15th Triennial World Congress, Barcelona, Spain

A NEW METHOD FOR THE ANALYSIS OF HIDDEN MARKOV MODEL ESTIMATES

László Gerencsér * **Gábor Molnár-Sáska** **

* *Computer and Automation Institute*
Hungarian Academy of Sciences
13-17 Kende u., Budapest 1111, Hungary
gerencser@sztaki.hu
** *Computer and Automation Institute*
Hungarian Academy of Sciences
13-17 Kende u., Budapest 1111, Hungary
and
Technical University of Budapest
1-3 Egry J. u., Budapest 1111, Hungary
molnar@math.bme.hu

Abstract.
The estimation of Hidden Markov Models has attracted a lot of attention recently, see results of (Le Gland and Mevel, 2000), (Leroux, 1992), (Mevel and Finesso, 2000). The purpose of this paper is to lay the foundation for a new approach for the analysis of the maximum-likelihood estimation of HMM-s, using representation of HMM-s due to (Borkar, 1993). Useful connection between the estimation theory of HMM-s and linear stochastic systems is established via the theory of L-mixing processes developed in (Gerencsér, 1988).

Key words. Hidden Markov Models, stochastic systems, random transformations, Doeblin condition, L-mixing processes, maximum-likelihood estimation,

1. INTRODUCTION

Hidden Markov Models have become a basic tool for modeling stochastic systems with a wide range of applicability in such diverse areas as robotics telecommunication, econometrics and protein research.

The estimation of the dynamic of a Hidden Markov Model is a basic problem in applications. A key element in statistical analysis of HMM-s is a strong law of large numbers for the log-likelihood function. In previous works stability theory of Markov chains and the subadditive ergodic theorem were used (Le Gland and Mevel, 2000), (Leroux, 1992), (Mevel and Finesso, 2000). Although these tools are very powerful, they do not yield a LNN with guaranteed rate of convergence. An alternative tool that can be widely used in system identification is theory of L-mixing processes. The relevance of this theory will be established in this paper using a random-transformation representation for Markov-processes (see (Kifer, 1986), (Borkar, 1993)). The advantage of this approach is that, potentially a more precise characterization of the estimation error-process can be obtained, which, in turn, is crucial for the analysis of the performance of adaptive prediction, see (Gerencsér, 1990).

2. HIDDEN MARKOV MODELS

Hidden Markov Models are based on a Markov chain $\{X_n\}$ which describes the evolution of the state of a system. Given a realized sequence of state variables $\{x_n\}$, observed variables $\{Y_n\}$ are conditionally independent, with the distribution of Y_n depending on the corresponding state x_n. In many estimation problems the distribution of Y_n is assumed to belong to a parametric family and the state space is assumed to finite. The original model was introduced in (Baum and Petrie, 1966).

We are going to introduce the exact definition of the Hidden Markov Modell (X_n, Y_n).

Definition 2.1. Let $(X_n) \in \mathcal{X}$ be a finite space homogenous Markov chain. We can observe an $(Y_n) \in \mathcal{Y}$ process, where $\mathcal{Y}$ is a Polish space. We assume that the probability of observations are conditionally independent

$$P(Y_n = y_n, \ldots Y_0 = y_o | X_n = x_n, \ldots X_0 = x_0) =$$

$$\prod_{i=0}^{n} P(Y_i = y_i | X_i = x_i).$$

Although the observed space can be any Polish space we are going to talk about the discrete or Euclidean cases only. Through the article we assume the observed space $\mathcal{Y}$ is discrete.

