Online Expectation Maximization based algorithms for inference in hidden Markov models

Sylvain Le Corff, Gersende Fort

Online Expectation Maximization based algorithms for inference in Hidden Markov Models

Sylvain Le Corff*† and Gersende Fort†

August 19, 2011

Abstract

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. An averaging procedure is also proposed to improve the rate of convergence. Finally, practical illustrations are presented as well as extensions to some online stochastic EM when Sequential Monte Carlo methods have to be used in combination, in order to make the E-step tractable.

1 Introduction

The Expectation Maximization (EM) algorithm is a well-known iterative algorithm to solve maximum likelihood estimation in incomplete data models [11]. In this context, model parameter estimates are obtained by maximizing the log-likelihood of the observations $Y_{0:T}$. Despite in incomplete data models the log-likelihood is not explicit, EM algorithm is generally simple to implement since it relies on complete data computations: each iteration consists in a E-step where the expectation of the complete log-likelihood under the conditional distribution of the latent data given the observations is computed; and a M-step, which updates the parameter estimate based on this conditional expectation.

In many situations of interest, the complete data likelihood belongs to the exponential family. In this case, the E-step consists in the computation of

*This work is partially supported by the French National Research Agency, under the program ANR-08-BLAN-0218 BigMC

†LCTI, CNRS and TELECOM ParisTech, 46 rue Barrault 75634 Paris Cedex 13, France

the expectation of the complete data sufficient statistic under the conditional distribution. In such case, the EM algorithm can be considered equivalently as an iterative algorithm in the space of the complete data sufficient statistics (instead of in the parameter space).

The EM algorithm has been successfully applied for maximum likelihood inference in general state-space models. Except for simple models 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 resp. [5, 16]) depending on the complexity of the model.

When processing large data sets or data streams however, 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. The first online procedure for parameter estimation was introduced in [29] by Titterington. This algorithm relies on a stochastic gradient approach which aims at incorporating the newly available observation. In Cappé and Moulines [4], the proposed algorithm is more closely related to the original EM recursion: in the case of an exponential complete-data likelihood, the E-step is replaced by a stochastic approximation step while the M-step remains unchanged.

More complex incomplete data models such as Hidden Markov Models (HMM) are of common use to represent time series in many fields such as statistics, information engineering and financial econometrics, see [14, 31]. An online version of the EM algorithm for inference in Hidden Markov Model 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 and Denève [23] (resp. Cappé [3]). In Cappé [3], the algorithm relies on the ability to compute approximations of the filtering distribution and on an intermediate quantity based on the sufficient statistics. In order to update these computations recursively, stochastic approximation procedures are introduced. This algorithm has been extended to the case of general state-space models by substituting deterministic approximation of the smoothing probabilities by Sequential Monte Carlo algorithms (see Cappé [2], Del Moral *et al.* [8] and Le Corff *et al.* [21]).

Despite the encouraging first results when applying these online EM algorithms, the convergence of these algorithms and the characterization of the limit points (when the number of observations tends to infinity) remain an open question. The convergence of the online variants of the EM algorithm for i.i.d. observations is addressed by Cappé and Moulines [4]: the limit points are the stationary points of the Kullback-Leibler divergence between the marginal distribution of the observation and the model distribution. There do not exist convergence results for the online EM algorithms for general state-space models (some insights on the asymptotic behavior are nevertheless given in Cappé [3]): 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 exponential complete-data likelihood. It sticks more 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 $Y_{T_k+1:T_{k+1}}$. 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 $\mathbf{Y}_t$, $t \in \{T_k + 1, \ldots, T_{k+1}\}$ is available. Such recursions are provided in recent works on online estimation in HMM, see [2, equation (1)], [3, Section 2.2] and [8, Proposition 2.1].

This new algorithm, called *Block Online EM* algorithm (BOEM) is derived in Section 2 together with an *averaged* version. Section 3 is devoted to practical applications: BOEM is used to perform parameter inference in HMM where the forward recursions mentioned above are available explicitly (this occurs e.g. for finite state-space HMM and linear Gaussian models). In the case of finite state-space HMM, BOEM is compared to a gradient-type recursive maximum likelihood procedure. The new algorithm is also extended to models where the E-step is intractable and has to be approximated by Sequential Monte Carlo algorithms; in this context, it is compared to online EM-type algorithms existing in the literature when applied to a stochastic volatility model. The convergence of BOEM is addressed in Section 4. BOEM is seen as a perturbation of a deterministic *limiting EM* algorithm, the limiting behavior of which is studied through a Lyapunov-function technique. The perturbation is shown to vanish (in some sense) as the number of observations increases thus implying that BOEM inherits the asymptotic behavior of the limiting-EM. Finally, in Section 5, we prove that the rate of convergence of BOEM strongly depends upon the block size sequence: this rate is optimal when the block size increases exponentially which is, quite unfortunately, of poor practical interest. Nevertheless, we prove that the averaged BOEM reaches this optimal rate of convergence for slowly increasing block size sequence. The proofs are postponed in Section 6; supplementary materials are provided in the supplement paper [20].

2 The Block Online EM algorithms

2.1 Notations and Model assumptions

For any $r \leq t$, $x_{r:t}$ is a shorthand notation for the sequence $(x_r, \cdots, x_t)$.

Let $\mathbf{Y} = \{Y_t\}_{t\in\mathbb{Z}}$ be the observation process defined on $(\Omega, \mathbb{P}_\star, \mathcal{F})$ and taking values in $\mathbb{Y}^{\mathbb{Z}}$ where $\mathbb{Y}$ is a general space endowed with a countably generated σ-field $\mathcal{B}(\mathbb{Y})$.

A HMM model parameterized by θ, for θ in a set $\Theta \subseteq \mathbb{R}^{d_\theta}$, is fitted to the observations: consider a family of transition kernels $\{m_\theta(x, x')\mathrm{d}\lambda(x')\}_{\theta\in\Theta}$ onto $\mathbb{X} \times \mathcal{B}(\mathbb{X})$ where $\mathbb{X}$ is a general state-space equipped with a countably generated σ-field $\mathcal{B}(\mathbb{X})$, and λ is a bounded non-negative measure on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Let $\{g_\theta(x, y)\mathrm{d}\nu(y)\}_{\theta\in\Theta}$ be a family of transition kernels on $(\mathbb{X} \times \mathcal{B}(\mathbb{Y}))$, where ν is a measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$.

For any initial distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, any $\theta \in \Theta$, any $r < s \leq t$ and any sequence $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, define the probability measure $\Phi_{\theta,s,t}^{\chi,r}(\cdot, \mathbf{y})$ by

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

for any bounded function h. Note that if $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is a HMM with transition kernels m_θ and g_θ, $\Phi_{\theta,s,t}^{\chi,r}(h, Y)$ is the conditional expectation of $h(X_{s-1}, X_s, Y_s)$ given $Y_{r+1:t}$ when $X_r \sim \chi$.

It is assumed that the HMM is *exponential* i.e.

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

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

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 $S(\mathbb{X} \times \mathbb{X} \times \mathbb{Y}^{\mathbb{Z}})$.

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

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

2.2 Block Online EM (BOEM)

Define

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \Phi_{\theta,t,T+\tau}^{\chi,T}(S, \mathbf{Y}) . \quad (2)$$

Once again, note that if $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is a HMM with transition kernels m_θ and g_θ, $\bar{S}_\tau^{\chi,T}(\theta, Y)$ is the conditional expectation of the additive functional $\sum_{t=T+1}^{T+\tau} S(X_{t-1}, X_t, Y_t)$ given $Y_{T+1:T+\tau}$ when $X_T \sim \chi$. BOEM updates the parameter estimates by using such integrals computed on non-overlapping block of observations; the expectation is with respect to (w.r.t.) a conditional distribution given the (random) observations $\mathbf{Y}_T, \cdots, \mathbf{Y}_{T+\tau}$. Consequently, it is a stochastic iterative algorithm.

Let $\{\tau_n\}_{n\geq 1}$ be a sequence of positive integers and set

$$T_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} \tau_k \quad \text{and} \quad T_0 \stackrel{\text{def}}{=} 0 ; \quad (3)$$

τ_n denotes the length of the block n. To ensure the stability of this stochastic iterative algorithm, we use a reprojection scheme adapted from [6]. Let $\{\Theta_n\}_{n\geq 0}$ be a sequence of compact subsets of Θ s.t.

$$\forall n \geq 0, \ \Theta_n \subset \Theta_{n+1} \quad \text{and} \quad \Theta = \bigcup_{n\geq 0} \Theta_n .$$

Given an initial value $\theta_0 \in \Theta_0$ and starting with $p_0 = 0$, the BOEM algorithm defines a sequence $\{\theta_n\}_{n\geq 1}$ by

$$\begin{aligned} \theta_{n-1/2} &\stackrel{\text{def}}{=} \bar{\theta}\left[\bar{S}_{\tau_n}^{\chi,T_{n-1}}(\theta_{n-1},\mathbf{Y})\right], \\ \theta_n &= \begin{cases} \theta_{n-1/2} & \text{if } \theta_{n-1/2} \in \Theta_{p_n} \\ \theta_0 & \text{otherwise and set } p_n = p_{n-1}+1\ . \end{cases} \end{aligned} \tag{4}$$

p_n counts the number of truncations; it is proved in Theorem 4.4 that $\{p_n\}_{n\geq 0}$ is finite w.p.1. i.e. w.p.1., $\theta_n = \theta_{n-1/2}$ for all n large enough.

For ease of notation, it is assumed in this recursion that the initial distribution χ is the same for all blocks even though it will be clear in Section 4 that the initial distribution can change over blocks. $\lim_{\tau\to\infty} \bar{S}_\tau^{\chi,T}(\theta,\mathbf{Y})$ exists $\mathbb{P}_\star -$ a.s (see Theorem 4.1 below): it is thus expected that BOEM applied with a sequence $\{\tau_n\}_{n\geq 1}$ increasing to infinity will have the same asymptotic behavior as the iterative procedure in which $\bar{S}_{\tau_n}^{\chi,T_{n-1}}(\theta_{n-1},\mathbf{Y})$ is replaced by its limit. We will give a rigorous proof of this intuition in section 4, as well as assumptions on $\{\tau_n\}_{n\geq 1}$ in order to prove such a result.

2.3 Averaged Block Online EM

When τ_n is large, $\bar{S}_{\tau_n}^{\chi,T}(\theta,\mathbf{Y})$ may be seen as an estimator of the a.s. limit $\lim_{\tau\to\infty} \bar{S}_\tau^{\chi,T}(\theta,\mathbf{Y})$. By analogy to the regression problem, an estimator with reduced variance can be obtained by averaging and weighting the successive estimates (see [25, 26, 18] for a discussion on the averaging procedures). Define $\Sigma_0 \stackrel{\text{def}}{=} 0$ and for $n \geq 1$,

$$\Sigma_n \stackrel{\text{def}}{=} \frac{1}{T_n}\sum_{j=1}^{n} \tau_j\, \bar{S}_{\tau_j}^{\chi,T_{j-1}}(\theta_{j-1},\mathbf{Y})\ ; \tag{5}$$

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

$$\widetilde{\theta}_n \stackrel{\text{def}}{=} \bar{\theta}\left(\Sigma_n\right)\ . \tag{6}$$

3 Applications to inverse problems in Hidden Markov Models

3.1 Linear Gaussian Model

Consider the Linear Gaussian model (LGM):

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

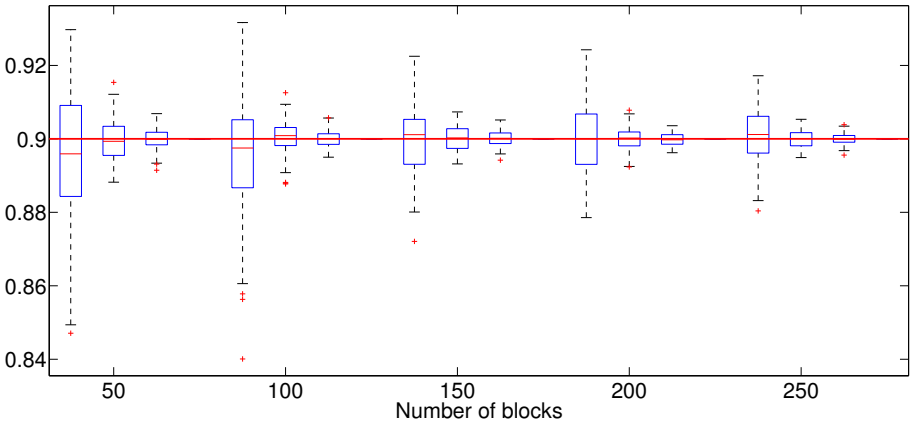
where $X_0 \sim \mathcal{N}\left(0, \sigma_u^2(1-\phi^2)^{-1}\right)$, $\{U_t\}_{t\geq 0}, \{V_t\}_{t\geq 0}$ are i.i.d. standard Gaussian r.v., independent from X_0. Data are sampled using $\phi = 0.9$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$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

We illustrate the convergence of the BOEM algorithms. We choose $\tau_n = a(n+1)$. We display in Figure 1 the box and whisker plots for the estimation of ϕ obtained with 100 independent Monte Carlo experiments; different values of a are also considered. Both the BOEM algorithm and the averaged one converge to the true value $\phi = 0.9$; and the averaging procedure clearly improves the variance of the estimation. In Figure 2, the estimates of the three parameters $(\phi, \sigma_u^2, \sigma_v^2)$ are given as a function of the number of blocks when $a = 10$, illustrating the performance of the BOEM algorithms (the first 10 iterations are not shown for a better clarity). Figures 1 and 2 show that the averaged procedure needs a few more iterations to converge but when compared to the non averaged one, the variance is much smaller.

