)^T \Gamma_z (y-\mu_z)\right) \\ &= \sum_{z=1}^{L} \exp\left(\sum_{\ell=1}^{L} 1_{z=\ell}\left\{\ln \alpha_\ell + \frac{1}{2}\ln\det(\Gamma_\ell) - \frac{1}{2}(y-\mu_\ell)^T \Gamma_\ell (y-\mu_\ell)\right\}\right)\end{aligned}$$

Therefore, $\mathcal{L}_i(\theta) = -\ln \sum_{z=1}^{L} \exp\left(\langle s_i(z), \phi(\theta)\rangle\right)$ with $s_i = (s_i^{(1)}, s_i^{(2)}, s_i^{(3)})$ and $\phi = (\phi^{(1)}, \phi^{(2)}, \phi^{(3)})$ given by

$$\begin{aligned}
&s_i^{(1)}(z) \stackrel{\text{def}}{=} (\mathbb{1}_{z=1}, \ldots, \mathbb{1}_{z=L}) \ , \\
&s_i^{(2)}(z) \stackrel{\text{def}}{=} (y_i \mathbb{1}_{z=1}, \ldots, y_i \mathbb{1}_{z=L}) \ , \\
&s_i^{(3)}(z) \stackrel{\text{def}}{=} \left(y_i y_i^T \mathbb{1}_{z=1}, \ldots, y_i y_i^T \mathbb{1}_{z=L}\right) \ , \\
&\phi^{(1)}(\theta) \stackrel{\text{def}}{=} \ln \alpha_\ell + 0.5 \ln\det(\Gamma_\ell) - 0.5 \mu_\ell^T \Gamma_\ell \mu_\ell, \quad \ell = 1, \ldots, L, \\
&\phi^{(2)}(\theta) \stackrel{\text{def}}{=} \Gamma_\ell \mu_\ell \quad \ell = 1, \ldots, L, \\
&\phi^{(3)}(\theta) \stackrel{\text{def}}{=} -\frac{1}{2}\Gamma_\ell \ .
\end{aligned}$$

Remember that $y^T A y = \mathrm{Tr}(A y y^T)$ is the scalar product between A and yy^T.

5.1.2 The expression of $p_i(z, \theta)$ and $\bar{s}_i(\theta)$

We have for any $u \in \{1, \ldots, L\}$,

$$p_i(u, \theta) \stackrel{\text{def}}{=} \frac{\alpha_u \mathcal{N}_d(\mu_u, \Sigma_u)[y_i]}{\sum_{\ell=1}^{L} \alpha_\ell \mathcal{N}_d(\mu_\ell, \Sigma_\ell)[y_i]}$$

so that $\bar{s}_i(\theta) = (\bar{s}_i^{(1)}(\theta), \bar{s}_i^{(2)}(\theta), \bar{s}_i^{(3)}(\theta))$ with

$$\begin{aligned}
&\bar{s}_i^{(1)}(z) \stackrel{\text{def}}{=} (p_i(1,\theta), \ldots, p_i(L,\theta)) \ , \\
&\bar{s}_i^{(2)}(z) \stackrel{\text{def}}{=} (y_i p_i(1,\theta), \ldots, y_i p_i(L,\theta)) \ , \\
&\bar{s}_i^{(3)}(z) \stackrel{\text{def}}{=} \left(y_i y_i^T p_i(1,\theta), \ldots, y_i y_i^T p_i(L,\theta)\right) \ .
\end{aligned}$$

5.1.3 The expression of T

We have for any symmetric matrix H

$$\ln \frac{\det(\Gamma + H)}{\det(\Gamma)} = \ln \det(I + \Gamma^{-1} H) = \ln(1 + \mathrm{Tr}(\Gamma^{-1} H) + o(\|H\|))$$
$$= \mathrm{Tr}(\Gamma^{-1} H) + o(\|H\|) = \left\langle H, \Gamma^{-1} \right\rangle + o(\|H\|)$$

$\mathsf{T}(s)$ depends on Γ_ℓ through the function

$$G(\Gamma_\ell) \stackrel{\text{def}}{=} \frac{s_\ell^{(1)}}{2} \ln \det(\Gamma_\ell) - \frac{1}{2} \left\langle \Gamma_\ell, s_\ell^{(3)} + s_\ell^{(1)} \mu_\ell \mu_\ell^T \right\rangle + \left\langle \Gamma_\ell, \mu_\ell (s_\ell^{(2)})^T \right\rangle \ .$$

Therefore

$$G(\Gamma_\ell + H) - G(\Gamma_\ell) = \frac{s_\ell^{(1)}}{2} \left\langle H, \Gamma^{-1} \right\rangle - \frac{1}{2} \left\langle H, s_\ell^{(3)} + s_\ell^{(1)} \mu_\ell \mu_\ell^T \right\rangle + \left\langle H, \mu_\ell (s_\ell^{(2)})^T \right\rangle + o(\|H\|) \ .$$

This yields as an update

$$\Sigma_\ell = \Gamma_\ell^{-1} = \frac{1}{s_\ell^{(1)}} \left(s_\ell^{(3)} + s_\ell^{(1)} \mu_\ell \mu_\ell^T - 2 \mu_\ell \, (s_\ell^{(2)})^T \right) \ .$$

Let us write $\langle s, \phi(\theta) \rangle = \sum_{j=1}^3 \left\langle s^{(j)}, \phi^{(j)}(\theta) \right\rangle$; we obtain $\mathsf{T}(s) = \{\alpha_\ell, \mu_\ell, \Sigma_\ell; \ell = 1, \ldots, L\}$ with

$$\alpha_\ell \stackrel{\text{def}}{=} \frac{s_\ell^{(1)}}{\sum_{u=1}^L s_u^{(1)}} \ ,$$
$$\mu_\ell \stackrel{\text{def}}{=} \frac{s_\ell^{(2)}}{s_\ell^{(1)}} \ ,$$
$$\Sigma_\ell \stackrel{\text{def}}{=} \frac{1}{s_\ell^{(1)}} \left(s_\ell^{(3)} + s_\ell^{(1)} \mu_\ell \mu_\ell^T - 2 \mu_\ell \, (s_\ell^{(2)})^T \right) = \frac{1}{s_\ell^{(1)}} \left(s_\ell^{(3)} - s_\ell^{(1)} \mu_\ell \, \mu_\ell^T \right) \ .$$

5.1.4 The EM algorithm in the parameter space.

Given the current value of the parameter θ_{curr}, compute the a posteriori distribution for all $i = 1, \ldots, n$ and $u = 1, \ldots, L$,

$$p_i(u, \theta_{\text{curr}}) \stackrel{\text{def}}{=} \frac{\alpha_u \mathcal{N}_d(\mu_u, \Sigma_u)[y_i]}{\sum_{\ell=1}^L \alpha_\ell \mathcal{N}_d(\mu_\ell, \Sigma_\ell)[y_i]} \ ;$$

then, compute the statistics for $\ell = 1, \ldots, L$

$$\bar{s}^{(1)}(\theta_{\text{curr}}) \stackrel{\text{def}}{=} \left(n^{-1} \sum_{i=1}^n p_i(u, \theta_{\text{curr}}), \qquad u = 1, \ldots, L \right) \ ,$$
$$\bar{s}^{(2)}(\theta_{\text{curr}}) \stackrel{\text{def}}{=} \left(n^{-1} \sum_{i=1}^n p_i(u, \theta_{\text{curr}}) y_i, \qquad u = 1, \ldots, L \right) \ ,$$
$$\bar{s}^{(3)}(\theta_{\text{curr}}) \stackrel{\text{def}}{=} \left(n^{-1} \sum_{i=1}^n p_i(u, \theta_{\text{curr}}) y_i y_i^T, \qquad u = 1, \ldots, L \right) \ .$$

Update the parameters by

$$\begin{aligned}
\alpha_\ell &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} p_i(\ell,\theta_{\text{curr}}) \ , \\
\mu_\ell &\stackrel{\text{def}}{=} \frac{\sum_{i=1}^{n} p_i(\ell,\theta_{\text{curr}}) y_i}{\sum_{i=1}^{n} p_i(\ell,\theta_{\text{curr}})} = \frac{1}{\alpha_\ell\, n}\sum_{i=1}^{n} p_i(\ell,\theta_{\text{curr}}) y_i \ , \\
\Sigma_\ell &\stackrel{\text{def}}{=} \frac{1}{\alpha_\ell\, n}\sum_{i=1}^{n} p_i(\ell,\theta_{\text{curr}})(y_i-\mu_\ell)(y_i-\mu_\ell)^T \ .
\end{aligned}$$

5.1.5 The SA (Online-EM) algorithm in the parameter space

Input: the current value of the parameter θ_{curr} and the current value of the statistics $\widehat{S}_{\text{curr}} = (\widehat{S}^{(1)}_{\text{curr}}, \widehat{S}^{(2)}_{\text{curr}}, \widehat{S}^{(3)}_{\text{curr}})$ where

$$\widehat{S}^{(1)}_{\text{curr}} \in \mathbb{R}^L, \qquad \widehat{S}^{(2)}_{\text{curr}} \in (\mathbb{R}^d)^L, \qquad \widehat{S}^{(3)}_{\text{curr}} \in (\mathbb{R}^{d\times d})^L \ .$$

Compute the a posteriori distribution for all $i = 1,\dots,n$ and $u = 1,\dots,L$,

$$p_i(u,\theta_{\text{curr}}) \stackrel{\text{def}}{=} \frac{\alpha_u \mathcal{N}_d(\mu_u,\Sigma_u)[y_i]}{\sum_{\ell=1}^{L}\alpha_\ell \mathcal{N}_d(\mu_\ell,\Sigma_\ell)[y_i]} \ ;$$

sample at random $I \in \{1,\dots,n\}$ and compute the statistics

$$\begin{aligned}
\widehat{S}^{(1)} &\stackrel{\text{def}}{=} (1-\gamma)\widehat{S}^{(1)}_{\text{curr}} + \gamma\left(p_I(1,\theta_{\text{curr}}),\dots,p_I(L,\theta_{\text{curr}})\right) \ , \\
\widehat{S}^{(2)} &\stackrel{\text{def}}{=} (1-\gamma)\widehat{S}^{(2)}_{\text{curr}} + \gamma\left(y_I p_I(1,\theta_{\text{curr}}),\dots,y_I p_I(L,\theta_{\text{curr}})\right) \ , \\
\widehat{S}^{(3)} &\stackrel{\text{def}}{=} (1-\gamma)\widehat{S}^{(3)}_{\text{curr}} + \gamma\left(y_I y_I^T p_I(1,\theta_{\text{curr}}),\dots,y_I y_I^T p_I(L,\theta_{\text{curr}})\right) \ .
\end{aligned}$$

