CONVERGENCE OF A STOCHASTIC APPROXIMATION VERSION OF THE EM ALGORITHM

BERNARD DELYON, MARC LAVIELLE, AND ERIC MOULINES

SUMMARY

The Expectation-Maximization (EM) algorithm is a powerful computational technique for locating maxima of functions. It is widely used in statistics for maximum likelihood or maximum a posteriori estimation in incomplete data models. In certain situations however, this method is not applicable because the expectation step cannot be performed in closed–form. To deal with these problems, a novel method is introduced, the Stochastic Approximation EM (SAEM), which consists in replacing the expectation step of the EM algorithm by one iteration of a stochastic approximation procedure. The convergence of the SAEM algorithm is established under conditions that are applicable to many practical situations. Moreover it is proved that, under mild additional conditions, the attractive stationary points of the SAEM algorithm correspond to the local maxima of the function.

Keywords : Incomplete data; Optimization; Maximum Likelihood; Missing data; Monte-Carlo algorithm; EM Algorithm; Simulation; Stochastic algorithm.

AMS-classification: Primary: 65U05, 62F10, Secondary 62M30,60K35

Affiliations

Dr. M. Lavielle, UFR de Mathématiques et Informatique, Université Paris V, and CNRS URA 743, Université Paris-Sud, E-mail: lavielle@stats.matups.fr

Dr. B. Delyon, IRISA/INRIA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France, E-mail: delyon@irisa.fr

Pr. E. Moulines (Corresponding author), Ecole Nationale Supérieure des Télécommunications, CNRS-URA 820, 46, rue Barrault, 75634 Paris CEDEX 13, E-mail: moulines@sig.enst.fr

Short Title: Convergence of the SAEM algorithm.

1. Introduction

The EM algorithm (Dempster et al. , 1977) is a very popular tool for maximum-likelihood (or maximum a posteriori) estimation. The common strand to problems where this approach is applicable is a notion of *incomplete-data*, which includes the conventional sense of missing data but is much broader than that. The EM algorithm demonstrates its strength in situations where some hypothetical experiments yields (complete) data that are related to the parameters more conveniently than the measurements are. Problems where the EM algorithm has proven to be useful include, among many others: maximum likelihood estimation of the parameters of mixture of densities (Titterington et al. (1985)) and of Hidden Markov Models (Baum and Welsh (1973)), maximum a posteriori estimation in censored data models (see Little and Rubin (1987), Tanner (1993) for a thorough reviews of these models). The EM algorithm has several appealing properties. Because it relies on complete-data computations, it is generally simple to implement: the E-step of each iteration only involves taking expectation over complete-data conditional distribution; the M-step only involves complete data maximum-likelihood estimation, which is often in simple closed–form. Moreover, it is numerically stable, in the sense that it increases the incomplete likelihood at each iteration. When either the maximization step (M-step) or the expectation step (E-step) involves intricate or even infeasible computations, the EM paradigm is no longer directly applicable. The problems raised by the potentially difficult global maximization involved in the M-step have recently been addressed successfully. A possible solution consists in replacing the global optimization by a chain of simpler conditional maximization, leading to the so-called ECM algorithm (see for example, Meng and Rubin (1993), Liu and Rubin (1994)). Another approach is to use a single iteration of an approximate Newton's method leading to the EM gradient algorithm (see Lange (1995)). On the contrary, only partial answers have been obtained to deal with problems for which the expectation of the complete likelihood cannot be done in closed–form. A possible solution to cope with this problem has been proposed by Wei and Tanner (1990) (see section 2). The basic idea consists in computing the expectation in the E-step by means of a Monte-Carlo method. In this contribution, a stochastic version of the EM algorithm (referred to as SAEM, standing for Stochastic Approximation EM) is presented, as an alternative to the MCEM algorithm. It makes use of a stochastic approximation procedure for estimating the conditional expectation of the complete data log-likelihood. Given the current approximation of the parameters, complete data are simulated under the *a posteriori* density; these simulated complete data are then used to update the current value of the conditional expectation of the complete data likelihood. A (decreasing) step size is used to control the allowed amount of update during the successive iterations of the algorithm (section 3). The convergence of the sequence of parameter estimates to a stationary point of the incomplete likelihood is established for a general class of complete data likelihood functions. It is further demonstrated that *only* the local maxima of the incomplete likelihood are attractive for the stochastic approximation algorithm, *i.e.* convergence toward saddle points are avoided with probability one.

2. The EM and the MCEM algorithms

In this section, we shall review the key properties of the EM algorithm that we shall need as derived by Dempster et al. (1977). Let μ be a σ-finite positive Borel measure on $\mathbb{R}^\ell$ and let $\mathcal{F} = \{f(z;\theta), \theta \in \Theta\}$

be a family of positive integrable Borel functions on $\mathbb{R}^\ell$, where Θ is a subset of $\mathbb{R}^p$. Define

$$g(\theta) \triangleq \int_{\mathbb{R}^\ell} f(z;\theta)\mu(dz), \tag{1}$$

$$l(\theta) \triangleq \log g(\theta), \tag{2}$$

$$p(z;\theta) \triangleq \begin{cases} f(z;\theta)/g(\theta) & \text{if } g(\theta) \neq 0 \\ 0 & \text{if } g(\theta) = 0 \end{cases} . \tag{3}$$

Note that $p(z;\theta)$ defines a probability density function w.r.t. to the measure μ. In the terminology introduced by Geyer (1994), $\mathcal{F}$ is a family of *unnormalized density*, $\mathcal{P} = \{p(z;\theta), \theta \in \Theta\}$ is the family of *normalized density* and $g : \Theta \to [0,\infty)$ is the *normalizing function* for the family $\mathcal{F}$. We wish to find the value $\hat{\theta} \in \Theta$ that maximizes $g(\theta)$ (conditions upon which such maximum exists are detailed later on). Many statistical inference problems fall into the framework (1). In the standard missing data problem

- $g(\theta)$ is the incomplete data likelihood, *i.e.* the likelihood of the observed data y (the dependence of $g(\theta)$ w.r.t y is here implicit),
- $f(z;\theta)$ is the complete data likelihood, *i.e.* the likelihood of the complete data x obtained by augmenting the observed data y with the missing data z: $x = (y, z)$,
- $p(z;\theta)$ is the posterior distribution of the missing data z given the observed data y (often referred to as the *predictive distribution*).

Other examples may be found in spatial statistics and stochastic geometry (Geyer (1994,1996)), or in Bayesian inference. We will find convenient to use in the sequel the classical terminology of the *missing data problem*, even though the approaches developed here apply to a more general context. Define

$$Q(\theta|\theta') = \int_{\mathbb{R}^\ell} \log f(z;\theta) p(z;\theta')\mu(dz). \tag{4}$$

The EM algorithm is useful in situations where maximization of $\theta \to Q(\theta|\theta')$ is much simpler than direct maximization of $\theta \to l(\theta)$. Indeed, EM is an iterative algorithm, which maximizes $l(\theta)$ by iteratively maximizing $Q(\theta|\theta')$. Each iteration may be formally decomposed in two steps: an E-step and a M-step. At iteration k, the E-step consists in evaluating:

$$Q(\theta|\theta_k) = \int_{\mathbb{R}^\ell} \log f(z;\theta) p(z;\theta_k)\mu(dz).$$

In the M-step, the value of θ maximizing $Q(\theta|\theta_k)$ is found. This yields the new parameter estimate θ_{k+1}. This two steps procedure is repeated until convergence is apparent. The essence of the EM algorithm is that increasing $Q(\theta|\theta_k)$ forces an increase of $l(\theta)$.

In situations where global maximization of $\theta \to Q(\theta|\theta')$ is not in simple closed-form, alternate solutions can be contemplated. Possible solutions are considered in Meng and Rubin (1993), where global maximization is replaced by a cycle of simpler maximization problem and in Lange (1995), where the M-step is replaced by a single iteration of a Newton's method (see section 8). In certain situations, the expectation step can be numerically involved or even intractable. To deal with these cases, Wei and Tanner (1990) (see also Tanner (1993)) propose to replace the expectation in the computation of $Q(\theta|\theta_k)$ by a Monte-Carlo integration, leading to the so-called Monte-Carlo EM (MCEM algorithm). Basically, the E-step at iteration k is replaced by the following procedure:

- *Simulation-step (S-step):* generate $m(k)$ realizations $z_k(j)$ $(j = 1, \cdots, m(k))$ of the missing data vector under the distribution function $p(z;\theta_k)$,

- *Monte-Carlo integration:* compute the current approximation of $Q(\theta|\theta_k)$ according to

$$\tilde{Q}_k(\theta) = m(k)^{-1} \sum_{j=1}^{m(k)} \log f(z_k(j); \theta). \tag{5}$$

The maximization step remains unchanged (see Tanner (1993) for implementation details). Simulation under the posterior density $p(z;\theta)$ is generally easy when $p(z;\theta)$ is a product of low dimensional marginal distributions: standard stochastic simulation methods such as importance sampling may in such case be used. When the posterior $p(z;\theta)$ cannot be decomposed, it is often convenient to resort to Markov chain Monte-Carlo simulation methods.

3. SAEM : The stochastic approximation EM algorithm

We propose in this contribution an alternative scheme, which shares most of the attractive behavior of the MCEM algorithm. Similar to the MCEM algorithm, the basic idea consists in splitting the E-step, into a simulation step and an integration step. Similar to the MCEM, the S-step consists in generating realizations of the missing data vector under the posterior distribution $p(z;\theta)$; the Monte-Carlo integration is substituted by a stochastic averaging procedure. The proposed algorithm may thus be summarized as follows:

- *Simulation* : generate $m(k)$ realizations $z_k(j)$ $(j = 1, \cdots, m(k))$ of the missing data under the posterior density $p(z;\theta_k)$
- *Stochastic approximation* : update $\hat{Q}_k(\theta)$ according to

$$\hat{Q}_k(\theta) = \hat{Q}_{k-1}(\theta) + \gamma_k \left(\frac{1}{m(k)} \sum_{j=1}^{m(k)} \log f(z_k(j); \theta) - \hat{Q}_{k-1}(\theta) \right) \tag{6}$$

  where $\{\gamma_k\}_{k\geq 1}$ is a sequence of positive step-size.
- *Maximization*: maximize $\hat{Q}_k(\theta)$ in the feasible set Θ, *i.e.* find $\theta_{k+1} \in \Theta$ such that:

$$\hat{Q}_k(\theta_{k+1}) \geq \hat{Q}_k(\theta) \quad \forall \theta \in \Theta.$$

The SAEM algorithm shares some similarity with the Stochastic EM (also referred to as the probabilistic teacher) algorithm developed by Celeux and Diebolt in a series of paper (see Celeux and Diebolt (1988,1992), Diebolt and Celeux (1993); see also Diebolt and Ip (1996)). The main difference w.r.t to the SEM approach is the use of a decreasing sequence of stepsize to approximate (using stochastic approximation) the EM auxiliary function. In the SEM algorithm, the stepsize is set to zero $\gamma_k = 0$ (no approximation) and the number of simulations by iteration is constant (most often, $m(k) = 1$); hence, under mild assumptions, $\{\theta_k\}_{k\geq 0}$ is an homogeneous Markov chain. Under appropriate conditions (see, e.g. Diebolt and Ip (1996)), it may be shown that this Markov chain is geometrically ergodic, and point estimate can be obtained, for example, by computing ergodic averages. The behavior of the stationary distribution of the chain has been characterized in some very simple scenarios; see Diebolt and Ip (1996). The SAEM algorithm is also related to the stochastic approximation procedure developed by Younes (see Younes (1989,1992)) for incompletely observed Gibbsian fields (more on this later).

It should be noted that the stochastic approximation is used here in a slightly unusual context, because the amount of data to process (i.e. the incomplete data) is fixed. The convergence of the SAEM algorithm depends on the choice of step-sizes γ_k and/or the specification of $m(k)$ used in the stochastic approximation. It is inappropriate to start with small values for step-size γ_k and/or large values for the number of simulations $m(k)$. Rather, it is recommended that one *decreases* γ_k and/or *increases* $m(k)$ as the current approximation of the parameter vector moves closer to a stationary point. These intuitions are formally developed in the next section, where convergence of the sequence $\{\theta_k\}_{k\geq 0}$ is discussed. When the maximization step is straightforward to implement and/or is, from the computational point of view, much faster than the simulation step, one may set the number of simulations $m(k) = 1$ for all the iterations.

In comparison with the MCEM algorithm, the SAEM makes a more efficient use of the imputed missing values. At each new iteration of the MCEM algorithm, a whole set of missing values needs to be simulated and all the missing values simulated during the previous iterations are dropped. In the SAEM algorithm, *all* the simulated missing values contribute to the evaluation of the auxiliary quantity $\hat{Q}_k(\theta)$: there are gradually discounted, with a forgetting factor inversely proportional to the step-size. As a result, the SAEM algorithm moves toward modal area more quickly than the MCEM in terms of the number of simulations. The computational advantage of the SAEM algorithm over the MCEM algorithm is striking in problems where maximization is much cheaper than simulation.

All the acceleration methods devised from the EM paradigm (such as the method proposed by Louis (1982) or the modified scoring developed by Meilijson (1989)) can be adapted to the SAEM algorithm following essentially the same lines than Wei and Tanner (1990) for the MCEM algorithm. Note in particular that the gradient (the Fisher score function) and the Hessian (observed Fisher information) of $l(\theta)$ can be obtained almost directly by using the values of the simulated missing data $z_k(j)$, $1 \leq j \leq m(k)$. Using the so-called Fisher identity, the Jacobian of the incomplete data log-likelihood $l(\theta)$ is equal to the conditional expectation of the complete data log-likelihood,

$$\partial_\theta l(\theta) \triangleq E_\theta[\partial_\theta \log f(Z;\theta)], \tag{7}$$

where ∂_θ denotes the differential with respect to θ and

$$E_\theta[\psi(Z)] \triangleq \int \psi(z)p(z;\theta)\mu(dz). \tag{8}$$

Eq. (7) suggests the following stochastic approximation scheme

$$\Delta_k = \Delta_{k-1} + \gamma_k \left(\frac{1}{m(k)} \sum_{j=1}^{m(k)} \partial_\theta \log f(z_k(j);\theta_k) - \Delta_{k-1} \right) \tag{9}$$

Using the Louis's missing information principle (1982), the Hessian of l at θ, $\partial^2_\theta l(\theta)$ (the observed Fisher information matrix) may be expressed as

$$-\partial^2_\theta l(\theta) = -E_\theta[\partial^2_\theta \log f(Z;\theta)] - \mathrm{Cov}_\theta[\partial_\theta \log f(Z;\theta)], \tag{10}$$

where $\mathrm{Cov}_\theta[\psi(Z)] \triangleq E_\theta[(\psi(Z) - E_\theta(\psi(Z)))(\psi(Z) - E_\theta(\psi(Z)))^t]$. Using this expression, it is possible to derive the following stochastic approximation procedure to approximate $\partial^2_\theta l(\theta)$:

(11)
$$\begin{aligned} G_k &= G_{k-1} + \gamma_k \left(\frac{1}{m(k)} \sum_{j=1}^{m(k)} \left(\partial^2_\theta \log f(z_k(j);\theta_k) + \partial_\theta \log f(z_k(j);\theta_k) \partial_\theta \log f(z_k(j);\theta_k)^t \right) - G_{k-1} \right), \\ H_k &= G_k - \Delta_k \Delta_k^t \end{aligned}$$

Provided the SAEM algorithm converges to a limiting value $\theta^\star$ (see section 6), and $\theta \to l(\theta)$ is sufficiently smooth, H_k converges to $-\partial_\theta^2 l(\theta^\star)$. When $l(\theta)$ is an incomplete data likelihood function of a *regular* statistical experiment (see, e.g., Ibragimov and Has'minski (1981)), the maximum likelihood estimator (i.e., the value of θ that maximizes $l(\theta)$ over the feasible set Θ) is asymptotically normal and the inverse of the observed Fisher information matrix $-\partial_\theta^2 l(\theta^\star)$ converges to the asymptotic covariance of the estimator. Hence, the limiting value of G_k can be used to assess the dispersion of the estimator.

