Stochastic Processes and their Applications 66 (1997) 79–96

stochastic
processes
and their
applications

On recursive estimation for hidden Markov models

Tobias Rydén *,[1]

Department of Statistics, University of California, Evans Hall, Berkeley, CA 94720, USA

Received September 1995; revised August 1996

Abstract

Hidden Markov models (HMMs) have during the last decade become a widespread tool for modelling sequences of dependent random variables. In this paper we consider a recursive estimator for HMMs based on the m-dimensional distribution of the process and show that this estimator converges to the set of stationary points of the corresponding Kullback–Leibler information. We also investigate averaging in this recursive scheme and show that conditional on convergence to the true parameter, and provided m is chosen large enough, the averaged estimator is close to optimal.

AMS classification: 62M09; 62L20

Keywords: Hidden Markov model; Incomplete data; Missing data; Recursive estimation; Stochastic approximation; Poisson equation

---

1. Introduction

A hidden Markov model (HMM) is a discrete-time stochastic process $\{(X_k, Y_k)\}$ such that (i) $\{X_k\}$ is a finite-state Markov chain, and (ii) given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables and the conditional distribution of Y_n depends on $\{X_k\}$ only through X_n. The Markov chain $\{X_k\}$ is sometimes called the *regime*. The name HMM is motivated by the assumption that $\{X_k\}$ is not observable, so that inference, etc. have to be based on $\{Y_k\}$ alone. HMMs have during the last decade become widespread for modelling sequences of weakly dependent random variables, with applications in areas like speech processing (Rabiner, 1989), biochemistry (Fredkin and Rice, 1992), and biology (Leroux and Puterman, 1992).

Inference for HMMs was first considered by Baum and Petrie who treated the case when $\{Y_k\}$ takes values in a finite set. In Baum and Petrie (1966), results on consistency and asymptotic normality of the maximum-likelihood estimate (MLE) are given, and the conditions for consistency are weakened in Petrie (1969). In the latter paper

---

* Present address: Department of Mathematical Statistics, Lund University, Box 118, S-221 00 Lund, Sweden. Fax: 46 4622 24 623; e-mail: tobias@maths.lth.se.

[1] Supported by the Swedish Natural Science Research Council (contract no. F-PD 10538-301), the Fulbright Commission, and Kungliga Fysiografiska Sällskapet i Lund.

PII S0304-4149(96)00114-7

the identifiability problem is also discussed, i.e. under what conditions there are no other parameters that induce the same law for $\{Y_k\}$ as the true parameter does, with the exception for permutations of states. For general HMMs, with Y_k, conditional on X_k, having density $f(\cdot;\theta_{X_k})$, Lindgren (1978) constructed consistent and asymptotically normal estimates of the θ-parameters, but he did not consider estimation of the transition probabilities. Later, Leroux (1992) proved consistency of the MLE for general HMMs under mild conditions, and local asymptotic normality has been proved by Bickel and Ritov (1996).

The topic of the present paper is recursive estimation for HMMs. Procedures of this kind have been considered by Holst and Lindgren (1991) and Krishnamurthy and Moore (1993). The difference between their approaches essentially concerns the scaling matrices that are used in the recursive procedure, and we comment further on this below. These papers are both simulation studies in the sense that no results on convergence, etc. are proved, but numerical results show that the algorithms often work out well in practice. In this paper we base the estimation on the m-dimensional density of $\{Y_k\}$ and give a recursive estimator which under suitable assumptions converges to the set of stationary points of the Kullback–Leibler information associated with this density. In Rydén (1994) a similar off-line estimator was shown to be consistent, asymptotically normal, and, provided m is large enough, almost efficient. Conditional on convergence to the true parameter, these asymptotic properties are shared by the recursive estimator of the present paper.

2. Notation and preliminaries

Let $\{X_k\}_{k=1}^{\infty}$ be a stationary Markov chain on $\{1,\dots,r\}$ with transition probability matrix $\{a_{ij}\}$. The transition probabilities will be parameterized by a parameter $\phi \in \Phi$, i.e. $a_{ij} = a_{ij}(\phi)$, where $\Phi \subseteq \mathbb{R}^q$ is the parameter space. The observable process $\{Y_k\}$ is assumed to take values in some space E and the conditional distributions of Y_k given X_k are all assumed to be dominated by some measure μ on E. Moreover, the corresponding conditional densities are assumed to belong to some parametric family $\{f(\cdot;\theta) : \theta \in \Theta\}$, and the parameter of this density is a function of X_k as well as of ϕ. Thus, the conditional density of Y_k given $X_k = i$ is $f(\cdot;\theta_i(\phi))$.

The order r of $\{X_k\}$ is assumed to be fixed and known, so that the statistical problem is to recursively estimate ϕ from an observation of $\{Y_k\}$. The most common parameterization is $\phi = (a_{11}, a_{12}, \dots, a_{rr}, \theta_1, \dots, \theta_r)$ with $a_{ij}(\cdot)$ and $\theta_i(\cdot)$ being the coordinate projections. We will refer to this case as the "usual parameterization" in the sequel. Having introduced this notation, the likelihood of a sequence of observations $y_1, \dots, y_m$, or the joint distribution of $Y_1, \dots, Y_m$, can be written

$$p^{(m)}(y_1,\dots,y_m;\phi) = \sum_{x_1=1}^{r} \cdots \sum_{x_m=1}^{r} \pi_{x_1}(\phi) f(y_1;\theta_{x_1}(\phi)) \prod_{k=2}^{m} a_{x_{k-1}x_k}(\phi) f(y_k;\theta_{x_k}(\phi)), \tag{1}$$

where π is the stationary distribution for $\{a_{ij}\}$, assuming this to be unique.

Let ϕ^0 denote the true parameter and let $K^{(m)}(\phi)$ denote the Kullback–Leibler divergence

$$K^{(m)}(\phi) = \mathbb{E}_{\phi^0}\left[\log \frac{p^{(m)}(Y_1,\dots,Y_m;\phi^0)}{p^{(m)}(Y_1,\dots,Y_m;\phi)}\right]$$

between ϕ^0 and some other parameter ϕ with respect to the m-dimensional distribution. It is well-known that $K^{(m)}(\phi)$ is well-defined and non-negative for all ϕ, even though it may be infinite. Moreover, $K^{(m)}(\phi) = 0$ if and only if $p^{(m)}(\cdot;\phi^0) = p^{(m)}(\cdot;\phi)$ μ^m-a.e., whence, assuming $p^{(m)}$ identifies ϕ^0, $\phi = \phi^0$. We will return to the question of identifiability.

We can estimate ϕ^0 by looking for an at least local minimum point of the Kullback–Leibler divergence $K^{(m)}(\phi)$, and this can be done with a recursive estimator of the form

$$\hat{\phi}_{n+1} = \hat{\phi}_n + \gamma_n S^{(m)}(Y_{nm+1},\dots,Y_{(n+1)m};\hat{\phi}_n), \tag{2}$$

where $S^{(m)} = \partial \log p^{(m)}/\partial\phi$ is the score function associated with $p^{(m)}$ and $\{\gamma_n\}$ is some positive sequence of numbers tending to zero. Obviously, this estimator is updated for each new group of m observations.

Typical conditions that are needed to ensure convergence of procedures of the form (2) include Lipschitz-continuity of $\partial K^{(m)}(\phi)/\partial\phi$, that ϕ^0 be a unique minimum of $K^{(m)}(\phi)$, and convexity of $K^{(m)}(\phi)$. In Rydén (1996) it was pointed out that in the case of i.i.d. observations from finite mixtures, these conditions are in general violated, and we cannot expect them to hold for HMMs either. The solution proposed in Rydén (1996), which we will adopt also in the present paper, is to constrain the recursive estimator to a compact convex set $G \subseteq \Phi$.

3. Consistency

In this section we give a general convergence result for a stochastic approximation procedure for HMMs and then apply this result to parameter estimation.

Before we state the result, we introduce some additional notation. Let $\{(X_k, Y_k)\}_{k=1}^{\infty}$ be a stationary HMM such that $\{X_k\}$ has state space $\{1,\dots,R\}$ and $\{Y_k\}$ takes values in some space $\mathscr{E}$.