Update the parameters by

$$\begin{aligned}
\alpha_\ell &\stackrel{\text{def}}{=} \frac{\widehat{S}^{(1)}_\ell}{\sum_{u=1}^{L}\widehat{S}^{(1)}_u} \ , \\
\mu_\ell &\stackrel{\text{def}}{=} \frac{\widehat{S}^{(2)}_\ell}{\widehat{S}^{(1)}_\ell} \ , \\
\Sigma_\ell &\stackrel{\text{def}}{=} \frac{1}{\widehat{S}^{(1)}_\ell}\left(\widehat{S}^{(3)}_\ell - \widehat{S}^{(1)}_\ell \mu_\ell \mu_\ell^T\right) \ .
\end{aligned}$$

5.1.6 The iEM algorithm in the parameter space

Input:

- the current value of the parameter θ_{curr}
- the current value of the statistics $\widehat{S}_{\text{curr}} = (\widehat{S}^{(1)}_{\text{curr}}, \widehat{S}^{(2)}_{\text{curr}}, \widehat{S}^{(3)}_{\text{curr}})$ where

$$\widehat{S}^{(1)}_{\text{curr}} \in \mathbb{R}^L, \qquad \widehat{S}^{(2)}_{\text{curr}} \in (\mathbb{R}^d)^L, \qquad \widehat{S}^{(3)}_{\text{curr}} \in (\mathbb{R}^{d\times d})^L \ .$$

– the current memory vector $\mathsf{S}_{\rm curr} = (\mathsf{S}^{(1)}_{\rm curr}, \mathsf{S}^{(2)}_{\rm curr}, \mathsf{S}^{(3)}_{\rm curr})$
– the current mean of this vector $\widetilde{S}_{\rm curr}$.

Compute the a posteriori distribution for all $i = 1, \ldots, n$ and $u = 1, \ldots, L$,

$$p_i(u, \theta_{\rm curr}) \stackrel{\rm def}{=} \frac{\alpha_u \mathcal{N}_d(\mu_u, \Sigma_u)[y_i]}{\sum_{\ell=1}^{L} \alpha_\ell \mathcal{N}_d(\mu_\ell, \Sigma_\ell)[y_i]} \; ;$$

Sample at random $I \in \{1, \ldots, n\}$ and update the `memory` quantities: for $i \neq I$, $\mathsf{S}_i = \mathsf{S}_{{\rm curr},i}$ and otherwise

$$\begin{aligned}
\mathsf{S}^{(1)}_I &\stackrel{\rm def}{=} (p_I(u, \theta_{\rm curr}), \qquad u = 1, \ldots, L) \; , \\
\mathsf{S}^{(2)}_I &\stackrel{\rm def}{=} (p_I(u, \theta_{\rm curr}) y_I, \qquad u = 1, \ldots, L) \; , \\
\mathsf{S}^{(3)}_I &\stackrel{\rm def}{=} \left(p_I(u, \theta_{\rm curr}) y_I y_I^T, \qquad u = 1, \ldots, L \right) \; ;
\end{aligned}$$

update the mean

$$\widetilde{S} \stackrel{\rm def}{=} \widetilde{S}_{\rm curr} + \frac{1}{n} \left(\mathsf{S}_I - \mathsf{S}_{{\rm curr},I} \right) \; .$$

Update the statistics

$$\widehat{S} \stackrel{\rm def}{=} (1 - \gamma) \widehat{S}_{\rm curr} + \gamma \widetilde{S} \; .$$

Update the parameters by

$$\begin{aligned}
\alpha_\ell &\stackrel{\rm def}{=} \frac{\widehat{S}^{(1)}_\ell}{\sum_{u=1}^{L} \widehat{S}^{(1)}_u} \; , \\
\mu_\ell &\stackrel{\rm def}{=} \frac{\widehat{S}^{(2)}_\ell}{\widehat{S}^{(1)}_\ell} \; , \\
\Sigma_\ell &\stackrel{\rm def}{=} \frac{1}{\widehat{S}^{(1)}_\ell} \left(\widehat{S}^{(3)}_\ell - \widehat{S}^{(1)}_\ell \mu_\ell \mu_\ell^T \right) \; .
\end{aligned}$$

5.1.7 The FIEM algorithm in the parameter space

Input:

– the current value of the parameter $\theta_{\rm curr}$
– the current value of the statistics $\widehat{S}_{\rm curr} = (\widehat{S}^{(1)}_{\rm curr}, \widehat{S}^{(2)}_{\rm curr}, \widehat{S}^{(3)}_{\rm curr})$ where

$$\widehat{S}^{(1)}_{\rm curr} \in \mathbb{R}^L, \qquad \widehat{S}^{(2)}_{\rm curr} \in (\mathbb{R}^d)^L, \qquad \widehat{S}^{(3)}_{\rm curr} \in (\mathbb{R}^{d \times d})^L \; .$$

– the current memory vector $\mathsf{S}_{\rm curr} = (\mathsf{S}^{(1)}_{\rm curr}, \mathsf{S}^{(2)}_{\rm curr}, \mathsf{S}^{(3)}_{\rm curr})$
– the current mean of this vector $\widetilde{S}_{\rm curr}$.

Compute the a posteriori distribution for all $i = 1, \dots, n$ and $u = 1, \dots, L$,

$$p_i(u, \theta_{\text{curr}}) \stackrel{\text{def}}{=} \frac{\alpha_u \mathcal{N}_d(\mu_u, \Sigma_u)[y_i]}{\sum_{\ell=1}^{L} \alpha_\ell \mathcal{N}_d(\mu_\ell, \Sigma_\ell)[y_i]} \ ;$$

Sample at random $I \in \{1, \dots, n\}$ and update the `memory` quantities: for $i \neq I$, $\mathsf{S}_i = \mathsf{S}_{\text{curr},i}$ and otherwise

$$\begin{aligned}
\mathsf{S}_I^{(1)} &\stackrel{\text{def}}{=} (p_I(u, \theta_{\text{curr}}), \qquad u = 1, \dots, L) \ , \\
\mathsf{S}_I^{(2)} &\stackrel{\text{def}}{=} (p_I(u, \theta_{\text{curr}}) y_I, \qquad u = 1, \dots, L) \ , \\
\mathsf{S}_I^{(3)} &\stackrel{\text{def}}{=} \left(p_I(u, \theta_{\text{curr}}) y_I y_I^T, \qquad u = 1, \dots, L \right) \ ;
\end{aligned}$$

update the mean

$$\widetilde{S} \stackrel{\text{def}}{=} \widetilde{S}_{\text{curr}} + \frac{1}{n} \left(\mathsf{S}_I - \mathsf{S}_{\text{curr},I} \right) \ .$$

Sample J at random in $\{1, \dots, n\}$. Compute

$$\begin{aligned}
\bar{s}_J^{(1)} &\stackrel{\text{def}}{=} (p_J(u, \theta_{\text{curr}}), \qquad u = 1, \dots, L) \ , \\
\bar{s}_J^{(2)} &\stackrel{\text{def}}{=} (p_J(u, \theta_{\text{curr}}) y_J, \qquad u = 1, \dots, L) \ , \\
\bar{s}_J^{(3)} &\stackrel{\text{def}}{=} \left(p_J(u, \theta_{\text{curr}}) y_J y_J^T, \qquad u = 1, \dots, L \right) \ ;
\end{aligned}$$

Update the statistics

$$\widehat{S} \stackrel{\text{def}}{=} \widehat{S}_{\text{curr}} + \gamma \left(\bar{s}_J - \widehat{S}_{\text{curr}} + \widetilde{S} - \mathsf{S}_J \right) \ .$$

Update the parameters by

$$\begin{aligned}
\alpha_\ell &\stackrel{\text{def}}{=} \frac{\widehat{S}_\ell^{(1)}}{\sum_{u=1}^{L} \widehat{S}_u^{(1)}} \ , \\
\mu_\ell &\stackrel{\text{def}}{=} \frac{\widehat{S}_\ell^{(2)}}{\widehat{S}_\ell^{(1)}} \ , \\
\Sigma_\ell &\stackrel{\text{def}}{=} \frac{1}{\widehat{S}_\ell^{(1)}} \left(\widehat{S}_\ell^{(3)} - \widehat{S}_\ell^{(1)} \mu_\ell \mu_\ell^T \right) \ .
\end{aligned}$$

6 Proof

6.1 Proof of section 2

6.1.1 Proof of Proposition 1

(proof of item 1). From the Jensen's inequality, it holds

$$\mathcal{L}_i(\theta) - \mathcal{L}_i(\theta') \leq - \int_{\mathsf{Z}} \left\langle s_i(z), \phi(\theta) - \phi(\theta') \right\rangle \, p_i(z; \theta') \, \mu(\mathrm{d}z) = - \left\langle \bar{s}_i(\theta'), \phi(\theta) - \phi(\theta') \right\rangle \ ;$$

which concludes the proof. *(proof of item 2)* From (3) and item 1, it holds

$$F(\theta) \leq -\left\langle \bar{s}(\theta'), \phi(\theta) \right\rangle + \frac{1}{n}\sum_{i=1}^{n} \mathcal{C}_i(\theta') + \mathsf{R}(\theta) \; .$$

(proof of item 3) From item 2 and the definition of T, it holds

$$F(\mathsf{T} \circ \bar{s}(\theta^k)) \leq \overline{F}(\mathsf{T} \circ \bar{s}(\theta^k), \theta^k) \leq \overline{F}(\theta^k, \theta^k) = F(\theta^k) \; .$$

6.1.2 Proof of Proposition 2

(Proof of item 1). The statements are trivial and we only prove the first claim: if $s^\star = \bar{s} \circ \mathsf{T}(s^\star)$ then by applying T (under the uniqueness assumption H3), we have $\mathsf{T}(s^\star) = (\mathsf{T} \circ \bar{s}) \circ \mathsf{T}(s^\star)$ and the proof follows.

(Proof of item 2). For $\theta \in \Theta^v$, set $D\phi(\theta) \stackrel{\text{def}}{=} \left(\dot{\phi}(\theta)\right)^T$. By H4-item 2 and a chain rule,

$$\dot{V}(s) = \left(\dot{\mathsf{T}}(s)\right)^T \left\{\dot{\mathsf{R}}(\mathsf{T}(s)) - D\phi(\mathsf{T}(s))\; \bar{s} \circ \mathsf{T}(s)\right\} \; .$$

Moreover, using H3 and H4-item 1, the minimum $\mathsf{T}(s)$ is a critical point of $\theta \mapsto -\left\langle s, \phi(\theta) \right\rangle + \mathsf{R}(\theta)$: we have for any $s \in \mathbb{R}^q$, $\dot{\mathsf{R}}(\mathsf{T}(s)) - D\phi(\mathsf{T}(s))\, s = 0$. Hence,

$$\dot{V}(s) = -\left(\dot{\mathsf{T}}(s)\right)^T \; D\phi(\mathsf{T}(s))\; h(s) = -\left(B(s)\right)^T \; h(s) \; .$$

H4-item 3 implies that $B^T = B$ and the zeros of h are the zeros of $\dot{V}$.

