New methods for the statistical analysis of Hidden Markov Models

László Gerencsér[1], Gábor Molnár-Sáska[1], György Michaletzky[2], Gábor Tusnády[3]

Abstract— **The estimation of Hidden Markov Models (HMM-s) has attracted a lot of attention recently, see results of [29], [30]. The purpose of this paper is to lay the foundation of a new approach for the statistical analysis of Hidden Markov Models (HMM-s), in particular for the analysis of the maximum-likelihood estimate, using a random mapping representation of HMM-s, see [6]. Our analysis is applicable to HMM-s with a general state-space and read-out space, assuming that the state process satisfies Doebin's condition. The key technical results are Theorem V.1 and VI.1 giving conditions for the functions of the input-output process of a non-linear stochastic systems to be L-mixing, see [18]. This is then applied to HMM-s extended by the filter process in Section VII. Three application are presented: first we state a strong approximation theorem for finite state HMM-s, complementing the results of [29] and [30]. Then the validity of Rissanen's tail condition is formulated, and finally the performance of adaptive encoders for HMM-s is stated in Section VIII.**

Keywords: Hidden Markov Models, random mappings, Doeblin-condition, L-mixing processes, maximum-likelihood estimation, strong approximation.

I. Introduction

HIDDEN Markov Models have become a basic tool for modeling stochastic systems with a wide range of applications in such diverse areas as nano-tecnology [26], quantized Gaussian linear regression [13], [14], telecommunication, [36], speech recognition [25], switching systems, [12], [16], financial mathematics [9], astronomy [3].

A good introduction to HMM-s, and stochastic systems in general is given in [37]. For a survey of recent results on HMM-s see [11]. An extension of HMM-s allowing dynamic memory is given in [35].

The estimation of the dynamics of a Hidden Markov Model is a basic problem in applications. The first basic result is due to Baum and Petrie for finite state Markov chains with finite-range read-outs [2]. Their analysis relies on the Shannon-Breiman-McMillan theorem, and exploits the finiteness of both the state-space $\mathcal{X}$ and the read-out space $\mathcal{Y}$. Strong consistency of the maximum-likelihood estimator for finite-state and binary read-out HMM-s has been established by Arapasthotis and Marcus in [1]. Strong consistency of the maximum-likelihood estimator for continuous read-out space has been first proven by Leroux in [30] using the subadditive ergodic theorem. An extensive study of HMM-s with finite state-space and continuous read-out-space has been carried out by LeGland and Mevel in [29] and [28] using the theory of geometric ergodicity for Markov chains. These results have been extended to compact state-space and continuous read-out-space by Douc and Matias in [7]. Strong consistency for the maximum-likelihood estimate for continous-time HMM-s with finite state-space and Gaussian read-out-space has been established by Moore and Elliott using martingale-theory in [10]. Adaptive control of HMM-s has been considered in Duncan et al. [8].

A key element in the statistical analysis of HMM-s is a strong law of large numbers for the log-likelihood function. All the listed tools are quite powerful and applicable under very weak conditions to derive strong laws of large number. The most fertile approach seems to be that of LeGland and Mevel, based on the use of geometric ergodicity, and leading to results such as CLT or convergence of recursive estimators.

Now it is known form the statistical theory of linear stochastic systems that these classical statistical results are not always sufficiently informative to answer natural questions like the performance of adaptive predictors. This has been pointed out by Gerencsér and Rissanen in [23]. In fact this very problem, the performance analysis of adaptive predictors and controllers has lead prompted research in deriving strong approximation results for estimators of linear stochastic systems, leading to a basic results in [19].

A main technical tool for deriving these results is the concept of L-mixing processes, developed in [18], a generalization of what is known as exponentially stable processes, introduced by Caines and Rissanen in [34] and Ljung [31]. This is a concept which, in its motivation, strongly exploits the linear algebraic structure of the underlying stochastic system.

A key observation of the present paper is that using a random mapping representation of HMM-s, which goes back to Borkar [6], see also [27], the concept of L-mixing naturally extends for HMM-s. Thus e.g. if the state-process satisfies the Doeblin-condition, then any fixed bounded measurable function of a Hidden Markov process will result in an L-mixing process, see Proposition IV.1.