We will use the following notations

$$P(Y_k = y_k | X_k = x_k) = P(y_k | x_k),$$
$$P(y|x) = b^x(y) \quad B(y) = \mathrm{diag}(b^i(y)).$$

An easy statement can be obtained

Proposition 2.1. If (X_n, Y_n) is a Hidden Markov process, then $Z_n = (X_n, Y_n)$ is a Markov process

For further notation let $Q > 0$ be the transition matrix of the unobserved Markov process (X_n),

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

$(p_{n+1} = (p^1_{n+1}, \ldots, p^N_{n+1})^T.$)

The filter process is generated by the Baum-equation introduced in (Baum and Petrie, 1966)

$$p_{n+1} = \pi(Q^* B(Y_n) p_n). \tag{1}$$

where π is the normalising operator to make p_{n+1} a probability vector.

In (Le Gland and Mevel, 2000) the following basic theorem is proved:

Theorem 2.1. Let $Q > 0$ and let q and q' are different starting points, which are compatible with Y_0. Then

$$\|p_n(q) - p_n(q')\| \le C(1-\delta)^n.$$

3. REPRESENTATION OF MARKOV PROCESS

The material of this section is based on (Borkar, 1993) and (Bhattacharya and Waymire, 1999).

Let the state space M be discrete, $\mathcal{X} : M \longrightarrow M$ the space of mappings. Let assume, that $\mathcal{X}$ is measureable with a probability measure m on it. Finally let T_n be i.i.d mappings according to m. In this case the process $X_0 \in M$, $X_{n+1} = T_{n+1} X_n$ is Markov.

Conversely if we have a Markov-process with transition probabilities $P(x, G)$ $(x \in M, \quad G \in \mathcal{B}(M)$, where $\mathcal{B}(M)$ is the algebra of Borel-sets, and $P(x, .)$ is a probability measure on $\mathcal{B}(M)$-n) then we can find its representation in the form $P(x, G) = m\{T : Tx \in G\}$ with a measure m on $\mathcal{X}$

$$P(x, G) = m\{T : Tx \in G\}$$

see (Kifer, 1986) . The representation can be given in a constructive way but it it should be noted that it is not unique.

Next we are going to introduce the notion of Doeblin condition (see (Bhattacharya and Waymire, 1999))

Definition 3.1. Given a Markov chain $(X_n) \in M$. If for $\forall x \in M$ and $A \subset \mathcal{B}(M)$ the inequality $P(x, A) \ge \delta\nu(A)$ is true, where $\delta > 0$ and ν probability measure we say that the Doeblin condition is satisfied.

In fact δ shows the weight of the i.i.d. factor of a Markov chain. The following lemma (see (Bhattacharya and Waymire, 1999)) shows the relation between the Doeblin condition and the representation of the Markov chain.

Lemma 3.1. The Doeblin condition is valid for an (X_n) Markov chain if and only if there exists an i.i.d. representation T_n with $P(T_n \in \Gamma_c) \ge \delta$, where Γ_c is the set of the constant mappings.

Proof. First let us assume that there exists a representation T_n. In this case $P(x, A) = P(T_1 x \in A) \ge P(T_1 x \in A | T_1 \in \Gamma_c) P(T_1 \in \Gamma_c) \ge m(A)\delta$, where m is the probability measure.

On the other hand assume that the Doeblin condition is valid. In this case we can choose an x or a T mapping with probability δ or $1-\delta$ respectively according to ν. T is received from a representation of a Markov chain with kernel function

$$\frac{P(x, A) - \delta\nu(A)}{(1-\delta)^{-1}} = \overline{P}(x, A).$$

□

Theorem 3.1. Let us assume that the Doeblin condition holds for a Markov chain X_n. In this case there exists an invariant distribution π, and the following inequality is valid

$$|P^n(x, A) - \pi(A)| \le (1-\delta)^n \quad \forall A \in \mathcal{B}(M).$$

Proof. see in (Bhattacharya and Waymire, 1999) □

Now let (X_n, Y_n) be a Hidden Markov process and assume that the state space X and the observed space Y are discretes.

Lemma 3.2. Let us assume that the Doebin condition is valid for the Markov chain X_n. In this case the Doeblin condition is valid for (X_n, Y_n) as well.

