Algorithmes de type Expectation-Maximization en ligne pour l'estimation dans les modèles de Markov cachés (article)

The Expectation Maximization (EM) algorithm is a versatile tool for model parameter estimation in latent data models. When processing large data sets or data stream however, EM becomes intractable since it requires the whole data set to be available at each iteration of the algorithm. In this contribution, a new generic online EM algorithm for model parameter inference in general Hidden Markov Model is proposed. This new algorithm updates the parameter estimate after a block of observations is processed (online). The convergence of this new algorithm is established, and the rate of convergence is studied showing the impact of the block-size sequence. An averaging procedure is also proposed to improve the rate of convergence. Finally, practical illustrations are presented to highlight the performance of these algorithms in comparison to other online maximum likelihood procedures.

6.1 Introduction

A hidden Markov model (HMM) is a stochastic process $\{X_k, Y_k\}_{k\geq 0}$ in $\mathbb{X} \times \mathbb{Y}$, where the state sequence $\{X_k\}_{k\geq 0}$ is a Markov chain and where the observations $\{Y_k\}_{k\geq 0}$ are independent conditionally on $\{X_k\}_{k\geq 0}$. Moreover, the conditional distribution of Y_k given the state sequence depends only on X_k. The sequence $\{X_k\}_{k\geq 0}$ being unobservable, any statistical inference task is carried out using the observations $\{Y_k\}_{k\geq 0}$. These HMM can be applied in a large variety of disciplines such as financial econometrics ([Mamon et Elliott, 2007]), biology ([Churchill, 1992]) or speech recognition ([Juang et Rabiner, 1991]).

The Expectation Maximization (EM) algorithm is an iterative algorithm used to solve maximum likelihood estimation in HMM. The EM algorithm is generally simple to implement since it relies on complete data computations. Each iteration is decomposed into two steps: the E-step computes the conditional expectation of the complete data log-likelihood given the observations and the M-step updates the parameter estimate based on this conditional expectation. In many situations of interest, the complete data likelihood belongs to the curved exponential family. In this case, the E-step boils down to the computation of the conditional expectation of the complete data sufficient statistic. Even in this case, except for simple models such as linear Gaussian models or HMM with finite state-spaces, the E-step is intractable and has to be approximated e.g. by Monte Carlo methods such as Markov Chain Monte Carlo methods or Sequential Monte Carlo methods (see [Carlin *et al.*, 1992] or [Cappé *et al.*, 2005, Doucet *et al.*, 2001] and the references therein).

However, when processing large data sets or data streams, the EM algorithm might become impractical. *Online* variants of the EM algorithm have been first proposed for independent and identically distributed (i.i.d.) observations, see [Cappé et Moulines, 2009]. When the complete data likelihood belongs to the cruved exponential family, the E-step is replaced by a stochastic approximation step while the M-step remains unchanged. The convergence of this online variant of the EM algorithm for i.i.d. observations is addressed by [Cappé et Moulines, 2009]: the limit points are the stationary points of the Kullback-Leibler divergence between the marginal distribution of the observation and the model distribution.

An online version of the EM algorithm for HMM when both the observations and the states take a finite number of values (resp. when the states take a finite number of values) was recently proposed by [Mongillo et Denève, 2008] (resp. by [Cappé, 2011a]). This algorithm has been extended to the case of general state-space models by substituting deterministic approximation of the smoothing probabilities for Sequential Monte Carlo algorithms (see for example [Cappé, 2009, Del Moral *et al.*, 2010a, Le Corff *et al.*, 2011b]). There do not exist convergence results for these online EM algorithms for general state-space models (some insights on the asymptotic behavior are nevertheless given in [Cappé, 2011a]): the introduction of many approximations at different steps of the algorithms makes the analysis quite challenging.

In this contribution, a new online EM algorithm is proposed for HMM with complete data likelihood belonging to the curved exponential family.

This algorithm sticks closely to the principles of the original batch-mode EM algorithm. The M-step (and thus, the update of the parameter) occurs at some deterministic times $\{T_k\}_{k\geq 1}$ i.e. we propose to keep a fixed parameter estimate for blocks of observations of increasing size. More precisely, let $\{T_k\}_{k\geq 0}$ be an increasing sequence of integers ($T_0 = 0$). For each $k \geq 0$, the parameter's value is kept fixed while accumulating the information brought by the observations $\{Y_{T_k+1}, \cdots, Y_{T_{k+1}}\}$. Then, the parameter is updated at the end of the block. This algorithm is an online algorithm since the sufficient statistics of the k-th block can be computed on the fly by updating an intermediate quantity when a new observation Y_t, $t \in \{T_k + 1, \dots, T_{k+1}\}$ becomes available. Such recursions are provided in recent works on online estimation in HMM, see [Cappé, 2009, Cappé, 2011a, Del Moral *et al.*, 2010a].

This new algorithm, called *Block Online EM* (BOEM) is derived in Section 6.2 together with an *averaged* version. Section 6.3 is devoted to practical applications: the BOEM algorithm is used to perform parameter inference in HMM where the forward recursions mentioned above are available explicitly. In the case of finite state-space HMM, the BOEM algorithm is compared to a gradient-type recursive maximum likelihood procedure and to the online EM algorithm of [Cappé, 2011a]. The convergence of the BOEM algorithm is addressed in Section 6.4. The BOEM algorithm is seen as a perturbation of a deterministic *limiting EM* algorithm which is shown to converge to the stationary points of the limiting relative entropy (to which the true parameter belongs if the model is well specified). The perturbation is shown to vanish (in some sense) as the number of observations increases thus implying that the BOEM algorithms inherits the asymptotic behavior of the *limiting EM* algorithm. Finally, in Section 6.5, we study the rate of convergence of the BOEM algorithm as a function of the block-size sequence. We prove that the averaged BOEM algorithm is rate-optimal when the block-size sequence grows polynomially. All the proofs are postponed to Section 6.6; supplementary proofs and comments are provided in Appendix A.

6.2 The Block Online EM algorithms

6.2.1 Notations and Model assumptions

Our model is defined as follows. Let Θ be a compact subset of $\mathbb{R}^{d_\theta}$. We are given a family of transition kernels $\{M_\theta\}_{\theta\in\Theta}$, $M_\theta : \mathbb{X} \times \mathcal{X} \to [0, 1]$, a positive σ-finite measure μ on $(\mathbb{Y}, \mathcal{Y})$, and a family of transition densities with respect to μ, $\{g_\theta\}_{\theta\in\Theta}$, $g_\theta : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}_+$. For each $\theta \in \Theta$, define the transition kernel K_θ on $\mathbb{X} \times \mathbb{Y}$ by

$$K_\theta\left[(x,y),C\right] \stackrel{\text{def}}{=} \int \mathbf{1}_C(x',y')\, g_\theta(x',y')\, \mu(\mathrm{d}y')\, M_\theta(x,\mathrm{d}x')\ .$$

Denote by $\{X_k, Y_k\}_{k\geq 0}$ the canonical coordinate process on the measurable space $\left((\mathbb{X}\times\mathbb{Y})^{\mathbb{N}}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes\mathbb{N}}\right)$. For any $\theta \in \Theta$ and any probability distribution χ on $(\mathbb{X},\mathcal{X})$, let $\mathbb{P}_\theta^\chi$ be the probability distribution on $((\mathbb{X}\times\mathbb{Y})^{\mathbb{N}}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes\mathbb{N}})$ such that $\{X_k, Y_k\}_{k\geq 0}$ is Markov chain with initial distribution $\mathbb{P}_\theta^\chi((X_0,Y_0)\in C) = \int \mathbf{1}_C(x,y)\, g_\theta(x,y)\, \mu(\mathrm{d}y)\, \chi(\mathrm{d}x)$ and transition kernel K_θ. The expectation with respect to $\mathbb{P}_\theta^\chi$ is denoted by $\mathbb{E}_\theta^\chi$. Throughout this paper, it is assumed that the Markov transition kernel K_θ has a unique invariant distribution π_θ (see below for further comments). For the stationary Markov chain with initial distribution π_θ, we write $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ instead of $\mathbb{P}_\theta^{\pi_\theta}$ and $\mathbb{E}_\theta^{\pi_\theta}$. Note also that the stationary Markov chain $\{X_k, Y_k\}_{k\geq 0}$ can be extended to a two-sided Markov chain $\{X_k, Y_k\}_{k\in\mathbb{Z}}$.

It is assumed that, for any $\theta\in\Theta$ and any $x\in\mathbb{X}$, $M_\theta(x,\cdot)$ has a density $m_\theta(x,\cdot)$ with respect to a finite measure λ on $(\mathbb{X},\mathcal{X})$. Define the complete data likelihood by

$$p_\theta(x_{0:T}, y_{0:T}) \stackrel{\text{def}}{=} g_\theta(x_0,y_0) \prod_{i=0}^{T-1} m_\theta(x_i,x_{i+1}) g_\theta(x_{i+1},y_{i+1})\ , \tag{6.1}$$

where, for any $u\leq s$, we will use the shorthand notation $x_{u:s}$ for the sequence $(x_u,\cdots,x_s)$. For any probability distribution χ on $(\mathbb{X},\mathcal{X})$, any $\theta\in\Theta$ and any $s\leq u\leq v\leq t$, we have

$$\mathbb{E}_\theta^\chi\left[f(X_{u:v})|Y_{s:t}\right] = \int f(x_{u:v})\phi^\chi_{\theta,u:v|s:t}(\mathrm{d}x_{u:v})\ ,$$

where $\phi^\chi_{\theta,u:v|s:t}$ is the so-called fixed-interval smoothing distribution. We also define the fixed-interval smoothing distribution when $X_s\sim\chi$:

$$\mathbb{E}_\theta^{\chi,s}\left[f(X_{u:v})|Y_{s+1:t}\right] \\ = \frac{\int\prod_{i=s+1}^t\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}f(x_{u:v})\chi(\mathrm{d}x_s)\lambda(\mathrm{d}x_{s+1:t})}{\int\prod_{i=s+1}^t\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}\chi(\mathrm{d}x_s)\lambda(\mathrm{d}x_{s+1:t})}\ .$$

Given an initial distribution χ on $(\mathbb{X},\mathcal{X})$ and $T+1$ observations $Y_{0:T}$, the EM algorithm maximizes the so-called incomplete data log-likelihood $\theta\mapsto\ell^\chi_{\theta,T}$ defined by

$$\ell^\chi_{\theta,T}(\mathbf{Y}) \stackrel{\text{def}}{=} \log\int p_\theta(x_{0:T},Y_{1:T})\chi(\mathrm{d}x_0)\lambda(\mathrm{d}x_{1:T})\ . \tag{6.2}$$

The central concept of the EM algorithm is that the intermediate quantity defined by

$$\theta\mapsto Q(\theta,\theta') \stackrel{\text{def}}{=} \mathbb{E}^\chi_{\theta'}\left[\log p_\theta(X_{0:T},Y_{0:T})|Y_{0:T}\right]$$

may be used as a surrogate for $\ell^{\chi}_{\theta,T}(Y_{0:T})$ in the maximization procedure. Therefore, the EM algorithm iteratively builds a sequence $\{\theta_n\}_{n\geq 0}$ of parameter estimates following the two steps:

i) Compute $\theta \mapsto Q(\theta, \theta_n)$.

ii) Choose θ_{n+1} as a maximizer of $\theta \mapsto Q(\theta, \theta_n)$.

In the sequel, it is assumed that there exist functions S, ϕ and ψ such that (see A1 for a more precise definition), for any $(x, x') \in \mathbb{X}^2$ and any $y \in \mathbb{Y}$,

$$m_\theta(x, x')g_\theta(x', y) = \exp\left\{\phi(\theta) + \left\langle S(x, x,', y), \psi(\theta)\right\rangle\right\} \ .$$

Therefore, the complete data likelihood belongs to the curved exponential family and the step i) of the EM algorithm amounts to computing

$$\theta \mapsto Q(\theta, \theta_n) = \phi(\theta) + \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}^{\chi}_{\theta_n}\left[S(X_{t-1}, X_t, Y_t)|Y_{0:T}\right], \psi(\theta) \right\rangle \ ,$$

where $\langle \cdot, \cdot \rangle$ is the scalar product on $\mathbb{R}^d$. It is also assumed that for any $s \in \mathcal{S}$, where $\mathcal{S}$ is an appropriately defined set, the function $\theta \mapsto \phi(\theta) + \langle s, \psi(\theta)\rangle$ has a unique maximum denoted by $\bar{\theta}(s)$. Hence, a step of the EM algorithm writes

$$\theta_n = \bar{\theta}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}^{\chi}_{\theta_{n-1}}\left[S(X_{t-1}, X_t, Y_t)|Y_{0:T}\right]\right) \ .$$

6.2.2 The Block Online EM (BOEM) algorithms

We now derive an online version of the EM algorithm. Define $\bar{S}^{\chi,T}_{\tau}(\theta, \mathbf{Y})$ as the intermediate quantity of the EM algorithm computed with the observations $Y_{T+1:T+\tau}$:

$$\bar{S}^{\chi,T}_{\tau}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau}\sum_{t=T+1}^{T+\tau} \mathbb{E}^{\chi,T}_{\theta}\left[S(X_{t-1}, X_t, Y_t)|Y_{T+1:T+\tau}\right] \ . \tag{6.3}$$

Let $\{\tau_n\}_{n\geq 1}$ be a sequence of positive integers such that $\lim_{n\to\infty} \tau_n = +\infty$ and set

$$T_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} \tau_k \quad \text{and} \quad T_0 \stackrel{\text{def}}{=} 0 \ ; \tag{6.4}$$

