

Recursive Identification of HMM's with Observations in a Finite Set*

François LeGland
IRISA / INRIA
Campus de Beaulieu
35042 Rennes Cédex, France
legland@irisa.fr

Laurent Mével
IRMAR
Campus de Beaulieu
35042 Rennes Cédex, France
lmevel@irisa.fr

Abstract : We consider the problem of identification of a partially observed finite-state Markov chain, based on observations in a finite set. We first investigate the asymptotic behaviour of the maximum likelihood estimate (MLE) for the transition probabilities, as the number of observations increases to infinity. In particular, we exhibit the associated contrast function, and we discuss consistency issues. Based on this expression, we design a recursive identification algorithm, which converges to the set of local minima of the contrast function.

Keywords : hidden Markov models, maximum likelihood estimate, recursive identification

1 INTRODUCTION

In this paper, we consider the problem of recursive identification of a partially observed finite-state Markov chain, based on observations in a finite set. In a first part, we investigate the asymptotic behaviour of the maximum likelihood estimate (MLE) for the transition probabilities, as the number of observations increases to infinity. The problem has already been considered by Petrie [6], and our main contribution is to exhibit a convenient expression for the associated contrast function. In a second part, we propose a recursive identification algorithm, based on the expression obtained for the contrast function. Similar algorithms have been already considered, based rather on the recursive minimization of the prediction error :

- An extensive motivation for the algorithms and interesting discussions of implementation issues can be found in Krishnamurthy and Moore [5] and Collings, Krishnamurthy and Moore [3], in the case where the observation is a function of the state in additive Gaussian white noise.
- A complete mathematical analysis of the algorithm can be found in Arapostathis and Marcus [1], in a very special case of observations in a

finite set, where the prediction error can still be defined.

In this paper, we try and prove results similar to those of [1] in the general case of observations in a finite set, and in fact many of our intermediate results are borrowed from [1]. Our main result is that the proposed recursive identification algorithm converges to the set of local minima of the contrast function.

The statistical model is as follows. Let $\{X_n, n \geq 0\}$ and $\{Y_n, n \geq 1\}$ be two sequences, with values in the finite sets $S = \{1, \dots, N\}$ and $O = \{1, \dots, M\}$ respectively. On the corresponding canonical space, a family $(P^\theta, \theta \in \Theta)$ of probability measures is considered, with $\Theta \subset \mathbf{R}^p$ compact, such that under P^θ :

- The unobserved state sequence $\{X_n, n \geq 0\}$ is a Markov chain with transition probability matrix $Q_\theta = (q_\theta^{ij})$, i.e. for any $i, j \in S$

$$q_\theta^{ij} = P^\theta[X_{n+1} = j \mid X_n = i],$$

and initial probability distribution $p_0 = (p_0^i)$ independent of $\theta \in \Theta$, i.e. for any $i \in S$

$$p_0^i = P^\theta[X_0 = i].$$

- The observations $\{Y_n, n \geq 1\}$ are mutually independent given the sequence of states of the Markov chain, i.e.

$$\begin{aligned} P^\theta[Y_n = \ell_n, \dots, Y_1 = \ell_1 \mid X_n = i_n, \dots, X_1 = i_1] &= \\ &= \prod_{k=1}^n P^\theta[Y_k = \ell_k \mid X_k = i_k]. \end{aligned}$$

For simplicity, the transition probabilities

$$b_i^\ell = P^\theta[Y_n = \ell \mid X_n = i],$$

are independent of the parameter $\theta \in \Theta$.

Notations. Let $\langle \cdot, \cdot \rangle$ denote the scalar product in \mathbf{R}^N . For any $\ell \in O$, let

$$b^\ell = (b_1^\ell, \dots, b_N^\ell)^* \quad \text{and} \quad B^\ell = \text{diag}(b^\ell).$$

*This work was partially supported by the Commission of the European Communities, under the SCIENCE project *System Identification*, project number SCI*-CT92-0779, and under the HCM project *Statistical Inference for Stochastic Processes*, contract number CHRX-CT92-0078, and by the Army Research Office, under grant DAAH04-95-1-0164.

For any $n \geq 1$, let

$$b(Y_n) = \sum_{\ell \in O} b^\ell \mathbf{1}_{[Y_n = \ell]},$$

and

$$B(Y_n) = \sum_{\ell \in O} B^\ell \mathbf{1}_{[Y_n = \ell]} = \text{diag}(b(Y_n)).$$

Let $e = (1, \dots, 1)^*$ denote the N -dimensional vector with all components equal to 1, and notice that $B^\ell e = b^\ell$ for any $\ell \in O$, hence $B(Y_n) e = b(Y_n)$ for any $n \geq 1$.

