



HAL
open science

Supplement paper to "Online Expectation Maximization based algorithms for inference in Hidden Markov Models"

Sylvain Le Corff, Gersende Fort

► **To cite this version:**

Sylvain Le Corff, Gersende Fort. Supplement paper to "Online Expectation Maximization based algorithms for inference in Hidden Markov Models". 2011. hal-00615615v1

HAL Id: hal-00615615

<https://hal.science/hal-00615615v1>

Preprint submitted on 19 Aug 2011 (v1), last revised 16 Oct 2012 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Supplement paper to “Online Expectation Maximization based algorithms for inference in Hidden Markov Models”

Sylvain Le Corff^{*†} and Gersende Fort[†]

August 19, 2011

This is a supplementary material to the paper [6].

It contains technical discussions and/or results adapted from published papers: in section 2, we show that geometrically ergodic Markov chains satisfy the assumptions H4b and H5; in sections 3 and 4, we provide results - useful for the proofs of some theorems in [6] - which are close to existing results in the literature.

It also contains in Section 5, additional plots for the illustration of comments in [6, Section 3].

To make this supplement paper as self-contained as possible, we decided to rewrite in Section 1 the model and the main definitions introduced in [6].

1 Assumptions and Model

Let $\mathbf{Y} = \{Y_t\}_{t \in \mathbb{Z}}$ be the observation process defined on $(\Omega, \mathbb{P}_*, \mathcal{F})$ 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 $\{m_\theta(x, x') d\lambda(x')\}_{\theta \in \Theta}$ 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 $\{g_\theta(x, y) d\nu(y)\}_{\theta \in \Theta}$ be a family of transition kernels on $(\mathbb{X} \times \mathcal{B}(\mathbb{Y}))$, where ν is a measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$.

It is assumed:

- H1** (a) There exist continuous functions $\phi : \Theta \rightarrow \mathbb{R}$, $\psi : \Theta \rightarrow \mathbb{R}^d$ and $S : \mathbb{X} \times \mathbb{X} \times \mathbb{Y}^{\mathbb{Z}} \rightarrow \mathbb{R}^d$ s.t.

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

^{*}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

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 $\bar{\theta} : \mathcal{S} \rightarrow \Theta$ s.t. for any $s \in \mathcal{S}$,

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

H2 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 $\rho \stackrel{\text{def}}{=} 1 - (\sigma_-/\sigma_+)$.

H3-(γ) $\mathbb{E}_* [\sup_{x, x' \in \mathbb{X}^2} |S(x, x', \mathbf{Y}_0)|^\gamma] < +\infty$.

- H4** (a) Under \mathbb{P}_* , \mathbf{Y} is a stationary sequence.
- (b) The shift operator is ergodic with respect to \mathbb{P}_* .
- (c) $\mathbb{E}_* [|\log b_-(\mathbf{Y}_0)| + |\log b_+(\mathbf{Y}_0)|] < +\infty$ where

$$b_-(y) \stackrel{\text{def}}{=} \inf_{\theta \in \Theta} \int g_\theta(x, y) \lambda(\mathrm{d}x) , \quad (1)$$

$$b_+(y) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \int g_\theta(x, y) \lambda(\mathrm{d}x) . \quad (2)$$

For any sequence of r.v. $Z \stackrel{\text{def}}{=} \{Z_t\}_{t \in \mathbb{Z}}$ on $(\Omega, \tilde{\mathbb{P}}, \mathcal{F})$, let

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

be σ -fields associated to Z . We also define the mixing coefficients by, see [3],

$$\beta^Z(n) = \sup_{u \in \mathbb{Z}} \sup_{B \in \mathcal{G}_{u+n}^Z} |\tilde{\mathbb{P}}(B | \mathcal{F}_u^Z) - \tilde{\mathbb{P}}(B)| , \forall n \geq 0 . \quad (3)$$

H5 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 (3).

H6-(γ) The block size sequence $\{\tau_n\}_{n \geq 1}$ satisfies $\sum_{k \geq 0} \tau_k^{-\gamma/2} < \infty$.

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

$$\bar{S}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_* [\mathbb{E}_\theta [S(X_{-1}, X_0, \mathbf{Y}_0) | \mathbf{Y}]] . \quad (4)$$

$$R(\theta) \stackrel{\text{def}}{=} \bar{\theta}(\bar{S}(\theta)) . \quad (5)$$

$$G(s) \stackrel{\text{def}}{=} \bar{S}(\bar{\theta}(s)) , \quad \forall s \in \mathcal{S} , \quad (6)$$

where $\bar{\theta}$ is given by H1(c).

- H7** (a) G is twice continuously differentiable on \mathcal{S} .
- (b) $s_* = G(s_*)$ and there exists $0 < \gamma < 1$ s.t. $\operatorname{sp}(\Gamma) \leq \gamma$ where sp denotes the spectral norm.

Set

$$T_n \stackrel{\text{def}}{=} \sum_{i=1}^n \tau_i, \quad T_0 \stackrel{\text{def}}{=} 0.$$

- H8** (a) $\{\tau_{n+1}/\tau_n\}_{n \geq 0}$ converges to q and $\gamma q < 1$.
 (b) $\limsup_n \sum_{k=1}^n \left\{ \left| \frac{\tau_{k+1}}{\tau_k} - q \right| \sqrt{\tau_k} + \log \tau_k \right\} / \sqrt{T_n} < \infty$.

2 Checking H4(b) and H5

2.1 Checking H4(b)

The following discussion has been suggested by R. Douc¹ and E. Moulines². The authors would like to thank them for the fruitful discussions.

Assumption H4(b) can be easily proved when \mathbf{Y} is a positive recurrent and ψ -irreducible Markov chain. Assume first that \mathbf{Y} is a one-sided Markov chain $\{\mathbf{Y}_t\}_{t \geq 0}$ with invariant probability π defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P}_\pi)$. We may choose $\Omega \stackrel{\text{def}}{=} \mathbb{Y}^{\mathbb{N}}$ and $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{B}(\mathbb{Y})^{\otimes \mathbb{N}}$, \mathbf{Y} being the canonical process. To prove H4(b), we show that \mathbf{Y} is mixing (see [3, Chapter 13]) i.e.

$$\lim_{k \rightarrow +\infty} \mathbb{P}_\pi \{ \mathbf{Y} \in A, \vartheta^k(\mathbf{Y}) \in B \} = \mathbb{P}_\pi \{ \mathbf{Y} \in A \} \mathbb{P}_\pi \{ \mathbf{Y} \in B \}, \forall A, B \in \mathcal{B}(\mathbb{Y})^{\otimes \mathbb{N}}. \quad (7)$$

This is sufficient to prove (7) when A is a cylinder, i.e. when there exists $p \in \mathbb{N}$ such that $A = \{ \omega \stackrel{\text{def}}{=} \{w_i\}_{i \geq 0} \in \mathbb{Y}^{\mathbb{N}}; (w_{i_1}, \dots, w_{i_p}) \in H \}$, where $H \in \mathcal{B}(\mathbb{Y})^{\otimes p}$ and (i_1, \dots, i_p) is a p -tuple of distinct non-negative integers. For all sufficiently large k , by the Markov property,

$$\begin{aligned} \mathbb{P}_\pi \{ \mathbf{Y} \in A, \vartheta^k(\mathbf{Y}) \in B \} &= \mathbb{E}_\pi \left[\mathbf{1}_A(\mathbf{Y}) \mathbf{1}_B(\vartheta^k(\mathbf{Y})) \right] \\ &= \mathbb{E}_\pi \left[\mathbf{1}_A(\mathbf{Y}) \mathbb{E}_{\mathbf{Y}_{i_p}} \left[\mathbb{E}_{\mathbf{Y}_{k-i_p}} \left[\mathbf{1}_B(\mathbf{Y}) \right] \right] \right]. \end{aligned}$$

Under the stated assumption on \mathbf{Y} , we can choose \mathbb{Y}_0 such that $\pi(\mathbb{Y}_0) = 1$ and for any $x \in \mathbb{Y}_0$, $\mathbb{E}_x \left[\mathbb{E}_{\mathbf{Y}_{k-i_p}} \left[\mathbf{1}_B(\mathbf{Y}) \right] \right] \xrightarrow[k \rightarrow +\infty]{} \mathbb{P}_\pi \{ B \}$. The proof is then concluded by the dominated convergence theorem. This result implies that if \mathbf{Y} is a two sided Markov chain $\{\mathbf{Y}_t\}_{t \in \mathbb{Z}}$, with a positive recurrent and ψ -irreducible transition kernel, then \mathbf{Y} is ergodic. Indeed, \mathbb{P}_π may be extended on $(\mathbb{Y}^{\otimes \mathbb{Z}}, \mathcal{B}(\mathbb{Y})^{\otimes \mathbb{Z}})$ where $\mathcal{B}(\mathbb{Y})^{\otimes \mathbb{Z}}$ is generated by $\cup_{i < j \in \mathbb{Z}} \mathcal{F}_{i,j}$, with, for any integers $i < j$, $\mathcal{F}_{i,j} \stackrel{\text{def}}{=} \sigma(\{\mathbf{Y}_k; i \leq k \leq j\})$. For any integers i, j and k s.t. $i < j$ and any $A, B \in \mathcal{F}_{i,j}$,

$$\mathbb{P}_\pi \{ A \cap \vartheta^{-k}(B) \} = \mathbb{P}_\pi \{ \vartheta^{-i}(A) \cap \vartheta^{-k}(\vartheta^{-i}(B)) \},$$

where $\vartheta^{-i}(A), \vartheta^{-i}(B) \in \sigma(\{\mathbf{Y}_k; k \geq 0\})$. Then, we can conclude as above.

¹randal.douc@it-sudparis.eu

²eric.moulines@telecom-paristech.fr

2.2 Checking H5

- 1) Combining [9, Chapter 9] and recent results on the control of the ergodicity of Markov chains by coupling technique, see [4], it can be proved that H5 holds for ψ -irreducible, aperiodic and geometrically ergodic Markov chains.
- 2) Upon noting that, for all $n \geq 0$, $\beta^{\mathbf{Y}}(n) \leq \beta^{(\mathbf{X}, \mathbf{Y})}(n)$, we can similarly prove that H4(b) (and H5) hold when \mathbf{Y} is the observation process of a joint ψ -irreducible and aperiodic Markov chain (\mathbf{X}, \mathbf{Y}) (e.g. (\mathbf{X}, \mathbf{Y}) is a hidden Markov model). In this case, irreducibility, aperiodicity and geometric ergodicity have to be established for the Markov transition kernel Q of the $\mathbb{X} \times \mathbb{Y}$ valued Markov chain $\{(X_t, Y_t)\}_{t \in \mathbb{Z}}$.