τ_n denotes the length of the n-th block. Given an initial value $\theta_0 \in \Theta$, the BOEM algorithm defines a sequence $\{\theta_n\}_{n\geq 1}$ by

$$\theta_n \stackrel{\text{def}}{=} \bar{\theta}\left[S_{n-1}\right] \ , \text{ and } S_{n-1} \stackrel{\text{def}}{=} \bar{S}^{\chi_{n-1}, T_{n-1}}_{\tau_n}(\theta_{n-1}, \mathbf{Y}) \ , \tag{6.5}$$

where $\{\chi_n\}_{n\geq 0}$ is a family of probability distributions on $(\mathbb{X}, \mathcal{X})$. By analogy to the regression problem, an estimator with reduced variance can be obtained by averaging and weighting the successive estimates (see for example [Kushner et Yin, 1997, Polyak et Juditsky, 1992] for a discussion on the averaging procedures). Define $\Sigma_0 \stackrel{\text{def}}{=} 0$ and for $n \geq 1$,

$$\Sigma_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{j=1}^{n} \tau_j \, S_{j-1} \; . \tag{6.6}$$

Note that this quantity can be computed iteratively and does not require to store the past statistics $\{S_j\}_{j=0}^{n-1}$. Given an initial value $\widetilde{\theta}_0$, the averaged BOEM algorithm defines a sequence $\{\widetilde{\theta}_n\}_{n\geq 1}$ by

$$\widetilde{\theta}_n \stackrel{\text{def}}{=} \bar{\theta}\,(\Sigma_n) \; . \tag{6.7}$$

The algorithm above relies on the assumption that S_n can be computed in closed form. In the HMM case, this property is satisfied only for linear Gaussian models or when the state-space is finite. In all other cases, S_n cannot be computed explicitly and will be replaced by a Monte Carlo approximation $\widetilde{S}_n$. Several Monte Carlo approximations can be used to compute $\widetilde{S}_n$. The convergence properties of the Monte Carlo BOEM algorithms rely on the assumption that the Monte Carlo error can be controlled on each block. [Le Corff et Fort, 2011a] provides examples of applications when Sequential Monte Carlo algorithms are used. Hereafter, we use the same notation $\{\theta_n\}_{n\geq 0}$ and $\{\widetilde{\theta}_n\}_{n\geq 0}$ for the original BOEM algorithm or its Monte Carlo approximation.

Our algorithms update the parameter after processing a block of observations. Nevertheless, the intermediate quantity S_n can be either exactly computed or approximated in such a way that the observations are processed online. In this case, the intermediate quantity S_n or $\widetilde{S}_n$ is updated online for each observation. Such an algorithm is described in [Cappé, 2011a, Section 2.2] and [Del Moral *et al.*, 2010b, Proposition 2.1] and can be applied either to finite state-space HMM or to linear Gaussian models. A Sequential Monte Carlo approximation to compute $\widetilde{S}_n$ online for more complex models is proposed in [Del Moral *et al.*, 2010a] (see also [Le Corff et Fort, 2011a]).

The classical theory of maximum likelihood estimation often relies on the assumption that the "true" distribution of the observations belongs to the specified parametric family of distributions. In many cases, it is doubtful that this assumption is satisfied. It is therefore natural to investigate the convergence of the BOEM algorithms and to identify the possible limit for misspecified models i.e. when the observations $\mathbf{Y}$ are from an ergodic process which is not necessarily an HMM.

6.3 Application to inverse problems in Hidden Markov Models

In Section 6.3.1, the performance of the BOEM algorithm and its averaged version are illustrated in a truncated linear Gaussian model. In Section 6.3.2, the BOEM algorithm is compared to online maximum likelihood procedures in the case of finite state-space HMM.

Other applications of the Monte Carlo BOEM algorithm to more complex models with online Sequential Monte Carlo methods can be found in [Le Corff et Fort, 2011a].

6.3.1 Linear Gaussian Model

Consider the linear Gaussian model:

$$X_{t+1} = \phi X_t + \sigma_u U_t \;, \qquad\qquad Y_t = X_t + \sigma_v V_t \;,$$

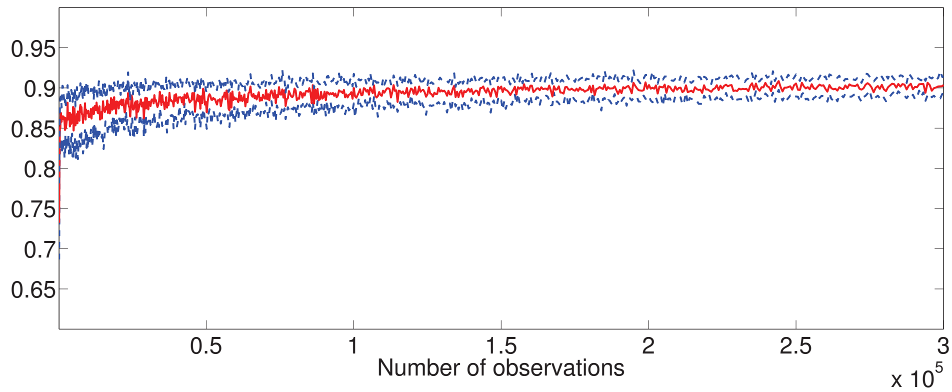
where $X_0 \sim \mathcal{N}\left(0, \sigma_u^2(1-\phi^2)^{-1}\right)$, $\{U_t\}_{t\geq 0}, \{V_t\}_{t\geq 0}$ are independent i.i.d. standard Gaussian r.v., independent from X_0. Data are sampled using $\phi = 0.9$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 2$.

We illustrate the convergence of the BOEM algorithms. We choose $\tau_n = n^{1.1}$. We display in Figure 6.1 the median and lower and upper quartiles for the estimation of ϕ obtained with 100 independent Monte Carlo experiments. Both the BOEM algorithm and its averaged version converge to the true value $\phi = 0.9$; the averaging procedure clearly improves the variance of the estimation.

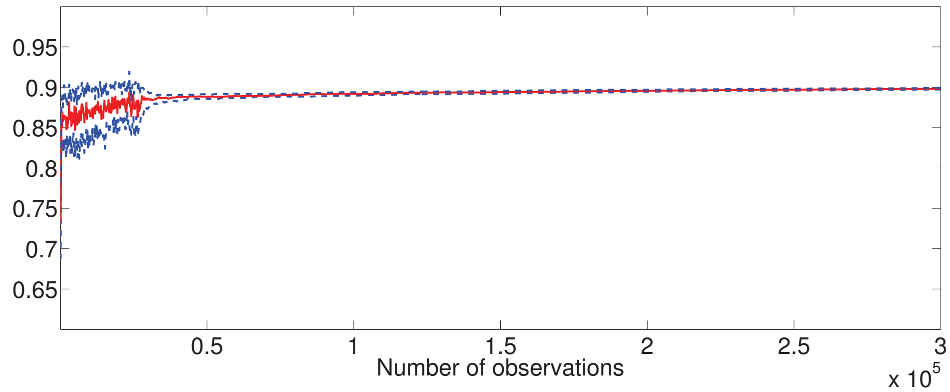
We now discuss the role of $\{\tau_n\}_{n\geq 0}$. Figure 6.2 displays the empirical variance, when estimating ϕ, computed with 100 independent Monte Carlo runs, for different numbers of observations and, for both the BOEM algorithm and its averaged version. We consider four polynomial rates $\tau_n \sim n^b$, $b \in \{1.2, 1.8, 2, 2.5\}$. Figure 6.2a shows that the choice of $\{\tau_n\}_{n\geq 0}$ has a great impact on the empirical variance of the (non averaged) BOEM path $\{\theta_n\}_{n\geq 0}$. To reduce this variability, a solution could consist in increasing the block sizes τ_n at a larger. The influence of the block size sequence τ_n is greatly reduced with the averaging procedure as shown in Figure 6.2b. We will show in Section 6.5 that averaging really improves the rate of convergence of the BOEM algorithm.

6.3.2 Finite state-space HMM

We consider a Gaussian mixture process with Markov dependence of the form: $Y_t = X_t + V_t$ where $\{X_t\}_{t\geq 0}$ is a Markov chain taking values in $\mathbb{X} \stackrel{\text{def}}{=} \{x_1, \ldots, x_d\}$, with initial distribution χ and a $d \times d$ transition matrix m. $\{V_t\}_{t\geq 0}$ are i.i.d. $\mathcal{N}(0, v)$ r.v., independent from $\{X_t\}_{t\geq 0}$, i.e., for all



(a) The BOEM algorithm without averaging.



(b) The BOEM algorithm with averaging.

Figure 6.1: Estimation of ϕ.

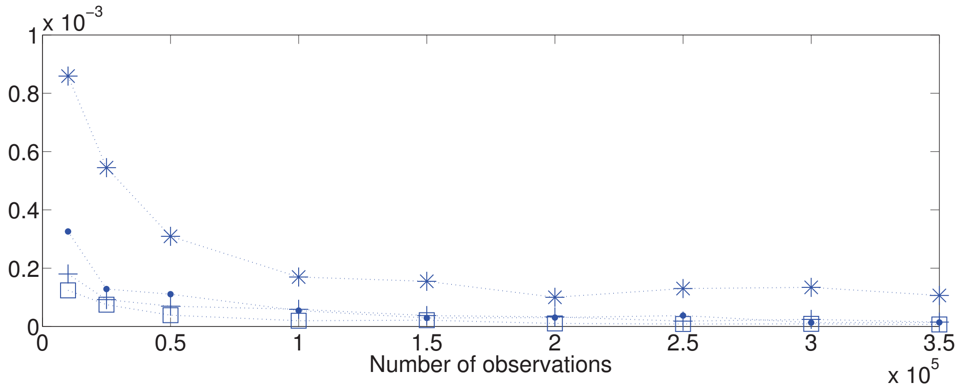
$(x, y) \in \mathbb{X} \times \mathbb{Y}$,

$$g_\theta(x, y) \stackrel{\text{def}}{=} (2\pi v)^{-1/2} \exp\left\{-\frac{(y-x)^2}{2v}\right\} ,$$

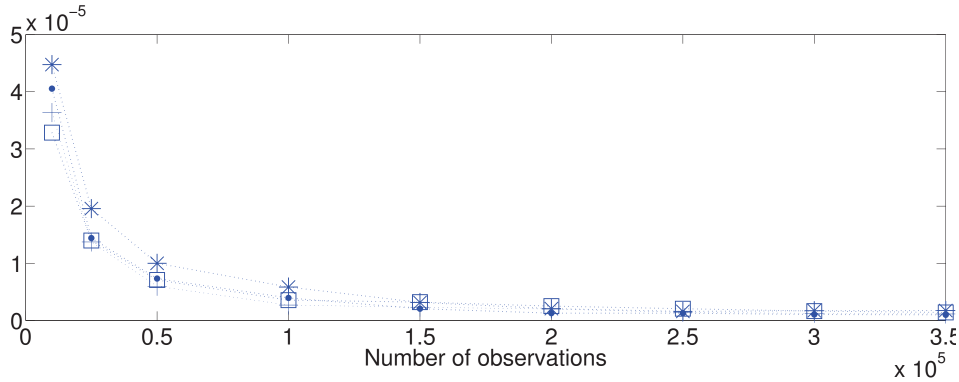
where $\theta \stackrel{\text{def}}{=} \left(v, x_{1:d}, (m_{i,j})_{i,j=1}^d\right)$. In the experiments below, the initial distribution below is chosen as the uniform distribution on $\mathbb{X}$. The statistics used to estimate θ are, for all $(i, j) \in \{1, \cdots, d\}$ and all $(x, x') \in \mathbb{X}^2$,

$$\begin{aligned} S^{i,0}(x, x', y) &= \mathbf{1}_{x_i}(x') , & S^{i,1}(x, x', y) &= y\mathbf{1}_{x_i}(x') , \\ S^{i,2}(x, x', y) &= y^2\mathbf{1}_{x_i}(x') , & S_{i,j}(x, x', y) &= \mathbf{1}_{x_i}(x)\mathbf{1}_{x_j}(x') . \end{aligned} \tag{6.8}$$

The online computation of these intermediate quantities can be found in [Cappé, 2011a, Section 2.2]. The computations below are performed for each statistic in (6.8). Define, for all $x \in \mathbb{X}$, $\phi_0(x) = \chi(x)$ and $\rho_0(x) = 0$.



