Improved estimation of the exponential stability of the predictive filter in Hidden Markov Models

László Gerencsér, Gábor Molnár-Sáska and György Michaletzky

***Abstract*— We consider finite state continuous read-out Hidden Markov Models. The exponential stability of the predictive filter was investigated in [16] when the transition probability matrix Q of the underlying Markov chain is primitive. We carry out further investigation of this exponential stability. Two important applications are derived: the strong approximation result has been extended for HMMs with primitive transition probability matrices and the validity of the recursive estimation of HMMs with primitive transition probability matrices has been shown.**

I. Introduction

Hidden Markov Models have become a basic tool for modelling stochastic systems with a wide range of applicability. For a general introduction see [6]. The estimation of the dynamics of a Hidden Markov Model is a basic problem in applications.

Let (X_n, Y_n) be a Hidden Markov process, i.e. (X_n) is a homogenous Markov process with state space $\mathcal{X}$ and the observation sequence (Y_n) is conditionally independent and identically distributed given the σ-field generated by the process (X_n).

Example 1.1: 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 $h : \mathcal{X} \to R$ is measurable.

Throughout the paper let the state space of the Hidden Markov Model be finite now, i.e. $|\mathcal{X}| = N$.

Let Q^* be the transition probability 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. Throughout the paper we deal with parametric problems, i.e. the unknown quantities depend on a parameter. The true value of the parameter (or the unknown quantities) is the one which is used to generate the process.

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 Morgan Stanley Budapest, Hungary and MTA SZTAKI, Computer and Automation Institute of the Hungarian Academy of Sciences, 1111 Budapest, Hungary gabor.molnar-saska@morganstanley.com

Gy. Michaletzky is with ELTE, Eotvos Lorand University, Budapest, Hungary michgy@ludens.elte.hu

If $\mathcal{Y}$ is finite, say $|\mathcal{Y}| = M$, then conditional independence can be written as

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

In this case we will use the following notation:

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

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

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

where λ is a fixed nonnegative, σ-finite measure. Let us introduce the following notations:

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

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

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

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

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

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

Writing $p^*_{n+1} = (p^{*1}_{n+1}, \dots, p^{*N}_{n+1})^T$, we know from [3] that the filter process satisfies the Baum-equation

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

both in discrete and continuous read-out cases, where π is the normalizing operator: for $x \in R^N$, $x \geq 0$, $x \neq 0$ set $\pi(x)^i = x^i / \sum_{j=1}^N x^j$. 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) as well as 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:

$$p_{n+1} = \pi(Q^T B(Y_n) p_n), \tag{3}$$

with initial condition $p_0 = q$, where $Q \in R^{N \times N}$ is a stochastic matrix, $B(y) = \mathrm{diag}(b^i(y))$ is a collection of conditional probabilities, and $q \in R^N$ is a probability vector, i.e. $q^i \geq 0$ for $i = 1, \dots N$ and $\sum_{i=1}^N q^i = 1$.

We will take an arbitrary probability vector q as initial condition, and the solution of the Baum equation will be denoted by $p_n(q)$.

A key element in the statistical analysis of HMM-s is the exponential stability of the predictive filter, i.e. the distance between iterates $p_n(q)$ and $p_n(q')$ goes to zero exponentially fast, where q, q' are arbitrary initializations. This has been established in [16] (Theorem 2.1) as follows: Let

$$\delta(y) = \frac{\max_x b^x(y)}{\min_x b^x(y)} \tag{4}$$

and

$$\epsilon = \min_{x,x'}{}^{+} q(x, x'), \tag{5}$$

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

Proposition 1.1: If the stochastic matrix Q is primitive, with index of primitivity r, then for any q, q', any integer $n \geq r$ and any sequence $y_1, \dots y_n \in \mathcal{Y}$ we have

$$\|p_n(q) - p_n(q')\| \leq \epsilon^{-r}\delta(y_0)\dots\delta(y_{r-1})$$
$$\prod_{k=1}^{\lfloor n/r \rfloor} (1 - \epsilon^r(\delta(y_{kr-r+1})\dots\delta(y_{kr-1}))^{-1})\|q - q'\|.$$

Proposition 1.1 implies the following Proposition, see also Theorem 2.2 in [16].