6.1.3 Auxiliary results

Lemma 1 *Assume that Θ and $\phi(\Theta)$ are open; and ϕ is continuously differentiable on Θ. Then for all $i \in \{1, \ldots, n\}$, $\mathcal{L}_i$ is continuously differentiable on Θ.*

If in addition H1, H2, H3 and H4-item 1 hold, then F (resp. $V \stackrel{\text{def}}{=} F \circ \mathsf{T}$) is continuously differentiable on Θ (resp. on $\mathbb{R}^q$) and for any $\theta \in \Theta$,

$$\dot{F}(\theta) = -\left(\dot{\phi}(\theta)\right)^T \; \bar{s}(\theta) + \dot{\mathsf{R}}(\theta) \; .$$

Proof H1 and (Sundberg, 2019, Proposition 3.8) (see also (Brown, 1986, Theorem 2.2.)) imply that $L_i : \tau \mapsto \int_Z h_i(z) \exp\left(\left\langle s_i(z), \tau \right\rangle\right) \mu(\mathrm{d}z)$ is continuously differentiable on the interior of the set $\{\tau \in \mathbb{R}^q, \int_Z h_i(z) \exp\left(\left\langle s_i(z), \tau \right\rangle\right) \mu(\mathrm{d}z) < \infty\}$ and its derivative is

$$\int_{\mathsf{Z}} s_i(z)\; h_i(z) \exp\left(\left\langle s_i(z), \tau \right\rangle\right) \mu(\mathrm{d}z) \; .$$

This set contains $\phi(\Theta)$ under H1. The equality $\mathcal{L}_i = -\log(L_i \circ \phi)$ and the differentiability of composition of functions conclude the proof of the first item. The second one easily follows.

pour Pierre: ne faut-il pas que $\phi(\Theta)$ soit un ouvert ? Personnellement, je ne sais faire la démo que si je peux caser une variation ϵ autour du point en lequel je dérive, pour contrôler des exponentielles. Le théo 2.2. de Brown ne s'applique qu'en un point de l'intérieur. Je n'ai pas accès électronique au livre de Sundberg, je ne peux donc pas vérifier la référence. J'ai donc rajouté les hypothèses en conséquence et modifié l'ensemble des résultats basés sur ce lemme.

Lemma 2 *Assume H1 and H3. Assume also that for any* $s \in \mathbb{R}^q$, $\tau \mapsto \mathsf{Q}(s,\tau) \stackrel{\text{def}}{=} -\langle s, \phi(\tau)\rangle + \mathsf{R}(\tau)$ *is twice continuously differentiable on* Θ^v *where* $\Theta^v \stackrel{\text{def}}{=} \Theta$ *if* Θ *is open, or* Θ^v *is a neighborhood of* Θ *otherwise. Then for any* $s \in \mathbb{R}^q$, $\dot{\phi \circ \mathsf{T}}$ *is a symmetric* $q \times q$ *matrix satisfying*

$$\dot{\phi \circ \mathsf{T}}(s) = \left(\dot{\mathsf{T}}(s)\right)^T \ \partial_\tau^2 \mathsf{Q}(s, \mathsf{T}(s)) \ \left(\dot{\mathsf{T}}(s)\right) .$$

Proof The proof is adapted from the one of (Delyon et al., 1999, Lemma 2). H3 and the regularity conditions on Q imply that for any $s \in \mathbb{R}^q$:

$$\partial_\tau \mathsf{Q}(s, \mathsf{T}(s)) = -\left(\dot{\phi}(\mathsf{T}(s))\right)^T s + \dot{\mathsf{R}}(\mathsf{T}(s)) = 0$$

and (from the uniqueness assumption) $s \mapsto \partial_\tau^2 \mathsf{Q}(s, \mathsf{T}(s)) \ \dot{\mathsf{T}}(s) - \left(\dot{\phi}(\mathsf{T}(s))\right)^T$ is positive-definite. This concludes the proof.

Lemma 3 *Assume H1, H2 and H3. Assume in addition that (i) there exists* $L_{i,p}$ *such that for any* $\theta, \theta' \in \Theta$

$$\sup_{z \in \mathsf{Z}} \left| p_i(z;\theta) - p_i(z;\theta') \right| \leq L_{i,p} \ \|\theta - \theta'\|;$$

(ii) T *is globally Lipschitz on* $\mathbb{R}^q$, *and (iii)* $\int_{\mathsf{Z}} \|s_i\| \mathrm{d}\mu < \infty$. *Then there exists a constant* $0 < L_i < \infty$ *such that for all* $s, s' \in \mathbb{R}^q$, $\|\bar{s}_i \circ \mathsf{T}(s) - \bar{s}_i \circ \mathsf{T}(s')\| \leq L_i \|s - s'\|$.

Proof Let $s, s' \in \mathbb{R}^q$. We have

$$\bar{s}_i \circ \mathsf{T}(s) - \bar{s}_i \circ \mathsf{T}(s') = \int_{\mathsf{Z}} s_i(z) \left[p_i(z; \mathsf{T}(s)) - p_i(z; \mathsf{T}(s')) \right] \mu(\mathrm{d}z)$$

so that

$$\begin{aligned} \|\bar{s}_i \circ \mathsf{T}(s) - \bar{s}_i \circ \mathsf{T}(s')\| &\leq \int_{\mathsf{Z}} \|s_i(z)\| \left| p_i(z; \mathsf{T}(s)) - p_i(z; \mathsf{T}(s')) \right| \mu(\mathrm{d}z) \\ &\leq L_p \ \|\mathsf{T}(s) - \mathsf{T}(s')\| \ \int_{\mathsf{Z}} \|s_i(z)\| \mu(\mathrm{d}z) . \end{aligned}$$

6.2 Proofs of section 3

For any $k \geq 0$ and $i \in \{1, \ldots, n\}$, we define $\widehat{S}^{<k,i}$ such that

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

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

$$\widehat{S}^{<k+1,i} = \widehat{S}^{\ell}, \begin{cases} \ell = k & \text{if } I_{k+1} = i, \\ 1 \leq \ell \leq k-1 & \text{if } I_{k+1} \neq i, I_k \neq i, \ldots, I_{\ell+1} = i, \\ \ell = 0 & \text{otherwise.} \end{cases} \tag{26}$$

Define the filtrations, for $k \geq 0$,

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

note that $\widehat{S}^k \in \mathcal{F}_k$ and $\mathsf{S}_{k+1,\cdot} \in \mathcal{F}_{k+1/2}$. Set

$$H_{k+1} \stackrel{\text{def}}{=} \bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k + \frac{1}{n} \sum_{i=1}^{n} \mathsf{S}_{k+1,i} - \mathsf{S}_{k+1,J_{k+1}} \; .$$

6.2.1 Proof of Theorem 1

By Proposition 2 and H5-item 3, $\dot{V}$ is $L_{\dot{V}}$-Lipschitz, and we have

$$\begin{aligned} V(\widehat{S}^{k+1}) &\leq V(\widehat{S}^k) + \left\langle \widehat{S}^{k+1} - \widehat{S}^k, \dot{V}(\widehat{S}^k) \right\rangle + \frac{L_{\dot{V}}}{2} \|\widehat{S}^{k+1} - \widehat{S}^k\|^2 \\ &\leq V(\widehat{S}^k) + \gamma_{k+1} \left\langle H_{k+1}, \dot{V}(\widehat{S}^k) \right\rangle + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \|H_{k+1}\|^2. \end{aligned}$$

Taking the expectation yields, upon noting that $\widehat{S}^k \in \mathcal{F}_k$

$$\begin{aligned} &\mathbb{E}\left[V(\widehat{S}^{k+1})\right] - \mathbb{E}\left[V(\widehat{S}^k)\right] \\ &\leq \gamma_{k+1} \mathbb{E}\left[\left\langle \mathbb{E}\left[H_{k+1}|\mathcal{F}_k\right], \dot{V}(\widehat{S}^k) \right\rangle\right] + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|H_{k+1}\|^2\right] \\ &\leq \gamma_{k+1} \mathbb{E}\left[\left\langle h(\widehat{S}^k), \dot{V}(\widehat{S}^k) \right\rangle\right] + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|H_{k+1}\|^2\right] \\ &\leq -\gamma_{k+1} v_{min} \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|H_{k+1}\|^2\right] \\ &\leq -\gamma_{k+1} \left(v_{min} - \gamma_{k+1} \frac{L_{\dot{V}}}{2}\right) \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|H_{k+1} - h(\widehat{S}^k)\|^2\right] \end{aligned}$$

where we used that $\mathbb{E}\left[H_{k+1}|\mathcal{F}_k\right] = h(\widehat{S}^k)$ and Proposition 3. Set

$$A_k \stackrel{\text{def}}{=} \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] \; , \qquad B_{k+1} \stackrel{\text{def}}{=} \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^k)\|^2\right] \; .$$