(a) The BOEM algorithm, without averaging



(b) The BOEM algorithm, with averaging

Figure 6.2: The BOEM algorithm: empirical variance of the estimation of ϕ after $n = 0.5\ell \cdot 10^5$ observations ($\ell \in \{1, \cdots, 7\}$) for different block size schemes $\tau_n \sim n^{1.2}$ (stars), $\tau_n \sim n^{1.8}$ (dots), $\tau_n \sim n^2$ (crosses) and $\tau_n \sim n^{2.5}$ (squares).

i) For $t \in \{1, \cdots, \tau\}$, compute, for any $x \in \mathbb{X}$,

$$\phi_t(x) = \frac{\sum_{x' \in \mathbb{X}} \phi_{t-1}(x') m_{x',x} g_\theta(x, Y_{t+T})}{\sum_{x',x'' \in \mathbb{X}} \phi_{t-1}(x') m_{x',x''} g_\theta(x'', Y_{t+T})} ,$$

and

$$r_t(x, x') = \frac{\phi_{t-1}(x') m_{x',x}}{\sum_{x'' \in \mathbb{X}} \phi_{t-1}(x'') m_{x'',x}} .$$

$$\rho_t(x) = \sum_{x' \in \mathbb{X}} \left[\frac{1}{t} S(x, x', Y_{t+T}) + \left(1 - \frac{1}{t}\right) \rho_{t-1}(x') \right] r_t(x, x') .$$

ii) Set

$$\bar{S}_\tau^{\mathbb{X},T}(\theta, \mathbf{Y}) = \sum_{x \in \mathbb{X}} \rho_\tau(x) \phi_\tau(x) .$$

At the end of the block, the new estimate is given, for all $(i, j) \in \{1, \cdots, d\}^2$ by (the dependence on $\mathbf{Y}$, θ, χ, T and τ is dropped from the notation)

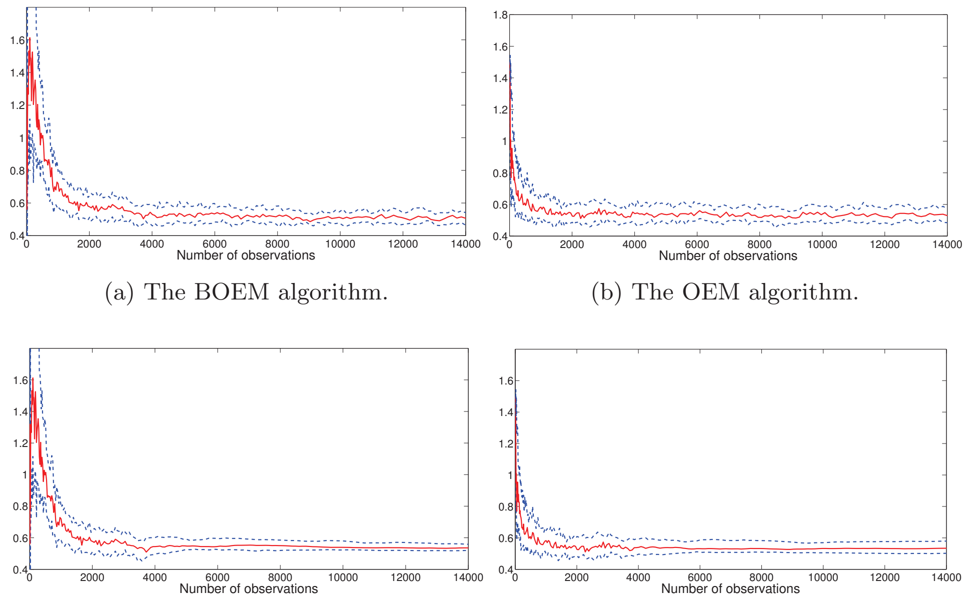
$$m_{i,j} = \frac{\bar{S}_{i,j}}{\sum_{j=1}^{d} \bar{S}_{i,j}}\ ,\ x_i = \frac{\bar{S}^{i,1}}{\bar{S}^{i,0}},\ v = \sum_{i=1}^{d} \bar{S}^{i,2} + \sum_{i=1}^{d} x_i^2 \bar{S}^{i,0} - 2\sum_{i=1}^{d} x_i \bar{S}^{i,1}\ .$$

Observations are sampled using $d = 6$, $v = 0.5$, $x_i = i\ , \forall i \in \{1, \ldots, d\}$ and the true transition matrix is given by

$$m = \begin{pmatrix} 0.5 & 0.05 & 0.1 & 0.15 & 0.15 & 0.05 \\ 0.2 & 0.35 & 0.1 & 0.15 & 0.05 & 0.15 \\ 0.1 & 0.1 & 0.6 & 0.05 & 0.05 & 0.1 \\ 0.02 & 0.03 & 0.1 & 0.7 & 0.1 & 0.05 \\ 0.1 & 0.05 & 0.13 & 0.02 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.13 & 0.12 & 0.1 & 0.45 \end{pmatrix}.$$

We first compare the averaged BOEM algorithm to the online EM (OEM) procedure of [Cappé, 2011a] combined with a Polyak-Ruppert averaging (see [Polyak et Juditsky, 1992]). Note that the convergence of the OEM algorithm is still an open problem. In this case, we want to estimate the variance v and the states $\{x_1, \ldots, x_d\}$. All the runs are started from $v = 2$ and from the initial states $\{-1; 0; .5; 2; 3; 4\}$. The algorithm in [Cappé, 2011a] follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n\geq 0}$. It is expected that the rate of convergence in L_2 after n observations is $\gamma_n^{1/2}$ (and $n^{-1/2}$ for its averaged version) - this assertion relies on classical results for stochastic approximation. We prove in Section 6.5 that the rate of convergence of the BOEM algorithm is $n^{-b/(2(b+1))}$ (and $n^{-1/2}$ for its averaged version) when $\tau_n \propto n^b$. Therefore, we set $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. Figure 6.3 displays the empirical median and first and last quartiles for the estimation of v with both algorithms and their averaged versions as a function of the number of observations. These estimates are obtained over 100 independent Monte Carlo runs. Both the BOEM and the OEM algorithms converge to the true value of v and the averaged versions reduce the variability of the estimation. Figure 6.4 shows the similar behavior of both averaged algorithms for the estimation of x_1 in the same experiment. Some supplementary graphs on the estimation of the states can be found in Appendix A.3.

We now compare the averaged BOEM algorithm to a recursive maximum likelihood (RML) procedure (see [Le Gland et Mevel, 1997, Tadić, 2010]) combined with Polyak-Ruppert averaging (see [Polyak et Juditsky, 1992]). We want to estimate the variance v and the transition matrix m. All the runs are started from $v = 2$ and from a matrix m with each entry equal to $1/d$. The RML algorithm follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n\geq 0}$ which is chosen in the same way

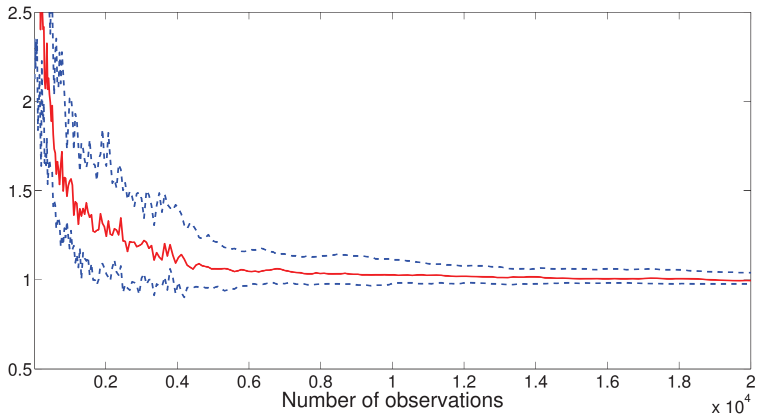


(a) The BOEM algorithm. (b) The OEM algorithm.

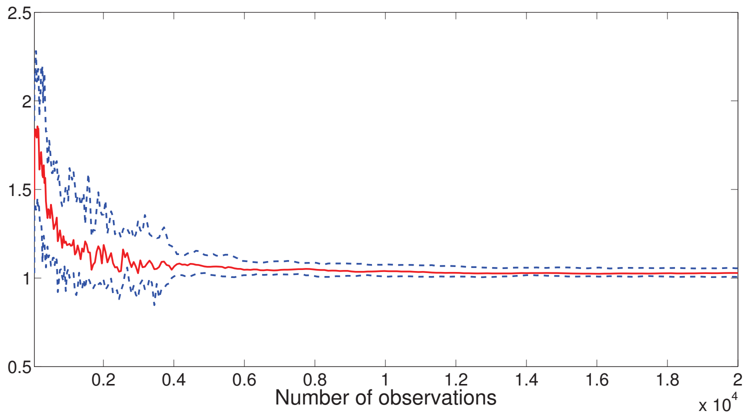
(c) The averaged BOEM algorithm. (d) The averaged OEM algorithm.

Figure 6.3: Estimation of v using the online EM and the BOEM algorithms (top) and their averaged versions (bottom). Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$.

as above. Therefore, for a fair comparison, the RML algorithm (resp. the BOEM algorithm) is run with $\gamma_n = n^{-0.53}$ (resp. $\tau_n = n^{1.1}$). Figure 6.5 displays the empirical median and empirical first and last quartiles of the estimation of $m(1,1)$ as a function of the number of observations over 100 independent Monte Carlo runs. For both algorithms, the bias and the variance of the estimation decrease as n increases. Nevertheless, the bias and/or the variance of the averaged BOEM algorithm decrease faster than those of the averaged RML algorithm (similar graphs have been obtained for the estimation of the other entries of the matrix m and for the estimation of v; see Appendix A.3). As a conclusion, it is advocated to use the averaged BOEM algorithm instead of the averaged RML algorithm.

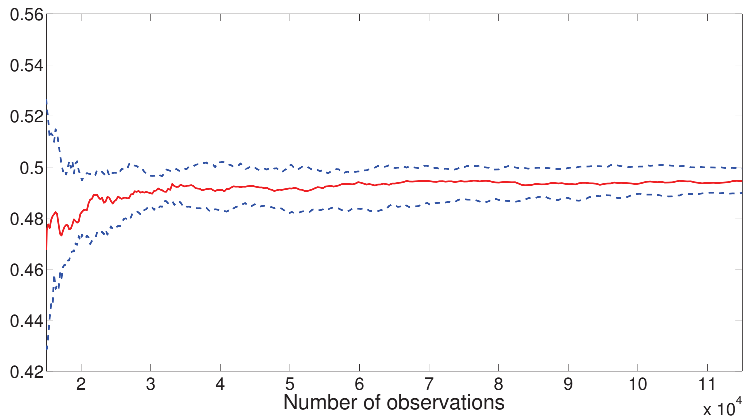


(a) The averaged BOEM algorithm.

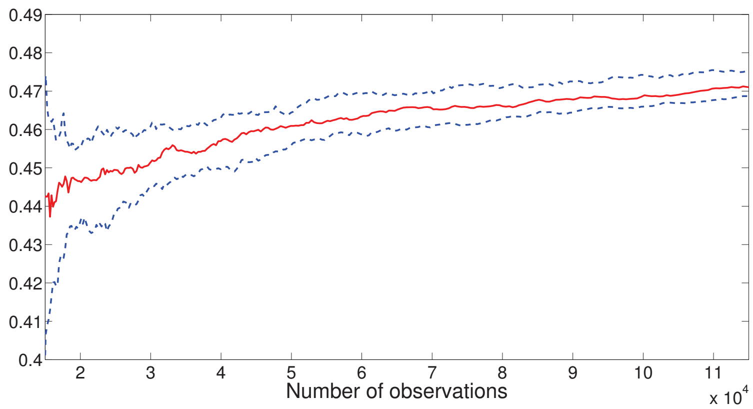


(b) The averaged OEM algorithm.

Figure 6.4: Estimation of x_1 using the averaged OEM and the averaged BOEM algorithms. Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. The first ten observations are omitted for a better visibility.



(a) The averaged BOEM algorithm.



(b) The averaged RML algorithm.

Figure 6.5: Empirical median (bold line) and first and last quartiles for the estimation of $m(1,1)$ using the averaged RML algorithm and the averaged BOEM algorithm (left). The true value is $m(1,1) = 0.5$ and the averaging procedure is starter after 10000 observations. The first 10000 observations are not displayed for a better clarity.

6.4 Convergence of the Block Online EM algorithms

Consider the following assumptions.

A1 (a) There exist continuous functions $\phi : \Theta \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}^d$ and $S : \mathbb{X} \times \mathbb{X} \times \mathbb{Y} \to \mathbb{R}^d$ s.t.

$$\log m_\theta(x, x') + \log g_\theta(x', y) = \phi(\theta) + \left\langle S(x, x,', y), \psi(\theta) \right\rangle ,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product on $\mathbb{R}^d$.

(b) There exists an open subset $\mathcal{S}$ of $\mathbb{R}^d$ that contains the convex hull of $S(\mathbb{X} \times \mathbb{X} \times \mathbb{Y})$.

(c) There exists a continuous function $\bar{\theta} : \mathcal{S} \to \Theta$ s.t. for any $s \in \mathcal{S}$,

$$\bar{\theta}(s) = \mathrm{argmax}_{\theta \in \Theta} \; \{\phi(\theta) + \langle s, \psi(\theta) \rangle\} \; .$$

A2 There exist σ_- and σ_+ s.t. for any $(x, x') \in \mathbb{X}^2$ and any $\theta \in \Theta$, $0 < \sigma_- \le m_\theta(x, x') \le \sigma_+$. Set $\rho \stackrel{\text{def}}{=} 1 - (\sigma_-/\sigma_+)$.

A2, often referred to as the strong mixing condition, is commonly used to prove the forgetting property of the initial condition of the filter, see e.g. [Del Moral et Guionnet, 1998, Del Moral *et al.*, 2003]. This assumption holds for example if $\mathbb{X}$ is finite or for linear state-spaces with truncated gaussian state and measurement noises. More generally, this condition holds when $\mathbb{X}$ is compact.

We now introduce assumptions on the observation process $\mathbf{Y}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let

$$\mathcal{F}_k^{\mathbf{Y}} \stackrel{\text{def}}{=} \sigma\left(\{Y_u\}_{u \le k}\right) \quad \text{and} \quad \mathcal{G}_k^{\mathbf{Y}} \stackrel{\text{def}}{=} \sigma\left(\{Y_u\}_{u \ge k}\right) \tag{6.9}$$

be σ-fields associated to $\mathbf{Y}$. We also define the β-mixing coefficients by, see [Davidson, 1994],

$$\beta^{\mathbf{Y}}(n) = \sup_{u \in \mathbb{Z}} \sup_{B \in \mathcal{G}_{u+n}^{\mathbf{Y}}} \mathbb{E}\left[|\mathbb{P}(B|\mathcal{F}_u^{\mathbf{Y}}) - \mathbb{P}(B)| \right] \;, \forall\, n \ge 0 \; . \tag{6.10}$$

A3-(p) $\mathbb{E}\left[\sup_{x, x' \in \mathbb{X}^2} |S(x, x', Y_0)|^p\right] < +\infty$.