3 Detailed proofs of [6]

Recall the following definition from [6]: for a distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, positive integers T, τ and $\theta \in \Theta$, set

$$\bar{S}_\tau^{\chi, T}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \Phi_{\theta, t, T+\tau}^{\chi, T}(S, \mathbf{Y}) , \quad (8)$$

where S is the function given by H1(a) and

$$\begin{aligned} & \Phi_{\theta, s, t}^{\chi, r}(S, \mathbf{y}) \\ \stackrel{\text{def}}{=} & \frac{\int \chi(dx_r) \{ \prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, y_{i+1}) \} S(x_{s-1}, x_s, y_s) d\lambda(x_{r+1:t})}{\int \chi(dx_r) \{ \prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, y_{i+1}) \} d\lambda(x_{r+1:t})} . \end{aligned} \quad (9)$$

3.1 Proof of [6, Theorem 4.4]

We check the assumptions of [5, Proposition 9] and [5, Proposition 11] with $T(\theta) \stackrel{\text{def}}{=} R(\theta)$ (see (5)), $F_n(\theta) \stackrel{\text{def}}{=} \bar{\theta} \left(\bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y}) \right)$ and $\mathcal{L} \stackrel{\text{def}}{=} \{\theta \in \Theta; R(\theta) = \theta\}$. We start by checking the conditions of [5, Proposition 11]. Under the stated assumptions, (a) holds. For (c) we prove that for any compact subset $\mathcal{K} \subset \Theta$,

$$\left| W \circ R(\theta_n) - W \circ \bar{\theta}(\bar{S}_{\tau_{n+1}}^{\chi, T_n}(\theta_n, \mathbf{Y})) \right| \mathbf{1}_{\theta_n \in \mathcal{K}} \xrightarrow{n \rightarrow +\infty} 0 \quad \mathbb{P}_* - \text{a.s.} \quad (10)$$

By Theorem 4.1, the function \bar{S} given by (4) is continuous on Θ and then $\bar{S}(\mathcal{K}) \stackrel{\text{def}}{=} \{s \in \mathcal{S}; \exists \theta \in \mathcal{K}, s = \bar{S}(\theta)\}$ is compact and, for any $\delta > 0$ (small enough), we can define the compact subset $\bar{S}(\mathcal{K}, \delta) \stackrel{\text{def}}{=} \{s \in \mathbb{R}^d; d(s, \bar{S}(\mathcal{K})) \leq \delta\}$ of \mathcal{S} , where $d(s, \bar{S}(\mathcal{K})) \stackrel{\text{def}}{=} \inf_{s' \in \bar{S}(\mathcal{K})} |s - s'|$. Let $\delta > 0$ (small enough) and $\varepsilon > 0$. Since $W \circ \bar{\theta}$ is continuous (see H1(c) and [6, Proposition 4.2]) and $\bar{S}(\mathcal{K}, \delta)$ is compact, $W \circ \bar{\theta}$ is uniformly continuous on $\bar{S}(\mathcal{K}, \delta)$ and there exists $\eta > 0$ s.t.,

$$\forall x, y \in \bar{S}(\mathcal{K}, \delta) , \quad |x - y| \leq \eta \Rightarrow |W \circ \bar{\theta}(x) - W \circ \bar{\theta}(y)| \leq \varepsilon . \quad (11)$$

Set $\alpha \stackrel{\text{def}}{=} \delta \wedge \eta$ and $\Delta S_n \stackrel{\text{def}}{=} |\bar{S}(\theta_n) - \bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y})| \mathbf{1}_{\theta_n \in \mathcal{K}}$. We write,

$$\begin{aligned} & \mathbb{P}_\star \left\{ \left| \mathbb{W} \circ \bar{\theta}(\bar{S}(\theta_n)) - \mathbb{W} \circ \bar{\theta}(\bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y})) \right| \mathbf{1}_{\theta_n \in \mathcal{K}} \geq \varepsilon \right\} \\ &= \mathbb{P}_\star \left\{ \left| \mathbb{W} \circ \bar{\theta}(\bar{S}(\theta_n)) - \mathbb{W} \circ \bar{\theta}(\bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y})) \right| \mathbf{1}_{\theta_n \in \mathcal{K}} \geq \varepsilon; \Delta S_n > \delta \right\} \\ &\quad + \mathbb{P}_\star \left\{ \left| \mathbb{W} \circ \bar{\theta}(\bar{S}(\theta_n)) - \mathbb{W} \circ \bar{\theta}(\bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y})) \right| \mathbf{1}_{\theta_n \in \mathcal{K}} \geq \varepsilon; \Delta S_n \leq \delta \right\} \\ &\leq \mathbb{P}_\star \{ \Delta S_n > \delta \} + \mathbb{P}_\star \{ \Delta S_n > \eta \} \leq 2\mathbb{P}_\star \{ \Delta S_n > \alpha \} . \end{aligned}$$

By the Markov inequality and [6, Proposition 6.5], since $2 < \bar{p}_1 < \bar{p}_2$, there exists a constant C s.t.

$$\begin{aligned} & \mathbb{P}_\star \left\{ \left| \mathbb{W} \circ \bar{\theta}(\bar{S}(\theta_n)) - \mathbb{W} \circ \bar{\theta}(\bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y})) \right| \mathbf{1}_{\theta_n \in \mathcal{K}} \geq \varepsilon \right\} \\ &\leq \frac{2}{\alpha^{\bar{p}_1}} \mathbb{E}_\star \left[|\bar{S}(\theta_n) - \bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y})|^{\bar{p}_1} \right] \leq C \left[\frac{1}{\tau_{n+1}} \right]^{\bar{p}_1/2} . \end{aligned}$$

(10) follows from H6- (\bar{p}_1) and the Borel-Cantelli lemma. The proof of the condition (b) follows the same lines. By [5, Proposition 11], this implies that $\limsup_n p_n < +\infty$ \mathbb{P}_\star -a.s and that $\{\theta_n\}_{n \geq 0}$ is a compact sequence \mathbb{P}_\star -a.s. For the other statements, we apply [5, Proposition 9]. $\mathcal{L} \cap \mathcal{K}$ is compact since \mathcal{L} is closed and \mathcal{K} is compact. We now prove that for any compact subset $\mathcal{K} \subset \Theta$,

$$|\mathbb{W}(\theta_{n+1}) - \mathbb{W} \circ \mathbb{R}(\theta_n)| \mathbf{1}_{\theta_n \in \mathcal{K}} \xrightarrow{n \rightarrow +\infty} 0 \quad \mathbb{P}_\star \text{-a.s.} \quad (12)$$

Since $\limsup_n p_n < +\infty$ \mathbb{P}_\star -a.s, it is sufficient to prove this convergence on the set $\{\omega \in \Omega; \limsup_n p_n(\omega) < +\infty\}$. For any ω s.t. $\limsup_n p_n(\omega) < +\infty$, there exists (a random) n_0 s.t., for any $n \geq n_0$, $p_n(\omega) = p_{n+1}(\omega)$ and then $\theta_{n+1}(\omega) = \theta_{n+1/2}(\omega)$, see [6, Eq.(4)]. Therefore (12) follows from (10).

3.2 Proof of [6, Proposition 6.6]

We start with rewriting some definitions and assumptions introduced in [6]. Define the sequences S_n, μ_n, ρ_n , $n \geq 0$ by

$$S_0 \stackrel{\text{def}}{=} \bar{S}_{\tau_1}^{X, 0}(\theta_0, \mathbf{Y}) \quad \text{and} \quad S_n \stackrel{\text{def}}{=} \bar{S}_{\tau_{n+1}}^{X, T_n}(\theta_n, \mathbf{Y}), \quad \forall n \geq 0, \quad (13)$$

where $\bar{S}_\tau^{X, T}$ is given by (8); $\mu_0 = 0$, $\rho_0 = S_0 - s_\star$ and

$$\mu_n \stackrel{\text{def}}{=} \Gamma \mu_{n-1} + e_n, \quad \rho_n \stackrel{\text{def}}{=} S_n - s_\star - \mu_n, \quad n \geq 1, \quad (14)$$

where,

$$e_n \stackrel{\text{def}}{=} S_n - \bar{S}(\theta_n), \quad n \geq 1, \quad (15)$$

and \bar{S} is given by (4).

(proof) Let $p \in (2, \bar{p}_2)$. By (14), for all $n \geq 1$, $\mu_n = \sum_{k=0}^{n-1} \Gamma^k e_{n-k}$. By H7 and the Minkowski inequality, for all $n \geq 1$, $\|\mu_n\|_{\star, p} \leq \sum_{k=0}^{n-1} \gamma^k \|e_{n-k}\|_{\star, p}$. By (13) and [6, Proposition 6.5], there exists a constant C s.t. for any $n \geq 1$,

$$\|\mu_n\|_{\star, p} \leq C \sum_{k=0}^{n-1} \gamma^k \sqrt{\frac{1}{\tau_{n+1-k}}}.$$

By [7, Result 178, p. 39] and H8(a) (upon noting that $q \geq 1$ so that $\sqrt{q}\gamma < 1$), this yields $\sqrt{\tau_n}\mu_n = \mathcal{O}_{L_p}(1)$.