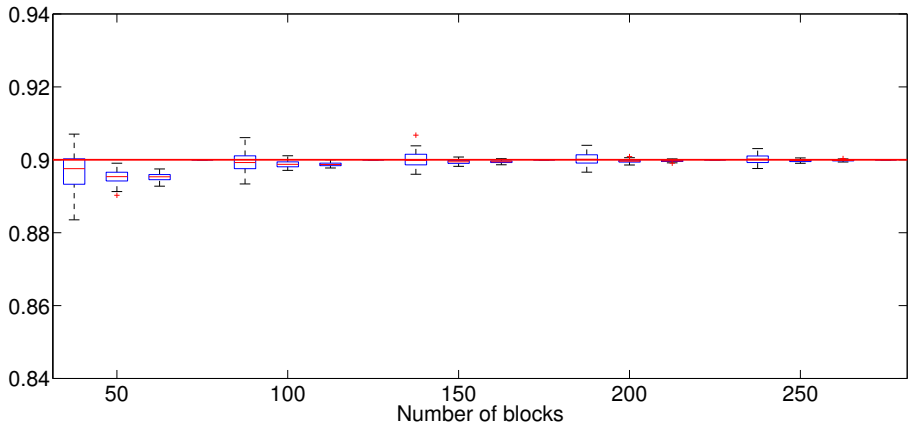
We now discuss the role of the initial distribution χ. The convergence results (see Section 4) show that our algorithms converge whatever χ. Figure 3 displays the estimation of ϕ by the averaged BOEM algorithm with $\tau_n \sim (n+99)^{1.2}$, over 100 independent Monte Carlo runs as a function of the number of blocks. We consider first the case when χ is the stationary distribution of the hidden process i.e. $\chi \equiv \mathcal{N}(0, (1-\phi^2)^{-1}\sigma_u^2)$, and the case when χ is the filtering distribution obtained at the end of the previous block, computed with the Kalman filter. In terms of the error of the estimation, the two strategies are similar. We observe the same phenomenon for different values of ϕ (see [20, section 5]). Therefore, it is advocated to choose χ as the filtering distribution obtained at the end of the previous block.

We now discuss the role of $\{\tau_n\}_{n\geq 0}$. Figure 4 displays the empirical variance, when estimating ϕ, computed with 100 independent Monte Carlo runs, for different numbers of observations and, for both the BOEM and its averaged version. We consider four polynomial rates $\tau_n \sim n^b$, $b \in \{1.2, 1.8, 2, 2.5\}$. Figure 4a 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 rate although this implies practical difficulties: when $\tau_n \sim n^2$, many observations are needed for each update of the parameter sequence. Then, the estimation process is highly dependent on the initialization of the algorithm, as illustrated by Figure 5. This phenomenon is all the more important than τ_n increases rapidly; therefore, geometrically increasing sequence τ_n is not at all advocated, at least in the first iterations of the algorithm. The influence of the block size sequence τ_n is greatly reduced with the averaging procedure as shown in Figure 4b. We will show in Section 5 that averaging really improves the rate of convergence of BOEM.

As a conclusion, it is advocated to use the averaged BOEM algorithm. In practice, one could use slowly increasing sequences τ_n for the first iterations, and then, use more rapidly increasing sequences after the burn-in period.



(a) BOEM without averaging, when $\tau_n = a(n+1)$.

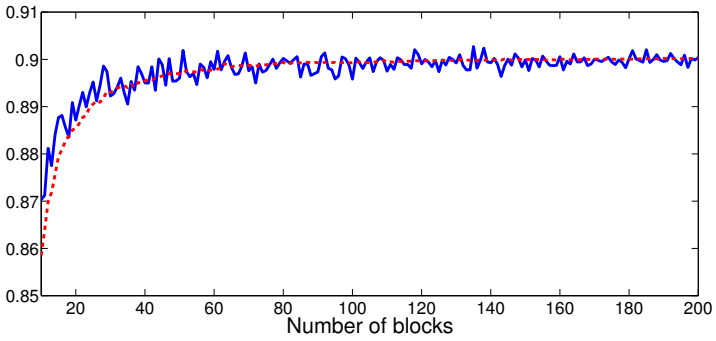


(b) BOEM with averaging, when $\tau_n = a(n+1)$.

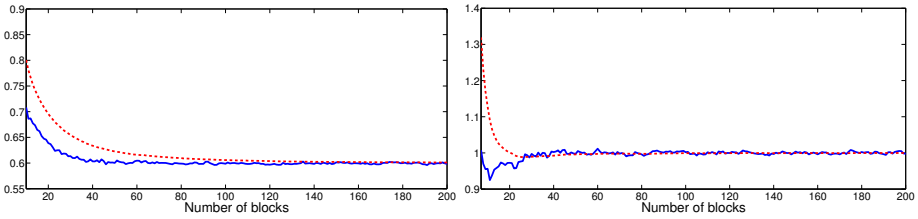
Figure 1: Estimation of ϕ for $a = 10$ (left), $a = 100$ (middle) and $a = 300$ (right) after $50, 100, 150, 200$ and 250 blocks.

3.2 Finite state-space HMM

We consider models where the unobservable states take a finite number of values. Mixture processes with Markov dependence, switching processes with Markov regime, communication channels driven by Hidden Markov processes, composite sources with switch controlled by a Markov chain are examples of finite state-space HMM found useful in many fields including biostatistics, genomics, information theory, speech processing, ... (see [15] for a review). In the numerical applications below, 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 $\{\mu(1), \ldots, \mu(d)\}$, with initial distribution ν and a $d \times d$ transition ma-



(a) Estimation of ϕ. The true value is $\phi = 0.9$.



(b) Estimation of σ_u^2 (left) and σ_v^2 (right). The true value is $(\sigma_u^2, \sigma_v^2) = (0.6, 1)$.

Figure 2: Estimation of the three parameters without averaging (bold line) and with averaging (dotted line), $a = 10$.

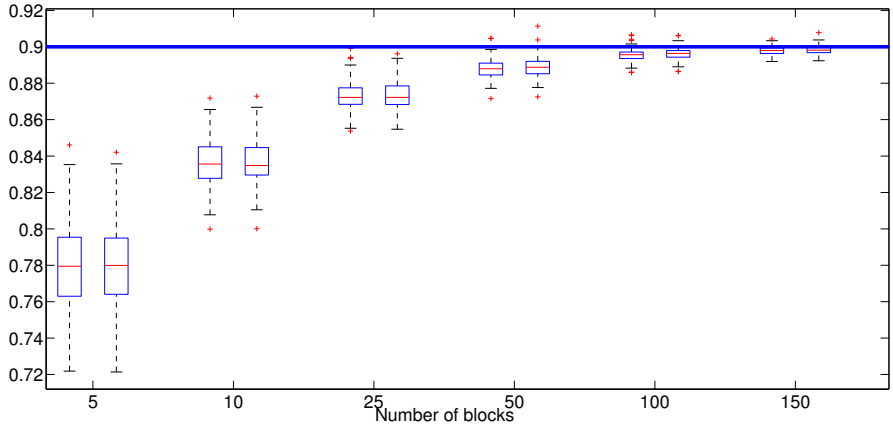
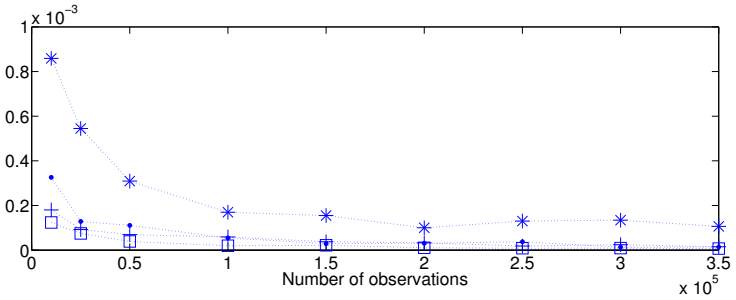
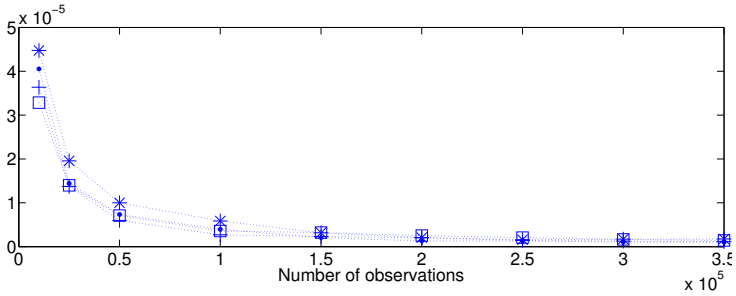


Figure 3: Estimation of ϕ after $5, 10, 25, 50, 100$ and 150 blocks, with two different initialization schemes: the stationary distribution (left) and the filtering distribution at the end of the previous block (right). The boxplots are computed with 100 Monte Carlo runs.



(a) BOEM, without averaging



(b) BOEM, with averaging

Figure 4: BOEM: Empirical variance of the estimation of ϕ after $n = 0.5\ell\, 10^5$ 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).

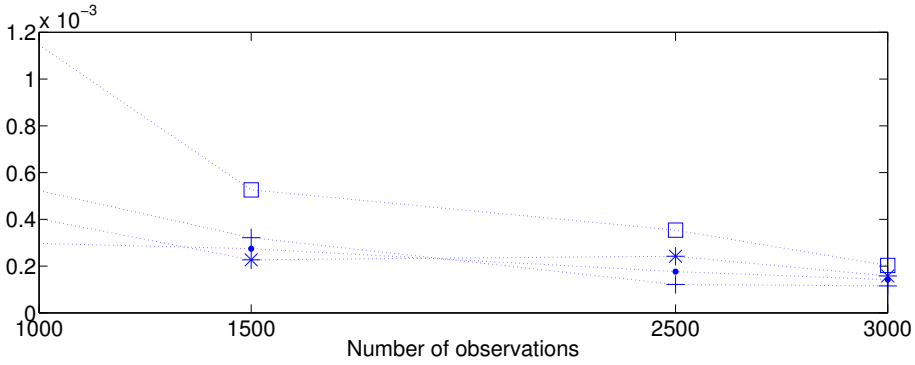


Figure 5: BOEM with averaging: Empirical variance of the estimation of ϕ after $n = 1000, 1500, 2500$ and 3000 observations for different block size schemes $\tau_n \sim n^{1.2}$ (stars), $\tau_n \sim n^{1.8}$ (dots), $\tau_n \sim n^2$ (crosses) and $\tau_n \sim n^{2.5}$ (squares).

trix m. $\{V_t\}_{t\geq 0}$ are i.i.d. $\mathcal{N}(0, v)$ r.v., independent from $\{X_t\}_{t\geq 0}$. Observations

are sampled using $d = 6$, $v = 0.5$, $\mu(i) = i\ , \forall i \in \{1, \ldots, d\}$ and

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

We 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 averaged BOEM is compared to a Polyak-Ruppert averaged (see [25]) recursive maximum likelihood (RML) procedure (see [22, 28]). RML follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n\geq 0}$. In the case of the RML, it is expected that the rate of convergence in L_2 after n observations is $\gamma_n^{1/2}$ (and $1/\sqrt{n}$ for its averaged version) - this assertion relies on classical results for stochastic approximation. We prove in Section 5 that the rate of convergence of BOEM is $n^{-b/(2(b+1))}$ (and $1/\sqrt{n}$ for its averaged version) when $\tau_n \propto n^b$. Therefore, for a fair comparison, RML (resp. BOEM) is run with $\gamma_n \propto n^{-0.6}$ (resp. $\tau_n \propto n^{3/2}$). Figure 6 displays boxplots of the estimations of v and $m(1,1)$ after different numbers of observations n; the boxplots are 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 decrease faster than those of the averaged RML (similar graphs have been obtained for the estimation of the other entries of the matrix m; some supplementary graphs can be found in [20, Section 5]). As a conclusion, it is advocated to use the averaged BOEM instead of the averaged RML.

