Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Recursive estimation of Hidden Markov Models

László Gerencsér, Gábor Molnár-Sáska and Zsanett Orlovits

Abstract— **A recursive estimation method for Hidden Markov Models has been proposed in [24]. As suggested there the proposed recursive algorithm could be analyzed via the theory of stochastic approximations developed in [4]. The purpose of this note is to verify the basic probabilistic conditions of [4], given in Part II, Chapter 1 of [4]. For this purpose we consider a general class of Markov models in which a simple Markov process is passed through an exponentially stable non-linear system. The general theory is relatively easily applied to HMMs extended by their filter process and their derivatives, see [1].**

I. INTRODUCTION

The estimation of the dynamics of a Hidden Markov Model is a basic problem in applications, see [14]. A key element in the statistical analysis of HMM-s is a strong law of large numbers for the log-likelihood function. The first basic result in this area is due to Baum and Petrie for finite state Markov chains with finite-range read-outs [3]. This result has been extended to continuous read-outs by Leroux in [26] and LeGland and Mevel in [23]. An alternative approach is introduced via the theory of L-mixing processes, see [19]. The techniques yield consistency and strong approximation results for off-line estimators, see [18].

A recursive estimation method for Hidden Markov Models has been proposed in [24] and [25]. As suggested the proposed recursive algorithm could be analyzed via the theory of stochastic approximations, see [4]. Krishnamurthy and Yin, see [21], investigated the convergence and rate of convergence of the recursive estimation of HMMs using the weak convergence approach of Kushner and Yin [22]. Recursive estimation for continuous time Hidden Markov Models was given by Elliott, see [12].

The purpose of this paper is to verify the basic probabilistic conditions for HMMs, given in Part II, Chapter 1 of [4]. For this purpose we consider a general class of Markov models in which a simple Markov process is passed through an exponentially stable non-linear system. The general theory is applied to HMMs extended by their filter process and their derivatives, see [1].

This work was supported by the National Research Foundation of Hungary (OTKA) under Grant no. T047193

L. Gerencsér is with MTA SZTAKI, Computer and Automation Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary gerencser@sztaki.hu

G. Molnár-Sáska is with MTA SZTAKI, Computer and Automation Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary molnarsg@sztaki.hu

Zs. Orlovits is with MTA SZTAKI, Computer and Automation Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary orlovits@sztaki.hu

II. HIDDEN MARKOV MODELS

We consider Hidden Markov Models with a general state space $\mathcal{X}$ and a general observation or read-out space $\mathcal{Y}$. Both are assumed to be Polish spaces, i.e. they are complete, separable metric spaces.

Definition 2.1: The pair (X_n, Y_n) is a Hidden Markov process if (X_n) is a homogenous Markov process, with state space $\mathcal{X}$ and the observations (Y_n) are conditionally independent and identically distributed given (X_n).

Example 2.1: Assume that the observations are of the form

$$Y_n = h(X_n) + \epsilon_n,$$

for any integer $n \geq 0$, where $\{\epsilon_n, n \geq 0\}$ is a Gaussian white noise sequence independent of the Markov process $\{X_n, n \geq 0\}$, and $h : \mathcal{X} \to R$.

To illustrate the basic concepts let the state space of the Hidden Markov Model be finite in this paper, i.e. $|\mathcal{X}|$=N. The results for compact state space are very similar.

Let Q^* be the transition matrix of the unobserved Markov process (X_n), i.e.

$$Q^*_{ij} = P(X_{n+1} = j | X_n = i),$$

where $*$ indicates that we take the true value of the corresponding unknown quantity. If $\mathcal{Y}$ is finite, say $|\mathcal{Y}| = M$, then we have

$$P(Y_n = y_n, \dots Y_0 = y_0 | X_n = x_n, \dots X_0 = x_0) = \prod_{i=0}^{n} P(Y_i = y_i | X_i = x_i).$$

In this case we will use the following notations

$$P(Y_k = y | X_k = x) = b^{*x}(y).$$

Continuous read-outs will be defined by taking the following conditional densities:

$$P(Y_n \in y + dy | X_n = x) = b^{*x}(y)\lambda(dy), \qquad (1)$$

where λ is a fixed nonnegative, σ-finite measure. Let

$$B^*(y) = \mathrm{diag}(b^{*i}(y)),$$

where $i = 1, \dots, N$ and

$$\mathbf{b}^*(y) = (b^1(y), \dots b^N(y))^T.$$

For notational convention we use $Q > 0$ if all the elements of the transition probability matrix are strictly positive.