We will consider the recursive procedure

$$\phi_{n+1} = P_G(\phi_n + \gamma_n h(Y_{n+1};\phi_n)), \tag{3}$$

where $h : \mathscr{E} \times \Phi \longrightarrow \mathbb{R}^q$ is some function, G is a closed, bounded, and convex subset of Φ, and P_G is the projection into G. The sequence $\{\gamma_n\}$ is defined by $\gamma_n = \gamma_0 n^{-\alpha}$ for some $\gamma_0 > 0$ and $\alpha \in (1/2, 1]$. We also assume that G is the closure of its interior, that G can be written in the form $G = \{\phi : g_i(\phi) \leqslant 0,\ i = 1,\dots,s\}$ for some finite set $\{g_i\}_{i=1}^{s}$ of continuously differentiable functions, and that at each $\phi \in \partial G$ the gradients of the active constraints are linearly independent. Here, the active constraints are those g-functions with $g(\phi) = 0$. The simplest G satisfying these requirements is a simplex, for which all g-functions are linear. The recursion is initialized at the point $\phi_0 = \phi^s$.

The symbol **1** denotes an indicator function. Finally we let $H(\phi) = \mathbb{E}h(Y_1;\phi)$ and $\xi_n = h(Y_{n+1};\phi_n) - H(\phi_n)$.

The following conditions will be used in the sequel.

C1. The transition probability matrix of $\{X_k\}$ is irreducible.

C2. For each $x \in \{1,\dots,R\}$, the function $\mathbb{E}[h(Y_1;\cdot)|X_1 = x]$ is finite and Lipschitz-continuous on G.

C3. $\mathbb{E}[|h(Y_1;\cdot)|^2]$ is bounded on G.

Remark. C1 guarantees that the stationary distribution of $\{X_k\}$, and hence the distribution of $\{(X_k, Y_k)\}$, is uniquely defined.

Lemma 1. *Let $\{(X_k, Y_k)\}$ be as above, let ϕ_n be defined by* (3), *and assume that* C1–C3 *are satisfied. Then $\sum_{k=1}^{\infty} k^{-\delta}\xi_k$ is convergent a.s. for every $\delta > \frac{1}{2}$.*

Proof. Note that C2 implies that $H(\cdot)$ is well-defined and Lipschitz-continuous on G, since

$$\begin{aligned}|H(\phi') - H(\phi)| &\leqslant \mathbb{E}\left|\mathbb{E}\left[h(Y_1;\phi') - h(Y_1;\phi)|X_1\right]\right| \\ &\leqslant \max_{1\leqslant x\leqslant R} \left|\mathbb{E}\left[h(Y_1;\phi') - h(Y_1;\phi)|X_1 = x\right]\right| \\ &\leqslant M_1|\phi' - \phi|,\end{aligned} \tag{4}$$

where M_1 is the largest of the Lipschitz-constants for the functions in C2.

The idea of the proof is to split the sum into three new ones and show that each of these three new sums is convergent. This technique was used also in Ma et al. (1990), but in the simpler context of stochastic approximation for finite-state Markov chains.

The HMM $\{(X_k, Y_k)\}$ is a Markov chain on $\{1,\dots,R\} \times \mathscr{E}$ and we let $\mathbf{P}$ denote its transition kernel. Note that the function h can be seen as a vector $(h_1,\dots,h_q)$ of real-valued functions defined on $\mathscr{E} \times \Phi$. Fix ϕ for a moment, fix $i \in \{1,\dots,q\}$, and consider the Poisson equation

$$u_i(x,y) = h_i(y;\phi) - H_i(\phi) + (\mathbf{P}u_i)(x,y), \quad (x,y) \in \{1,\dots,R\} \times \mathscr{E}, \tag{5}$$

where H_i is the ith component of H. In order to solve this equation we shall exploit the regenerative properties of $\{(X_k, Y_k)\}$. Indeed, by the very definition of an HMM, this process regenerates whenever $X_k = x_0$, where x_0 is some fixed but arbitrary state, and the regeneration cycles are independent. Let $\tau = \inf\{k > 1 : X_k = x_0\}$ be the first regeneration time after time one. By Proposition 5.2 in Asmussen (1994) (see also Eq. (3.2) in Glynn, 1994), the solution of the Poisson equation (5) is given by

$$\begin{aligned}u_i(x,y) &= \mathbb{E}\left[\sum_{k=1}^{\tau-1}\{h_i(Y_k;\phi) - H_i(\phi)\}\,\middle|\,(X_1, Y_1) = (x,y)\right] \\ &= h_i(y;\phi) + \mathbb{E}\left[\sum_{k=2}^{\tau-1} h_i(Y_k;\phi)\,\middle|\,X_1 = x\right] - H_i(\phi)\mathbb{E}[\tau - 1|X_1 = x] \\ &= h_i(y;\phi) + v_i(x;\phi) - H_i(\phi)w(x),\end{aligned} \tag{6}$$

say. Note that only the first term on the right-hand side depends on y, the reason being the conditional independence of $\{Y_k\}$ given $\{X_k\}$. Of course u_i depends also on ϕ, and we stress this by writing $u_i = u_i(x,y;\phi)$. For each i and x, $v_i(x;\cdot)$ is bounded on G, cf. (16) below, and $w(x)$ is finite for all x because $\{X_k\}$ is a finite-state irreducible Markov chain.

We proceed to the splitting of ξ_n. Defining the σ-algebras $\mathscr{F}_n = \sigma(X_1,\dots,X_n, Y_1,\dots,Y_n)$, we may write

$$\begin{aligned}\xi_n &= h(Y_{n+1};\phi_n) - H(\phi_n)\\ &= u(X_{n+1},Y_{n+1};\phi_n) - \mathbb{E}\left[u(X_{n+2},Y_{n+2};\phi_n)|\mathscr{F}_{n+1}\right]\\ &= u(X_{n+1},Y_{n+1};\phi_n) - \mathbb{E}\left[u(X_{n+1},Y_{n+1};\phi_n)|\mathscr{F}_n\right]\\ &\quad +\mathbb{E}\left[u(X_{n+1},Y_{n+1};\phi_n)|\mathscr{F}_n\right] - \mathbb{E}\left[u(X_{n+2},Y_{n+2};\phi_{n+1})|\mathscr{F}_{n+1}\right]\\ &\quad +\mathbb{E}\left[u(X_{n+2},Y_{n+2};\phi_{n+1})|\mathscr{F}_{n+1}\right] - \mathbb{E}\left[u(X_{n+2},Y_{n+2};\phi_n)|\mathscr{F}_{n+1}\right]\\ &= \xi_n^{(1)} + \xi_n^{(2)} + \xi_n^{(3)}.\end{aligned} \tag{7}$$

In Section 7 we show that each of the three sums so obtained is convergent. □

In order to proceed we need to introduce the following condition.

C4. There is a real-valued function L, defined on some open set $\mathcal{O} \supset G$, such that $H(\phi) = -\partial L(\phi)/\partial\phi$.

Define also the set of Kuhn–Tucker points for the problem of minimizing $L(\phi)$ over G,

$$KT = \left\{\phi \in G : \text{ there are } \lambda_i > 0, \quad i = 1,\dots,s, \text{ such that} \right.$$
$$\left. -H(\phi) + \sum_{i:g_i(\phi)=0} \lambda_i \frac{\partial}{\partial\phi} g_i(\phi) = 0\right\}.$$

Lemma 1 in particular shows that $\sum_{k=1}^{\infty} \gamma_k\xi_k$ is convergent a.s., and the following result is then an immediate consequence of Theorem 5.3.1 in Kushner and Clark (1978).

Corollary 1. *Let* $\{(X_k, Y_k)\}$ *be as above, let* ϕ_n *be defined by* (3), *and assume that* C1–C4 *are satisfied. Then* $\phi_n \to KT$ *a.s.*

Remark. Provided that $L(\phi)$ is twice continuously differentiable and that KT consists of a finite number of isolated components which are disjoint from ∂G, it can be shown that ϕ_n cannot fluctuate indefinitely between different components of KT. See Ljung (1978) for details.

