Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models*

François LeGland and Laurent Mevel
IRISA / INRIA and IRMAR
Campus de Beaulieu
35042 Rennes Cédex, France
e–mail : {legland,lmevel}@irisa.fr

Abstract

We consider a hidden Markov model with multidimensional observations, and with misspecification, i.e. the *assumed* coefficients (transition probability matrix, and observation conditional densities) are possibly different from the *true* coefficients. Under mild assumptions on the coefficients of both the true and the assumed models, we prove that : (i) the prediction filter forgets almost surely their initial condition exponentially fast, and (ii) the extended Markov chain, whose components are : the unobserved Markov chain, the observation sequence, and the prediction filter, is geometrically ergodic, and has a unique invariant probability distribution.

1 Introduction

Let $\{X_n\,,\,n \geq 0\}$ and $\{Y_n\,,\,n \geq 0\}$ be two random sequences defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P}_\bullet)$, with values in the finite set $S = \{1, \cdots, N\}$ and in $\mathbf{R}^d$ respectively. It is assumed that :

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

$$\mathbf{P}_\bullet[X_{n+1} = j \mid X_n = i] = q_\bullet^{i,j} \;,$$

  and initial probability distribution $p_\bullet = (p_\bullet^i)$, i.e. for any $i \in S$

$$\mathbf{P}_\bullet[X_0 = i] = p_\bullet^i \;.$$

- The observations $\{Y_n\,,\,n \geq 0\}$ are mutually independent given the sequence of states of the Markov chain, i.e. for any integer $n \geq 0$, and for any $i_0, \cdots, i_n \in S$

$$\mathbf{P}_\bullet[Y_n \in dy_n, \cdots, Y_0 \in dy_0 \mid X_n = i_n, \cdots, X_0 = i_0]$$

$$= \prod_{k=0}^{n} \mathbf{P}_\bullet[Y_k \in dy_k \mid X_k = i_k] \;.$$

  For any integer $n \geq 0$, and for any $i \in S$, the conditional probability distribution of the observation Y_n given that $(X_n = i)$ is absolutely continuous with respect to a non–negative and σ–finite measure λ on $\mathbf{R}^d$, i.e.

$$\mathbf{P}_\bullet[Y_n \in dy \mid X_n = i] = b_\bullet^i(y)\;\lambda(dy) \;,$$

  with a λ–a.e. positive density. For any $y \in \mathbf{R}^d$, let

$$b_\bullet(y) = [b_\bullet^1(y), \cdots, b_\bullet^N(y)]^* \;,$$

$$B_\bullet(y) = \mathrm{diag}\,[b_\bullet^1(y), \cdots, b_\bullet^N(y)] \;.$$

  Here and throughout the paper, the notation * denotes the transpose of a matrix.

Example 1.1 [conditionally Gaussian observations] Assume that the observations are of the form

$$Y_n = h_\bullet(X_n) + V_n \;,$$

for any integer $n \geq 0$, where $\{V_n\,,\,n \geq 0\}$ is a Gaussian white noise sequence, with identity covariance matrix. The mapping $h_\bullet$ from S to $\mathbf{R}^d$ is equivalently defined as $h_\bullet = (h_\bullet^i)$ where $h_\bullet^i \in \mathbf{R}^d$ for any $i \in S$. In this case, λ is the Lebesgue measure on $\mathbf{R}^d$, the mutual independence condition is satisfied, and

$$b_\bullet^i(y) = (2\pi)^{-d/2}\;\exp\left\{\,-\tfrac{1}{2}\,|y - h_\bullet^i|^2\,\right\} \;,$$

for any $i \in S$, and any $y \in \mathbf{R}^d$. Here and throughout the paper, the notation $|\cdot|$ denotes the Euclidean norm.