A key quantity in estimation theory is the predictive filter defined by

$$p^{*j}_{n+1} = P(X_{n+1} = j | Y_n, \dots, Y_0).$$

0-7803-9568-9/05/$20.00 ©2005 IEEE

Writing $\mathbf{p}^*_{n+1} = (p^{*1}_{n+1}, \dots, p^{*N}_{n+1})^T$, the filter process satisfies the Baum-equation

$$\mathbf{p}^*_{n+1} = \pi(Q^{*T} B^*(Y_n)\mathbf{p}^*_n), \tag{2}$$

both in discrete and continuous read-out cases, where π is the normalizing operator: for $x \geq 0,\ x \neq 0$ set $\pi(x)^i = x^i / \sum_j x^j$, see [3]. Here $p^{*j}_0 = P(X_0 = j)$.

In practice, the transition probability matrix Q^* and the initial probability distribution p^*_0 of the unobserved Markov chain (X_n) and the conditional probabilities $b^{*i}(y)$ of the observation sequence (Y_n) are possibly unknown. For this reason we consider the Baum-equation in a more general sense

$$\mathbf{p}_{n+1} = \pi(Q^T B(Y_n)\mathbf{p}_n), \tag{3}$$

with initial condition $\mathbf{p}_0 = \mathbf{q}$, where Q is a stochastic matrix, $\mathbf{p}_n$ is a probability vector on $\mathcal{X}$, and $B(y) = \mathrm{diag}(b^i(y))$ is a collection of conditional probabilities.

We will take an arbitrary probability vector $\mathbf{q}$ as initial condition, and the solution of the Baum equation will be denoted by $\mathbf{p}_n(\mathbf{q})$.

A key property of the Baum equation is its exponential stability with respect to the initial condition. This has been established in [23] for finite state space with continuous read-outs and in [11] for compact state space with continuous read-outs. Here we state the result only for HMMs with positive transition probability matrix:

Proposition 2.1: Assume that $Q > 0$ and $b^x(y) > 0$ for all x, y. Let $\mathbf{q}$, $\mathbf{q}'$ be any two initializations. Then

$$\|\mathbf{p}_n(\mathbf{q}) - \mathbf{p}_n(\mathbf{q}')\|_{TV} \leq C(1-\delta)^n \|\mathbf{q} - \mathbf{q}'\|_{TV}, \tag{4}$$

where $\|\quad\|_{TV}$ denotes the total variation norm and $0 < \delta < 1$.

That is, the filter forgets its initial condition with exponential rate. An essential feature of the result is that, $\|\mathbf{q} - \mathbf{q}'\|_{TV}$ shows up in the upper bound, see [2]. We note that Proposition 2.1 is a non-probabilistic statement.

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (4) holds with a random C.

Consider the following estimation problem: let Q and $\mathbf{b}$ be parameterized by $\theta \in D \subset R^r$, and let

$$Q^* = Q(\theta^*), \quad \mathbf{b}^* = \mathbf{b}(\theta^*).$$

Usually the entries of Q are part of θ. Assume that Q and $\mathbf{b}$ are smooth functions of θ.

The log-likelihood function $\log L(y_0, \dots y_n, \theta)$ is

$$\sum_{k=1}^{n-1} \log p(y_k | y_{k-1}, \dots y_0, \theta) + \log p(y_0, \theta). \tag{5}$$

The k-th term in (5) for $k \geq 1$ can be written as

$$\log \sum_i b^i(y_k, \theta) P(X_k = i | y_{k-1}, \dots, y_0, \theta) =$$

$$\log \sum_i b^i(y_k, \theta) p^i_k(\theta). \tag{6}$$

For the off-line maximum-likelihood estimation we should solve the following equation

$$\frac{\partial}{\partial \theta} \log L(y_0, \dots y_n, \theta) = 0.$$

The aim of this paper is to investigate the on-line estimation procedure.

III. On-line estimation

In the on-line estimation procedure we define a stochastic algorithm with Markovian dynamics, see Benveniste, Metivier and Priouret [4], as follows.

Let us denote the on-line estimation of the parameter at step n by $\overline{\theta}_n$. Consider the parameter-dependent Baum-equation

$$\mathbf{p}_{n+1}(\theta) = \frac{Q^T(\theta) B(y_n, \theta)\mathbf{p}_n(\theta)}{\mathbf{b}(y_n, \theta)^T \mathbf{p}_n(\theta)} = \Phi_1(y_n, \mathbf{p}_n, \theta), \tag{7}$$