By Lemma 4 and Proposition 7, we have for any $k \geq 0$:

$$\mathbb{E}\left[V(\widehat{S}^{k+1})\right] - \mathbb{E}\left[V(\widehat{S}^{k})\right] \leq -\gamma_{k+1}\left(v_{min} - \gamma_{k+1}\frac{L_{\dot{V}}}{2}\right) A_k - \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} B_{k+1} + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}}\|^2\right] \leq T_{1,k} + T_{2,k+1}$$

by setting

$$T_{1,k} \stackrel{\text{def}}{=} -\gamma_{k+1}\left(v_{min} - \gamma_{k+1}\frac{L_{\dot{V}}}{2}\right) A_k + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \sum_{j=0}^{k-1} \tilde{\Lambda}_{j+1,k} A_j$$

$$T_{2,k+1} \stackrel{\text{def}}{=} -\gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \left\{ B_{k+1} + \sum_{j=0}^{k-1} \tilde{\Lambda}_{j+1,k} (1+\beta_{j+1}^{-1})^{-1} B_{j+1} \right\} ;$$

by convention, $\sum_{j=0}^{1} a_j = 0$. By summing from $k=0$ to $k = K_{\max} - 1$, we have

$$\sum_{k=0}^{K_{\max}-1} \gamma_{k+1}\left(v_{min} - \gamma_{k+1}\frac{L_{\dot{V}}}{2}\right) A_k - \frac{L_{\dot{V}} L^2}{2} \sum_{k=0}^{K_{\max}-2} \gamma_{k+1}^2 \Lambda_k \; A_k \leq \Delta V - \frac{L_{\dot{V}}}{2} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1}^2 \left(L^2 \Xi_k + 1\right) B_{k+1},$$

where for $0 \leq k \leq K_{\max} - 2$ and with the convention $\Lambda_{K_{\max}-1} = \Xi_{K_{\max}-1} = 0$,

$$\Lambda_k \stackrel{\text{def}}{=} \left(1 + \frac{1}{\beta_{k+1}}\right) \sum_{j=k+1}^{K_{\max}-1} \gamma_{j+1}^2 \left(\frac{n-1}{n}\right)^{j-k} \prod_{\ell=k+2}^{j} \left(1 + \beta_\ell + \gamma_\ell^2 L^2\right)$$

$$\leq \left(1 + \frac{1}{\beta_{k+1}}\right) \sum_{j=k+1}^{K_{\max}-1} \gamma_{j+1}^2 \prod_{\ell=k+2}^{j} \left(1 - \frac{1}{n} + \beta_\ell + \gamma_\ell^2 L^2\right),$$

$$\Xi_k \stackrel{\text{def}}{=} \left(1 + \frac{1}{\beta_{k+1}}\right)^{-1} \Lambda_k = \frac{\Lambda_k \beta_{k+1}}{1 + \beta_{k+1}} .$$

Hence,

$$\sum_{k=0}^{K_{\max}-1} \{\gamma_{k+1}\left(v_{min} - \gamma_{k+1}\frac{L_{\dot{V}}}{2}\right) - \gamma_{k+1}^2 \Lambda_k \frac{L_{\dot{V}} L^2}{2}\} \; A_k + \sum_{k=0}^{K_{\max}-1} \gamma_{k+1}^2 \{1 + \Xi_k L^2\} \frac{L_{\dot{V}}}{2} B_{k+1} \leq \Delta V .$$

6.2.2 Proof of Proposition 4

It is a follow-up of Theorem 1; the quantities $\alpha_k, \Lambda_k, \delta_k$ introduced in the statement of Theorem 1 are used below without being defined again. We consider the case when for $\ell = 1, \ldots, K_{\max}$,

$$\beta_\ell \stackrel{\text{def}}{=} \frac{1-\lambda}{n^{\mathsf{b}}} , \qquad \gamma_\ell^2 \stackrel{\text{def}}{=} \frac{C}{L^2 n^{2\mathsf{c}} K_{\max}^{2\mathsf{d}}} ,$$

for some $\lambda \in (0,1)$, $C > 0$ and $\mathsf{b}, \mathsf{c}, \mathsf{d}$ to be defined in the proof in such a way that *(i)* $\alpha_k \geq 0$, *(ii)* $\sum_{k=0}^{K_{\max}-1} \alpha_k$ is positive and as large as possible. Since there will be a discussion on (n, C, λ), we make more explicit the dependence of some constants upon these quantities: α_k will be denoted by $\alpha_k(n, C, \lambda)$.

With theses definitions, we have

$$1 - \frac{\rho_n}{n} \stackrel{\text{def}}{=} 1 - \frac{1}{n} + \beta_\ell + \gamma_\ell^2 L^2 = 1 - \frac{1}{n}\left(1 - \frac{1-\lambda}{n^{\mathsf{b}-1}} - \frac{C}{n^{2\mathsf{c}-1} K_{\max}^{2\mathsf{d}}}\right) ,$$

and choose $(\mathsf{b}, \mathsf{c}, \mathsf{d}, \lambda, C)$ such that

$$\frac{1-\lambda}{n^{\mathsf{b}-1}} + \frac{C}{n^{2\mathsf{c}-1} K_{\max}^{2\mathsf{d}}} < 1 , \tag{27}$$

which ensures that $\rho_n \in (0,1)$. Hence, for any $0 \leq k \leq K_{\max} - 2$,

$$\begin{aligned} \Lambda_k &\leq n^{\mathsf{b}} \left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right) \frac{C}{L^2 n^{2\mathsf{c}} K_{\max}^{2\mathsf{d}}} \sum_{j=k+1}^{K_{\max}-1} \left(1 - \frac{\rho_n}{n}\right)^{j-k-1} \\ &\leq \left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right) \frac{C}{L^2 \rho_n} \frac{1}{n^{2\mathsf{c}-\mathsf{b}-1} K_{\max}^{2\mathsf{d}}} . \end{aligned}$$

From this upper bound, we deduce for any $0 \leq k \leq K_{\max} - 1$: $\alpha_k(n, C, \lambda) \geq \underline{\alpha}_n(C, \lambda)$ where

$$\begin{aligned} \underline{\alpha}_n(C, \lambda) \stackrel{\text{def}}{=} \frac{\sqrt{C}}{L n^{\mathsf{c}} K_{\max}^{\mathsf{d}}} \Bigg(v_{\min} - \frac{L_{\dot V}}{2L} \frac{\sqrt{C}}{n^{\mathsf{c}} K_{\max}^{\mathsf{d}}} \\ - \frac{L_{\dot V}}{2L} \frac{C^{3/2}}{\rho_n \, n^{3\mathsf{c}-\mathsf{b}-1} K_{\max}^{3\mathsf{d}}} \left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\Bigg) . \end{aligned} \tag{28}$$

From (27) and (28), we choose $\mathsf{b} = 1$, $\mathsf{c} = 2/3$, $\mathsf{d} = 0$; which yields for $n \geq 1$, since $\rho_n = \lambda - C n^{-1/3}$

$$n^{2/3} \underline{\alpha}_n(C, \lambda) \geq \mathcal{L}_n(C, \lambda) ,$$

with

$$\begin{aligned} \mathcal{L}_n(C, \lambda) &\stackrel{\text{def}}{=} \frac{L_{\dot V} \sqrt{C}}{2L^2} \left(v_{\min} \frac{2L}{L_{\dot V}} - \sqrt{C} f_n(C, \lambda) \right) , \\ f_n(C, \lambda) &\stackrel{\text{def}}{=} \frac{1}{n^{2/3}} + \frac{C}{\lambda - C n^{-1/3}} \left(\frac{1}{n} + \frac{1}{1-\lambda}\right) . \end{aligned}$$

Let $\mu \in (0,1)$. Fix $\lambda \in (0,1)$ and $C > 0$ such that (see (27) for the second condition)

$$\sqrt{C} f_n(C, \lambda) = 2\mu v_{\min} \frac{L}{L_{\dot V}} , \qquad \frac{1}{n^{1/3}} < \frac{\lambda}{C} . \tag{29}$$

This implies that $n^{2/3} \alpha_k(n, C, \lambda) \geq n^{2/3} \underline{\alpha}_n(C, \lambda) \geq n^{2/3} \alpha_\star(C) \stackrel{\text{def}}{=} \sqrt{C}(1-\mu) v_{\min}/L$. We obtain an upper bound on E_1 by

$$\mathsf{E}_1 \leq \frac{1}{K_{\max} \; \alpha_\star(C)} \sum_{k=0}^{K_{\max}-1} \alpha_k(n, C, \lambda) \, \mathbb{E}\left[\|h(\widehat{S}^k)\|^2 \right] .$$

For E_2, since $\delta_k \geq L_{\dot V}\gamma_{k+1}^2/2$,

$$\frac{L_{\dot V}\sqrt{C}}{2L(1-\mu)n^{2/3}}\frac{1}{v_{\min}}\mathsf{E}_2 \leq \frac{L_{\dot V}\,C}{2L^2n^{4/3}}\frac{1}{K_{\max}\;\alpha_\star(C)}\sum_{k=0}^{K_{\max}-1}\mathbb{E}\left[\|\widetilde{S}^{k+1}-\bar{s}\circ\mathsf{T}(\widehat{S}^k)\|^2\right]$$
$$\leq \frac{1}{K_{\max}\;\alpha_\star(C)}\sum_{k=0}^{K_{\max}-1}\delta_k\mathbb{E}\left[\|\widetilde{S}^{k+1}-\bar{s}\circ\mathsf{T}(\widehat{S}^k)\|^2\right]\;.$$

We then conclude by

$$\frac{1}{K_{\max}\,\alpha_\star(C)} = \frac{n^{2/3}}{K_{\max}}\frac{L}{\sqrt{C}(1-\mu)v_{\min}}\;, \tag{30}$$

and use $\sqrt{C}f_n(C,\lambda) = 2\mu v_{\min}L/L_{\dot V}$.

• *The choice $C=\lambda$.* Since $n\geq 2$, the second condition in (29) is satisfied with $\lambda = C$. (30) is a decreasing function of C so that by the first condition in (29), C solves

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

A solution exists in $(0,1)$ and is unique (see Lemma 6); it is denoted by $C^\star$. Since the LHS is lower bounded by $C\mapsto(1-C)^{-1}$ on $(0,1)$, $C^\star$ is upper bounded by $C^+\in(0,1)$ solving

$$\sqrt{C} = 2\mu v_{\min}\frac{L}{L_{\dot V}}(1-C)\;.$$

This yields $C^+ = (\sqrt{1+4A^2}-1)/(2A)$ with $A \stackrel{\text{def}}{=} 2\mu v_{\min}L/L_{\dot V}$. Note that $f_n(C,C)\leq f_2(C,C)\leq f_2(C^+,C^+)$ for any $C\in(0,1)$.