Proposition 1.2: Assume that $Q, Q^* > 0$ and $b^x(y), b^{x*}(y) > 0$ for all x, y. Let q, q' be any two initializations. Then for some $0 < \rho < 1$ and a fix constant C

$$\|p_n(q) - p_n(q')\|_{TV} \leq C\rho^n\|q - q'\|_{TV}, \tag{6}$$

where $\|\cdot\|_{TV}$ denotes the total variation norm.

That is, the filter forgets its initial condition with an exponential rate. An essential feature of the result is that $\|q - q'\|_{TV}$ shows up in the upper bound, see [2]. We note that Proposition 1.2 is a purely linear algebraic statement, i.e. there is no need for probability.

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (6) holds with a random C. The main result of this paper is the following:

Theorem 1.1: Assume that Q, Q^* are primitive matrices, with index of primitivity r, and $b^x(y), b^{x*}(y) > 0$ for all x, y. Furthermore, assume that

$$\int |\delta(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{7}$$

Let q, q' be any two initializations. Then for an arbitrary $s > 1$ there exists a $0 < \rho < 1$ and a random variable C, such that

$$\|p_n(q) - p_n(q')\|_{TV} \leq C\rho^n\|q - q'\|_{TV}, \tag{8}$$

and the s-th moment of C does exist.
This is an improvement of Theorem 2.2 of [16].

The rest of the paper is organized as follows. In section II we introduce the main technical tools. In section III we prove (8) for finite state continuous read-outs Hidden Markov Models. For this we prove that if X_n is a Markov process satisfying the Doeblin condition and g is a measurable function with negative values, then for any $s > 1$ there exists an $\alpha > 0$ and a random variable C such that

$$\sum_{k=1}^{n}(g(X_k) + \alpha) \leq \log C \tag{9}$$

and the s-th moment of C does exist, see Theorem 3.1.

In section IV two applications of the result of Theorem 1.1 are given. In section IV-A we extend the results of [7] to finite state continuous read-out Hidden Markov Models with primitive transition probability matrix, see also [9]. In section IV-B we present the results of [10] for Hidden Markov Models with primitive transition probability matrix.

II. Representation of Markov processes

Consider a Polish space $\mathcal{X}$ and a sequence of independent, $[0, 1]$-uniform random variables (U_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

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

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

A converse result is given in the following proposition:

Proposition 2.1: Let (X_n) be a Markov process on a Polish space $\mathcal{X}$ with transition probabilities $P(x, G)$, $x \in \mathcal{X}$, $G \in \mathcal{B}(\mathcal{X})$. Then there exists a Borel measurable function $f : \mathcal{X} \times [0, 1] \longrightarrow \mathcal{X}$ such that, with U being uniform in $[0, 1]$ over some probability space $(\Omega, \mathcal{F}, \mathcal{Q})$, for all $x \in \mathcal{X}$ and $G \in \mathbf{B}(\mathcal{X})$ we have

$$P(x, G) = \mathcal{Q}\{f(x, U) \in G\}.$$

For the proof see [11]. In the following we will denote the random mapping $f(\cdot, U_{n-1})$ by T_n, i.e. for $x \in \mathcal{X}$

$$T_n x = f(x, U_{n-1}).$$

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 [5]:

Definition 2.1: 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 [5], shows the relation between the Doeblin-condition and the representation of the Markov chain.

Lemma 2.1: 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 $\mathbf{Q}(T_n \in \Gamma_c) \geq \delta$, where Γ_c is the set of constant mappings.

Proposition 2.2: Assume that the Doeblin-condition holds with $m = 1$ for a Markov chain (X_n). Then there exists an invariant distribution π, and

$$|P^n(x, A) - \pi(A)| \leq (1 - \delta)^n \quad \text{for} \quad \forall A \in \mathcal{B}(\mathcal{X}). \quad (10)$$

Now let (X_n, Y_n) be a Hidden Markov process and assume that the state space $\mathcal{X}$ and the observed space $\mathcal{Y}$ are Polish. The following lemma is given in [8].