Returning to our original estimation problem, let m be a fixed positive integer, let $\mathbf{X}_n = (X_{(n-1)m+1},\dots,X_{nm})$, $\mathbf{Y}_n = (Y_{(n-1)m+1},\dots,Y_{nm})$, $L(\phi) = K^{(m)}(\phi)$, $H(\phi) = -\partial K^{(m)}(\phi)/\partial\phi$, and $h(\mathbf{y};\phi) = S^{(m)}(\mathbf{y};\phi)$. Provided the conditions C1–C4 are satisfied for this choice, we obtain an estimator $\hat{\phi}_n$, defined by

$$\hat{\phi}_{n+1} = P_G\left(\hat{\phi}_n + \gamma_n S_m(Y_{n+1}; \hat{\phi}_n)\right), \tag{8}$$

which converges to the set KT defined in terms of $K^{(m)}(\phi)$ and G. Condition C1 is satisfied if $\{X_k\}$ is irreducible and aperiodic under ϕ^0, and C2–C4 are satisfied for many important parametric families $\{f(\cdot;\theta)\}$, for example the family of normal distributions (with variances bounded away from zero).

Clearly, it is of importance to have an idea of what the set KT looks like. If the m-dimensional distribution of $\{Y_k\}$ identifies ϕ^0, then the only global minima of $K^{(m)}(\phi)$ are ϕ^0 itself and possibly also parameters equal to ϕ^0 up to a permutation of states (observe that we can always permute the numbering of the states of $\{X_k\}$ leaving the distribution of $\{Y_k\}$ unchanged). In particular, this is true for any $m > 1$ if we have the usual parameterization, finite mixtures of the parametric family $\{f(y;\theta)\}$ are identifiable, and all θ_i^0 are distinct. The problem of identifiability is further discussed in Rydén (1995). Besides the location of the global minima, essentially no other properties of $K^{(m)}(\phi)$ are known, however, and KT may well contain other points. Indeed, estimation for mixtures is usually considered an inherently difficult problem.

Generally speaking, one may expect that $\hat{\phi}_n$ converges to ϕ^0, at least with high probability, if the initial estimate ϕ^s is reasonably close to ϕ^0, but more precise statements are usually very difficult to make. For this reason, we will *assume* that $\hat{\phi}_n \to \phi^0$ and proceed under this condition. Formally, fix the initial value ϕ^s, let $(\Omega, \mathscr{F})$ be the basic measure space, and assume that there is a set $\Omega^* \in \mathscr{F}$ with $\mathbb{P}_{\phi^0}(\Omega^*) > 0$ such that $\hat{\phi}_n \to \phi^0$ (or possibly some permutation of it) $\mathbb{P}_{\phi^0}$-a.s. on Ω^*. This set obviously depends on ϕ^0 as well as on ϕ^s, α, and γ_0.

We write $Z_k = \mathrm{O}(c_k)$ $(\mathrm{o}(c_k))$ if $\{c_k\}$ is a sequence of real numbers and $Z_k(\omega) = \mathrm{O}(c_k)$ $(\mathrm{o}(c_k))$ for all ω in a set of $\mathbb{P}_{\phi^0}$-probability one. An analogous terminology is used on subsets of Ω.

4. Averaging and asymptotic normality

For recursive estimation in the i.i.d. setting, it is well-known that if $\gamma_n = \mathrm{O}(n^{-1})$, i.e. $\alpha = 1$, then the sequence $\{\hat{\phi}_n\}$ (with $m = 1$) in general converges to ϕ^0 at rate $n^{-1/2}$, provided the eigenvalues of the Fisher information matrix are all larger than $\frac{1}{2}$, see Theorem 2.2 in Fabian (1968). This recursive estimator is never efficient, though, but can be made efficient if the score function is premultiplied by an adaptive matrix which estimates the inverse information matrix, see e.g. Fabian (1978). On the other hand, if $\alpha < 1$ then $\{\hat{\phi}_n\}$ will in general converge to ϕ^0 at the slower rate $n^{-\alpha/2}$, but the averaged estimate

$$\bar{\phi}_n = \frac{1}{n}\sum_{k=1}^{n} \hat{\phi}_k \tag{9}$$

will under mild conditions converge at rate $n^{-1/2}$ and in addition be efficient. The idea of using step lengths of order larger than n^{-1} and then averaging the so obtained sequence of estimates was proposed independently by Ruppert (1988) and Polyak (1990).

In our case $\{Y_k\}$ is not i.i.d., but the results below show that averaging indeed works also for the present problem. More precisely, we will show that, on Ω^*, the averaged estimator $\overline{\phi}_n$ converges at rate $n^{-1/2}$ and has the same asymptotic covariance matrix as an off-line estimator, the so-called maximum split-data likelihood estimator (MSDLE), that maximizes a "false" likelihood obtained by grouping the data into blocks of size m, see Rydén (1994) for details. For large m, the MSDLE is almost efficient, and hence so is $\overline{\phi}_n$.

Before stating these results, however, we need some additional assumptions. In the sequel, $m > 1$ is a fixed integer and the definitions of X_n, Y_n, h, etc., are as in the end of the previous section.

A1. $\{X_k\}$ is irreducible under all $\phi \in G$ and aperiodic under ϕ^0. The true parameter ϕ^0 is an interior point of G.

A2. There is some open set $\mathcal{O} \supset G$ such that $\log p^{(m)}(y;\cdot)$ is continuously differentiable on $\mathcal{O}$ for all $y \in E^m$ and $K^{(m)}(\phi)$ is continuously differentiable under the expectation operator on $\mathcal{O}$. In addition, the function $\phi \mapsto \mathbb{E}_{\phi^0}[S^{(m)}(Y_1,\ldots,Y_m;\phi)| X_1 = x_1,\ldots,X_m = x_m]$ is Lipschitz-continuous for all $x_1,\ldots,x_m \in \{1,\ldots,r\}$ such that $\mathbb{P}_{\phi^0}(X_1 = x_1,\ldots,X_m = x_m) > 0$.

A3. $\mathbb{E}_{\phi^0}|S^{(m)}(Y_1,\ldots,Y_m;\phi)|^2$ is bounded on G.

A4. $S^{(m)}(y;\cdot)$ is continuously differentiable in some neighbourhood $\mathcal{U}$ of ϕ^0 for each $y \in E^m$, and

$$\mathbb{E}_{\phi^0}\left[\sup_{\phi\in\mathcal{U}}\left|\frac{\partial}{\partial\phi}S^{(m)}(Y_1,\ldots,Y_m;\phi)\right|^2\right] < \infty.$$

A5. The information matrix $\mathcal{J}_0$ associated with $p^{(m)}$ at ϕ^0, i.e. $\partial^2K^{(m)}(\phi)/\partial\phi^2$ at ϕ^0, is positive definite.

A6. $K^{(m)}(\phi)$ is three times continuously differentiable in a neighbourhood of ϕ^0.

A7. $\frac{2}{3} < \alpha < 1$.

A8. For some $\delta > 2(\alpha^{-1} - 1)$, $\mathbb{E}_{\phi^0}|S^{(m)}(Y_1,\ldots,Y_m;\phi^0)|^{2+\delta} < \infty$.

Remarks. (i) The irreducibility assumption in A1 guarantees that $p^{(m)}$, $S^{(m)}$, etc., are uniquely defined on G. (ii) Note that A4 implies that $K(\phi)$ is twice continuously differentiable under the expectation operator in a neighbourhood of ϕ^0, so that the matrix $\mathcal{J}_0$ defined in A5 is indeed a covariance matrix and hence always positive semi-definite. (iii) Because of A7, A8 is always satisfied if the third moment of $S^{(m)}$ exists at ϕ^0. (iv) A1–A3 guarantee that C1–C4 hold with h, etc. defined as in the end of the previous section. The aperiodicity assumption asserts that the Markov chain $\{X_k\}$ is irreducible.

Lemma 2. *Assume* A1–A8. *Then there exists an* $\varepsilon > 0$ *such that*

$$\overline{\phi}_n - \phi^0 = \mathcal{J}_0^{-1}\frac{1}{n}\sum_{k=1}^{n}\xi_k + o\left(n^{-1/2-\varepsilon}\right)$$

on Ω^*, *where* ξ_n *is as in Lemma* 1.

The proof is found in Section 7. We now give the main result of the paper.

Theorem 1. *Assume* A1–A8. *Then the following two representations are valid on* Ω^*,

$$\text{(i)}\ n^{1/2}(\bar{\phi}_n - \phi^0) = \mathscr{I}_0^{-1}\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\{u(X_{k+1}, Y_{k+1}; \phi^0) - \mathbb{E}_{\phi^0}\left[u(X_{k+1}, Y_{k+1}; \phi^0)|\mathscr{F}_k\right]\} + \mathrm{o}(1),$$

$$\text{(ii)}\ n^{1/2}(\bar{\phi}_n - \phi^0) = \mathscr{I}_0^{-1}\frac{1}{\sqrt{n}}\sum_{k=1}^{n} S_m(Y_{k+1}; \phi^0) + \mathrm{o}(1),$$

where u is as in the proof of Lemma 1.