Throughout the paper, the *true* value of the parameter will be denoted by $\alpha \in \Theta$, and we make the following assumptions :

Assumption A : For the *true* value $\alpha \in \Theta$, the transition probability matrix $Q_\alpha = (q_\alpha^{ij})$ is an irreducible and aperiodic stochastic matrix.

Assumption B : For all $i \in S$, $\ell \in O$, $b_i^\ell > 0$ and we define

$$\delta_B = \frac{\max_{i \in S, \ell \in O} b_i^\ell}{\min_{i \in S, \ell \in O} b_i^\ell} < \infty. \quad (1)$$

Model parametrization. The parameter θ to be estimated is the set of transition probabilities of the Markov chain $\{X_n, n \geq 0\}$. We parametrize any $N \times N$ stochastic matrix by the collection of its off-diagonal entries. As a result, the parameter space is a closed convex subset of \mathbf{R}^p , $p = N(N-1)$, defined as the N -fold product $\Theta = \Delta_0 \times \dots \times \Delta_0 \subset \mathbf{R}^p$, where $\Delta_0 \subset \mathbf{R}^{N-1}$ is the $(N-1)$ -simplex

$$\Delta_0 = \{u \in \mathbf{R}^{N-1} : \sum_{i=1}^{N-1} u_i \leq 1,$$

$$\text{and } u_i \geq 0, \text{ for all } i = 1, \dots, N-1\}.$$

Let $\theta = (\theta^1, \dots, \theta^N) \in \Theta$: then, for all $i = 1, \dots, N$

$$q_\theta^{ij} = \begin{cases} \theta_j^i, & \text{if } j = 1, \dots, i-1 \\ 1 - \sum_{j=1}^{N-1} \theta_j^i, & \text{if } j = i \\ \theta_{j-1}^i, & \text{if } j = i+1, \dots, N. \end{cases}$$

On the parameter space Θ , we define the following distance :

$$\|\theta - \theta'\| \triangleq \max_{i \in S} \left\{ \sum_{j \in S} |q_\theta^{ij} - q_{\theta'}^{ij}| \right\}.$$

As in [1], we consider for each $\varepsilon > 0$ the set Θ_ε of stochastic matrices with all entries larger than ε , defined as the N -fold product $\Theta_\varepsilon = \Delta_\varepsilon \times \dots \times \Delta_\varepsilon$ where $\Delta_\varepsilon \subset \Delta_0$ is

$$\Delta_\varepsilon = \{u \in \mathbf{R}^{N-1} : \sum_{i=1}^{N-1} u_i \leq 1 - \varepsilon,$$

$$\text{and } u_i \geq \varepsilon, \text{ for all } i = 1, \dots, N-1\}.$$

The sets $\Delta_\varepsilon \subset \mathbf{R}^{N-1}$ are closed convex polytopes, and we define $\pi_\varepsilon(u)$ as the *unique* closest point to $u \in \mathbf{R}^{N-1}$ in Δ_ε . We consider also the set Θ^+ of stochastic matrices with positive entries, defined as the N -fold product $\Theta^+ = \Delta^+ \times \dots \times \Delta^+$ where $\Delta^+ \subset \Delta_0$ is

$$\Delta^+ = \bigcup_{\varepsilon > 0} \Delta_\varepsilon = \{u \in \mathbf{R}^{N-1} : \sum_{i=1}^{N-1} u_i < 1,$$

$$\text{and } u_i > 0, \text{ for all } i = 1, \dots, N-1\}.$$

Instead of Assumption A, we make the stronger assumption :

Assumption A' : For the *true* value $\alpha \in \Theta$, the transition probability matrix $Q_\alpha = (q_\alpha^{ij})$ has positive entries. (In other words, $\alpha \in \Theta^+$, i.e. $\alpha \in \Theta_\varepsilon$ for some unknown $\varepsilon > 0$).

2 MAXIMUM LIKELIHOOD ESTIMATE

Let the probability distribution of (Y_1, \dots, Y_n) under P^θ be denoted as

$$p^\theta\{\ell_n, \dots, \ell_1\} = P^\theta[Y_n = \ell_n, \dots, Y_1 = \ell_1].$$

By definition, the log-likelihood function (suitably normalized) for the estimation of the unknown parameter θ based on observations (Y_1, \dots, Y_n) is

$$\ell_n(\theta) = \frac{1}{n} \log p^\theta[Y_n, \dots, Y_1],$$

and the maximum likelihood estimate (MLE) satisfies

$$\widehat{\theta}_n^{\text{MLE}} \in \underset{\theta \in \Theta_{\varepsilon_0}}{\text{argmax}} \ell_n(\theta), \quad (2)$$

for some $\varepsilon_0 > 0$.