Proof. Let T_n be the representation of the Markov chain as in lemma (3.1). It means there exists a sequence of i.i.d. mappings T_n such that $X_{n+1} = T_{n+1}X_n$ with $P(T_n \in \Gamma_c) \geq \delta > 0$ and T_n is independent from the starting point X_0.

Let us look now at the observations. Let $P(x, G)$ be the transition kernel of the original Markov chain X, where $x \in X$ and $G \subset \mathcal{Y}$. In this case just like in the previous cases there is an m probability measure on the space $X \longrightarrow Y$ for which $P(x, G) = m\{U : Ux \in G\}$.

With the notation $Y_n = U_n X_n$ we get $X_{n+1} = T_{n+1}X_n$, and so $Y_{n+1} = U_{n+1}T_{n+1}X_n$. So if $T_n \in \Gamma_c(X \to X)$, then $U_n T_n \in \Gamma_c(X \to Y)$, and the lemma is proved. □

The Doeblin condition can be defined in more general form.

Definition 3.2. If there exists $m \geq 1$ such that $P^m(x, A) \geq \delta\nu(A)$ is valid for $\forall x \in M \quad$ and $\quad A \subset \mathcal{B}(M)$ with some probability measure ν then we say that the general Doeblin condition is valid in order m.

Proposition 3.1. (Bhattacharya, Waymire) 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 $P(T_m \ldots T_1 \in \Gamma_c) \geq \delta$ and T_n is the representation of X_n.

4. L-MIXING PROCESSES

Now we are going to introduce a class of processes (see (Gerencsér, 1988)) called L-mixing processes, which have proved to be extremely useful in the statistical theory of linear stochastic systems (see e.g. (Gerencsér, 1990)). First of all we need the definition of M-boundedness.

Definition 4.1. The real stochastic process u_n $(n \geq 0)$ is M-bounded if for $\forall q \geq 1$

$$M_q(u) = \sup_{n \geq 0} E^{1/q}|u_n|^q < \infty$$

Let $(\mathcal{F}_n)$ and $(\mathcal{F}_n^+)$ be two sequences of monoton increasing and monoton decreasing σ-algebras, respectively such that $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent for $\forall n$.

Definition 4.2. The stochastic process u_n is L-mixing, if it is M-bounded and with

$$\gamma_q(\tau) = \sup_{n \geq \tau} E^{1/q}|X_n - E(X_n|\mathcal{F}_{n-\tau}^+)|^q,$$

$$\Gamma(q) = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty.$$

holds.

The following lemma is very useful to check whether a process is L-mixing or not.

Lemma 4.1. Let X be a random variable with $E|X|^q < \infty \quad$ for $\quad \forall q, \mathcal{G} \subset \mathcal{F}$ a σ-algebra and η is a $\mathcal{G}$ measurable random variable. In this case we have

$$E^{1/q}|X - E(X|\mathcal{G})|^q \leq 2E^{1/q}|X - \eta|^q.$$

The following lemma shows the importance of the L-mixing processes.

Proposition 4.1. Let $X_n \in X$ a Markov chain (X is a discrete space), and assume that the Doeblin condition is valid for $m = 1$, further let $g : X \longrightarrow R$ be a bounded, measureable function. In this case $g(X_n)$ is an L-mixing process.

Proof. Let $n > m$ and $n - m = \tau$. Our aim is to approximate the process $g(X_n)$. Let $F_n = \sigma\{X_0, T_k : k \leq n\}$ and $F_n^+ = \sigma\{T_k : k \geq n+1\}$. First of all we approximate X_n with $X_{n,m}^+$, where $X_{n,m}^+ = T_n \ldots T_{m+1}X^*$ and X^* is a constant. Certainly $X_{n,m}^+$ is F_m^+ measurable. It is easy to see that with the help of the previous lemma the process $g(X_n)$ is L-mixing. □

Next we consider an extension of the original Markov chain, similar to the extension of (x_n, y_n) by (p_n)