Lemma 2.2: Assume that the Doeblin-condition holds with $m = 1$ for the Markov chain (X_n). Then the Doeblin-condition holds for (X_n, Y_n) as well.

The Doeblin-condition can be defined in more general form.

Definition 2.2: If there exists $m \geq 1$ such that $P^m(x, A) \geq \delta\nu(A)$ is valid for $\forall x \in \mathcal{X}$ and $A \subset \mathcal{B}(\mathcal{X})$ with some probability measure ν then we say that the general Doeblin-condition is valid in order m.

Proposition 2.3: (see [5]) Let (X_n) be a Markov chain. The general Doeblin-condition is valid if and only if there exists a sequence of i.i.d. mappings (T_n) such that $Q(T_m \dots T_1 \in \Gamma_c) \geq \delta$

Remark 2.1: Let (X_n) be a Markov chain. The Doeblin-condition is valid with $m \geq 1$ if and only if there exists a representation such that $\mathbf{Q}(T_n \dots T_{n-m+1} \in \Gamma_c) \geq \delta$, where Γ_c is the set of constant mappings. Thus Proposition 2.2 and Lemma 2.2 also valid if the Doeblin-condition holds for $m \geq 1$.

III. Exponential stability of the predictive filter

Consider a Markov chain (X_n) on an arbitrary abstract measurable space $\mathcal{X}$. Assume that the Doeblin condition is satisfied for this Markov chain with $m = 1$. Furthermore, assume that $g : \mathcal{X} \to R^-$ is a strictly negative measurable function. Then we have the following theorem.

Theorem 3.1: Let X_k be a Markov process satisfying the Doeblin condition with $m = 1$. Assume that $g(X_k) < 0$. Then for any $s > 1$ there exist $\alpha > 0$ and a random variable C such that

$$\sum_{k=1}^{n} (g(X_k) + \alpha) \leq \log C \quad (11)$$

and the s-th moment of C does exist.

First we prove Theorem 3.1 in the special case when the Markov process is an i.i.d. sequence.

Theorem 3.2: Let ξ_k be i.i.d random variables such that $\xi_k < 0$. Then for any $s > 1$ there exist $\alpha > 0$ and a random variable C such that

$$\sum_{k=1}^{n} (\xi_k + \alpha) \leq \log C$$

and the s-th moment of C does exist.

Note that, these kind of results are very important in risk theory. Let $Y_k = \xi_k + \alpha$ be a sequence of independent, identically distributed random variables with negative expected values. Assume that Y_k denotes the net payment of an insurance company for the k-th period (net payment is the difference between the fees received from the clients and the amount due to damages). To determine the reserve fund of the insurance company it is crucial to estimate the expression $\sup_n \sum_{k=1}^{n} Y_k$, see [17].

Proof: (Theorem 3.2) Let us define the function Ψ as follows

$$\Psi(u) = P(\sup_n \sum_{k=1}^{n} (\xi_k + \alpha) > u).$$

It is enough to prove that if $\alpha > 0$ is small enough, then

$$\Psi(u) < e^{-su}. \quad (12)$$

For this, consider the following proposition.

Proposition 3.1: Let $g(r) = E(e^{r(\xi_k + \alpha)})$. If $E(\xi_k) + \alpha < 0$, then there can be at most one solution of the equation

$$g(R) = 1. \quad (13)$$

If $\alpha > 0$, then there exists a solution of (13).

Proof: Let $Z_k = \xi_k + \alpha$. $g(r)$ is trivially a continuous function of r. Furthermore $g(0) = 1$ and $g'(0) = E(Z_k) < 0$. Since $g(r)$ is a mixture of strictly convex functions (namely e^{rx}), it is also a strictly convex function. Thus there can be at most one solution of (13).