Proof. Let $\zeta_n^{(1)} = u(X_{k+1}, Y_{k+1}; \phi^0) - \mathbb{E}_{\phi^0}[u(X_{k+1}, Y_{k+1}; \phi^0)|\mathscr{F}_k]$. By the preceding lemma and its proof it follows that for some $\varepsilon > 0$, $\bar{\phi}_n - \phi^0 = \mathscr{I}_0^{-1} n^{-1}\sum_1^n \xi_k + \mathrm{o}(n^{-1/2-\varepsilon}) = \mathscr{I}_0^{-1} n^{-1}\sum_1^n \zeta_k^{(1)} + \mathrm{o}(n^{-1/2-\varepsilon})$, proving (i).

For (ii), split $h(Y_{n+1}; \phi^0) - H(\phi^0)$ similarly to (7) as

$$\begin{aligned} h(Y_{n+1}; \phi^0) - H(\phi^0) &= u(X_{n+1}, Y_{n+1}; \phi^0) - \mathbb{E}_{\phi^0}[u(X_{n+2}, Y_{n+2}; \phi^0)|\mathscr{F}_{n+1}] \\ &= u(X_{n+1}, Y_{n+1}; \phi^0) - \mathbb{E}_{\phi^0}[u(X_{n+1}, Y_{n+1}; \phi^0)|\mathscr{F}_n] \\ &\quad + \mathbb{E}_{\phi^0}[u(X_{n+1}, Y_{n+1}; \phi^0)|\mathscr{F}_n] \\ &\quad - \mathbb{E}_{\phi^0}[u(X_{n+2}, Y_{n+2}; \phi^0)|\mathscr{F}_{n+1}] \\ &= \zeta_n^{(1)} + \zeta_n^{(2)}. \end{aligned}$$

The sum $\sum_1^n \zeta_k^{(2)}$ telescopes and equals $\mathrm{O}(1)$ (cf. part B of the proof of Lemma 1), and since $H(\phi^0) = 0$, (ii) follows. □

The above theorem is powerful, since from the strong representations one can derive (functional) central limit theorems, large deviation results, etc. As an example, we state the following CLT. Let $\bar{\phi}_n^*$ be the restriction of $\bar{\phi}_n$ to Ω^*, define $\mathscr{F}^* = \{A \cap \Omega^* : A \in \mathscr{F}\}$, and define the probability measure $\mathbb{P}^*$ on $(\Omega^*, \mathscr{F}^*)$ by $\mathbb{P}^*(A) = \mathbb{P}_{\phi^0}(A)/\mathbb{P}_{\phi^0}(\Omega^*)$.

Let $i_0 \in \{1, \ldots, r\}$ be an arbitrary state and define $\tau' = \min\{k > 1 : X_{(k-1)m+1} = i_0\}$ and

$$V_{ij} = \mathbb{E}_{\phi^0}\left[\sum_{k=1}^{\tau'-1} S_i^{(m)}(Y_{(k-1)m+1}, \ldots, Y_{km}; \phi^0) \times \sum_{k=1}^{\tau'-1} S_j^{(m)}(Y_{(k-1)m+1}, \ldots, Y_{km}; \phi^0) \,\middle|\, X_1 = i_0\right].$$

Theorem 2. *Assume* A1–A8. *Then*

$$n^{1/2}(\bar{\phi}_n^* - \phi^0) \to \mathscr{N}(0, \pi_{i_0}^0 \mathscr{I}_0^{-1} V \mathscr{I}_0^{-1}) \quad \mathbb{P}_{\phi^0}^*\text{-weakly}.$$

Proof. By the central limit theorem for martingales, see e.g. Corollary 3.6 in Rootzén (1983), the sum in Theorem 1(i) converges $\mathbb{P}_{\phi^0}$-weakly to a normal distribution with covariance matrix V'. Moreover, this convergence is (Rényi) stable, see Theorem 4.2 in Rootzén (1983), so that the sequence restricted to Ω^* converges $\mathbb{P}^*_{\phi^0}$-weakly to the same normal law. It remains to show that $V' = \pi^0_{i_0} V$. By an argument based on regenerative theory, the sum in Theorem 1(ii) converges $\mathbb{P}_{\phi^0}$-weakly to a normal law with covariance matrix $\pi^0_{i_0} V$, see Rydén (1994) for details. Thus $V' = \pi^0_{i_0} V$ and the proof is complete. □

5. Numerical results

The procedure presented in the previous section was applied to the problem of estimating the transition probabilities and means in a mixture of two normal distributions with known variance $\sigma^2 = 1$ and Markov regime. The parameter vector can thus be taken as $\phi = (a_{12}, a_{21}, \mu_1, \mu_2)$, where the first two parameters are the probabilities of switches in the Markov chain and μ_i is the mean of the normal distribution when the chain is in state i.

Concerning the choice of the group size m, it was observed in Rydén (1994) that for the problems considered in that paper, the asymptotic variances of the estimates are about as small as they can be if m is chosen larger than some threshold. This is true also for the present problem, and the value $m = 20$ is well above this threshold.

If the initial parameter ϕ^s is far from the true parameter, then $\hat{\phi}_k$ is a poor estimate for small k and the contributions from these terms in (9) decay only as n^{-1}. Thus it may be favourable not to start the averaging at once, but rather wait until the basic recursion (8) has been run for n_a steps. This was done in the simulations presented below, with $n_a = 200$. The step lengths γ_n were chosen as $\gamma_n = \gamma_0 n^{-0.7}$.

Table 1 shows simulation results for six different initial parameters ϕ^s and $\gamma_0 = 0.01$ or 0.05. True parameter was $\phi^0 = (0.25, 0.25, 0, 2)$ in all cases, so that the hidden chain has somewhat slow dynamics and the marginal distribution of $\{Y_k\}$ is the mixture $0.5\,\mathcal{N}(0,1)+0.5\,\mathcal{N}(2,1)$. This density is unimodal with a flat peak. The following cases are referred to in the table:

(A) $\phi^s = (0.25, 0.25, 0, 2)$, (B) $\phi^s = (0.25, 0.25, -0.5, 2.5)$,
(C) $\phi^s = (0.25, 0.25, 0.5, 1.5)$, (D) $\phi^s = (0.10, 0.10, 0, 2)$,
(E) $\phi^s = (0.40, 0.40, 0, 2)$, (F) $\phi^s = (0.10, 0.10, 0.5, 1.5)$.

Case A is obviously the easiest one, with $\phi^s = \phi^0$, and case F is the most difficult one.

The results are based on 500 simulated replicates of $N = 20\,000$ samples each, so that each replicate contained $N/m = 1000$ groups. The replicates were identical for all cases, i.e. exactly the same data was used in all cases. Table 1 shows, for each case, bias and normalized sampled variances over the 500 replicates. The maximum split-data likelihood estimate is also shown. The row labelled "variance bounds" gives lower bounds on the (normalized) asymptotic variances, i.e. the diagonal elements of the matrix $\pi_{i_0} \mathscr{I}_0^{-1} V \mathscr{I}_0^{-1}$ in Theorem 2, obtained by simulation as described in Rydén (1994). The MSDLE conforms reasonably well to these bounds, actually it does a little

Table 1
Simulation results for the averaged recursion $\overline{\phi}_n$. For each case, bias and normalized sample variances (Ns^2) are shown