In order to derive convenient equivalent expressions for $\ell_n(\theta)$, we introduce the following standard framework for state estimation in HMM's.

Let $\mathcal{Y}_n = \sigma(Y_1, \dots, Y_n)$ denote the σ -algebra generated by the observations, and let $p_n(\theta)$ denote the *prediction* probability distribution under P^θ of the state X_n given \mathcal{Y}_{n-1} , i.e.

$$p_n^i(\theta) = P^\theta[X_n = i | \mathcal{Y}_{n-1}],$$

for any $i \in S$. The sequence $\{p_n(\theta), n \geq 0\}$ satisfies

$$p_{n+1}(\theta) = \frac{Q_\theta^* B(Y_n) p_n(\theta)}{\langle b(Y_n), p_n(\theta) \rangle}, \quad (3)$$

with initial condition $p_0(\theta) = p_0$ independent of $\theta \in \Theta$. Notice that

$$\begin{aligned} \langle e, Q_\theta^* B(Y_n) p_n(\theta) \rangle &= \langle B(Y_n) Q_\theta e, p_n(\theta) \rangle \\ &= \langle b(Y_n), p_n(\theta) \rangle, \end{aligned}$$

hence the probability distribution defined by the right-hand side of (3) is properly normalized. Let $\mathcal{P}(S)$ denote the set of probability distributions on S .

Notations. For any $\theta \in \Theta$, any $\ell \in O$, and any $p \in \mathcal{P}(S)$, let

$$f_{\theta}^{\ell}(p) = \frac{Q_{\theta}^{*} B^{\ell} p}{\langle b^{\ell}, p \rangle},$$

hence for any $n \geq 1$

$$f_{\theta}[Y_n, p] = \sum_{\ell \in O} f_{\theta}^{\ell}(p) \mathbf{1}_{[Y_n = \ell]} = \frac{Q_{\theta}^{*} B(Y_n) p}{\langle b(Y_n), p \rangle},$$

so that equation (3) reads also

$$p_{n+1}(\theta) = f_{\theta}[Y_n, p_n(\theta)].$$

For any parameter sequence $(\theta_1, \dots, \theta_n) \in \Theta$, any observation sequence $(\ell_1, \dots, \ell_n) \in O$, and any initial condition $p_0 \in \mathcal{P}(S)$, we consider the sequence $\{p_n, n \geq 1\}$ with

$$p_{n+1} = f_{\theta_n}^{\ell_n}(p_n),$$

and we introduce the notation

$$p_n = p_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(p_0),$$

so as to express the dependency upon the parameter sequence, the sequence of observations, and the initial condition.

We can now provide the following expression for the log-likelihood function :

Proposition 2.1 For any $\theta \in \Theta$

$$\ell_n(\theta) = \frac{1}{n} \sum_{k=1}^n \log \langle b(Y_k), p_k(\theta) \rangle.$$

PROOF. By successive conditioning

$$p^{\theta}[Y_n, \dots, Y_1] = \prod_{k=1}^n p^{\theta}[Y_k | \mathcal{Y}_{k-1}].$$

On the other hand, under the mutual independence condition

$$\begin{aligned} P^{\theta}[Y_k = \ell | \mathcal{Y}_{k-1}] &= \\ &= \sum_{i \in S} P^{\theta}[Y_k = \ell | X_k = i] P^{\theta}[X_k = i | \mathcal{Y}_{k-1}] \\ &= \sum_{i \in S} b_i^{\ell} p_k^i(\theta) = \langle b^{\ell}, p_k(\theta) \rangle, \end{aligned} \quad (4)$$

hence

$$p^{\theta}[Y_k | \mathcal{Y}_{k-1}] = \langle b(Y_k), p_k(\theta) \rangle,$$

which concludes the proof. \square

For any $\theta \in \Theta$, we define $W_n(\theta) = (X_n, Y_n, p_n(\theta))$, and we consider the extended sequence $\{W_n(\theta), n \geq 1\}$, which is a Markov chain under P^{α} , with values in $S \times O \times \mathcal{P}(S)$.

The following key estimate can be proved exactly as in [1, Corollary 2.1] :

Lemma 2.2 Under Assumption B, define for any $\varepsilon \in (0, \frac{1}{2})$

$$\delta_{\varepsilon} = \delta_B (1 - \varepsilon) / \varepsilon, \quad r_{\varepsilon} = 1 - \delta_{\varepsilon}^{-3}, \quad K_{\varepsilon} = N \delta_{\varepsilon}.$$

Then, for any $(\theta_1, \dots, \theta_n) \in \Theta_{\varepsilon}$, any $(\ell_1, \dots, \ell_n) \in O$, and any $p_0, p'_0 \in \mathcal{P}(S)$

$$\|p_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(p_0) - p_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(p'_0)\|_1 \leq K_{\varepsilon} r_{\varepsilon}^n \|p_0 - p'_0\|_1.$$

From this, we can prove the following geometric ergodicity property :

Proposition 2.3 Under Assumption B, for any $\theta \in \Theta^{+}$ the Markov chain $\{W_n(\theta), n \geq 1\}$ has a unique invariant probability measure under P^{α} , which is a probability distribution $\mu^{\theta, \alpha}$ on $S \times O \times \mathcal{P}(S)$.