Furthermore, if $\alpha > 0$, then $P(Z_1 \leq 0) < 1$, thus there exists a positive constant x_0 such that $P(Z_1 \leq x_0) < 1$. Hence

$$E(e^{rZ_1}) \geq e^{rx_0}(1 - P(Z_1 \leq x_0)).$$

Since $e^{rx_0}(1 - P(Z_1 \leq x_0)) \to \infty$ if $r \to \infty$ there exists one solution of (13). ∎

Let us continue the proof of Theorem 3.2. Choose $\alpha > 0$ such that for the solution of (13) we have $R > q$. By Proposition 3.1 we have that the solution of (13) exists for any $\alpha > 0$. Furthermore, $E(e^{s(\xi_k + \alpha)})$ is a continuous and monotone decreasing function of α ($s > 1$ fix), and $E(e^{s\xi_k}) < 1$ since ξ_k is a negative random variable. Thus there exists a positive α such that $E(e^{s(\xi_k + \alpha)}) < 1$. Let us denote it by α^*. For this α^* consider the solution of (13). Let the solution be R^*. We have that

$$R^* > s. \quad (14)$$

Consider the following statement of Sparre Andersen [1]
Proposition 3.2:

$$\Psi(u) \leq e^{-R^* u}.$$

Using the relation (14) and Proposition 3.2 we get (12) and the proof of Theorem 3.2. ∎

Let us turn to the proof of Theorem 3.1.

Proof: Let T_n be the representation of the Markov process (X_n) defined in Proposition 2.1. By Lemma 2.1 we have that $P(T_n \in \Gamma_c) > \delta$. Let us define the following sequence of stopping times

$$\tau_0 = 0$$

and for $k \geq 1$

$$\tau_k = \min_n \{n > \tau_{k-1} : T_n \in \Gamma_c\}.$$

Observe that $\xi_k = \sum_{j=\tau_k}^{\tau_{k+1}-1} g(X_j)$ is a sequence of i.i.d. random variables.

Using that for all $k > 0$ $g(X_k) < 0$, we have that

$$\sup_n \sum_{k=1}^{n} g(X_k) < \sup_t \sum_{k=\tau_1}^{\tau_t} g(X_k) = \sup_t \sum_{k=1}^{t} \xi_k \qquad (15)$$

Using Theorem 3.2 we have that for $2s$ there exist an $\alpha_1 > 0$ and a random variable C_1 such that

$$\sum_{k=1}^{t} \xi_k \leq \log C_1 - \alpha_1 t, \qquad (16)$$

and the $2s$-th moment of C_1 does exists. Note that α_1 can be chosen such that $\alpha_1 < 1$.

Here

$$t = \sum_{k=1}^{n} \chi_{\Gamma_c}(T_k).$$

Since $\chi_{\Gamma_c}(T_k)$ are independent, identically distributed random variables, from Theorem 3.2 we have that there exist an $\alpha_2 > 0$ and a random variable C_2 such that

$$t = \sum_{k=1}^{n} \chi_{\Gamma_c}(T_k) > \alpha_2 n - \log C_2, \qquad (17)$$

and the $2s$-th moment of C_2 does exists.

From (16) and (17) we get that

$$\sum_{k=1}^{t} \xi_k \leq \log C_1 - \alpha_1(\alpha_2 n - \log C_2) \qquad (18)$$

By (15) we have that

$$\sup_n \sum_{k=1}^{n} g(X_k) < \log C_1 + \alpha_1 \log C_2 - (\alpha_1\alpha_2)n.$$

Choose $\alpha = \alpha_1\alpha_2$ and $C = C_1 C_2^{\alpha_1}$. Using the Hölder inequality we have

$$E|C_1 C_2^{\alpha_1}|^s \leq (E|C_1|^{2s})^{1/2} (E|C_2|^{2s\alpha_1})^{1/2}.$$

Since the $2s$-th moment of C_1 and C_2 exist and $\alpha_1 < 1$, the s-th moment of C does exist. Thus we have finished the proof of Theorem 3.1. ■

Remark 3.1: If X_k is a Markov process satisfying the Doeblin condition with $m > 1$, then Theorem 3.1 is also valid.

Let us turn to the main result of this paper, i.e. to the proof of Theorem 1.1. First, we consider an important technical result.

Lemma 3.1: Let X_n be a Markov process with a primitive transition probability matrix Q with index of primitivity r. Then the process $U_n \in \mathcal{X}^r$ defined by

$$U_n = (X_{(n-1)r+1}, \dots X_{nr})$$

satisfies the Doeblin condition with $m = 2$.