• *Another choice, for any n large enough.* We have when $n\to\infty$

$$\mathcal{L}_n(C,\lambda)\uparrow\mathcal{L}_\infty(C,\lambda) \stackrel{\text{def}}{=} \frac{L_{\dot V}\sqrt{C}}{2L^2}\left(v_{\min}\frac{2L}{L_{\dot V}}-\frac{C^{3/2}}{\lambda}\frac{1}{1-\lambda}\right)\;.$$

By Lemma 7 applied with $A\leftarrow v_{\min}/L$ and $B\leftarrow 2L_{\dot V}/L^2$, we have $\mathcal{L}_\infty(C,\lambda)\leq\mathcal{L}_\infty(C_\star,\lambda_\star)$ where

$$\lambda_\star \stackrel{\text{def}}{=} \frac{1}{2}\;,\qquad C_\star \stackrel{\text{def}}{=} \frac{1}{4}\left(\frac{v_{\min}L}{L_{\dot V}}\right)^{2/3}\;,\qquad \mathcal{L}_\infty(C_\star,\lambda_\star) = \frac{3}{8}\frac{v_{\min}}{L}\left(\frac{v_{\min}L}{L_{\dot V}}\right)^{1/3}\;.$$

Set $N_\star \stackrel{\text{def}}{=} (v_{\min}L/L_{\dot V})^2/8$; for any $n\geq N_\star$, the second condition in (27) is satisfied and we have

$$\lim_n n^{2/3}\alpha_k(n,C_\star,\lambda_\star)\geq\lim_n n^{2/3}\underline{\alpha}_n(C_\star,\lambda_\star)\geq\mathcal{L}_\infty(C_\star,\lambda_\star)>0\;,$$

thus showing that for any n large enough (with a bound which only depends upon $L, L_{\dot V}, v_{\min}$), we have

$$\frac{1}{K_{\max}\sum_{k=0}^{K_{\max}-1}\alpha_k(n,C_\star,\lambda_\star)}\leq\frac{n^{2/3}}{K_{\max}\mathcal{L}_\infty(C_\star,\lambda_\star)} = \frac{n^{2/3}}{K_{\max}}\frac{8}{3}\frac{L}{v_{\min}}\left(\frac{L_{\dot V}}{v_{\min}L}\right)^{1/3}\;.$$

6.2.3 Proof of Proposition 5

It is a follow-up of Theorem 1; the quantities $\alpha_k, \Lambda_k, \delta_k$ introduced in the statement of Theorem 1 are used below without being defined again.

We consider the case when, for $\ell = 1, \ldots, K_{\max}$,

$$\beta_\ell \stackrel{\text{def}}{=} \frac{1-\lambda}{n^{\mathsf{b}}}, \qquad \gamma_\ell^2 \stackrel{\text{def}}{=} \frac{C}{L^2 n^{2\mathsf{c}} K_{\max}^{2\mathsf{d}}}$$

for some $\lambda \in (0,1)$, $C > 0$ and $\mathsf{b}, \mathsf{c}, \mathsf{d}$ to be defined in the proof in such a way that *(i)* $\alpha_k \geq 0$, *(ii)* $\sum_{k=0}^{K_{\max}-1} \alpha_k$ is positive and as large as possible. Since there will be a discussion on (n, C, λ), we make more explicit the dependence of some constants upon these quantities: α_k will be denoted by $\alpha_k(n, C, \lambda)$.

With theses definitions, we have

$$\rho \stackrel{\text{def}}{=} 1 - \frac{1}{n} + \beta_\ell + L^2\gamma_\ell^2 = 1 - \frac{1}{n}\left(1 - \frac{1-\lambda}{n^{\mathsf{b}-1}} - \frac{C}{n^{2\mathsf{c}-1}K_{\max}^{2\mathsf{d}}}\right) ,$$

and choose $(\mathsf{b}, \mathsf{c}, \mathsf{d}, \lambda, C)$ such that

$$\frac{1-\lambda}{n^{\mathsf{b}-1}} + \frac{C}{n^{2\mathsf{c}-1}K_{\max}^{2\mathsf{d}}} \leq 1 , \tag{31}$$

which ensures that $\rho \in (0,1]$. Hence, for any $0 \leq k \leq K_{\max} - 2$,

$$\begin{aligned}\Lambda_k &\leq n^{\mathsf{b}}\left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\frac{C}{L^2 n^{2\mathsf{c}} K_{\max}^{2\mathsf{d}}} \sum_{j=k+1}^{K_{\max}-1} \rho^{j-k-1} \\ &\leq \left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\frac{C}{L^2 n^{2\mathsf{c}-\mathsf{b}} K_{\max}^{2\mathsf{d}-1}} .\end{aligned}$$

From this upper bound, we obtain the following lower bound for any $0 \leq k \leq K_{\max} - 1$: $\alpha_k(n, C, \lambda) \geq \underline{\alpha}_n(C, \lambda)$ where

$$\begin{aligned}(n^{\mathsf{c}} K_{\max}^{\mathsf{d}})\, \underline{\alpha}_n(C,\lambda) \stackrel{\text{def}}{=} \frac{\sqrt{C}}{L}\Bigg(v_{\min} - \sqrt{C}\frac{L_{\dot V}}{2L}\Bigg\{&\frac{1}{n^{\mathsf{c}} K_{\max}^{\mathsf{d}}} \\ &+ \frac{C}{n^{3\mathsf{c}-\mathsf{b}} K_{\max}^{3\mathsf{d}-1}}\left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\Bigg\}\Bigg) .\end{aligned}$$

Based on this inequality and on (31), we choose $\mathsf{b} = 1$ and $\mathsf{c} = \mathsf{d} = 1/3$; which yields for $n \geq 1$,

$$\begin{aligned}(nK_{\max})^{1/3}\, \underline{\alpha}_n(C,\lambda) = \mathcal{L}_n(C,\lambda) &\stackrel{\text{def}}{=} \frac{\sqrt{C} L_{\dot V}}{2L^2}\left(v_{\min}\frac{2L}{L_{\dot V}} - \sqrt{C}\tilde f_n(C,\lambda)\right) , \\ \tilde f_n(C,\lambda) &\stackrel{\text{def}}{=} \frac{1}{(nK_{\max})^{1/3}} + C\left(\frac{1}{n} + \frac{1}{1-\lambda}\right) .\end{aligned}$$

Let $\mu \in (0,1)$. Fix $\lambda \in (0,1)$ and $C > 0$ such that (see (31) for the second condition)

$$\sqrt{C}\tilde f_n(C,\lambda) = 2\mu v_{\min}\frac{L}{L_{\dot V}} , \qquad \frac{n^{1/3}}{K_{\max}^{2/3}} \leq \frac{\lambda}{C} . \tag{32}$$

This implies that

$$(nK_{\max})^{1/3}\alpha_k(n,C,\lambda) \geq (nK_{\max})^{1/3}\underline{\alpha}_n(C,\lambda) \\ \geq (nK_{\max})^{1/3}\alpha_\star(C) \stackrel{\text{def}}{=} \sqrt{C}(1-\mu)v_{\min}/L \; .$$

We obtain the upper bound on E_1 by

$$\mathsf{E}_1 \leq \frac{1}{K_{\max}\ \alpha_\star(C)} \sum_{k=0}^{K_{\max}-1} \alpha_k(n,C,\lambda)\ \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] \; .$$

For E_2 and since $\delta_k \geq L_{\dot{V}}\gamma_{k+1}^2/2$

$$\frac{L_{\dot{V}}\sqrt{C}}{2(1-\mu)Ln^{1/3}} \frac{1}{K_{\max}^{1/3}v_{\min}}\mathsf{E}_2 \\ \leq \frac{L_{\dot{V}}\,C}{2L^2n^{2/3}K_{\max}^{2/3}} \frac{1}{K_{\max}\ \alpha_\star(C)} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}^k)\|^2\right] \\ \leq \frac{1}{K_{\max}\ \alpha_\star(C)} \sum_{k=0}^{K_{\max}-1} \delta_k\mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s}\circ\mathsf{T}(\widehat{S}^k)\|^2\right] \; .$$

We then conclude by

$$\frac{1}{K_{\max}\alpha_\star(C)} = \frac{n^{1/3}}{K_{\max}^{2/3}} \frac{L}{\sqrt{C}(1-\mu)v_{\min}} \; , \tag{33}$$

and use $\sqrt{C}\tilde{f}_n(C,\lambda) = 2\mu v_{\min}L/L_{\dot{V}}$.

Complexity. For $\tau > 0$, set $C = \lambda\tau$. Then for any $\lambda \in (0,1)$,

$$\sqrt{\lambda\tau}\tilde{f}_n(\lambda\tau,\lambda) = \frac{\sqrt{\lambda}\sqrt{\tau}}{(nK_{\max})^{1/3}} + \lambda^{3/2}\tau^{3/2}\left(\frac{1}{n} + \frac{1}{1-\lambda}\right) \; ,$$

which is a continuous increasing function of λ, which tends to zero when $\lambda \to 0$ and to $+\infty$ when $\lambda \to 1$. Hence, there exists an unique $\lambda_\star \in (0,1)$, depending upon $L, L_{\dot{V}}, v_{\min}, \tau, \mu$ and $n, K_{\max}$ such that $\sqrt{\lambda_\star\tau}\tilde{f}_n(\lambda_\star\tau,\lambda_\star) = 2\mu v_{\min}L/L_{\dot{V}}$. Note however that since $\sqrt{\lambda\tau}\tilde{f}_n(\lambda\tau,\lambda) \geq \lambda^{3/2}\tau^{3/2}/(1-\lambda)$ for any $\lambda \in (0,1)$, then $\lambda_\star$ is upper bounded by the unique solution $\lambda^+ \in (0,1)$ satisfying $L_{\dot{V}}\lambda^{3/2}\tau^{3/2}/(2L(1-\lambda)) = \mu v_{\min}$ (see Lemma 8). Such a solution λ^+ only depends upon $L, L_{\dot{V}}, v_{\min}, \tau, \mu$. Hence, for any $\tau > 0$,

$$\tilde{f}_n(\lambda\tau,\lambda) \leq \sup_{n,K_{\max}} \tilde{f}_n(\lambda^+(\tau)\tau,\lambda^+(\tau))$$

and the RHS does not depend on $n, K_{\max}$. There exists $M > 0$ depending upon $L, L_{\dot{V}}, v_{\min}, \tau, \mu$ such that for any $\varepsilon > 0$,

$$K_{\max} \geq \left(\tau^{3/2}\sqrt{n}\right) \vee \left(M\sqrt{n}\varepsilon^{-3/2}\right) \Longrightarrow \frac{n^{1/3}}{K_{\max}^{2/3}} \frac{L\,\tilde{f}_n(\lambda\tau,\lambda)}{\mu(1-\mu)v_{\min}^2} \leq \varepsilon \; .$$

Another choice of (λ, C)*, for any* n *large enough.* In this section, we consider that there exists $\tau > 0$ such that $\sup_{n, K_{\max}} n^{1/3} K_{\max}^{-2/3} \leq \tau$, and $n \to \infty$, $nK_{\max} \to \infty$. In this asymptotic, we have $\mathcal{L}_n(C, \lambda) \uparrow \mathcal{L}_\infty(C, \lambda)$ where

$$\mathcal{L}_\infty(C, \lambda) \stackrel{\text{def}}{=} \frac{\sqrt{C}}{L} \left(v_{\min} - \frac{L_{\dot{V}}}{2L} \frac{C^{3/2}}{1-\lambda} \right) .$$