A4 (a) $\mathbf{Y}$ is a β-mixing stationary sequence such that there exist $C \in [0, 1)$ and $\beta \in (0, 1)$ satisfying, for any $n \ge 0$, $\beta^{\mathbf{Y}}(n) \le C\beta^n$, where $\beta^{\mathbf{Y}}$ is defined in (6.10).

(b) $\mathbb{E}\left[|\log b_-(Y_0)| + |\log b_+(Y_0)|\right] < +\infty$ where

$$\begin{aligned} b_-(y) &\stackrel{\text{def}}{=} \inf_{\theta \in \Theta} \textstyle\int g_\theta(x, y) \lambda(\mathrm{d}x) \; , \\ b_+(y) &\stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \textstyle\int g_\theta(x, y) \lambda(\mathrm{d}x) \; . \end{aligned}$$

Upon noting that, for all $n \geq 0$, $\beta^{\mathbf{Y}}(n) \leq \beta^{(\mathbf{X},\mathbf{Y})}(n)$, we can prove that A4(a) holds when $\mathbf{Y}$ is the observation process of a an HMM under classical geometric ergodicity conditions [Meyn et Tweedie, 1993, Chapter 15] and [Cappé *et al.*, 2005, Chapter 14].

A5 There exists $c > 0$ and $a > 1$ such that for all $n \geq 1$, $\tau_n = \lfloor cn^a \rfloor$.

For $p > 0$ and Z a random variable measurable with respect to the σ-algebra $\sigma(Y_n, n \in \mathbb{Z})$, set $\|Z\|_p \stackrel{\text{def}}{=} (\mathbb{E}[|Z|^p])^{1/p}$.

A6 -(p) There exists $b \geq (a+1)/2a$ (where a is defined in A5) such that, for any $n \geq 0$,

$$\left\| S_n - \widetilde{S}_n \right\|_p = O(\tau_{n+1}^{-b}) ,$$

where $\widetilde{S}_n$ is the Monte Carlo approximation of S_n defined by (6.5).

A6 gives a L_p control of the Monte Carlo error on each block. Such bounds are given for Sequential Monte Carlo algorithms in [Dubarry et Le Corff, 2011, Theorem 1]. [Le Corff et Fort, 2011a] provides practical conditions to ensure A6 in the case of Sequential Monte Carlo methods. In the sequel, $\mathcal{M}(\mathbb{X})$ denotes the set of all probability distributions on $(\mathbb{X}, \mathcal{X})$.

Theorem 6.1. *Let $\bar{p} > 2$. Assume that A1-2, A3-($\bar{p}$) and A4 hold.*

i) For any $\theta \in \Theta$, there exists a r.v. $\mathsf{S}(\theta, \mathbf{Y})$ s.t.

$$\sup_{\theta \in \Theta,\, \chi \in \mathcal{M}(\mathbb{X})} \left| \mathbb{E}_\theta^\chi \left[S(X_{-1}, X_0, Y_0) | Y_{-\tau:\tau} \right] - \mathsf{S}(\theta, \mathbf{Y}) \right| \\ \leq C \rho^\tau \sup_{(x,x') \in \mathbb{X}^2} \left| S(x, x', Y_0) \right| , \quad \mathbb{P} - \text{a.s.} , \tag{6.11}$$

where C is a finite constant. Define for all $\theta \in \Theta$,

$$\bar{\mathrm{S}}(\theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathsf{S}(\theta, \mathbf{Y}) \right] . \tag{6.12}$$

ii) $\theta \mapsto \bar{\mathrm{S}}(\theta)$ is continuous on Θ and for any $T > 0$,

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \underset{\tau \to +\infty}{\longrightarrow} \bar{\mathrm{S}}(\theta) , \quad \mathbb{P} - \text{a.s.} , \tag{6.13}$$

where $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ is defined by (6.3).

iii) Assume in addition that A6-($\bar{p}$) holds. For any $p \in (2, \bar{p})$, there exists a constant C s.t. for any $n \geq 1$,

$$\left\| \widetilde{S}_n - \bar{\mathrm{S}}(\theta_n) \right\|_p \leq \frac{C}{\sqrt{\tau_{n+1}}} ,$$

where $\widetilde{S}_n$ is the Monte Carlo approximation of S_n defined by (6.5).

Theorem 6.1 allows to introduce the limiting EM algorithm, defined as the deterministic iterative algorithm $\check{\theta}_n = \mathrm{R}(\check{\theta}_{n-1})$ where

$$\mathrm{R}(\theta) \stackrel{\mathrm{def}}{=} \bar{\theta}\left(\bar{\mathrm{S}}(\theta)\right) . \tag{6.14}$$

The limiting EM can be seen as an EM algorithm applied as if the whole trajectory $\mathbf{Y}$ was observed instead of $Y_{0:T}$. For this limiting EM, the so-called sufficient statistics depend on the observations only through the mean $\mathbb{E}\left[\mathsf{S}(\theta, \mathbf{Y})\right]$. The stationary points of the limiting EM are defined as

$$\mathcal{L} \stackrel{\mathrm{def}}{=} \{\theta \in \Theta;\ \mathrm{R}(\theta) = \theta\} . \tag{6.15}$$

We show that there exists a Lyapunov function W w.r.t. to the map R and the set $\mathcal{L}$ *i.e.*, a continuous function W satisfying the two conditions:

(i) for all $\theta \in \Theta$,
$$\mathrm{W} \circ \mathrm{R}(\theta) - \mathrm{W}(\theta) \geq 0 , \tag{6.16}$$

(ii) for all compact set $\mathcal{K} \subset \Theta \setminus \mathcal{L}$,
$$\inf_{\theta \in \mathcal{K}} \{\mathrm{W} \circ \mathrm{R}(\theta) - \mathrm{W}(\theta)\} > 0 . \tag{6.17}$$

Recall that, for such a function, the sequence $\{\mathrm{W}(\check{\theta}_k)\}_{k \geq 0}$ is nondecreasing and $\{\check{\theta}_k\}_{k\geq 0}$ converges to $\mathcal{L}$.

Define, for any $m \geq 0$, $\theta \in \Theta$ and probability distribution χ on $(\mathbb{X}, \mathcal{X})$,

$$p_\theta^\chi\left(Y_1 | Y_{-m:0}\right) \stackrel{\mathrm{def}}{=} \frac{\int \chi(\mathrm{d}x_{-m}) g_\theta(x_{-m}, Y_m) \prod_{i=-m+1}^{1} \left\{m_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i)\right\} \lambda(\mathrm{d}x_{-m+1:1})}{\int \chi(\mathrm{d}x_{-m}) g_\theta(x_{-m}, Y_m) \prod_{i=-m+1}^{0} \left\{m_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i)\right\} \lambda(\mathrm{d}x_{-m+1:0})} .$$

By [Douc *et al.*, 2004b, Lemma 2 and Proposition 1], under A1-4, for any $\theta \in \Theta$, there exists a random variable $\log p_\theta\left(Y_1 | Y_{-\infty:0}\right)$, such that for any probability distribution χ on $(\mathbb{X}, \mathcal{X})$, $\log p_\theta\left(Y_1 | Y_{-\infty:0}\right)$ is the a.s. limit of $\log p_\theta^\chi\left(Y_1 | Y_{-m:0}\right)$ as $m \to +\infty$ and

$$T^{-1} \ell_{\theta,T}^\chi(\mathbf{Y}) \underset{T \to +\infty}{\longrightarrow} \ell(\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}\left[\log p_\theta\left(Y_1 | Y_{-\infty:0}\right)\right] , \ \mathbb{P} - \text{a.s.} , \tag{6.18}$$

where $\ell_{\theta,T}^\chi(\mathbf{Y})$ is the log-likelihood defined by (6.2). The function $\theta \mapsto \ell(\theta)$ may be interpreted as the limiting log-likelihood. We consider the function W, given, for all $\theta \in \Theta$, by

$$\mathrm{W}(\theta) \stackrel{\mathrm{def}}{=} \exp\left\{\ell(\theta)\right\} . \tag{6.19}$$

To identify the stationary points of the limiting EM algorithm as the stationary points of ℓ, we introduce an additional assumption.

A7 (a) For any $y \in \mathbb{Y}$ and for all $(x, x') \in \mathbb{X}^2$, $\theta \mapsto g_\theta(x, y)$ and $\theta \mapsto m_\theta(x, x')$ are continuously differentiable on Θ.

(b) $\mathbb{E}\left[\phi(\mathbf{Y}_0)\right] < +\infty$ where

$$\phi(y) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \sup_{(x,x') \in \mathbb{X}^2} \left| \nabla_\theta \log m_\theta(x, x') + \nabla_\theta \log g_\theta(x', y) \right| .$$

Proposition 6.1. *Assume that A1-2, A3-*(1) *and A4 hold. Then, the function* W *given by* (6.19) *is a Lyapunov function for* $(\mathrm{R}, \mathcal{L})$. *Assume in addition that A7 holds. Then,* $\theta \mapsto \ell(\theta)$ *is continuously differentiable and*

$$\mathcal{L} = \{\theta \in \Theta;\ \mathrm{R}(\theta) = \theta\} = \{\theta \in \Theta;\ \nabla \ell(\theta) = 0\} .$$

Proposition 6.1 is proved in Section 6.6.2.

Remark. In the case where $\{Y_k\}_{k \geq 0}$ is the observation process of the stationary HMM $\{(X_k, Y_k)\}_{k \geq 0}$ parameterized by $\theta_\star \in \Theta$, we can build a two-sided stationary extension of this process to obtain a sequence of observations $\{Y_k\}_{k \in \mathbb{Z}}$. Following [Douc *et al.*, 2004b, Proposition 3], the quantity $\ell(\theta)$ can be written as

$$\begin{aligned} \ell(\theta) &= \mathbb{E}_{\theta_\star}\left[\lim_{m \to +\infty} \log p_\theta(Y_1 | Y_{-m:0})\right] \\ &= \lim_{m \to +\infty} \mathbb{E}_{\theta_\star}\left[\log p_\theta(Y_1 | Y_{-m:0})\right] \\ &= \lim_{m \to +\infty} \mathbb{E}_{\theta_\star}\left[\mathbb{E}_{\theta_\star}\left[\log p_\theta(Y_1 | Y_{-m:0}) | Y_{-m:0}\right]\right] , \end{aligned}$$

where $p_\theta(Y_1 | Y_{-m:0})$ is the conditional distribution under the stationary distribution. Since

$$\mathbb{E}_{\theta_\star}\left[\log p_{\theta_\star}(Y_1 | Y_{-m:0}) | Y_{-m:0}\right] - \mathbb{E}_{\theta_\star}\left[\log p_\theta(Y_1 | Y_{-m:0}) | Y_{-m:0}\right]$$

is the Kullback-Leibler divergence between $p_{\theta_\star}(Y_1 | Y_{-m:0})$ and $p_\theta(Y_1 | Y_{-m:0})$, for any $\theta \in \Theta$, $\ell(\theta_\star) - \ell(\theta) \geq 0$ and $\theta_\star$ is a maximizer of $\theta \mapsto \ell(\theta)$. If in addition $\theta_\star$ lies in the interior of Θ, then $\theta_\star \in \mathcal{L}$.

The following proposition gives sufficient conditions for the convergence of the limiting EM algorithm and the Monte Carlo BOEM algorithm to the set $\mathcal{L}$.

Theorem 6.2. *Let* $\bar{p} > 2$. *Assume that A1-2, A3-*($\bar{p}$) *and A4 hold. Assume that* $\mathrm{W}(\mathcal{L})$ *has an empty interior. For any initial value* $\check{\theta}_0 \in \Theta$, *there exists* $w_\star$ *s.t.* $\{\check{\theta}_k\}_{k \geq 0}$ *converges to* $\{\theta \in \mathcal{L};\ \mathrm{W}(\theta) = w_\star\}$. *If in addition A5 and A6-*($\bar{p}$) *hold, then the sequence* $\{\theta_n\}_{n \geq 0}$ *converges* $\mathbb{P} - \text{a.s.}$ *to the same stationary points.*

Theorem 6.2 is a direct application of Proposition 6.4 for the limiting EM algorithm. The proof for the Monte Carlo BOEM algorithm is detailed in Section 6.6.3.

By Sard's theorem if W is at least d_θ (where $\Theta \subset \mathbb{R}^{d_\theta}$) continuously differentiable, then $\mathrm{W}(\mathcal{L})$ has Lebesgue measure 0 and hence has an empty interior.

6.5 Rate of convergence of the Block Online EM algorithms

We address the rate of convergence of the Monte Carlo BOEM algorithms to a point $\theta_\star \in \mathcal{L}$. It is assumed that

A8 (a) $\bar{\mathrm{S}}$ and $\bar{\theta}$ are twice continuously differentiable on Θ and $\mathcal{S}$.

(b) There exists $0 < \gamma < 1$ s.t. the spectral radius of $\nabla_s(\bar{\mathrm{S}} \circ \bar{\theta})_{s=\bar{\mathrm{S}}(\theta_\star)}$ is lower than γ.

Hereafter, for any sequence of random variables $\{Z_n\}_{n\geq 0}$, write $Z_n = O_{\mathrm{L}_p}(1)$ if $\sup_n \mathbb{E}\left[|Z_n|^p\right] < \infty$ and $Z_n = O_{\mathrm{a.s}}(1)$ if $\sup_n |Z_n| < +\infty$ $\mathbb{P}-$a.s.