Proof: (Theorem 1.1) From Proposition 1.1 we have that

$$\|p_n(q) - p_n(q')\| \leq \epsilon^{-r} \delta(y_0) \dots \delta(y_{r-1}) \prod_{k=1}^{\lfloor n/r \rfloor} (1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1}) \|q - q'\|.$$

First, we are looking for an upper bound for the product term

$$\prod_{k=1}^{\lfloor n/r \rfloor} (1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1}) \leq \widetilde{C}\rho^n, \qquad (19)$$

such that $0 < \rho < 1$ and the s-th moment of the random variable $\widetilde{C}$ does exist. Taking the logarithm of both sides we have

$$\sum_{k=1}^{\lfloor n/r \rfloor} \log(1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1}) \leq \log \widetilde{C} + n \log \rho.$$

Observe that $\log(1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1})$ is a sequence of bounded random variables. Indeed, $\delta(y) \geq 1$ and $\epsilon > 0$, thus we have that

$$1 - \epsilon^r \leq 1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1} < 1.$$

By Lemma 3.1 and Lemma 2.2 we have that the process

$$U_k = ((X_{kr-r+1}, Y_{kr-r+1}), \dots (X_{kr}, Y_{kr}))^T$$

satisfies the Doeblin condition. Thus using Theorem 3.1 with $g(U_k) = \log(1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1}$ we get that there exist $\alpha > 0$ and a random variable $\widetilde{C}$ such that

$$\sum_{k=1}^{\lfloor n/r \rfloor} \log(1 - \epsilon^r (\delta(y_{kr-r+1}) \dots \delta(y_{kr-1}))^{-1}) \leq \log \widetilde{C} + \lfloor n/r \rfloor \alpha$$

and the s-th moment of $\widetilde{C}$ does exist. Let $\rho = \exp(-\alpha/r)$ and we get (19). ■

IV. Applications

In this section we give two important applications of the result of Theorem 1.1. In section IV-A we extend the results of [7] to finite state continuous read-out Hidden Markov Models with primitive transition probability matrix, see also [9]. In section IV-B we present the results of [10] for Hidden Markov Models with primitive transition probability matrix.

A. Strong approximation

Let $G \subset R^r$ be an open set, $D \subset G$ be a compact set, and $D^* \subset \mathrm{int}D$ be another compact set, where $\mathrm{int}D$ denotes the interior of D. Assume that for the true value of the parameter we have $\theta^* \in D^*$. Furthermore, assume that for an estimation of the parameter of the Hidden Markov Model we have $\theta \in D$. We will refer to D^* and D as compact domains.

Consider the following estimation problem: let Q and b be parameterized by $\theta \in D$ and let

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

In this paragraph we always consider finite state-space and continuous read-out space. Although the results of this section are valid for a general read-out space, we will always assume that $\mathcal{Y}$ is a measurable subset of R^d and λ is the Lebesgue-measure. Assume that the densities $b^x(y,\theta)$ are with respect to the Lebesgue measure λ.

In the finite case (when both $\mathcal{X}$ and $\mathcal{Y}$ are finite) θ is often the parameter of the model parameterizing the transition matrix Q and the conditional read-out probabilities $b^i(y)$. Usually the entries of Q are included in θ.

Consider a Hidden Markov Model (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is a measurable subset of R^d. Let $Q(\theta), Q^*$ be primitive transition probability matrices with index of primitivity r and $b^i(y,\theta), b^{*i}(y) > 0$ for all i, y. Let the initialization of the process (X_n, Y_n) be random, where the Radon-Nikodym derivative of the initial distribution π_0 w.r.t the stationary distribution π is bounded, i.e.

$$\frac{d\pi_0}{d\pi} \leq K. \tag{20}$$

Assume that for all $i, j \in \mathcal{X}$, $\theta \in D$ and $q \geq 1$

$$\int |\log b^j(y,\theta)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{21}$$

To estimate the unknown parameter we use the maximum-likelihood (ML) method. Let the log-likelihood function be

$$L_N = \sum_{n=1}^{N} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta).$$