If $\theta \in \Theta_{\varepsilon}$, then for any real-valued measurable function $g = (g_{i, \ell})$ defined on $S \times O \times \mathcal{P}(S)$, such that the coordinate functions $p \mapsto g_{i, \ell}(p)$ are locally Lipschitz continuous on $\mathcal{P}(S)$, there exist constants $C_{\varepsilon} > 0$ and $0 < \rho_{\varepsilon} < 1$ such that

$$|\mathbf{E}^{\alpha}[g(W_n(\theta))] - \mu^{\theta, \alpha}(g)| \leq C_{\varepsilon} \rho_{\varepsilon}^n. \quad (5)$$

In addition, the following strong law of large numbers holds :

Proposition 2.4 For any $\theta \in \Theta^{+}$

$$\ell_n(\theta) \longrightarrow \ell(\theta, \alpha), \quad P^{\alpha}\text{-a.s.}$$

as $n \rightarrow \infty$, where

$$\ell(\theta, \alpha) = \sum_{\ell \in O} \int_{\mathcal{P}(S)} \log \langle b^{\ell}, p \rangle \nu_{\ell}^{\theta, \alpha}(dp),$$

and $\nu^{\theta, \alpha}$ denotes the marginal of $\mu^{\theta, \alpha}$ on $O \times \mathcal{P}(S)$.

For any $\theta \in \Theta^{+}$, define the contrast function

$$K(\theta, \alpha) = -[\ell(\theta, \alpha) - \ell(\alpha, \alpha)], \quad (6)$$

which is nonnegative. Define also

$$\begin{aligned} M(\alpha) &= \operatorname{argmin}_{\theta \in \Theta^{+}} K(\theta, \alpha) \\ &= \{\theta \in \Theta^{+} : K(\theta, \alpha) = 0\} \supseteq \{\alpha\}. \end{aligned}$$

In order to characterize the set $M(\alpha)$, we introduce the following assumption, which turns out to be a sufficient condition for identifiability :

Assumption C : There exist $\ell \in O$ such that

$$b_i^{\ell} = b_j^{\ell} \quad \text{if and only if} \quad i = j.$$

Indeed, it follows from [6, Theorem 1.3] that $M(\alpha) = \{\alpha\}$ under Assumptions A', B and C.

Finally, the following consistency result holds :

Theorem 2.5 Assume that $\varepsilon_0 > 0$ in the definition (2) of the MLE is such that, for some $\varepsilon > \varepsilon_0$

$$\alpha \in \Theta_\varepsilon \subset \Theta_{\varepsilon_0}.$$

Then, under Assumptions B and C, any MLE sequence $\{\hat{\theta}_n^{\text{MLE}}, n \geq 1\}$ converges P^α -a.s. to the true value α as $n \rightarrow \infty$.

In the next section, we design a recursive identification algorithm, based on the expression of the contrast function obtained in Proposition 2.4.

3 RECURSIVE IDENTIFICATION ALGORITHM

Recall that the log-likelihood function is defined by

$$\ell_n(\theta) = \frac{1}{n} \sum_{k=1}^n \log \langle b(Y_k), p_k(\theta) \rangle,$$

for any $\theta \in \Theta$. Its gradient w.r.t. the parameter $\theta = (\theta^1, \dots, \theta^N)$, or score function, is defined by

$$S_n^i(\theta) = \frac{\partial}{\partial \theta^i} \ell_n(\theta) = \frac{1}{n} \sum_{k=1}^n \frac{e^* B(Y_k) \zeta_k^i(\theta)}{\langle b(Y_k), p_k(\theta) \rangle},$$

where

$$\zeta_n^i(\theta) = \frac{\partial}{\partial \theta^i} p_n(\theta),$$

for any $\theta \in \Theta$, and any $i = 1, \dots, N$. The sequence $\{\zeta_n^i(\theta), n \geq 1\}$ satisfies

$$\begin{aligned} \zeta_{n+1}^i(\theta) &= \\ &= \frac{Q_\theta^* B(Y_n)}{\langle b(Y_n), p_n(\theta) \rangle} \left[I - \frac{p_n(\theta) \otimes b(Y_n)}{\langle b(Y_n), p_n(\theta) \rangle} \right] \zeta_n^i(\theta) \\ &\quad + R_i \frac{b_i(Y_n) p_n^i(\theta)}{\langle b(Y_n), p_n(\theta) \rangle}, \end{aligned} \quad (7)$$

where the $N \times (N-1)$ matrix

$$R_i = \left(\frac{\partial q_\theta^{ij}}{\partial \theta^j} \right)_{jj'},$$

can be computed explicitly, and does not depend on θ .