Throughout the paper, we make the following assumption :

*This work was partially supported by the Commission of the European Communities, under the SCIENCE project *System Identification*, project number SC1*–CT92–0779, and under the HCM project *Statistical Inference for Stochastic Processes*, project number CHRX–CT92–0078, and by the Army Research Office, under grant DAAH04–95–1–0164.

Assumption A : The stochastic matrix $Q_\bullet = (q_\bullet^{i,j})$ is primitive (or equivalently, irreducible and aperiodic).

Remark 1.2 Under Assumption A, there exist constants $0 < \rho_\bullet < 1$ and $K_\bullet > 1$ such that, for any integer $n \geq 1$

$$\frac{1}{2} \max_{i,i' \in S} \sum_{j \in S} | q_\bullet^{i,j}(n) - q_\bullet^{i',j}(n) | \leq K_\bullet \, \rho_\bullet^n \ ,$$

and the Markov chain $\{X_n \, , \, n \geq 0\}$ is ergodic, with a unique invariant probability distribution $\mu_\bullet = (\mu_\bullet^i)$ on S. Here and throughout the paper $(q_\bullet^{i,j}(n))$ denote the entries of the stochastic matrix $Q_\bullet^n$, i.e. for any $i, j \in S$

$$\mathbf{P}_\bullet[X_n = j \mid X_0 = i] = q_\bullet^{i,j}(n) \ .$$

For any integer $n \geq 1$, let $p_n^\bullet = (p_n^i)$ denote the *prediction filter*, i.e. the conditional probability distribution under $\mathbf{P}_\bullet$ of the state X_n given observations $(Y_0, \cdots, Y_{n-1})$: for any $i \in S$

$$p_n^i = \mathbf{P}_\bullet[X_n = i \mid Y_0, \cdots, Y_{n-1}] \ .$$

The random sequence $\{p_n^\bullet \, , \, n \geq 0\}$ takes values in the set $\mathcal{P}(S)$ of probability distributions over the finite set S, and satisfies the forward Baum equation

$$p_{n+1}^\bullet = \frac{Q_\bullet^* B_\bullet(Y_n) \, p_n^\bullet}{b_\bullet^*(Y_n) \, p_n^\bullet} \ , \qquad (1)$$

for any integer $n \geq 0$, with initial condition $p_0^\bullet = p_\bullet$.

In practice, the transition probability matrix $Q_\bullet$ and the initial probability distribution $p_\bullet$ of the unobserved Markov chain $\{X_n \, , \, n \geq 0\}$, and the conditional densities $b_\bullet(\cdot)$ of the observation sequence $\{Y_n \, , \, n \geq 0\}$ are possibly unknown. For this reason, we consider instead of (1) the more general equation

$$p_{n+1} = \frac{Q^* B(Y_n) \, p_n}{b^*(Y_n) \, p_n} = f[Y_n, p_n] \ , \qquad (2)$$

for any integer $n \geq 0$, with initial condition $p_0 = p$, where $Q = (q^{i,j})$ is a $N \times N$ stochastic matrix, $p = (p^i)$ is a probability vector on S, and $b(\cdot) = (b^i(\cdot))$ are λ–a.e. positive densities on $\mathbf{R}^d$.

To make explicit the dependency w.r.t. the initial condition and the observations, we introduce the notation

$$p_{n+1} = f[Y_n, \cdots, Y_m, p_m] \ ,$$

for any integers n, m such that $n \geq m$.