By H7, using a Taylor expansion with integral form of the remainder term,

$$\begin{aligned} & \mathbf{G}(S_{n-1}) - \mathbf{G}(s_\star) - \Gamma(S_{n-1} - s_\star) \\ &= \sum_{i,j=1}^d (S_{n-1,i} - s_{\star,i})(S_{n-1,j} - s_{\star,j}) R_{n-1}(i, j) \\ &= \sum_{i,j=1}^d (\mu_{n-1,i} + \rho_{n-1,i})(\mu_{n-1,j} + \rho_{n-1,j}) R_{n-1}(i, j), \end{aligned}$$

where $x_{n,i}$ denotes the i -th component of $x_n \in \mathbb{R}^d$ and

$$R_n(i, j) \stackrel{\text{def}}{=} \int_0^1 (1-t) \frac{\partial^2 \mathbf{G}}{\partial s_i \partial s_j} (s_\star + t(S_n - s_\star)) dt, \quad n \in \mathbb{N}, 1 \leq i, j \leq d.$$

Observe that under H7, $\limsup_n |R_n| \mathbf{1}_{\lim_n \theta_n = \theta_\star} < \infty$ w.p.1. Define for $n \geq 1$ and $k \leq n$,

$$H_n \stackrel{\text{def}}{=} \sum_{i=1}^d (2\mu_{n,i} + \rho_{n,i}) R_n(i, \cdot), \quad r_n \stackrel{\text{def}}{=} \sum_{i,j=1}^d R_n(i, j) \mu_{n,i} \mu_{n,j}, \quad (16)$$

$$\delta_n \stackrel{\text{def}}{=} \bar{\mathbf{S}}(\theta_n) - \mathbf{G}(S_{n-1}), \quad \psi(n, k) \stackrel{\text{def}}{=} (\Gamma + H_n) \cdots (\Gamma + H_k) \quad (17)$$

with the convention $\psi(n, n+1) \stackrel{\text{def}}{=} \text{Id}$. By (14),

$$\rho_n = \psi(n-1, 0) \rho_0 + \sum_{k=0}^{n-1} \psi(n-1, k+1) r_k + \sum_{k=1}^n \psi(n-1, k) \delta_k. \quad (18)$$

Since $\sqrt{\tau_n}\mu_n = \mathcal{O}_{L_p}(1)$, H6-(\bar{p}_1) implies that $\mu_n \xrightarrow[n \rightarrow +\infty]{} 0$ \mathbb{P}_\star -a.s. Then, by (14), $\rho_n \mathbf{1}_{\lim_n S_n = s_\star} \xrightarrow[n \rightarrow +\infty]{} 0$ \mathbb{P}_\star -a.s and by (16) $\lim_{n \rightarrow +\infty} |H_n| \mathbf{1}_{\lim_n S_n = s_\star} = 0$ \mathbb{P}_\star -a.s. Let $\tilde{\gamma} \in (\gamma, q^{-1})$, where γ is given by H7 and q by H8(a). Since $\lim_{n \rightarrow +\infty} |H_n| \mathbf{1}_{\lim_n S_n = s_\star} = 0$, there exists a \mathbb{P}_\star -a.s finite random variable Z_1 s.t., for all $0 \leq k \leq n-1$,

$$|\psi(n-1, k)| \mathbf{1}_{\lim_n S_n = s_\star} \leq \tilde{\gamma}^{n-k} Z_1 \mathbf{1}_{\lim_n S_n = s_\star}. \quad (19)$$

Therefore, $|\psi(n-1, 0)\rho_0| \mathbf{1}_{\lim_n S_n = s_*} \leq \tilde{\gamma}^n Z_1 |\rho_0| \mathbb{P}_* - \text{a.s.}$ and, by H3-(\bar{p}_2), (9), (8) and (13) $\mathbb{E}_* [|\rho_0|^{\bar{p}_2}] < +\infty$ which implies that $\rho_0 < +\infty \mathbb{P}_* - \text{a.s.}$ Since $q\tilde{\gamma} < 1$, the first term in the RHS of (18) is $\tau_n^{-1} \mathcal{O}_{L^p}(1) \mathcal{O}_{\text{a.s.}}(1)$.

We now consider the second term in the RHS of (18). From equation (19),

$$\left| \sum_{k=0}^{n-1} \psi(n-1, k+1) r_k \right| \mathbf{1}_{\lim_n S_n = s_*} \leq Z_1 \sum_{k=0}^{n-1} \tilde{\gamma}^{n-k-1} |r_k| \mathbf{1}_{\lim_n S_n = s_*} \mathbb{P}_* - \text{a.s.}$$

By (16) and H7, there exists a $\mathbb{P}_* - \text{a.s.}$ finite random variable Z_2 s.t.

$$|r_k| \mathbf{1}_{\lim_n S_n = s_*} \leq Z_2 \sum_{i,j=1}^d \mu_{k,i} \mu_{k,j}, \mathbb{P}_* - \text{a.s.}$$

In addition, since $\sqrt{\tau_n} \mu_n = \mathcal{O}_{L^p}(1)$, there exists a constant C s.t.

$$\left\| \sum_{k=0}^{n-1} \tilde{\gamma}^{n-k-1} \sum_{i,j=1}^d \mu_{k,i} \mu_{k,j} \right\|_{\star, p/2} \leq C \sum_{k=0}^{n-1} \frac{\tilde{\gamma}^{n-k-1}}{\tau_k}.$$

Applying again [7, Result 178, p. 39] yields that the second term in the RHS of (18) is $\tau_n^{-1} \mathcal{O}_{\text{a.s.}}(1) \mathcal{O}_{L^{p/2}}(1)$. We finally consider the third term in the RHS of (18). By H1(c), on the set $\{\omega \in \Omega; \lim_{n \rightarrow +\infty} S_n(\omega) = s_*\}$ we have $\lim_{n \rightarrow +\infty} \bar{\theta}(S_n(\omega)) = \bar{\theta}(s_*)$. Hence, for any $\omega \in \Omega_*$, the set $\{\bar{\theta}(S_n(\omega))\}_{n \geq 0}$ is compact and $\theta_{n+1}(\omega) = \theta_{n+1/2}(\omega)$ for all large n . By (6) and (4), there exists a random integer $n_0(\omega)$ s.t. for all $n \geq n_0(\omega)$, $\delta_n(\omega) = 0$. Then, there exists a $\mathbb{P}_* - \text{a.s.}$ -finite random variable Z_3 s.t. for all $n \geq 1$,

$$\left| \sum_{k=1}^n \psi(n-1, k) \delta_k \mathbf{1}_{\lim_n S_n = s_*} \right| \leq \tilde{\gamma}^n Z_3.$$

Since under H7 $\lim_n \tau_n \tilde{\gamma}_n = 0$, this implies that the third term in the RHS of (18) is $\tau_n^{-1} \mathcal{O}_{\text{a.s.}}(1)$.

4 General results on HMM

In this section, we derive results on the forgetting properties of HMM (section 4.1), on the asymptotic behavior of the normalized log-likelihood (section 4.2) and on the normalized score (section 4.3).

We consider a HMM with kernels $m(x, x') d\lambda(x')$ onto $\mathbb{X} \times \mathcal{B}(\mathbb{X})$ and $g(x, y) d\nu(y)$ on $(\mathbb{X} \times \mathcal{B}(\mathbb{Y}))$. \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 $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$; ν is a measure on $(\mathbb{Y}, \mathcal{B}(\mathbb{Y}))$.

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

$$\begin{aligned} & \Phi_{s,t}^{\chi,r}(h, \mathbf{y}) \\ \stackrel{\text{def}}{=} & \frac{\int \chi(dx_r) \{ \prod_{i=r}^{t-1} m(x_i, x_{i+1}) g(x_{i+1}, y_{i+1}) \} h(x_{s-1}, x_s, y_s) d\lambda(x_{r+1:t})}{\int \chi(dx_r) \{ \prod_{i=r}^{t-1} m(x_i, x_{i+1}) g(x_{i+1}, y_{i+1}) \} d\lambda(x_{r+1:t})}, \end{aligned} \quad (20)$$

for any bounded function h .

For any $s \in \mathbb{Z}$ and any $A \in \mathcal{B}(\mathbb{X})$, define

$$L_s(x, A) \stackrel{\text{def}}{=} \int m(x, x') g(x', y_{s+1}) \mathbb{1}_A(x') \lambda(dx'), \quad (21)$$

and, for any $s \leq t$ denote by $L_{\theta,s:t}$ the composition of the kernels defined by

$$L_{s:s} \stackrel{\text{def}}{=} L_s, \quad L_{s:u+1}(x, A) \stackrel{\text{def}}{=} \int L_{s:u}(x, dx') L_{u+1}(x', A).$$

By convention, $L_{s:s-1}$ is the identity kernel: $L_{s:s-1}(x, A) = \delta_x(A)$. For any sequence $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and any function $h : \mathbb{X}^2 \times \mathbb{Y} \rightarrow \mathbb{R}$, denote by h_s the function on $\mathbb{X}^2 \rightarrow \mathbb{R}$ given by

$$h_s(x, x') \stackrel{\text{def}}{=} h(x, x', y_s). \quad (22)$$

With these notations, equation (20) becomes

$$\Phi_{s,t}^{\chi,r}(h, \mathbf{y}) = \frac{\int \chi(dx_r) L_{r:s-2}(x_s, dx_{s-1}) h_s(x_{s-1}, x_s) L_{s:t-1}(x_s, \mathbb{X})}{\int \chi(dx_r) L_{r:t-1}(x_r, \mathbb{X})}. \quad (23)$$

4.1 Forward and Backward forgetting

For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, any probability distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and for any integers such that $r \leq s < t$, let us define two Markov kernels on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ by

$$F_{s,t}(x, A) \stackrel{\text{def}}{=} \frac{\int L_s(x, dx_{s+1}) \mathbb{1}_A(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X})}{L_{s:t-1}(x, \mathbb{X})}, \quad (24)$$

$$B_s^{\chi,r}(x, A) \stackrel{\text{def}}{=} \frac{\int \phi_{s|r:s}^{\chi,r}(dx_s) \mathbb{1}_A(x_s) m(x_s, x)}{\int \phi_{s|r:s}^{\chi,r}(dx_s) m(x_s, x)}, \quad (25)$$

where

$$\phi_{s|r:s}^{\chi,r}(A) \stackrel{\text{def}}{=} \frac{\int \chi(dx_r) L_{r:s-1}(x_r, dx_s) \mathbb{1}_A(x_s)}{\int \chi(dx_r) L_{r:s-1}(x_r, \mathbb{X})}.$$

Finally, the Dobrushin coefficient of a Markov kernel $F : (\mathbb{X}, \mathcal{B}(\mathbb{X})) \rightarrow [0, 1]$ is defined by:

$$\delta(F) \stackrel{\text{def}}{=} \frac{1}{2} \sup_{(x, x') \in \mathbb{X}^2} \|F(x, \cdot) - F(x', \cdot)\|_{\text{TV}}.$$

Lemma 4.1. *Assume that there exist positive numbers σ_-, σ_+ such that $\sigma_- \leq m(x, x') \leq \sigma_+$ for any $x, x' \in \mathbb{X}$. Then for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, $\delta(\mathbb{F}_{s,t}) \leq \rho$ and $\delta(\mathbb{B}_s^{\chi,r}) \leq \rho$ where $\rho \stackrel{\text{def}}{=} \sigma_-/\sigma_+$.*

Proof. Let r, s, t be such that $r \leq s < t$. Under the stated assumptions,

$$\begin{aligned} & \int L_s(x_s, dx_{s+1}) \mathbb{1}_A(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \\ & \geq \sigma_- \int g(x_{s+1}, y_{s+1}) \mathbb{1}_A(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \lambda(dx_{s+1}) \end{aligned}$$

and

$$L_{s:t-1}(x_s, \mathbb{X}) \leq \sigma_+ \int g(x_{s+1}, y_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \lambda(dx_{s+1}).$$