Let $\Sigma = \mathbf{R}^{N \times p}$, with $p = N(N-1)$. For any $\theta \in \Theta$, we define $u_n(\theta) = (p_n(\theta), \zeta_n^1(\theta), \dots, \zeta_n^N(\theta)) \in \mathcal{P}(S) \times \Sigma$.

Notations. For any $\theta \in \Theta$, any $\ell \in O$, and any $p \in \mathcal{P}(S)$, recall that

$$f_\theta^\ell(p) = \frac{Q_\theta^* B^\ell p}{\langle b^\ell, p \rangle},$$

and let

$$\Phi_\theta^\ell(p) = \frac{Q_\theta^* B^\ell}{\langle b^\ell, p \rangle} \left[I - \frac{p \otimes b^\ell}{\langle b^\ell, p \rangle} \right], \quad r_i^\ell(p) = R_i \frac{b_i^\ell p^i}{\langle b^\ell, p \rangle}.$$

For any $\theta \in \Theta$, any $\ell \in O$, and any $u = (p, \zeta^1, \dots, \zeta^N) \in \mathcal{P}(S) \times \Sigma$, let also

$$H_i^\ell(u) = \frac{e^* B^\ell \zeta^i}{\langle b^\ell, p \rangle}. \quad (8)$$

Hence, for any $n \geq 1$

$$\begin{aligned} \Phi_\theta[Y_n, p] &= \sum_{\ell \in O} \Phi_\theta^\ell(p) \mathbf{1}_{[Y_n = \ell]} \\ &= \frac{Q_\theta^* B(Y_n)}{\langle b(Y_n), p \rangle} \left[I - \frac{p \otimes b(Y_n)}{\langle b(Y_n), p \rangle} \right], \end{aligned}$$

$$r_i[Y_n, p] = \sum_{\ell \in O} r_i^\ell(p) \mathbf{1}_{[Y_n = \ell]} = R_i \frac{b_i(Y_n) p^i}{\langle b(Y_n), p \rangle},$$

and

$$H_i[Y_n, u] = \sum_{\ell \in O} H_i^\ell(u) \mathbf{1}_{[Y_n = \ell]} = \frac{e^* B(Y_n) \zeta^i}{\langle b(Y_n), p \rangle}, \quad (9)$$

so that equations (3) and (7) read also

$$p_{n+1}(\theta) = f_\theta[Y_n, p_n(\theta)],$$

$$\zeta_{n+1}^i(\theta) = \Phi_\theta[Y_n, p_n(\theta)] \zeta_n^i(\theta) + r_i[Y_n, p_n(\theta)],$$

and the score function reads also

$$S_n^i(\theta) = \frac{1}{n} \sum_{k=1}^n H_i[Y_k, u_k(\theta)].$$

More generally, for any parameter sequence $(\theta_1, \dots, \theta_n) \in \Theta$, any observation sequence $(\ell_1, \dots, \ell_n) \in O$, and any initial condition $u_0 \in \mathcal{P}(S) \times \Sigma$, we consider the sequence $\{u_n, n \geq 1\}$ with $u_n = (p_n, \zeta_n^1, \dots, \zeta_n^N)$ and

$$p_{n+1} = f_{\theta_n}^{\ell_n}(p_n),$$

$$\zeta_{n+1}^i = \Phi_{\theta_n}^{\ell_n}(p_n) \zeta_n^i + r_i^{\ell_n}(p_n),$$

for all $i = 1, \dots, N$, and we introduce the notation

$$u_n = u_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(u_0),$$

so as to express the dependency of the sequence $\{u_n, n \geq 1\}$ upon the parameter sequence, the sequence of observations, and the initial condition.