Under the *true* probability measure $\mathbf{P}_\bullet$, the extended Markov chain $\{Z_n = (X_n, Y_n, p_n) \, , \, n \geq 0\}$ with values in $S \times \mathbf{R}^d \times \mathcal{P}(S)$, has the following transition probability matrix / kernel

$$\Pi^{i,j}(y, p, dy', dp') \ = \ q_\bullet^{i,j} \, b_\bullet^j(y') \, \lambda(dy') \, \delta_{f[y,p]}(dp') \ .$$

Notice that both

- the initial condition for the prediction filter,
- the transition matrix, and the observation conditional densities,

are possibly unknown. We expect that a wrong initial condition for the prediction filter is rapidly forgotten, so that we could use any initial condition with practically the same effect. On the other hand, we expect that two different transition probability matrices, and two different observation conditional densities will have a significantly different effect, i.e. will produce significantly different values for some related Kullback–Leibler information, so that we could estimate the unknown transition probability matrix and the unknown observation conditional densities accurately, by accumulating observations — this estimation problem is addressed in LeGland and Mevel [10, 11]. Indeed, it can be shown that the log–likelihood function for the estimation of the unknown transition probability matrix and the unknown observation conditional densities, can be easily expressed as an additive functional of the extended Markov chain $\{Z_n = (X_n, Y_n, p_n) \, , \, n \geq 0\}$. Consequently, an asymptotic expression can be obtained for the corresponding Kullback–Leibler information provided some ergodicity property holds.

These statements are made rigorous in this paper, with the proof of the following two properties :

- the exponential forgetting of the initial condition for the sequence $\{p_n \, , \, n \geq 0\}$ defined in (2),
- and the geometric ergodicity of the extended Markov chain $\{Z_n = (X_n, Y_n, p_n) \, , \, n \geq 0\}$.

In the case of conditionally Gaussian observations, the exponential forgetting property has been investigated by Atar and Zeitouni [2] in terms of Lyapunov exponents, and the existence of a unique invariant probability distribution for a related extended Markov chain (based on the filter itself instead of the prediction filter) has been proved by DiMasi and Stettner [4] and by Stettner [14], under some additional restrictive assumption. In this paper, we adopt the approach initiated, in a special case of HMM with observations in a finite set, by Arapostathis and Marcus [1] — see also LeGland and Mevel [8] for a generalization.

Here is a short overview of our results, under the assumption that the stochastic matrix Q is primitive — full details can be found in LeGland and Mevel [9] :

- We obtain in Theorem 2.2 a bound for the difference between the solutions of equation (2) starting from two different initial conditions. As a corollary, we obtain in Proposition 2.3, an upper bound for the $\mathbf{P}_\bullet$–a.s. exponential rate of forgetting of the initial condition for equation (2).
- Using the estimate of Theorem 2.2 we prove in Theorem 3.3, under some integrability assumption on the observation densities $b(\cdot)$ and $b_\bullet(\cdot)$, the geometric ergodicity of the extended Markov chains $\{Z_n = (X_n, Y_n, p_n)\,,\, n \geq 0\}$.

We notice that

$$f_{n,m}[y_n, \cdots, y_m, p] = \frac{Q_n^* B(y_n) \cdots Q_m^* B(y_m)\, p}{e^*\, [Q_n^* B(y_n) \cdots Q_m^* B(y_m)\, p]} = \frac{M_{n,m}\, p}{e^*\, M_{n,m}\, p}\,,$$

if we define

$$M_{n,m} = Q_n^* B(y_n) \cdots Q_m^* B(y_m)\,,$$

and our results will be based on auxiliary estimates for products of non–negative matrices, which improve earlier estimates in Furstenberg and Kesten [5]. See in particular Proposition A.1, which is stated without proof in Appendix A.

Remark 1.3 Most of this work could be generalized easily, under suitable assumptions, to the case where the state space of the Markov chain $\{X_n\,,\, n \geq 0\}$ is compact.