This yields to

$$\mathbb{F}_{s,t}(x_s, A) \geq \frac{\sigma_-}{\sigma_+} \frac{\int g(x_{s+1}, y_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \mathbb{1}_A(x_{s+1}) \lambda(dx_{s+1})}{\int g(x_{s+1}, y_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \lambda(dx_{s+1})}.$$

Similarly, the assumption implies

$$\mathbb{B}_s^{\chi,r}(x_{s+1}, A) \geq \frac{\sigma_-}{\sigma_+} \phi_{s|r;s}^{\chi,r}(A),$$

which gives the upper bound for the Dobrushin coefficients, see [2, Lemma 4.3.13]. \square

Lemma 4.2. *Assume that there exist positive numbers σ_-, σ_+ such that $\sigma_- \leq m(x, x') \leq \sigma_+$ for any $x, x' \in \mathbb{X}$. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$.*

(i) *for any bounded function h , any probability distributions χ and $\tilde{\chi}$ and any integers $r \leq s \leq t$*

$$\left| \frac{\int \chi(dx_r) L_{r:s-1}(x_r, dx_s) h(x_s) L_{s:t-1}(x_s, \mathbb{X})}{\int \chi(dx_r) L_{r:t-1}(x_r, \mathbb{X})} - \frac{\int \tilde{\chi}(dx_r) L_{r:s-1}(x_r, dx_s) h(x_s) L_{s:t-1}(x_s, \mathbb{X})}{\int \tilde{\chi}(dx_r) L_{r:t-1}(x_r, \mathbb{X})} \right| \leq \rho^{s-r} \text{osc}(h), \quad (26)$$

(ii) *for any bounded function h , for any non-negative functions f and \tilde{f} and any integers $r \leq s \leq t$*

$$\left| \frac{\int \chi(dx_s) h(x_s) L_{s:t-1}(x_s, dx_t) f(x_t)}{\int \chi(dx_s) L_{s:t-1}(x_s, dx_t) f(x_t)} - \frac{\int \chi(dx_s) h(x_s) L_{s:t-1}(x_s, dx_t) \tilde{f}(x_t)}{\int \chi(dx_s) L_{s:t-1}(x_s, dx_t) \tilde{f}(x_t)} \right| \leq \rho^{t-s} \text{osc}(h). \quad (27)$$

Proof of (i). See [2, Proposition 4.3.23].

Proof of (ii) When $s = t$, then (ii) is equal to

$$\left| \frac{\int \chi(\mathrm{d}x_t) h(x_t) f(x_t)}{\int \chi(\mathrm{d}x_t) f(x_t)} - \frac{\int \chi(\mathrm{d}x_t) h(x_t) \tilde{f}(x_t)}{\int \chi(\mathrm{d}x_t) \tilde{f}(x_t)} \right|.$$

This is of the form $(\eta - \tilde{\eta}) h$ where η and $\tilde{\eta}$ are probability distributions on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. Then,

$$|(\eta - \tilde{\eta}) h| \leq \frac{1}{2} \|\eta - \tilde{\eta}\|_{\mathrm{TV}} \mathrm{osc}(h) \leq \mathrm{osc}(h).$$

Let $s < t$. By definition of the backward smoothing kernel, see (25),

$$\mathrm{B}_s^{\chi, s}(x_{s+1}, A) = \frac{\int \chi(\mathrm{d}x_s) 1_A(x_s) m(x_s, x_{s+1})}{\int \chi(\mathrm{d}x_s) m(x_s, x_{s+1})}.$$

Therefore,

$$\begin{aligned} & \int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t) \\ &= \int \chi(\mathrm{d}x_s) L_s(x_s, \mathrm{d}x_{s+1}) \mathrm{B}_s^{\chi, s} h(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathrm{d}x_t) f(x_t). \end{aligned}$$

By repeated application of the backward smoothing kernel we have

$$\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t) = \int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \mathrm{B}_{t-1:s}^{\chi, s} h(x_t) f(x_t),$$

where we denote by $\mathrm{B}_{t-1:s}^{\chi, s}$ the composition of the kernels defined by induction for $s \leq u$

$$\mathrm{B}_{s:s}^{\chi, s} \stackrel{\mathrm{def}}{=} \mathrm{B}_s^{\chi, s}, \quad \mathrm{B}_{u:s}^{\chi, s}(x, A) \stackrel{\mathrm{def}}{=} \int \mathrm{B}_u^{\chi, s}(x, \mathrm{d}x') \mathrm{B}_{u-1:s}^{\chi, s}(x', A).$$

Finally, by definition of $\phi_{t|s:t}^{\chi, s}$

$$\begin{aligned} & \left| \frac{\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)}{\int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)} - \frac{\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \tilde{f}(x_t)}{\int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \tilde{f}(x_t)} \right| \\ &= \left| \frac{\phi_{t|s:t}^{\chi, s} [(\mathrm{B}_{t-1:s}^{\chi, s} h) f]}{\phi_{t|s:t}^{\chi, s} [f]} - \frac{\phi_{t|s:t}^{\chi, s} [(\mathrm{B}_{t-1:s}^{\chi, s} h) \tilde{f}]}{\phi_{t|s:t}^{\chi, s} [\tilde{f}]} \right|. \end{aligned}$$

This is of the form $(\eta - \tilde{\eta}) \mathrm{B}_{t-1:s}^{\chi, s} h$ where η and $\tilde{\eta}$ are probability distributions on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. The proof of the second statement is completed upon noting that

$$\begin{aligned} |\eta \mathrm{B}_{t-1:s}^{\chi, s} h - \tilde{\eta} \mathrm{B}_{t-1:s}^{\chi, s} h| &\leq \frac{1}{2} \|\eta - \tilde{\eta}\|_{\mathrm{TV}} \mathrm{osc}(\mathrm{B}_{t-1:s}^{\chi, s} h) \\ &\leq \frac{1}{2} \|\mu - \tilde{\mu}\|_{\mathrm{TV}} \delta(\mathrm{B}_{t-1:s}^{\chi, s}) \mathrm{osc}(h) \leq \rho^{t-s} \mathrm{osc}(h), \end{aligned}$$

where we used Lemma 4.1 in the last inequality. \square

4.2 Limiting normalized log-likelihood

Define for any $r \leq s$,

$$\delta_{\theta,s}^{\chi,r}(\mathbf{y}) \stackrel{\text{def}}{=} \ell_{\theta,s+1}^{\chi,r}(\mathbf{y}) - \ell_{\theta,s}^{\chi,r}(\mathbf{y}), \quad (28)$$

where $\ell_{\theta,s+1}^{\chi,r}(\mathbf{y})$ is defined by

$$\ell_{\theta,s+1}^{\chi,r}(\mathbf{Y}) \stackrel{\text{def}}{=} \log \left(\int \chi(\mathrm{d}x_r) \prod_{u=r+1}^{s+1} m_{\theta}(x_{u-1}, x_u) g_{\theta}(x_u, \mathbf{Y}_u) \lambda(\mathrm{d}x_{r+1}) \cdots \lambda(\mathrm{d}x_{s+1}) \right). \quad (29)$$

For any $T > 0$ and any probability distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, we thus have

$$\ell_{\theta,T}^{\chi,0}(\mathbf{y}) = \sum_{s=0}^{T-1} \left(\ell_{\theta,s+1}^{\chi,0}(\mathbf{y}) - \ell_{\theta,s}^{\chi,0}(\mathbf{y}) \right) = \sum_{s=0}^{T-1} \delta_{\theta,s}^{\chi,0}(\mathbf{y}). \quad (30)$$

It is established in Lemma 4.3 that for any $\theta \in \Theta$, $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, $s \geq 0$ and any initial distribution χ , the sequence $\{\delta_{\theta,s}^{\chi,s-r}(\mathbf{y})\}_{r \geq 0}$ is a Cauchy sequence and its limit does not depend upon χ . Regularity conditions on this limit are given in Lemmas 4.4 and 4.5. Finally, Theorem 4.6 shows that for any θ , $\lim_T T^{-1} \ell_{\theta,T}^{\chi,0}(\mathbf{Y})$ exists w.p.1. and this limit is a (deterministic) continuous function in θ .

Lemma 4.3. *Assume H2.*

(i) *For any $\ell, r, s \geq 0$, any initial distributions χ, χ' on \mathbb{X} and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$*

$$\sup_{\theta \in \Theta} \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_{\theta,s}^{\chi',s-r-\ell}(\mathbf{y}) \right| \leq \frac{2}{1-\rho} \rho^r.$$

(ii) *For any $\theta \in \Theta$, there exists a function $\mathbf{y} \mapsto \delta_{\theta}(\mathbf{y})$ such that for any initial distribution χ , any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and any $r, s \geq 0$,*

$$\sup_{\theta \in \Theta} \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_{\theta}(\vartheta^s \circ \mathbf{y}) \right| \leq \frac{2}{1-\rho} \rho^r.$$

Proof. *Proof of (i).* Let $s \geq 0$ and r and r' be such that $r' > r$. By (29) and (28), we have $|\delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y})| = |\log \alpha - \log \beta|$ where

$$\begin{aligned} \alpha &\stackrel{\text{def}}{=} \frac{\int \chi(\mathrm{d}x_{s-r}) \prod_{i=s-r+1}^{s+1} m_{\theta}(x_{i-1}, x_i) g_{\theta}(x_i, y_i) \lambda(\mathrm{d}x_i)}{\int \chi(\mathrm{d}x_{s-r}) \prod_{i=s-r+1}^s m_{\theta}(x_{i-1}, x_i) g_{\theta}(x_i, y_i) \lambda(\mathrm{d}x_i)}, \\ \beta &\stackrel{\text{def}}{=} \frac{\int \chi'(\mathrm{d}x_{s-r'}) \prod_{i=s-r'+1}^{s+1} m_{\theta}(x_{i-1}, x_i) g_{\theta}(x_i, y_i) \lambda(\mathrm{d}x_i)}{\int \chi'(\mathrm{d}x_{s-r'}) \prod_{i=s-r'+1}^s m_{\theta}(x_{i-1}, x_i) g_{\theta}(x_i, y_i) \lambda(\mathrm{d}x_i)}. \end{aligned} \quad (31)$$

We prove that

$$\alpha \wedge \beta \geq \sigma_- \int g_\theta(x_{s+1}, y_{s+1}) \lambda(dx_{s+1}), \quad (32)$$

$$|\alpha - \beta| \leq 2\rho^r \sigma_+ \int g_\theta(x_{s+1}, y_{s+1}) \lambda(dx_{s+1}), \quad (33)$$

and the proof is concluded since $|\log \alpha - \log \beta| \leq |\alpha - \beta|/(\alpha \wedge \beta)$.