Parameter	a_{12}		a_{21}		μ_1		μ_2	
True values	0.25		0.25		0		2	
Bounds	10^{-6}, $1\text{-}10^{-6}$		10^{-6}, $1\text{-}10^{-6}$		-10^6, 10^6		-10^6, 10^6	
	Bias	Ns^2	Bias	Ns^2	Bias	Ns^2	Bias	Ns^2
Variance bounds		0.93		0.93		4.92		4.92
MSDLE	1.3E−4	0.83	4.4E−4	0.99	1.5E−4	4.46	4.2E−4	4.60
Case A, $\gamma_0 = 0.01$	1.5E−3	1.01	2.0E−3	1.22	3.5E−4	5.16	1.2E−3	5.65
Case A, $\gamma_0 = 0.05$	1.7E−3	1.02	2.3E−3	1.36	1.0E−3	8.98	1.2E−3	8.62
Case B, $\gamma_0 = 0.01$	1.8E−2	1.38	1.8E−2	1.57	−1.0E−1	12.1	1.1E−1	12.7
Case B, $\gamma_0 = 0.05$	2.0E−3	1.12	2.3E−3	1.38	−1.4E−4	10.9	1.1E−3	10.2
Case C, $\gamma_0 = 0.01$	−2.4E−2	0.93	−2.4E−2	1.15	1.4E−1	7.91	−1.4E−1	8.34
Case C, $\gamma_0 = 0.05$	1.7E−3	1.10	2.0E−3	1.37	1.7E−3	10.4	−6.7E−4	10.1
Case D, $\gamma_0 = 0.01$	3.9E−3	6.61	5.3E−3	9.46	7.3E−3	24.4	−3.2E−3	25.4
Case D, $\gamma_0 = 0.05$	1.7E−3	1.04	2.2E−3	1.29	1.0E−3	8.47	9.3E−4	8.13
Case E, $\gamma_0 = 0.01$	9.2E−3	1.17	9.8E−3	1.30	−4.3E−3	4.71	6.0E−3	5.06
Case E, $\gamma_0 = 0.05$	1.9E−3	1.04	2.2E−3	1.28	4.2E−4	8.23	7.9E−4	7.82
Case F, $\gamma_0 = 0.01$	3.9E−3	135	2.0E−3	120	1.6E−1	94.7	−1.6E−1	112
Case F, $\gamma_0 = 0.05$	1.5E−3	1.16	1.9E−3	1.34	3.0E−3	10.0	−1.2E−3	10.3

better, a finite sample effect. The compact convex set G used in the projection P_G was of the simple form $\prod_1^d[low_i, high_i]$, where the values of the bounds low_i and $high_i$ are shown in Table 1.

Comments on the results:

(i) For $\gamma_0 = 0.01$, a non-negligible bias occurs in cases B, C, and F. This γ_0 is simply too small to let the steps taken by $\hat{\phi}$ be large enough. Choosing $\gamma_0 = 0.05$ removes this problem and improves the overall performance, with case E being a notable exception in which both choices lead to small bias and $\gamma_0 = 0.01$ gives the smallest variances (at least for the μ-parameters). Case A is another exception, but since $\phi^s = \phi^0$ here, it is of less interest.

The value $\gamma_0 = 0.25$ was also tested, but then the recursive procedure not always seemed to converge to ϕ^0. For the other two (smaller) values of γ_0, $\overline{\phi}_n$ always seemed to converge to ϕ^0, except possibly in case F with $\gamma_0 = 0.01$. Convergence was assessed by plotting the estimates. Normal probability plots of the estimates revealed no gross deviations from normality except for the cases D and F with $\gamma_0 = 0.01$.

(ii) Other values of n_a were tested as well, namely $n_a = 0$, 10, 25, 100, and 400. Of the values up to 200, $n_a = 200$ consistently gave the smallest variances and among the smallest biases. Hence, one should not hesitate to exclude a substantial portion of the sample, here 20%, from the averaged estimator. For $n_a = 400$, the sample variances of the estimates of μ_1 and μ_2 continued to decrease, while the sample variances of the estimates of a_{12} and a_{21} increased slightly.

(iii) For $\gamma_0 = 0.05$, which gave the best performance, the normalized samples variances for a_{12} and a_{21} are up to 50% larger than the asymptotic lower bounds, and the normalized sample variances for μ_1 and μ_2 are up to 120% larger than the lower bounds. This is not completely satisfactory, but Ruppert (1991) wrote, "In stochastic approximation, finite-sample distributions are often quite different from what asymptotic theory suggests, even for moderately large sample sizes." Moreover, the algorithm seems stable as in case F the initial point is quite far from the true parameters.

(iv) The choice $\alpha = 0.9$ was also tested, resulting in bias and variances that were consistently larger than those for $\alpha = 0.7$. These results are therefore not reported.

6. Comparison with the Holst–Lindgren procedure

Holst and Lindgren (1991) proposed a recursive estimator for HMMs of the form

$$\tilde{\phi}_{n+1} = \tilde{\phi}_n + \frac{1}{n+1} H_n h_{n+1}(Y_{n+1}; \tilde{\phi}_n), \tag{10}$$

where, essentially,

$$h_{n+1}(Y_{n+1}; \phi) = \mathbb{E}_\phi \left[\frac{\partial \ell_{n+1}}{\partial \phi} \,\middle|\, Y_1, \ldots, Y_{n+1} \right], \tag{11}$$

and

$$\ell_{n+1} = \sum_{i,j} \mathbf{1}\{X_n = i, X_{n+1} = j\}\{\log a_{ij}(\phi) + \log f(Y_{n+1}; \theta_j(\phi))\} \tag{12}$$

is the conditional loglikelihood for (X_{n+1}, Y_{n+1}) given X_n, and H_n is an estimate of the inverse of $\mathbb{E}_{\phi^0}[h_n(Y_n; \phi^0) h_n^T(Y_n; \phi^0)]$. Holst and Lindgren write that h_{n+1} is the "score function for a new observation", but h_{n+1} is not equal to $\partial \log p(Y_{n+1} \mid Y_n, \ldots, Y_1; \phi)/\partial \phi$. The recursive procedure (10) is rather a scheme for finding roots of the function $Q(\phi) = \lim_{n\to\infty} \mathbb{E}_{\phi^0}[h_n(Y_n; \phi)]$. This function can be given a strict definition, cf. Leroux (1992, p. 133), and it indeed has a zero at ϕ^0.

Moreover, provided that $\tilde{\phi}_n \to \phi^0$, the limiting behaviour of $n^{1/2}(\tilde{\phi}_{n+1} - \phi^0)$ is the same as that of

$$\mathscr{I}_{\mathrm{HL}}^{-1} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} h_{n+1}(Y_{n+1}; \phi^0), \tag{13}$$

where $\mathscr{I}_{\mathrm{HL}} = \lim_{n\to\infty} \mathbb{E}_{\phi^0}[h_n(Y_n; \phi^0) h_n^{\mathrm{T}}(Y_n; \phi^0)]$.

This follows from a strong representation similar to the one employed in the previous section, cf. Ruppert (1982) and Schwabe (1986). Now, it is relatively easy to see that under $\mathbb{P}_{\phi^0}$, $h_{n+1}(Y_{n+1}; \phi^0)$ is a martingale difference sequence with respect to $\sigma(Y_1, \ldots, Y_{n+1})$, and the martingale central limit theorem shows that the limiting distribution of (13) is a centered multivariate normal distribution with covariance matrix $\mathscr{I}_{\mathrm{HL}}^{-1} \mathscr{I}_{\mathrm{HL}} \mathscr{I}_{\mathrm{HL}}^{-1} = \mathscr{I}_{\mathrm{HL}}^{-1}$. Simulation results in Holst and Lindgren (1991) confirm this result.

The matrix $\mathscr{I}_{\mathrm{HL}}$ was simulated as described in Holst and Lindgren (1991) for the estimation problem considered in the previous section using 400 000 samples of the h-vector. The resulting diagonal elements of $\mathscr{I}_{\mathrm{HL}}^{-1}$ were 1.62, 1.64, 5.49, and 5.48, respectively, and these should be compared to the lower bounds given in Table 1. For the transition probabilities a_{12} and a_{21}, the Holst–Lindgren procedure gives asymptotic variances about 75% larger than those of the averaged estimator $\bar{\phi}_n$, and the corresponding figure for the normal means μ_1 and μ_2 is 10%.

Krishnamurthy and Moore (1993) proposed a recursive procedure similar to the one given by Holst and Lindgren. Their recursion (3.18) has the form (10), but their scaling matrix H_n is the inverse of the conditional expectation (given $Y_1, \ldots, Y_n$) of the information matrix for the complete observation including also the hidden Markov chain.