2 Exponential forgetting for the prediction filter

Throughout the paper, $\|\cdot\|$ will denote the L_1–norm, i.e. for any $u = (u^i)$ in $\mathbf{R}^N$

$$\|u\| = \sum_{i \in S} |u^i|\,.$$

The sequence $\{p_n\,,\, n \geq 0\}$ satisfies (2), i.e.

$$p_{n+1} = \frac{Q^* B(Y_n)\, p_n}{b^*(Y_n)\, p_n} = f[Y_n, p_n] = f[Y_n, \cdots, Y_m, p_m]\,,$$

for any integers n, m such that $n \geq m$. For the stochastic matrix $Q = (q^{i,j})$, and for any $y \in \mathbf{R}^d$, we define

$$\delta(y) = \frac{\max\limits_{i \in S} b^i(y)}{\min\limits_{i \in S} b^i(y)} < \infty \quad \text{and} \quad \varepsilon = \min_{i,j \in S}{}^{+}\, q^{i,j} > 0\,,$$

where the notation $\min^+$ denotes the minimum over positive elements.

Proposition 2.1 *If the stochastic matrix Q is positive, then for any $p, p' \in \mathcal{P}(S)$, any integers n, m such that $n \geq m$, and any sequence $y_m, \cdots, y_n \in \mathbf{R}^d$*

$$\|f[y_n, \cdots, y_m, p] - f[y_n, \cdots, y_m, p']\| \leq 2\,\varepsilon^{-1}\,\delta(y_m)\,(1-\varepsilon)^{n-m+1}\,\|p - p'\|\,.$$

For any $p, p' \in \mathcal{P}(S)$ such that $p \neq p'$, and for any infinite sequence $y_m, \cdots, y_n, \cdots \in \mathbf{R}^d$, the difference $\|f[y_n, \cdots, y_m, p] - f[y_n, \cdots, y_m, p']\|$ goes to zero at exponential rate

$$\limsup_{n \to \infty} \frac{1}{n}\, \log \|f[y_n, \cdots, y_m, p] - f[y_n, \cdots, y_m, p']\| \leq \log(1-\varepsilon)\,.$$

Theorem 2.2 *If the stochastic matrix Q is primitive, with index of primitivity r, then for any $p, p' \in \mathcal{P}(S)$, any integers n, m such that $n \geq m + r - 1$, and any sequence $y_m, \cdots, y_n \in \mathbf{R}^d$*

$$\|f[y_n, \cdots, y_m, p] - f[y_n, \cdots, y_m, p']\| \leq 2\,\varepsilon^{-r}\,\delta(y_m) \cdots \delta(y_{m+r-1})\,\|p - p'\| \prod_{\kappa=0}^{[n,m]} (1 - \varepsilon^r\,[\delta(y_{m+\kappa r+1}) \cdots \delta(y_{m+(\kappa+1)r-1})]^{-1})\,,$$

where $[n, m] = \lfloor \frac{n-m+1}{r} \rfloor - 1$.

To obtain an estimate of the almost sure exponential rate of forgetting in this case, we define

$$\Delta_{-1} = \min_{i \in S} \int_{\mathbf{R}^d} \delta^{-1}(y)\, b_\bullet^i(y)\, \lambda(dy)\,.$$

Notice that $0 < \Delta_{-1} \leq 1$, hence for any sequence $i_1, \cdots, i_r \in S$

$$0 \leq \int_{\mathbf{R}^d} \cdots \int_{\mathbf{R}^d} (1 - \varepsilon^r\,[\delta(y_2) \cdots \delta(y_r)]^{-1})\, b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_r}(y_r)\, \lambda(dy_1) \cdots \lambda(dy_r)$$

$$= 1 - \varepsilon^r \prod_{k=2}^{r} \int_{\mathbf{R}^d} \delta^{-1}(y)\, b_\bullet^{i_k}(y)\, \lambda(dy) \leq 1 - \varepsilon^r\, \Delta_{-1}^{r-1} = 1 - R < 1\,,$$

with $R = \varepsilon^r\, \Delta_{-1}^{r-1} > 0$. Notice that when the matrix Q is positive, i.e. when $r = 1$, then R reduces to $R = \varepsilon$.