For any $(C, \lambda) \in \mathbb{R}^+ \times (0,1)$ s.t. $\tau \leq \lambda / C$, we have $\mathcal{L}_\infty(C, \lambda) \leq \mathcal{L}_\infty(C_\star(\lambda), \lambda)$ where

$$C_\star(\lambda) \stackrel{\text{def}}{=} \left(\frac{v_{\min} L}{2 L_{\dot{V}}} \right)^{2/3} (1-\lambda)^{2/3} ;$$

see Lemma 7. The condition $C\tau \leq \lambda$ implies that this inequality holds for any $\lambda \in [\lambda_\star, 1)$ where $\lambda_\star$ is the unique solution of (see Lemma 8)

$$\left(\frac{v_{\min} L}{2 L_{\dot{V}}} \right)^2 (1-\lambda_\star)^2 = \lambda_\star^3 / \tau^3 .$$

Since $\mathcal{L}_\infty(C_\star(\lambda), \lambda) = \frac{3}{4} \left(\frac{v_{\min}^4}{2L^2 L_{\dot{V}}} \right)^{1/3} (1-\lambda)^{1/3}$, this quantity is maximal by choosing $\lambda = \lambda_\star$. Therefore, we have for any $(C, \lambda) \in \mathbb{R}^+ \times (0,1)$, s.t. $\tau \leq \lambda / C$, we have

$$\lim_n n^{1/3} K_{\max}^{1/3} \underline{\alpha}_n(C_\star(\lambda_\star), \lambda_\star) = \mathcal{L}_\infty(C_\star(\lambda_\star), \lambda_\star) > 0 .$$

For any n large enough (with a bound which only depends upon $L, L_{\dot{V}}, v_{\min}, \tau$), we have

$$\frac{1}{K_{\max} \alpha_\star(C_\star, \lambda_\star)} = \frac{n^{1/3}}{K_{\max}^{2/3}} \frac{4}{3} \left(\frac{2L^2 L_{\dot{V}}}{v_{\min}^4} \right)^{1/3} (1-\lambda_\star)^{-1/3} .$$

6.2.4 Proof of Proposition 6

It is a follow-up of Theorem 1; the quantities $\alpha_k, \Lambda_k, \delta_k$ introduced in the statement of Theorem 1 are used below without being defined again.

Let $p_0, \ldots, p_{K_{\max}-1}$ be positive real numbers such that $\sum_{k=0}^{K_{\max}-1} p_k = 1$. We consider the case when

$$\beta_\ell \stackrel{\text{def}}{=} \frac{1-\lambda}{n^{\mathsf{b}}} , \qquad \gamma_\ell^2 \stackrel{\text{def}}{=} \frac{C_\ell}{L^2 n^{2\mathsf{c}} K_{\max}^{2\mathsf{d}}} ,$$

for $\lambda \in (0,1)$, $C_\ell > 0$, and $\mathsf{b}, \mathsf{c}, \mathsf{d}$ to be defined in the proof.

The first step consists in the definition of a function F and of a family $\mathcal{C}$ of vectors $\underline{C} = (C_1, \ldots, C_{K_{\max}}) \in \mathbb{R}_+^{K_{\max}}$ such that $\alpha_k \geq F(C_{k+1}) \geq 0$, and $\sum_{\ell=0}^{K_{\max}-1} F(C_{\ell+1}) > 0$. The second step proves that we can find $\underline{C} \in \mathcal{C}$ such that $p_k = F(C_{k+1}) / \sum_{\ell=0}^{K_{\max}-1} F(C_{\ell+1})$ for any $k = 0, \ldots, K_{\max} - 1$.

Such a pair $(F, \underline{C})$ is not unique, and among the possible ones, we indicate two strategies, all motivated by making the sum $\sum_{\ell=0}^{K_{\max}-1} F(C_{\ell+1})$ as large as possible.

Step 1- Definition of the function F. With the definition of the sequences γ_ℓ and β_ℓ, we have

$$1 - \frac{\rho_{n,\ell}}{n} \stackrel{\text{def}}{=} 1 - \frac{1}{n} + \beta_\ell + \gamma_\ell^2 L^2 = 1 - \frac{1}{n}\left(1 - \frac{1-\lambda}{n^{\mathsf{b}-1}} - \frac{C_\ell}{n^{2\mathsf{c}-1}K_{\max}^{2\mathsf{d}}}\right)$$

and choose $(\mathsf{b}, \mathsf{c}, \mathsf{d}, \lambda, C_\ell)$ such that

$$\frac{1-\lambda}{n^{\mathsf{b}-1}} + \frac{C_{\max}}{n^{2\mathsf{c}-1}K_{\max}^{2\mathsf{d}}} < 1 \ , \text{ where } C_{\max} \stackrel{\text{def}}{=} \max_\ell C_\ell \ , \tag{34}$$

which ensures that $\rho_{n,\ell} \in (0,1)$. Define

$$\rho_n \stackrel{\text{def}}{=} \min_\ell \rho_{n,\ell} = 1 - \frac{1-\lambda}{n^{\mathsf{b}-1}} - \frac{C_{\max}}{n^{2\mathsf{c}-1}K_{\max}^{2\mathsf{d}}} \ .$$

Hence, for any $0 \le k \le K_{\max} - 2$,

$$\begin{aligned}\Lambda_k &\le n^{\mathsf{b}}\left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\frac{1}{L^2 n^{2\mathsf{c}} K_{\max}^{2\mathsf{d}}} \sum_{j=k+1}^{K_{\max}-1} C_{j+1}\left(1 - \frac{\rho_n}{n}\right)^{j-k-1} \\ &\le \left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\frac{C_{\max}}{L^2\rho_n}\frac{1}{n^{2\mathsf{c}-\mathsf{b}-1}K_{\max}^{2\mathsf{d}}} \ .\end{aligned}$$

From this upper bound, we obtain the following lower bound on α_k, for any $0 \le k \le K_{\max} - 1$,

$$\alpha_k \ge \frac{\sqrt{C_{k+1}}}{L n^{\mathsf{c}} K_{\max}^{\mathsf{d}}}\left(v_{\min} - \frac{L_{\dot V}}{2L}\frac{\sqrt{C_{k+1}}}{n^{\mathsf{c}}K_{\max}^{\mathsf{d}}} - \frac{L_{\dot V}}{2L}\frac{C_{\max}\sqrt{C_{k+1}}}{\rho_n\, n^{3\mathsf{c}-\mathsf{b}-1}K_{\max}^{3\mathsf{d}}}\left(\frac{1}{n^{\mathsf{b}}} + \frac{1}{1-\lambda}\right)\right) \ .$$

Based on this inequality and on (34), we choose $\mathsf{b} = 1$, $\mathsf{c} = 2/3$, $\mathsf{d} = 0$: this yields $\rho_n = \lambda - C_{\max} n^{-1/3}$ and $\alpha_k \ge \underline{\alpha}_k$ with (see (16) for the definition of f_n)

$$\underline{\alpha}_k \stackrel{\text{def}}{=} \frac{\sqrt{C_{k+1}}L_{\dot V}}{2L^2 n^{2/3}}\left(v_{\min}\frac{2L}{L_{\dot V}} - \sqrt{C_{k+1}} f_n(C_{\max}, \lambda)\right) \ ; \tag{35}$$

the condition (34) gets into $n^{-1/3} < \lambda / C_{\max}$.

Define the quadratic function $x \mapsto F(x) \stackrel{\text{def}}{=} Ax(v_{\min} - Bx)$ where

$$A \stackrel{\text{def}}{=} \frac{1}{Ln^{2/3}} \ , \quad B \stackrel{\text{def}}{=} f_n(C,\lambda)\frac{L_{\dot V}}{2L} \ ; \tag{36}$$

we have $\underline{\alpha}_k = F(\sqrt{C_{k+1}})$. By Lemma 5, F is increasing on $(0, v_{\min}/(2B)]$, reaches its maximum at $x_\star \stackrel{\text{def}}{=} v_{\min}/(2B)$ and its maximal value is $F_\star \stackrel{\text{def}}{=} Av_{\min}^2/(4B)$. In addition, its inverse F^{-1} exists on $(0, F_\star]$.

Step 2- Choice of $C_1, \ldots, C_{K_{\max}}$. We are now looking for $C_1, \ldots, C_{K_{\max}}$ such that $p_k = F(\sqrt{C_{k+1}}) / \sum_{\ell=0}^{K_{\max}-1} F(\sqrt{C_{\ell+1}})$ or equivalently

$$\frac{p_k}{p_I} = \frac{F(\sqrt{C_{k+1}})}{F(\sqrt{C_I})}, \qquad I \in \operatorname{argmax}_k p_k \ . \tag{37}$$

It remains to fix $F(\sqrt{C_I})$ in such a way that F is invertible on $\left(0, \sqrt{C_I}\right]$. Since we also want $\sum_\ell F(\sqrt{C_{\ell+1}}) = F(\sqrt{C_I})/p_I$ as large as possible, and F is increasing on $(0, x_\star]$, we choose

$$\sqrt{C_I} = \sqrt{C_{\max}} = x_\star = \frac{v_{\min}}{2B} \; . \tag{38}$$

Therefore, $C_{\max}$ solves the equation $\sqrt{C_{\max}} = v_{\min}/(2B)$ or equivalently

$$\frac{v_{\min} L}{L_{\dot V}} = \sqrt{C_{\max}} f_n(C_{\max}, \lambda) \; , \tag{39}$$

under the constraint that $\lambda \in (0,1)$ and $n^{-1/3} < \lambda/C_{\max}$. When $C_{\max}$ is fixed, we set

$$\sqrt{C_{k+1}} \stackrel{\text{def}}{=} F^{-1}\left(\frac{p_k}{\max_\ell p_l} F(\sqrt{C_{\max}})\right) .$$

With these definitions, we have (see (37))

$$\frac{1}{\sum_{k=0}^{K_{\max}-1} F(\sqrt{C_{k+1}})} = \frac{\max_\ell p_\ell}{F(\sqrt{C_{\max}})} .$$

Remember that $F(\sqrt{C_{\max}}) = F(x_\star) = v_{\min}\sqrt{C_{\max}}/(2Ln^{2/3})$.

Step 3. Lower bound on δ_k We write

$$\delta_k \geq \frac{L_{\dot V}}{2} \gamma_{k+1}^2 \; ,$$

so that

$$\frac{\delta_k}{\sum_{k=0}^{K_{\max}-1} F(\sqrt{C_{k+1}})} \geq \frac{L_{\dot V} L}{v_{\min}} n^{2/3} \frac{\max_\ell p_\ell}{\sqrt{C_{\max}}} \gamma_{k+1}^2 \; .$$

Case $\lambda = C$. A simple strategy is to choose $n \geq 2$ and $C_{\max} = \lambda$ solution of $v_{\min}/2 = \sqrt{C} f_n(C, C)$. This solution exists and is unique, and it is upper bounded by a quantity C^+ which depends only on $L, L_{\dot V}, v_{\min}$ - the same discussion is proved in subsubsection 6.2.2.