The algorithm is defined as follows : For all $i = 1, \dots, N$

$$\hat{\theta}_{n+1}^i = \pi_{\varepsilon_0}(\hat{\theta}_n^i + \gamma_{n+1} H_i[Y_n, \hat{u}_n]), \quad (10)$$

where $\gamma_n = 1/n$, π_{ε_0} denotes the projection on the convex polytope Δ_{ε_0} for some $\varepsilon_0 > 0$, and the sequence $\{\hat{u}_n, n \geq 1\}$ is defined by $\hat{u}_n = (\hat{p}_n, \hat{\zeta}_n^1, \dots, \hat{\zeta}_n^N)$, and

$$\hat{p}_{n+1} = f_{\hat{\theta}_{n+1}}[Y_n, \hat{p}_n],$$

$$\hat{\zeta}_{n+1}^i = \Phi_{\hat{\theta}_{n+1}}[Y_n, \hat{p}_n] \hat{\zeta}_n^i + r_i[Y_n, \hat{p}_n].$$

For any $\theta \in \Theta$, we define $Z_n(\theta) = (X_n, Y_n, u_n(\theta))$, and we consider the extended sequence $\{Z_n(\theta), n \geq 1\}$, which is a Markov chain under P^α , with values in

$S \times O \times \mathcal{P}(S) \times \Sigma$. Let $\Pi_{\theta, \alpha}$ denote the corresponding transition probability matrix/kernel. Then

$$\begin{aligned} P^\alpha[\widehat{Z}_{n+1} \in B \mid \widehat{\theta}_0, \dots, \widehat{\theta}_{n+1}, \widehat{Z}_0, \dots, \widehat{Z}_n] &= \\ &= \Pi_{\widehat{\theta}_{n+1}, \alpha}(\widehat{Z}_n, B), \end{aligned}$$

i.e. the algorithm (10) belongs to the class of stochastic algorithms with Markovian dynamics, see Benveniste, Métivier and Priouret [2].

The first step in the study of the algorithm (10) is to study the Markov chain $\{Z_n(\theta), n \geq 1\}$, for a fixed value $\theta \in \Theta$. The following compactness result has been proved in [1, Proposition 2.2].

Lemma 3.1 *Under Assumption B, there exist a compact neighbourhood $U_\varepsilon \subset \Sigma$ of the origin, constants $K_\varepsilon > 0$ and $0 < r_\varepsilon < 1$ such that :*

- (i) *For any $(\theta_1, \dots, \theta_n) \in \Theta_\varepsilon$, any $(\ell_1, \dots, \ell_n) \in O$, and any $u_0, u'_0 \in \mathcal{P}(S) \times U_\varepsilon$*

$$u_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(u_0) \in \mathcal{P}(S) \times U_\varepsilon,$$

and

$$\|u_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(u_0) - u_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(u'_0)\|_1 \leq K_\varepsilon r_\varepsilon^n \|u_0 - u'_0\|_1,$$

- (ii) *For each $M > 0$, there exists an integer $n_0 = n_0(M)$ such that for any $(\theta_1, \dots, \theta_n) \in \Theta_\varepsilon$, any $(\ell_1, \dots, \ell_n) \in O$, and any $u_0 \in \mathcal{P}(S) \times B(0, M)$*

$$u_{\theta_n, \dots, \theta_1}^{\ell_n, \dots, \ell_1}(u_0) \in \mathcal{P}(S) \times U_\varepsilon, \quad \text{for all } n \geq n_0.$$

From this, we can prove the following geometric ergodicity property, as in [1, Proposition 3.2] :

Proposition 3.2 *Under Assumption B, for any $\theta \in \Theta^+$, the Markov chain $\{Z_n(\theta), n \geq 1\}$ with an initial distribution such that $\zeta_0(\theta)$ is bounded a.s., has a unique invariant probability measure under P^α , which is a probability distribution $\rho^{\theta, \alpha}$ in $S \times O \times \mathcal{P}(S) \times \Sigma$.*

If $\theta \in \Theta_\varepsilon$, then the support of $\rho^{\theta, \alpha}$ is contained in $S \times O \times \mathcal{P}(S) \times U_\varepsilon$, and for any real-valued measurable function $\gamma = (\gamma_{i, \ell})$ defined on $S \times O \times \mathcal{P}(S) \times \Sigma$, such that the coordinate functions $(p, \zeta) \mapsto \gamma_{i, \ell}(p, \zeta)$ are locally Lipschitz continuous on $\mathcal{P}(S) \times \Sigma$, there exist constants $C_\varepsilon > 0$ and $0 < \rho_\varepsilon < 1$ such that

$$|\mathbf{E}^\alpha[\gamma(Z_n(\theta))] - \rho^{\theta, \alpha}(\gamma)| \leq C_\varepsilon \rho_\varepsilon^n.$$

Finally, as in [1, Corollary 3.1]

Proposition 3.3 *Under Assumption B, for any $\theta \in \Theta_\varepsilon$, and any real-valued measurable function $g = (g_{i, \ell})$ defined on $S \times O \times \mathcal{P}(S)$, such that the coordinate functions $p \mapsto g_{i, \ell}(p)$ have Lipschitz continuous derivatives on $\mathcal{P}(S)$, there exist constants $C_\varepsilon > 0$ and $0 < \rho_\varepsilon < 1$ such that, in addition to (5)*

$$\left\| \frac{\partial}{\partial \theta^i} \mathbf{E}^\alpha[g(W_n(\theta))] - \frac{\partial}{\partial \theta^i} \mu^{\theta, \alpha}(g) \right\|_1 \leq C_\varepsilon \rho_\varepsilon^n,$$

for all $i = 1, \dots, N$.