The minorization on α and β is a consequence of H2 upon noting that α and β are of the form $\int \mu(dx_s) m_\theta(x_s, x_{s+1}) g_\theta(x_{s+1}, y_{s+1}) \lambda(dx_{s+1})$ for some probability measure μ . The upper bound on $|\alpha - \beta|$ is a consequence of Lemma 4.2(i) applied with

$$\tilde{\chi}(dx_{s-r}) \leftarrow \int_{\mathbb{X}^{r'-r}} \chi'(dx_{s-r'}) \left\{ \prod_{i=s-r'}^{s-r-1} g_\theta(x_i, y_i) m_\theta(x_i, x_{i+1}) \right\} \lambda(dx_{s-r'+1:s-r})$$

and $h(u) \leftarrow \int g_\theta(x_{s+1}, y_{s+1}) m_\theta(u, x_{s+1}) \lambda(dx_{s+1})$.

Proof of (ii). By (i), for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, the sequence $\{\delta_{\theta,0}^{\chi,-r}(\mathbf{y})\}_{r \geq 0}$ is a Cauchy sequence uniformly in θ : there exists a limit denoted by $\delta_\theta(\mathbf{y})$ - which does not depend upon χ - such that

$$\lim_{r \rightarrow +\infty} \sup_{\theta \in \Theta} \left| \delta_{\theta,0}^{\chi,-r}(\mathbf{y}) - \delta_\theta(\mathbf{y}) \right| = 0. \quad (34)$$

We write for $r \leq r'$

$$\left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \leq \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_{\theta,s}^{\chi,s-r'}(\mathbf{y}) \right| + \left| \delta_{\theta,s}^{\chi,s-r'}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right|.$$

Observe that by definition, $\delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) = \delta_{\theta,0}^{\chi,-r}(\vartheta^s \circ \mathbf{y})$. This property, combined with Lemma 4.3(i), yield

$$\sup_{\theta \in \Theta} \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \leq \frac{2}{1-\rho} \rho^r + \sup_{\theta \in \Theta} \left| \delta_{\theta,0}^{\chi,-r'}(\vartheta^s \circ \mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right|.$$

When $r' \rightarrow +\infty$, the second term in the rhs tends to zero by (34) - for fixed \mathbf{y}, s and χ -. This concludes the proof. \square

Lemma 4.4. *Assume H2. For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and $s \geq 0$,*

$$\sup_{r \geq 0} \sup_{\theta \in \Theta} \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) \right| \leq |\log \sigma_+ b_+(\mathbf{y}_{s+1})| + |\log \sigma_- b_-(\mathbf{y}_{s+1})|,$$

and, for any $r \geq 0$,

$$\sup_{\theta \in \Theta} |\delta_\theta(\mathbf{y})| \leq \frac{2}{(1-\rho)} \rho^r + |\log \sigma_+ b_+(\mathbf{y}_1)| + |\log \sigma_- b_-(\mathbf{y}_1)|,$$

where b_+ and b_- are defined by (1).

Proof. For any $0 < m \leq A/B \leq M$, $|\log(A/B)| \leq |\log M| + |\log m|$. Note that by definition, $\delta_{\theta,s}^{X,0}(\mathbf{y})$ is of the form $\log(A/B)$ and under H4(c), $\sigma_- b_-(y_{s+1}) \leq A/B \leq \sigma_+ b_+(y_{s+1})$. The second upper bound is a consequence of Lemma 4.3(ii). \square

Lemma 4.5. *Assume H1, H2, H4(a) and H4(c). Then, $\theta \mapsto \mathbb{E}_*[\delta_\theta(\mathbf{Y})]$ is continuous on Θ . If in addition Θ is compact,*

$$\lim_{\eta \rightarrow 0} \mathbb{E}_* \left[\sup_{\{\theta, \theta' \in \Theta; |\theta - \theta'| < \eta\}} |\delta_\theta(\mathbf{Y}) - \delta_{\theta'}(\mathbf{Y})| \right] = 0 \quad \mathbb{P}_* - \text{a.s.} \quad (35)$$

Proof. By the dominated convergence theorem, Lemma 4.4 and H4(c), $\theta \mapsto \mathbb{E}_*[\delta_\theta(\mathbf{Y})]$ is continuous if $\theta \mapsto \delta_\theta(\mathbf{y})$ is continuous for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$. By Lemma 4.3(ii), $\lim_{r \rightarrow +\infty} \sup_{\theta \in \Theta} |\delta_{\theta,0}^{X,-r}(\mathbf{y}) - \delta_\theta(\mathbf{y})| = 0$. Therefore, $\theta \mapsto \delta_\theta(\mathbf{y})$ is continuous provided for any $r \geq 0$, $\theta \mapsto \delta_{\theta,0}^{X,-r}(\mathbf{y})$ is continuous (for fixed \mathbf{y} and χ). By definition of $\delta_{\theta,0}^{X,-r}(\mathbf{y})$, see (28), it is sufficient to prove that $\theta \mapsto \ell_{\theta,s}^{X,-r}(\mathbf{y})$ is continuous for $s \in \{0, 1\}$. By definition of $\ell_{\theta,s}^{X,-r}(\mathbf{y})$, see (29),

$$\ell_{\theta,s}^{X,-r}(\mathbf{y}) = \log \int \chi(dx_{-r}) \prod_{i=-r+1}^s m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i) \lambda(dx_i).$$

Under H1(a), $\theta \mapsto \prod_{i=-r+1}^s m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i)$ is continuous on Θ , for any $x_{-r:s}$ and \mathbf{y} . In addition, under H1, for any $\theta \in \Theta$,

$$\begin{aligned} & \left| \prod_{i=-r+1}^s m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, \mathbf{y}_{i+1}) \right| \\ &= \exp \left((s+r)\phi(\theta) + \left\langle \psi(\theta), \sum_{i=-r+1}^s S(x_i, x_{i+1}, \mathbf{y}_{i+1}) \right\rangle \right). \end{aligned}$$

Let \mathcal{K} be a compact subset of Θ . Since by H1 ϕ and ψ are continuous, there exist constants C_1 and C_2 such that,

$$\begin{aligned} & \sup_{\theta \in \mathcal{K}} \left| \prod_{i=-r+1}^s m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, \mathbf{y}_{i+1}) \right| \\ & \leq C_1 \exp \left(C_2 \sum_{i=-r+1}^s \sup_{x, x'} |S(x, x', \mathbf{y}_{i+1})| \right). \end{aligned}$$

Since the measure $\chi(dx_{-r}) \prod_{i=-r+1}^s \lambda(dx_i)$ is finite, the dominated convergence theorem now implies that $\ell_{\theta,s}^{X,-r}(\mathbf{y})$ is continuous on Θ .

For the proof of (35), let us apply the dominated convergence theorem again. Since Θ is compact, for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, $\theta \mapsto \delta_\theta(\mathbf{y})$ is uniformly continuous and

$\lim_{\eta \rightarrow 0} \sup_{|\theta - \theta'| < \eta} |\delta_\theta(\mathbf{y}) - \delta_{\theta'}(\mathbf{y})| = 0$. In addition, we have by Lemma 4.4

$$\begin{aligned} & \sup_{\{\theta, \theta' \in \Theta; |\theta - \theta'| < \eta\}} |\delta_\theta(\mathbf{y}) - \delta_{\theta'}(\mathbf{y})| \\ & \leq 2 \sup_{\theta \in \Theta} |\delta_\theta(\mathbf{y})| \leq \frac{4}{(1 - \rho)} + 2 \{|\log \sigma_+ b_+(y_1)| + |\log \sigma_- b_-(y_1)|\} . \end{aligned}$$

Under H4(a) and H4(c), this upper bound is \mathbb{P}_* -integrable. This concludes the proof. \square

Theorem 4.6. *Assume H1, H2, H4. Define the function $c_* : \Theta \rightarrow \mathbb{R}$ by $c_*(\theta) \stackrel{\text{def}}{=} \mathbb{E}_*[\delta_\theta(\mathbf{Y})]$, where $\delta_\theta(\mathbf{y})$ is defined in Lemma 4.4.*

(i) *The function $\theta \mapsto c_*(\theta)$ is continuous on Θ .*

(ii) *For any initial distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$*

$$\left| \frac{1}{T} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) - c_*(\theta) \right| \xrightarrow{T \rightarrow +\infty} 0 \quad \mathbb{P}_* - \text{a.s.} \quad (36)$$

where $\ell_{\theta, T}^{\chi, 0}(\mathbf{Y})$ is defined in (29).

(iii) *If in addition Θ is compact, for any initial distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$*

$$\sup_{\theta \in \Theta} \left| \frac{1}{T} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) - c_*(\theta) \right| \xrightarrow{T \rightarrow +\infty} 0 \quad \mathbb{P}_* - \text{a.s.} \quad (37)$$

Proof. (i) is proved in Lemma 4.5.

(ii) By (30), for any $T > 0$, we have, for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$:

$$\begin{aligned} \frac{1}{T} \ell_{\theta, T}^{\chi, 0}(\mathbf{y}) &= \frac{1}{T} \sum_{s=0}^{T-1} \delta_{\theta, s}^{\chi, 0}(\mathbf{y}) \\ &= \frac{1}{T} \sum_{s=0}^{T-1} \left(\delta_{\theta, s}^{\chi, 0}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right) + \frac{1}{T} \sum_{s=0}^{T-1} \delta_\theta(\vartheta^s \circ \mathbf{y}) . \end{aligned}$$

By Lemma 4.3(ii), for any $0 \leq s \leq T-1$, $\left| \delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_\theta(\vartheta^s \circ \mathbf{Y}) \right| \leq 2 \frac{\rho^s}{1 - \rho}$. Since $\rho \in (0, 1)$,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{s=0}^{T-1} \left(\delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_\theta(\vartheta^s \circ \mathbf{Y}) \right) = 0 \quad \mathbb{P}_* - \text{a.s.}$$

By Lemma 4.4

$$\mathbb{E}_*[\delta_\theta(\mathbf{Y})] \leq \frac{2}{(1 - \rho)} + \mathbb{E}_* [|\log \sigma_+ b_+(\mathbf{Y}_1)| + |\log \sigma_- b_-(\mathbf{Y}_1)|] ,$$

and the rhs is finite under assumption H4(c). By H4(a-b), the ergodic theorem, see [1, Theorem 24.1, p.314], concludes the proof.