Theorem 6.3. *Let $\bar{p} > 2$. Assume that A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 hold. Then, for any $p \in (2, \bar{p})$,*

$$\sqrt{\tau_n}\ \left[\theta_n - \theta_\star\right] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{1}{\sqrt{\tau_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1)\ . \tag{6.20}$$

By a direct application of (6.20),

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\left\{\sqrt{\tau_n}\ \|\theta_n - \theta_\star\| \mathbf{1}_{\lim_n \theta_n = \theta_\star} \geq M\right\} = 0\ .$$

The rate of convergence of the Monte Carlo BOEM algorithm is closely related to the choice of the number of observations per block. In (6.20), the rate is a function of the number of updates (i.e. the number of iteration of the algorithm). Theorem 6.4 shows that the averaging procedure reduces the influence of the block-size schedule: the rate of convergence is proportional to $T_n^{1/2}$ i.e. to the inverse of the square root of the total number of observations up to iteration n.

Theorem 6.4. *Let $\bar{p} > 2$. Assume that A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 hold. Then, for any $p \in (2, \bar{p})$,*

$$\sqrt{T_n}\left[\widetilde{\theta}_n - \theta_\star\right] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{n}{\sqrt{T_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1)\ . \tag{6.21}$$

In this case, (6.21) yields

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\left\{\sqrt{T_n}\ \|\widetilde{\theta}_n - \theta_\star\| \mathbf{1}_{\lim_n \theta_n = \theta_\star} \geq M\right\} = 0\ .$$

Theorems 6.3 and 6.4 give the rates of convergence as a function of the number of updates but they can also be studied as a function of the number of observations. Let $\{\theta_k^{\text{int}}\}_{k\geq 0}$ (resp. $\{\widetilde{\theta}_k^{\text{int}}\}_{k\geq 0}$) be such that, for any $k \geq 0$, θ_k^{int} (resp. $\widetilde{\theta}_k^{\text{int}}$) is the value θ_n (resp. $\widetilde{\theta}_n$), where n is the only integer such that $k \in [T_n + 1, T_{n+1}]$. The sequences $\{\theta_k^{\text{int}}\}_{k\geq 0}$ and $\{\widetilde{\theta}_k^{\text{int}}\}_{k\geq 0}$ are piecewise constant and their values are updated at times $\{T_n\}_{n\geq 1}$.

By Theorem 6.3, the rate of convergence of $\{\theta_k^{\text{int}}\}_{k\geq 0}$ is given (up to a multiplicative constant) by $k^{-a/(2(a+1))}$, where a is given by A5. This rates is slower than $k^{-1/2}$ and depends on the block-size sequence (through a). On the contrary, by Theorem 6.4, the rate of convergence of $\{\widetilde{\theta}_k^{\text{int}}\}_{k\geq 0}$ is given (up to a multiplicative constant) by $k^{-1/2}$, for any value of a. Therefore, this rate of convergence does not depend on the block-size sequence.

6.6 Proofs

Define, for any initial density χ on $(\mathbb{X}, \mathcal{X})$, any $\theta \in \Theta$, any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and any $r < u \leq s \leq t$,

$$\Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\int \chi(x_r)\{\prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, y_{i+1})\}\, h(x_{s-1}, x_s, y_s)\, \lambda(\mathrm{d}x_{r:t})}{\int \chi(x_r)\{\prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, y_{i+1})\}\, \lambda(\mathrm{d}x_{r:t})}, \tag{6.22}$$

for any bounded function h on $\mathbb{X}^2 \times \mathbb{Y}$. Then, the intermediate quantity of the Block online EM algorithm is (see (6.3)),

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \Phi_{\theta,t,T+\tau}^{\chi,T}(S, \mathbf{Y}) \,. \tag{6.23}$$

Lemma 6.1. *Assume A1-2. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ s.t. $\sup_{x,x'} |S(x, x', y_i)| < +\infty$ for any $i \in \mathbb{Z}$. Then for any $r > 0$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$, $\theta \mapsto \Phi_{\theta,0,r}^{\chi,-r}(S, \mathbf{y})$ is continuous on Θ.*

Proof. Set $K_\theta(x, x', y) \stackrel{\text{def}}{=} m_\theta(x, x') g_\theta(x', y)$. Let $r > 0$ and χ be a distribution on $(\mathbb{X}, \mathcal{X})$. By definition of $\Phi_{\theta,0,r}^{\chi,-r}(S, \mathbf{y})$ (see (6.22)) we have to prove that

$$\theta \mapsto \int \chi(\mathrm{d}x_{-r}) \left(\prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, y_{i+1}) \right) h(x_{-1}, x_0, y_0)\, \mathrm{d}\lambda(x_{-r+1:r})$$

is continuous for $h(x, x', y) = 1$ and $h(x, x', y) = S(x, x', y)$. By A1(a), the function $\theta \mapsto \prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, y_{i+1})\, h(x_{-1}, x_0, y_0)$ is continuous. In addition, under A1, for any $\theta \in \Theta$,

$$\left|\prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, y_{i+1})\, h(x_{-1}, x_0, y_0)\right| \\ = |h(x_{-1}, x_0, y_0)| \exp\left(2r\phi(\theta) + \left\langle \psi(\theta), \sum_{i=-r}^{r-1} S(x_i, x_{i+1}, y_{i+1}) \right\rangle\right) .$$

Since Θ is compact, by A1, there exist constants C_1 and C_2 s.t. the supremum in $\theta \in \Theta$ of this expression is bounded above by

$$C_1 \sup_{x,x'} |h(x, x', y_0)| \exp\left(C_2 \sum_{i=-r}^{r-1} \sup_{x,x'} |S(x, x', y_{i+1})|\right) .$$

Since χ is a distribution and λ is a finite measure, the continuity follows from the dominated convergence theorem. □

Let us introduce the following shorthand $S_s(x, x') \stackrel{\text{def}}{=} S(x, x', Y_s)$. Define the shift operator ϑ onto $\mathbb{Y}^{\mathbb{Z}}$ by $(\vartheta \mathbf{y})_k = \mathbf{y}_{k+1}$ for any $k \in \mathbb{Z}$; and by induction, define the s-iterated shift operator $\vartheta^{s+1}\mathbf{y} = \vartheta(\vartheta^s \mathbf{y})$, with the convention that ϑ^0 is the identity operator. For a function h, define $\mathrm{osc}(h) \stackrel{\text{def}}{=} \sup_{z,z'} |h(z) - h(z')|$.

6.6.1 Proof of Theorem 6.1

The proof of Theorem 6.1 relies on auxiliary results about the forgetting properties of HMM. Most of them are close to published results and their proof is provided in the Appendix A. The main novelty is the forgetting property of the bivariate smoothing distribution.

Proof of i) Note that under A3-(1), $\mathbb{E}[\mathrm{osc}(S_0)] < +\infty$. Under A2, Proposition 6.5(ii) implies that for any $\theta \in \Theta$, there exists a r.v. $\Phi_\theta(S, \mathbf{Y})$ s.t. for any $r < s \le T$,

$$\sup_{\theta \in \Theta} \left|\Phi^{\chi,r}_{\theta,s,T}(S, \mathbf{Y}) - \Phi_\theta(S, \vartheta^s \mathbf{Y})\right| \le \left(\rho^{T-s} + \rho^{s-r-1}\right) \mathrm{osc}(S_s) . \qquad (6.24)$$

This concludes the proof of (6.11).

Proof of ii) We introduce the following decomposition: for all $T > 0$,

$$\bar{S}^{\chi,T}_\tau(\theta, \mathbf{Y}) = \frac{1}{\tau}\sum_{t=1}^{\tau}\left[\Phi_\theta\left(S, \vartheta^{t+T}\mathbf{Y}\right) + \left\{\Phi^{\chi,0}_{\theta,t,\tau}\left(S, \vartheta^T \mathbf{Y}\right) - \Phi_\theta\left(S, \vartheta^{t+T}\mathbf{Y}\right)\right\}\right] ,$$

upon noting that by (6.23), $\bar{S}^{\chi,T}_\tau(\theta, \mathbf{Y}) = \tau^{-1}\sum_{t=1}^{\tau} \Phi^{\chi,0}_{\theta,t,\tau}\left(S, \vartheta^T \mathbf{Y}\right)$. By (6.22), (6.24) and A3-(1) $\mathbb{E}[|\Phi_\theta(S, \mathbf{Y})|] < +\infty$. Under A4, the ergodic theorem (see e.g. [Billingsley, 1995, Theorem 24.1, p.314]) states that, for any fixed T,

$$\lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=1}^{\tau}\Phi_\theta\left(S,\vartheta^{t+T}\mathbf{Y}\right)=\mathbb{E}\left[\Phi_\theta(S,\mathbf{Y})\right]\;,\quad \mathbb{P}-\text{a.s.}$$

By (6.24),

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\left|\Phi^{\chi,0}_{\theta,t,\tau}\left(S,\vartheta^{T}\mathbf{Y}\right)-\Phi_\theta\left(S,\vartheta^{t+T}\mathbf{Y}\right)\right|\le\frac{1}{\tau}\sum_{t=1}^{\tau}\left(\rho^{\tau-t}+\rho^{t-1}\right)\operatorname{osc}(S_{t+T})\;. \tag{6.25}$$

Set $Z_t\stackrel{\text{def}}{=}\frac{1}{t}\sum_{s=1}^{t}\operatorname{osc}(S_{s+T})$ and $Z_0\stackrel{\text{def}}{=}0$. Then, by an Abel transform,

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\rho^{t-1}\operatorname{osc}(S_{t+T})=\rho^{\tau-1}Z_\tau+\frac{1-\rho}{\tau}\sum_{t=1}^{\tau-1}t\rho^{t-1}Z_t\;. \tag{6.26}$$

By A3-(1) and A4, the ergodic theorem implies that $\lim_{\tau\to\infty}Z_\tau=\mathbb{E}\left[\operatorname{osc}(S_0)\right]$, $\mathbb{P}-$ a.s. Therefore, $\limsup_\tau Z_\tau<\infty$, $\mathbb{P}-$ a.s. Since $\sum_{t\ge1}t\rho^{t-1}<\infty$, this implies that $\tau^{-1}\sum_{t=1}^{\tau}\rho^{t-1}\operatorname{osc}(S_{t+T})\underset{\tau\to+\infty}{\longrightarrow}0$, $\mathbb{P}-$ a.s. Similarly,

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\rho^{\tau-t}\operatorname{osc}(S_{t+T})=Z_\tau-(1-\rho)\sum_{t=1}^{\tau-1}\rho^{\tau-t-1}Z_t+\frac{1-\rho}{\tau}\sum_{t=1}^{\tau-1}t\rho^{t-1}Z_{\tau-t}\;.$$

Using the same arguments as for the second term in (6.26), we can state that $\lim_{\tau\to\infty}\tau^{-1}\sum_{t=1}^{\tau-1}t\rho^{t-1}Z_{\tau-t}=0$, $\mathbb{P}-$ a.s. Furthermore,

$$\left|\sum_{t=1}^{\tau-1}\frac{\rho^{\tau-t-1}}{1-\rho}Z_t-\mathbb{E}\left[\operatorname{osc}(S_0)\right]\right|\le\sum_{t=1}^{\tau-1}\frac{\rho^{\tau-t-1}}{1-\rho}\left|Z_t-\mathbb{E}\left[\operatorname{osc}(S_0)\right]\right|\\ +\;\mathbb{E}\left[\operatorname{osc}(S_0)\right]\rho^{\tau-1}\;.$$

Since, $\mathbb{P}-$ a.s., $Z_\tau\underset{\tau\to+\infty}{\longrightarrow}\mathbb{E}\left[\operatorname{osc}(S_0)\right]$, the RHS converges $\mathbb{P}-$ a.s. to 0 and

$$\lim_{\tau\to+\infty}\left|Z_\tau-(1-\rho)\sum_{t=1}^{\tau-1}\rho^{\tau-t-1}Z_t\right|=0\;,\quad\mathbb{P}-\text{a.s.}$$

Hence, the RHS in (6.25) converges $\mathbb{P}-$a.s. to 0 and this concludes the proof of (6.13). We now prove that the function $\theta\mapsto\mathbb{E}\left[\Phi_\theta\left(S,\mathbf{Y}\right)\right]$ is continuous by application of the dominated convergence theorem. By Proposition 6.5(ii), for any $\mathbf{y}$ s.t. $\operatorname{osc}(S_0)<\infty$,

$$\lim_{r\to+\infty}\sup_{\theta\in\Theta}\left|\Phi^{\chi,-r}_{\theta,0,r}(S,\mathbf{y})-\Phi_\theta\left(S,\mathbf{y}\right)\right|=0\;.$$

Then, by Lemma 6.1, $\theta \mapsto \Phi_\theta(S, \mathbf{y})$ is continuous for any $\mathbf{y}$ such that $\mathrm{osc}(S_0) < +\infty$. In addition, $\sup_{\theta\in\Theta} |\Phi_\theta(S, \mathbf{Y})| \leq \sup_{x,x'} |S(x, x', Y_0)|$. We then conclude by A3-(1).