Proposition 2.3 *If the stochastic matrix Q is primitive, with index of primitivity r, and if Assumption A holds, then for any $p, p' \in \mathcal{P}(S)$ such that $p \neq p'$, and any integer m*

$$\limsup_{n\to\infty} \frac{1}{n} \log \| f[Y_n, \cdots, Y_m, p] - f[Y_n, \cdots, Y_m, p'] \|$$

$$\leq \frac{1}{r} \log(1 - R) , \qquad \mathbf{P}_\bullet\text{-}a.s.$$

where $R = \varepsilon^r \, \Delta_{-1}^{r-1}$.

Remark 2.4 In Atar and Zeitouni [2, Corollary 2.1] the exact exponential rate of forgetting is expressed, in the case of an HMM with conditionally Gaussian observations, without misspecification, as the difference between the two top Lyapunov exponents of the equation (1). Since no explicit expression is available in general for these Lyapunov exponents, estimates for the exponential rate are given in Theorems 1.3 and 1.4 of [2] when the signal–to–noise ratio is large or small respectively.

In this paper, we not only obtain explicit estimates for the exponential rate of forgetting, but we also obtain non–logarithmic and non–asymptotic bounds, in the more general case of a misspecified HMM with arbitrary observation conditional densities, and with primitive transition probability matrices (for both the true and the assumed models). Our proof of the geometric ergodicity of the extended Markov chain $\{Z_n = (X_n, Y_n, p_n) , n \geq 0\}$ is based on these explicit bounds.

3 Geometric ergodicity of the Markov chain $\{X_n, Y_n, p_n\}$

Under the probability measure $\mathbf{P}_\bullet$ corresponding to the *true* transition probability matrix $Q_\bullet$ and the *true* observation densities $b_\bullet(\cdot)$, the extended Markov chain $\{Z_n = (X_n, Y_n, p_n) , n \geq 0\}$ has the following transition probability matrix / kernel

$$\Pi^{i,j}(y, p, dy', dp') = q_\bullet^{i,j} \, b_\bullet^j(y') \, \lambda(dy') \, \delta_{f[y,p]}(dp') .$$

For any real–valued function g defined on $S \times \mathbf{R}^d \times \mathcal{P}(S)$, which is equivalently defined as a collection $g = (g^i)$ of real–valued functions defined on $\mathbf{R}^d \times \mathcal{P}(S)$, we have

$$(\Pi\, g)^i(y, p) =$$

$$= \mathbf{E}_\bullet[g(X_{n+1}, Y_{n+1}, p_{n+1}) \mid X_n = i, Y_n = y, p_n = p]$$

$$= \sum_{j\in S} \int_{\mathbf{R}^d \times \mathcal{P}(S)} g^j(y', p') \, \Pi^{i,j}(y, p, dy', dp') ,$$

for any $i \in S$, any $y \in \mathbf{R}^d$, and any $p \in \mathcal{P}(S)$.

In addition to the Assumption A on the *true* transition probability matrix $Q_\bullet$, we shall need the following *integrability* assumption on the observation densities $b(\cdot)$ and $b_\bullet(\cdot)$:

$$\Delta = \max_{i\in S} \int_{\mathbf{R}^d} \delta(y) \, b_\bullet^i(y) \, \lambda(dy) < \infty .$$

Example 3.1 If the observation conditional densities are Gaussian for both the true and the assumed models, i.e. in particular

$$b^i(y) = (2\pi)^{-d/2} \exp\{ -\tfrac{1}{2} |y - h^i|^2 \} ,$$

for any $i \in S$, and any $y \in \mathbf{R}^d$, then Δ is finite. Indeed, for any $i, j \in S$, and any $y \in \mathbf{R}^d$

$$\frac{b^i(y)}{b^j(y)} = \exp\{ -\tfrac{1}{2} |y - h^i|^2 + \tfrac{1}{2} |y - h^j|^2 \}$$

$$= \exp\{ y^* (h^i - h^j) - \tfrac{1}{2} (h^i + h^j)^* (h^i - h^j) \} ,$$

hence

$$\delta(y) \leq \exp\{ \max_{i,j\in S} |h^i - h^j| \, [\, |y| + \max_{i\in S} |h^i| \,] \} .$$

We introduce next the following suitable class of test functions defined on $S \times \mathbf{R}^d \times \mathcal{P}(S)$.