(iii) Since Θ is compact, (37) holds if for any $\varepsilon > 0$, any $\theta' \in \Theta$, there exists $\eta > 0$ such that

$$\lim_{T \rightarrow +\infty} \sup_{\{\theta; |\theta - \theta'| < \eta\} \cap \Theta} \left| T^{-1} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) - T^{-1} \ell_{\theta', T}^{\chi, 0}(\mathbf{Y}) \right| \leq \varepsilon, \quad \mathbb{P}_\star - \text{a.s.} \quad (38)$$

Let $\varepsilon > 0$ and $\theta' \in \Theta$. Choose $\eta > 0$ such that

$$\mathbb{E}_\star \left[\sup_{\{\theta \in \Theta; |\theta - \theta'| < \eta\}} |\delta_\theta(\mathbf{Y}) - \delta_{\theta'}(\mathbf{Y})| \right] \leq \varepsilon; \quad (39)$$

such an η exists by Lemma 4.5. By (30), we have, for any $\theta \in \Theta$ such that $|\theta - \theta'| < \eta$

$$\left| \frac{1}{T} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) - \frac{1}{T} \ell_{\theta', T}^{\chi, 0}(\mathbf{Y}) \right| \leq \frac{1}{T} \sum_{s=0}^{T-1} \left| \delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_{\theta', s}^{\chi, 0}(\mathbf{Y}) \right|. \quad (40)$$

In addition, by Lemma 4.3(ii)

$$\begin{aligned} & \sum_{s=0}^{T-1} \left| \delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_{\theta', s}^{\chi, 0}(\mathbf{Y}) \right| \\ & \leq 2 \sum_{s=0}^{T-1} \sup_{\theta \in \Theta} \left| \delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_\theta(\vartheta^s \circ \mathbf{Y}) \right| + \sum_{s=0}^{T-1} \left| \delta_\theta(\vartheta^s \circ \mathbf{Y}) - \delta_{\theta'}(\vartheta^s \circ \mathbf{Y}) \right| \\ & \leq \frac{4}{(1-\rho)^2} + \sum_{s=0}^{T-1} \Xi(\vartheta^s \circ \mathbf{Y}) \end{aligned}$$

where $\Xi(\mathbf{y}) \stackrel{\text{def}}{=} \sup_{\{\theta \in \Theta; |\theta - \theta'| < \eta\}} |\delta_\theta(\mathbf{y}) - \delta_{\theta'}(\mathbf{y})|$. This implies that

$$\lim_{T \rightarrow +\infty} \sup_{\{\theta \in \Theta; |\theta - \theta'| < \eta\}} \frac{1}{T} \sum_{s=0}^{T-1} \left| \delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_{\theta', s}^{\chi, 0}(\mathbf{Y}) \right| \leq \lim_{T \rightarrow +\infty} \frac{1}{T} \sum_{s=0}^{T-1} \Xi(\vartheta^s \circ \mathbf{Y}).$$

Under H4(a-b), the ergodic theorem implies that the rhs converges \mathbb{P}_\star a.s to $\mathbb{E}_\star[\Xi(\mathbf{Y})]$, see [1, p.314]. Then, using again (39),

$$\lim_{T \rightarrow +\infty} \sup_{\{\theta \in \Theta; |\theta - \theta'| < \eta\}} \frac{1}{T} \sum_{s=0}^{T-1} \left| \delta_{\theta, s}^{\chi, 0}(\mathbf{Y}) - \delta_{\theta', s}^{\chi, 0}(\mathbf{Y}) \right| \leq \varepsilon \quad \mathbb{P}_\star - \text{a.s.}$$

Then, (38) holds and this concludes the proof. \square

4.3 Limit of the normalized score

This section is devoted to the proof of the \mathbb{P}_\star -a.s convergence of the normalized score $T^{-1}\nabla_\theta \ell_{\theta,T}^{\chi,0}(\mathbf{Y})$ to $\nabla_\theta c_\star(\theta)$. This result is established under additional assumptions on the model.

- S1** (a) For any $y \in \mathbb{Y}$ and for all $(x, x') \in \mathbb{X}^2$, $\theta \mapsto g_\theta(x, y)$ and $\theta \mapsto m_\theta(x, x')$ are continuously differentiable on Θ .
- (b) We assume that $\mathbb{E}_\star[\phi(\mathbf{Y}_0)] < +\infty$ where

$$\phi(y) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \sup_{(x, x') \in \mathbb{X}^2} |\nabla_\theta \log m_\theta(x, x') + \nabla_\theta \log g_\theta(x', y)|. \quad (41)$$

Lemma 4.7. *Assume S1. For any initial distribution χ , any integers $s, r \geq 0$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$, the function $\theta \mapsto \ell_{\theta,s}^{\chi,s-r}(\mathbf{y})$ is continuously differentiable on Θ and*

$$\nabla_\theta \ell_{\theta,s}^{\chi,s-r}(\mathbf{y}) = \sum_{u=s-r}^s \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_\theta, \mathbf{y}),$$

where Υ_θ is the function defined on $\mathbb{X}^2 \times \mathbb{Y}$ by

$$\Upsilon_\theta : (x, x', y) \mapsto \nabla_\theta \log \{m_\theta(x, x')g_\theta(x', y)\}.$$

Proof. Under S1, the dominated convergence theorem implies that the function $\theta \mapsto \ell_{\theta,s}^{\chi,s-r}(\mathbf{y})$ is continuously differentiable and its derivative is obtained by permutation of the gradient and integral operators. \square

Lemma 4.8. *Assume H2 and S1.*

- (i) *There exists a function $\xi : \mathbb{Y}^{\mathbb{Z}} \rightarrow \mathbb{R}_+$ such that for any $s \geq 0$ and any $r, r' \geq s$, any initial distribution χ, χ' on \mathbb{X} and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$,*

$$\sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \right| \leq \frac{16\rho^{-1/4}}{1-\rho} \rho^{(r' \wedge r)/4} \xi(\mathbf{y}),$$

where

$$\xi(\mathbf{y}) \stackrel{\text{def}}{=} \sum_{u \in \mathbb{Z}} \phi(\mathbf{y}_u) \rho^{|u|/4}. \quad (42)$$

- (ii) *For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ satisfying $\xi(\mathbf{y}) < +\infty$, the function $\theta \mapsto \delta_\theta(\mathbf{y})$ given by Lemma 4.3(ii) is continuously differentiable on Θ ; and, for any $\theta \in \Theta$, any initial distribution χ and any integers $r \geq s \geq 0$,*

$$\sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_\theta(\mathbf{y} \circ \vartheta^s) \right| \leq \frac{16\rho^{-1/4}}{1-\rho} \rho^{r/4} \xi(\mathbf{y}).$$

Proof. (i) By definition of $\delta_{\theta,s}^{\chi,s-r}(\mathbf{y})$, see (28) and Lemma 4.7,

$$\begin{aligned}
& \nabla_{\theta} \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_{\theta} \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \\
&= \nabla_{\theta} \ell_{\theta,s+1}^{\chi,s-r}(\mathbf{y}) - \nabla_{\theta} \ell_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_{\theta} \ell_{\theta,s+1}^{\chi',s-r'}(\mathbf{y}) + \nabla_{\theta} \ell_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \\
&= \sum_{u=s-r}^s \left(\Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right) \\
&\quad - \sum_{u=s-r'}^s \left(\Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) \right) \\
&\quad + \Phi_{\theta,s+1,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,s+1,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}).
\end{aligned}$$

We can assume without loss of generality that $r' \leq r$ so that

$$\begin{aligned}
& \nabla_{\theta} \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_{\theta} \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \\
&= \sum_{u=s-r}^{s-r'-1} \left\{ \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right\} + \Phi_{\theta,s+1,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,s+1,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) \\
&+ \sum_{u=s-r'}^s \left\{ \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) + \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) \right\}.
\end{aligned}$$

Under H2 and S1, Remark A.2 can be applied and for any $s-r \leq u \leq s-r'-1$,

$$\left| \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right| \leq 2\rho^{s-u} \phi(\mathbf{y}_u),$$

where $\phi_u(\mathbf{y})$ is defined in (41). Similarly, by Remark A.2

$$\left| \Phi_{\theta,s+1,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,s+1,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) \right| \leq 2\rho^{r'+1} \phi(\mathbf{y}_{s+1}).$$

For any $s-r' \leq u \leq s$, by Remark A.2,

$$\begin{aligned}
& \left| \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) + \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right| \\
&\leq \left| \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) \right| + \left| \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right| \\
&\leq 4\rho^{u+r'-s} \phi(\mathbf{y}_u)
\end{aligned}$$

and by Remark A.2,

$$\begin{aligned}
& \left| \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) + \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right| \\
&\leq \left| \Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_{\theta}, \mathbf{y}) \right| + \left| \Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_{\theta}, \mathbf{y}) \right| \\
&\leq 4\rho^{s-u} \phi(\mathbf{y}_u).
\end{aligned}$$

Hence,

$$\left| \nabla_{\theta} \delta_{\theta, s}^{\chi, s-r}(\mathbf{y}) - \nabla_{\theta} \delta_{\theta, s}^{\chi', s-r'}(\mathbf{y}) \right| \leq 2 \sum_{u=s-r}^{s-r'-1} \rho^{s-u} \phi(\mathbf{y}_u) + 4 \sum_{u=s-r'}^{s+1} \left(\rho^{u+r'-s} \wedge \rho^{s-u} \right) \phi(\mathbf{y}_u).$$

Furthermore,

$$\begin{aligned} & \sum_{u=s-r'}^{s+1} \phi(\mathbf{y}_u) \left(\rho^{u+r'-s} \wedge \rho^{s-u} \right) \\ & \leq \sum_{s-r' \leq u \leq \lfloor s-r'/2 \rfloor} \rho^{s-u} \phi(\mathbf{y}_u) + \sum_{u \geq \lfloor s-r'/2 \rfloor} \rho^{u+r'-s} \phi(\mathbf{y}_u) \\ & \leq \rho^{r'/2} \sum_{u \in \mathbb{Z}} \phi(\mathbf{y}_u) \rho^{|u|/4} \dots \\ & \quad \times \left(\sum_{u \leq \lfloor s-r'/2 \rfloor} \rho^{s-u-r'/2-|u|/4} + \sum_{\lfloor s-r'/2 \rfloor + 1 \leq u \leq s+1} \rho^{u+r'/2-s-|u|/4} \right) \\ & \leq 2 \frac{\rho^{(r'-1)/4}}{1-\rho} \sum_{u \in \mathbb{Z}} \phi(\mathbf{y}_u) \rho^{|u|/4}, \end{aligned}$$

where we used that $\sup_{s-r' \leq u \leq \lfloor s-r'/2 \rfloor} |u| \leq r'$ and $\sup_{\lfloor s-r'/2 \rfloor + 1 \leq u \leq s+1} |u| \leq r' + 1$. Moreover, upon noting that $-u/2 + (s+1)/2 \leq s-u-r'/2$ when $u \leq s-r'-1$,

$$\begin{aligned} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u) \rho^{s-u} & \leq \rho^{r'/2} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u) \rho^{s-u-r'/2} \\ & \leq \rho^{r'/2} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u) \rho^{-u/2+(s+1)/2} \\ & \leq \rho^{r'/2} \rho^{(s+1)/2} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u) \rho^{|u|/2}, \end{aligned}$$

where we used that $s-r'-1 \leq 0$ in the last inequality.