Proof of iii) Let m_n, v_n be positive integers s.t. $1 \leq m_n \leq \tau_{n+1}$ and $\tau_{n+1} = 2v_n m_n + r_n$, where $0 \leq r_n < 2m_n$. Set $\Delta p \stackrel{\text{def}}{=} p^{-1} - \bar{p}^{-1}$. By the Minkowski inequality combined with Lemmas 6.5, 6.6 applied with $q_n \stackrel{\text{def}}{=} 2v_n m_n$, there exists a constant C s.t.

$$\left\|S_n - \bar{\mathrm{S}}(\theta_n)\right\|_p \leq C\left[\rho^{m_n} + \frac{m_n}{\tau_{n+1}} + \beta^{m_n\Delta p} + \frac{1}{\sqrt{\tau_{n+1}}}\right] .$$

The proof is concluded by choosing $m_n = \lfloor -\log \tau_{n+1}/\left(\log\rho \vee \Delta p \log\beta\right)\rfloor$ and by A6-($\bar{p}$) (since b in A6-($\bar{p}$) is such that $b \geq 1/2$).

6.6.2 Proof of Proposition 6.1

Continuity of R *and* W

By A1(c) and Theorem 6.1, the function R is continuous. Under A1-2 and A4, there exists a continuous function ℓ on Θ s.t. $\lim_T T^{-1}\ell_\chi\theta, T(\mathbf{Y}) = \ell(\theta)$ $\mathbb{P}-$a.s. for any distribution χ on $(\mathbb{X}, \mathcal{X})$ and any $\theta \in \Theta$, (see [Douc *et al.*, 2004b, Lemma 2 and Propositions 1 and 2], see also Theorem A.1). Therefore, W is continuous.

Proof of the Lyapunov property (6.16)

Under Assumption A1(a)

$$\frac{1}{T}\log p_\theta(x_{0:T}, Y_{1:T}) = \phi(\theta) + \left\langle \left\{\frac{1}{T}\sum_{t=1}^{T} S(x_{t-1}, x_t, Y_t)\right\}, \psi(\theta)\right\rangle ,$$

where $p_\theta(x_{0:T}, Y_{1:T})$ is defined by (6.1). Upon noting that

$$\int S(x_{t-1}, x_t, Y_t)\frac{p_\theta(x_{0:T}, Y_{1:T})}{\int p_\theta(z_{0:T}, Y_{1:T})\lambda(\mathrm{d}z_{1:T})\chi(\mathrm{d}z_0)}\lambda(\mathrm{d}x_{1:T})\chi(\mathrm{d}x_0) \\ = \Phi^{\chi,0}_{\theta,t,T}(S, \mathbf{Y}) ,$$

the Jensen inequality gives, $\mathbb{P}-$a.s.,

$$\frac{1}{T}\ell^\chi_{\mathrm{R}(\theta),T}(\mathbf{Y}) - \frac{1}{T}\ell^\chi_{\theta,T}(\mathbf{Y}) \geq \phi(\mathrm{R}(\theta)) + \left\langle \frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S, \mathbf{Y}), \psi(\mathrm{R}(\theta))\right\rangle \\ - \phi(\theta) - \left\langle \frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S, \mathbf{Y}), \psi(\theta)\right\rangle . \quad (6.27)$$

Under A1-4, it holds by Theorem 6.1 and [Douc *et al.*, 2004b, Lemma 2 and Proposition 1] (see also Theorem A.1) that for all $\theta \in \Theta$, $\mathbb{P} -$ a.s.,

$$\frac{1}{T}\sum_{t=1}^{T} \Phi^{\chi,0}_{\theta,t,T}(S,\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \bar{\mathrm{S}}(\theta) , \qquad \frac{1}{T}\ell^{\chi}_{\theta,T}(\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \ln \mathrm{W}(\theta) .$$

Therefore, when $T \to +\infty$, (6.27) implies

$$\ln\left(\mathrm{W}(\mathrm{R}(\theta))/\mathrm{W}(\theta)\right) \geq \phi(\mathrm{R}(\theta)) + \left\langle \bar{\mathrm{S}}(\theta), \psi(\mathrm{R}(\theta))\right\rangle - \phi(\theta) - \left\langle \bar{\mathrm{S}}(\theta), \psi(\theta)\right\rangle . \tag{6.28}$$

By definition of $\bar{\theta}$ and R (see A1(c) and (6.14)), the RHS is non negative. This concludes the proof of Proposition 6.1(6.16).

Proof of the Lyapunov property (6.17)

We prove that $\mathrm{W}\circ\mathrm{R}(\theta) - \mathrm{W}(\theta) = 0$ if and only if $\theta \in \mathcal{L}$. Since $\mathrm{W}\circ\mathrm{R} - \mathrm{W}$ is continuous, this implies that $\inf_{\theta\in\mathcal{K}} \mathrm{W}\circ\mathrm{R}(\theta) - \mathrm{W}(\theta) > 0$ for all compact set $\mathcal{K} \subset \Theta \setminus \mathcal{L}$. Let $\theta \in \Theta$ be s.t. $\mathrm{W}\circ\mathrm{R}(\theta) - \mathrm{W}(\theta) = 0$. Then, the RHS in (6.28) is equal to zero. By definition of $\bar{\theta}$, $\mathrm{R}(\theta) = \theta$ and thus $\theta \in \mathcal{L}$. The converse implication is immediate from the definition of $\mathcal{L}$.

Stationary points If in addition A7 holds, Theorem A.2 proves that, for any initial distribution χ on $(\mathbb{X}, \mathcal{X})$,

$$\frac{1}{T}\nabla_\theta \ell^{\chi}_{\theta,T}(\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \nabla_\theta \ell(\theta) \quad \mathbb{P} - \text{a.s.}$$

Therefore,

$$\frac{1}{T}\nabla_\theta \ell^{\chi}_{\theta,T}(\mathbf{Y}) = \nabla_\theta \phi(\theta) + \nabla_\theta \psi'(\theta) \left\{ \frac{1}{T}\sum_{t=1}^{T} \Phi^{\chi,0}_{\theta,t,T}(S,\mathbf{Y}) \right\} ,$$

where A' is the transpose matrix of A. Theorem 6.1 yield,

$$\nabla_\theta \ell(\theta) = \nabla_\theta \phi(\theta) + \nabla_\theta \psi'(\theta)\bar{S}(\theta) .$$

The proof follows upon noting that by definition of $\bar{\theta}$, the unique solution to the equation $\nabla_\theta \phi(\tau) + \nabla_\theta \psi'(\tau)\bar{S}(\theta) = 0$ is $\tau = \mathrm{R}(\theta)$.

6.6.3 Proof of Theorem 6.2

The proof of Theorem 6.2 relies on Proposition 6.4 applied with $T(\theta) \stackrel{\text{def}}{=} \mathrm{R}(\theta)$ and with $\theta_{n+1} = \bar{\theta}(\widetilde{S}_n)$. The key ingredient for this proof is the control of the L_p-mean error between the Monte Carlo Block Online EM algorithm and the *limiting EM*. The proof of this bound is derived in Theorem 6.1 and relies on preliminary lemmas given in Appendix 6.7. The proof of (6.38) is now close to the proof of [Fort et Moulines, 2003, Proposition 11] and is postponed to the Appendix A.

6.6.4 Proof of Theorem 6.3

Define $s_\star \stackrel{\text{def}}{=} \bar{\mathrm{S}}(\theta_\star)$ and write

$$\bar{\theta}(\widetilde{S}_n) - \bar{\theta}(s_\star) = \Upsilon(\widetilde{S}_n - s_\star) + \bar{\theta}(\widetilde{S}_n) - \bar{\theta}(s_\star) - \Upsilon(\widetilde{S}_n - s_\star) \;, \tag{6.29}$$

where $\Upsilon \stackrel{\text{def}}{=} \nabla\bar{\theta}(s_\star)$. We now derive the rate of convergence of the quantity $\widetilde{S}_n - s_\star$. Set $\mathrm{G}(s) \stackrel{\text{def}}{=} \bar{\mathrm{S}} \circ \bar{\theta}(s)$. Note that under A8(b), the spectral radius of Γ is lower than γ, where $\Gamma \stackrel{\text{def}}{=} \nabla \mathrm{G}(s_\star)$. Since $\mathrm{G}(s_\star) = s_\star$, we write

$$\widetilde{S}_n - s_\star = \Gamma\left(\widetilde{S}_{n-1} - s_\star\right) + \widetilde{S}_n - \mathrm{G}(\widetilde{S}_{n-1}) + \mathrm{G}(\widetilde{S}_{n-1}) - \mathrm{G}(s_\star) - \Gamma\left(\widetilde{S}_{n-1} - s_\star\right) \;.$$

Define $\{\mu_n\}_{n\geq 0}$ and $\{\rho_n\}_{n\geq 0}$ s.t. $\mu_0 = 0$, $\rho_0 = \widetilde{S}_0 - s_\star$ and

$$\mu_n \stackrel{\text{def}}{=} \Gamma\mu_{n-1} + e_n \;, \qquad \rho_n \stackrel{\text{def}}{=} \widetilde{S}_n - s_\star - \mu_n \;, \qquad n \geq 1 \;, \tag{6.30}$$

where,

$$e_n \stackrel{\text{def}}{=} \widetilde{S}_n - \bar{\mathrm{S}}(\theta_n) \;, \qquad n \geq 1 \;. \tag{6.31}$$

Proposition 6.2. *Assume A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 for some $\bar{p} > 2$. Then for any $p \in (2, \bar{p})$,*

$$\sqrt{\tau_n}\mu_n = O_{\mathrm{L}_p}(1) \quad \textit{and} \quad \tau_n \rho_n \mathbf{1}_{\lim_n S_n = s_\star} = O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) \;.$$

The proof of Proposition 6.2 follows the same lines as the proof of [Fort et Moulines, 2003, Theorem 6]. The main ingredient is the control of $\|\mu_n\|_p$ which is a consequence of [Pólya et Szegő, 1976, Result 178, p. 39] and Theorem 6.1. The detailed proof is thus omitted and postponed to the Appendix A.

By Proposition 6.2, the first term in (6.29) gives

$$\sqrt{\tau_n}\Upsilon(S_n - s_\star)\mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{1}{\sqrt{\tau_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) \;.$$

A Taylor expansion with integral remainder term gives the rate of convergence of the second term. This concludes the proof of Theorem 6.3, Eq. (6.20).

6.6.5 Proof of Theorem 6.4

In the sequel, for all function Ξ on $\Theta \times \mathbb{Y}^{\mathbb{Z}}$ and all $\upsilon \in \Theta$, we denote by $\mathbb{E}\left[\Xi(\theta, \mathbf{Y})\right]_{\theta=\upsilon}$ the function $\theta \mapsto \mathbb{E}\left[\Xi(\theta, \mathbf{Y})\right]$ evaluated at $\theta = \upsilon$. We preface the proof by the following lemma.

Lemma 6.2. *Assume A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 for some $\bar{p} > 2$. For any $p \in (2, \bar{p})$,*

$$\limsup_{n\to+\infty} \frac{1}{\sqrt{T_{n+1}}} \left\| \sum_{k=1}^{n} \tau_{k+1} e_k \right\|_p < \infty \,,$$

where e_n is given by (6.31).

Proof. By A5 and A6-($\bar{p}$), we have

$$\limsup_{n\to+\infty} \frac{1}{\sqrt{T_{n+1}}} \sum_{k=1}^{n} \tau_{k+1} \left\| \widetilde{S}_k - S_k \right\|_p < \infty \,.$$

Then, it is sufficient to prove that

$$\limsup_{n\to+\infty} \frac{1}{\sqrt{T_{n+1}}} \left\| \sum_{k=1}^{n} \tau_{k+1} \left(\bar{S}(\theta_k) - S_k \right) \right\|_p < \infty \,.$$

Let $p \in (2, \bar{p})$. In the sequel, C is a constant independent on n and whose value may change upon each appearance. Let $1 \le m_n \le \tau_{n+1}$ and set $v_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$. By Lemma 6.6 applied with $q_k \stackrel{\text{def}}{=} 2v_k m_k$, we have,

$$\left\| \sum_{k=1}^{n} \tau_{k+1} \left(\bar{S}(\theta_k) - S_k \right) \right\|_p \le C \left(\sum_{k=1}^{n} \{ \tau_{k+1} \rho^{m_k} + m_k \} + \left\| \sum_{k=1}^{n} \{ \delta_k + \zeta_k \} \right\|_p \right) \,,$$

where δ_k and ζ_k are defined by

$$\delta_k \stackrel{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \right\} \,,$$

$$\zeta_k \stackrel{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] - \mathbb{E}\left[\Phi^{\chi,-m_k}_{\theta,0,m_k}(S, \mathbf{Y}) \right]_{\theta=\theta_k} \right\}$$

and where $F_{t,k}(\theta_k, \mathbf{Y}) \stackrel{\text{def}}{=} \Phi^{\chi,t-m_k}_{\theta_k,t,t+m_k}(S, \vartheta^{T_k}\mathbf{Y})$ and $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k}$ is given by (6.42). We will prove below that there exists C s.t.

$$\|\zeta_k\|_p \le C\, \beta^{m_k/pb} \tau_{k+1} \,, \qquad \forall k \ge 1 \tag{6.32}$$

$$\left\| \sum_{k=1}^{n} \delta_k \right\|_p \le C\sqrt{T_{n+1}} + C \sum_{k=1}^{n} \tau_{k+1} \beta^{m_k/pb} \,, \qquad \forall n \ge 1 \tag{6.33}$$

so that the proof is concluded by choosing $m_k = \lfloor \eta \log \tau_{k+1} \rfloor$, $\eta \stackrel{\mathrm{def}}{=} (-1/\log \rho) \vee (-pb/\log \beta)$ and by using A5.