To simplify the notations we drop the dependence on the parameter θ. Differentiating $\mathbf{p}_{n+1}$ with respect to θ we have

$$W_{n+1} = Q^T \left(I - \frac{B(y_n)\mathbf{p}_n \mathbf{e}^T}{\mathbf{b}^T(y_n)\mathbf{p}_n} \right) \frac{B(y_n) W_n}{\mathbf{b}^T(y_n)\mathbf{p}_n} + F, \tag{8}$$

where

$$F = \frac{Q^T_\theta B(y_n)\mathbf{p}_n}{\mathbf{b}^T(y_n)\mathbf{p}_n} + Q^T \left(I - \frac{B(y_n)\mathbf{p}_n \mathbf{e}^T}{\mathbf{b}^T(y_n)\mathbf{p}_n} \right) \frac{\beta(y_n)\mathbf{p}_n}{\mathbf{b}^T(y_n)\mathbf{p}_n},$$

$W_n = \frac{\partial \mathbf{p}_n}{\partial \theta}$, $\beta(y_n) = \frac{\partial B(y_n)}{\partial \theta}$ and $e = (1, \dots, 1)^T$.

In a compact form

$$W_{n+1} = \Phi_2(y_n, \mathbf{p}_n, W_n, \theta).$$

Thus for a fix θ, $u_n = (X_n, Y_n, \mathbf{p}_n, W_n, \theta)$ is a Markov chain.

Let the score function be

$$\varphi_n(\theta) = \frac{\partial}{\partial \theta} \log p(y_n | y_{n-1}, \dots, y_0, \theta).$$

Using (6) we get

$$\varphi_n = \frac{\beta(y_n)\mathbf{p}_n + W_n \mathbf{b}(y_n)}{\mathbf{b}(y_n)^T \mathbf{p}_n}.$$

Let

$$H(\theta, u) = H(\theta, x, y, \mathbf{p}, W) = \frac{\beta(y, \theta)\mathbf{p} + W\mathbf{b}(y, \theta)}{\mathbf{b}(y, \theta)^T \mathbf{p}}, \tag{9}$$

and consider the following adaptive algorithm.

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n} H(\overline{\theta}_n, x_n, y_n, \overline{\mathbf{p}}_n, \overline{W}_n), \tag{10}$$

$$\overline{\mathbf{p}}_{n+1} = \Phi_1(y_n, \overline{\mathbf{p}}_n, \overline{\theta}_n), \tag{11}$$

$$\overline{W}_{n+1} = \Phi_2(y_n, \overline{\mathbf{p}}_n, \overline{W}_n, \overline{\theta}_n). \tag{12}$$

For the convergence of this algorithm we use the approach of Benveniste, Metivier and Priouret, see [4]. In the next section we summarize the results therein.

IV. THE BMP SCHEME

In this section we present the basics of the theory of recursive estimation developed by Benveniste, Metivier and Priouret, BMP henceforth (see Chapter 2, Part II. of [4]).

Let a family of transition probabilities $\{\Pi_\theta,\ \theta \in D \subset R^d\}$ on $\mathcal{U}$ be given, where $\mathcal{U}$ is a Polish space. Let us denote the metric by d. Note that in [4] $\mathcal{U}$ is R^n, but the results can be generalized for complete separable metric space. Let D be an open set. Assume that for any $\theta \in D$ there exists a unique invariant probability measure, say μ_θ. Let $(U_n(\theta))$ be a Markov-chain such that its initial state $U_0(\theta)$ has distribution μ_θ. Let $H(\theta, u)$ be a mapping from $R^d \times \mathcal{U}$ to R^d. Then the basic estimation problem of the BMP-theory is to solve the equation

$$E_{\mu_\theta} H(\theta, U(\theta)) = 0.$$

Assume that a solution $\theta^* \in D$ exists.

The BMP-scheme. The recursive estimation procedure to solve the above equation is then defined as

$$\theta_{n+1} = \theta_n + \frac{1}{n} H(\theta_n, U_n), \tag{13}$$

where U_n is the time-varying process defined by

$$P(U_{n+1} \in A | \mathcal{F}_n) = \Pi_{\theta_n}(U_n, A).$$

Here $\mathcal{F}_n$ is the σ-field of events generated by the random variables $U_0, \ldots, U_n$ and A is any Borel subset of $\mathcal{U}$.

To specify the class of functions H for which the theory is developed consider a Lyapunov function $V : \mathcal{U} \to R^+$ and define for real-valued functions g on $\mathcal{U}$ and any $p \geq 0$ the norms

$$||g||_p := \sup_u \frac{|g(u)|}{1 + V(u)^p},$$

and

$$||\Delta g||_p = \sup_{u_1 \neq u_2} \frac{|g(u_1) - g(u_2)|}{d(u_1, u_2)(1 + V(u_1)^p + V(u_2)^p)}.$$

Introduce the class of functions

$$C(p) = \{\ g : g \text{ is continuous and } ||g||_p < \infty\}.$$

and

$$Li(p) = \{\ g\ :\ ||\Delta g||_p < +\infty\}.$$

Note that $Li(p) \subseteq C(p+1)$ for any $p \geq 0$.