Definition 3.2 *Let L denote the set of functions $g = (g^i)$ defined on $S \times \mathbf{R}^d \times \mathcal{P}(S)$, such that for any $i \in S$, and any $y \in \mathbf{R}^d$ the partial mapping $p \mapsto g^i(y, p)$ is Lipschitz continuous (hence bounded since $\mathcal{P}(S)$ is compact), i.e.*

$$|g^i(y, p) - g^i(y, p')| \leq \mathrm{Lip}(g^i, y) \, \|p - p'\| ,$$

for any $p, p' \in \mathcal{P}(S)$, and such that

$$\mathrm{Lip}(g) = \max_{i\in S} \int_{\mathbf{R}^d} \mathrm{Lip}(g^i, y) \, b_\bullet^i(y) \, \lambda(dy) < \infty ,$$

$$\mathrm{K}(g) = \max_{i\in S} \int_{\mathbf{R}^d} \mathrm{K}(g^i, y) \, b_\bullet^i(y) \, \lambda(dy) < \infty ,$$

where by definition

$$\mathrm{K}(g^i, y) = \sup_{p\in\mathcal{P}(S)} |g^i(y, p)| .$$

Theorem 3.3 *If the stochastic matrix Q is primitive, with index of primitivity r, if Assumption A holds, and if Δ is finite, then there exist constants $0 < \rho < 1$, with $\rho > \max(\rho_\bullet, (1 - R)^{1/r})$, and $C > 0$ such that, for any $z, z' \in S \times \mathbf{R}^d \times \mathcal{P}(S)$, and for any function $g = (g^i)$ in L*

$$|\Pi^n \, g(z) - \Pi^n \, g(z')| \leq C \, \varepsilon^{-r} \, [\, \mathrm{Lip}(g) + \mathrm{K}(g) \,] \, \rho^n ,$$

where the constant C depends only on r, Δ, and $K_\bullet$.

The constants $\rho_\bullet$ and $K_\bullet$ are defined in Remark 1.2 above. The following corollary holds, whose proof is similar to the proof of Proposition 2 in Benveniste, Métivier and Priouret [3, Part II, Chapter 2].

Corollary 3.4 *With the assumptions of Theorem 3.3, the Markov chain* $\{Z_n = (X_n, Y_n, p_n), n \geq 0\}$ *has, under the true probability measure* $\mathbf{P}_\bullet$*, a unique invariant probability distribution* $\mu = (\mu^i)$ *on* $S \times \mathbf{R}^d \times \mathcal{P}(S)$*. For any* $z \in S \times \mathbf{R}^d \times \mathcal{P}(S)$*, and for any function* $g = (g^i)$ *in* L

$$|\Pi^n g(z) - \lambda| \leq C \varepsilon^{-r} \left[\mathrm{Lip}(g) + \mathrm{K}(g) \right] \frac{\rho^n}{1-\rho} ,$$

and there exist a unique solution $V = (V^i)$ *defined on* $S \times \mathbf{R}^d \times \mathcal{P}(S)$ *of the* Poisson equation

$$[I - \Pi] V(z) = g(z) - \lambda ,$$

where λ *is defined as*

$$\lambda = \sum_{i \in S} \int_{\mathbf{R}^d \times \mathcal{P}(S)} g^i(y,p) \, \mu^i(dy, dp) .$$

The proof of the theorem is based on the next proposition.