4. Convergence of the EM algorithm for curved exponential families

The convergence of the EM algorithm has been addressed by many different authors, starting with the seminal paper by Dempster et al (1977). Convergence under very general conditions has been established by Wu (1983). Before embarking on the more subtle task of investigating the convergence property of the SAEM algorithm, we briefly recall the basic ingredients needed to prove the convergence of the EM algorithm. We mainly base our discussion on the recent work by Lange (1995). In the sequel, we restrict attention to models for which the unnormalized density $f(z;\theta)$ belongs to the curved exponential family:

- **(M1)** the parameter space Θ is an open subset of $\mathbb{R}^p$. The complete data likelihood function is given by:

$$f(z;\theta) = \exp\left\{-\psi(\theta) + \left\langle \tilde{S}(z), \phi(\theta)\right\rangle\right\} \tag{12}$$

  where $\langle \cdot, \cdot \rangle$ denotes the scalar product, $\tilde{S}(\cdot)$ is a Borel function on $\mathbb{R}^\ell$ taking its values in an open subset $\mathcal{S}$ of $\mathbb{R}^m$. Moreover, the convex hull of $\tilde{S}(\mathbb{R}^\ell)$ is included in $\mathcal{S}$, and, for all $\theta \in \Theta$,

$$\int_{\mathbb{R}^\ell} |\tilde{S}(z)| p(z;\theta)\mu(dz) < \infty. \tag{13}$$

The choice of $\tilde{S}(z)$ is of course not unique. It is generally guided by considerations on the practical implementation of the algorithm. This assumption is often met in situations where the EM algorithm is employed. It implies that the expectation step reduces to the computation of:

$$E_\theta[\tilde{S}(Z)] \triangleq \int_{\mathbb{R}^\ell} \tilde{S}(z) p(z;\theta)\mu(dz).$$

Note that $E_\theta[\tilde{S}(Z)]$ belongs to the convex hull of $\tilde{S}(\mathbb{R}^\ell) \subset \mathcal{S}$. To develop further the theory, some regularity conditions are needed. Define $L : \mathcal{S} \times \Theta \to \mathbb{R}$ as:

$$L(s;\theta) = -\psi(\theta) + \langle s, \phi(\theta)\rangle . \tag{14}$$

- **(M2)** The functions, $\psi(\theta)$ and $\phi(\theta)$ are twice continuously differentiable on Θ.
- **(M3)** The function $\bar{s} : \Theta \to \mathcal{S}$ defined as

$$\bar{s}(\theta) \triangleq \int_{\mathbb{R}^\ell} \tilde{S}(z) p(z;\theta)\mu(dz) \tag{15}$$

  is continuously differentiable on Θ.
- **(M4)** The function $l(\theta)$ is continuously differentiable on Θ and

$$\partial_\theta \int f(z;\theta)\ \mu(dz) = \int \partial_\theta f(z;\theta)\ \mu(dz).$$

- **(M5)** There exists a function $\hat{\theta} : \mathcal{S} \to \Theta$, such that:

$$\forall \theta \in \Theta, \quad \forall s \in \mathcal{S}, \quad L(s; \hat{\theta}(s)) \geq L(s; \theta).$$

Moreover, the function $\hat{\theta}(s)$ is continuously differentiable on $\mathcal{S}$.

In many models of practical interest, the function $\theta \to L(s;\theta)$ has a unique global maximum, and the existence and the differentiability of $s \to \hat{\theta}(s)$ is a direct consequence of the implicit function theorem. Using the notations introduced above, the EM-reestimation functional $Q(\theta|\theta')$ may be expressed as:

$$Q(\theta|\theta') = L(\bar{s}(\theta'); \theta). \tag{16}$$

Using these assumptions, the iteration $\theta_k \to \theta_{k+1}$ of the EM algorithm is defined as:

$$\theta_{k+1} = T(\theta_k) = \hat{\theta}(\bar{s}(\theta_k)). \tag{17}$$

Adaptations of the algorithm to situations where the maximum of $L(s;\theta)$ cannot be computed explicitly will be considered in section 8. The following properties are simple consequences of the basic properties of the EM algorithm (see Dempster et al. (1977)):

- $\theta \to T(\theta) = \hat{\theta}(\bar{s}(\theta))$ is continuously differentiable on the feasible set Θ. Moreover, $l(T(\theta)) \geq l(\theta)$, with equality occuring only when θ is a fixed point of T,
- The set of fixed points of $T(\theta)$ coincides with the set $\mathcal{L}$ of *stationary points* of $l(\theta)$:

$$\mathcal{L} = \{\theta \in \Theta \; ; \partial_\theta l(\theta) = 0\} = \{\theta \in \Theta \; ; \theta = T(\theta)\}. \tag{18}$$

Note that $\mathcal{L}$ is closed. In the terminology of Dempster et al. (1977), $T(\theta)$ is a continuous EM algorithm. The EM recursion may equivalently be written in that case in terms of expectation of the complete data sufficient statistics $s_k \triangleq \bar{s}(\theta_{k-1})$ as:

$$s_{k+1} = \bar{s}(\hat{\theta}(s_k)) \triangleq G(s_k), \tag{19}$$

this form coming up more naturally in the analysis of the SAEM algorithm. Under (M1-5), G also is continuously differentiable on $\mathcal{S}$. For the model above, the convergence of the EM algorithm can be established upon the additional condition that the sequence $\theta_{k+1} = T(\theta_k)$ stays within some compact set in Θ. The following theorem is adapted from Lange (1995) (Proposition 4 and 5; see also Theorems 1-2 and 4 in Wu (1983)). For $\theta \in \Theta$, we denote $\mathcal{L}(\theta)$ the set of limit points of the sequence: $\theta_{k+1} = T(\theta_k)$, starting at $\theta_0 = \theta$. We denote $\mathrm{clos}(A)$ and $\mathrm{int}(A)$ the closure and the interior of the set A, and, when A is a closed set, $d(x, A)$ the distance of x to A.

Theorem 1. *Assume that (M1-5) hold. Assume in addition that for any $\theta \in \Theta$, $\mathrm{clos}(\mathcal{L}(\theta))$ is a compact subset of Θ. Then, for any initial point $\theta_0 = \theta$, the sequence $l(\theta_k)$ is increasing and $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$.*

Note that the compactness of the sequence $\{\theta_k\}_{k\geq 0}$ plays an essential role in the proof. In Lange (1995) (see also Wu (1983)) this property is guaranteed by assuming that $l(\theta)$ is continuous and lower compact in the sense that the level set

$$\{\theta \in \Theta \; : \; -l(\theta) \leq c\} \tag{20}$$

is compact for any $c \geq 0$.

5. General results on Robbins-Monro type stochastic approximation procedures

The SAEM algorithm is a Robbins-Monro type stochastic approximation procedure. These procedures have received considerable attention in the literature (see, for example, Kushner and Clark (1978), Duflo (1996), Kushner and Yin (1997) and the references therein). The most commonly used tool for proving w.p.1 convergence for such algorithm is by far Theorem 2.3.1 by Kushner and Clark (1978) (see Fort and Pagès (1996)). However some conditions of this theorem prove intractable, because the mean field (see definition below) associated with the SAEM algorithm is multimodal in many situations of interest. In the sequel, we use an alternate technique to prove the convergence, which extends results obtained earlier in Delyon (1996). The type of assumptions on which these results are based are of course well suited for the analysis of the SAEM algorithm. However, these results are applicable in a much wider context: this is why we present these results for a general Robbins-Monro (RM) stochastic approximation procedure. We will apply these results to the SAEM algorithm in the next section. Throughout this section, we consider the following RM stochastic approximation procedure:

$$s_n = s_{n-1} + \gamma_n h(s_{n-1}) + \gamma_n e_n + \gamma_n r_n \tag{21}$$

where $\{e_n\}_{n\geq 1}$ and $\{r_n\}_{n\geq 1}$ are random processes defined on the same probability space taking their values in an open subset $\mathcal{X} \subset \mathbb{R}^m$; h is referred to as the mean field of the algorithm; $\{r_n\}_{n\geq 1}$ is a remainder term and $\{e_n\}_{n\geq 1}$ is the stochastic excitation.

Theorem 2. *Assume that:*

- **(SA0)** *w.p. 1, for all* $n \geq 0$, $s_n \in \mathcal{X}$,
- **(SA1)** $\{\gamma_n\}_{n\geq 1}$ *is a decreasing sequence of positive number such that* $\sum_{n=1}^{\infty} \gamma_n = \infty$,
- **(SA2)** *The vector field h is continuous on $\mathcal{X}$ and there exists a continuously differentiable function* $V : \mathcal{X} \to \mathbb{R}$ *such that:*
  1. *for all* $s \in \mathcal{X}$, $F(s) = \langle \partial_s V(s), h(s) \rangle \leq 0$,
  2. $\operatorname{int}(V(\mathcal{L})) = \emptyset$, *where* $\mathcal{L} \triangleq \{s \in \mathcal{X} \ : \ F(s) = 0\}$.
- **(SA3)** *w.p. 1,* $\operatorname{clos}(\{s_n\}_{n\geq 0})$ *is a compact subset of* $\mathcal{X}$,
- **(SA4)** *w.p. 1,* $\lim_{p\to\infty} \sum_{n=1}^{p} \gamma_n e_n$ *exists and is finite,* $\lim_{n\to\infty} r_n = 0$.

Then, w.p. 1, $\overline{\lim}\, d(s_n, \mathcal{L}) = 0$.

The proof of this theorem is given in appendix. Note that this theorem shares most of the assumptions of Theorem 2.3.1 by Kushner and Clark (1978) and Theorem 2.1 in Kushner and Yin (1997). The main difference lies in assumption (SA2), which replaces the following recurrence assumption: "the sequence $\{s_k\}_{k\geq 0}$ is w.p. 1 (or with a probability $\geq \rho$) infinitely often in some compact set in the domain of attraction of an asymptotically stable point $s^\star$ (in the sense of Lyapunov) of h[1]." This assumption is almost impossible to verify in situations where there are more than one stationary point (which is by far the most common situation when dealing with likelihoods). The boundedness of the sequence $\{s_k\}_{k\geq 0}$ does not imply in that case that $\{s_k\}_{k\geq 0}$ is infinitely often in a compact subset of the domain of attraction of a stationary point. Assumption (SA3) implies that, along every trajectory of (21), the sequence $\{s_k\}_{k\geq 0}$ stays in a compact subset (depending upon the trajectory) of $\mathcal{X}$. This assumption (compactness) is often difficult to check, because the behavior of the logarithm of the normalizing function $l(\theta)$ on the boundary

[1] A stationary point of the vector field h is a point $s^\star$ such that $h(s^\star) = 0$. The *domain of attraction* of $s^\star$ is the set $D(s^\star) \subset \mathcal{X}$ such that for every $x \in D(s^\star)$ the ODE $ds(t)/dt = h(s(t))$, with initial condition $s(0) = x$ has a solution for $t \geq 0$ verifying $\lim_{t\to\infty} s(t) = s^\star$.

of the feasible set is most often pathological. This assumption can be replaced by a recurrence condition, provided there exists a Lyapunov function controlling the excursion outside the compact sets of $\mathcal{X}$.

Lemma 1. *Assume (SA0-SA2). Assume in addition:*

- **(STAB1)** *there exists a continuously differentiable function* $W : \mathcal{X} \to \mathbb{R}$ *and a compact set* $\mathcal{K} \subset \mathcal{X}$ *such that:*
  (1) *For all* $c \geq 0$*, the level set* $\{s \in \mathcal{X} \ : W(s) \leq c\}$ *is a compact subset of* $\mathcal{X}$*, and for* $c < c'$*,* $\{s \in \mathcal{X} \ : W(s) \leq c\} \subset \mathrm{int}(\{s \in \mathcal{X} \ : W(s) \leq c'\})$
  (2) $\langle \partial_s W(s), h(s) \rangle < 0$*, for all* $s \in \mathcal{X} \setminus \mathcal{K}$*.*
- **(STAB2)** *for any positive integer* M*, the sequence* $\{\epsilon_n\}_{n \geq 1}$ *defined as:*
$$\epsilon_n = \sum_{k=1}^{n} \gamma_k e_k \, \mathbb{I}(W(s_{k-1}) \leq M)$$
  converges w.p. 1 to a finite limit. In addition, w.p.1,
$$\overline{\lim}_{k \longrightarrow \infty} |r_k| \ \mathbb{I}(W(s_{k-1}) \leq M) = 0.$$
- **(STAB3)** *w.p.1, the sequence* $\{s_k\}_{k \geq 0}$ *admits a limit point in* $\mathcal{X}$*,*

Then, $\overline{\lim}\, W(s_k) < \infty$ *w.p. 1 and assumptions (SA3–4) are satisfied.*

The proof of this result may be adapted from the proof of Theorem 1 in Delyon (1996) (it is omitted for brevity). The function W does not necessarily coincide with the Lyapunov function V used in Theorem 2 (assumption (SA2)). The condition (STAB3) is a recurrence condition: w.p. 1, the sequence $\{s_k\}_{k \geq 0}$ returns infinitely often in a compact set. In some cases the recurrence condition (STAB3) is not verified. A possible solution to overcome this problem consists in implementing an explicit stabilization device. To that purpose, we will present a simple modification of (21), adapting the procedure presented in Chen et al. (1988) (see also Andradottir (1995)). The procedure consists in truncating the original RM recursion (21) : every time s_k is outside a specific set, it is re-initialized at a point chosen at random in a compact set of $\mathcal{X}$. In the technique proposed by Chen et al. (1988), the truncation bounds are random functions of the recursion index k. The advantage of this approach is that the truncation *does not* modify the mean field (and hence, the set of stationary points of the recursion) of the original RM recursion: this is contrast with the traditional projection or truncation approach, where the truncation set is fixed. This is also advantageous from the practical point of view, because the truncation set is selected automatically. The procedure goes as follows. Choose a sequence of compact subsets such that

$$\mathcal{K}_n \subset \mathrm{int}(\mathcal{K}_{n+1}), \quad \mathcal{X} = \bigcup_{n=0}^{\infty} \mathcal{K}_n. \tag{22}$$

Every time s_k wanders out the compact subset $\mathcal{K}_{n_k}$, the sequence is reset at an arbitrary point inside $\mathcal{K}_0$, and the index n_k is increased (n_k is thus the number of projections up to the k-th iteration). More precisely :

$$\begin{aligned} &\hat{s}_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k e_k + \gamma_k r_k, \\ &\text{if } \hat{s}_k \in \mathcal{K}_{n_{k-1}} \ : \begin{cases} s_k = \hat{s}_k, \\ n_k = n_{k-1} \end{cases} \\ &\text{if } \hat{s}_k \notin \mathcal{K}_{n_{k-1}} \ : \begin{cases} s_k = s'_k, \\ n_k = n_{k-1} + 1 \end{cases} \end{aligned} \tag{23}$$

where $\{s'_k\}_{k \geq 0}$ is an arbitrary sequence of random variables taking their values in $\mathcal{K}_0$.