Now let X_n be a Markov chain on X and the Doeblin condition holds with $m = 1$. Let $f : X \times N \longrightarrow N$ a function, where N is a normal space. Let us look at the recursion $z_{n+1} = f(x_n, z_n)$ ($z_0 = \xi$, x_n arbitrary) and denote the solution of it by $z_n(\xi)$. First of all we introduce a definition for exponential stability:

Definition 4.3. The mapping $f(x, z)$ is uniformly exponential stable if for every sequence $\{x_n\} \quad z_n$ is bounded (independent from $\{x_n\}$) and

$$\|z_n(\xi) - z_n(\xi')\| \leq C(1-\delta)^n\|\xi - \xi'\|,$$

where $C, \delta > 0$ are independent from the sequence $\{x_n\}$.

We notice that we get a special case if N is the k dimensional Euclidean space and x_n is a $k \times k$ matrix, in this case the recursion is $Z_{n+1} = A_n Z_n$, where $Z_0 = \xi$ and $A_n \in \mathcal{A}$. The question how to choose $\mathcal{A}$ to get a uniformly exponential stable process was answered nowadays.

Theorem 4.1. Consider the process $\{X_n,\ Z_n\}$, where $Z_{n+1} = f(X_n, Z_n)$, $\quad Z_0 = \xi$, and X_0, Z_0 are independent from $\{T_n\}$ $(n \geq 1)$ ($\{T_n\}$ is obtained from the representation of the Markov chain

X_n). Let $g(x,z)$ be a bounded, measurable Lipschitz-continuous function in z. In this case $v_n = g(x_n, z_n)$ is an L-mixing process.

Proof. Let $n > m$, $\tau = n - m$, $\mathcal{F}_n, \mathcal{F}_n^+$, and $X_{n,m}^+$ be the same as before. Let the approximation of Z_n be the following: $Z_{k+1,m}^+ = f(X_{k,m}^+, Z_{k,m}^+)$, where $Z_{m,m}^+ = z^*$ is constant. Certainly, in this case $Z_{n,m}^+$ is $\mathcal{F}_m^+$ measurable.

Let $m' = n - [\frac{\tau}{2}]$. The probability that there is no coupling until m' is small (in other words there is no constant mapping in the representation), because $P(X_{m',m}^+ \neq X_{m'}) \leq (1-\delta)^{[\frac{\tau}{2}]}$. Let us denote the event, when there is no coupling until n by B, so

$$B \subset \{\omega : \quad X_{n,m}^+ \neq X_n\}$$

Now let us look at the other case, namely B^C.In this case $X_{k,m}^+ = X_k$ for all $k \geq m'$.

Consider now the following two processes:

$$Z_{k+1,m}^+ = f(X_k, Z_{k,m}^+) \quad \text{with starting point} Z_{m',m}$$

$$Z_{k+1} = f(X_k, Z_k) \quad \text{with starting point} Z_{m'}$$

Using the lemma 4.1 it is easy to see the statement of the theorem. □

At the end let us apply these general results for our Hidden Markov Model. The state space X is finite and for the transition matrix Q $Q > 0$ as we mentioned before. In this case the Doeblin condition is valid for X_n and also for the pair (X_n, Y_n), where $Y_n \in \mathcal{Y}$ is an arbitrary observable space (Polish space). With the notation $p_n^i = P(X_n = i | Y_{n-1}, \ldots, Y_0)$ we have the Baum-equation

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

.

Using the theorem 2.1 in (Le Gland and Mevel, 2000) we get

Proposition 4.2. If $Q > 0$, then (x_n, y_n, p_n) is an L-mixing process.