Case $\lambda = 1/2$. $f_n(C, \lambda)$ controls the errors E_i and we can choose $\lambda \in (0,1)$ and then $C > 0$ such that this quantity is minimal; to make the computations easier, we minimize w.r.t. λ the function $\lim_n f_n(C, \lambda)$: it behaves like $\lambda^{-1}(1-\lambda)^{-1}$ so that we set $\lambda = 1/2$. The equation $\sqrt{C} f_n(C, 1/2) = v_{\min} L / L_{\dot V}$ possesses an unique solution in $\left(0, \lambda n^{1/3}\right)$.

Upon noting that $x \mapsto \sqrt{x} f_n(x, 1/2)$ is lower bounded by $x \mapsto 4x^{3/2}$, the solution to the equation $\sqrt{C} f_n(C, 1/2) = v_{\min} L / L_{\dot V}$ satisfies

$$C \leq \left(\frac{v_{\min} L}{4 L_{\dot V}}\right)^{2/3} \; ,$$

thus showing that the constraint $n^{-1/3} < \lambda/C = 1/(2C)$ is satisfied for any n such that $8n > (v_{\min} L / L_{\dot V})^2$.

6.2.5 Auxiliary results

Lemma 4 *Assume H1, H2 and H3. For any $k \geq 0$,*

$$\mathbb{E}\left[\|H_{k+1}\|^2\right] = \mathbb{E}\left[\|H_{k+1} - h(\widehat{S}^k)\|^2\right] + \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right],$$

and

$$\mathbb{E}\left[\|H_{k+1} - h(\widehat{S}^k)\|^2\right] + \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^k)\|^2\right] = \mathbb{E}\left[\|\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}}\|^2\right] .$$

Proof Since $\mathbb{E}\left[H_{k+1}\middle|\mathcal{F}_{k+1/2}\right] = h(\widehat{S}^k)$, we have

$$\mathbb{E}\left[\|H_{k+1}\|^2\right] = \mathbb{E}\left[\|H_{k+1} - h(\widehat{S}^k)\|^2\right] + \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] .$$

In addition, upon noting that $\mathsf{S}_{k+1,i} \in \mathcal{F}_{k+1/2}$ for any i,

$$\begin{aligned} H_{k+1} - h(\widehat{S}^k) &= \bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}} - \bar{s} \circ \mathsf{T}(\widehat{S}^k) + \widetilde{S}^{k+1} \\ &= \bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}} - \mathbb{E}\left[\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}}\middle|\mathcal{F}_{k+1/2}\right] , \end{aligned}$$

we have

$$\mathbb{E}\left[\|H_{k+1} - h(\widehat{S}^k)\|^2\right] + \mathbb{E}\left[\|\widetilde{S}^{k+1} - \bar{s} \circ \mathsf{T}(\widehat{S}^k)\|^2\right] = \mathbb{E}\left[\|\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}}\|^2\right] .$$

Proposition 7 *Assume H1, H2, H3 and H5-item 2. Set $L^2 \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^n L_i^2$. Then*

$$\mathbb{E}\left[\|\bar{s}_{J_1} \circ \mathsf{T}(\widehat{S}^0) - \mathsf{S}_{1,J_1}\|^2\right] = 0 ,$$

and for any $k \geq 1$ and $\beta_1, \ldots, \beta_k > 0$,

$$\begin{aligned} &\mathbb{E}\left[\|\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,J_{k+1}}\|^2\right] \\ &\quad \leq \sum_{j=1}^k \widetilde{\Lambda}_{j,k} \left\{ \mathbb{E}\left[\|h(\widehat{S}^{j-1})\|^2\right] - \left(1 + \frac{1}{\beta_j}\right)^{-1} \mathbb{E}\left[\|\widetilde{S}^j - \bar{s} \circ \mathsf{T}(\widehat{S}^{j-1})\|^2\right] \right\} , \end{aligned}$$

where

$$\widetilde{\Lambda}_{j,k} \stackrel{\text{def}}{=} L^2 \left(\frac{n-1}{n}\right)^{k-j+1} \gamma_j^2 \left(1 + \frac{1}{\beta_j}\right) \prod_{\ell=j+1}^k \left(1 + \beta_\ell + \gamma_\ell^2 L^2\right) .$$

By convention, $\prod_{\ell=k+1}^k a_\ell = 1$.

Proof For $k = 0$,

$$\mathbb{E}\left[\|\bar{s}_{J_1} \circ \mathsf{T}(\widehat{S}^0) - \mathsf{S}_{1,J_1}\|^2\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\bar{s}_i \circ \mathsf{T}(\widehat{S}^0) - \mathsf{S}_{1,i}\|^2\right] = 0\,.$$

Let $k \geq 1$. We write (see (11))

$$\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}\mathbb{1}_{I_{k+1}\neq i} + \bar{s}_i \circ \mathsf{T}(\widehat{S}^k)\mathbb{1}_{I_{k+1}=i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}^{<k,i})\mathbb{1}_{I_{k+1}\neq i} + \bar{s}_i \circ \mathsf{T}(\widehat{S}^k)\mathbb{1}_{I_{k+1}=i},$$

where $\widehat{S}^{<\ell,i}$ is defined by (26). This yields, by H5-item 2

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\bar{s}_i \circ \mathsf{T}(\widehat{S}^k) - \mathsf{S}_{k+1,i}\|^2\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\bar{s}_i \circ \mathsf{T}(\widehat{S}^k) - \bar{s}_i \circ \mathsf{T}(\widehat{S}^{<k,i})\|^2 \mathbb{1}_{I_{k+1}\neq i}\right] \\ &\leq \Delta_k \stackrel{\text{def}}{=} \frac{n-1}{n^2}\sum_{i=1}^{n} L_i^2\, \mathbb{E}\left[\|\widehat{S}^k - \widehat{S}^{<k,i}\|^2\right]. \end{aligned} \tag{40}$$

We have

$$\Delta_k = \frac{n-1}{n^2}\sum_{i=1}^{n} L_i^2\, \mathbb{E}\left[\|\widehat{S}^k - \widehat{S}^{k-1} + \left(\widehat{S}^{k-1} - \widehat{S}^{<k-1,i}\right)\mathbb{1}_{I_k\neq i}\|^2\right]$$

where we used in the last inequality that $\widehat{S}^{<k,i} = \widehat{S}^{k-1}\mathbb{1}_{I_k=i} + \widehat{S}^{<k-1,i}\mathbb{1}_{I_k\neq i}$. Upon noting that $2\left\langle \tilde{U}, V\right\rangle \leq \beta^{-1}\|\tilde{U}\|^2 + \beta\|V\|^2$ for any $\beta > 0$, we have for any $\mathcal{G}$-measurable r.v. V

$$\mathbb{E}\left[\|U+V\|^2\right] \leq \mathbb{E}\left[\|U\|^2\right] + \beta^{-1}\mathbb{E}\left[\|\mathbb{E}\left[U|\mathcal{G}\right]\|^2\right] + (1+\beta)\mathbb{E}\left[\|V\|^2\right].$$

Applying this inequality with $\beta \leftarrow \beta_k$, $U \leftarrow \widehat{S}^k - \widehat{S}^{k-1} = \gamma_k H_k$ and $\mathcal{G} \leftarrow \mathcal{F}_{k-1/2}$ yields

$$\begin{aligned}\Delta_k \leq \gamma_k^2 \frac{n-1}{n} L^2\, \mathbb{E}\left[\|H_k\|^2\right] + \frac{\gamma_k^2}{\beta_k}\frac{n-1}{n} L^2\, \mathbb{E}\left[\|\mathbb{E}\left[H_k|\mathcal{F}_{k-1/2}\right]\|^2\right] \\ + (1+\beta_k)\frac{n-1}{n^2}\sum_{i=1}^{n} L_i^2\, \mathbb{E}\left[\|\widehat{S}^{k-1} - \widehat{S}^{<k-1,i}\|^2 \mathbb{1}_{I_k\neq i}\right].\end{aligned}$$

By Lemma 4 and (40), we have

$$\mathbb{E}\left[\|H_k\|^2\right] \leq \mathbb{E}\left[\|h(\widehat{S}^{k-1})\|^2\right] + \Delta_{k-1} - \mathbb{E}\left[\|\widetilde{S}^k - \bar{s} \circ \mathsf{T}(\widehat{S}^{k-1})\|^2\right];$$

for the second term, we use again $\mathbb{E}\left[H_k|\mathcal{F}_{k-1/2}\right] = h(\widehat{S}^{k-1})$; for the third term, since $I_k \in \mathcal{F}_{k-1/2}$, $\widehat{S}^{k-1} \in \mathcal{F}_{k-1}$, $\widehat{S}^{<k-1,i} \in \mathcal{F}_{k-1}$, then

$$\sum_{i=1}^{n} L_i^2\, \mathbb{E}\left[\|\widehat{S}^{k-1} - \widehat{S}^{<k-1,i}\|^2 \mathbb{1}_{I_k\neq i}\right] = n\Delta_{k-1}.$$

Therefore, we established

$$\begin{aligned}\Delta_k \leq \left(1 + \beta_k + \gamma_k^2 L^2\right)\frac{n-1}{n}\Delta_{k-1} + \gamma_k^2(1+\frac{1}{\beta_k})L^2\,\frac{n-1}{n}\,\mathbb{E}\left[\|h(\widehat{S}^{k-1})\|^2\right] \\ - \gamma_k^2 L^2 \frac{n-1}{n}\mathbb{E}\left[\|\widetilde{S}^k - \bar{s} \circ \mathsf{T}(\widehat{S}^{k-1})\|^2\right].\end{aligned}$$

The proof is then concluded by standard algebra upon noting that $\Delta_0 = 0$.

6.2.6 Technical lemmas

Lemma 5 *Let $A, B, v > 0$ and define $F(x) \stackrel{\text{def}}{=} Ax(v - Bx)$ on $\mathbb{R}$. Then the roots of F are $\{0, v/B\}$; F is positive on $(0, v/B)$; the maximal value of F is $Av^2/(4B)$ and it is reached at $x_\star \stackrel{\text{def}}{=} v/2B$.*

Lemma 6 *Let $a, b > 0$ and define F on $(0, 1)$ by $F(x) = \sqrt{x}(a + b/(1 - x))$. For any $v > 0$, there exists an unique $x \in (0, 1)$ such that $F(x) = v$.*

Proof $x \mapsto F(x)$ is continuous and increasing on $(0, 1)$, tends to zero when $x \to 0$ and to $+\infty$ when $x \to 1$; therefore for any $v > 0$, there exists an unique $x \in (0, 1)$ such that $F(x) = v$.