The modified recursion (23) automatically satisfies the recurrence condition (STAB3), because by construction, all the sequences that do not stay in a compact set of $\mathcal{X}$ are infinitely often in the set $\mathcal{K}_0$ and hence have a limit point in that set, since that set is compact. We have the following convergence result:

Theorem 3. *Consider $\{s_k\}_{k\geq 0}$ as given by the truncated RM procedure (23) and assume that (SA0–2) and (STAB1–2) hold. Then, w.p. 1, $\lim d(s_k, \mathcal{L}) = 0$.*

5.1. Rate of convergence, step-size selection and averaging of iterates. The difficulty of selecting an appropriate step-size sequence has long been considered as a serious handicap for practical applications. In a path-breaking paper, Juditski and Polyak (1990) showed that if the step-size γ_n goes to zero slower than $1/n$ (yet fast enough to ensure convergence at a given rate), then the averaged sequence, defined as $n^{-1}\sum_{i=1}^{n} s_i$ converges to its limit at an *optimum* rate (see the comments below). This result implies that we should use *large* step-size (larger than $\gamma_n = 1/n$, typically $\gamma_n = n^{-2/3}$) and an off-line averaging controls the increased noise effect. The practical value of the averaging method have been reported for many different approximations (see Kushner and Yin (1997), chapter 11 for a thorough investigation of the averaging method; see also Delyon and Juditski (1992)). In particular, it happens that this approach tends to robustify the overall procedure, in the sense that the *primary* approximation algorithm (which uses large step-size) is less likely to get stuck at an early stage in a local minimum and more likely to have a faster initial convergence.

The use of an averaging procedure makes sense when the stochastic approximation procedure converges to a *regular stable stationary point*, defined as follows:

Definition: We say that $s^\star$ is a *regular stable stationary point* of the stochastic approximation procedure (21), if **(i)** $h(s^\star) = 0$ and **(ii)** h is twice differentiable in a neighborhood of $s^\star$ and **(iii)** $H(s^\star)$, the Jacobian matrix of h at $s^\star$, is an Hurwitz matrix, i.e., the real parts of the eigenvalues of $H(s^\star)$ are negative.

As evidenced by Kushner and Yin (1997), averaging improves the asymptotic convergence whenever s_n converges to $s^\star$ with a given *rate of convergence*. We say that the *rate of convergence* of s_n about $s^\star$ is $\gamma_n^{1/2}$ if the sequence $\gamma_n^{-1/2}(s_n - s^\star)$ is bounded in probability, *i.e.*

$$\lim_{M\to\infty} \overline{\lim}_{n\to\infty} P(\gamma_n^{-1/2}\|s_n - s^\star\| \ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) \geq M) = 0. \tag{24}$$

To that purpose, we need strengthen the assumptions made on the step size sequence $\{\gamma_n\}_{n\geq 1}$ and on the disturbance processes $\{e_n\}_{n\geq 1}$, $\{r_n\}_{n\geq 1}$. Denote $\mathcal{F} = \{\mathcal{F}_n, n \geq 1\}$, the increasing family of σ-algebra generated by $s_0, e_1, \cdots, e_n, r_1, \cdots, r_n$. For $\rho > 0$, we define $e_n(\rho) \triangleq e_n \mathbb{I}(\|s_{n-1} - s^\star\| \leq \rho)$. The process $\{e_n(\rho)\}$ is adapted to $\mathcal{F}$.

- **(AVE1)** For some $\rho > 0$ such that $\{s \in \mathcal{X} : \|s - s^\star\| \leq \rho\} \subset \mathcal{X}$, there exists a constant $C(\rho) < \infty$, such that, for any deterministic sequence of matrices $\{\alpha_k\}_{k\geq 1}$, and any $n \geq 1$, we have

$$E\left(\left\|\sum_{k=1}^{n} \alpha_k e_k(\rho)\right\|^2\right) \leq C(\rho)\sum_{k=1}^{n}\|\alpha_k\|^2. \tag{25}$$

  In addition, w.p. 1, $\lim_{p\to\infty}\sum_{n=1}^{p}\gamma_n e_n(\rho)$ exists and is finite.
- **(AVE2)** w.p.1, $\overline{\lim}_{n\to\infty}\gamma_n^{-1/2}\|r_n\| \ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) < \infty$,

- **(AVE3(α))** ($1/2 < \alpha < 1$) $\{\gamma_n\}_{n\geq 1}$ is a non-increasing sequence of positive real numbers; in addition, $\lim_{n\to\infty} n^\alpha \gamma_n = \gamma_\star$ and $\gamma_{n+1}/\gamma_n = 1 + O(n^{-1})$.

If $\{e_k\}_{k\geq 1}$ is a L^2-martingale increment (w.r.t $\mathcal{F}$), then $\{e_k(\rho)\}_{k\geq 0}$ also is a L^2-martingale increment and (AVE1) is verified with $C(\rho) = \sup_{k\geq 0} E(\|e_k(\rho)\|^2)$, provided that $\sup_{k\geq 0} E(\|e_k(\rho)\|^2) < \infty$. Assumption (AVE1) is also verified for a large-class of weak-dependent disturbance process $\{e_k\}$ (including Markov chain), by application of the Rosenthal's inequality (see, e.g. Doukhan (1994)).

Under these assumptions, it may be shown that $\gamma_n^{-1/2}(s_n - s^\star)\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0)$ is bounded in probability, provided $s^\star$ is a regular stable stationary point (see Lemma 6). Next, we define the averaged version of the algorithm:

$$\begin{cases} s_n = s_{n-1} + \gamma_n h(s_{n-1}) + \gamma_n e_n + \gamma_n r_n, \\ \bar{s}_n = \bar{s}_{n-1} + n^{-1}(s_{n-1} - \bar{s}_{n-1}). \end{cases} \tag{26}$$

where $\bar{s}_0 = 0$. The basic result for the convergence of the average of the iterates is given below. We denote $\to_P$ and $\to_{\mathcal{L}}$ the convergences in probability and in distribution, and $\mathcal{N}(\mu, \Gamma)$ the multivariate normal distribution with mean μ and covariance matrix Γ.

Theorem 4. *Assume that (AVE1-AVE3(α)) hold, with $1/2 < \alpha < 1$. Assume in addition that $s^\star$ is a stable regular stationary point. Then, $\gamma_n^{-1/2}(s_n - s^\star)\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0)$ is bounded in probability and*

$$\sqrt{n}\left(\bar{s}_n - s^\star + n^{-1}\sum_{k=1}^{n} \nu_k\right)\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) \to_P 0 \tag{27}$$

where $\nu_k \triangleq e_k + r_k$. If, in addition

$$1/\sqrt{n}\sum_{k=1}^{n}\nu_k\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) \to_{\mathcal{L}} \mathcal{N}(0, S)\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) \tag{28}$$

then

$$\sqrt{n}(\bar{s}_n - s^\star)\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) \to_{\mathcal{L}} \mathcal{N}(0, H(s^\star)^{-1} S H(s^\star)^{-1})\ \mathbb{I}(\lim_{n\to\infty}\|s_n - s_\star\| = 0). \tag{29}$$

Remark 1. *As shown in Chickin and Poznyak (1984) and Chickin (1988), the rate obtained in (29) is the best asymptotic rate of convergence. This rate may be achieved using a Gauss-Newton type stochastic approximation algorithm; however, in this latter case, the knowledge of $H(s^\star)$ is needed; this is not required when the averaging of iterates is used.*

Remark 2. *The Central Limit Theorem is checked under a wide variety of conditions. We apply the result in the case where $\{e_n\}_{n\geq 0}$ is a martingale increment; the CLT is thus verified under standard Lindeberg's type conditions (see Hall and Heyde (1980)).*

Remark 3. *When coupled with the stabilization device (Eq. (23)), the averaging procedure should be modified. The modification consists in resetting to zero the averaged value at every time instants the primary stochastic approximation algorithm is restarted, and computing averages from that point. As shown in Theorem 3, the number of time the algorithm is restarted is finite w.p. 1, (27) still holds for the modified algorithm.*

The proof of the Theorem 4 is given in Appendix B. It should be stressed that our result is obtained under much less stringent conditions than those used in Delyon and Juditski (1992) (this contribution deals uniquely with gradient type algorithms; moreover, assumption (A3) of this contribution involves condition on the growth of the Lyapunov function and of the mean field outside the compact sets of $\mathcal{X}$). It also weakens some of the assumptions of Theorem 3.1 (pp. 338) in Kushner and Yin (1997); in particular it avoids the tightness condition (A3.1), which is most often difficult to check (see the remark in appendix B.

6. Convergence of the SAEM algorithm

To avoid cumbersome notations, we set: $m(k) = 1$. The k-th iteration of the SAEM algorithm boils down to:

$$S_k = S_{k-1} + \gamma_k(\tilde{S}(Z_k) - S_{k-1}) \qquad \theta_k = \hat{\theta}(S_k) \tag{30}$$

where $\{Z_k\}_{k\geq 1}$ is the missing value simulated at step k, under the posterior density $p(z;\theta_{k-1})$ (see Eq. (4)). It is assumed that the random variables (r.v.) $S_0, Z_1, Z_2, \cdots$ are defined on the same probability space $(\Omega, \mathcal{A}, P)$; we denote $\mathcal{F} = \{\mathcal{F}_n\}_{n\geq 0}$ the increasing family of σ-algebra generated by the r.v. $S_0, Z_1, Z_2, \cdots, Z_n$. In addition, we assume that:

- **(SAEM1)** For all $k > 0$, $0 \leq \gamma_k \leq 1$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$,
- **(SAEM2)** $l : \Theta \to \mathbb{R}$ and $\hat{\theta} : \mathcal{S} \to \Theta$ are m times differentiable,
- **(SAEM3)** For all positive Borel function ϕ:
$$E[\phi(Z_{k+1})|\mathcal{F}_k] = \int \phi(z)p(z;\theta_k)\mu(dz),$$
- **(SAEM4)** For all $\theta \in \Theta$, $\int ||\tilde{S}(z)||^2 p(z;\theta)\mu(dz) < \infty$, and the function
$$\Gamma(\theta) \triangleq \mathrm{Cov}_\theta(\tilde{S}(Z)) \tag{31}$$
is continuous w.r.t θ.

Since the convex hull of $\tilde{S}(\mathbb{R}^\ell) \subset \mathcal{S}$, (SAEM1) implies that, for all $k \geq 0$, $S_k \in \mathcal{S}$. Assumption (SAEM3) is equivalent to saying that given $\theta_0, \cdots, \theta_k$ the simulated missing observations $Z_1, \cdots, Z_k$ are conditionally independent. This assumption can be relaxed to allow for Markovian dependence, a situation which is typical when Markov Chain Monte Carlo method is used for simulation. We may express the recursion (30) into the Robbins-Monro form (Kushner and Clark (1978)) :

$$S_k = S_{k-1} + \gamma_k h(S_{k-1}) + \gamma_k E_k \tag{32}$$

where $h(s)$ stands for the *mean field* of the algorithm and E_k is a random perturbation

$$h(s) = E_{\hat{\theta}(s)}(\tilde{S}(Z)) - s = \bar{s}(\hat{\theta}(s)) - s \tag{33}$$

$$E_k = \tilde{S}(Z_k) - E[\tilde{S}(Z_k)|\mathcal{F}_{k-1}] = \tilde{S}(Z_k) - \bar{s}(\hat{\theta}(S_{k-1})). \tag{34}$$

Proving the convergence amounts to verify that the assumptions of theorem 2 are satisfied.

Lemma 2. *Assume that (M1-5) and (SAEM2) hold. Then (SA2) is satisfied with $V(s) = -l(\hat{\theta}(s))$. Moreover,*

$$\{s \in \mathcal{S} \ : F(s) = 0\} = \{s \in \mathcal{S} \ : \partial_s V(s) = 0\}, \tag{35}$$

$$\hat{\theta}(\{s \in \mathcal{S} \ : F(s) = 0\}) = \{\theta \in \Theta \ : \partial_\theta l(\theta) = 0\} = \mathcal{L}. \tag{36}$$

where $F(s) \triangleq \langle \partial_s V(s), h(s)\rangle$.

Proof Lemma 2: Under (M3), $\bar{s}(\theta)$ is continuously differentiable on Θ. Under (SAEM2), $\hat{\theta}(s)$ is (m times) continuously differentiable on $\mathcal{S}$ and $h(s) = \bar{s}(\hat{\theta}(s)) - s$ is continuously differentiable on $\mathcal{S}$. Hence, $h(s)$ is bounded on every compact subset of $\mathcal{S}$. Under assumption (M5), for all s in $\mathcal{S}$, the function $\hat{\theta}(s)$ satisfies $\partial_\theta L(s; \hat{\theta}(s)) = 0$; thus

$$- \partial_\theta \psi(\hat{\theta}(s)) + s^t \partial_\theta \phi(\hat{\theta}(s)) = 0 \tag{37}$$

where the superscript ^{t} denotes the transposition. Under (M2-M5), we may differentiate the relation $\partial_\theta L(s; \hat{\theta}(s)) = 0$ with respect to s:

$$\partial_\theta^2 L(s; \hat{\theta}(s)) \ \partial_s \hat{\theta}(s) = -\partial_\theta \phi(\hat{\theta}(s))^t. \tag{38}$$

where ∂_θ^2 denotes the second order derivative w.r.t. θ. On the other hand, for all $\theta \in \Theta$, assumption (M4) implies:

$$\partial_\theta l(\theta) = -\partial_\theta \psi(\theta) + \bar{s}(\theta)^t \ \partial_\theta \phi(\theta).$$

Plugging (37) in the previous expression yields:

$$\begin{aligned}
\partial_\theta l(\hat{\theta}(s)) &= \left(-s + \bar{s}(\hat{\theta}(s))\right)^t \ \partial_\theta \phi(\hat{\theta}(s)) \\
&= h(s)^t \ \partial_\theta \phi(\hat{\theta}(s)) \\
&= -h(s)^t \ \partial_s \hat{\theta}(s)^t \ \partial_\theta^2 L(s; \hat{\theta}(s)) \\
\partial_s l(\hat{\theta}(s)) &= -h(s)^t \ \partial_s \hat{\theta}(s)^t \ \partial_\theta^2 L(s; \hat{\theta}(s)) \ \partial_s \hat{\theta}(s).
\end{aligned}$$

Since, under (M5) $\partial_\theta^2 L(s; \hat{\theta}(s)) \leq 0$, we have:

$$F(s) \triangleq \langle \partial_s V(s), h(s) \rangle = -\left\langle \partial_s l(\hat{\theta}(s)), h(s) \right\rangle = h(s)^t \ \partial_s \hat{\theta}(s)^t \ \partial_\theta^2 L(s; \hat{\theta}(s)) \ \partial_s \hat{\theta}(s) \ h(s) \leq 0, \tag{39}$$

which proves (SA2.1). Note that obviously $\{s : \partial_s V(s) = 0\} \subset \{s : F(s) = 0\}$. On the other hand, let $s^\star \in \mathcal{S}$ be such that $F(s^\star) = 0$; since under (M5) $\partial_\theta^2 L(s^\star; \hat{\theta}(s^\star))$ is a non-positive matrix, (39) implies that $\partial_\theta l(\hat{\theta}(s^\star)) = 0$, and thus $\partial_s V(s^\star) = 0$, showing the reverse inclusion $\{s : F(s) = 0\} \subset \{s : \partial_s V(s) = 0\}$. By (M4-M5), function V is m times continuously differentiable. Sard's theorem (Brocker (1975)) implies that $V(\{s : \partial_s V(s) = 0\}) = 0$ has zero Lebesgue measure, showing (SA2.2). ■

We may now formulate a Theorem which is the stochastic counterpart of the basic convergence Theorem 1 for continuous EM algorithm:

Theorem 5. *Assume that the assumptions (M1-M5) and (SAEM1-SAEM4) hold. Assume in addition, that w.p. 1,* **(A)** $\mathrm{clos}(\{S_k\}_{k \geq 1})$ *is a compact subset of* $\mathcal{S}$. *Then, w.p. 1,* $\lim_{k \to \infty} d(S_k, \{s \in \mathcal{S} : \partial_s V(s) = 0\}) = 0$ *and* $\lim_{k \to \infty} d(\theta_k, \mathcal{L}) = 0$.