Hence,

$$\sup_{\theta \in \Theta} \left| \nabla_{\theta} \delta_{\theta, s}^{\chi, s-r}(\mathbf{y}) - \nabla_{\theta} \delta_{\theta, s}^{\chi', s-r'}(\mathbf{y}) \right| \leq \frac{16}{1-\rho} \rho^{(r'-1)/4} \sum_{u \in \mathbb{Z}} \phi(\mathbf{y}_u) \rho^{|u|/4}. \quad (43)$$

(ii) Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\xi(\mathbf{y}) < +\infty$. Then for any $u \in \mathbb{Z}$, $\phi(\mathbf{y}_u) < +\infty$. By Lemma 4.7 and Eq. (28), the functions $\{\theta \mapsto \delta_{\theta, 0}^{\chi, -r}(\mathbf{y})\}_{r \geq 0}$ are C^1 functions on Θ . By (i), there exists a function $\theta \mapsto \tilde{\delta}_{\theta}(\mathbf{y})$ such that

$$\lim_{r \rightarrow +\infty} \sup_{\theta \in \Theta} \left| \nabla_{\theta} \delta_{\theta, 0}^{\chi, -r}(\mathbf{y}) - \tilde{\delta}_{\theta}(\mathbf{y}) \right| = 0.$$

Furthermore, by Lemma 4.3,

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

Then, $\theta \mapsto \delta_{\theta}(\mathbf{y})$ is C^1 on Θ and for any $\theta \in \Theta$, $\tilde{\delta}_{\theta}(\mathbf{y}) = \nabla_{\theta} \delta_{\theta}(\mathbf{y})$.

We thus proved that for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\xi(\mathbf{y}) < +\infty$ and for any initial distribution χ ,

$$\lim_{r \rightarrow +\infty} \sup_{\theta \in \Theta} \left| \nabla_{\theta} \delta_{\theta,0}^{\chi, -r}(\mathbf{y}) - \nabla_{\theta} \delta_{\theta}(\mathbf{y}) \right| = 0. \quad (44)$$

Observe that by definition, $\nabla_{\theta} \delta_{\theta,s}^{\chi, s-r}(\mathbf{y}) = \nabla_{\theta} \delta_{\theta,0}^{\chi, -r}(\vartheta^s \circ \mathbf{y})$. This property, combined with Lemma 4.8(i), yields

$$\begin{aligned} & \sup_{\theta \in \Theta} \left| \nabla_{\theta} \delta_{\theta,s}^{\chi, s-r}(\mathbf{y}) - \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{y}) \right| \\ & \leq \frac{16\rho^{-1/4}}{1-\rho} \rho^{r/4} \xi(\mathbf{y}) + \sup_{\theta \in \Theta} \left| \nabla_{\theta} \delta_{\theta,0}^{\chi, -r'}(\vartheta^s \circ \mathbf{y}) - \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{y}) \right|. \end{aligned}$$

Since $\xi(\vartheta^s \circ \mathbf{y}) < +\infty$, when $r' \rightarrow +\infty$, the second term tends to zero by (44) - for fixed \mathbf{y} , s and χ -. This concludes the proof. \square

Lemma 4.9. (i) *Assume S1. For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$, for any integers $r, s \geq 0$,*

$$\sup_{\theta \in \Theta} \left| \nabla_{\theta} \delta_{\theta,s}^{\chi, s-r}(\mathbf{y}) \right| \leq 2 \sum_{u=s-r}^{s+1} \phi(\mathbf{y}_u).$$

(ii) *Assume H2 and S1. Then, for any $y \in \mathbb{Y}^{\mathbb{Z}}$ such that $\xi(\mathbf{y}) < +\infty$ and for any $r \geq 0$,*

$$\sup_{\theta \in \Theta} |\nabla_{\theta} \delta_{\theta}(\mathbf{y})| \leq 2 \sum_{u=-r}^1 \phi(\mathbf{y}_u) + \frac{16\rho^{-1/4}}{1-\rho} \xi(\mathbf{y}) \rho^{r/4},$$

where $\xi(\mathbf{y})$ is defined in Lemma 4.8.

Proof. (i) By (28) and Lemma 4.7,

$$\begin{aligned} & \left| \nabla_{\theta} \delta_{\theta,s}^{\chi, s-r}(\mathbf{y}) \right| = \left| \nabla_{\theta} \ell_{\theta, s+1}^{\chi, s-r}(\mathbf{y}) - \nabla_{\theta} \ell_{\theta, s}^{\chi, s-r}(\mathbf{y}) \right| \\ & \leq 2 \sum_{u=s-r}^{s+1} \left| \frac{\int \chi(dx_{s-r}) L_{\theta, s-r:u-1}(x_{s-r}, dx_u) \nabla_{\theta} \log [m_{\theta}(x_{u-1}, x_u) g_{\theta}(x_u, \mathbf{y}_u)] L_{\theta, u:s-1}(x_u, \mathbb{X})}{\int \chi(dx_{s-r}) L_{\theta, s-r:s-1}(x_{s-r}, \mathbb{X})} \right|. \end{aligned}$$

The proof is concluded upon noting that for any $s-r \leq u \leq s+1$,

$$\left| \frac{\int \chi(dx_{s-r}) g_{\theta}(x_{s-r}, y_{s-r}) L_{\theta, s-r:u-1}(x_{s-r}, dx_u) \nabla_{\theta} \log g_{\theta}(x_u, \mathbf{y}_u) L_{\theta, u:s-1}(x_u, \mathbb{X})}{\int \chi(dx_{s-r}) g_{\theta}(x_{s-r}, \mathbf{y}_{s-r}) L_{\theta, s-r:s-1}(x_{s-r}, \mathbb{X})} \right|$$

is upper bounded by $\phi(\mathbf{y}_u)$.

(ii) is a consequence of Lemma 4.8(ii) and Lemma 4.9(i). \square

Theorem 4.10. *Assume H2, H4(a-b) and S1.*

- (i) *For any $T \geq 0$ and any distribution χ on \mathbb{X} , the functions $\theta \mapsto \ell_{\theta,T}^{\chi,0}(\mathbf{Y})$ and $\theta \mapsto c_*(\theta)$ are continuously differentiable $\mathbb{P}_* - a.s.$*
- (ii) *For any initial distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$,*

$$\frac{1}{T} \nabla_{\theta} \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) \xrightarrow{T \rightarrow +\infty} \nabla_{\theta} c_*(\theta) \quad \mathbb{P}_* - a.s. \quad (45)$$

Proof. By (30) and Lemma 4.7, for any \mathbf{y} such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$, $\ell_{\theta,T}^{\chi,0}(\mathbf{y})$ and $\delta_{\theta,s}^{\chi,0}(\mathbf{y})$ are continuously differentiable and (30) implies

$$\nabla_{\theta} \ell_{\theta,T}^{\chi,0}(\mathbf{y}) = \sum_{s=0}^{T-1} \nabla_{\theta} \delta_{\theta,s}^{\chi,0}(\mathbf{y}).$$

This decomposition leads to

$$\frac{1}{T} \nabla_{\theta} \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) = \frac{1}{T} \sum_{s=0}^{T-1} \left(\nabla_{\theta} \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{Y}) \right) + \frac{1}{T} \sum_{s=0}^{T-1} \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{Y}). \quad (46)$$

Consider the first term of the rhs of (46). Since \mathbf{Y} is a stationary process, assumption S1(b) implies that $\mathbb{E}_* [\xi(\mathbf{Y})] < +\infty$, where ξ is defined by (42). Then, $\xi(\mathbf{Y}) < +\infty \quad \mathbb{P}_* - a.s.$ and by Lemma 4.8(ii), for any $0 \leq s \leq T-1$,

$$\left| \nabla_{\theta} \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{Y}) \right| \leq \xi(\mathbf{Y}) \frac{16\rho^{-1/4}}{1-\rho} \rho^{s/4}.$$

Therefore

$$\frac{1}{T} \sum_{s=0}^{T-1} \left| \nabla_{\theta} \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{Y}) \right| \leq \frac{1}{T} \xi(\mathbf{Y}) \frac{16\rho^{-1/4}}{1-\rho} \frac{1}{1-\rho^{1/4}},$$

and

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{s=0}^{T-1} \left(\nabla_{\theta} \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{Y}) \right) = 0 \quad \mathbb{P}_* - a.s.$$

Finally, consider the second term of the rhs of (46). By Lemma 4.9 (applied with $r = 1$), $\mathbb{E}_* [|\nabla_{\theta} \delta_{\theta}(\mathbf{Y})|] < +\infty$. Under H4(b), the ergodic theorem (see [1, Theorem 24.1, p.314]) states that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{s=0}^{T-1} \nabla_{\theta} \delta_{\theta}(\vartheta^s \circ \mathbf{Y}) = \mathbb{E}_* [\nabla_{\theta} \delta_{\theta}(\mathbf{Y})] \quad \mathbb{P}_* - a.s.$$

Then, by (46) and the above discussion,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \nabla_{\theta} \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) = \mathbb{E}_* [\nabla_{\theta} \delta_{\theta}(\mathbf{Y})] \quad \mathbb{P}_* - a.s.$$

By Lemma 4.9, applied with $r = 0$,

$$\sup_{\theta \in \Theta} |\nabla_{\theta} \delta_{\theta}(\mathbf{Y})| \leq 2 [\phi(Y_0) + \phi(Y_1)] + \xi(\mathbf{Y}) \rho^{1/2},$$

and the rhs is integrable under the stated assumptions. Therefore, by the dominated convergence theorem, $\mathbb{E}_{\star} [\nabla_{\theta} \delta_{\theta}(\mathbf{Y})] = \nabla_{\theta} \mathbb{E}_{\star} [\delta_{\theta}(\mathbf{Y})] = \nabla_{\theta} c_{\star}(\theta)$. This concludes the proof. \square