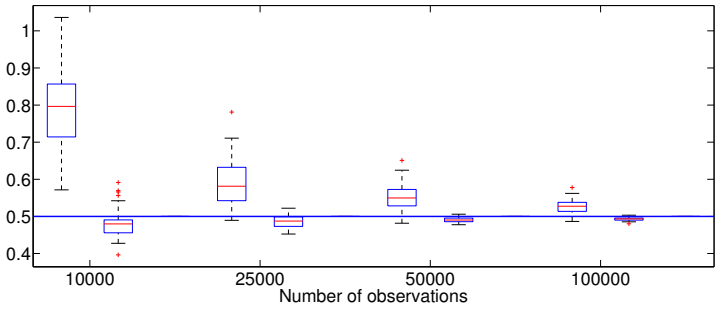
3.3 Stochastic Block Online EM algorithms

Consider the following stochastic volatility model (SVM):

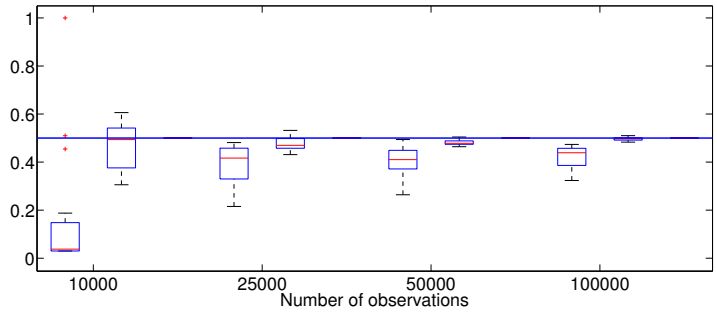
$$X_{t+1} = \phi X_t + \sigma U_t\ , \qquad\qquad Y_t = \beta \mathrm{e}^{\frac{X_t}{2}} V_t\ ,$$

where $X_0 \sim \mathcal{N}\left(0, (1-\phi^2)^{-1}\sigma^2\right)$ and $(U_t)_{t\geq 0}$ and $(V_t)_{t\geq 0}$ are two sequences of i.i.d. standard Gaussian r.v., independent from X_0. Data are sampled using $\phi = 0.8$, $\sigma^2 = 0.2$ and $\beta^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma^2 = 0.6$ and $\beta^2 = 2$.

In this model, the smoothed sufficient statistics $\{\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})\}_{n\geq 1}$ can not be computed explicitly. We thus propose to replace the exact computation by a Monte Carlo approximation based on particle filtering. The performance of the Stochastic BOEM is compared to the online EM algorithm given in [3] (see also [8]). To our best knowledge, there do not exist results on the asymptotic behavior of the algorithms by [3, 8]; these algorithms rely on many approximations that make the proof quite difficult (some insights on the asymptotic



(a) Estimation of v.



(b) Estimation of $m(1,1)$.

Figure 6: Estimation of v and $m(1,1)$ using the averaged RML algorithm (left) and the averaged BOEM algorithm (right), based on $n = \{10k, 25k, 50k, 100k\}$ observations.

behavior are given in [3]). Despite there are no results in the literature on the rate of convergence of the Online EM algorithm by [3] we choose the step size γ_n in [3] and the block size τ_n s.t. $\gamma_n = n^{-0.6}$ and $\tau_n \propto n^{3/2}$ (see section 3.2 for a discussion on this choice). 50 particles are used for the approximation of the filtering distribution by Particle filtering. We report in Figure 7, the boxplots for the estimation of the three parameters (β, ϕ, σ^2) for the Polyak-Ruppert [25] averaged Online EM and the averaged BOEM. Both average versions are started after 20000 observations. Figure 7 displays the estimation of ϕ. The estimation of σ^2 and β^2 are given in the supplement paper [20]. This figure shows that both algorithms have the same behavior. Similar conclusions are obtained by considering other true values for ϕ (such as $\phi = 0.95$); these analyzes are provided in [20, Section 5]). Therefore, the intuition is that online EM and Stochastic BOEM have the same asymptotic behavior. The main advantage of the second approach is that it relies on approximations which can be controlled in such a way that we are able to show that the limiting points of the particle version of the Stochastic BOEM algorithms are the stationary points of the limiting nor-

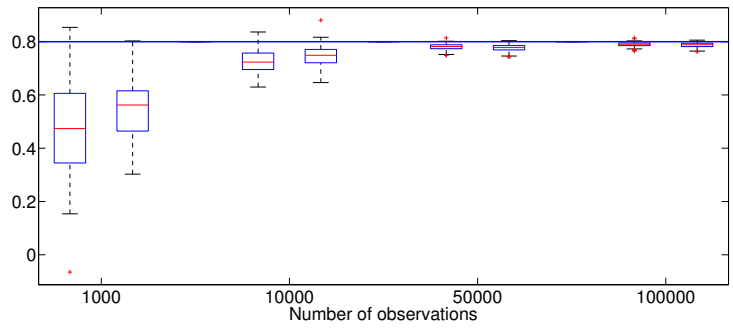


Figure 7: Estimation of ϕ using the averaged online EM algorithm (left) and the averaged BOEM algorithm (right), after $n = \{1000, 10k, 50k, 100k\}$ observations.

malized log-likelihood of the observations. The convergence of this stochastic BOEM is out of the scope of this paper and is addressed in the paper [19].

4 Convergence of the Block Online EM algorithms

In this section, it is shown that for any $T > 0$ and any initial distribution χ, the quantity $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ converges $\mathbb{P}_\star -$ a.s when $\tau \to +\infty$, to a deterministic quantity $\bar{\mathrm{S}}(\theta)$ that does not depend on T and χ (see Theorem 4.1). Therefore, the BOEM algorithm can be seen as a perturbation of the so-called *limiting EM* algorithm, defined as a deterministic iterative algorithm $\check{\theta}_n = \mathrm{R}(\check{\theta}_{n-1})$ where

$$\mathrm{R}(\theta) \stackrel{\text{def}}{=} \bar{\theta}\left(\bar{\mathrm{S}}(\theta)\right) \ . \tag{7}$$

We identify the limiting points of the limiting EM algorithm (see section 4.3) and show that BOEM inherits this limiting behavior provided the perturbation can be set small enough (see section 4.4). We start with introducing the assumptions to address such a convergence result.

4.1 Assumptions

Consider the following assumptions

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

This assumption is known in the literature as the *strong mixing condition*. It is commonly used to prove the forgetting property of the initial condition of the filter, see e.g. [9, 10]. This assumption holds for example in $\mathbb{X}$ is finite and for any $(x, x') \in \mathbb{X}^2$, $0 < \inf_\theta m_\theta(x, x') \leq \sup_\theta m_\theta(x, x') < +\infty$. Under regularity conditions on the kernels $\{m_\theta; \theta \in \Theta\}$, it also holds when $\mathbb{X}$ is compact.

Nevertheless, it fails to hold in standard situations s.t. linear and Gaussian state-space models. It has been weakened in recent works: in [12], the exponential forgetting of the initial condition of the filter is proved with a local Doeblin property; [30] gives an uniform time average convergence of some particle filters. The approach in [12] could be adapted to the present paper but at a quite technical cost. For pedagogical purposes, we think that weakening A2 is out of the scope of this paper.

We now introduce assumptions on the observation process. Define the shift operator ϑ onto $\mathbb{Y}^{\mathbb{Z}}$ by $(\vartheta \circ \mathbf{y})_k = \mathbf{y}_{k+1}$ for any $k \in \mathbb{Z}$; and by induction, define the s-iterated shift operator

$$\vartheta^{s+1} \circ \mathbf{y} = \vartheta \circ (\vartheta^s \circ \mathbf{y}) \,, \qquad \forall s \geq 0 \,, \tag{8}$$

with the convention that ϑ^0 is the identity operator. The shift operator is said to be ergodic for $\mathbb{P}_\star$ if for each set A in $\{A \in \mathcal{B}(\mathbb{Y})^{\otimes\mathbb{Z}}; A = \vartheta^{-1}(A)\}$, $\mathbb{P}_\star(A) \in \{0,1\}$ (see [1, p.314]).

A3-(γ) $\mathbb{E}_\star \left[\sup_{x,x' \in \mathbb{X}^2} |S(x,x',\mathbf{Y}_0)|^\gamma\right] < +\infty$.

A4 (a) Under $\mathbb{P}_\star$, $\mathbf{Y}$ is a stationary sequence.

(b) The shift operator is ergodic with respect to $\mathbb{P}_\star$.

(c) $\mathbb{E}_\star \left[|\log b_-(\mathbf{Y}_0)| + |\log b_+(\mathbf{Y}_0)|\right] < +\infty$ where

$$b_-(y) \stackrel{\text{def}}{=} \inf_{\theta\in\Theta} \int g_\theta(x,y)\lambda(\mathrm{d}x) \,, \qquad b_+(y) \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} \int g_\theta(x,y)\lambda(\mathrm{d}x) \,. \tag{9}$$

Finally, assumptions on the forgetting properties of the observations $\mathbf{Y}$ are required. For any sequence of r.v. $Z \stackrel{\text{def}}{=} \{Z_t\}_{t\in\mathbb{Z}}$ on $(\Omega, \widetilde{\mathbb{P}}, \mathcal{F})$, let

$$\mathcal{F}_k^Z \stackrel{\text{def}}{=} \sigma\left(\{Z_u\}_{u\leq k}\right) \quad \text{and} \quad \mathcal{G}_k^Z \stackrel{\text{def}}{=} \sigma\left(\{Z_u\}_{u\geq k}\right) \tag{10}$$

be σ-fields associated to Z. We also define the mixing coefficients by, see [7],

$$\beta^Z(n) = \sup_{u\in\mathbb{Z}} \sup_{B\in\mathcal{G}_{u+n}^Z} |\widetilde{\mathbb{P}}(B|\mathcal{F}_u^Z) - \widetilde{\mathbb{P}}(B)| \,, \forall\, n \geq 0 \,. \tag{11}$$

A5 There exist $C \in [0,1)$ and $\beta \in (0,1)$ s.t. for any $n \geq 0$, $\beta^{\mathbf{Y}}(n) \leq C\beta^n$, where $\beta^{\mathbf{Y}}$ is defined in (11).

Under A4(a), the shift operator preserves the measure $\mathbb{P}_\star$ on $\left(\mathbb{Y}^{\mathbb{Z}}, \mathcal{B}(\mathbb{Y})^{\otimes\mathbb{Z}}\right)$. A5 is used to control the L_p-mean error between the deterministic map $\mathrm{R}(\theta)$ and a BOEM iteration $\bar{\theta}\left(S_{\tau_n}^{\chi,T_{n-1}}(\theta,\mathbf{Y})\right)$ both started from the same point θ. Examples of observation processes satisfying A4(b) and A5 include geometrically ergodic Markov chains as discussed in [20, Section 2.1].

We conclude this set of assumptions by a condition on the block size sequence.

A6 -(γ) The block size sequence $\{\tau_n\}_{n\geq 1}$ satisfies $\sum_{k\geq 0} \tau_k^{-\gamma/2} < \infty$.