Conditions of BMP. All but one condition will be formulated in terms of the Markov chain $\{U_n(\theta) : n \geq 0\}$ for a fixed $\theta \in D$ with an arbitrary non-random initial value $U_0(\theta) = u$. The conditions are as follows. The real number $p \geq 0$ is fixed all over the conditions A1.-A3. below.

A1. For any compact subset $Q \subset D$ there exists a constant $K = K(Q)$ such that for all $\theta \in Q$, $n \geq 0$ and $U_0(\theta) = u \in \mathcal{U}$:

$$\int \Pi_\theta^n(u, dy)(1 + V(y)^{p+1}) \leq K(1 + V(u)^{p+1}).$$

A2. For any compact subset Q of D there exist constants $K = K(Q)$ and $0 < \rho < 1$ such that for all $g \in Li(p)$, any $\theta \in Q, n \geq 0$ and $u, u' \in \mathcal{U}$:

$$|\Pi_\theta^n g(u) - \Pi_\theta^n g(u')| \leq \\ \leq K||\Delta g||_p\ \rho^n d(u, u')(1 + V(u)^p + V(u')^p).$$

Conditions A1 and A2 imply geometric ergodicity of the Markov chains in the following sense: for any $\theta \in D, u \in \mathcal{U}$ and any $g \in C(p+1)$ there exists a $\Gamma_\theta g$ such that

$$|\Pi_\theta^n g(u) - \Gamma_\theta g| \leq ||g||_{p+1} \rho^n (1 + V(u)^{p+1}).$$

A key contribution of the BMP theory is that the above geometric ergodicity is derived by verifying conditions on a much more convenient class of test functions, namely $Li(p)$. It follows that that there exists a unique invariant measure μ_θ such that

$$\Gamma_\theta g = \int g(u) d\mu_\theta(du)$$

for $g \in C(p+1)$.

A3. For any compact subset Q of D there exists a constant $K = K(Q)$ such that for all $g \in Li(p)$, any $\theta, \theta' \in Q$ and $n \geq 0, u \in \mathcal{U}$:

$$|\Pi_\theta^n g(u) - \Pi_{\theta'}^n g(u)| \leq K||\Delta g||_p\ |\theta - \theta'|(1 + V(u)^{p+1}).$$

In other words the kernels Π_θ^n are supposed to be Lipschitz-continuous, uniformly in n, with respect to the parameter θ when applied to a small set of test functions $Li(p)$.

Let $D_0 \subset D$ be a fixed compact truncation domain such that $\theta^* \in \text{int} D_0$. Define the stopping time

$$\tau = \inf\{n : \theta_{n+1} \notin D_0\}.$$

In addition let ϵ be a fixed small positive number, and define

$$\sigma = \inf\{n : |\theta_n - \theta_{n-1}| > \epsilon\}.$$

The stability of the time-varying process U_n is enforced by stopping it at $\tau \wedge \sigma$.

A4. For any compact subset Q of D there exists a constant $K = K(Q)$ such that for any $n \geq 0$ and arbitrary starting values $\theta \in Q,\ u \in \mathcal{U}$

$$E_{\theta,u}\{I(n < \tau \wedge \sigma)(1 + V(U_{n+1})^{p+1}\} \leq K(1 + V(u)^{p+1})$$

Regularity of the function H is required in the next condition:

A5. For any compact subset Q of D there exists a constant $K = K(Q)$ such that for all $\theta, \theta' \in Q$

$$|H(\theta, u)| \leq K(1 + V(u)^{p+1}) \\ |H(\theta, u) - H(\theta', u)| \leq K|\theta - \theta'|(1 + V(u)^{p+1}) \\ ||\Delta H(\theta, \cdot)||_p \leq K.$$

Remark: In fact it is sufficient to require the above condition for $\Pi_\theta H_\theta$, thus H may be discontinuous.

Since $H(\theta, \cdot) \in Li(p)$ we may set as above

$$h(\theta) = \lim_{n \to \infty} \Pi_\theta^n H(\theta, U_n(\theta)) = E_{\mu_\theta} H(\theta, U(\theta)).$$

The associated ODE is then given by

$$\dot{\theta}_s = h(\theta_s). \tag{14}$$

To ensure the convergence of the SA-procedure we require global asymptotic stability of the associated ODE by assuming the existence of a Lyapunov function:

A6. There exists a real-valued C^2-function $\widetilde{U}$ on D such that

(i) $\widetilde{U}(\theta^*) = 0,\ \widetilde{U}(\theta) > 0$ for all $\theta \in D\backslash\{\theta^*\}$

(ii) $\widetilde{U}'(\theta)h(\theta) < 0$ for all $\theta \in D\backslash\{\theta^*\}$

(iii) $\widetilde{U}(\theta) \to \infty$ if $\theta \to \partial D$ or $|\theta| \to \infty$.