7. Proofs of Lemmas 1 and 2

Proof of Lemma 1 (*continued*)

(A) Trivially, $\{\xi_n^{(1)}\}$ is a martingale difference sequence. The conditional variance of its ith component $\xi_{n,i}^{(1)}$ may be bounded as

$$\begin{aligned}
\mathbb{E}\left[\left(\xi_{n,i}^{(1)}\right)^2 \middle| \mathscr{F}_n\right] &= \mathbb{E}\left[\{u_i(X_{n+1}, Y_{n+1}; \phi) - \mathbb{E}[u_i(X_{n+1}, Y_{n+1}; \phi) \mid \mathscr{F}_n]\}^2 \middle| \mathscr{F}_n\right]\Big|_{\phi=\phi_n} \\
&\leqslant \mathbb{E}\left[u_i^2(X_{n+1}, Y_{n+1}; \phi) \middle| \mathscr{F}_n\right]\Big|_{\phi=\phi_n} \\
&\leqslant \sup_{\phi \in G} \mathbb{E}\left[u_i^2(X_{n+1}, Y_{n+1}; \phi) \middle| \mathscr{F}_n\right] \qquad (14) \\
&\leqslant \sup_{\phi \in G} \mathbb{E}\left[3h_i^2(Y_{n+1}; \phi) + 3v_i^2(X_{n+1}; \phi) + 3H_i^2(\phi) w^2(X_{n+1}) \middle| \mathscr{F}_n\right] \\
&\leqslant 3 \sup_{\phi \in G} \max_{1 \leqslant x \leqslant R} \mathbb{E}\left[h_i^2(Y_{n+1}; \phi) \middle| X_{n+1} = x\right] + 3 \sup_{\phi \in G} \max_{1 \leqslant x \leqslant R} v_i^2(x; \phi) \\
&\quad + 3 \sup_{\phi \in G} H_i^2(\phi) \max_{1 \leqslant x \leqslant R} w^2(x).
\end{aligned}$$

C1 and C3 imply that $\mathbb{E}[h_i^2(Y_1; \phi) | X_1 = x]$ is bounded in x and $\phi \in G$, so that the right-hand side of (14) is finite. Since $\sum_1^\infty k^{-2\delta} < \infty$, it follows by a standard theorem of martingale theory, see e.g. Theorem 2.15 in Hall and Heyde (1980), that $\sum_1^\infty k^{-\delta} \xi_k^{(1)}$ is convergent a.s.

(B) First note that

$$\begin{aligned}
\sum_{k=1}^n k^{-\delta} \xi_{k,i}^{(2)} &= \mathbb{E}\left[u_i(X_2, Y_2; \phi_1) \middle| \mathscr{F}_1\right] \\
&\quad + \sum_{k=2}^n \{k^{-\delta} - (k-1)^{-\delta}\} \mathbb{E}\left[u_i(X_{k+1}, Y_{k+1}; \phi_k) \middle| \mathscr{F}_k\right] \\
&\quad - n^{-\delta} \mathbb{E}\left[u_i(X_{n+2}, Y_{n+2}; \phi_{n+1}) \middle| \mathscr{F}_{n+1}\right]. \qquad (15)
\end{aligned}$$

It follows as above that $\mathbb{E}[|u_i(X_{k+1}, Y_{k+1}; \phi_k)| \, |\mathscr{F}_k]$ is bounded, and since $k^{-\delta} - (k-1)^{-\delta} = O(k^{-1-\delta})$ the sum on the right-hand side of (15) is absolutely convergent, and thus also convergent, a.s. It is also immediate that the last term on the right-hand side of (15) tends to zero, and hence $\sum_1^\infty k^{-\delta} \xi_k^{(2)}$ is convergent a.s.

(C) By (5) and (6),

$$\begin{aligned}\mathbb{E}[u_i(X_{n+2},Y_{n+2};\phi)|\mathscr{F}_{n+1}] &= u_i(X_{n+1},Y_{n+1};\phi)+H_i(\phi)-h_i(Y_{n+1};\phi)\\ &= v_i(X_{n+1};\phi)-H_i(\phi)(w(X_{n+1})-1),\end{aligned}$$

whence

$$\xi^{(3)}_{n,i} = v_i(X_{n+1};\phi_{n+1})-v_i(X_{n+1},\phi_n)-\{H_i(\phi_{n+1})-H_i(\phi_n)\}(w(X_{n+1})-1).$$

Moreover,

$$\begin{aligned}&|v_i(x;\phi')-v_i(x;\phi)|\\ &= \left|\mathbb{E}\left[\sum_{k=2}^{\infty}\mathbf{1}\{k<\tau\}[h_i(Y_k;\phi')-h_i(Y_k;\phi)]\,\middle|\,X_1=x\right]\right|\\ &\leqslant \sum_{k=2}^{\infty}\mathbb{E}\left[\mathbf{1}\{k<\tau\}\left|\mathbb{E}\left[h_i(Y_k;\phi')-h_i(Y_k;\phi)\,|\,\{X_\ell\}_{\ell=2}^{\infty},X_1=x\right]\right|\,\middle|\,X_1=x\right]\\ &= \sum_{k=2}^{\infty}\mathbb{E}\left[\mathbf{1}\{k<\tau\}\left|\mathbb{E}\left[h_i(Y_k;\phi')-h_i(Y_k;\phi)\,|\,X_k\right]\right|\,\middle|\,X_1=x\right]\\ &\leqslant \sum_{k=2}^{\infty}\mathbb{E}\left[\mathbf{1}\{k<\tau\}\max_{1\leqslant x'\leqslant R}\left|\mathbb{E}\left[h_i(Y_k;\phi')-h_i(Y_k;\phi)\,|\,X_k=x'\right]\right|\,\middle|\,X_1=x\right]\\ &\leqslant \max_{1\leqslant x\leqslant R}w(x)\max_{1\leqslant x'\leqslant R}\left|\mathbb{E}\left[h_i(Y_1;\phi')-h_i(Y_1;\phi)\,|\,X_1=x'\right]\right|. \qquad (16)\end{aligned}$$

By C2, the right-hand side of (16) is bounded by

$$\max_{1\leqslant x\leqslant R} w(x)M_1|\phi'-\phi| = M_2|\phi'-\phi|,$$

where M_1 is as in (4). Thus,

$$\left|\xi^{(3)}_{n,i}\right| \leqslant M_2\,|\phi_{n+1}-\phi_n| + M_1\max_{1\leqslant x\leqslant R}|w(x)-1|\,|\phi_{n+1}-\phi_n| = M_3\,|\phi_{n+1}-\phi_n|.$$

Now, since the projection P_G is a contraction, $|\phi_{n+1}-\phi_n|\leqslant\gamma_n|h(Y_{n+1};\phi_n)|$, and it follows that

$$n^{-\delta}|\xi^{(3)}_{n,i}| \leqslant n^{-\delta}\gamma_n M_3|h(Y_{n+1};\phi_n)|.$$

By C1 and C3, $E[|h(Y_{n+1};\phi_n)|\,|\mathscr{F}_n]$ is bounded by some finite constant, so that the corresponding unconditional expectation is bounded by the same constant as well. We conclude that $\sum_1^\infty k^{-\delta}\xi^{(3)}_k$ is (absolutely) convergent a.s. □

In the next proof, we say that a sequence $\{Z_k\}$ of random variables has a property eventually if there exists a set $B\subseteq\Omega$ with $\mathbb{P}_{\phi^0}(B) = 1$ such that for all $\omega\in B$, $\{Z_k(\omega)\}$ has the property in question for $k\geqslant K(\omega)$. An analogous terminology is used on subsets of Ω.

Proof of Lemma 2. Since $\mathscr{I}_0$ is a covariance matrix it is symmetric and can be thus be diagonalized by a real orthogonal matrix. Write this diagonalization as $\mathscr{I}_0 = R\Lambda R^{-1}$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. By A5, $\lambda_i > 0$ for each i. Define $\tilde{\phi}_n = R^{-1}\hat{\phi}_n$ etc., and define the q-variate process $\{\tilde{Z}_k\}$ by $\tilde{Z}_0 = 0$ and

$$\tilde{Z}_{n+1} = \tilde{Z}_n - \gamma_n(\Lambda \tilde{Z}_n - \tilde{\xi}_n), \tag{17}$$

where $\tilde{\xi}_n = R^{-1}\xi_n$.