Proof of Theorem 5: (SA0) is verified under (M1) and (SAEM1) because $0 < \gamma_k < 1$ and the convex hull of $\tilde{S}(\mathbb{R}^m)$ is included in $\mathcal{S}$. (SA1) is implied by (SAEM1) and (SA3) by (A). Note that under (A), there exists w.p. 1 a compact set $\mathcal{K}$, such that $S_k \in \mathcal{K}$ for a all $k \geq 0$. Denote $M_n = \sum_{k=1}^n \gamma_k E_k$. $\{M_n\}_{n \geq 1}$ is a $\mathcal{F}$-martingale which satisfies under (SAEM1-SAEM3),

$$\sum_{n=1}^{\infty} E[\|M_{n+1} - M_n\|^2 | \mathcal{F}_n] \leq \sum_{n=1}^{\infty} \gamma_{n+1}^2 \int \|\tilde{S}(z)\|^2 p(z; \hat{\theta}(S_n)) \mu(dz) < \infty \quad \text{(w.p.1)}$$

since, by (A(i)-A(ii)) and (M5), w.p. 1 $\hat{\theta}(S_n)$ is in a compact set $\hat{\theta}(\mathcal{K})$ of Θ. This property implies that w.p. 1, $\lim_{n \to \infty} M_n$ exists (see Hall and Heyde (1980), Theorem 2.15 p.33) proving (SA4). Finally, Lemma 2 shows that (SA2) is satisfied and Theorem 2 holds. Since the function $\hat{\theta} : \mathcal{S} \to \Theta$ is continuous, the proof is concluded by applying Lemma 2 (Eq. (35)). ■

Note that, since the function $\hat{\theta} : \mathcal{S} \to \Theta$ is continuous, Theorem 5 and Lemma 2 (Eq. (35)) imply that

$$d(\theta_k, \mathcal{L}) = d(\hat{\theta}(S_k), \hat{\theta}(\{s \in \mathcal{S} : F(s) = 0\})) \to 0, \text{ w.p. } 1 \tag{40}$$

In certain models, the compactness condition (A) of theorem 5 is trivially satisfied. This is the case when the complete data are bounded and the unnormalized density is well behaved. In others situations, checking (A) may prove intractable; we apply in these cases the stabilization procedure presented in section 5. The truncated SAEM algorithm takes the form:

$$\begin{aligned} &\hat{S}_k = S_{k-1} + \gamma_k(\tilde{S}(Z_k) - S_{k-1}), \\ &\text{if } \hat{S}_k \in \mathcal{K}_{n_{k-1}} : \begin{cases} S_k = \hat{S}_k, \ \theta_k = \hat{\theta}(S_k), \\ n_k = n_{k-1} \end{cases} \\ &\text{if } \hat{S}_k \notin \mathcal{K}_{n_{k-1}} : \begin{cases} S_k = S'_k, \ \theta_k = \hat{\theta}(S_k), \\ n_k = n_{k-1} + 1 \end{cases} \end{aligned} \tag{41}$$

where $\{\mathcal{K}_n\}_{n \geq 0}$ is a sequence of compact sets, such that

$$\mathcal{K}_n \subset \text{int}(\mathcal{K}_{n+1}) \quad \text{and} \quad \bigcup_{n=0}^{\infty} \mathcal{K}_n = \mathcal{S} \tag{42}$$

and S'_k is an arbitrary sequence of random variables taking their values in $\mathcal{K}_0$. The convergence of the truncated SAEM algorithm follows from Theorem 3 and Theorem 6.

7. Convergence to local maxima

The results obtained in the previous section demonstrate that, under appropriate conditions, the sequence $\{\theta_k\}_{k \geq 1}$ (defined in (30) or (41)) converges to a connected component of the set $\mathcal{L}$ of stationary points of $l(\theta)$. Assume that the connected components $\mathcal{L}$ are reduced to points: $\mathcal{L} = \bigcup\{\theta^\star\}$, so that $\theta_k \to \theta^\star$. Depending upon the values of the Hessian of l, these stationary points may correspond either to local maxima, local minima or saddle points. Of course, it would be of interest to find conditions upon which the convergence toward local maxima is guaranteed. Such conditions are worked out in this section.

- **(LOC1)** The stationary points of $l(\theta)$ are isolated: any compact subset of $\mathcal{X}$ contains only a finite number of such points.
- **(LOC2)** For every stationary point $\theta^\star \in \mathcal{L}$, the matrices:

  $$E_{\theta^\star}\left[\partial_\theta L(\tilde{S}(Z); \theta^\star)^t \partial_\theta L(\tilde{S}(Z); \theta^\star)\right] \text{ and } \partial_\theta^2 L(E_{\theta^\star}[\tilde{S}(Z)]; \theta^\star)$$

  are positive definite.

Assumption (LOC1) is of course satisfied if the Hessian of $l(\theta)$ at $\theta^\star$ is non-singular. Under the conditions of Theorem 5 or 3, the sequence $\{\theta_k\}_{k \geq 1}$ converges w.p. 1 to a compact and connected subset of $\mathcal{L}$. Under

(LOC1), the connected components of $\mathcal{L}$ are reduced to points so that the sequence $\{\theta_k\}_{k\geq 1}$ converges (w.p.1) pointwise to some point $\theta^\star \in \mathcal{L}$. We may associate to $\theta^\star$ the point in $\mathcal{S}$:

$$s^\star = \bar{s}(\theta^\star) = E_{\theta^\star}\left[\tilde{S}(Z)\right].$$

Heuristically, since $\theta_k \to \theta^\star$, the missing data Z_k (for large enough k) are all simulated under approximately the same posterior distribution $p(z;\theta^\star)$, it is expected that $\{S_k\}_{k\geq 1}$ converges w.p. 1 to $s^\star$.

Theorem 6. *Under the assumptions of Theorem 5 and (LOC1) the sequence $\{S_k\}_{k\geq 1}$ defined by (30) converges to $s^\star = \bar{s}(\theta^\star)$, where $\theta^\star = \lim_{k\to\infty}\theta_k$. $\theta^\star$ and $s^\star$ are fixed points of the EM-mappings $T : \Theta \to \Theta$ (Eq. (17)) and $G : \mathcal{S} \to \mathcal{S}$ (Eq. (19)).*

Proof of Theorem 6: Denote $\delta_k = S_k - s^\star$. We have:

$$\delta_{k+1} = (1-\gamma_{k+1})\delta_k + \gamma_{k+1}(\tilde{S}(Z_{k+1}) - s^\star), \tag{43}$$

$$= \delta_k - \gamma_{k+1}\delta_k + \gamma_{k+1}(\bar{s}(\theta_k) - s^\star) + \gamma_{k+1}E_{k+1} \tag{44}$$

where $E_{k+1} = \tilde{S}(Z_{k+1}) - \bar{s}(\theta_k)$. Since $\{\theta_k\}_{k\geq 1}$ converges w.p.1, the sequence $\{\theta_k\}_{k\geq 1}$ is w.p. 1 bounded and w.p. 1, the martingale $\sum_{k=0}^n \gamma_k E_k$ converges. On the other hand, the continuity of $\bar{s}(.)$ implies that

$$\lim_{k\to\infty} \bar{s}(\theta_k) - s^\star = 0.$$

The sequence $\{\delta_k\}$ converges to zero w.p.1 by application of Theorem 2 with the Lyapunov function $V(x) = -||x||^2$. This proves the convergence of S_k to $s^\star$. The relation $\theta_k = \hat{\theta}(S_k)$ implies that $\theta^\star = \hat{\theta}(s^\star)$ and since $\bar{s}(\theta^\star) = s^\star$, $\theta^\star$ and $s^\star$ are fixed points of T and G. ∎

A similar result can be obtained for the stabilized version of the SAEM algorithm. The regularity of the stationary points (see the definition above) is connected with the stability of the EM-mappings $T : \Theta \to \Theta$ and $G : \mathcal{S} \to \mathcal{S}$ at $\theta^\star \in \mathcal{L}$, and $s^\star = \bar{s}(\theta^\star)$.

Lemma 3. *Assume that (M1-M5) and (LOC1-LOC2) hold. Let $\theta^\star$ be any fixed point of T, and let $s^\star = \bar{s}(\theta^\star)$. Then,*

- **(i)** *$s^\star$ is a fixed point of the mapping G.*
- **(ii)** *$\partial_\theta T(\theta^\star)$ is diagonalizable and its eigenvalues are positive real numbers.*
- **(iii)** *The fixed point $\theta^\star$ is* stable *if it is a proper maximizer of the normalizing function $l(\theta)$. It is hyperbolic if it is a saddle point of $l(\theta)$. It is unstable if it is a proper local maximizer of $l(\theta)$.*
- **(iv)** *$s^\star$ is a* regular stable stationary point *if and only if $\theta^\star$ is a proper maximizer of $\theta \to l(\theta)$.*

Proof of Lemma 3: (i) follows from:

$$G(s^\star) = \bar{s}(\hat{\theta}(\bar{s}(\theta^\star))) = \bar{s}(T(\theta^\star)) = s^\star.$$

(ii) The differential of the mapping T at $\theta^\star$ is given by (see Lange (1995), Eq. 7)

$$\partial_\theta T(\theta^\star) = \partial_\theta^2 L(s^\star;\theta^\star)^{-1}\left[\partial_\theta^2 L(s^\star;\theta^\star) - \partial_\theta^2 l(\theta^\star)\right], \tag{45}$$

$$\partial_\theta^2 l(\theta^\star) - \partial_\theta^2 L(s^\star;\theta^\star) = E_{\theta^\star}\left[\partial_\theta L(\tilde{S}(Z);\theta^\star)^t \partial_\theta L(\tilde{S}(Z);\theta^\star)\right]. \tag{46}$$

Because matrices $-\partial_\theta^2 L(s^\star;\theta^\star)$ and $\partial_\theta^2 l(\theta^\star) - \partial_\theta^2 L(s^\star;\theta^\star)$ are positive definite, $\partial_\theta T(\theta^\star)$ is diagonalizable with strictly positive eigenvalues.

(iii) $\partial_\theta T(\theta^\star)$ has the same eigenvalues (counting multiplicities) than the matrix $A^\star$:

$$A^\star = I + B^\star \quad B^\star = (-\partial_\theta^2 L(s^\star;\theta^\star))^{-1/2}\partial_\theta^2 l(\theta^\star)(-\partial_\theta^2 L(s^\star;\theta^\star))^{-1/2}$$

The proof is concluded by applying the Sylvester law of inertia (Horn and Johnson (1985)), showing that $B^\star$ has the same inertia (number of positive, negative and zero eigenvalues) than $\partial_\theta^2 l(\theta^\star)$.

(iv) Using the defintions (17) and (19) of the mappings T and G, we have:

$$\partial_\theta T(\theta^\star) = \partial_s\hat{\theta}(s^\star)\partial_\theta \bar{s}(\theta^\star) \quad \partial_s G(s^\star) = \partial_\theta \bar{s}(\theta^\star)\partial_s\hat{\theta}(s^\star). \tag{47}$$

It follows from Theorem 1.3.20 in Horn and Johnson (1985) that $\partial_s G(s^\star)$ has the same eigenvalues than $\partial_\theta T(\theta^\star)$ counting multiplicities, together with an additional $(m-p)$ eigenvalues equal to 0. The proof is concluded by noting that $\partial_s h(s^\star) = \partial_s G(s^\star) - I$. ■

Hence, $\theta^\star$ is a stable stationary point of T if and only if $\theta^\star$ is a proper maximizer of l. Otherwise, $\theta^\star$ is hyperbolic or unstable. This is why most often the EM sequences do not converge toward saddle points or local minima (see Murray (1977)). Of course, in the stochastic context, such ill-convergence are avoided automatically: roughly speaking, the stochastic approximation noise prevents the convergence toward hyperbolic or unstable points. The following additional assumption is needed:

- **(LOC3)** : The minimum eigenvalue of the covariance matrix:

  $$R(\theta) = E_\theta\left[(\tilde{S}(Z) - \bar{s}(\theta))(\tilde{S}(Z) - \bar{s}(\theta)^t)\right]$$

  is bounded away from zero for θ in any compact subset $\mathcal{K} \subset \Theta$.

By application of Brandiere and Duflo (1996), the sequence $\{\theta_k\}$ converges w.p. 1 to a proper maximizer of $l(\theta)$, under the assumptions of Theorem 5 and (LOC1-LOC3).

In the situations where the S_k converge to a regular stable stationary point (which is guaranteed under (LOC1-3)), it makes sense to use an averaging procedure. In the application considered here, it is more interesting to compute the average on the parameters themselves. The averaging procedure takes the form:

$$S_k = S_{k-1} + \gamma_k(\tilde{S}(Z_k) - S_{k-1}) \quad \theta_k = \hat{\theta}(S_k), \tag{48}$$

$$\bar{\theta}_{k+1} = \bar{\theta}_k + k^{-1}(\theta_k - \bar{\theta}_k). \tag{49}$$

where γ_k is a sequence such that $\lim_{k\to\infty} k^\alpha \gamma_k = \gamma_\star$ and $\gamma_k/\gamma_{k+1} = 1 + O(k^{-1})$, for some $1/2 < \alpha < 1$. Adaptations of the procedure for the stabilized algorithm is along the same lines. To apply Theorem 4, we need to strengthen the assumptions on the stochastic perturbation $\{E_k\}$. Under (SAEM3), $\{E_k\}_{k\geq 1}$ is a martingale increment; CLT for martingale increments hold under Lyapunov type assumptions.

- **(SAEM4')** For some $\alpha > 0$, $\theta \in \Theta$, $E_\theta[||\tilde{S}(Z)||^{2+\alpha}] < \infty$, and $\Gamma(\theta) \triangleq \mathrm{Cov}_\theta(\tilde{S}(Z))$ is a continuous function of θ.