4.2 Block Online EM and Limiting EM algorithms

Theorem 4.1. *Assume A2 and A4(a-b). Let $S : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ be a measurable function s.t. A3-(1) holds. For any $\theta \in \Theta$, there exists a $\mathbb{P}_\star$-integrable r.v. $\mathbb{E}_\theta\left[S(X_{-1}, X_0, \mathbf{Y}_0)|\mathbf{Y}\right]$ s.t. for any probability distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,*

$$\sup_{\theta\in\Theta} \left| \Phi_{\theta,0,n}^{\chi,-n}(S, \mathbf{Y}) - \mathbb{E}_\theta\left[S(X_{-1}, X_0, \mathbf{Y}_0)|\mathbf{Y}\right] \right| \leq 2\left(\rho^n + \rho^{n-1}\right) \sup_{(x,x')\in\mathbb{X}^2} |S(x, x', \mathbf{Y}_0)| \quad \mathbb{P}_\star - a.s\,. \tag{12}$$

Define for all $\theta \in \Theta$,

$$\bar{\mathrm{S}}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\star\left[\mathbb{E}_\theta\left[S(X_{-1}, X_0, \mathbf{Y}_0)|\mathbf{Y}\right]\right]\,. \tag{13}$$

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

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \underset{\tau\to+\infty}{\longrightarrow} \bar{\mathrm{S}}(\theta) \quad \mathbb{P}_\star - a.s\,. \tag{14}$$

The proof of Theorem 4.1 is given in Section 6.1. Eqs (1) and (12) show that when $\{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ is a HMM with transition kernels m_θ and g_θ, the limiting statistic $\mathbb{E}_\theta\left[S(X_{-1}, X_0, Y_0)|Y_\mathbb{Z}\right]$ is the a.s. limit of the conditional expectation of $S(X_{-1}, X_0, Y_0)$ given $Y_{-n+1:n}$ when $X_{-n} \sim \chi$, whatever χ is.

As a consequence of (14), when τ is large, the quantity $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ is an approximation of $\bar{\mathrm{S}}(\theta)$. Therefore, the BOEM algorithm (4) is a perturbation of the *Limiting EM* algorithm defined by (7). This remark will be central to address the convergence of BOEM in Section 4.4. We thus start by addressing the convergence of the limiting EM algorithm.

4.3 Asymptotic behavior of the Limiting EM

The convergence of the limiting EM is addressed following the same approach as in [32] for the convergence of the EM algorithm. It relies on a Lyapunov function W w.r.t. to the map R and a set $\mathcal{L}$. The existence of such a Lyapunov function is the key ingredient to identify the limiting points of the algorithm (7).

It is well known that for the EM algorithm, a natural Lyapunov function is based on the (normalized) log-likelihood of the observations. We prove a similar result for the limiting EM: we show that the limiting normalized log-likelihood is related to a Lyapunov function. [13, Lemma 2 and Proposition 1] shows that the normalized log-likelihood converges and this limit, hereafter denoted by $c_\star(\theta)$, is deterministic and does not depend on the initial distribution χ of the Hidden chain. $\theta \to c_\star(\theta)$ will be referred to as the *contrast function*. Proposition 4.2 states that the function $\theta \mapsto \exp(c_\star(\theta))$ is a Lyapunov function for the map R and the set $\mathcal{L} \stackrel{\text{def}}{=} \{\theta \in \Theta;\ \mathrm{R}(\theta) = \theta\}$.

Proposition 4.2. *Assume A1-2, A3-(1) and A4. Then* R *given by (7) and* $\mathrm{W} : \theta \mapsto \exp(c_\star(\theta))$ *are continuous on Θ and satisfy*

(i) *For all* $\theta \in \Theta$, $\mathrm{W} \circ \mathrm{R}(\theta) - \mathrm{W}(\theta) \geq 0$.

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

The proof of Proposition 4.2 is given in Section 6.2. It can be proved that under regularity conditions on the HMM, the set $\mathcal{L}$ is the set of the stationary points of the contrast function $c_\star$; this discussion is detailed in [20, Theorem 4.11]. The following proposition gives a set of sufficient conditions for the convergence of the limiting EM algorithm $\check{\theta}_n = \mathrm{R}(\check{\theta}_{n-1})$ to the set $\mathcal{L}$ (see [16, Proposition 9] for the proof).

Proposition 4.3. *Assume A1-2, A3-(1) and A4. Assume in addition that for any* $M > 0$, *the set* $\mathcal{K}_M \stackrel{\text{def}}{=} \{\theta \in \Theta;\ \mathrm{W}(\theta) \geq M\}$ *is a compact subset of* Θ. *Then, for any initial value* $\check{\theta}_0$,

(i) $\{\mathrm{W}(\check{\theta}_k)\}_{k \geq 0}$ *converges to a connected component of* $\mathrm{W}(\mathcal{K}_{\mathrm{W}(\check{\theta}_0)} \cap \mathcal{L})$.

(ii) *If* $\mathrm{W}(\mathcal{K}_{\mathrm{W}(\check{\theta}_0)} \cap \mathcal{L})$ *has an empty interior, there exists* $w_\star$ *such that* $\{\mathrm{W}(\check{\theta}_k)\}_{k \geq 0}$ *converges to* $w_\star$ *and* $\{\check{\theta}_k\}_{k \geq 0}$ *converges to* $\mathcal{K}_{\mathrm{W}(\check{\theta}_0)} \cap \{\theta \in \mathcal{L};\ \mathrm{W}(\theta) = w_\star\}$.

4.4 Asymptotic behavior of the Block Online EM algorithms

Theorem 4.4 establishes the convergence of BOEM. Let $\mathrm{Cl}(A)$ be the closure of the set A.

Theorem 4.4. *Assume A1-2, A3-(*$\bar{p}_2$*), A4-5 and A6-(*$\bar{p}_1$*) for some* $2 < \bar{p}_1 < \bar{p}_2$. *Assume in addition that* $\mathrm{W}(\mathcal{L})$ *is compact and, for any* $M > 0$, *the level set* $\{\theta \in \Theta; \mathrm{W}(\theta) \geq M\}$ *is compact. Then,*

(a) $\limsup_n p_n < +\infty$ $\mathbb{P}_\star - \mathrm{a.s}$ *where* p_n *is defined in (4).*

(b) $\{\mathrm{W}(\theta_n)\}_{n \geq 0}$ *converges to a connected component of* $\mathrm{W}(\mathcal{L})$.

(c) *If* $\mathrm{W}(\mathcal{L} \cap \mathrm{Cl}(\{\theta_n\}_{n \geq 0}))$ *has an empty interior, there exists* $w_\star$ *s.t.* $\{\mathrm{W}(\theta_n)\}_{n \geq 0}$ *converges almost surely to* $w_\star$ *and* $\{\theta_n\}_{n \geq 0}$ *converges to* $\{\theta \in \mathcal{L}; \mathrm{W}(\theta) = w_\star\}$.

Theorem 4.4 implies that the number of truncations p_n in (4) is almost surely finite so that for a (random) sufficently large n, $\theta_n = \theta_{n-1/2}$. It shows that the BOEM algorithm and the limiting EM have the same asymptotic behavior under the assumptions of Theorem 4.4. The proof is detailed in Section 6.3: it consists in applying the results of [16] on the convergence of a sequence generated by iterated random maps, which are perturbations of a point-to-point map associated to a Lyapunov function. The key ingredient is to prove that the perturbation vanishes when the number of iterations tends to infinity; in our case, this is done through the control of the L_p-mean error when replacing the limiting quantity $\bar{S}(\theta_{n-1})$ by $\bar{S}_{\tau_n}^{\chi, T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ (see Proposition 6.5 in Section 6.3).

A convergence result for the averaged BOEM algorithm can be obtained following the same lines as in the proof of Theorem 4.4. The main ingredient for this proof is the control of the L_p-mean error when replacing $\bar{S}(\theta_{n-1})$ by Σ_n (see Lemma 6.7 below). It can be proved that, along any converging BOEM path, the averaged BOEM algorithm and the limiting EM have the same asymptotic behavior. Details are omitted for brevity.

5 Rate of convergence of the Block Online EM algorithm

We now address the rate of convergence of the BOEM algorithm. To that goal, we consider a converging path $\{\theta_n\}_{n\geq 0}$ with limiting point $\theta_\star \in \mathcal{L}$. First, observe that BOEM can equivalently be defined by recursions on the space of sufficient statistics. Let $\mathrm{G} : \mathcal{S} \to \mathcal{S}$ be the limiting EM map defined on the space of sufficient statistics by

$$\mathrm{G}(s) \stackrel{\text{def}}{=} \bar{\mathrm{S}}(\bar{\theta}(s)) \;, \qquad \forall s \in \mathcal{S} \;, \tag{15}$$

where $\bar{\theta}$ and $\bar{\mathrm{S}}$ are given by A1(c) and (13). The following proposition shows that the convergence of the sequence $\{\theta_n\}_{n\geq 0}$ is equivalent to the convergence of the sufficient statistics $\{\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y})\}_{n\geq 0}$.

Proposition 5.1. *Assume A1, A2, A3-($\bar{p}_2$), A4(a-b), A5 and A6-($\bar{p}_1$) for some $2 < \bar{p}_1 < \bar{p}_2$.*

(i) Let $\theta_\star \in \mathcal{L}$. Set $s_\star \stackrel{\text{def}}{=} \bar{\mathrm{S}}(\theta_\star) = \mathrm{G}(s_\star)$. Then $\mathbb{P}_\star$ − a.s,

$$\lim_{n\to+\infty} \left|\bar{S}^{\chi,T_{n-1}}_{\tau_n}(\theta_{n-1}, \mathbf{Y}) - s_\star\right| \mathbf{1}_{\lim_n \theta_n = \theta_\star} = 0 \;.$$

(ii) Let $s_\star \in \mathcal{S}$ s.t. $\mathrm{G}(s_\star) = s_\star$. Set $\theta_\star \stackrel{\text{def}}{=} \bar{\theta}(s_\star) = \mathrm{R}(\theta_\star)$. Then $\mathbb{P}_\star$ − a.s,

$$\lim_{n\to+\infty} |\theta_n - \theta_\star| \, \mathbf{1}_{\lim_n \bar{S}^{\chi,T_{n-1}}_{\tau_n}(\theta_{n-1},\mathbf{Y}) = s_\star} = 0 \;.$$

The proof of Proposition 5.1 is given in Section 6.4. We thus address equivalently the rate of convergence of the statistics $\{\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y})\}_{n\geq 0}$ to some fixed point of G. Define

$$S_0 \stackrel{\text{def}}{=} \bar{S}^{\chi,0}_{\tau_1}(\theta_0, \mathbf{Y}) \quad \text{and} \quad S_n \stackrel{\text{def}}{=} \bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) \;, \; \forall n \geq 0 \;. \tag{16}$$

It is assumed that

A7 (a) G is twice continuously differentiable on $\mathcal{S}$.

(b) $s_\star = \mathrm{G}(s_\star)$ and there exists $0 < \gamma < 1$ s.t. $\mathrm{sp}(\Gamma) \leq \gamma$ where sp denotes the spectral norm.

A8 (a) $\{\tau_{n+1}/\tau_n\}_{n\geq 0}$ converges to q and $\gamma q < 1$.

(b) $\limsup_n \sum_{k=1}^n \{\left|\frac{\tau_{k+1}}{\tau_k} - q\right| \sqrt{\tau_k} + \log \tau_k\}/\sqrt{T_n} < \infty$.

Under A6, $\lim_n \tau_n = +\infty$. A8 strengthens A6. A8(a) is satisfied for geometric rates of the form $\tau_n \sim a\tau^n$ with $\tau \in (1, \gamma^{-1})$, for polynomial rates $\tau_n \sim cn^b$ with $b > 0$ and sub-exponential rates $\log \tau_n \sim cn^b$ with $c > 0, b \in (0, 1)$, and more generally with sub-geometric rates. A8(b) is satisfied for geometric rates of the form $\tau_n \sim a\tau^n$ with $\tau > 1$, for polynomial rates of the form $\tau_n \sim cn^b$ with $b \geq 1$ and with any sub-exponential rates.

Hereafter, for any sequence of random variables $\{Z_n\}_{n\geq 0}$, write $Z_n = \mathcal{O}_{\mathrm{L}_p}(1)$ if $\limsup_n \mathbb{E}_\star [|Z_n|^p] < \infty$; $Z_n = \mathcal{O}_{\mathrm{a.s}}(1)$ if $\sup_n |Z_n| < +\infty$ $\mathbb{P}_\star$ − a.s and $Z_n = o_{\mathrm{a.s}}(1)$ if $\lim_{n\to+\infty} |Z_n| = 0$ $\mathbb{P}_\star$ − a.s.