Theorem 13, p. 236 of [4] yields the following convergence result.

Theorem 4.1: Assume that Conditions A1 - A6 are satisfied, and ϵ is sufficiently small. Let $\theta \in \mathrm{int}D_0,\ U_m = u \in \mathcal{U}$, and consider the stopped process $\theta_n^\circ = \theta_{n\wedge\tau\wedge\sigma}$. Then for any $0 < \lambda < 1$ there exist constants B and s such that for all $m \geq 0$ we have $\lim \theta_n^\circ = \theta^*$ with probability at least

$$1 - B(1 + V(u)^s) \sum_{n=m+1}^{+\infty} n^{-1-\lambda}.$$

V. Exponentially stable nonlinear systems

In the rest of the paper conditions (**A1**)-(**A4**) are verified for Hidden Markov Models. For this we consider general results, then conclude for HMMs. For the sake of readability, all proofs are relegated to the Appendix.

Consider a Polish space $\mathcal{X}$ and a sequence of independent, $[0,1]$-uniform random variables (E_n) on a probability space $(\Omega, \mathcal{F}, \mathcal{Q})$. Let f be a Borel measurable deterministic function $f:\ \mathcal{X} \times [0,1] \longrightarrow \mathcal{X}$. Then the sequence (X_n) defined by

$$X_n = f(X_{n-1}, E_{n-1}), \quad X_0 = x \tag{15}$$

is a Markov chain, where $x \in \mathcal{X}$ is an arbitrary initialization.

Conversely, any Markov chain can be represented by (15), see [20]. In the following we will denote the random mapping $f(\cdot, E_{n-1})$ by T_n, i.e. for $x \in \mathcal{X}$

$$T_n x = f(x, E_{n-1}). \tag{16}$$

The process defined by $X_{n+1} = T_{n+1}X_n$ is Markov.

The representation can be given in a constructive way but it should be noted that it is not unique. This representation plays a key role in subsequent analysis.

Next we are going to introduce the notion of Doeblin-condition, see [7]:

Definition 5.1: Given a Markov chain (X_n) with state space $\mathcal{X}$. If there exists an integer $m \geq 1$ such that

$$P^m(x, A) \geq \delta\nu(A)$$

is valid for all $x \in \mathcal{X}$ and $A \subset \mathcal{B}(\mathcal{X})$ with $\delta > 0$ and some probability measure ν, then we say that the Doeblin-condition is satisfied.

Here δ can be interpreted as the weight of the i.i.d. factor of the Markov chain. The following lemma, see [7], shows the relation between the Doeblin-condition and the representation of the Markov chain.

Lemma 5.1: Let (X_n) be a Markov chain. The Doeblin-condition is valid with $m = 1$ if and only if there exists a representation such that $\mathcal{Q}(T_n \in \Gamma_c) \geq \delta$, where Γ_c is the set of constant mappings.

Proposition 5.1: Assume that the Doeblin-condition holds with $m = 1$ for a Markov chain (X_n). Then there exists an invariant distribution π, and

$$|P^n(x, A) - \pi(A)| \leq (1 - \delta)^n \quad \text{for} \quad \forall A \in \mathcal{B}(\mathcal{X}). \tag{17}$$

Now we formulate a general concept of exponential stability motivated by Proposition 2.1. Let $\mathcal{X}$ be a Polish space and let us denote the metric on $\mathcal{X}$ by $d_{\mathcal{X}}$. Furthermore let $\mathcal{Z}$ be a Banach space. Let $f : \mathcal{X} \times \mathcal{Z} \longrightarrow \mathcal{Z}$ be a Borel-measurable function, and for a fixed sequence $(x_n)_{n\geq 0}$, $x_n \in \mathcal{X}$ consider the recursion

$$z_{n+1} = f(x_n, z_n, \theta), \quad z_0 = \xi. \tag{18}$$

Let the solution be denoted by $z_n(\xi)$. To simplify the notations we drop the dependence on the sequence (x_n) and the parameter θ.

Definition 5.2: The mapping f is uniformly exponentially stable if for every sequence $(x_n)_{n\geq 0}$, $x_n \in \mathcal{X}$

$$\|z_n(\xi) - z_n(\xi')\| \leq C(1 - \varrho)^n \|\xi - \xi'\|, \tag{19}$$

where $C > 0, 1 > \varrho > 0$ are independent of the sequence (x_n).