Proposition 3.5 *If the stochastic matrix* Q *is primitive, with index of primitivity* r*, and if* Δ *is finite, then for any* $p, p' \in \mathcal{P}(S)$*, for any integers* n, m *such that* $n \geq m + 2r - 1$*, and for any function* $g = (g^i)$ *in* L

$$\max_{i_m, \cdots, i_{n+1} \in S} \int_{\mathbf{R}^d} \cdots \int_{\mathbf{R}^d} \left| g^{i_{n+1}}(y_{n+1}, f[y_n, \cdots, y_m, p]) \right.$$

$$\left. - g^{i_{n+1}}(y_{n+1}, f[y_n, \cdots, y_m, p']) \right|$$

$$b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \, \lambda(dy_m) \cdots \lambda(dy_{n+1})$$

$$\leq C \varepsilon^{-r} \, \mathrm{Lip}(g) \, \rho_*^{n-m+1-2r} .$$

where $\rho_* = (1 - R)^{1/r}$*, and where the constant* C *depends only on* r*, and* Δ*.*

Remark 3.6 In DiMasi and Stettner [4], the existence of a unique invariant probability distribution is proved for another extended Markov chain based on the filter itself instead of the prediction filter, in the case of a misspecified HMM with conditionally Gaussian observations, with positive transition probability matrices (for both the true and the assumed models), and under the somewhat restrictive assumption that the mapping $h_\bullet = (h_\bullet^i)$ from S to $\mathbf{R}^d$ is *injective*.

Acknowledgement

The authors gratefully acknowledge Jan van Schuppen for bringing the paper [1] to their attention, Ofer Zeitouni for providing a preprint of the paper [2], and Dan Ocone and Amarjit Budhiraja for interesting discussion, in particular for their suggestion of using the Birkhoff contraction coefficient.

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] R. ATAR and O. ZEITOUNI. Lyapunov exponents for finite state nonlinear filtering. *SIAM Journal on Control and Optimization*, 35(1):36–55, January 1997.

[3] A. BENVENISTE, M. MÉTIVIER, and P. PRIOURET. *Adaptive Algorithms and Stochastic Approximations*, volume 22 of *Applications of Mathematics*. Springer Verlag, New York, 1990.

[4] G. DI MASI and L. STETTNER. On adaptive control of a partially observed Markov chain. *Applicationes Mathematicæ*, 22(2):165–180, 1994.

[5] H. FURSTENBERG and H. KESTEN. Product of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, June 1960.

[6] T. KAIJSER. A limit theorem for partially observed Markov chains. *The Annals of Probability*, 3(4):677–696, August 1975.

[7] T. KAIJSER. A limit theorem for Markov chains with applications to products of random matrices. *Duke Mathematical Journal*, 45(2):311–349, June 1978.

[8] F. LE GLAND and L. MEVEL. Recursive identification of HMM's with observations in a finite set. In *Proceedings of the 34th Conference on Decision and Control, New Orleans 1995*, pages 216–221. IEEE–CSS, December 1995.

[9] F. LE GLAND and L. MEVEL. Geometric ergodicity in hidden Markov models. Publication Interne 1028, IRISA, July 1996. `ftp://ftp.irisa.fr/techreports/1996/PI-1028.ps.gz`.

[10] F. LE GLAND and L. MEVEL. Asymptotic behaviour of the MLE in hidden Markov models. In *Proceedings of the 4th European Control Conference, Bruxelles 1997*, July 1997. Paper FRA–F6.

[11] F. LE GLAND and L. MEVEL. Recursive identification in HMM's. In *Proceedings of the 36th Conference on Decision and Control, San Diego 1997*. IEEE–CSS, December 1997. Paper TP18–5.

[12] E. SENETA. Coefficients of ergodicity : structure and applications. *Advances in Applied Probability*, 11:576–590, 1979.

[13] E. SENETA. *Non–negative Matrices and Markov Chains*. Springer Series in Statistics. Springer Verlag, New York, 2nd edition, 1981.