Theorem 5.2. *Assume A2, A3-($\bar{p}_2$), A4-5, A6-($\bar{p}_1$), A7 and A8*(a) *for some* $2 < \bar{p}_1 < \bar{p}_2$. *Then, for any* $p \in (2, \bar{p}_2)$,

$$\sqrt{\tau_n}\ [S_n - s_\star]\, \mathbf{1}_{\lim_n S_n = s_\star} = \mathcal{O}_{\mathrm{L}_p}(1) + \sqrt{\tau_n}^{-1} \mathcal{O}_{\mathrm{L}_{p/2}}(1) \mathcal{O}_{\mathrm{a.s}}(1)\ . \tag{17}$$

If in addition A8(b) holds, then for any $p \in (2, \bar{p}_2)$,

$$\sqrt{T_n}\, [\Sigma_n - s_\star]\, \mathbf{1}_{\lim_n S_n = s_\star} = \mathcal{O}_{\mathrm{L}_p}(1) + \frac{n}{\sqrt{T_n}} \mathcal{O}_{\mathrm{L}_{p/2}}(1) \mathcal{O}_{\mathrm{a.s}}(1)\ . \tag{18}$$

The proof of Theorem 5.2 is given in Section 6.4. Eq. (17) shows that the error $S_n - s_\star$ is decomposed into two terms and the L_p-norm of the leading term is inversely proportional to $\tau_n^{1/2}$. Hence, the rate of the BOEM algorithm is closely related to the choice of the number of observations per block. The first column of Table 1 gives explicit rates of convergence for different block-sizes.

In (17), the rate is a function of the number of updates (i.e. the number of iteration of the algorithm). This rate could also be interpreted as a function of the total number of observations up to iteration n. To that goal, let $\phi(n)+1$ be the index of the block the n-th observation belongs to, i.e. $\phi(n)$ is the largest integer s.t.

$$\sum_{k=0}^{\phi(n)} \tau_k < n \leq \sum_{k=0}^{\phi(n)+1} \tau_k\ ,\ (\text{by convention},\ \sum_{k=0}^{-1} \tau_k = 0)\ .$$

The interpolated sequence $\{\theta_n^i\}_{n\geq 0}$ deduced from $\{\theta_n\}_{n\geq 0}$ is thus defined by $\theta_n^i = \theta_{\phi(n)}$ (the value of the interpolated sequence is kept fixed within each block). The second column of Table 1 gives the rate of convergence of this interpolated sequence (deduced from the square root of $\tau_{\phi(n)}$) up to a multiplicative constant. This rate of convergence is slower than $n^{-1/2}$, except in the geometric case. Note however that the geometric case is of weak practical interest, since the parameter is hardly ever updated thus yielding to algorithms which are really sensible to the initial value θ_0 (see Section 3).

Eq. (18) addresses the rate of convergence of the averaged BOEM algorithm. It shows that when the condition A8 is strengthened in such a way that $\lim_n n/\sqrt{T_n} = 0$, averaging reduces the influence of the block-size schedule: the error $\Sigma_n - s_\star$ has a rate of convergence 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. The last column of Table 1 shows that this averaging procedure gives an optimal rate of convergence, whatever the block-size sequence.

τ_n	$\tau_n^{1/2}$	$\tau_{\phi(n)}^{1/2}$	$T_n^{1/2}$	$T_{\phi(n)}^{1/2}$
$c\,n^b$, $(b>1)$	$n^{b/2}$	$n^{b/(2(b+1))}$	$n^{(b+1)/2}$	$n^{1/2}$
$\exp(cn^b)$, $(b\in(0,1))$	$\exp(0.5c\,n^b)$	$n^{1/2}(\ln n)^{(b-1)/(2b)}$	$n^{(1-b)/2}\exp(0.5cn^b)$	$n^{1/2}$
$c\tau^n$, $(\tau\in(1,\gamma^{-1}))$	$\tau^{n/2}$	$n^{1/2}$	$\tau^{n/2}$	$n^{1/2}$

Table 1: Rate of convergence of both algorithms for different block sizes (up to a multiplicative constant)

As a conclusion, the averaged BOEM algorithm reaches the optimal rate of convergence even when the block size sequence $\{\tau_n\}_{n\geq 0}$ slowly increases, thus allowing polynomially increasing size of blocks.

6 Proofs

For $p > 0$ and Z a random variable measurable w.r.t. the σ-algebra $\sigma(Y_n, n \in \mathbb{Z})$, set $\|Z\|_{\star,p} \stackrel{\text{def}}{=} \left(\mathbb{E}_\star\left[|Z|^p\right]\right)^{1/p}$.

6.1 Proof of Theorem 4.1

The proof of Theorem 4.1 relies on auxiliary results about the forgetting properties of HMM. Most of them are really close to published results and their proof is provided in the supplementary material [20, Section 4]. The main novelty is the forgetting property of the bivariate smoothing distribution, and the proof is given in Appendix A.

Lemma 6.1. *Assume A1-2. Let* $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ *s.t.* $\sup_{x,x'} |S(x,x',\mathbf{y}_i)| < +\infty$ *for any* $i \in \mathbb{Z}$. *Then for any* $r > 0$ *and any distribution* χ *on* $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, $\theta \mapsto \Phi_{\theta,0,r}^{\chi,-r}(S,\mathbf{y})$ *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 $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. By definition of $\Phi_{\theta,0,r}^{\chi,-r}(S,\mathbf{y})$ (see (1)) we have to prove that

$$\theta \mapsto \int \chi(\mathrm{d}x_{-r}) \left(\prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, \mathbf{y}_{i+1}) \right) h(x_{-1}, x_0, \mathbf{y}_0)\, \mathrm{d}\lambda(x_{-r+1:r})$$

is continuous for $h(x,x',y) = 1$ and $h(x,x',y) = S(x,x',y)$. By A1(a), the function $\theta \mapsto \prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, \mathbf{y}_{i+1})\, h(x_{-1}, x_0, \mathbf{y}_0)$ is continuous. In addition, under A1, for any $\theta \in \Theta$,

$$\left| \prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, \mathbf{y}_{i+1})\, h(x_{-1}, x_0, \mathbf{y}_0) \right| = |h(x_{-1}, x_0, \mathbf{y}_0)| \exp\left(2r\phi(\theta) + \left\langle \psi(\theta), \sum_{i=-r}^{r-1} S(x_i, x_{i+1}, \mathbf{y}_{i+1}) \right\rangle \right) .$$

Let $\mathcal{K}$ be a compact subset of Θ. By A1, there exist constants C_1 and C_2 s.t.

$$\sup_{\theta\in\mathcal{K}} \left| \prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, \mathbf{y}_{i+1})\, h(x_{-1}, x_0, \mathbf{y}_0) \right| \leq C_1 \sup_{x,x'} |h(x,x',\mathbf{y}_0)| \exp\left(C_2 \sum_{i=-r}^{r-1} \sup_{x,x'} |S(x,x,',\mathbf{y}_{i+1})| \right) .$$

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 $S_s(x,x') \stackrel{\text{def}}{=} S(x,x',\mathbf{Y}_s)$. For a function h, define $\mathrm{osc}(h) \stackrel{\text{def}}{=} \sup_{z,z'} |h(z) - h(z')|$. Note that under A3-(1), $\mathbb{E}_\star[\mathrm{osc}(S_0)] < +\infty$. Under A2, Proposition A.1(ii) implies that for any $\theta \in \Theta$, there exists a r.v. $\Phi_\theta(S,\mathbf{Y})$ s.t. for any $r < s \leq T$,

$$\sup_{\theta\in\Theta} \left| \Phi^{\chi,r}_{\theta,s,T}(S,\mathbf{Y}) - \Phi_\theta(S, \vartheta^s \circ \mathbf{Y}) \right| \leq \left(\rho^{T-s} + \rho^{s-r-1}\right) \mathrm{osc}(S_s) . \tag{19}$$

This concludes the proof of (12). For the proof of (14), we introduce the following decomposition: for all $T > 0$,

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

upon noting that by (2), $\bar{S}^{\chi,T}_\tau(\theta,\mathbf{Y}) = \tau^{-1}\sum_{t=1}^{\tau} \Phi^{\chi,0}_{\theta,t,\tau}\left(S, \vartheta^T \circ \mathbf{Y}\right)$. By (1), (19) and A3-(1) $\mathbb{E}_\star[|\Phi_\theta(S,\mathbf{Y})|] < +\infty$. Under A4(a-b), the ergodic theorem (see e.g. [1, Theorem 24.1, p.314]) states that

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

for any fixed T. By (19),

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\left|\Phi_{\theta,t,\tau}^{\chi,0}\left(S,\vartheta^{T}\circ\mathbf{Y}\right)-\Phi_{\theta}\left(S,\vartheta^{t+T}\circ\mathbf{Y}\right)\right| \leq \frac{1}{\tau}\sum_{t=1}^{\tau}\left(\rho^{\tau-t}+\rho^{t-1}\right)\mathrm{osc}(S_{t+T})\ . \quad (20)$$

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

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\rho^{t-1}\mathrm{osc}(S_{t+T}) = \rho^{\tau-1}Z_\tau + \frac{1-\rho}{\tau}\sum_{t=1}^{\tau-1}t\rho^{t-1}Z_t\ . \quad (21)$$

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

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

We have $\lim_{\tau\to\infty}\tau^{-1}\sum_{t=1}^{\tau-1}t\rho^{t-1}Z_{\tau-t}=0,\ \mathbb{P}_\star-$a.s by using the same arguments as for the second term in (21). Furthermore,

$$\left|(1-\rho)\sum_{t=1}^{\tau-1}\rho^{\tau-t-1}Z_t - \mathbb{E}_\star\left[\mathrm{osc}(S_0)\right]\right| \leq (1-\rho)\sum_{t=1}^{\tau-1}\rho^{\tau-t-1}\left|Z_t-\mathbb{E}_\star\left[\mathrm{osc}(S_0)\right]\right| + \mathbb{E}_\star\left[\mathrm{osc}(S_0)\right]\rho^{\tau-1}\ .$$

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

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

Hence, the RHS in (20) converges $\mathbb{P}_\star-\mathrm{a.s}$ to 0 and this concludes the proof of (14). We now prove that the function $\theta\mapsto\mathbb{E}_\star\left[\Phi_\theta\left(S,\mathbf{Y}\right)\right]$ is continuous by application of the dominated convergence theorem. From Proposition A.1(ii), for any $\mathbf{y}$ s.t. $\mathrm{osc}(S(\cdot,\cdot,\mathbf{y}_0))<\infty$,

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

Then, by Lemma 6.1, $\theta\mapsto\Phi_\theta\left(S,\mathbf{y}\right)$ is continuous for any $\mathbf{y}$ such that $\mathrm{osc}(S(\cdot,\cdot,\mathbf{y}_0)) < +\infty$. In addition, $\sup_{\theta\in\Theta}\left|\Phi_\theta\left(S,\mathbf{Y}\right)\right| \leq \sup_{x,x'}\left|S(x,x',\mathbf{Y}_0)\right|$. We then conclude by A3-(1).

6.2 Proof of Proposition 4.2

Set

$$\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) \stackrel{\text{def}}{=} \log\left(\int \chi(\mathrm{d}x_0)\left\{\prod_{t=1}^{T} m_\theta(x_{t-1},x_t)g_\theta(x_t,\mathbf{Y}_t)\right\}\,\lambda(\mathrm{d}x_1)\cdots\lambda(\mathrm{d}x_T)\right)\,.$$

(Continuity of R *and* W*)* By A1(c) and Theorem 4.1, the function R is continuous. Under A1-2 and A4, there exists a continuous function $c_\star$ on Θ s.t. $\lim_T T^{-1}\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) = c_\star(\theta)$ $\mathbb{P}_\star$ − a.s for any distribution χ on $(\mathbb{X},\mathcal{B}(\mathbb{X})$ and any $\theta\in\Theta$, (see [13, Lemma 2 and Proposition 1], see also [20, Theorem 4.6]). Therefore, W is continuous.