We say that the process z_n is exponentially stable if (19) holds. Under reasonable technical conditions this condition is satisfied for the Baum-equation.

Define the process (Z_n) by

$$Z_{n+1} = f(X_n, Z_n, \theta), \quad Z_0 = \xi, \tag{20}$$

where (X_n) is a Markov chain which satisfies the Doeblin condition. Let

$$X_{n+1} = T_n X_n, \tag{21}$$

where (T_n) is a sequence of i.i.d. random mappings, see (16). Let $U_n = (X_n, Z_n) \in \mathcal{X} \times \mathcal{Z} = \mathcal{U}$. Define the metric on $\mathcal{U}$ by

$$d(u, u') = \|z - z'\| + d_{\mathcal{X}}(x, x'), \tag{22}$$

where $u = (x, z)$ and $u' = (x', z')$, and let the Lyapunov function be

$$V(u) = \|z\|. \tag{23}$$

In the following subsection conditions (**A1**)-(**A3**) are verified for the process U_n defined above.

A. Verification of BMP conditions

By Proposition 5.1 a stationary distribution of X_n exists. Let us denote it by π. For assumption (**A1**) we need two conditions: the first one ensures that there are no states in "large distances", the second one is (**A1**) for one-step when X_0 has an invariant distribution.

Condition 5.1: Let the distribution of X_1 be π_1. Assume

$$\frac{d\pi_1}{d\pi} \leq C_1.$$

Condition 5.2: Assume for all $\xi \in \mathcal{Z}$ and for $p \geq 1$

$$E_\pi \|Z_1(\xi)\|^p \leq K_1(1 + \|\xi\|^p),$$

or equivalently

$$\int_{\mathcal{X}} \|f(x,\xi)\|^p d\pi(x) \leq K_1(1 + \|\xi\|^p). \quad (24)$$

Theorem 5.1: Consider a process $U_n = (X_n, Z_n)$ defined by (20), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 5.1 and 5.2 are satisfied. Then assumption (**A1**) holds.

Proof: Following the route of the proof of Lemma V.1 in [19] we get that

$$E\|Z_n\|^p \leq K(1 + \|\xi\|^p),$$

where $Z_0 = \xi$ is a fixed constant initialization. Using the definition of the function V, see (23) we get (**A1**). ■

Note that in Theorem 5.1 $\mathcal{X}$ can be any abstract set, we do not use the metric property here. Furthermore we do not use the Doeblin property of the Markov chain X_n.

For assumption (**A2**) we need two more conditions for the stability of the process (X_n).

Condition 5.3: Assume that f is Lipschitz continuous in x, i.e.

$$\|f(x_1, z) - f(x_2, z)\| \leq L d_{\mathcal{X}}(x_1, x_2)$$

Condition 5.4: Assume that for the process (X_n) we have

$$E d_{\mathcal{X}}(X_n, X_n') \leq K d_{\mathcal{X}}(X_0, X_0')$$

Theorem 5.2: Consider a process $U_n = (X_n, Z_n)$ defined by (20), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 5.1, 5.2, 5.3 and 5.4 are satisfied. Then assumption (**A2**) holds.

Proof: Using the definition of $d(u, u')$ and the idea of Lemma V.1 in [19] we get that Conditions 5.3 and 5.4 imply

$$E d(U_n, U_n') \leq K d(u_0, u_0'), \quad (25)$$

where K is independent of n.

For $g \in Li(p)$ we have

$$|g(u_n) - g(u_n')| \leq$$
$$\|\Delta g\|_p d(u_n, u_n')(1 + |V(u_n)|^p + |V(u_n')|^p). \quad (26)$$

Let $A = \{\omega : \ T_k(\omega) \in \Gamma_c \text{ for } k \leq n/2\}$. From Lemma 5.1 we have $P(A) = 1 - (1-\delta)^{n/2}$. On A we have $x_k = x_k'$ for all $n/2 \leq k \leq n$. Thus from the definition of d and the exponential stability of the mapping f we have on the set A

$$d(u_n, u_n') = |z_n - z_n'| \leq C\rho^{n/2}|z_{n/2} - z_{n/2}'| =$$
$$C\rho^{n/2} d(u_{n/2}, u_{n/2}').$$

Taking the expectation of both sides of (26) and considering (25) we have

$$E\chi_A |g(U_n) - g(U_n')| \leq$$
$$\|\Delta g\|_p C\rho^{n/2} d(u, u')(1 + |V(u)|^p + |V(u')|^p). \quad (27)$$

Consider now the complement of A. We have $P(A^c) = (1-\delta)^{n/2}$. Taking the expectation of (26) on the set A^c and using (25) we have

$$E\chi_{A^c} |g(U_n) - g(U_n')| \leq$$
$$(1-\delta)^{n/2} \|\Delta g\|_p d(u, u')(1 + |V(u)|^p + |V(u')|^p) \quad (28)$$

Adding (27) and (28) we finish the proof. ■

For assumption (**A3**) we need the smoothness of f with respect to the parameter θ.