The first claim we want to show is

$$n^\beta \tilde{Z}_n \to 0 \ \mathbb{P}_{\phi^0}\text{-a.s. on } \Omega^* \text{ implies that } n^\beta(\tilde{\phi}_n - \tilde{\phi}^0) \to 0 \ \mathbb{P}_{\phi^0}\text{-a.s. on } \Omega^*. \tag{18}$$

To prove this, note that since ϕ^0 is an interior point of G, the projection in (8) is applied only a finite (but random) number of times on Ω^*, and on this event the recursion is thus eventually given by

$$\tilde{\phi}_{n+1} = \tilde{\phi}_n - \gamma_n(-\tilde{H}(\tilde{\phi}_n)) - \gamma_n(-\tilde{\xi}_n),$$

where $\tilde{H}(\tilde{\varphi}) = R^{-1}H(R\tilde{\varphi})$. The Jacobian of $-\tilde{H}$ at $\tilde{\phi}^0$ is Λ, whence for each coordinate i, $1 \leqslant i \leqslant q$, there exist positive numbers $\lambda_i^{(2)} \geqslant \lambda_i^{(1)}$ and a neighbourhood $\mathscr{U}_i$ of $\tilde{\phi}^0$ such that

$$\lambda_i^{(1)}|\tilde{\varphi}_i - \tilde{\phi}_i^0| \leqslant |\tilde{H}_i(x)| \leqslant \lambda_i^{(2)}|\tilde{\varphi}_i - \tilde{\phi}_i^0|$$

for $\tilde{\varphi} \in \mathscr{U}_i$. The claim (18) can now be proved by a coordinate-wise application of the proof of Theorem 2 in Schwabe and Walk (1996).

Now, provided that we can show

$$n^\beta \tilde{Z}_n \to 0 \ \mathbb{P}_{\phi^0}\text{-a.s.} \quad \text{on } \Omega^* \quad \text{for all } \beta < \alpha/2, \tag{19}$$

we obtain

$$\bar{\tilde{\phi}}_{n,i} - \tilde{\phi}_i^0 = \lambda_i^{-1}\frac{1}{n}\sum_{k=1}^{n}\tilde{\xi}_{k,i} + \mathrm{o}(n^{-1/2-\varepsilon})$$

on Ω^* for each i and some $\varepsilon > 0$, cf. Eq. (17) in Schwabe (1993) and Theorem 3 in Schwabe and Walk (1996). This is the point where A6 is needed. By a change of variables back to the original coordinate system, the theorem then follows.

It thus remains to prove (19). Since Λ is diagonal, by Lemma 5 in Schwabe and Walk (1996), (19) follows if we can show that for each i, the process $\{\tilde{\xi}_{k,i}\}$ satisfies a strong invariance principle of the form

$$\sum_{k=1}^{n}\tilde{\xi}_{k,i} = \sigma B_i(n) + \mathrm{o}(n^{\alpha/2}) \quad \mathbb{P}_{\phi^0}\text{-a.s. on } \Omega^*, \tag{20}$$

where $\sigma > 0$ and $\{B_i(t)\}_{t \geqslant 0}$ is a standard Brownian motion defined on $(\Omega, \mathscr{F}, \mathbb{P}_{\phi^0})$. Possibly one needs to enlarge this probability space in order to define $\{B_i(t)\}$, but this can always be done without changing the distributions of the original processes, cf. Philipp and Stout (1975).

Split $\tilde{\xi}_n$ into the sum $\tilde{\xi}_n = \tilde{\xi}_n^{(1)} + \tilde{\xi}_n^{(2)} + \tilde{\xi}_n^{(3)}$ as in the proof of Lemma 1. First we have, with $\tilde{u} = R^{-1}u$,

$$\sum_{k=1}^{n}\tilde{\xi}_{k,i}^{(2)} = \mathbb{E}_{\phi^0}\left[\tilde{u}_i(X_2, Y_2; \hat{\phi}_1)\middle|\mathscr{F}_1\right] - \mathbb{E}_{\phi^0}\left[\tilde{u}_i(X_{n+2}, Y_{n+2}; \hat{\phi}_{n+1})\middle|\mathscr{F}_{n+1}\right] = \mathrm{O}(1).$$

Secondly, by part C of the proof of Lemma 1,

$$\left|\sum_{k=1}^{n} \tilde{\xi}_{k,i}^{(3)}\right| \leqslant M_3 \sum_{k=1}^{n} \gamma_k |h(Y_{k+1};\hat{\phi}_k)|.$$

By Kronecker's lemma this expression is $o(n^{\alpha/2})$ provided $\sum_1^\infty k^{-3\alpha/2}|h(Y_{k+1};\hat{\phi}_k)| < \infty$ $\mathbb{P}_{\phi^0}$-a.s. But this is true, since the sum indeed has finite expectation. This follows by again using an argument as in part C of the proof of Lemma 1 and the assumption $\alpha > 2/3$.

Hence, it suffices to show (20) with $\tilde{\xi}_{k,i}$ replaced by $\tilde{\xi}_{k,i}^{(1)}$, and we proceed by doing that. The coordinate-wise solution of the linear difference equation (17) can be written

$$\tilde{Z}_{n+1,i} = d_{n,i}^{-1} \sum_{k=1}^{n} \gamma'_k d_{k,i} \tilde{\xi}_{k,i} + d_{n,i}^{-1} \tilde{Z}_{n_0,i}, \qquad n \geqslant n_0, \tag{21}$$

where n_0 is a number such that $\lambda_i \gamma_n < 1$ for all $n \geqslant n_0$ and all i, $\gamma'_n = 0$ if $n < n_0$ and $\gamma'_n = \gamma_n$ otherwise, and $d_{n,i}$ is defined by

$$d_{n,i} = \prod_{k=1}^{n} (1 - \lambda_i \gamma'_k)^{-1},$$

see Schwabe and Walk (1996). Using (21) and Kronecker's lemma it is clear that $n^\beta \tilde{Z}_n \to 0$ if $\sum_1^\infty k^\beta \gamma'_k \tilde{\xi}_k$ is convergent. But since $k^\beta \gamma'_k = \gamma_0 k^{-\alpha+\beta}$ for large k, by Lemma 1 this sum is convergent if $-\alpha + \beta < -\frac{1}{2}$. Thus $n^\beta \tilde{Z}_n \to 0$ for $\beta < \alpha - \frac{1}{2}$, and by (18) we obtain $n^\beta(\hat{\phi}_n - \phi^0) \to 0$ $\mathbb{P}_{\phi^0}$-a.s. on Ω^* for such β.

Now we want to show that the solution u of the Poisson equation (5) as a function of ϕ is continuously differentiable on $\mathscr{U}$ (see A4) for each (x,y), and for this we use the representation (6). By A4, $h(y;\cdot) = S^{(m)}(y;\cdot)$ and $H(\cdot) = \mathbb{E}_{\phi^0}[S^{(m)}(Y_1;\cdot)]$ are continuously differentiable on $\mathscr{U}$, whence it is sufficient to prove this property for v. By A4 and in analogy with (16) we obtain

$$\mathbb{E}_{\phi^0}\left[\sup_{\phi\in\mathscr{U}}\left|\sum_{k=2}^{\tau-1} \frac{\partial}{\partial\phi} h(Y_k;\phi)\right| \,\middle|\, X_1 = x\right]$$

$$\leqslant \mathbb{E}_{\phi^0}\left[\sum_{k=2}^{\tau-1} \sup_{\phi\in\mathscr{U}}\left|\frac{\partial}{\partial\phi} h(Y_k;\phi)\right| \,\middle|\, X_1 = x\right] < \infty,$$

so that the expression in (6) for v can be differentiated inside the expectation, yielding a continuous derivative on $\mathscr{U}$.