Proof of Proposition 4.2 (i) For all $T>0$ and all $\theta\in\Theta$, define

$$p_\theta(x_{0:T},\mathbf{Y}_{1:T}) \stackrel{\text{def}}{=} \prod_{i=1}^{T} m_\theta(x_{i-1},x_i)g_\theta(x_i,\mathbf{Y}_i)\,. \tag{22}$$

It holds, $\mathbb{P}_\star$ − a.s,

$$\begin{aligned}
&\ell^{\chi,0}_{\mathrm{R}(\theta),T}(\mathbf{Y}) - \ell^{\chi,0}_{\theta,T}(\mathbf{Y}) = \log\frac{\int p_{\mathrm{R}(\theta)}(x_{0:T},\mathbf{Y}_{1:T})\lambda(\mathrm{d}x_{1:T})\chi(\mathrm{d}x_0)}{\int p_\theta(x_{0:T},\mathbf{Y}_{1:T})\lambda(\mathrm{d}x_{1:T})\chi(\mathrm{d}x_0)}\\
&\geq \int \log\left[p_{\mathrm{R}(\theta)}(x_{0:T},\mathbf{Y}_{1:T})\right]\frac{p_\theta(x_{0:T},\mathbf{Y}_{1:T})}{\int p_\theta(z_{0:T},\mathbf{Y}_{1:T})\lambda(\mathrm{d}z_{1:T})\chi(\mathrm{d}z_0)}\lambda(\mathrm{d}x_{1:T})\chi(\mathrm{d}x_0)\\
&- \int \log\left[p_\theta(x_{0:T},\mathbf{Y}_{1:T})\right]\frac{p_\theta(x_{0:T},\mathbf{Y}_{1:T})}{\int p_\theta(z_{0:T},\mathbf{Y}_{1:T})\lambda(\mathrm{d}z_{1:T})\chi(\mathrm{d}z_0)}\lambda(\mathrm{d}x_{1:T})\chi(\mathrm{d}x_0)\,,
\end{aligned}$$

where we used the Jensen inequality. Under Assumption A1(a)

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

Upon noting that

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

this yields

$$\begin{aligned}
\frac{1}{T}\ell^{\chi,0}_{\mathrm{R}(\theta),T}(\mathbf{Y}) - \frac{1}{T}\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) \geq \phi(\mathrm{R}(\theta)) + \left\langle\frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S,\mathbf{Y}),\psi(\mathrm{R}(\theta))\right\rangle\\
- \phi(\theta) - \left\langle\frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S,\mathbf{Y}),\psi(\theta)\right\rangle\,.
\end{aligned}\tag{23}$$

Under A1-4, it holds by Theorem 4.1 and [13, Lemma 2 and Proposition 1] (see also [20, Theorem 4.6(ii)]) that for all $\theta\in\Theta$, $\mathbb{P}_\star$ − a.s,

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

Therefore, when $T \to +\infty$, (23) implies

$$\ln\left(\mathrm{W}(\mathrm{R}(\theta))/\mathrm{W}(\theta)\right) \geq \phi(\mathrm{R}(\theta)) + \left\langle \bar{\mathrm{S}}(\theta), \psi(\mathrm{R}(\theta)) \right\rangle - \phi(\theta) - \left\langle \bar{\mathrm{S}}(\theta), \psi(\theta) \right\rangle \ . \quad (24)$$

By definition of $\bar{\theta}$ and R (see A1(c) and (7)), the RHS is non negative. This concludes the proof of Proposition 4.2(i).

Proof of Proposition 4.2 (ii) 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 (24) 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 $\mathcal{L}$.

6.3 Proof of Theorem 4.4

The proof of Theorem 4.4 follows the same lines as the proof of [16, Theorem 3]. The key ingredient for this proof is the control of the L_p-mean error between the Block Online EM algorithm and the limiting EM. This is the crucial difference with [16]. The proof of this bound is derived in Proposition 6.5 and relies on preliminary lemmas; the detailed proof of Theorem 4.4 is given in [20, Section 3.1].

In the sequel, for all function Ξ on $\Theta \times \mathbb{Y}^{\mathbb{Z}}$ and all $\theta_\star \in \Theta$, we denote by $\mathbb{E}_\star\left[\Xi(\theta, \mathbf{Y})\right]_{\theta=\theta_\star}$ the function $\theta \mapsto \mathbb{E}_\star\left[\Xi(\theta, \mathbf{Y})\right]$ evaluated at $\theta = \theta_\star$. Finally, for any $L \geq 1$, $m \geq 1$ and any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, define

$$\kappa^{\chi}_{L,m}(\boldsymbol{\theta}, \mathbf{Y}) \stackrel{\mathrm{def}}{=} \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \mathbb{E}_\star\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon=\boldsymbol{\theta}} \ . \quad (25)$$

Lemma 6.2. *Assume A2, A3-($\bar{p}$), A4*(a) *and A5 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, any $m \geq 1$, $k, \ell \geq 0$ and any Θ-valued $\mathcal{F}_0^{\mathbf{Y}}$-measurable r.v. $\boldsymbol{\theta}$,*

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

where $\Delta p \stackrel{\mathrm{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$ and β is given by A5.

Proof. For ease of notation χ is dropped from the notation $\kappa^{\chi}_{2um,m}$. By the Berbee Lemma (see [27, Chapter 5]), for any $m \geq 1$, there exists a Θ-valued r.v. $\boldsymbol{\theta}^\star$ on $(\Omega, \mathcal{F}, \mathbb{P}_\star)$ independent from $\mathcal{G}_m^{\mathbf{Y}}$ (see Eq.(10)) s.t.

$$\mathbb{P}_\star\left\{\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star\right\} = \sup_{B \in \mathcal{G}_m^{\mathbf{Y}}} \left|\mathbb{P}_\star(B|\sigma(\boldsymbol{\theta})) - \mathbb{P}_\star(B)\right| \ . \quad (26)$$

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

$$\sum_{u=1}^{k} \kappa_{L_u,m}(\boldsymbol{\theta}, \mathbf{Y}) = \sum_{u=1}^{k} \left\{ \Phi^{\chi, L_u - m}_{\boldsymbol{\theta}, L_u, L_u+m}(S, \mathbf{Y}) - \Phi^{\chi, L_u - m}_{\boldsymbol{\theta}^\star, L_u, L_u+m}(S, \mathbf{Y}) \right\}$$
$$+ \sum_{u=1}^{k} \kappa_{L_u,m}(\boldsymbol{\theta}^\star, \mathbf{Y}) + k \left\{ \mathbb{E}_\star \left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y}) \right]_{\upsilon=\boldsymbol{\theta}^\star} - \mathbb{E}_\star \left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y}) \right]_{\upsilon=\boldsymbol{\theta}} \right\} . \tag{27}$$

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

$$\left\| \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \Phi^{\chi, L-m}_{\boldsymbol{\theta}^\star, L, L+m}(S, \mathbf{Y}) \right\|_{\star,p}$$
$$= \left\| \left(\Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \Phi^{\chi, L-m}_{\boldsymbol{\theta}^\star, L, L+m}(S, \mathbf{Y}) \right) \mathbf{1}_{\boldsymbol{\theta} \neq \boldsymbol{\theta}^\star} \right\|_{\star,p}$$
$$\leq \left\| \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \vartheta^T \circ \mathbf{Y}) - \Phi^{\chi, L-m}_{\boldsymbol{\theta}^\star, L, L+m}(S, \mathbf{Y}) \right\|_{\star,\bar{p}} \mathbb{P}_\star \left\{ \boldsymbol{\theta} \neq \boldsymbol{\theta}^\star \right\}^{\Delta p} .$$

By A4(a), A3-($\bar{p}$), A5, (1), (26) and (11), there exists a constant C_1 s.t. for any $m, L \geq 1$, any distribution χ and any Θ-valued $\mathcal{F}_0^{\mathbf{Y}}$-measurable r.v. $\boldsymbol{\theta}$,

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

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

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

Let us consider the second term in (27). 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 $\mathcal{G}_m^{\mathbf{Y}}$-measurable. $\boldsymbol{\theta}^\star$ is independent from $\mathcal{G}_m^{\mathbf{Y}}$ so that:

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

Define the strong mixing coefficient (see [7])

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

Then, [7, 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 [27, Theorem 2.5],

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

where $N_{(m)}(t) \stackrel{\text{def}}{=} \sum_{i\geq 1} \mathbf{1}_{\alpha^{\kappa_{(\mathbf{m})}(i)}>t}$ and $\mathcal{Q}_{v,m}$ denotes the inverse of the tail function $t \mapsto \mathbb{P}_\star(|\kappa_{L_u,m}(v,\mathbf{Y})| \geq t)$. The sequence $\mathbf{Y}$ being stationary, this inverse function does not depend on u. By A5 and the inequality $\alpha^{\mathbf{Y}}(r) \leq \beta^{\mathbf{Y}}(r)$ (see e.g. [7, 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 $C\beta^{2mb} < 1$. Then, by the Holder inequality applied with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

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

Since U is uniform on $[0,1]$, $\mathcal{Q}_{v,m}(U)$ and $|\kappa_{L_u,m}(v,\mathbf{Y})|$ have the same distribution, see [27]. Then, by Lemma A.3 and A3-($\bar{p}$), there exists a constant C s.t. for any $v \in \Theta$, any $m \geq 1$,

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

which concludes the proof. □

Lemma 6.3. *Assume A2, A3-($\bar{p}$), A4*(a) *and A5 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 $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,*

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

where $\kappa^\chi_{L,m}$ and β are defined by (25) *and A5, $v_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$ and $\Delta p \stackrel{\text{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$.*

Proof. We write,

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

Observe that by definition $\theta_n \in \mathcal{F}^{\mathbf{Y}}_{T_n}$. Then, by Lemma 6.2, there exists a constant C s.t. for any $m_n \geq 1$ and any $\ell \geq 0$,

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

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

Lemma 6.4. *Assume A2, A3-($\bar{p}$) and A4*(a) *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 $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,*

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

where

$$\widetilde{\rho}_n \stackrel{\mathrm{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=2m_n}^{q_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n} \mathbf{Y}) , \tag{28}$$

and $\kappa^{\chi}_{L,m}$ is defined by (25).

Proof. By (1) and (2), $\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) - \bar{\mathrm{S}}(\theta_n) - \widetilde{\rho}_n = \sum_{i=1}^{4} g_{i,n}$ where

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

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

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

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

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

$$\tau_{n+1} |g_{1,n}| \leq \sum_{t=\tau_{n+1}-m_n+1}^{\tau_{n+1}} \left(\rho^{m_n - 1} + \rho^{\tau_{n+1} - t} \right) \mathrm{osc}\{ S(\cdot, \cdot, \mathbf{Y}_{t+T_n}) \}$$

$$+ \sum_{t=1}^{m_n} \left(\rho^{m_n} + \rho^{t-1} \right) \mathrm{osc}\{ S(\cdot, \cdot, \mathbf{Y}_{t+T_n}) \} + 2\rho^{m_n - 1} \sum_{t=m_n+1}^{\tau_{n+1}-m_n} \mathrm{osc}\{ S(\cdot, \cdot, \mathbf{Y}_{t+T_n}) \} ,$$

where we used Proposition A.1(i) and Remark A.2 in the last inequality. By A3-($\bar{p}$) and A4(a), there exists C s.t. $\|g_{1,n}\|_{\star,p} \leq C \left(\rho^{m_n} + \tau^{-1}_{n+1} \right)$ In the case $\tau_{n+1} \leq 2m_n$, it can be proved along the same lines that $\|g_{1,n}\|_{\star,p} \leq C \left(\rho^{\tau_{n+1} - m_n} + \tau^{-1}_{n+1} \right)$. For $g_{2,n}$ and $g_{3,n}$, we use the bounds

$$\left|\Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S,\vartheta^{T_n}\circ\mathbf{Y})-\mathbb{E}_\star\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S,\mathbf{Y})\right]_{\theta=\theta_n}\right|$$
$$\leq \sup_{(x,x')\in\mathbb{X}^2}|S(x,x',\mathbf{Y}_{T_n+t})|+\mathbb{E}_\star\left[\sup_{(x,x')\in\mathbb{X}^2}|S(x,x',\mathbf{Y}_0)|\right] .$$

Then, by A4(a),

$$\left\|\Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S,\vartheta^{T_n}\circ\mathbf{Y})-\mathbb{E}_\star\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S,\mathbf{Y})\right]_{\theta=\theta_n}\right\|_{\star,p}$$
$$\leq 2\left\|\sup_{(x,x')\in\mathbb{X}^2}|S(x,x',\mathbf{Y}_0)|\right\|_{\star,p} ,$$