We shall refer to this as the cost function associated with the ML estimation of the parameter. The right hand side depends on θ^* through the sequence (Y_n). To stress the dependence of L_N on θ and θ^* we shall write $L_N = L_N(\theta, \theta^*)$. The ML estimation $\widehat{\theta}_N$ of θ^* is defined as the solution of the equation

$$\frac{\partial}{\partial\theta} L_N(\theta, \theta^*) = L_{\theta N}(\theta, \theta^*) = 0 \tag{22}$$

More exactly $\widehat{\theta}_N$ is a random vector such that $\widehat{\theta}_N \in D$ for all ω and if the equation (22) has a unique solution in D, then $\widehat{\theta}_N$ is equal to this solution. By the measurable selection theorem such a random variable does exist.

Let us introduce the asymptotic cost function

$$W(\theta, \theta^*) = \lim_{n\to\infty} E_{\theta^*} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta). \tag{23}$$

In [8] we have proved that this limit exists for all $\theta \in D$. Assume that the function $W(\theta, \theta^*)$ is smooth in the interior of D, i.e. the third derivative exists. Furthermore, assume that the following technical conditions are satisfied.

Condition 4.1:

$$\int \left| \frac{\max_x \|\partial^k b^x(y)/\partial\theta^k\|}{(\min_x b^x(y))^{2^l}} \right|^q b^{*i}(y)\lambda(dy) < \infty$$

is satisfied where $k = 0, 1, 2, 3$ and $l = 0, 1, 2$

Under Condition 4.1 we have

$$W_\theta(\theta, \theta^*) = \lim_{n\to\infty} E_{\theta^*} \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta), \tag{24}$$

and for the Fisher-information matrix we have

$$I^* = W_{\theta\theta}(\theta^*, \theta^*) =$$

$$\lim_{n\to\infty} E_{\theta^*} \left((\phi_n)^T (\phi_n) \right),$$

where $\phi_n = (\frac{\partial}{\partial\theta} \log p(y_n|y_{n-1}, \ldots, \theta^*))$, see [7] for details.

Remark 4.1: Note that $W_\theta(\theta^*, \theta^*) = 0$.

Consider the following identifiability condition:

Condition 4.2: The equation

$$W_\theta(\theta, \theta^*) = 0$$

has exactly one solution in D, namely θ^*.

The following characterization theorem for the error term of the off-line ML estimation is an extended version of [7] for Hidden Markov Models with primitive transition probability matrix, see also [9].

Theorem 4.1: Consider a Hidden Markov Model (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is a measurable subset of R^d. Let Q, Q^* be primitive transition probability matrices and $b^i(y), b^{*i}(y) > 0$ for all i, y. Assume that Condition 4.1 is satisfied. Let $\hat{\theta}_N$ be the ML estimate of θ^*. Furthermore assume that the identifiability Condition 4.2 is satisfied. Then

$$\hat{\theta}_N - \theta^* =$$

$$-(I^*)^{-1} \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta^*) + O_M(N^{-1}), \tag{25}$$

where I^* is the Fisher-information matrix.

A key point in Theorem 4.1 is that the error term is $O_M(N^{-1})$. This ensures that all basic limit theorems, that are known for the dominant term, which is a martingale, are also valid for $\hat{\theta}_N - \theta^*$.

Let us consider now the case when the read-out space is finite.

Theorem 4.2: Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let $Q(\theta), Q^*$ be primitive matrices and $b^i(y,\theta), b^{*i}(y) > 0$ for all i, y. Assume that Q and b are smooth in θ, i.e. the third derivatives exist. Let $\hat{\theta}_N$ be the ML estimate of θ^*. Assume that the identifiability condition 4.2 is satisfied. Then

$$\hat{\theta}_N - \theta^* =$$

$$-(I^*)^{-1} \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta^*) + O_M(N^{-1}), \tag{26}$$

where I^* is the Fisher-information matrix.

B. Recursive estimation

A recursive estimation method for Hidden Markov Models has been proposed in [14] and [15]. As suggested the proposed recursive algorithm could be analyzed via the theory of stochastic approximations, see [4]. Krishnamurthy and Yin, see [12], investigated the convergence and rate of convergence of the recursive estimation of HMMs using the weak convergence approach of Kushner and Yin [13].