4.4 Contrast and the limit set \mathcal{L}

Theorem 4.11. *Assume H2, H3-(1), H4(a-b) and S1. Then, $\theta \in \mathcal{L}$ if and only if $\nabla_{\theta} c_{\star}(\theta) = 0$ where $c_{\star}(\theta) = \lim_{T \rightarrow +\infty} T^{-1} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y})$ \mathbb{P}_{\star} -a.s for any initial distribution χ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$.*

Proof. For any initial distribution χ , all $\theta \in \Theta$ and all $T > 0$,

$$\frac{1}{T} \nabla_{\theta} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) = \frac{1}{T} \int \nabla_{\theta} \log p_{\theta}^{\chi}(x_{0:T}, \mathbf{Y}_{1:T}) \frac{p_{\theta}^{\chi}(x_{0:T}, \mathbf{Y}_{1:T})}{\int p_{\theta}^{\chi}(z_{0:T}, \mathbf{Y}_{1:T}) \lambda(dz_{0:T})} \lambda(dx_{0:T}),$$

where p_{θ}^{χ} is defined in (22) in [6]. Under Assumption H1(a)

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

and then, by definition of $\Phi_{\theta, t, T}^{\chi, 0}(S, \mathbf{Y})$ (see (9)),

$$\frac{1}{T} \nabla_{\theta} \ell_{\theta, T}^{\chi, 0}(\mathbf{Y}) = \nabla_{\theta} \phi(\theta) + \nabla_{\theta} \psi^T(\theta) \left\{ \frac{1}{T} \sum_{t=1}^T \Phi_{\theta, t, T}^{\chi, 0}(S, \mathbf{Y}) \right\}. \quad (47)$$

Under the stated assumptions, Theorem 4.1 in [6] and Theorem 4.10 can be applied. Therefore, (47) becomes, as $T \rightarrow +\infty$,

$$\nabla_{\theta} c_{\star}(\theta) = \nabla_{\theta} \phi(\theta) + \nabla_{\theta} \psi^T(\theta) \{ \mathbb{E}_{\star} [\mathbb{E}_{\theta} [S(X_{-1}, X_0, \mathbf{Y}_0) | \mathbf{Y}]] \}.$$

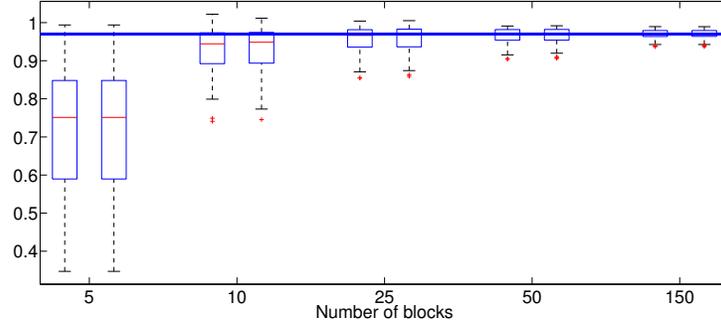
The proof follows upon noting that by definition of $\bar{\theta}$, the unique solution to the equation $\nabla_{\theta} \phi(\tau) + \nabla_{\theta} \psi^T(\tau) \{ \mathbb{E}_{\star} [\mathbb{E}_{\theta} [S(X_{-1}, X_0, \mathbf{Y}_0) | \mathbf{Y}]] \} = 0$ is $\tau = \bar{\theta}$. \square

5 Additional experiments

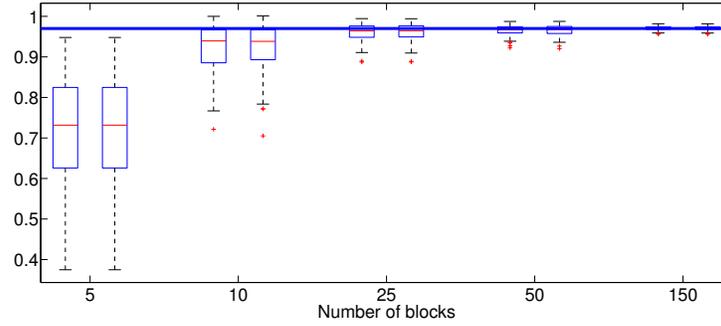
In this section, we provide additional plots for the applications studied in [6, section 3].

5.1 Linear Gaussian model

Figure 1 illustrates the fact that the convergence properties of the BOEM do not depend on the initial distribution χ used in each block. Data are sampled using $\phi = 0.97$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 2$. Figure 1 displays the estimation of ϕ by the averaged BOEM algorithm with $\tau_n \sim n$ and $\tau_n \sim n^{1.5}$, 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. The estimation error is similar for both initialization schemes, even when ϕ is close to 1 and for any choice of $\{\tau_n\}_{n \geq 1}$.



(a) $\tau_n \sim n$



(b) $\tau_n \sim n^{1.5}$

Figure 1: Estimation of ϕ after 5, 10, 25, 50 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.

5.2 Finite state-space HMM

In the numerical applications below, we give a supplementary graph to compare the convergence of the averaged BOEM with the convergence of the Polyak-Ruppert averaged RML procedure. The experiment is the same as the one in [6, section 3]. Figure 2 displays boxplots of the estimation of $m(1, 4)$ after different numbers of observations n ; the boxplots are over 100 independent Monte Carlo runs. The variance of the estimation given by the averaged BOEM is smaller than the one given by the averaged RML. Same conclusions can be drawn for the estimation of the other entries of the transition matrix m .

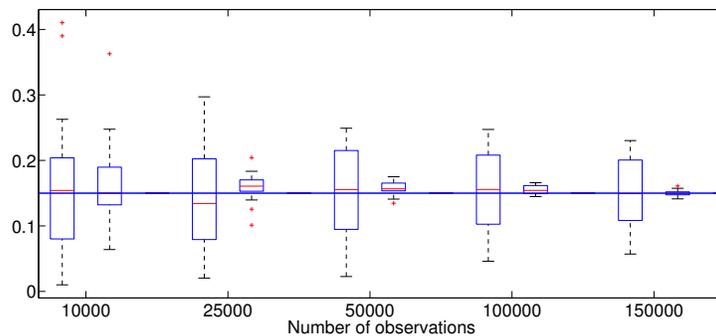


Figure 2: Estimation of $m(1, 4)$ using the averaged RML algorithm (left) and the averaged BOEM algorithm (right), based on $n = \{10k, 25k, 50k, 100k, 150k\}$ observations.

5.3 Stochastic volatility model

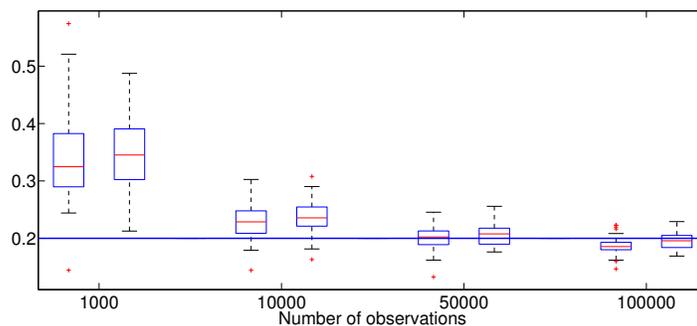
We report in Figure 3, the boxplots for the estimation of the parameters β^2 and σ^2 for the Polyak-Ruppert [8] averaged Online EM and the averaged BOEM. Both average versions are started after 20000 observations. Data are sampled using $\phi = 0.8$, $\sigma^2 = 0.2$ and $\beta^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma^2 = 0.6$ and $\beta^2 = 2$. This figure illustrates the similar behavior of both algorithms.

We now compare the two algorithms when the true value of ϕ is (in absolute value) closer to 1: we choose $\phi = 0.95$, β^2 and σ^2 being the same as in the previous experiment.

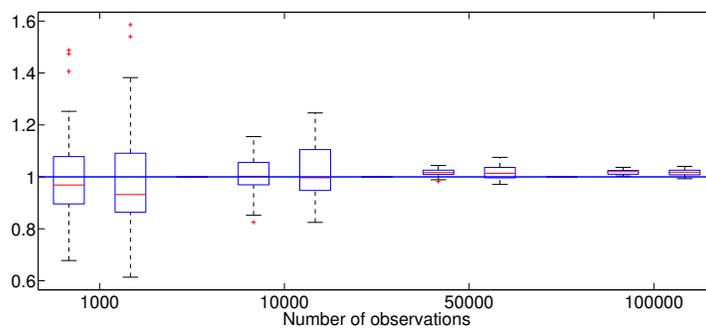
As illustrated on Figure 4, the same conclusions are drawn for greater values of ϕ .

References

- [1] P. Billingsley. *Probability and Measure*. Wiley, New York, 3rd edition, 1995.
- [2] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.



(a) Estimation of σ^2 .



(b) Estimation of β^2 .

Figure 3: Estimation of σ^2 and β^2 using the averaged online EM algorithm (left) and the averaged BOEM (right), after $n = \{1000, 10k, 50k, 100k\}$ observations. The true value of ϕ is 0.8.

- [3] J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.
- [4] R. Douc, E. Moulines, and J. Rosenthal. Quantitative bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 14(4):1643–1665, 2004.
- [5] G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259, 2003.
- [6] S. Le Corff and G. Fort. Online Expectation Maximization based algorithms for inference in Hidden Markov Models. Technical report, arXiv, 2011.
- [7] G. Pólya and G. Szegő. *Problems and Theorems in Analysis. Vol. II*. Springer, 1976.

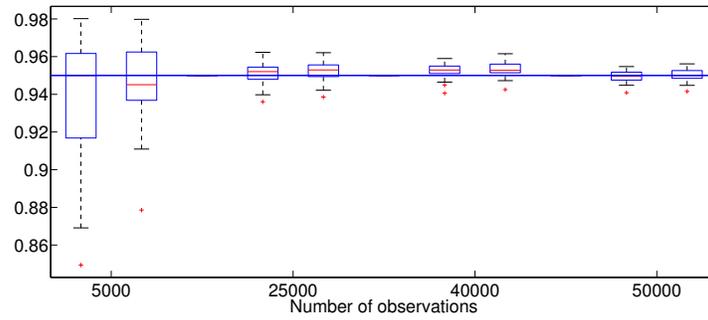


Figure 4: Estimation of ϕ using the averaged online EM algorithm (left) and the averaged BOEM algorithm (right), after $n = \{5k, 25k, 40k, 50k\}$ observations. The true value of ϕ is 0.95.

- [8] B. T. Polyak. A new method of stochastic approximation type. *Autom. Remote Control*, 51:98–107, 1990.
- [9] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer, 1990.