Under (SAEM1) and (SAEM4'), $n^{-1/2}\sum_{k=1}^n E_k \to_{\mathcal{L}} \mathcal{N}(0, \Gamma(\theta^\star))$, and, by a direct application of Theorem 4, it holds that

$$\sqrt{n}(\bar{\theta}_n - \theta^\star) \to_{\mathcal{L}} \mathcal{N}(0, \Sigma(\theta^\star)), \tag{50}$$

$$\Sigma(\theta^\star) = \partial_s\hat{\theta}(s^\star)H(s^\star)^{-1}\Gamma(\theta^\star)H(s^\star)^{-1}\partial_s\hat{\theta}(s^\star)^t. \tag{51}$$

where $H(s^\star) = \partial_s h(s^\star)$. Some insight may be gained by working-out the previous expression. To that purpose, note that, by application of (38) and (45):

$$\begin{aligned}
\partial_s \hat{\theta}(s^\star) H(s^\star)^{-1} &= \partial_s \hat{\theta}(s^\star) \left(\partial_\theta \bar{s}(\theta^\star) \partial_s \hat{\theta}(s^\star) - I\right)^{-1}, \\
&= \left(\partial_s \hat{\theta}(s^\star) \partial_\theta \bar{s}(\theta^\star) - I\right)^{-1} \partial_s \hat{\theta}(s^\star), \\
&= -[\partial_\theta^2 l(\theta^\star)]^{-1} \partial_\theta^2 L(s^\star, \theta^\star) \partial_s \hat{\theta}(s^\star), \\
&= [\partial_\theta^2 l(\theta^\star)]^{-1} \partial_\theta \phi(\theta^\star)^t.
\end{aligned}$$

Note that, by (46)

$$\begin{aligned}
\mathrm{Cov}_{\theta^\star}(\partial_\theta L(\tilde{S}(Z); \theta^\star)) &= \partial_\theta \phi(\theta^\star)^t \mathrm{Cov}_{\theta^\star}(\tilde{S}(Z)) \partial_\theta \phi(\theta^\star), &(52)\\
&= \partial_\theta^2 l(\theta^\star) - \partial_\theta^2 L(s^\star; \theta^\star) &(53)
\end{aligned}$$

which finally implies that

$$\Sigma(\theta^\star) = [\partial_\theta^2 l(\theta^\star)]^{-1} [\partial_\theta^2 l(\theta^\star) - \partial_\theta^2 L(s^\star; \theta^\star)] [\partial_\theta^2 l(\theta^\star)]^{-1}. \tag{54}$$

In the missing data terminology, the asymptotic covariance of the averaged estimate is thus related to the *missing information*, the difference between the incomplete vs. complete data Fisher information matrix. Note that an estimate of $\Sigma(\theta^\star)$ can be recursively obtained from (11). We may summarize the above results as follow

Theorem 7. *Assume that (SAEM1-3)-(SAEM4') and (LOC1)-(LOC3) hold. Then, the sequence $\{\theta_n\}$ converges to some proper maximizer $\theta^\star$ of $\theta \to l(\theta)$. If in addition, $\lim_{k\to\infty} k^\alpha \gamma_k = \gamma_\star$ and $\gamma_k/\gamma_{k+1} = 1 + O(k^{-1})$, then $\sqrt{n}(\bar{\theta}_n - \theta^\star)\mathbb{I}(\lim_{n\to\infty} \|\theta_n - \theta^\star\| = 0)$ has a limiting distribution :*

$$\sqrt{n}(\bar{\theta}_n - \theta^\star)\mathbb{I}(\lim_{n\to\infty} \|\theta_n - \theta^\star\| = 0) \to_{\mathcal{L}} \mathcal{N}(0, \Sigma(\theta^\star))\ \mathbb{I}(\lim_{n\to\infty} \|\theta_n - \theta^\star\| = 0) \tag{55}$$

where $\Sigma(\theta^\star) = [\partial_\theta^2 l(\theta^\star)]^{-1} [\partial_\theta^2 l(\theta^\star) - \partial_\theta^2 L(s^\star; \theta^\star)] [\partial_\theta^2 l(\theta^\star)]^{-1}$.

In situations where the dispersion of the estimator $\theta^\star$ is given by $\partial_\theta^2 l(\theta^\star)$ (this quantity can be approximated by (11)), the expression (55) gives a practical stopping rule, consisting in comparing an estimate of the simulation variance $\Sigma(\theta^\star)/n$ (where n is the number of stochastic approximation cycles), with $\partial_\theta^2 l(\theta^\star)$.

8. Extensions and discussions

8.1. A stochastic approximation version of the ECM algorithm. It has long been noticed that there is a variety of problems where the maximum of $\theta \to L(s;\theta)$ is hard to compute. In many cases, however, the maximum of $L(s;\theta)$ restricted to particular subsets of Θ are in closed form and relatively easy to obtain. Motivated by this observation, Meng and Rubin (1993) have proposed a modified version of the EM algorithm, named ECM, which maintains the E-step but replaces the M-step by a cycle of conditional maximizations (CM)-steps. For $c = 1, \cdots, C$ and $s \in \mathcal{S}$ and $\eta \in \Theta$, define $\tilde{\theta}_c(s)$ as the value of θ that maximizes $\theta \to L(s;\theta)$ subject to the constraint $g_c(\theta) = g_c(\eta)$ (it is admitted at this point that such value exists and is unique). A full-cycle of CM-step is recursively defined by composing the conditional maximization: starting from a point $(s,\theta) \in \mathcal{S} \times \Theta$, this iteration takes the form for $1 \le c \le C$

$$\hat{\theta}_c(s;\theta) \triangleq \tilde{\theta}_c(s; \hat{\theta}_{c-1}(s;\theta)), \quad \hat{\theta}_0(s;\theta) = \theta. \tag{56}$$

We denote $\hat{\theta}'(s;\theta) \triangleq \hat{\theta}_C(s;\theta)$, that is the value of θ after a full-cycle of conditional maximization. From a global point of view, we have replaced the M-step consisting in the determination of the global maximum $\hat{\theta}(s)$ of the function $L(s;\theta)$ by a cycle of conditional maximizations, which, starting from some point $(s;\theta) \in \mathcal{S} \times \Theta$ leads us to some point $\hat{\theta}'(s;\theta) \in \Theta$. Note that, contrary to $\hat{\theta}'(s;\theta)$, $\hat{\theta}(s)$ does not depend on the *starting point*: this simplifies the analysis of the EM algorithm; not surprisingly, this has of course some implications on the specification and on the convergence analysis of the stochastic version of the EM algorithm. The ECM algorithm for the curved exponential model may be compactly written as

$$\theta_{k+1} = T'(\theta_k) = \hat{\theta}'(\bar{s}(\theta_k);\theta_k). \tag{57}$$

The ECM algorithm belongs to the class of Generalized EM algorithm. The ECM algorithm is monotone in the sense that, at each step the incomplete data likelihood function is increased: $l(\theta_{k+1}) \geq l(\theta_k)$. The convergence of the ECM has been thoroughly analyzed by Meng (1994) (see also the Meng and Rubin (1997), where several extensions of this scheme are investigated); it is shown in these contributions that the ECM sequence is convergent as soon as the set of constraints verify a *space-filling* property: at any point θ, the convex hull of all feasible directions determined by the constraint spaces is the whole Euclidean space $\mathbb{R}^p$ (the resulting maximization by repeated conditional maximizations is over the whole space and not a subspace of it). This space filling condition in fact guarantees that the sequence of iterates $\hat{\theta}^{(i)}(s;\theta)$, defined recursively as

$$\hat{\theta}^{(i)}(s;\theta) = \hat{\theta}'(s;\hat{\theta}^{(i-1)}(s;\theta)), \quad \hat{\theta}^{(0)}(s;\theta) = \theta \tag{58}$$

converges to $\hat{\theta}(s)$, and that the order of convergence is linear (see Luenberger (1984), chapter 7.9, for related results for cyclic maximization procedures). To state the results in its full generality, we will start from this point, rather than from the *space filling* assumption. We will consider the following assumptions

- **(MAX1)** For all $s \in \mathcal{S}$, $\theta \to L(s;\theta)$ has a unique global maximum on Θ, denoted $\hat{\theta}(s)$; moreover, $\hat{\theta}(s)$ is continuously differentiable on $\mathcal{S}$.
- **(MAX2)** There exists a function $\hat{\theta}' : \mathcal{S} \times \Theta \to \Theta$, a function $C : \mathcal{S} \times \Theta \to [0,\infty)$ and a function $\rho : \mathcal{S} \to [0,1)$, such that, for all $(s;\theta) \in \mathcal{S} \times \Theta$, and all $i \in \mathbb{N}$

$$||\hat{\theta}^{(i)}(s;\theta) - \hat{\theta}(s)|| \leq C(s;\theta)\rho(s)^i, \tag{59}$$

  where $\{\hat{\theta}^{(i)}(s;\theta)\}$ are the iterates of $\hat{\theta}(s;\theta)$ defined recursively as

$$\hat{\theta}^{(i)}(s;\theta) = \hat{\theta}'(s;\hat{\theta}^{(i-1)}(s;\theta)) \text{ for } i > 0, \quad \hat{\theta}^{(0)}(s;\theta) = \theta. \tag{60}$$

  Moreover, for all $i \geq 0$, the functions $\hat{\theta}^{(i)} : \mathcal{S} \times \Theta \to \Theta$ are continuous and the functions C and ρ are bounded on the compact subsets of $\mathcal{S} \times \Theta$ and $\mathcal{S}$, respectively.

In many models of practical interests, the function L is for a given $s \in \mathcal{S}$ strictly convex w.r.t θ, and the feasible set Θ also is convex: in that case, (MAX1) is generically verified, while (MAX2) holds, under mild assumptions, for a large variety of iterative maximization procedure, including generalized coordinate ascent methods (as above), steepest ascent methods and / or Gauss-Newton type algorithms. In the latter case, $\hat{\theta}^{(i)}(s;\theta)$ are the iterates of the function $\hat{\theta}(s;\theta)$ defined as

$$\hat{\theta}(s;\theta) = \theta + \alpha(s;\theta)d(s;\theta), \quad d(s;\theta) \triangleq [\partial_\theta^2 L(s;\theta)]^{-1}\partial_\theta L(s;\theta), \tag{61}$$

where $\alpha(s;\theta)$ maximizes $\alpha \to L(s + \alpha d(s;\theta))$ on the interval $[0,1]$. Gauss-Newton type iterates have been considered, in incomplete data problems by Lange (1995), who proposed a gradient-type algorithm equivalent to the EM algorithm using basically (61).

In such context, the SAEM algorithm may be adapted as follows:

$$S_k = S_{k-1} + \gamma_k(\tilde{S}(Z_k) - S_{k-1}), \quad \theta_k = \hat{\theta}(S_k; \theta_{k-1}) \tag{62}$$

where $\{Z_k\}$ is simulated under $p(z;\theta_{k-1})$ conditionally to θ_k independently from the past. The previous equation may be re-expressed as (see (33) and (34)) as

$$S_k = S_{k-1} + \gamma_k h(S_{k-1}; \theta_{k-1}) + \gamma_k E_k \tag{63}$$

where,

$$h(s;\theta) = E_\theta[\tilde{S}(Z)] - s = \bar{s}(\theta) - s, \tag{64}$$

$$E_k = \tilde{S}(Z_k) - E_{\theta_{k-1}}(\tilde{S}(Z)) = \tilde{S}(Z_k) - \bar{s}(\theta_{k-1}). \tag{65}$$

Note that the mean field $h(s)$ of the original SAEM algorithm (see (33)) is given by $h(s) \triangleq h(s;\hat{\theta}(s))$. Because of the dependence in θ, (63) is not in the standard RM form, and the results obtained previously cannot be directly applied. We will use to prove the convergence of this scheme a *state perturbation approach*. Define, for $(s,\theta) \in \mathcal{S} \times \Theta$ and $m \geq n \geq 0$

$$v_{n,m}(s;\theta) \triangleq \sum_{i=0}^{m} \gamma_{i+n+1}[h(s;\hat{\theta}^{(i)}(s;\theta)) - h(s)], \tag{66}$$

and denote $v_n(s;\theta) = \lim_{m\to\infty} v_{n,m}(s;\theta)$ (the convergence of the series is shown in Lemma 7). Define now the *perturbed state* as

$$\tilde{S}_n = S_n + v_n(S_n;\theta_n). \tag{67}$$

By the definition of (63) and (66), the *perturbed state algorithm* may be written as

$$\tilde{S}_k = \tilde{S}_{k-1} + \gamma_k h(S_{k-1};\theta_{k-1}) + \gamma_k E_k + v_k(S_k;\theta_k) - v_{k-1}(S_{k-1};\theta_{k-1}). \tag{68}$$

The last term on the right can be expanded as

$$\begin{aligned} v_k(S_k;\theta_k) - v_{k-1}(S_{k-1};\theta_{k-1}) = \\ -\gamma_k[h(S_{k-1};\theta_{k-1}) - h(S_{k-1})] + v_k(S_k;\theta_k) - v_k(S_{k-1};\hat{\theta}(S_{k-1};\hat{\theta}(S_{k-1};\theta_{k-1})). \end{aligned} \tag{69}$$

Using this latter expression, we may express the perturbed state equation (68) as:

$$\tilde{S}_k = \tilde{S}_{k-1} + \gamma_k h(\tilde{S}_{k-1}) + \gamma_k E_k + \gamma_k R_k \tag{70}$$

where the remainder term R_k is defined as follows

$$R_k = h(S_{k-1}) - h(\tilde{S}_{k-1}) + \gamma_k^{-1}\left[v_k(S_k;\theta_k) - v_k(S_{k-1};\hat{\theta}(S_{k-1};\theta_{k-1}))\right]. \tag{71}$$

This form illustrates the value of the perturbed state approach. The use of the perturbation removes $h(S_{k-1};\theta_{k-1})$ and replaces it by $h(\tilde{S}_{k-1})$. If we may prove that $R_k = o(1)$, then the perturbed state equation will have the same mean field than the original SAEM algorithm, and will thus converge to the same set of stationary points. Then, by adaptating Theorem 6.2 (using Theorem 5.1), we will show that

$$\lim_{n\to\infty} d(\tilde{S}_n, \{s \in \mathcal{S} : \partial_s V(s) = 0\}) = 0, \quad \text{w.p. } 1. \tag{72}$$

This property will in turn implies the convergence of $\{S_n\}$, because Lemma 7 implies that $\tilde{S}_n - S_n = O(\gamma_n)$ w.p. 1 (under the compactness assumption). The discussion is summarized in the following theorem:

Theorem 8. *Assume that the assumptions (M1-M4), (SAEM1-SAEM4) and (MAX1-MAX2) hold. Assume in addition, that w.p. 1,* **(A)** $\mathrm{clos}(\{S_k\}_{k\geq 1})$ *is a compact subset of* $\mathcal{S}$*. Then, w.p. 1,* $\lim_{k\to\infty} d(S_k, \{s \in \mathcal{S} : \partial_s V(s) = 0\}) = 0$ *and* $\lim_{k\to\infty} d(\hat{\theta}(S_k), \mathcal{L}) = 0$.

The proof is in section C. All the issues on the stability, the convergence, and the control of the algorithm can be adapted (using the state perturbation approach). We do not pursue this discussion here.