We conclude this section with the following result, which identifies the mean vector field of the algorithm as the opposite of the gradient of contrast function.

Proposition 3.4 *For any $\theta \in \Theta^+$*

$$\mathbf{E}^\alpha[H_i(Z_n(\theta))] \longrightarrow h_i(\theta, \alpha),$$

as $n \rightarrow \infty$, where

$$h_i(\theta, \alpha) = \sum_{\ell \in O} \int_{\mathcal{P}(S) \times \Sigma} \frac{e^* B^\ell \zeta^i}{\langle b^\ell, p \rangle} \lambda_\ell^{\theta, \alpha}(dp, d\zeta),$$

for all $i = 1, \dots, N$, and $\lambda^{\theta, \alpha}$ denotes the marginal of $\rho^{\theta, \alpha}$ on $O \times \mathcal{P}(S) \times \Sigma$.

In addition, for any $\theta \in \Theta^+$

$$h_i(\theta, \alpha) = \frac{\partial}{\partial \theta^i} \ell(\theta, \alpha) = - \frac{\partial}{\partial \theta^i} K(\theta, \alpha),$$

for all $i = 1, \dots, N$.

PROOF. By definition $H_i(Z_n(\theta)) = H_i[Y_n, u_n(\theta)]$. It follows from Propositions 2.3 and 3.2 that

$$\mathbf{E}^\alpha[\log \langle b(Y_n), p_n(\theta) \rangle] \longrightarrow \ell(\theta, \alpha),$$

and

$$\begin{aligned} \mathbf{E}^\alpha[H_i(Z_n(\theta))] &= \\ &= \frac{\partial}{\partial \theta^i} \mathbf{E}^\alpha[\log \langle b(Y_n), p_n(\theta) \rangle] \longrightarrow \frac{\partial}{\partial \theta^i} \ell(\theta, \alpha) = \\ &= - \frac{\partial}{\partial \theta^i} K(\theta, \alpha), \end{aligned}$$

as $n \rightarrow \infty$. On the other hand, it follows from Proposition 3.3 that

$$\mathbf{E}^\alpha[H_i(Z_n(\theta))] = \mathbf{E}^\alpha \left[\frac{e^* B(Y_n) \zeta_n^i(\theta)}{\langle b(Y_n), p_n(\theta) \rangle} \right] \longrightarrow h_i(\theta, \alpha),$$

as $n \rightarrow \infty$, and the proof is complete. \square

For any $\theta \in \Theta^+$, define

$$L(\alpha) = \{\theta \in \Theta^+ : h(\theta, \alpha) = 0\}.$$

We consider the recursive identification algorithm defined in (10), where the initial condition \widehat{p}_0 can be chosen arbitrarily in $\mathcal{P}(S)$, and the initial condition $\widehat{\zeta}_0$ is set to $\widehat{\zeta}_0 = 0 \in U_{\varepsilon_0}$, in such a way that $(\widehat{p}_n, \widehat{\zeta}_n) \in \mathcal{P}(S) \times U_{\varepsilon_0}$ for all $n \geq 0$, according to Lemma 3.1.

The following convergence result holds :

Theorem 3.5 *Assume that $\varepsilon_0 > 0$ in the definition (10) of the recursive algorithm is such that, for some $\varepsilon > \varepsilon_0 > \varepsilon'$*

$$L(\alpha) \subset \Theta_\varepsilon \subset \Theta_{\varepsilon_0} \subset \Theta_{\varepsilon'},$$

and for any $\theta \in \Theta_{\varepsilon'} \setminus \Theta_{\varepsilon_0}$

$$\sum_{i \in S} \langle h_i(\theta, \alpha), \pi_{\varepsilon_0}(\theta^i) - \theta^i \rangle > 0. \quad (11)$$

Then, under Assumption B, the recursive estimate sequence $\{\widehat{\theta}_n, n \geq 1\}$ converges P^α -a.s. to the deterministic set $L(\alpha)$ as $n \rightarrow \infty$.

Remark 3.6 The condition (11) roughly means that outside of Θ_{ε_0} , but close enough, the mean vector field of the algorithm is pointing towards Θ_{ε_0} .