We turn to the proof of (6.32). By the Berbee Lemma (see [Rio, 1990, Chapter 5]) and A4, there exist $C \in [0,1)$ and $\beta \in (0,1)$ s.t. for all $k \geq 1$, there exists a random variable $Y^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ independent from $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k}$ with the same distribution as $Y_{T_k+m_k:T_{k+1}+m_k}$ and

$$\mathbb{P}\left\{ Y^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k} \neq Y_{T_k+m_k:T_{k+1}+m_k} \right\} \leq C\beta^{m_k} \; . \tag{6.34}$$

Upon noting that $\mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}^{\star,(k)}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] = \mathbb{E}\left[F_{t,k}(\theta, \mathbf{Y}) \right]_{\theta=\theta_k}$, we have

$$\zeta_k = \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] - \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}^{\star,(k)}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \right\} \; . \tag{6.35}$$

Therefore, by setting $\mathcal{A}_k \stackrel{\mathrm{def}}{=} \{ Y^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k} \neq Y_{T_k+m_k:T_{k+1}+m_k} \}$,

$$|\zeta_k| \leq \sum_{t=2m_k}^{2v_k m_k} \mathbb{E}\left[\sup_{\theta \in \Theta} \left| F_{t,k}(\theta, \mathbf{Y}) - F_{t,k}(\theta, \mathbf{Y}^{\star,(k)}) \right| \mathbf{1}_{\mathcal{A}_k} \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \; .$$

Minkowski and Holder (with $a \stackrel{\mathrm{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\mathrm{def}}{=} 1 - a^{-1}$) inequalities, combined with (6.34), A4, Lemma 6.3 and A3-($\bar{p}$) yield (6.32).

We now prove (6.33). Upon noting that δ_k is $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_{k+1}}$-measurable and δ_k is a martingale increment, the Rosenthal inequality (see [Hall et Heyde, 1980, Theorem 2.12, p.23]) states that $\|\sum_{k=1}^n \delta_k\|_p \leq C \left(\sum_{k=1}^n I_k^{(1)} \right)^{1/p} + C I_n^{(2)}$ where

$$I_k^{(1)} \stackrel{\mathrm{def}}{=} \mathbb{E}\left[|\delta_k|^p \right] \quad \text{and} \quad I_n^{(2)} \stackrel{\mathrm{def}}{=} \left\| \left(\sum_{k=1}^n \mathbb{E}\left[|\delta_k|^2 \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \right)^{1/2} \right\|_p \; .$$

Using again $\mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}^{\star,(k)}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] = \mathbb{E}\left[F_{t,k}(\theta, \mathbf{Y}) \right]_{\theta=\theta_k}$ and (6.35)

$$I_k^{(1)} \leq C \left\| \sum_{t=2m_k}^{2v_k m_k} \left\{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}\left[F_{t,k}(\theta, \mathbf{Y}) \right]_{\theta=\theta_k} \right\} \right\|_p^p + C \; \|\zeta_k\|_p^p \; .$$

By Lemma 6.5 and (6.32), there exists C s.t. for any $k \geq 1$

$$I_k^{(1)} \leq C \left(\tau_{k+1}^{p/2} + \tau_{k+1}^p \beta^{m_k/b} \right) \; , \tag{6.36}$$

and since $2/p < 1$, convex inequalities yield $\left(\sum_{k=1}^n I_k^{(1)}\right)^{1/p} \leq C\sqrt{T_{n+1}} + C\sum_{k=1}^n \tau_{k+1}\beta^{m_k/pb}$. By the Minkowski and Jensen inequalities, it holds $I_n^{(2)} \leq \left(\sum_{k=1}^n \{I_k^{(1)}\}^{2/p}\right)^{1/2}$. Hence, by (6.36),

$$I_n^{(2)} \leq C\sqrt{T_{n+1}} + C\sum_{k=1}^n \tau_{k+1}\beta^{m_k/pb} \; .$$

This concludes the proof of (6.33). □

We write $\Sigma_n - s_\star = \bar{\mu}_n + \bar{\rho}_n$ with

$$\bar{\mu}_n \stackrel{\text{def}}{=} \frac{1}{T_n}\sum_{k=1}^n \tau_k \mu_{k-1} \quad \text{and} \quad \bar{\rho}_n \stackrel{\text{def}}{=} \frac{1}{T_n}\sum_{k=1}^n \tau_k \rho_{k-1} \; . \tag{6.37}$$

Proposition 6.3. *Assume A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 for some $\bar{p} > 2$. For any $p \in (2, \bar{p})$,*

$$\sqrt{T_n}\bar{\mu}_n = O_{\mathrm{L}_p}(1) \; , \qquad \frac{T_n}{n}\,\bar{\rho}_n \mathbf{1}_{\lim_n S_n = s_\star} = O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) \; .$$

Proof. Set $A \stackrel{\text{def}}{=} (I - \Gamma)$. Under A8, A^{-1} exists. By (6.30) and (6.37),

$$A\sqrt{T_n}\bar{\mu}_n = -\frac{\tau_{n+1}\mu_n}{\sqrt{T_n}} + \frac{1}{\sqrt{T_n}}\sum_{k=1}^n \tau_{k+1} e_k + \frac{1}{\sqrt{T_n}}\sum_{k=1}^n \tau_k \left(\frac{\tau_{k+1}}{\tau_k} - 1\right)\Gamma\mu_{k-1} \; .$$

The result now follows from Proposition 6.2, Lemma 6.2 and A5. The proof of the second assertion follows from (6.37) and Proposition 6.2. □

Upon noting that $\theta_\star = \bar{\theta}(s_\star)$, we may write, for the averaged sequence,

$$\widetilde{\theta}_n - \theta_\star = \Upsilon(\Sigma_n - s_\star) + \bar{\theta}(\Sigma_n) - \bar{\theta}(s_\star) - \Upsilon(\Sigma_n - s_\star) \; .$$

The first term in this decomposition gives

$$\sqrt{T_n}\Upsilon(\Sigma_n - s_\star)\mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{n}{\sqrt{T_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) \; .$$

By A8(b), as for the non averaged sequence, a Taylor expansion with integral remainder term gives the result for the second term. This concludes the proof of Theorem 6.4, Eq.(6.21).

6.7 Technical results

Proposition 6.4 is exactly [Fort et Moulines, 2003, Proposition 9] applied with a compact set Θ.

Proposition 6.4. *Let $T : \Theta \to \Theta$ and W be a continuous Lyapunov function relatively to T and to $\mathcal{L} \subset \Theta$. Assume $\mathrm{W}(\mathcal{L})$ has an empty interior and that $\{\theta_n\}_{n\geq 0}$ is a sequence lying in Θ such that*

$$\lim_{n\to+\infty} |\mathrm{W}(\theta_{n+1}) - \mathrm{W} \circ T(\theta_n)| = 0 \; . \tag{6.38}$$

Then, there exists $w_\star$ such that $\{\theta_n\}_{n\geq 0}$ converges to $\{\theta \in \mathcal{L};\ \mathrm{W}(\theta) = w_\star\}$.

The proof of Proposition 6.5 is given in Section A.2.

Proposition 6.5. *Assume A2. Let χ, $\widetilde{\chi}$ be two distributions on $(\mathbb{X}, \mathcal{X})$. For any measurable function $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\sup_{x,x'} |h(x, x', y_s)| < +\infty$ for any $s \in \mathbb{Z}$*

(i) For any $r < s \leq t$ and any $\ell_1, \ell_2 \geq 1$,

$$\sup_{\theta\in\Theta} \left| \Phi^{\widetilde{\chi},r}_{\theta,s,t}\left(h, \mathbf{y}\right) - \Phi^{\chi,r-\ell_1}_{\theta,s,t+\ell_2}\left(h, \mathbf{y}\right) \right| \leq \left(\rho^{s-1-r} + \rho^{t-s}\right) \mathrm{osc}(h_s) \; . \tag{6.39}$$

(ii) For any $\theta \in \Theta$, there exists a function $\mathbf{y} \mapsto \Phi_\theta(h, \mathbf{y})$ s.t. for any distribution χ on $(\mathbb{X}, \mathcal{X})$ and any $r < s \leq t$

$$\sup_{\theta\in\Theta} \left| \Phi^{\chi,r}_{\theta,s,t}\left(h, \mathbf{y}\right) - \Phi_\theta\left(h, \vartheta^s \mathbf{y}\right) \right| \leq \left(\rho^{s-1-r} + \rho^{t-s}\right) \mathrm{osc}(h_s) \; . \tag{6.40}$$

Remark. (a) If $\chi = \widetilde{\chi}$, $\ell_1 = 0$ and $\ell_2 \geq 1$, (6.39) becomes

$$\sup_{\theta\in\Theta} \left| \Phi^{\chi,r}_{\theta,s,t}\left(h, \mathbf{y}\right) - \Phi^{\chi,r}_{\theta,s,t+\ell_2}\left(h, \mathbf{y}\right) \right| \leq \rho^{t-s} \mathrm{osc}(h_s) \; .$$

(b) if $\ell_2 = 0$ and $\ell_1 \geq 1$, (6.39) becomes

$$\sup_{\theta\in\Theta} \left| \Phi^{\widetilde{\chi},r}_{\theta,s,t}\left(h, \mathbf{y}\right) - \Phi^{\chi,r-\ell_1}_{\theta,s,t}\left(h, \mathbf{y}\right) \right| \leq \rho^{s-1-r} \mathrm{osc}(h_s) \; .$$

Lemma 6.3 is a consequence of (6.22) and of Proposition 6.5(ii).

Lemma 6.3. *Assume A2. Let $r < s \leq t$ be integers, $\theta \in \Theta$ and $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, and $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ s.t. for any $s \in \mathbb{Z}$, $\sup_{x,x'} |h(x, x', y_s)| < \infty$. Then*

$$\left| \Phi^{\chi,r}_{\theta,s,t}\left(h, \mathbf{y}\right) \right| \leq \sup_{(x,x')\in\mathbb{X}^2} \left| h(x, x', y_s) \right| \; , \; \left| \Phi_\theta\left(h, \vartheta^s \mathbf{y}\right) \right| \leq \sup_{(x,x')\in\mathbb{X}^2} \left| h(x, x', y_s) \right| \; .$$

For any $L \geq 1$, $m \geq 1$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$, define

$$\kappa^{\chi}_{L,m}(\boldsymbol{\theta}, \mathbf{Y}) \stackrel{\mathrm{def}}{=} \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon = \boldsymbol{\theta}} . \tag{6.41}$$

We introduce the σ-algebra $\widetilde{\mathcal{F}}_{T_n}$ defined by

$$\widetilde{\mathcal{F}}_{T_n} \stackrel{\mathrm{def}}{=} \sigma\{\mathcal{F}^{\mathbf{Y}}_{T_n}, \mathcal{H}_{T_n}\} , \tag{6.42}$$

where $\mathcal{F}_{T_n}$ is given by (6.9) and where $\mathcal{H}_{T_n}$ is independent from $\mathbf{Y}$ (the σ-algebra $\mathcal{H}_{T_n}$ is generated by the random variables independent from the observations $\mathbf{Y}$ used to produce the Monte Carlo approximation of $\{S_{k-1}\}_{k=1}^{n}$). Hence, for any positive integer m and any $B \in \mathcal{G}^{\mathbf{Y}}_{T_n+m}$, since $\mathcal{H}_{T_n}$ is independent from B and from $\mathcal{F}^{\mathbf{Y}}_{T_n}$, $\mathbb{P}(B|\widetilde{\mathcal{F}}_{T_n}) = \mathbb{P}(B|\mathcal{F}^{\mathbf{Y}}_{T_n})$. Therefore, the mixing coefficients defined in (6.10) are such that

$$\beta(\mathcal{G}^{\mathbf{Y}}_{T_n+m}, \widetilde{\mathcal{F}}_{T_n}) = \beta(\mathcal{G}^{\mathbf{Y}}_{T_n+m}, \mathcal{F}_{T^{\mathbf{Y}}_n}) .$$

Note that θ_n is $\widetilde{\mathcal{F}}_{T_n}$- measurable and that $\widetilde{S}_n$ is $\widetilde{\mathcal{F}}_{T_{n+1}}$-measurable.

Lemma 6.4. *Assume A2, A3-($\bar{p}$) and A4 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any distribution χ on $(\mathbb{X}, \mathcal{X})$, any $m \geq 1$, $k, \ell \geq 0$ and any Θ-valued $\widetilde{\mathcal{F}}^{\mathbf{Y}}_0$-measurable r.v. $\boldsymbol{\theta}$,*

$$\left\| \sum_{u=1}^{k} \kappa^{\chi}_{2um+\ell,m}(\boldsymbol{\theta}, \mathbf{Y}) \right\|_p \leq C \left[\sqrt{\frac{k}{m}} + k\beta^{m\,\Delta p} \right] ,$$

where $\Delta p \stackrel{\mathrm{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$ and β is given by A4.