8.2. Stochastic approximation EM versus Stochastic gradient approaches. One of the oldest and most widely known methods for maximizing a function of several variables is the method of steepest ascent (often referred to as the gradient method). The method of steepest ascent is defined by the iterative algorithm,

$$\theta_k = \theta_{k-1} + \alpha_k g_k, \quad g_k = \partial_\theta l(\theta_{k-1}), \tag{73}$$

where α_k is a non-negative scalar (generally chosen as the maximum of the function $l(\theta_k + \alpha g_k)$). In the incomplete data models, the basic steepest ascent method (73) cannot be directly applied, because $\partial_\theta l(\theta)$ is not in closed form and/or is hard to compute. A stochastic approximation version of the steepest ascent has been proposed for incomplete data models by Younes (1989,1992); the basic version of this algorithm may be written as

$$\theta_k = \theta_{k-1} + \gamma_k \, \partial_\theta \log f(Z_k; \theta_{k-1}) \tag{74}$$

where the *missing* data Z_k is imputed from the current predictive distribution $p(z; \theta_{k-1})$ (conditionally to θ_{k-1} independently from the past; see (SAEM3)). The algorithm (74) shares with the SAEM algorithm the same imputation step. It differs however in the way this imputation step is used to update the parameters. The algorithm Eq. (74) may be written in RM form as

$$\theta_k = \theta_{k-1} + \gamma_k h(\theta_{k-1}) + \gamma_k E_k, \tag{75}$$

with the following definitions

$$\begin{aligned} h(\theta) &= E_\theta[\partial_\theta \log f(Z; \theta)], \\ E_k &= \partial_\theta \log f(Z_k; \theta_{k-1}) - h(\theta_{k-1}). \end{aligned}$$

Note that (under the stated assumptions), E_k is a martingale difference. Under standard regularity assumptions, we have

$$h(\theta) = \partial_\theta l(\theta), \tag{76}$$

a relation which is often referred as the *Fisher's identity.* In that sense, the algorithm (74) is a *stochastic gradient algorithm*, since its mean field is nothing but the gradient of the objective function to maximize. The analysis of the stochastic gradient algorithm (74) can be done using the tools presented in this contribution. The Lyapunov function $V(\theta)$ (needed to check (SA2)) trivially is the negated incomplete data log-likelihood $V(\theta) \triangleq -l(\theta)$:

$$\langle \partial_\theta V(\theta), h(\theta) \rangle = -||\partial_\theta l(\theta)||^2.$$

The set of stationary points of (74) $\{\theta \in \Theta : \langle \partial_\theta V(\theta), h(\theta)\rangle = 0\}$ coincides with the set of stationary points of the incomplete data log-likelihood: $\mathcal{L} = \{\theta \in \Theta : \partial_\theta l(\theta) = 0\}$. By application of the Sard's Theorem, $\mathrm{int}\,(l(\mathcal{L})) = \emptyset$ provided that $l : \Theta \to \mathbb{R}$ is p times continuously differentiable. A basic convergence theorem similar to Theorem 5 can be adapted from Theorem 2. The other issues (stability, re-projection, rate of convergence...) can be addressed along the same lines as above.

When the stationary points of $\theta \to l(\theta)$ are isolated, the procedure (74) converges pointwise to a proper maximizer of $\theta \to l(\theta)$, under mild assumptions (similar to (LOC1-LOC3)). In such case, it makes sense to consider the averaged sequence,

$$\bar{\theta}_n = \bar{\theta}_{n-1} + n^{-1}(\theta_n - \bar{\theta}_{n-1}). \tag{77}$$

where θ_n is given by (74), the step size sequence $\{\gamma_k\}$ being chosen in such a way that $\lim_{k\to\infty} k^\alpha \gamma_k = \gamma_\star$ and $\gamma_k/\gamma_{k+1} = 1+O(k^{-1})$. It follows under basically the same assumptions than Theorem 7 that $\sqrt{n}(\bar\theta_n - \theta^\star)\mathbb{I}(\lim \|\theta_n - \theta^\star\| = 0)$ has a limiting distribution. This limiting distribution can be deduced almost directly from Theorem 4. Note that $\partial_\theta h(\theta^\star) = \partial^2_\theta l(\theta^\star)$; Eq. (46) implies that $\mathrm{Cov}_{\theta^\star}(\partial_\theta L(\tilde S(Z);\theta^\star)) = \partial^2_\theta l(\theta^\star) - \partial^2_\theta L(s^\star;\theta^\star)$ and thus the asymptotic covariance of the averaged estimates is identical to Eq. (54). Thus, the averaged version of the MCEM algorithm and of the gradient algorithm have the same asymptotic rate of covariance $\sqrt{n}$ and the same asymptotic covariance matrix. The choice between these two algorithms is dictated by the about same considerations than for their deterministic counterparts.

Appendix A. Convergence of stochastic approximation

Proof of Theorem 2 We prove the results on a sample path by sample path basis. In all the statements below, the qualifier 'w.p.1' is implicit. The function F is upper semi-continuous and non-positive: the set $\mathcal{L} \triangleq \{s \in \mathcal{S} : F(s) = 0\}$ is closed. Define

$$s'_n = s_n + \sum_{i=n+1}^{\infty} \gamma_i e_i .$$

Rewritting (21) yields to

$$s'_n = s'_{n-1} + \gamma_n h(s_{n-1}) + \gamma_n r_n . \tag{78}$$

Condition (SA3) implies that it exists a compact set $\mathcal{K} \subset \mathcal{X}$ such that $s_n \in \mathcal{K}$, for all $n \geq 0$ (note that this compact set depends upon the trajectory) For $\alpha > 0$, define $\mathcal{K}_\alpha = \{s \in \mathbb{R}^m : d(s,\mathcal{K}) \leq \alpha\}$, where $d(s,\mathcal{K})$ denotes the distance from s to $\mathcal{K}$. Since $\mathcal{X}$ is an open set, there exists $\rho > 0$ such that $\mathcal{K}_\rho \subset \mathcal{X}$. Since, **(i)** under (SA4) $\sum_{i=n+1}^\infty \gamma_i e_i \to 0$ and $r_n \to 0$, **(ii)** under (SA2), the mean field h is bounded on $\mathcal{K}$, this implies :

$$\|s'_n - s'_{n-1}\| = O(\gamma_n). \tag{79}$$

and the segment $[s'_n, s'_{n-1}] \subset \mathcal{K}_\rho$, for n sufficiently large. Under (SA2), the Lyapunov function V is continuously differentiable on $\mathcal{X}$. For n sufficiently large, there exists $s''_{n-1} \in [s'_n, s'_{n-1}]$ such that

$$\begin{aligned} V(s'_n) &= V(s'_{n-1}) + \left\langle \partial_s V(s''_{n-1}), (s'_n - s'_{n-1}) \right\rangle , \\ &= V(s'_{n-1}) + \gamma_n F(s'_{n-1}) + \gamma_n r'_n \end{aligned}$$

with

$$r'_n = \left\langle \partial_s V(s'_{n-1}), h(s_{n-1}) - h(s'_{n-1}) \right\rangle + \left\langle \partial_s V(s''_{n-1}) - \partial_s V(s'_{n-1}), h(s_{n-1}) \right\rangle + \left\langle \partial_s V(s''_{n-1}), r_n \right\rangle .$$

Since h and $\partial_s V(s)$ is continuous on $\mathcal{K}_\rho$ and $\mathcal{K}_\rho$ is compact, h and $\partial_s V(s)$ are uniformly continuous on $\mathcal{K}_\rho$; thus, $r'_n = o(1)$ and

$$V(s'_n) = V(s'_{n-1}) + \gamma_n F(s'_{n-1}) + o(\gamma_n) \tag{80}$$

Note that function F is bounded on $\mathcal{K}_\rho$ and (80) in particular implies that there exists $C_V < \infty$ such that for n sufficiently large:

$$|V(s'_n) - V(s'_{n-1})| \leq C_V \gamma_n . \tag{81}$$

The proof proceeds in two steps: we first show that the sequence $\{V(s'_n)\}_{n\geq 0}$ converges to some point of $V(\mathcal{L})$ **(step 1)**. We then show that the sequence $\{s'_n\}_{n\geq 0}$ converges to $\mathcal{L}$ **(step 2)**.

Step 1: For $\alpha > 0$, define $\mathcal{A}_\alpha = \{x \in \mathbf{R} : \ d(x, V(\mathcal{L} \cap \mathcal{K}_\rho)) \leq \alpha\}$. $V(\mathcal{L} \cap \mathcal{K}_\rho)$ is a compact subset of $\mathbb{R}$, and $\mathcal{A}_\alpha$ is compact: $\mathcal{A}_\alpha$ is a finite union of disjoints closed intervals (of length greater than 2α).

$$\mathcal{A}_\alpha = \bigcup_{i=1}^{N_\alpha} [a_\alpha^{(i)}, b_\alpha^{(i)}], \text{ where } \ a_\alpha^{(1)} < b_\alpha^{(1)} < a_\alpha^{(2)} < \cdots < b^{(N_\alpha)}. \tag{82}$$

Set $\mathcal{N}_\alpha = V^{-1}(\mathrm{int}(\mathcal{A}_\alpha))$. Since V is continuous, $\mathcal{N}_\alpha$ is an open neighborhood of the set $\mathcal{L} \cap \mathcal{K}_\rho$. Since F is upper semi-continuous (any upper semi-continuous function reaches its maximum on any compact sets), there exists $\epsilon_\alpha > 0$ such that for n sufficiently large:

$$s'_{n-1} \in \mathcal{K}_\rho \setminus \mathcal{N}_\alpha \implies F(s'_{n-1}) \leq -2\epsilon_\alpha. \tag{83}$$

This implies by using (80) and (81):

$$V(s'_n) \leq V(s'_{n-1}) - \gamma_n \epsilon_\alpha + \gamma_n (C_V + \epsilon_\alpha) \mathbb{I}(V(s'_{n-1}) \in \mathrm{int}(\mathcal{A}_\alpha)) \tag{84}$$

Since $\sum \gamma_n = \infty$, (84) implies that $\{V(s'_n)\}$ is infinitely often in $\mathcal{A}_\alpha$. $\mathcal{A}_\alpha$ is a finite union of disjoint intervals: thus, $\{V(s'_n)\}$ is infinitely often in a given interval of the partitition (82), say $[a_\alpha, b_\alpha]$ (the superscript is implicit).

Let $\delta > 0$ be such that $[a_\alpha - 2\delta, b_\alpha + 2\delta] \cap \mathcal{A}_\alpha = [a_\alpha, b_\alpha]$ (such δ exists because $\mathcal{A}_\alpha$ is a finite union of closed disjoints intervals), and let N_V be such that, for all $n \geq N_V$, $\gamma_n \leq \delta / C_V$. Finally, let N be the first index greater than N_V such that $V(s'_N) \in [a_\alpha, b_\alpha]$ (such index exists because $V(s'_n)$ is i.o. in $[a_\alpha, b_\alpha]$). Eq. (84) implies that, for all $n \geq N$, $V(s'_n) \in [a_\alpha - \delta, b_\alpha + \delta]$. Hence, δ being arbitrary, the set of limit points $\mathcal{I}$ of the sequence $\{V(s'_n)\}_{n \geq 0}$ is included in the interval $[a_\alpha, b_\alpha]$. Take now $\alpha' < \alpha$. Proceeding as above, we may show that $\mathcal{I} \subset [a_{\alpha'}, b_{\alpha'}]$, where $[a_{\alpha'}, b_{\alpha'}]$ is a given element of the partition (82) of $\mathcal{A}_{\alpha'}$. Of course, we must have $[a_{\alpha'}, b_{\alpha'}] \subset [a_\alpha, b_\alpha]$. Applying the above result with a nonincreasing sequence $\alpha_n \to 0$, we may show $\mathcal{I}$ is included in the intersection of a decreasing sequence of closed intervals, and $\mathcal{I}$ is thus itself a closed interval. Since all the intervals $[a_{\alpha_n}, b_{\alpha_n}]$ are included in $\mathcal{A}_{\alpha_n}$ and $\bigcap_n \mathcal{A}_{\alpha_n} = V(\mathcal{L} \cap \mathcal{K}_\rho)$, $\mathcal{I}$ also is a subset of $V(\mathcal{L} \cap \mathcal{K}_\rho)$. The set $V(\mathcal{L} \cap \mathcal{K}_\rho)$ has an empty interior, and the connected component of $V(\mathcal{L} \cap \mathcal{K}_\rho)$ are reduced to points. Since $\mathcal{I}$ is connected it must be reduced to a point, which implies that $\lim_{n\to\infty} V(s'_n)$ exists. The proof of step 1 is concluded by noting that $\lim_{n\to\infty} V(s'_n) = \lim_{n\to\infty} V(s_n)$.

Step 2 Let $\mathcal{N}$ be an arbitrary open neighborhood of $\mathcal{L}$. By (80),(83), there exists $\epsilon_\mathcal{N} > 0$ such that for n sufficiently large:

$$V(s'_n) \leq V(s'_{n-1}) - \gamma_n \epsilon_\mathcal{N} + \gamma_n (C_V + \epsilon_\mathcal{N}) \mathbb{I}(s'_{n-1} \in \mathcal{N}). \tag{85}$$

Since $V(s'_n)$ converges and $\gamma_n = o(1)$, for any $\tau > 0$, there exists $N_\tau < \infty$ such that

$$|V(s'_n) - V(s'_p)| \leq \tau, \ N_\tau \leq n \leq p \ \text{ and } \ \ \gamma_k < \tau / \epsilon_\mathcal{N}. \ \ N_\tau \leq k. \tag{86}$$

Eqs. (85) and (86) imply that:

$$\epsilon_\mathcal{N} \sum_{k=n+1}^{p} \gamma_k - (C_V + \epsilon_\mathcal{N}) \sum_{k=n+1}^{p} \gamma_k \, \mathbb{I}(s'_{k-1} \in \mathcal{N}) \leq \tau, \ N_\tau \leq n \leq p. \tag{87}$$

Choose $n > N_\tau$ and let $p(n)$ be the first integer larger than n such that:

$$\frac{\tau}{\epsilon_\mathcal{N}} < \sum_{k=n+1}^{p(n)} \gamma_k \leq 2 \frac{\tau}{\epsilon_\mathcal{N}}.$$

Eq. (87) implies that there exists $k(n) \in [n, p(n)]$ such that $s'_{k(n)} \in \mathcal{N}$. By (79), there exists a finite constant K such that, for all $n \geq 1$ $\|s'_n - s'_{n-1}\| \leq K\gamma_k$, which implies:

$$\|s'_{k(n)} - s'_n\| \leq K \sum_{l=n+1}^{k(n)} \gamma_l \leq K \sum_{l=n+1}^{p(n)} \gamma_l \leq 2K \frac{\tau}{\epsilon_{\mathcal{N}}}$$

By construction $s'_{k(n)} \in \mathcal{N}$, and the later relation implies that $d(s'_n, \mathrm{clos}(\mathcal{N})) \leq 2\tau K/\epsilon_{\mathcal{N}}$. Since τ is arbitrary, this implies that

$$\lim_{n\to\infty} d(s'_n, \mathcal{N}) = 0.$$

Since $\mathcal{N}$ is arbitrary, $\{s'_n\}$ tends to $\mathcal{L}$, and so does $\{s_n\}$. The set of limit points of $\{s_n\}$ is compact because it is bounded and closed; it is connected because by (79) $\|s_{n+1} - s_n\| \to 0$. ∎