PROOF. We follow Delyon [4]. Let $K(\theta, \alpha) \geq 0$ be defined as in (6). It was proved in Proposition 3.4 that

$$\frac{\partial}{\partial \theta^i} K(\theta, \alpha) = -h_i(\theta, \alpha),$$

for all $i = 1, \dots, N$. Therefore

$$\sum_{i \in S} \left\langle \frac{\partial}{\partial \theta^i} K(\theta, \alpha), h_i(\theta, \alpha) \right\rangle = - \sum_{i \in S} |h_i(\theta, \alpha)|^2 < 0,$$

for any $\theta \notin L(\alpha)$, which is Assumption (A) of [4].

Moreover, it follows from the condition (11) that

$$\begin{aligned} \sum_{i \in S} \left\langle \frac{\partial}{\partial \theta^i} K(\theta, \alpha), \pi_{\varepsilon_0}(\theta^i) - \theta^i \right\rangle &= \\ &= - \sum_{i \in S} \langle h_i(\theta, \alpha), \pi_{\varepsilon_0}(\theta^i) - \theta^i \rangle < 0, \end{aligned}$$

for any $\theta \in \Theta_{\varepsilon'} \setminus \Theta_{\varepsilon_0}$, which is Assumption (Proj) of [4].

Finally, the following decomposition holds for all $i = 1, \dots, N$

$$\begin{aligned} H_i(\widehat{Z}_n) &= H_i[Y_n, \widehat{u}_n] \\ &= h_i(\widehat{\theta}_n, \alpha) + V_{\widehat{\theta}_n, \alpha}^i(\widehat{Z}_n) - \Pi_{\widehat{\theta}_n, \alpha} V_{\widehat{\theta}_n, \alpha}^i(\widehat{Z}_n) \\ &= h_i(\widehat{\theta}_n, \alpha) + V_{\widehat{\theta}_n, \alpha}^i(\widehat{Z}_n) - \Pi_{\widehat{\theta}_n, \alpha} V_{\widehat{\theta}_n, \alpha}^i(\widehat{Z}_{n-1}) \\ &\quad + \Pi_{\widehat{\theta}_n, \alpha} V_{\widehat{\theta}_n, \alpha}^i(\widehat{Z}_{n-1}) - \Pi_{\widehat{\theta}_{n+1}, \alpha} V_{\widehat{\theta}_{n+1}, \alpha}^i(\widehat{Z}_n) \\ &\quad + \Pi_{\widehat{\theta}_{n+1}, \alpha} V_{\widehat{\theta}_{n+1}, \alpha}^i(\widehat{Z}_n) - \Pi_{\widehat{\theta}_n, \alpha} V_{\widehat{\theta}_n, \alpha}^i(\widehat{Z}_n) \\ &= h_i(\widehat{\theta}_n, \alpha) + e_{n,i}^{(1)} + e_{n,i}^{(2)} + e_{n,i}^{(3)}, \end{aligned}$$

where, for any $\theta \in \Theta_{\varepsilon_0}$, the function $V_{\theta, \alpha}^i$ denotes a bounded solution of the Poisson equation associated with the function H_i .

Following [4, Corollary 1], it is then enough to prove that for all $i = 1, \dots, N$

$$\sum_{n=1}^{\infty} \frac{1}{n} [e_{n,i}^{(1)} + e_{n,i}^{(2)} + e_{n,i}^{(3)}] \text{ converges.}$$

The details will appear elsewhere. \square

ACKNOWLEDGEMENT

The authors gratefully acknowledge Jan van Schuppen for bringing the paper [1] to their attention.

4 REFERENCES

- [1] A. ARAPOSTATHIS and S.I. MARCUS. Analysis of an identification algorithm arising in the adaptive estimation of Markov chains. *Mathematics of Control, Signals, and Systems*, 3(1):1–29, 1990.

- [2] A. BENVENISTE, M. MÉTIVIER, and P. PRIOURET. *Adaptive Algorithms and Stochastic Approximations*. Volume 22 of *Applications of Mathematics*, Springer Verlag, New York, 1990.
- [3] I.B. COLLINGS, V. KRISHNAMURTHY, and J.B. MOORE. On-line identification of hidden Markov models via recursive prediction error techniques. *IEEE Transactions on Signal Processing*, SP-42(12):3535–3539, December 1994.
- [4] B. DELYON. *General results on the convergence of stochastic algorithms*. Publication Interne 890, IRISA, December 1994.
- [5] V. KRISHNAMURTHY and J.B. MOORE. On-line estimation of hidden Markov model parameters based on the Kullback–Leibler information measure. *IEEE Transactions on Signal Processing*, SP-41(8):2557–2573, August 1993.
- [6] T. PETRIE. Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969.