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

$$|g_{4,n}|\leq\sup_{\theta\in\Theta}\left|\mathbb{E}_\star\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S,\mathbf{Y})-\mathbb{E}_\theta\left[S(X_{-1},X_0,\mathbf{Y}_0)|\mathbf{Y}\right]\right]\right|$$
$$\leq 2\rho^{m_n-1}\mathbb{E}_\star\left[\mathrm{osc}\{S(\cdot,\cdot,\mathbf{Y}_0)\}\right] ,$$

where we used Theorem 4.1 in the last inequality. This concludes the proof. □

Proposition 6.5. *Assume A2, A3-($\bar{p}$), A4*(a) *and A5 for some $\bar{p}>2$. Let $p\in(2,\bar{p})$. There exists a constant C s.t. for any $n\geq 1$ and any distribution χ on $(\mathbb{X},\mathcal{B}(\mathbb{X}))$,*

$$\left\|\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n,\mathbf{Y})-\bar{\mathrm{S}}(\theta_n)\right\|_{\star,p}\leq\frac{C}{\sqrt{\tau_{n+1}}} .$$

Proof. Let m_n, v_n be positive integers s.t. $1\leq m_n\leq\tau_{n+1}$ and $\tau_{n+1}=2v_nm_n+r_n$, where $0\leq r_n<2m_n$. Set $\Delta p\stackrel{\text{def}}{=}1/p-1/\bar{p}$. By the Minkowski inequality combined with Lemmas 6.3, 6.4 applied with $q_n\stackrel{\text{def}}{=}2v_nm_n$, there exists a constant C s.t.

$$\left\|\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n,\mathbf{Y})-\bar{\mathrm{S}}(\theta_n)\right\|_{\star,p}\leq C\left[\rho^{m_n}+\frac{m_n}{\tau_{n+1}}+\beta^{m_n\Delta p}+\frac{1}{\sqrt{\tau_{n+1}}}\right] .$$

The proof is concluded by choosing $m_n=\lfloor-\log\tau_{n+1}/\left(\log\rho\vee\Delta p\log\beta\right)\rfloor$. □

6.4 Proof of Section 5

6.4.1 Proof of Proposition 5.1

Let $\bar{\mathrm{S}}$ be given by (13). By Proposition 6.5 and A6-($\bar{p}_1$), $\lim_n\left(\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n,\mathbf{Y})-\bar{\mathrm{S}}(\theta_n)\right)=0$ $\mathbb{P}_\star$−a.s. By Theorem 4.1, $\bar{\mathrm{S}}$ is continuous. Hence, $\lim_n\left|\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n,\mathbf{Y})-\bar{\mathrm{S}}(\theta_\star)\right|\mathbf{1}_{\lim_n\theta_n=\theta_\star}=0$ $\mathbb{P}_\star$ − a.s and the proof of (i) follows. Since $\bar{\theta}$ is continuous, (ii) follows.

6.4.2 Proof of Theorem 5.2, Eq. (17)

Since $\mathrm{G}(s_\star) = s_\star$, we write

$$S_n - s_\star = \Gamma\left(S_{n-1} - s_\star\right) + S_n - \mathrm{G}(S_{n-1}) + \mathrm{G}(S_{n-1}) - \mathrm{G}(s_\star) - \Gamma\left(S_{n-1} - s_\star\right) \ .$$

Define $\{\mu_n\}_{n\geq 0}$ and $\{\rho_n\}_{n\geq 0}$ s.t. $\mu_0 = 0$, $\rho_0 = S_0 - s_\star$ and

$$\mu_n \overset{\text{def}}{=} \Gamma\mu_{n-1} + e_n \ , \qquad \rho_n \overset{\text{def}}{=} S_n - s_\star - \mu_n \ , \qquad n \geq 1 \ , \tag{29}$$

where,

$$e_n \overset{\text{def}}{=} S_n - \bar{\mathrm{S}}(\theta_n) \ , \qquad n \geq 1 \ . \tag{30}$$

Proposition 6.6. *Assume A2, A3-($\bar{p}_2$), A4-5, A6-($\bar{p}_1$), A7 and A8*(a) *for some $2 < \bar{p}_1 < \bar{p}_2$. Then for any $p \in (2, \bar{p}_2)$, $\sqrt{\tau_n}\mu_n = \mathcal{O}_{\mathrm{L}_p}(1)$ and $\tau_k\rho_k\mathbf{1}_{\lim_n S_n = s_\star} = \mathcal{O}_{\mathrm{L}_{p/2}}(1)\mathcal{O}_{\mathrm{a.s}}(1)$.*

The proof of Proposition 6.6 is on the same lines as the proof of [16, Theorem 6]. The main ingredient is the control of $\|\mu_n\|_{\star,p}$ which is a consequence of [24, Result 178, p. 39] and Proposition 6.5. The detailed proof is thus omitted and postponed to the supplementary material [20, Section 3.2].

6.4.3 Proof of Theorem 5.2, Eq.(18)

We preface the proof by the following lemma.

Lemma 6.7. *Assume A2, A3-($\bar{p}_2$), A4-5, A7, A8*(b) *for some $\bar{p}_2 > 2$. For any $p \in (2, \bar{p}_2)$,*

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

where e_n is given by (30).

Proof. Let $p \in (2, \bar{p}_2)$. In the sequel, C is a constant independent on n and whose value may change upon each appearance. Let $1 \leq m_n \leq \tau_{n+1}$ and set $v_n \overset{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$. By Lemma 6.4 applied with $q_k \overset{\text{def}}{=} 2v_k m_k$, we have,

$$\left\| \sum_{k=1}^{n} \tau_{k+1} e_k \right\|_{\star,p} \leq C \left(\sum_{k=1}^{n} \{\tau_{k+1}\rho^{m_k \wedge (\tau_{k+1} - m_k)} + m_k\} + \left\| \sum_{k=1}^{n} \{\delta_k + \zeta_k\} \right\|_{\star,p} \right) \ ,$$

where δ_k and ζ_k are defined by

$$\begin{aligned}
F_{t,k}(\theta_k, \mathbf{Y}) &\overset{\text{def}}{=} \Phi^{\chi, t-m_k}_{\theta_k, t, t+m_k}(S, \vartheta^{T_k} \circ \mathbf{Y}) \ , \\
\delta_k &\overset{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}_\star\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \mathcal{F}^{\mathbf{Y}}_{T_k} \right] \right\} \ , \\
\zeta_k &\overset{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}_\star\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \mathcal{F}^{\mathbf{Y}}_{T_k} \right] - \mathbb{E}_\star\left[\Phi^{\chi, -m_k}_{\theta, 0, m_k}(S, \mathbf{Y}) \right]_{\theta=\theta_k} \right\} \ ,
\end{aligned}$$

and $\mathcal{F}^{\mathbf{Y}}_{T_k}$ is given by (10). We will prove below that there exists C s.t.

$$\|\zeta_k\|_{\star,p} \leq C\,\beta^{m_k/pb}\tau_{k+1}\ , \qquad \forall k \geq 1 \tag{31}$$

$$\left\|\sum_{k=1}^{n}\delta_k\right\|_{\star,p} \leq C\sqrt{T_{n+1}} + C\sum_{k=1}^{n}\tau_{k+1}\beta^{m_k/pb}\ , \qquad \forall n \geq 1 \tag{32}$$

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

We turn to the proof of (31). By the Berbee Lemma (see [27, Chapter 5]) and A5, there exist $C \in [0,1)$ and $\beta \in (0,1)$ s.t. for all $k \geq 1$, there exists a random variable $\mathbf{Y}^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k}$ on $(\Omega, \mathcal{F}, \mathbb{P}_\star)$ independent from $\mathcal{F}^{\mathbf{Y}}_{T_k}$ with the same distribution as $\mathbf{Y}_{T_k+m_k:T_{k+1}+m_k}$ and

$$\mathbb{P}_\star\left\{\mathbf{Y}^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k} \neq \mathbf{Y}_{T_k+m_k:T_{k+1}+m_k}\right\} \leq C\beta^{m_k}\ . \tag{33}$$

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

$$\zeta_k = \sum_{t=2m_k}^{2v_km_k}\left\{\mathbb{E}_\star\left[F_{t,k}(\theta_k,\mathbf{Y})\middle|\mathcal{F}^{\mathbf{Y}}_{T_k}\right] - \mathbb{E}_\star\left[F_{t,k}(\theta_k,\mathbf{Y}^{\star,(k)})\middle|\mathcal{F}^{\mathbf{Y}}_{T_k}\right]\right\}\ . \tag{34}$$

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

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

The Minkowski inequality followed by the Holder inequality with $a \stackrel{\text{def}}{=} \bar{p}_2/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$, combined with (33), A4(a), Lemma A.3 and A3-($\bar{p}_2$) yield (31). We now prove (32). Upon noting that δ_k is $\mathcal{F}^{\mathbf{Y}}_{T_{k+1}}$-measurable and δ_k is a martingale increment, the Rosenthal inequality (see [17, Theorem 2.12, p.23]) states that $\|\sum_{k=1}^n \delta_k\|_{\star,p} \leq C\left(\sum_{k=1}^n I_k^{(1)}\right)^{1/p} + CI_n^{(2)}$ where

$$I_k^{(1)} \stackrel{\text{def}}{=} \mathbb{E}_\star\left[\left|\sum_{t=2m_k}^{2v_km_k}\left\{F_{t,k}(\theta_k,\mathbf{Y}) - \mathbb{E}_\star\left[F_{t,k}(\theta_k,\mathbf{Y})\middle|\mathcal{F}^{\mathbf{Y}}_{T_k}\right]\right\}\right|^p\right]$$

$$I_n^{(2)} \stackrel{\text{def}}{=} \left\|\left(\sum_{k=1}^n\mathbb{E}_\star\left[\left|\sum_{t=2m_k}^{2v_km_k}\left\{F_{t,k}(\theta_k,\mathbf{Y}) - \mathbb{E}_\star\left[F_{t,k}(\theta_k,\mathbf{Y})\middle|\mathcal{F}^{\mathbf{Y}}_{T_k}\right]\right\}\right|^2\middle|\mathcal{F}^{\mathbf{Y}}_{T_k}\right]\right)^{1/2}\right\|_{\star,p}\ .$$

Using again $\mathbb{E}_\star\left[F_{t,k}(\theta_k,\mathbf{Y}^{\star,(k)})\middle|\mathcal{F}^{\mathbf{Y}}_{T_k}\right] = \mathbb{E}_\star\left[F_{t,k}(\theta,\mathbf{Y})\right]_{\theta=\theta_k}$ and (34)

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

By Lemma 6.3 and (31), there exists C s.t. for any $k \geq 1$

$$I_k^{(1)} \leq C \left(\tau_{k+1}^{p/2} + \tau_{k+1}^{p} \beta^{m_k/b} \right) , \tag{35}$$

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 (35), $I_n^{(2)} \leq C \sqrt{T_{n+1}} + C \sum_{k=1}^n \tau_{k+1} \beta^{m_k/pb}$. This concludes the proof of (32). □

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

$$\bar{\mu}_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{k=1}^n \tau_k \mu_{k-1} \quad \text{and} \quad \bar{\rho}_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{k=1}^n \tau_k \rho_{k-1} . \tag{36}$$

Proposition 6.8. *Assume A2, A3-($\bar{p}_2$), A4-5, A6-($\bar{p}_1$), A7 and A8 for some $2 < \bar{p}_1 < \bar{p}_2$. For any $p \in (2, \bar{p}_2)$,*

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

Proof. Set $A \stackrel{\text{def}}{=} (I - q\Gamma)$. Under A7, A^{-1} exists. By (29) and (36),

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

The result now follows from Proposition 6.6, Lemma 6.7 and A8. The proof of the second assertion follows from (36) and Proposition 6.6. □

A Bivariate smoothing distribution

For a function $g : \mathbb{X}^2 \to \mathbb{R}$, set $\quad \mathrm{osc}(g) \stackrel{\text{def}}{=} \sup_{z, z' \in \mathbb{X}^2} |g(z) - g(z')| \quad .$

Proposition A.1. *Assume A2. Let χ, $\widetilde{\chi}$ be two distributions on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. For any measurable function $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ s.t. $\sup_{x,x'} |h|(x, x', \mathbf{y}_s) < +\infty$ for any $s \in \mathbb{Z}$*

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

$$\sup_{\theta \in \Theta} \left| \Phi_{\theta,s,t}^{\widetilde{\chi},r} (h, \mathbf{y}) - \Phi_{\theta,s,t+\ell_2}^{\chi,r-\ell_1} (h, \mathbf{y}) \right| \leq \left(\rho^{s-1-r} + \rho^{t-s} \right) \mathrm{osc}(h(\cdot, \cdot, y_s)) . \tag{37}$$

(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 $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and any $r < s \leq t$

$$\sup_{\theta \in \Theta} \left| \Phi_{\theta,s,t}^{\chi,r} (h, \mathbf{y}) - \Phi_\theta (h, \vartheta^s \circ \mathbf{y}) \right| \leq \left(\rho^{s-r-1} + \rho^{t-s} \right) \mathrm{osc}(h(\cdot, \cdot, y_s)) . \tag{38}$$

Remark A.2. (a) If $\chi = \widetilde{\chi}$, $\ell_1 = 0$ and $\ell_2 \geq 1$, (37) becomes

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

(b) if $\ell_2 = 0$ and $\ell_1 \geq 1$, (37) becomes

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

Proof. We will use the shorthand h_s for $h_s(x,x') \stackrel{\mathrm{def}}{=} h(x,x',y_s)$.