[14] L. STETTNER. Invariant measures for the pair : state, approximate filtering process. *Colloquium Mathematicum*, LXII(2):347–351, 1991.

A Product of non–negative matrices

Let $S = \{1, \cdots, N\}$, and let $\| \cdot \|$ denote the L_1–norm, i.e. for any $u = (u^i)$ in $\mathbf{R}^N$

$$\|u\| = \sum_{i \in S} |u^i| \ .$$

We recall some results on coefficients of ergodicity, taken from Seneta [12] and [13, Chapter 3].

For any column–allowable nonnegative $N \times N$ matrix $T = (T^{j,i})$ (i.e. with at least one positive element in each column) the Birkhoff contraction coefficient $\tau_B(T)$ is defined by

$$\tau_B(T) = \frac{1 - \sqrt{\Phi(T)}}{1 + \sqrt{\Phi(T)}} \leq 1 - \sqrt{\Phi(T)} \ ,$$

where

- if each row of T is either zero–free or all–zero

$$\Phi(T) = \min_{i_1, i_2, j_1, j_2 \in S} \frac{T^{j_1,i_1} \, T^{j_2,i_2}}{T^{j_1,i_2} \, T^{j_2,i_1}} \ ,$$

  where the indices $j_1, j_2 \in S$ are restricted to the zero–free rows,

- otherwise (i.e. if there is a row with both zero and positive elements), $\Phi(T) = 0$,

see Seneta [13, Theorems 3.10 and 3.12].

For any stochastic matrix $P = (P^{i,j})$, and any probability vectors $p, q \in \mathcal{P}(S)$

$$\|P^* (p - q)\| \leq \tau_1(P) \, \|p - q\| \ ,$$

where the coefficient of ergodicity $\tau_1(P)$ is defined by

$$\tau_1(P) = \tfrac{1}{2} \max_{i, i' \in S} \sum_{j \in S} |P^{i,j} - P^{i',j}| \ ,$$

see Seneta [12]. Moreover $\tau_1(P) \leq \tau_B(P^*)$, and if in addition the stochastic matrix P is allowable (i.e. both row– and column–allowable), then $\tau_1(P) \leq \tau_B(P^*) = \tau_B(P)$, see Seneta [13, Theorem 3.13].

For any integer $n \geq 0$, let $A_n = (A_n^{i,j})$ be a row–allowable nonnegative $N \times N$ matrix. For any integers n, m such that $n \geq m$, define the backward product

$$A_{m,n} = (A_{m,n}^{i,j}) = A_m \cdots A_n \ ,$$

and the forward product

$$M_{n,m} = (M_{n,m}^{j,i}) = M_n \cdots M_m = A_{m,n}^* \ ,$$

and for any $i \in S$, define the sum of the i–th column entries of $M_{n,m}$, or equivalently the sum of the i–th row entries of $A_{m,n}$, as

$$M_{n,m}^{\bullet,i} = \sum_{j \in S} M_{n,m}^{j,i} = \sum_{j \in S} A_{m,n}^{i,j} = A_{m,n}^{i,\bullet} \ .$$

Define also the ratio

$$\omega(M_{n,m}) = \frac{\max_{i \in S} M_{n,m}^{\bullet,i}}{\min_{i \in S} M_{n,m}^{\bullet,i}} \ .$$

Proposition A.1 *For any integers n, m such that $n \geq m$, and any $p, q \in \mathcal{P}(S)$*

$$\left\| \frac{M_{n,m} \, p}{e^* \, M_{n,m} \, p} - \frac{M_{n,m} \, q}{e^* \, M_{n,m} \, q} \right\|$$

$$\leq 2 \, \omega(M_{n,m}) \, \tau_B(M_{n,m}) \, \|p - q\| \ .$$

The proof follows the same lines as the proof of Lemma 2.2 in Arapostathis and Marcus [1], see also Lemma 6.2 in Kaijser [6], and Lemma 8.1 in Kaijser [7].