Proof Theorem 3: The truncated RM procedure (23) can be written in the RM form (21)

$$s_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k e_k + \gamma_n r'_k \tag{88}$$

where the remainder term r'_k is defined as :

$$\text{if } \hat{s}_k \in \mathcal{K}_{n_{k-1}} \ : \begin{cases} r'_k = r_k, \\ n_k = n_{k-1} \end{cases} \tag{89}$$

$$\text{if } \hat{s}_k \notin \mathcal{K}_{n_{k-1}} \ : \begin{cases} r'_k = \gamma_k^{-1}(s'_k - s_{k-1}) - h(s_{k-1}) - e_k \\ n_k = n_{k-1} + 1 \end{cases} \tag{90}$$

We use Lemma 1 to check that the assumptions of Theorem 2 are satisfied. Note that the stochastic approximation procedure defined by Eq. (88) satisfies (SA0–2), (STAB1) (because the projected algorithm has the same mean field h than the original procedure) and also (STAB3) (because the algorithm is by construction recurrent in a compact subset). Also the first condition (STAB2) is directly verified; we have only to check that, for any integer M,

$$\lim_{k\to\infty} \|r'_k\| \mathbb{I}(W(s_{k-1}) \leq M) = 0. \tag{91}$$

If $\hat{s}_k \in \mathcal{K}_{n_{k-1}}$ then $r'_k = r_k$, and (89) shows that

$$\overline{\lim}_{k\to\infty} \|r'_k\| \mathbb{I}(\{W(s_{k-1}) \leq M\}) \mathbb{I}(\hat{s}_k \in \mathcal{K}_{n_{k-1}}) = 0.$$

Thus, we must show that $\mathbb{I}(\hat{s}_k \notin \mathcal{K}_{n_{k-1}}) \mathbb{I}(W(s_{k-1}) \leq M) = 0$ for all but a finite number of indexes k. Since $\{s : W(s) \leq (M+1)\}$ is a compact subset and $\bigcup \mathcal{K}_n = \mathcal{X}$, there exists $N_M < \infty$ such that $\{s : W(s) \leq (M+1)\} \subset \mathcal{K}_n$ for all $n \geq N_M$. Under the stated assumptions, $\{s : W(s) \leq M\}$ is a compact set; thus, there exists w.p. 1 an index K_M such that, for all $k \geq K_M$,

$$W(s_{k-1}) \leq M \ \to \ W(\hat{s}_k) > (M+1).$$

Since, for all $n \geq N_M$, we have $\{s \notin \mathcal{K}_n\} \subset \{W(s) > (M+1)\}$, the previous relation implies that, w.p. 1, for all $k \geq K_M$ and all $n \geq N_M$,

$$W(s_{k-1}) \leq M\} \ \to \ \hat{s}_k \notin \mathcal{K}_n\}.$$

Thus, w.p. 1, $n_k \leq N_M$, and $\mathbb{I}(\{\hat{s}_k \notin \mathcal{K}_{n_{k-1}}) \mathbb{I}(W(s_{k-1}) \leq M) = 0$ only for a finite number of index. This completes the proof of equation (91). ∎

Appendix B. Rate of convergence and averaging of estimates

This appendix is devoted to some technical results needed to derive the limiting distribution of the averaging estimates. These topics are intimately related, because basically, the averaging procedure can be fruitfully applied as soon as the *primary algorithm* converges to a regular stationary point with a given rate of convergence. Rate of convergence and averaging of estimates are extensively studied in Kushner and Yin (1997) Chapter 10 and 11 (see also the references therein). Our main contribution is to weaken some of the assumptions used in this reference, and in particular the tightness conditon ((A3.1), Kushner and Yin (1997), Chapter 11, pp 336). The improvement mainly stems from the decomposition of the error used in Lemma 6 (Eqs. (92) and (93)). We preface this appendix by some additional notations and definitions, and technical results.

Definitions and Notations: For A a square matrix denote $\|A\|_2$ the spectral norm. Let $\{a_n\}_{n\geq 0}$ be a sequence of positive real numbers. We say that $X_n = O_P(a_n)$ if $a_n^{-1}X_n$ is bounded in probability and $X_n = o_P(a_n)$ if $a_n^{-1}X_n$ converges to zero in probability. We say that $X_n = O_q(a_n)$ (where $q > 0$) if $\sup_{n\in\mathbb{N}} \|a_n^{-1}X_n\|_q < \infty$. We say that $X_n \overset{\text{w.p.1}}{=} O(a_n)$ (resp. $X_n \overset{\text{w.p.1}}{=} o(a_n)$ w.p. 1) if $\sup_{n\in\mathbb{N}} a_n^{-1}\|X_n\|$ is finite w.p. 1 (resp. $\overline{\lim}\, a_n^{-1}\|X_n\| = 0$). Similarly, let Y_n be a sequence of random variables. We say that $X_n \overset{\text{w.p.1}}{=} O(Y_n)$ if $Y_n^{-1}X_n \overset{\text{w.p.1}}{=} O(1)$.

The two following lemmas play a key role in the sequel (proofs are omitted for brevity).

Lemma 4. *Let $\{\gamma_k\}_{k\geq 0}$ be a positive sequence such that $\lim_{k\to\infty}\gamma_k = 0$, $\lim_{k\to\infty}\gamma_k^{-1} - \gamma_{k-1}^{-1} = 0$. Let $\{\epsilon_k\}_{k\geq 0}$ be a non-negative sequence. For $b > 0$ and $p \geq 0$ define*

$$\sigma_n = \sum_{i=1}^{n} \gamma_i^{p+1} e^{-b\bar{\gamma}(i,n)}\epsilon_i,$$

where $\bar{\gamma}(i,n) = \sum_{j=i+1}^{n}\gamma_j$ for $0 \leq i < n$. Then, $\overline{\lim}\ \gamma_k^{-p}\sigma_k \leq b^{-1}\overline{\lim}\ \epsilon_k$.

Lemma 5. *Let $\{A_k\}_{k\geq 0}$ be a sequence of square matrix such that $\lim_{k\to\infty}\|A_k - A\|_2 = 0$ where A is an Hurwitz matrix. Let $\{\gamma_k\}_{k\geq 0}$ be a sequence of positive number such that $\lim_{k\to\infty}\gamma_k = 0$. Then, there exist $\beta > 0$ and a constant $C_\beta < \infty$, such that, for all $m > n \geq 0$ it holds that*

$$\|\prod_{k=m+1}^{n}(I + \gamma_k A_k)\|_2 \leq C_\beta e^{-\beta\bar{\gamma}(m,n)}.$$

The crux for obtaining results on the rate of convergence and on the averaged sequence is to approximate the original non-linear difference equation by a linear one, and to bound the error incurred by the linearization. This is the purpose of the following Lemma.

Lemma 6. *Assume that (AVE1), (AVE2) and (AVE3(α)) hold for some $0 < \alpha < 1$, and let $s^\star$ be a regular stable stationary point. Define $\nu_n = e_n + r_n$ and*

$$\mu_n = (I + \gamma_n H(s^\star))\mu_{n-1} + \gamma_n\nu_n, \quad \mu_0 = 0, \tag{92}$$

$$\rho_n = s_n - s^\star - \mu_n. \tag{93}$$

Then

$$\mu_n\ \mathbb{1}(\lim_{n\to\infty}\|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O(\gamma_n^{1/2})O_2(1) \quad \textit{and} \quad \rho_n \overset{\text{w.p.1}}{=} O(\gamma_n)O_1(1). \tag{94}$$

Remark 4. *The expression* $\mu_n \ \mathbb{1}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O(\gamma_n^{1/2}) O_2(1)$ *is a shorthand for* $\|\mu_n\| \leq XY_n$ *where* X *is (non negative) random variable w.p.1 finite and* $\{Y_n\}_{n\geq 0}$ *is a sequence of r.v. bounded in* L^2*. The same notational convention holds for* ρ_n*. Note that* μ_n *can be seen as a leading term in the expansion of the error* $s_n - s^\star$*, whereas* ρ_n *is a remainder term.*

Proof Lemma 6: By Taylor expansion at $s^\star$ and since $h(s^\star) = 0$, it holds, for n sufficiently large,

$$s_n - s^\star = s_{n-1} - s^\star + \gamma_n H^\star (s_{n-1} - s^\star) + \gamma_n \nu_n + \gamma_n \epsilon_n$$

where $H^\star \triangleq H(s^\star)$ and $\epsilon_n \triangleq [\epsilon_{n,1}, \cdots, \epsilon_{n,m}]^t$ is defined component-wise as

$$\epsilon_{n,i} = \sum_{k,l=1}^{m} R_{n,i}(k,l)(s_{n-1,k} - s_k^\star)(s_{n-1,l} - s_l^\star),$$

$$R_{n,i}(k,l) = \int_0^1 \frac{(1-t)^2}{2!} \frac{\partial^2 h_i}{\partial s_k \partial s_l}(s_{n-1} + t(s^\star - s_{n-1}))dt.$$

Using the definition (92), we have

$$\rho_n = (I + \gamma_n H^\star)\rho_{n-1} + \gamma_n \epsilon_n = (I + \gamma_n H_n)\rho_{n-1} + \gamma_n \epsilon_n',$$

where $H_n = ([H_n(i,j)])_{1\leq i,j\leq m}$ and $\epsilon_n' = [\epsilon_{n,1}', \cdots \epsilon_{n,m}']^t$ are respectively defined as

$$H_n(i,j) \triangleq H^\star(i,j) + \sum_{k=1}^{m} (2R_{n,i}(j,k)\mu_{n-1,k} + R_{n,i}(j,k)\rho_{n-1,k}),$$

$$\epsilon_{n,i}' \triangleq \sum_{k,l=1}^{m} R_{n,i}(k,l)\mu_{n-1,k}\mu_{n-1,l}.$$

Denote for $n > k$, $\psi_\star(n,k) = (I + \gamma_n H^\star)\cdots(I + \gamma_{k+1}H^\star)$ ($\psi_\star(n,n) = I$ and $\psi_\star(n,k) = 0$ if $k > n$). By Lemma 5, there exists $\beta > 0$ and $C'(H^\star, \beta) < \infty$, such that, for all $n \geq k \geq 0$:

$$\|\psi_\star(n,k)\|_2 \leq C'(H^\star, \beta) e^{-\beta\bar{\gamma}(n,k)}$$

Decompose μ_n (see (92)) as $\mu_n = \mu_n^{(0)} + \mu_n^{(1)}$ where

$$\mu_n^{(0)} = \sum_{k=1}^{n} \gamma_k \psi_\star(n,k) e_k(\rho), \tag{95}$$

$$\mu_n^{(1)} = \sum_{k=1}^{n} \gamma_k \psi_\star(n,k) r_k'. \tag{96}$$

where $e_k(\rho) = e_k \mathbb{1}(\|s_{k-1} - s^\star\| \leq \rho)$ and $r_k' = r_k + e_k \mathbb{1}(\|s_{k-1} - s^\star\| > \rho)$. Note that r_k' verifies (AVE2). Under (AVE1), we have

$$[E(\|\mu_n^{(0)}\|^2)]^{1/2} \leq \ C'(H^\star, \beta)\ C^{1/2}(\rho)\ \left(\sum_{k=1}^{n} \gamma_k^2 e^{-2\beta\bar{\gamma}(k,n)}\right)^{1/2} \tag{97}$$

and thus $\mu_n^{(0)} = O_2(\gamma_n^{1/2})$, by application of Lemma 4. An application of the Abel's transform shows that

$$\mu_n^{(0)} = \psi_\star(n,1)E(n,1) + \sum_{k=2}^{n} \gamma_k \psi_\star(n,k) H^\star E(n,k) \tag{98}$$

where $E(n,k) \triangleq \sum_{j=k}^{n} \gamma_j e_j(\rho)$ for $k \leq n$, $E(n,k) = 0$ otherwise. Let δ be such that $\alpha < \delta < 1$. Under (AVE1), it holds that, w.p. 1, $\sup_{(n,k)\in\mathbb{N}\times\mathbb{N}} \|E(n,k)\| < \infty$ and $\lim_{n\to\infty} \sup_{n-[n^\delta]\leq k\leq n} \|E(n,k)\| = 0$. Under (AVE3($\alpha$)), we have

$$\sum_{k=1}^{n-[n^\delta]} \gamma_k \|\psi_\star(n,k)\| \leq C'(H^\star,\beta)\ (n-[n^\delta]) e^{-\beta[n^\delta]\gamma_{n-[n^\delta]}}$$

and thus, w.p.1,

$$\left\| \sum_{k=1}^{n-[n^\delta]} \gamma_k \psi_\star(n,k) H^\star E(n,k) \right\| \to_{n\to\infty} 0,$$

$$\left\| \sum_{k=n-[n^\delta]+1}^{n} \gamma_k \psi_\star(n,k) H^\star E(n,k) \right\| \leq \sup_{n-[n^\delta]\leq k\leq n} \|E(n,k)\|\ \|H^\star\|_2 \sum_{k=1}^{n} \gamma_k e^{-\beta\bar{\gamma}(k,n)} \to_{n\to\infty} 0$$

showing that $\mu_n^{(0)} \overset{\text{w.p.1}}{=} o(1)$. Similarly, we have

$$\|\mu_n^{(1)}\| \leq C'(H^\star,\beta) \sum_{k=1}^{n} \gamma_k^{3/2} e^{-\beta\bar{\gamma}(k,n)} \gamma_k^{-1/2} \|r_k'\| \tag{99}$$

and, by application of Lemma 4, $\mu_n^{(1)}\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O(\gamma_n^{1/2})$. Eqs. (97) and (99) together imply that $\mu_n\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O_2(\gamma_n^{1/2}) + O(\gamma_n^{1/2})$.

Define $\psi(n,k) = (I + \gamma_n H_n)\cdots(I + \gamma_{k+1} H_{k+1})$ for $n > k$ ($\psi(n,n) = I$, $\psi(n,k) = 0$, for $n < k$). Since $\mu_n\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} o(1)$, we have $\|H_n - H(s^\star)\|_2\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} o(1)$. By Lemma 5, there exist $\beta > 0$ and a w.p.1 finite r.v. $C(H,\beta) < \infty$, such that, for all $0 \leq k \leq n$,

$$\|\psi(n,k)\|_2\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \leq C(H,\beta)\ e^{-\beta\bar{\gamma}(k,n)}, \quad \text{w.p.1.} \tag{100}$$

Notice that

$$\rho_n = \sum_{k=1}^{n} \gamma_k \psi(n,k) \epsilon_k' + \psi(n,0) s_0$$

and $\epsilon_k'\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O(\|\mu_{k-1}\|^2)$; by applying (100) there exists a r.v. $K(\beta,\rho)$ w.p. 1 finite such that, for all $n \geq 1$,

$$\begin{aligned}\|\rho_n\|\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \leq K(\beta,\rho) \sum_{k=1}^{n} \gamma_k e^{-\beta\bar{\gamma}(k,n)} \|\mu_{k-1}\|^2\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) +\\ K(\beta,\rho) e^{-\beta\bar{\gamma}(0,n)} \|s_0 - s^\star\|\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0).\end{aligned} \tag{101}$$

By application of Lemma 4, we have

$$E\left(\sum_{k=1}^{n} \gamma_k e^{-\beta\bar{\gamma}(k,n)} \|\mu_{k-1}^{(0)}\|^2 \right) = O(\gamma_n), \tag{102}$$

$$\sum_{k=1}^{n} \gamma_k e^{-\beta\bar{\gamma}(k,n)} \|\mu_{k-1}^{(1)}\|^2\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O(\gamma_n), \tag{103}$$

$$e^{-\beta\bar{\gamma}(0,n)} \|s_0 - s^\star\|\ \mathbb{I}(\lim_{n\to\infty} \|s_n - s_\star\| = 0) \overset{\text{w.p.1}}{=} O(\gamma_n) \tag{104}$$

which concludes the proof. ■.