Proof. For ease of notation χ is dropped from the notation $\kappa^{\chi}_{2um,m}$. By the Berbee Lemma (see [Rio, 1990, Chapter 5]), for any $m \geq 1$, there exists a Θ-valued r.v. $\boldsymbol{\upsilon}^{\star}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ independent from $\mathcal{G}^{\mathbf{Y}}_m$ (see (6.9)) s.t.

$$\mathbb{P}\left\{\boldsymbol{\theta} \neq \boldsymbol{\upsilon}^{\star}\right\} = \sup_{B \in \mathcal{G}^{\mathbf{Y}}_m} |\mathbb{P}(B|\sigma(\boldsymbol{\theta})) - \mathbb{P}(B)| . \tag{6.43}$$

Set $L_u \stackrel{\mathrm{def}}{=} 2um + \ell$. We write

$$\sum_{u=1}^{k} \kappa_{L_u,m}(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{u=1}^{k} \left\{ \Phi^{\chi, L_u - m}_{\boldsymbol{\theta}, L_u, L_u+m}(S, \mathbf{Y}) - \Phi^{\chi, L_u - m}_{\boldsymbol{\upsilon}^{\star}, L_u, L_u+m}(S, \mathbf{Y}) \right\}$$
$$+ \sum_{u=1}^{k} \kappa_{L_u,m}(\boldsymbol{\upsilon}^{\star}, \mathbf{Y}) + k \left\{ \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon = \boldsymbol{\upsilon}^{\star}} - \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon = \boldsymbol{\theta}} \right\} . \tag{6.44}$$

By the Holder's inequality with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

$$\left\| \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \Phi^{\chi, L-m}_{\boldsymbol{\upsilon}^\star, L, L+m}(S, \mathbf{Y}) \right\|_p \leq \left\| \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \vartheta^T\mathbf{Y}) - \Phi^{\chi, L-m}_{\boldsymbol{\upsilon}^\star, L, L+m}(S, \mathbf{Y}) \right\|_{\bar{p}} \mathbb{P}\left\{\boldsymbol{\theta} \neq \boldsymbol{\upsilon}^\star\right\}^{\Delta p} .$$

By A3-($\bar{p}$), A4, (6.10) and (6.43), there exists a constant C_1 s.t. for any $m, L \geq 1$, any distribution χ and any Θ-valued $\widetilde{\mathcal{F}}^{\mathbf{Y}}_0$-measurable r.v. $\boldsymbol{\theta}$,

$$\left\| \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \Phi^{\chi, L-m}_{\boldsymbol{\upsilon}^\star, L, L+m}(S, \mathbf{Y}) \right\|_{\bar{p}} \leq C_1 \beta^{m\Delta p} .$$

Similarly, there exists a constant C_2 s.t. for any $m \geq 1$, any distribution χ and any Θ-valued $\widetilde{\mathcal{F}}^{\mathbf{Y}}_0$-measurable r.v. $\boldsymbol{\theta}$,

$$\left\| \mathbb{E}\left[\Phi^{\chi, -m}_{\upsilon, 0, m}(S, \mathbf{Y})\right]_{\upsilon = \boldsymbol{\upsilon}^\star} - \mathbb{E}\left[\Phi^{\chi, -m}_{\upsilon, 0, m}(S, \mathbf{Y})\right]_{\upsilon = \boldsymbol{\theta}} \right\|_p \leq C_2 \beta^{m\Delta p} .$$

Let us consider the second term in (6.44). For any $u \geq 1$ and any $\upsilon \in \Theta$, the r.v. $\kappa_{L_u, m}(\upsilon, \mathbf{Y})$ is a measurable function of $\mathbf{Y}_i$ for all $L_u - m + 1 \leq i \leq L_u + m$. Since $L_u \geq 2um$, for any $\upsilon \in \Theta$, $\sum_{u=1}^k \kappa_{L_u, m}(\upsilon, \mathbf{Y})$ is $\mathcal{G}^{\mathbf{Y}}_m$-measurable. $\boldsymbol{\upsilon}^\star$ is independent from $\mathcal{G}^{\mathbf{Y}}_m$ so that:

$$\left\| \sum_{u=1}^k \kappa_{L_u, m}(\boldsymbol{\upsilon}^\star, \mathbf{Y}) \right\|_p = \mathbb{E}\left[\mathbb{E}\left[\left| \sum_{u=1}^k \kappa_{L_u, m}(\upsilon, \mathbf{Y}) \right|^p \right]_{\upsilon = \boldsymbol{\upsilon}^\star} \right]^{1/p} .$$

Define the strong mixing coefficient (see [Davidson, 1994])

$$\alpha^{\mathbf{Y}}(r) \stackrel{\text{def}}{=} \sup_{u \in \mathbb{Z}} \sup_{(A,B) \in \mathcal{F}^{\mathbf{Y}}_u \times \mathcal{G}^{\mathbf{Y}}_{u+r}} \left| \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \right| , r \geq 0 .$$

Then, [Davidson, 1994, Theorem 14.1, p.210] implies that for any $m \geq 1$, the strong mixing coefficients of the sequence $\kappa_{(\mathbf{m})} \stackrel{\text{def}}{=} \{\kappa_{L_u, m}(\upsilon, \mathbf{Y})\}_{u \geq 1}$ satisfies $\alpha^{\kappa_{(\mathbf{m})}}(i) \leq \alpha^{\mathbf{Y}}(2(i-1)m + 1)$. Furthermore, by [Rio, 1990, Theorem 2.5],

$$\left\| \sum_{u=1}^k \kappa_{L_u, m}(\upsilon, \mathbf{Y}) \right\|_p \leq (2kp)^{1/2} \left(\int_0^1 \left[N_{(m)}(t) \wedge k\right]^{p/2} \mathcal{Q}^p_{\upsilon, m}(t) \mathrm{d}t \right)^{1/p} ,$$

where $N_{(m)}(t) \stackrel{\text{def}}{=} \sum_{i \geq 1} \mathbf{1}_{\alpha^{\kappa_{(\mathbf{m})}}(i) > t}$ and $\mathcal{Q}_{\upsilon, m}$ denotes the inverse of the tail function $t \mapsto \mathbb{P}(|\kappa_{L_u, m}(\upsilon, \mathbf{Y})| \geq t)$. The sequence $\mathbf{Y}$ being stationary, this inverse function does not depend on u. By A4 and the inequality $\alpha^{\mathbf{Y}}(r) \leq \beta^{\mathbf{Y}}(r)$ (see e.g. [Davidson, 1994, Chapter 13]), there exist $\beta \in [0, 1)$ and $C \in (0, 1)$ s.t. for any $u, m \geq 1$,

$$N_{(m)}(u) \leq \sum_{i \geq 1} \mathbf{1}_{\alpha^{\mathbf{Y}}(2(i-1)m+1) > u} \leq \sum_{i \geq 1} \mathbf{1}_{C\beta^{2(i-1)m} > u} \leq \left(\frac{\log u - \log C}{2m \log \beta} \right) \vee 0 .$$

Let U be a uniform r.v. on $[0,1]$. Observe that $C\beta^{2mk} < 1$. Then, by the Holder inequality applied with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

$$\left\| \left[N_{(m)}(U) \wedge k\right]^{1/2} \mathcal{Q}_{\upsilon,m}(U) \right\|_p \stackrel{\text{def}}{=} \left(\int_0^1 \left[N_{(m)}(u) \wedge k\right]^{p/2} \mathcal{Q}^p_{\upsilon,m}(u)\mathrm{d}u \right)^{1/p}$$
$$\leq \left[\frac{-1}{2m\log\beta}\right]^{1/2} \left\| \mathcal{Q}_{\upsilon,m}(U) \left(-\log\frac{U}{C}\right)^{1/2} \mathbf{1}_{(C\beta^{Cmk},C)}(U) \right\|_p$$
$$+ k^{1/2} \left\| \mathcal{Q}_{\upsilon,m}(U)\mathbf{1}_{U\leq C\beta^{2mk}} \right\|_p ,$$
$$\leq \left\{ (C\beta^{2mk})^{\Delta p} k^{1/2} + \left[\frac{-1}{2m\log\beta}\right]^{1/2} \left\| \left(-\log\frac{U}{C}\right)^{1/2} \mathbf{1}_{(C\beta^{Cmk},C)}(U) \right\|_{pb} \right\}$$
$$\times \left\| \mathcal{Q}_{\upsilon,m}(U) \right\|_{\bar{p}} .$$

Since U is uniform on $[0,1]$, $\mathcal{Q}_{\upsilon,m}(U)$ and $|\kappa_{L_u,m}(\upsilon,\mathbf{Y})|$ have the same distribution, see [Rio, 1990]. Then, by Lemma 6.3 and A3-($\bar{p}$), there exists a constant C s.t. for any $\upsilon \in \Theta$, any $m \geq 1$,

$$\sup_{\upsilon\in\Theta} \left\| \mathcal{Q}_{\upsilon,m}(U) \right\|_{\bar{p}} \leq C \left\| \sup_{x,x'\in\mathbb{X}^2} |S(x,x',\mathbf{Y}_0)| \right\|_{\bar{p}} ,$$

which concludes the proof. □

Lemma 6.5. *Assume A2, A3-($\bar{p}$) and A4 for some $\bar{p} > 2$. Let $p \in (2,\bar{p})$. There exists a constant C s.t. for any $n \geq 1$, any $1 \leq m_n \leq \tau_{n+1}$ and any distribution χ on $(\mathbb{X},\mathcal{X})$,*

$$\left\| \frac{1}{\tau_{n+1}} \sum_{t=2m_n}^{2\upsilon_n m_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n}\mathbf{Y}) \right\|_p \leq C \left[\frac{1}{\sqrt{\tau_{n+1}}} + \beta^{m_n \Delta p} \right] ,$$

where $\kappa^{\chi}_{L,m}$ and β are defined by (6.41) and A4, $\upsilon_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$ and $\Delta p \stackrel{\text{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$.

Proof. We write,

$$\left\| \sum_{t=2m_n}^{2\upsilon_n m_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n}\mathbf{Y}) \right\|_p \leq \sum_{\ell=0}^{2m_n-1} \left\| \sum_{u=1}^{\upsilon_n-1} \kappa^{\chi}_{2um_n+\ell,m_n}(\theta_n, \vartheta^{T_n}\mathbf{Y}) \right\|_p .$$

Observe that by definition θ_n is $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_n}$-measurable. Then, by Lemma 6.4, there exists a constant C s.t. for any $m_n \geq 1$ and any $\ell \geq 0$,

$$\left\| \sum_{u=1}^{\upsilon_n-1} \kappa^{\chi}_{2um_n+\ell,m_n}(\theta_n, \vartheta^{T_n}\mathbf{Y}) \right\|_p \leq C \left[\sqrt{\frac{\upsilon_n}{m_n}} + \upsilon_n \beta^{m_n\Delta p} \right] .$$

The proof is concluded upon noting that $\tau_{n+1} \geq 2m_n\upsilon_n$. □

Lemma 6.6. *Assume A2, A3-($\bar{p}$) and A4 for some $\bar{p} > 2$. For any $p \in (2, \bar{p}]$, there exists a constant C s.t. for any $n \geq 1$, any $1 \leq m_n \leq q_n \leq \tau_{n+1}$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$,*

$$\left\| \bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) - \bar{\mathrm{S}}(\theta_n) - \widetilde{\rho}_n \right\|_p \leq C \left[\rho^{m_n} + \frac{m_n}{\tau_{n+1}} + \frac{\tau_{n+1} - q_n}{\tau_{n+1}} \right] ,$$

where $\widetilde{\rho}_n \stackrel{\text{def}}{=} \tau_{n+1}^{-1} \sum_{t=2m_n}^{q_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n}\mathbf{Y})$ and $\kappa^{\chi}_{L,m}$ is defined by (6.41).

Proof. By (6.3) and (6.22), $\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) - \bar{\mathrm{S}}(\theta_n) - \widetilde{\rho}_n = \sum_{i=1}^{4} g_{i,n}$ where

$$g_{1,n} \stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=1}^{\tau_{n+1}} \left(\Phi^{\chi,0}_{\theta_n,t,\tau_{n+1}}(S, \vartheta^{T_n}\mathbf{Y}) - \Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) \right) ,$$

$$g_{2,n} \stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=1}^{2m_n-1} \left(\Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right) ,$$

$$g_{3,n} \stackrel{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=q_n+1}^{\tau_{n+1}} \left(\Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right) ,$$

$$g_{4,n} \stackrel{\text{def}}{=} \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} - \bar{\mathrm{S}}(\theta_n) .$$

In the case $\tau_{n+1} > 2m_n$, it holds

$$\begin{aligned} \tau_{n+1} |g_{1,n}| \leq & \sum_{t=\tau_{n+1}-m_n+1}^{\tau_{n+1}} \left(\rho^{m_n-1} + \rho^{\tau_{n+1}-t} \right) \mathrm{osc}(S_{t+T_n}) \\ & + \sum_{t=1}^{m_n} \left(\rho^{m_n} + \rho^{t-1} \right) \mathrm{osc}(S_{t+T_n}) + 2\rho^{m_n-1} \sum_{t=m_n+1}^{\tau_{n+1}-m_n} \mathrm{osc}(S_{t+T_n}) , \end{aligned}$$

where we used Proposition 6.5(i) and Remark 6.7 in the last inequality. By A3-($\bar{p}$) and A4, there exists C s.t. $\|g_{1,n}\|_p \leq C\left(\rho^{m_n} + \tau_{n+1}^{-1}\right)$. The same bound hold in the case $\tau_{n+1} \leq 2m_n$. For $g_{2,n}$ and $g_{3,n}$, we use the bounds

$$\begin{aligned} & \left| \Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right| \\ & \qquad \leq \sup_{(x,x')\in\mathbb{X}^2} \left| S(x, x', Y_{T_n+t}) \right| + \mathbb{E}\left[\sup_{(x,x')\in\mathbb{X}^2} \left| S(x, x', Y_0) \right| \right] . \end{aligned}$$

Then, by A4,

$$\begin{aligned} & \left\| \Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right\|_p \\ & \qquad \leq 2 \left\| \sup_{(x,x')\in\mathbb{X}^2} \left| S(x, x', Y_0) \right| \right\|_p , \end{aligned}$$

and the RHS is finite under A3-($\bar{p}$). Finally,

$$|g_{4,n}| \leq 2\rho^{m_n - 1}\mathbb{E}\left[\mathrm{osc}(S_0)\right] ,$$

where we used Theorem 6.1. This concludes the proof. □