Theorem 5.3: Consider a process $U_n = (X_n, Z_n)$ defined by (20), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 5.1 and 5.2 are satisfied. Then assumption (**A3**) holds.

Proof: Observe that if f is a uniformly exponentially stable mapping, then the derivative process $w_n = \frac{\partial z_n}{\partial \theta}$ is also exponentially stable, i.e. we have

$$\|w_n(\eta) - w_n(\eta')\| \leq C(1-\varrho)^n \|\eta - \eta'\|, \quad (29)$$

where $\eta = \frac{\partial \xi}{\partial \theta}$. Consider the derivative of $g(x_n, z_n)$ with respect to the parameter θ:

$$\frac{\partial g(x_n, z_n)}{\partial \theta} = \frac{\partial g}{\partial z_n} \frac{\partial z_n}{\partial \theta}.$$

Using (29) and the fact that $g \in Li(p)$ we have for a fix $\omega \in \Omega$

$$\|\frac{\partial g(x_n, z_n)}{\partial \theta}\| \leq \|\Delta g\|(1 + |V(u_n)|)K$$

Taking the expectation of both sides and using Theorem 5.1 we get the proof. ■

We conclude this section with the following theorem.

Theorem 5.4: Consider a process $U_n = (X_n, Z_n)$ defined by (20), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 5.1, 5.2, 5.3 and 5.4 are satisfied. Then assumptions (**A1**)-(**A3**) hold.

Thus we get that if assumption (**A5**) is satisfied for a function H, and we have a Lyapunov function satisfying (**A6**) then convergence result Theorem 4.1 holds for the algorithm (13).

VI. Verification of BMP conditions for HMMs

In this section we apply the results of Theorem 5.4 and 4.1 for Hidden Markov Models. Let us turn back to the notations of Section III. Assume that $Q(\theta)$ and $b(\theta)$ are smooth functions of the parameter, i.e. the second derivatives exist.

Theorem 6.1: Consider a Hidden Markov Model with finite state space and finite read-out space. Assume that $Q^* > 0$, $b^{*x}(y) > 0$ and for all θ we have $Q(\theta) > 0$ and $b^x(y, \theta) > 0$ for all x, y. Then assumptions (**A1**)-(**A3**) are satisfied.

Proof: Identify X_n of Theorem 5.4 with (X_n, Y_n) and Z_n of Theorem 5.4 with (p_n, W_n) and use the results of Theorem 5.4. ■

By the smoothness of $b(y,\theta)$ and $Q(\theta)$ assumption (**A5**) is satisfied under the conditions of Theorem 6.1. Furthermore, note that if the state space and the read-out space are finite then assumption (**A4**) is trivially satisfied.

Assumption (**A6**) is very hard even for linear stochastic systems. Let us define

$$h(\theta) = \lim_{n\to\infty} E\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)$$

This limit exists, see e.g. [19]. The following local identifiability condition is assumed.

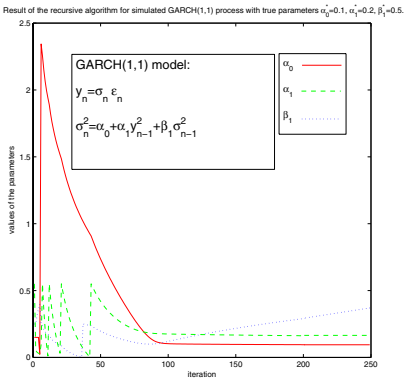
Condition 6.1: (Identifiability condition)

$$\frac{\partial}{\partial\theta}h(\theta^*) > 0.$$

Condition 6.1 implies assumption (**A6**) in a small domain. Thus we conclude with the following theorem.

Theorem 6.2: Consider a Hidden Markov Model with finite state space and finite read-out space. Assume that $Q^* > 0$, $b^{*x}(y) > 0$ and for all θ we have $Q(\theta) > 0$ and $b^x(y,\theta) > 0$ for all x, y. Assume Condition 6.1. Then the algorithm defined by (10), (11), (12) converges to the true value θ^* with probability arbitrary close to 1.

The above techniques can be used for recursive estimation in so called ARCH or GARCH processes playing an important role in mathematical finance. These processes introduced by Engle in [13] and Bollerslev in [8] are used to model the log-return of a stock-price process. For the best recent results on the estimation of GARCH processes see Berkes et al. [5]



VII. Conclusion

In the paper we have provided a description for the convergence of the recursive estimation procedure introduced in [24] for Hidden Markov Models using the theory of [4]. For this purpose we considered a general class of Markov models in which a simple Markov process was passed through an exponentially stable non-linear system.