Let us now turn to the maximum likelihood estimation of HMM-s. Write

$$\log P(y_{n-1}, \ldots y_0, \theta) =$$

$$\sum_{k=1}^{n-1} \log P(y_k | y_{k-1}, \ldots y_0, \theta) + \log P(y_0, \theta),$$

and

$$\log P(y_k | y_{k-1}, \ldots y_0, \theta) =$$

$$\sum_x \log b^x(y_k) P(x | y_{k-1}, \ldots, y_0, \theta) =$$

$$\sum_x \log b^x(y_k) p_k^x.$$

Write

$$g(y,p) = \sum_x \log b^x(y) p^x. \tag{2}$$

First we ask, under what condition does the limit

$$\frac{1}{N} \log P(Y_n, \ldots, Y_0, \Theta) = \frac{1}{N} \sum_{k=1}^{N} g(y_k, p_k)$$

exist. Although Proposition 4.1 is not applicable since g is not bounded extension to a class of unbounded function is possible. Ultimately it is hoped that an extension to g given by (2) is possible, and then the argument of (Gerencsér, 1992) are applicable. Thus it is conjectured and strongly supported that we have:

let $\hat{\theta}_N$ be the ML estimate of θ^*, then under suitable technical conditions

$$\hat{\theta}_N - \theta^* = O_M(N^{-1}) -$$

$$-(R^*)^{-1} \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial \theta} \log P(Y_n | Y_{n-1}, \ldots, Y_0, \theta^*),$$

where R^* is the Fisher-information matrix.

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

5. CONCLUSION

We have established a link between the statistical theory of Hidden Markov Models and linear stochastic systems via the concept of L-mixing processes. This has been made possible by using a random transformation representation of HMM-s. To demonstrate the usefulness of this connection a very precise characterization of the error of the parameter-estimations of HMM-s has been formulated, as a strongly supported conjecture.

6. ACKNOWLEDGEMENT

This work was partially supported by the National Research Foundation of Hungary (OTKA) under Grant no. T 032932. The first author expresses his thanks to Jan H. van Schuppen for valuable discussions on stochastic systems.

7. REFERENCES

F. Le Gland, L. Mevel, Basic Properties of the Projective Product with Application to Products of Column-Allowable Nonnegative Matrices, Math. Control Signals Systems 13. (2000) 41-62.

F. Le Gland, L. Mevel, Exponential Forgetting and Geometric Ergodi city in Hidden Markov Models, Math. Control Signals Systems 13. (2000) 63-93.

F. Le Gland, Stability and Approximation of Nonlinear Filters: an Information Theoretic Approach, Proceedings of the 38th Conference on Decision & Control, Phoenix, Arizona (1999)

L. Gerencsér, On a Class of Mixing Processes, Stochastics 26. 165-191.

L. Gerencsér, Rate of convergence of recursive estimators, SIAM J. Control and Optimization, Vol 30, No. 5 (1992) 1200-1227.

L. Gerencser, On the martingale approximation of the estimation error of ARMA parameters, Szstems & Control Letters 15 (1990) 417-423. North-Holland

F. Le Gland, L. Mevel, Asymptotic Properties of the MLE in Hidden Markov Models, Proceedings of the 4th European Control Conference, Bruxelles 1997

L. Mevel, L. Finesso, Convergence Rates of the Maximum Likelihood Estimator of Hidden Markov Models,2000

R. Bhattacharya, E. C, Waymire, An Approach to the Existence of Unique Invariant Probabilities for Markov Processes, 1999

B. G. Leroux, Maximum-likelihood estimation for hidden Markov models, Stochastic Processes and their Applications 40 (1992) 127-143. North-Holland

R. J. Elliott, J. van der Hoek, An application of hidden Markov models to asset allocation problems, Finance Stochast. 1. (1997) 229-238.

L. E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. 37 (1966) 1559-1563.

Y. Kifer, Ergodic Theory of Random Transformation, Progress in Probability and Statistics Vol 10, 1986

S. Karlin, H. M. Taylor, Stochastic Processes, Academic Press, New York

V. S. Borkar, On white noise representations in stochastic realization theory, SIAM J. Control and Optim. 31 (1993) 1093-1102.