Remark 5. *This Lemma serves for Theorem 4 the same purpose than assumptions (A3.1)-(A3.8) for Theorem 3.1 in Kushner and Yin (1997) (pp. 336-338); Note that the relation (94) does not imply (A3.1) (restated with our notations, (A3.1) requires that* $(s_n - s^\star)\mathbb{I}(\|s_n - s^\star\| \le \rho) = O_2(\gamma_n^{1/2})$*). The assumptions used in the previous lemma are weaker than those required to show (A3.1).*

Proof of Theorem 4: Write $s_n - s^\star = \mu_n + \rho_n$ where μ_n and ρ_n are defined in (92) and (93), respectively. Note that $\bar{s}_n - s^\star = \bar{\mu}_n + \bar{\rho}_n$, where $\bar{\mu}_n$ and $\bar{\rho}_n$ denote the Cesaro's mean of μ_n and ρ_n,

$$\bar{\mu}_n \triangleq n^{-1} \sum_{k=1}^{n} \mu_k, \text{ and, } \|\bar{\rho}_n\| \triangleq n^{-1} \sum_{k=1}^{n} \rho_k .$$

The proof is in two steps: **(i)** we show that $\sqrt{n}\bar{\rho}_n = o_P(1)$ and **(ii)** we then show that $\sqrt{n}(\bar{\mu}_n - H(s^\star)^{-1}\bar{\nu}_n) = o_P(1)$, where $\bar{\nu}_n$ is the Cesaro's mean of ν_n, $\bar{\nu}_n \triangleq n^{-1}\sum_{k=1}^n \nu_k$.

(i) Applying Lemma 6 (Eq. (94)), there exist a non-negative w.p. 1 bounded r.v. X, and a sequence of non-negative r.v. $\{Y_n\}_{n\in\mathbb{N}}$ such that $Y_n = O_1(1)$ and $\rho_k \overset{\text{w.p.1}}{\le} \gamma_k X Y_k$. Property **(i)** follows from

$$\sqrt{n}\bar{\rho}_n \overset{\text{w.p.1}}{\le} X\left(n^{-1/2}\sum_{i=1}^{n} \gamma_i Y_i\right) = O(1)O_1(n^{1/2-\alpha}).$$

(ii) Using Eq. (92), we may write, for $n \ge 1$: $\mu_{n-1} = H(s^\star)^{-1}(\gamma_n^{-1}(\mu_n - \mu_{n-1}) - \nu_n)$. Thus,

$$\begin{aligned}\bar{\mu}_n &= H(s^\star)^{-1}\left(-n^{-1}\sum_{k=1}^{n}\nu_{k+1} + n^{-1}\sum_{k=1}^{n}\gamma_{k+1}^{-1}(\mu_{k+1}-\mu_k)\right),\\ &= H(s^\star)^{-1}\left(-n^{-1}\sum_{k=1}^{n}\nu_{k+1} - n^{-1}\gamma_2^{-1}\mu_1 + n^{-1}\gamma_{n+1}^{-1}\mu_{n+1} + n^{-1}\sum_{k=2}^{n}(\gamma_k^{-1}-\gamma_{k+1}^{-1})\mu_k\right).\end{aligned}$$

The proof is concluded by showing that $n^{-1/2}\sum_{k=2}^n(\gamma_k^{-1} - \gamma_{k+1}^{-1})\mu_k = o_P(1)$. Using Lemma 6, Eq. (94), there exists a non-negative r.v. X and a sequence of non-negative r.v.s $\{Y_n\}$ such that $Y_n = O_2(1)$ and, for all $k \ge 1$, $\|\mu_k\| \overset{\text{w.p.1}}{\le} \gamma_k^{1/2} X Y_k$. Under (AVE3($\alpha$)), $\gamma_k^{1/2}|\gamma_k^{-1} - \gamma_{k+1}^{-1}| \le C_\gamma k^{\alpha/2-1}$, for some constant $C_\gamma < \infty$. Thus,

$$n^{-1/2}\sum_{k=2}^{n}(\gamma_k^{-1}-\gamma_{k+1}^{-1})\|\mu_k\| \overset{\text{w.p.1}}{\le} C_\gamma\, X n^{-1/2}\sum_{k=2}^{n} k^{\alpha/2-1}Y_k$$

and the proof is concluded by noting that $n^{-1/2}\sum_{k=2}^n k^{\alpha/2-1}Y_k = O_1(n^{(\alpha-1)/2}) = o_P(1)$.

Appendix C. Miscellaneous results

Lemma 7. *Assume (M1-M4) and (MAX1-2). Then, for all* $(s,\theta) \in \mathcal{S}\times\Theta$, $\sum_{i=0}^\infty \gamma_{i+n+1}\|h(s;\hat{\theta}^{(i)}(s;\theta)) - h(s)\| < \infty$ *and*

$$\|v_n(s;\theta)\| \le C(s;\theta)\,[1-\rho(s)]^{-1} \sup_{\{\theta'\in\Theta:\|\theta'-\hat{\theta}(s)\|\le C(s;\theta)\}} \|\partial_\theta \bar{s}(\theta)\|_2\, \gamma_{n+1} \tag{105}$$

where $v_n(s;\theta) = \lim_{m\to\infty} v_{n,m}(s;\theta)$*. Moreover, the set of functions* $\{\gamma_n^{-1}v_n(s;\theta)\}_{n\ge 0}$ *is equicontinuous,* i.e., *for any compact subsets* $\mathcal{K}_\mathcal{S} \subset \mathcal{S}$ *and* $\mathcal{K}_\Theta \subset \Theta$ *and for any* $\epsilon > 0$*, there is* $\delta > 0$*, such that for all* $n \ge 0$,

$$\sup_{\{s,s'\in\mathcal{K}_\mathcal{S},\|s-s'\|\le\delta\}}\ \sup_{\{\theta,\theta'\in\mathcal{K}_\Theta,\|\theta-\theta'\|\le\delta\}} \left[\gamma_n^{-1}\|v_n(s;\theta) - v_n(s';\theta')\|\right] \le \epsilon. \tag{106}$$

The proof of this Lemma is straightforward and is omitted.

Proof of Theorem 8: As mentionned above, we need to prove that **(i)** the remainder term $R_n = o(1)$ w.p. 1. Under the compactness assumption (A(i)-(ii)), there exists w.p. 1 a compact set $\mathcal{K}_\mathcal{S} \subset \mathcal{S}$ such that the sequence $\{S_n\}$ is in $\mathcal{K}_\mathcal{S}$ for all n (note that this compact set *depends on the trajectory*). Assumptions (MAX1)-(MAX2) then implies that there exists w.p. 1 a compact set $\mathcal{K}_\Theta \subset \Theta$ such that $\{\theta_n\}_{n\in\mathbb{N}} \subset \mathcal{K}_\Theta$. Under (SAEM1-3), $\sum_{k=0}^n \gamma_k E_k$ converges w.p. 1 (see theorem 6.2). This implies that, w.p. 1,

$$S_n - S_{n-1} = o(1), \tag{107}$$

$$\theta_n - \hat{\theta}(S_{n-1};\theta_{n-1}) = \hat{\theta}(S_n;\theta_{n-1}) - \hat{\theta}(S_{n-1};\theta_{n-1}) = o(1). \tag{108}$$

Using Lemma 7, we thus have under (MAX2) that $v_n(s_n;\theta_n) = O(\gamma_n)$ w.p. 1, which implies, under (M3) and (MAX1) $h(S_n) - h(\tilde{S}_n) = O(\gamma_n)$ w.p. 1 ($h(s)$ is continuously differentiable on $\mathcal{S}$). The proof of **(i)** then follows from the equicontinuity of the set of functions $\{\gamma_n^{-1} v_n(s;\theta)\}_{n\in\mathbb{N}}$ (Lemma 7). Eqs. (107) and (108) together with (106) imply indeed that

$$\gamma_n^{-1}(v_n(S_n;\theta_n) - v_n(S_{n-1};\hat{\theta}(S_{n-1};\theta_{n-1}))) = o(1).$$

which concludes the proof. ∎

References

[1] S. Andradottir, A stochastic approximation algorithm with varying bounds, *Oper. Res.*, pp. 1037–1048, 1995.

[2] J.C. Biscarat. Almost sure convergence of a class of stochastic approximation algorithms. *Stoch. Proc. Appl.*, 50:83–100, 1994.

[3] O. Brandiere and M. Duflo. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'Institut H. Poincaré*, 32(3):395–427, 1996.

[4] T. Brocker. *Differentiable germs and catastrophes.* L.M.S. Lecture Note Series 17, Cambridge University Press, 1975.

[5] G. Celeux and J. Diebolt, A probabilistic teacher algorithm for iterative maximum likelihood estimation. Classification and Related Methods of Data Analysis, (H. H. Bock, edr.), 617-623. Amsterdam: North-Holland, 1988.

[6] G. Celeux and J. Diebolt. A stochastic aprroximation type EM algorithm for the mixture problem. *Stochastics and Stochastics reports*, 41:127–146, 1992.

[7] H.F. Chen, L. Guo, and A.J. Gao. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stoch. Proc. and their Applications*, 27:217–231, 1988.

[8] D.O. Chickin and A.S. Poznyak, On the asymptotical properties of the stochastic approximation procedure under dependent noise, *Autom. and Remote Control*, 44(2), 1984

[9] D.O. Chickin, On the convergence of the stochastic approximation procedure under dependent noise, *Autom. and Remote Control*, 48(3), 1988

[10] P. de Jong and N. Shephard. The simulation smoother for time series model. *Biometrika*, 82(2):339–350, 1995.

[11] B. Delyon. General results on stochastic approximation. to appear in the IEEE Trans. on Automatic Control, 1996.

[12] B. Delyon and A. Juditski, Stochastic approximation with averaging of trajectories, *Stochastics Stochastics Rep.*, 39:107–118, 1992.

[13] P. Doukhan, *Mixing: properties and examples* Lecture Notes in Statistics, Springer-Verlag Eds., 1994

[14] J. Diebolt and G. Celeux. Asymptotic properties of a stochastic EM algorithm for estimating mixture proportions. *Stochastic Models* 9, 599-613, 1993

[15] J. Diebolt and E.H.S. Ip, A stochastic EM algorithm for approximating the maximum likelihood estimate. In: Markov Chain Monte Carlo in Practice (W.R. Gilks, S.T. Richardson, D.J. Spiegelhalter, eds.). Chapman and Hall.

[16] M. Duflo, *Random Iterative Models,* Application of Mathematics, 34, Springer-Verlag Berlin Heidelberg, 1997.

[17] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38, 1977.

[18] G. Fort, E. Moulines, P. Soulier, On the convergence of iterated random maps, with applications to the MCEM algorithm, To appear in the proceedings of the COMPSTAT conference,1998

[19] J.C. Fort and G. Pagès. Convergence of stochastic algorithms: from Kushner-Clark theorem to the Lyapounov functional method. *Adv .appl. prob.*, 28:1072–1094, 1996.

[20] C.J. Geyer and E.A. Thompson, Constrained Monte-Carlo maximum likelihood for dependent data, *J.R. Statist. Soc. B*, 54:657–699, 1992.

[21] C.J. Geyer, On the convergence of Monte-Carlo maximum likelihoods calculations *J. R. Statist. Soc. B*, 56:261–274, 1994.

[22] C.J. Geyer, Likelihood Inference for Spatial Point Processes. To appear in Current Trends in Stochastic Geometry and its Applications, eds. W. S. Kendall, O. E. Barndorff-Nielsen and M. C. van Lieshout, London: Chapman and Hall.

[23] P. Green. On the use of the EM algorithm for penalized likelihood estimation. *J.R. Statist. Soc. B*, 52(3):443–452, 1990.

[24] P. Hall and C.C. Heyde. *limit theory and its applications.* New-York: Academic Press, 1980.

[25] R. Horn and C. Johnson. *Matrix analysis.* Cambridge University Press, 1985.

[26] I. Ibragimov and R. Has'minski, *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New-York, 1981

[27] H. Kushner and D. Clark. *Stochastic approximation for constrained and unconstrained systems.* Applied math. science series. Springer-Verlag, 1978.

[28] H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Applications of Mathematics, Stochastic Modelling and Applied Probability, Springer-Verlag, 1997

[29] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. R. Statist. Soc. B*, 75(2):425–437, 1995.

[30] C. Liu and D. Rubin. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81:633–648, 1994.

[31] P. Loti-Viaud. Random perturbations of recursive sequences with an application to an epidemic model. *J. Applied. Prob*, 32:559–578, 1995.

[32] T.A. Louis. Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. B*, 44:226–233, 1982.

[33] I. Meilijson. A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc. B*, 51:127–138, 1989.

[34] X. Meng and D. Rubin. Maximum likelihood via the ECM algorithm: a general framework. *Biometrika*, 80:267–278, 1993.

[35] X. Meng. On the rate of convergence of the ECM algorithm. *The Annals of Stat.*, 22(1), pp. 326–339, 1994.

[36] G. Murray. in discussion on the paper by. Pr. Dempster et al: Maximum-Likelihood from incomplete data via the EM algorithm, 1977. J. R. Statist. Soc., pp. 27–28.

[37] J. Ortega. *Numerical analysis: a second course.* Soc. for Industrial and Applied Math., 1990.

[38] B.T. Polyak. New stochastic approximation type procedures. *Automatica i Telemekh.*, pages 98–107, 1990. translated in Autom. & Remote Contr. vol. 51(7).

[39] B.T. Polyak and A.B. Juditski, Acceleration of stochastic approximation by averaging *SIAM J. Control Optim.*, 30:838–855, 1992.

[40] C. Rao. *Linear statistical inference and its applications.* Wiley, New York, 1965.

[41] N. Shephard. Partial non-gaussian state space. *Biometrika*, 81(1):115–131, 1994.

[42] M. Tanner. *Tools for statistical inference: methods for exploration of posterior distributions and likelihood functions.* Springer-Verlag: Springer series in statistics, 1993.

[43] G. Wei and M. Tanner. A Monte-Carlo implementation of the EM algorithm and the Poor's Man's data augmentation algorithm. *J. Amer. Stat. Assoc.*, 85:699–704, 1990.

[44] C. Wu. On the convergence property of the EM algorithm. *The Annals of Stat.*, 11:95–103, 1983.

[45] L. Younes. Parameter estimation for imperfectly observed Gibbsian fields. *Prob. Theory and Rel. fields*, 82:625-645, 1989

[46] L .Younes. Parameter estimation for imperfectly observed Gibbs fields and some comments on Chalmond's EM Gibssian algorithm. *Proc. of Stoch. models, statistical methods and algorithms in image analysis*, Ed. P. Barone and A. Frigessi, Lecture Notes in Statistics, Springer.