VIII. Acknowledgements

The authors acknowledge the support of the National Research Foundation of Hungary (OTKA) under Grant no. T 047193.

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. Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincar Probab. Statist.*, 33 (6):697–725, 1997.

[3] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37:1559–1563, 1966.

[4] A. Benveniste, M. Metivier and P. Priouret. Adaptive Algorithms and Stochastic Approximations. Volume 22 of Application of Mathematics, Springer Verlag, New York, 1990.

[5] I. Berkes, J. Horváth, and P. Kokoszka. GARCH processes: structure and estimation, *Bernoulli*, 9: 201–217, 2003.

[6] R. Bhattacharya, E. C, Waymire, An Approach to the Existence of Unique Invariant Probabilities for Markov Processes, *Limit theorems in probability and statistics, János Bolyai Math. Soc.*, I (Balatonlelle 1999), 181–200, 2002.

[7] R. Bhattacharya and E. C. Waymire. An approach to the existence of unique invariant probabilities for Markov processes. *Limit theorems in probability and statistics, János Bolyai Math. Soc.*, Vol. I:181–200, 2002.

[8] T. Bollerslev. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[9] P. Bougerol, N. Picard (1992). Stationarity of GARCH processes and of some nonnegative time series. *Journal of Econometrics* 52: 115-127

[10] M. Chen, H. An (1998). A Note on the Stationarity and the Existence of Moments of the GARCH model. *Statistica Sinica* 8: 505-510

[11] R. Douc and C. Matias. Asymptotics of the Maximum likelihood estimator for general Hidden Markov Models. *Bernoulli*, 7:381–420, 2001.

[12] R.J. Elliott. Recursive estimation for hidden Markov models: a dependent case. *Stochastic Anal. Appl.* 13:(4), 437–460, 1995.

[13] R.F. Engle. Autoregressive Conditional Heteroskedasticity with Estimates of The Variance of the UK Inflation. *Econometrica*, 50:987–1008, 1982.

[14] Y. Ephraim and N. Merhav. Hidden Markov Processes. *IEEE Transactions on Information Theory*, 48:1508–1569., 2002.

[15] S. Geman. Some averaging and stability results for random differential equations. *SIAM Journal of Applied Mathematics*, 36:87–105, 1979.

[16] L. Gerencsér. On a class of mixing processes. *Stochastics*, 26:165–191, 1989.

[17] L. Gerencsér, L. Rate of convergence of recursive estimators. *SIAM J. Control and Optimization*, 30 (5):1200–1227, 1992.

[18] L. Gerencsér and G. Molnár-Sáska Estimation and Strong Approximation of Hidden Markov Models. Lecture Notes in Control and Information Sciences, Springer, 294:313–320, 2003.

[19] L. Gerencsér, G. Molnár-Sáska, Gy. Michaletzky, and G. Tusnády. New methods for the statistical analysis of Hidden Markov Models. *IEEE Trans. on Automatic Control*, 2003. submitted.

[20] Y. Kifer. Ergodic Theory of Random Transformation. *Progress in Probability ans Statistics*, 10, 1986.

[21] V. Krishnamurthy and G. Yin Recursive algorithms for estimation of hidden Markov models and autoregressive models with Markov regime. *IEEE Trans. Inform. Theory*, 48(2):458–476, 2002.

[22] H.J. Kushner and G. Yin Stochastic Approximation Algorithms and Applications *Springer-Verlag, New York*, 1997.

[23] F. LeGland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems*, 13:63–93, 2000.

[24] F. LeGland and L. Mevel. Recursive Identification of HMM's with Observation in a Finite Set. In *Proc. of the 34th IEEE CDC*, 216-221, 1995.

[25] F. LeGland and L. Mevel. Recursive Estimation in Hidden Markov Models. In *Proc. of the 36th IEEE CDC*, 3468-3473, 1997.

[26] B. G. Leroux. Maximum-likelihood estimation for hidden Markov-models. *Stochastic Processes and their Applications*, 40:127–143, 1992.

[27] L. Ljung. On consistency and identifiability. *Mathematical Programming Study*, 5:169–190., 1976.

[28] J. Rissanen and P.E. Caines. The strong consistency of maximum likelihood estimators for ARMA processes. *Ann. Statist.*, 7:297 – 315., 1979.

[29] T. Rydén. On recursive estimation for hidden Markov models. *Stochastic Process. Appl.* 66 (1):79–96., 1997.