The purpose of [10] is to verify the basic probabilistic conditions for HMMs, given in Part II, Chapter 1 of [4]. Theorem 1.1 allows us to extend the results of [10] for a wider class of Hidden Markov Models.

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

$$\mathbf{p}_{n+1}(\theta) = \frac{Q^T(\theta)B(y_n,\theta)\mathbf{p}_n(\theta)}{\mathbf{b}(y_n,\theta)^T\mathbf{p}_n(\theta)} = \Phi_1(y_n,\mathbf{p}_n,\theta). \quad (27)$$

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

$$W_{n+1} = Q^T\left(I - \frac{B(y_n)\mathbf{p}_n\mathbf{e}^T}{\mathbf{b}^T(y_n)\mathbf{p}_n}\right)\frac{B(y_n)W_n}{\mathbf{b}^T(y_n)\mathbf{p}_n} + F, \quad (28)$$

where

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

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

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

$$H(\theta,u) = H(\theta,x,y,\mathbf{p},W) = \frac{\beta(y,\theta)\mathbf{p} + W\mathbf{b}(y,\theta)}{\mathbf{b}(y,\theta)^T\mathbf{p}}, \quad (29)$$

and consider the following adaptive algorithm.

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n}H(\overline{\theta}_n, x_n, y_n, \overline{\mathbf{p}}_n, \overline{W}_n), \quad (30)$$

$$\overline{\mathbf{p}}_{n+1} = \Phi_1(y_n, \overline{\mathbf{p}}_n, \overline{\theta}_n), \quad (31)$$

$$\overline{W}_{n+1} = \Phi_2(y_n, \overline{\mathbf{p}}_n, \overline{W}_n, \overline{\theta}_n). \quad (32)$$

For the convergence of this algorithm the approach of Benveniste, Metivier and Priouret, see [4], was used in [10]. Here we restate the main theorem of [10] for Hidden Markov Models with primitive transition probability matrix.

Theorem 4.3: Consider a Hidden Markov Model with finite state space and finite read-out space. Assume that Q^* is a primitive transition probability matrix, $b^{*x}(y) > 0$ and for all θ we have that $Q(\theta)$ is primitive and $b^x(y,\theta) > 0$ for all x, y. Assume Condition 4.2. Then the algorithm defined by (30), (31), (32) converges to the true value θ^* with probability arbitrary close to 1.

Acknowledgement

This research was supported by the the National Research Foundation of Hungary (OTKA) under Grant no. T 047193.

REFERENCES

[1] E. Sparre Andersen. On the collective theory of risk in the case of contagion between the claims. In *Transaction XV-th Int. Congress of Actuaries*, pages 219–229, 1957.

[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. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*. Springer-Verlag, Berlin, 1990.

[5] 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.

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

[7] L. Gerencsér and G. Molnár-Sáska. Strong approximation of Hidden Markov Models. *IEEE Trans. on Automatic Control*, submitted.

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

[9] L. Gerencsér, G. Molnár-Sáska, Gy. Michaletzky and G. Tusnády New methods for the statistical analysis of Hidden Markov Models In *Proceedings of the 41th IEEE Conference on Decision & Control*, 2272-2277, 2002.

[10] L. Gerencsér, G. Molnár-Sáska, and Zs. Orlovits. Recursive estimation of Hidden Markov Models. In *Proceedings of the 44th IEEE Conference on Decision & Control*, page MoB16.3, 2005.

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

[12] 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.

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

[14] F. LeGland and L. Mevel. Recursive Identification of HMM's with Observation in a Finite Set. In *Proceedings of the 34th IEEE Conference on Decision & Control*, pages 216–221, 1995.

[15] F. LeGland and L. Mevel. Recursive Estimation in Hidden Markov Models. In *Proceedings of the 36th IEEE Conference on Decision & Control*, pages 3468–3473, 1997.

[16] F. LeGland and L. Mevel. Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models. *Mathematics of Control, Signals and Systems*, 13:63–93, 2000.

[17] H.H. Panjer and G.E. Willmot. *Insurance Risk Models*. Society of Actuaries, 1992.