Lemma 7 *Let $A, B > 0$. The function $F : x \mapsto Ax - Bx^4$ defined on $(0, \infty)$ reaches its unique maximum at $x_\star \stackrel{\text{def}}{=} A^{1/3}B^{-1/3}4^{-1/3}$ and $F(x_\star) = 3A^{4/3}/(B4^4)^{1/3}$.*

Proof $F'(x) = A - 4Bx^3$ and $F''(x) = -12Bx^2 < 0$; hence, F' is decreasing. $F'(x) = 0$ iff $x^3 = A/(4B)$, showing $F' > 0$ on $(0, x_\star)$ with $x_\star \stackrel{\text{def}}{=} A^{1/3}/(4B)^{1/3}$. Hence, F is increasing on $[0, x_\star]$ and then decreasing.

Lemma 8 *For any $v > 0$, the function $x \mapsto (1 - x)^2/x^3$ is decreasing on $(0, 1)$ and there exists an unique $x \in (0, 1)$ solving $(1 - x)^2/x^3 = v$.*

Proof The derivative of $x \mapsto (1 - x)^2/x^3$ is $-x^{-4}(x - 3)(x - 1)$ thus showing that the function is decreasing on $(0, 1)$; it tends to $+\infty$ when $x \to 0$ and to 0 when $x \to 1$. This concludes the proof.

References

Agarwal A, Bottou L (2015) A lower bound for the optimization of finite sums. In: Bach F, Blei D (eds) Proceedings of the 32nd International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 37, pp 78–86

Allassonnière S, Kuhn E, Trouvé A (2010) Construction of bayesian deformable models via a stochastic approximation algorithm: A convergence study. Bernoulli 16(3):641–678

Allen-Zhu Z, Hazan E (2016) Variance reduction for faster non-convex optimization. In: Balcan M, Weinberger K (eds) Proceedings of The 33rd International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 48, pp 699–707

Balakrishnan S, Wainwright MJ, Yu B (2017) Statistical guarantees for the EM algorithm: From population to sample-based analysis. Ann Statist 45(1):77–120

Benveniste A, Métivier M, Priouret P (1990) Adaptive Algorithms and Stochastic Approximations. Springer Verlag

Bertsekas DP (2011) Incremental proximal methods for large scale convex optimization. Math Program 129(2, Ser. B):163–195

Borkar VS (2008) Stochastic approximation. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, a dynamical systems viewpoint

Bottou L, Le Cun Y (2004) Large scale online learning. In: Thrun S, Saul LK, Schölkopf B (eds) Advances in Neural Information Processing Systems 16, MIT Press, pp 217–224

Brown LD (1986) Fundamentals of statistical exponential families with applications in statistical decision theory, Institute of Mathematical Statistics Lecture Notes—Monograph Series, vol 9. Institute of Mathematical Statistics, Hayward, CA

Cappé O, Moulines E (2009) On-line Expectation Maximization algorithm for latent data models. J Roy Stat Soc B Met 71(3):593–613

Carmon Y, Duchi JC, Hinder O, Sidford A (2018) Accelerated Methods for Non-Convex Optimization. SIAM J Optim 28(2):1751–1772

Cartis C, Gould NIM, Toint PL (2010) On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems. SIAM J on Optimization 20(6):2833–2852

Celeux G, Diebolt J (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly 2:73–82

Chen J, Zhu J, Teh Y, Zhang T (2018) Stochastic Expectation Maximization with Variance Reduction. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems 31, Curran Associates, Inc., pp 7967–7977

Csiszár I, Tusnády G (1984) Information geometry and alternating minimization procedures. In: Recent results in estimation theory and related topics, suppl. 1, Statist. Decisions, pp 205–237

Defazio A, Bach F, Lacoste-Julien S (2014) SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) Advances in Neural Information Processing Systems 27, Curran Associates, Inc., pp 1646–1654

Delyon B, Lavielle M, Moulines E (1999) Convergence of a Stochastic Approximation version of the EM algorithm. Ann Statist 27(1):94–128

Dempster A, Laird N, Rubin D (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J Roy Stat Soc B Met 39(1):1–38

Donnet S, Samson A (2007) Estimation of parameters in incomplete data models defined by dynamical systems. J Statist Plann Inference 137(9):2815 – 2831

Fang C, Li C, Lin Z, Zhang T (2018) SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems 31, Curran Associates, Inc., pp 689–699

Fort G, Moulines E (2003) Convergence of the Monte Carlo Expectation Maximization for curved exponential families. Ann Statist 31(4):1220–1259

Fort G, Ollier E, Samson A (2019) Stochastic proximal-gradient algorithms for penalized mixed models. Stat Comput 29(2):231–253

Frühwirth-Schnatter S, Celeux G, Robert CP (eds) (2019) Handbook of mixture analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, FL

Ghadimi S, Lan G (2013) Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. SIAM J Optimiz 23(4):2341–2368

Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Math Program 155(1-2, Ser. A):267–305

Glasserman P (2004) Monte Carlo methods in financial engineering. Springer, New York

Gunawardana A, Byrne W (2005) Convergence theorems for generalized alternating minimization procedures. J Mach Learn Res 6:2049–2073

Johnson R, Zhang T (2013) Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26, Curran Associates, Inc., pp 315–323

Karimi B, Lavielle M, Moulines E (2019a) On the Convergence Properties of the Mini-Batch EM and MCEM Algorithms. Tech. rep., hal-02334485

Karimi B, Miasojedow B, Moulines E, Wai HT (2019b) Non-asymptotic Analysis of Biased Stochastic Approximation Scheme. In: COLT

Karimi B, Wai HT, Moulines E, Lavielle M (2019c) On the Global Convergence of (Fast) Incremental Expectation Maximization Methods. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems 32, Curran Associates, Inc., pp 2837–2847

Kuhn E, Lavielle M (2004) Coupling a stochastic approximation version of EM with an MCMC procedure. ESAIM: Probability and Statistics 8:115–131

Kuhn E, Matias C, Rebafka T (2019) Properties of the Stochastic Approximation EM Algorithmwith Mini-batch Sampling. Tech. rep., arXiv.1907.09164

Lange K (1995) A Gradient Algorithm Locally Equivalent to the EM Algorithm. JRSS B 57(2):425–437

Le Corff S, Fort G (2013a) Online Expectation Maximization based algorithms for inference in Hidden Markov Models. Electron J Statist 7:763–792

Le Corff S, Fort G (2013b) Online Expectation Maximization based algorithms for inference in Hidden Markov Models. Electron J Statist 7:763–792

Lei L, Ju C, Chen J, Jordan M (2017) Non-convex Finite-Sum Optimization Via SCSG Methods. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, Curran Associates, Inc., pp 2348–2358

Little RJA, Rubin D (2002) Statistical analysis with missing data, 2nd edn. Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ

Mairal J (2015) Incremental majorization-minimization optimization with application to large-scale machine learning. SIAM J Optim 25(2):829–855

McLachlan G, Krishnan T (2008) The EM algorithm and extensions. Wiley series in probability and statistics, Wiley

Murty K, Kabadi S (1987) Some NP-complete problems in quadratic and nonlinear programming. Mathematical Programming 39:117–129

Neal RM, Hinton GE (1998) A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants. In: Jordan MI (ed) Learning in Graphical Models, Springer Netherlands, Dordrecht, pp 355–368

Nesterov Y (2004) Introductory lectures on convex optimization, Applied Optimization, vol 87. Kluwer Academic Publishers, Boston, MA, a basic course

Nettleton D (1999) Convergence properties of the EM algorithm in constrained parameter spaces. Canad J Statist 27(3):639–648

Ng SK, McLachlan GJ (2003) On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. Stat Comput 13(1):45–55

Nguyen H, Forbes F, McLachlan G (2020) Mini-batch learning of exponential family finite mixture models. Stat Comput

Nowlan S (1991) Soft Competitive Adaptation: Neural Network Learning Algorithms based on Fitting Statistical Mixtures. PhD thesis, School of Computer Science, Carnegie Mellon Univ., Pittsburgh

Parisi S, He K, Aghajani R, Sclaroff S, Felzenswalb P (2019) Generalized Majorization-Minimization. Tech. rep., arXiv:1506.07613-v3

Rahmani D, Niranjan M, Fay D, Takeda A, Brodzki J (2020) Estimation of gaussian mixture models via tensor moments with application to online learning. Pattern Recognition Letters 131:285 – 292

Reddi S, Sra S, Póczos B, Smola A (2016) Fast Incremental Method for Smooth Nonconvex Optimization. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp 1971–1977

Schmidt M, Le Roux N, Bach F (2017) Minimizing finite sums with the stochastic average gradient. Mathematical Programming 162(1-2):83–112

Springer T, Urban K (2014) Comparison of the EM algorithm and alternatives. Numerical Algorithms 67:335–364

Srivastava S, DePalma G, Liu C (2019) An Asynchronous Distributed Expectation Maximization Algorithm for Massive Data: The DEM Algorithm. J Comput Graph Stat 28(2):233–243

Sundberg R (2019) Statistical Modelling by Exponential Families. Cambridge University Press

Takai K (2012) Constrained EM algorithm with projection method. Comput Statist 27(4):701–714

Thiesson B, Meek C, Heckerman D (2001) Accelerating EM for Large Databases. Machine Learning 45:279–299

Wang X, Ma S, Yuan YX (2017) Penalty methods with stochastic approximation for stochastic nonlinear programming. Math Comp 86(306):1793–1820

Wei G, Tanner M (1990) A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. J Am Stat Assoc 85(411):699–704

Wu C (1983) On the Convergence Properties of the EM Algorithm. Ann Statist 11(1):95–103

Xu L, Jordan M (1996) On Convergence Properties of the EM Algorithm for Gaussian Mixtures. Neural Comput 8(1):129–151

Zangwill WI (1967) Non-linear programming via penalty functions. Management Sci 13:344–358

Zhao R, Li Y, Sun Y (2020) Statistical convergence of the em algorithm on gaussian mixture models. Electron J Statist 14(1):632–660

Zhou D, Xu P, Gu Q (2018) Stochastic nested variance reduced gradient descent for nonconvex optimization. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems 31, Curran Associates, Inc., pp 3921–3932

Zhu R, Wang L, Zhai C, Gu Q (2017) High-Dimensional Variance-Reduced Stochastic Gradient Expectation-Maximization Algorithm. In: Precup D, Teh YW (eds) Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 70, pp 4180–4188