Let $\eta \in (0, \alpha - \frac{1}{2})$ and define the functions $u^{(n)}$, $n = 1,2,\ldots$, by

$$u^{(n)}(x,y;\phi) = \begin{cases} u(x,y;\phi) & \text{if } |\phi - \phi^0| \leqslant M_4 n^{-\alpha+1/2+\eta}, \\ u(x,y;\phi^0) & \text{otherwise,} \end{cases}$$

where $M_4 > 0$ is chosen small enough that the sphere centered at ϕ^0 with radius M_4 is contained in $\mathscr{U}$. Since $n^\beta(\hat{\phi}_n - \phi^0) \to 0$ $\mathbb{P}_{\phi^0}$-a.s. on Ω^* for all $\beta < \alpha - \frac{1}{2}$, we have that $u^{(n)}(\cdot; \hat{\phi}_n) = u(\cdot; \hat{\phi}_n)$ eventually on Ω^*, whence it is sufficient to prove (20) with $\tilde{\xi}_{k,i}$ replaced by $\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k) - \mathbb{E}_{\phi^0}[\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k)|\mathscr{F}_k]$. Moreover, by Kronecker's lemma and Theorem 2.15 in Hall and Heyde (1980), it follows that

$$\begin{aligned}&\sum_{k=1}^{n}\left\{\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k) - \mathbb{E}_{\phi^0}[\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k)|\mathscr{F}_k]\right\}\\ &\quad= \sum_{k=1}^{n}\left\{\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0) - \mathbb{E}_{\phi^0}[\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)|\mathscr{F}_k]\right\} + o(n^{\alpha/2})\end{aligned} \tag{22}$$

holds if

$$\begin{aligned}&\sum_{k=1}^{\infty}\mathbb{E}_{\phi^0}\left[\left\{\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k) - \tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)\right.\right.\\ &\quad\left.\left.\left.-\mathbb{E}_{\phi^0}[\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k)|\mathscr{F}_k] + \mathbb{E}_{\phi^0}[\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)|\mathscr{F}_k]\right\}^2 \right| \mathscr{F}_k\right] \bigg/ k^\alpha < \infty.\end{aligned} \tag{23}$$

This sum of conditional variances is bounded by the corresponding sum of conditional second moments, and using the fact that $\tilde{u}$ is continuously differentiable in ϕ at ϕ^0, we can bound (23) by

$$\begin{aligned}&\sum_{k=1}^{\infty}\mathbb{E}_{\phi^0}\left[\left|\tilde{u}_i^{(k)}(X_{k+1}, Y_{k+1}; \hat{\phi}_k) - \tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)\right|^2 \middle| \mathscr{F}_k\right] \bigg/ k^\alpha\\ &\quad\leqslant \sum_{k=1}^{\infty} M_4^2 k^{-2\alpha+1+2\eta}\mathbb{E}_{\phi^0}\left[\sup_{\phi\in\mathscr{U}}\left|\frac{\partial}{\partial\phi}\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi)\right|^2 \middle| \mathscr{F}_k\right] \bigg/ k^\alpha.\end{aligned}$$

The expectation on the right-hand side above is finite by A4 and (6), and thus the right-hand side is bounded by a constant times $\sum_1^\infty k^{-3\alpha+1+2\eta}$. Since $\alpha > \frac{2}{3}$, this sum is finite if η is chosen small enough, and thus (22) is true. Consequently, it is sufficient to prove (20) with $\tilde{\xi}_{k,i}$ replaced by $\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0) - \mathbb{E}_{\phi^0}[\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)|\mathscr{F}_k]$, and this is the final step of the proof.

First note that $\mathbb{E}_{\phi^0}[\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)|\mathscr{F}_k] = \mathbb{E}_{\phi^0}[\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)|X_k]$. By an argument entirely similar to that in Lindgren (1978, p. 87), the mixing coefficients of the bivariate process $\{(X_k, Y_k)\}$ are bounded by four times the mixing coefficients of $\{X_k\}$. The latter are geometrically decaying, see Ibragimov and Linnik (1971, p. 366), so that $\{(X_k, Y_k)\}$ has geometrically decaying mixing coefficients as well. Hence by A8 and Theorem 4.1 in Qiman and Chuanrong (1987), (20) indeed holds if $\tilde{\xi}_{k,i}$ is replaced by $\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0) - \mathbb{E}_{\phi^0}[\tilde{u}_i(X_{k+1}, Y_{k+1}; \phi^0)|X_k]$, and the proof is complete. □

Acknowledgements

I would like to thank Dr. Rainer Schwabe at the Freie Universität, Berlin, for an enlightening discussion on averaged stochastic approximation and a critical reading of an earlier draft of this paper. I would also like to thank the referees for some constructive comments that improved the presentation of the paper.

References

S. Asmussen, Markov chains and related topics. A Short Second Course, Inst. for Electronic Systems, Aalborg Univ. (Aalborg, 1994).

L.E. Baum and T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Statist. 37 (1966) 1554–1563.

P.J. Bickel and Y. Ritov, Inference in hidden Markov models I: Local asymptotic normality in the stationary case, Bernoulli 2 (1996) 199–228.

V. Fabian, On asymptotic normality in stochastic approximation, Ann. Math. Statist. 39 (1968) 1327–1332.

V. Fabian, On asymptotically efficient recursive estimation, Ann. Statist. 6 (1978) 854–866.

D.R. Fredkin and J.A. Rice, Maximum likelihood estimation and identification directly from single-channel recordings, Proc. Roy. Soc. London B 249 (1992) 125–132.

P.W. Glynn, Poisson's equation for the recurrent M/G/1 queue, Adv. Appl. Probab. 26 (1994) 1044–1062.

P. Hall and C.C. Heyde, Martingale Limit Theory and its Applications (Academic Press, New York, 1980).

U. Holst and G. Lindgren, Recursive estimation in mixture models with Markov regime, IEEE Trans. Inform. Theory 37 (1991) 1683–1690.

I.A. Ibragimov and Y.V. Linnik, Independent and Stationary Sequences of Random Variables (Wolters-Noordhoff, Groningen, 1971).

V. Krishnamurthy and J.B. Moore, On-line estimation of hidden Markov model parameters based on the Kullback–Leibler information measure, IEEE Trans. Signal Process. 41 (1993) 2557–2573.

H.J. Kushner and D.S. Clark, Stochastic Approximation Methods for Constrained and Unconstrained Systems (Springer, New York, 1978).

B.G. Leroux, Maximum-likelihood estimation for hidden Markov models, Stochastic. Process. Appl. 40 (1992) 127–143.

B.G. Leroux and M.L. Puterman, Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models, Biometrics 48 (1992) 545–558.

G. Lindgren, Markov regime models for mixed distributions and switching regressions, Scand. J. Statist. 5 (1978) 81–91.

L. Ljung, Strong convergence of a stochastic approximation algorithm, Ann. Statist. 6 (1978) 680–696.

D.-J. Ma, A.M. Makowski, and A. Shwartz, Stochastic approximations for finite-state Markov chains, Stochastic. Process. Appl. 35 (1990) 27–45.

T. Petrie, Probabilistic functions of finite state Markov chains, Ann. Math. Statist. 40 (1969) 97–115.

W. Philipp and W. Stout, Almost sure invariance principles for partial sums of weakly dependent random variables, Mem. Amer. Math. Soc. 161 (1975).

B.T. Polyak, New method of stochastic approximation type, Autom. Remote Control 51 (1990) 937–946.

S. Qiman and L. Chuanrong, Strong approximations for partial sums of weakly dependent random variables, Scienta Sinica A 30 (1987) 575–587.

L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–284.

H. Rootzén, Central limit theory for martingales via random change of time, in: A. Gut and L. Holst, eds., Probability and Mathematical Statistics. Essays in Honour of Carl-Gustav Esseen (Uppsala University, Uppsala, 1983).

D. Ruppert, Almost sure approximations to the Robbins–Monro and Kiefer–Wolfowitz processes with dependent noise, Ann. Probab. 10 (1982) 178–187.

D. Ruppert, Efficient estimators from a slowly convergent Robbins–Monro process, Tech. Rept. No. 781, School of Operations Research and Industrial Engineering, Cornell Univ. (Ithaca, NY, 1988).

D. Ruppert, Stochastic approximation, in: B.K. Ghosh and P.K. Sen, eds., Handbook of Sequential Analysis (Marcel Dekker, New York, 1991).

T. Rydén, Consistent and asymptotically normal parameter estimates for hidden Markov models, Ann. Statist. 22 (1994) 1884–1895.

T. Rydén, Estimating the order of hidden Markov models, Statistics 26 (1995) 345–354.

T. Rydén, Asymptotically efficient recursive estimation for incomplete data models using the observed information, Tech. Rept., Dept. of Math. Statist., Lund Univ. (Lund, 1996).

R. Schwabe, Strong representation of an adaptive stochastic approximation procedure, Stochastic. Process. Appl. 23 (1986) 115–130.

R. Schwabe, Stability results for smoothed stochastic approximation procedures, Z. angew. Math. Mech. 73 (1993) T639–T643.

R. Schwabe and H. Walk, On a stochastic approximation procedure based on averaging, Metrika 44 (1996) 165–180.