(i) Let r, s, t such that $r < s \leq t$, $\ell_1, \ell_2 \geq 1$, and $\theta \in \Theta$. Define the distribution $\chi_{\theta,r-\ell_1:r}$ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ by

$$\chi_{\theta,r-\ell_1:r}(A) \stackrel{\mathrm{def}}{=} \frac{\int \chi(\mathrm{d}x_{r-\ell_1}) L_{\theta,r-\ell_1:r-1}(x_{r-\ell_1}, \mathrm{d}x_r) 1_A(x_r)}{\int \chi(\mathrm{d}x_{r-\ell_1}) L_{\theta,r-\ell_1:r-1}(x_{r-\ell_1}, \mathbb{X})} \ , \quad \forall A \in \mathcal{B}(\mathbb{X}) \ .$$

We write $\left| \Phi^{\widetilde{\chi},r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r-\ell_1}_{\theta,s,t+\ell_2}(h,\mathbf{y}) \right| \leq \widetilde{T}_1 + \widetilde{T}_2$ where, by using (1),

$$\widetilde{T}_1 \stackrel{\mathrm{def}}{=} \left| \frac{\int \widetilde{\chi}_r(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r, \mathrm{d}x_{s-1}) h_s(x_{s-1}, x_s) L_{\theta,s-1}(x_{s-1}, \mathrm{d}x_s) L_{\theta,s:t-1}(x_s, \mathbb{X})}{\int \widetilde{\chi}_r(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r, \mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:p}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r, \mathrm{d}x_{s-1}) h_s(x_{s-1}, x_s) L_{\theta,s-1}(x_{s-1}, \mathrm{d}x_s) L_{\theta,s:t-1}(x_s, \mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r, \mathbb{X})} \right| \ ,$$

and

$$\widetilde{T}_2 \stackrel{\mathrm{def}}{=} \left| \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r, \mathrm{d}x_{s-1}) h_s(x_{s-1}, x_s) L_{\theta,s-1}(x_{s-1}, \mathrm{d}x_s) L_{\theta,s:t-1}(x_s, \mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r, \mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r, \mathrm{d}x_{s-1}) h_s(x_{s-1}, x_s) L_{\theta,s-1}(x_{s-1}, \mathrm{d}x_s) L_{\theta,s:t+\ell_2-1}(x_s, \mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t+\ell_2-1}(x_r, \mathbb{X})} \right| \ .$$

Set $\bar{h}_{s,t} : x \mapsto \int F_{\theta,s-1,t}(x, \mathrm{d}x_s) h_s(x, x_s)$ where $F_{\theta,s-1,t}$ is the forward smoothing kernel (see e.g. [20, Eq.(24)]). Then,

$$\widetilde{T}_1 = \left| \frac{\int \widetilde{\chi}_r(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r, \mathrm{d}x_{s-1}) \bar{h}_{s,t}(x_{s-1}) L_{\theta,s-1:t-1}(x_{s-1}, \mathbb{X})}{\int \widetilde{\chi}_r(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r, \mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r, \mathrm{d}x_{s-1}) \bar{h}_{s,t}(x_{s-1}) L_{\theta,s-1:t-1}(x_{s-1}, \mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r, \mathbb{X})} \right| \ .$$

By [20, Lemma 4.2(i)],

$$\widetilde{T}_1 \leq \rho^{s-1-r}\mathrm{osc}(\bar{h}_{s,t}) \leq 2\rho^{s-1-r} \sup_{x\in\mathbb{X}} |\bar{h}_{s,t}(x)| \leq 2\rho^{s-1-r} \sup_{(x,x')\in\mathbb{X}^2} |h_s(x,x')| \ .$$

Set $\widetilde{h}_s : x \mapsto \int \mathrm{B}_{\theta,s-1}^{\chi_{\theta,r-\ell_1:s-1},s-1}(x, \mathrm{d}x_{s-1}) h_s(x_{s-1}, x)$, where $\mathrm{B}_{\theta,s-1}^{\chi_{\theta,r-\ell_1:s-1},s-1}$ is the backward smoothing kernel (see e.g. [20, Eq.(25)]). Then,

$$\widetilde{T}_2 = \left| \frac{\int \chi_{\theta,r-\ell_1:s}(\mathrm{d}x_s)\widetilde{h}_s(x_s) L_{\theta,s:t-1}(x_s, \mathrm{d}x_t) L_{\theta,t:t+\ell_2-1}(x_t, \mathbb{X})}{\int \chi_{\theta,r-\ell_1:s}(\mathrm{d}x_s) L_{\theta,s:t-1}(x_s, \mathrm{d}x_t) L_{\theta,t:t+\ell_2-1}(x_t, \mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:s}(\mathrm{d}x_s)\widetilde{h}_s(x_s) L_{\theta,s:t-1}(x_s, \mathbb{X})}{\int \chi_{\theta,r-\ell_1:s}(\mathrm{d}x_s) L_{\theta,s:t-1}(x_s, \mathbb{X})} \right| .$$

Then, by [20, Lemma 4.2(ii)],

$$\widetilde{T}_2 \leq \rho^{t-s} \mathrm{osc}(\widetilde{h}_s) \leq 2\rho^{t-s} \sup_{x \in \mathbb{X}} |\widetilde{h}_s(x)| \leq 2\rho^{t-s} \sup_{(x,x') \in \mathbb{X}^2} |h_s(x, x')| .$$

Hence, for any constant $c \in \mathbb{R}$,

$$\left| \Phi_{\theta,s,t}^{\widetilde{\chi},r}(h, \mathbf{y}) - \Phi_{\theta,s,t+\ell_2}^{\chi,r-\ell_1}(h, \mathbf{y}) \right| = \left| \Phi_{\theta,s,t}^{\widetilde{\chi},r}(h - c, \mathbf{y}) - \Phi_{\theta,s,t+\ell_2}^{\chi,r-\ell_1}(h - c, \mathbf{y}) \right|$$
$$\leq 2\left(\rho^{s-1-r} + \rho^{t-s}\right) \sup_{(x,x') \in \mathbb{X}^2} |h_s(x, x') - c| .$$

Since $\mathrm{osc}(h) = 2 \inf_{c \in \mathbb{R}} \left\{ \sup_{(x,x') \in \mathbb{X}^2} |h_s(x, x') - c| \right\}$, this concludes the proof.

(ii) By (37), for any increasing sequence of non negative integers $(r_\ell)_{\ell \geq 0}$, $(t_\ell)_{\ell \geq 0}$ s.t. $\lim r_\ell = \lim t_\ell = +\infty$, the sequence $\{\Phi_{\theta,0,t_\ell}^{\chi,-r_\ell}(h, \mathbf{y})\}_{\ell \geq 0}$ is a Cauchy sequence uniformly in θ and χ. Then, there exists a limit $\Phi_\theta(h, \mathbf{y})$ s.t.

$$\lim_{\ell \to +\infty} \sup_{\chi} \sup_{\theta \in \Theta} \left| \Phi_{\theta,0,t_\ell}^{\chi,-r_\ell}(h, \mathbf{y}) - \Phi_\theta(h, \mathbf{y}) \right| = 0 . \tag{39}$$

We write, for any $r < s \leq t$ and any $\ell \geq 1$

$$\left| \Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{y}) - \Phi_\theta(h, \vartheta^s \circ \mathbf{y}) \right|$$
$$\leq \left| \Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{y}) - \Phi_{\theta,s,t+\ell}^{\chi,r-\ell}(h, \mathbf{y}) \right| + \left| \Phi_{\theta,s,t+\ell}^{\chi,r-\ell}(h, \mathbf{y}) - \Phi_\theta(h, \vartheta^s \circ \mathbf{y}) \right| .$$

Since $\Phi_{\theta,s,t+\ell}^{\chi,r-\ell}(h, \mathbf{y}) = \Phi_{\theta,0,t+\ell-s}^{\chi,r-\ell-s}(h, \vartheta^s \circ \mathbf{y})$, Proposition A.1(i) yields

$$\left| \Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{y}) - \Phi_\theta(h, \vartheta^s \circ \mathbf{y}) \right| \leq \left(\rho^{s-r-1} + \rho^{t-s}\right) \mathrm{osc}(h_s)$$
$$+ \left| \Phi_{\theta,0,t+\ell-s}^{\chi,r-\ell-s}(h, \vartheta^s \circ \mathbf{y}) - \Phi_\theta(h, \vartheta^s \circ \mathbf{y}) \right| .$$

The proof is concluded by (39). □

Lemma A.3 is a consequence resp. of Eq. (1) and Proposition A.1(ii).

Lemma A.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*

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

References

[1] P. Billingsley. *Probability and Measure*. Wiley, New York, 3rd edition, 1995.

[2] O. Cappé. Online sequential Monte Carlo EM algorithm. In *IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.

[3] O. Cappé. Online EM algorithm for Hidden Markov Models. *To appear in J. Comput. Graph. Statist.*, 2011.

[4] O. Cappé and E. Moulines. Online Expectation Maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593–613, 2009.

[5] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

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

[7] J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

[8] M. Del Moral, A. Doucet, and S.S Singh. Forward smoothing using sequential Monte Carlo. Preprint, Dec 2010.

[9] P. Del Moral and A. Guionnet. Large deviations for interacting particle systems: applications to non-linear filtering. *Stoch. Proc. App.*, 78:69–95, 1998.

[10] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probab. Theory Related Fields*, 126(3):395–420, 2003.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

[12] R. Douc, G. Fort, E. Moulines, and P. Priouret. Forgetting the initial distribution for hidden Markov models. *Stochastic Processes and their Applications*, 119(4):1235–1256, 2009.

[13] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32(5):2254–2304, 2004.

[14] J. Durbin and S. J. Koopman. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *J. Roy. Statist. Soc. B*, 62:3–29, 2000.

[15] Y. Ephraim and N. Merhav. Hidden Markov Processes. *IEEE Trans. on information theory*, 18(6):1518–1570, 2002.

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

[17] P. Hall and C. C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, London, 1980.

[18] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, 1997.

[19] S. Le Corff. Convergence of a stochastic block online Expectation Maximization algorithm. Technical report, 2011.

[20] S. Le Corff and G. Fort. Supplementary to "Online Expectation Maximization based algorithms for inference in Hidden Markov Models". Technical report, arXiv, 2011.

[21] S. Le Corff, G. Fort, and E. Moulines. Online EM algorithm to solve the SLAM problem. In *IEEE Workshop on Statistical Signal Processing (SSP)*, 2011.

[22] F. Le Gland and L. Mevel. Recursive estimation in HMMs. In *Proc. IEEE Conf. Decis. Control*, pages 3468–3473, 1997.

[23] G. Mongillo and S. Denève. Online learning with hidden Markov models. *Neural Computation*, 20(7):1706–1716, 2008.

[24] G. Pólya and G. Szegő. *Problems and Theorems in Analysis. Vol. II.* Springer, 1976.

[25] B. T. Polyak. A new method of stochastic approximation type. *Autom. Remote Control*, 51:98–107, 1990.

[26] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

[27] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer, 1990.

[28] Vladislav B. Tadić. Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden Markov models. *IEEE Trans. Inf. Theor.*, 56:6406–6432, December 2010.

[29] D. M. Titterington. Recursive parameter estimation using incomplete data. *J. Roy. Statist. Soc. B*, 46(2):257–267, 1984.

[30] R. Van Handel. Uniform time average consistency of Monte Carlo particle filters. *Stoch. Proc. Appl.*, 119:3835–3861, 2009.

[31] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1989.

[32] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103, 1983.