Arapasthotis and Marcus have shown in [1], that extending a HMM process with its filter process we get a new HMM which plays a major role in the statistical analysis of the original process. The exponential forgetting of the filter-process has also been established for finite-state and binary-read-out HMM-s. An extensive analysis of the extended process has been carried out by LeGland and Mevel in [29] and [28] in the framework of geometric ergodicity. In the present paper the mixing properties of the extended process will be studied.

[1]MTA SZTAKI, Computer and Automation Institute, Hungarian Academy of Sciences, 13-17 Kende u., Budapest 1111, Hungary (email: gerencser@sztaki.hu, molnarsg@sztaki.hu)

[2]Eötvös Loránd University, 1/c Pázmány Péter, Budapest 1117, Hungary (email:michgy@ludens.elte.hu)

[3]Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, 13-15. Reáltanoda u., Budapest 1053, Hungary (email: tusnady@math-inst.hu).

It is well-known that the filter process is generated by a non-linear recursion, known as the Baum-equation, with the observation process as the input process. Uniform exponential stability (see Definition V.1) of this non-linear dynamic system has been shown in [29]. The key technical results given as Theorem V.1 and VI.1 are formulated for general non-linear stochastic systems that exhibit uniform exponential stability, driven by a Markov-process, giving conditions under which a fixed static function of the input-output process will be L-mixing. The application of Theorems V.1 and VI.1 to HMM-s with finite state-space and general read-out space, under Doeblin-condition for the state-process, is relatively easy and will be given in Section VII. Finally in Section VIII we compare our conditions with those of [29] and [28] and present three applications of the new results in the statistical analysis of HMM-s. First we state a strong approximation theorem for finite state HMM-s with finite read-outs, considerably strengthening the classic result of [2]. This easily generalizes to continuous read-outs, thus complementing the results of [29] and [30]. The next application is the verification of Rissanen's tail condition (see [32]) for finite state HMM-s with finite read-outs. Finally the performance of adaptive encoders for HMM-s is stated in Section VIII in analogy with the results of [20].

sssssss

Now L-mixing processes play a prominent role in modern theory of linear stochastic systems, and thus the latter result is directly applicable to derive a simple proof of the result of Baum and Petrie, see [2]. But it also provides the basic technical conditions, under which a very detailed characterization of the estimator process can be given in analogy with [19]. In particular we prove that for finite state-finite read-out HMM-s, parametrized by θ, the ML estimate of the true parameter θ^*, denoted by $\hat{\theta}_N$ satisfies, under simple technical conditions,

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

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

where $r_n = O_M(N^{-1})$ and R^* is the Fisher-information matrix.

A key point here 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 $\hat{\theta}_N - \theta^*$.

The finer characterization of the estimator process is not of purely academic interest: it plays a key role in adaptive prediction and model selection, see e.g. [20].

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 II.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).

If $\mathcal{X}$ and $\mathcal{Y}$ are finite, say $|\mathcal{X}| = N$, $|\mathcal{Y}| = M$, 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

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

where $i = 1, \dots, N$, and $*$ indicates that we take the true value of the corresponding unknown quantity.

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

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

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

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

where π is the normalizing operator: for $x \geq 0$, $x \neq 0$ set $\pi(x)^i = x^i / \sum_j x^j$, see [2]. 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

$$p_{n+1} = \pi(Q^T B(Y_n) p_n), \tag{2}$$

with initial condition $p_0 = q$, where Q is a stochastic matrix, p_n is a probability vector on $\mathcal{X}$, and $B(y) = \mathrm{diag}(b^i(y))$ is a collection of conditional probabilities.

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

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

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

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

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

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 property of the Baum equation is its exponential stability with respect to the initial condition. This has been established in [29] for continuous read-outs. Here we state the result for HMM-s with positive transition probability matrix:

Proposition II.1: Assume that $Q > 0$, i.e. the elements of the matrix Q are positive and $b^x(y) > 0$ for all x, y. Let q, q' be any two initializations. Then

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(1-\delta)^n \|q - q'\|_{TV}, \tag{3}$$

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

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (3) holds with a random C.

This basic property of the prediction filter will be used to introduce the concept of exponential stability and will be used to derive general mixing properties of the extended process (X_n, Y_n, p_n). This is formulated in the main result of the paper, Theorem VI.1.

Remark II.1: The existence of an invariant initialization and the ergodicity of the filter process has been proved under quite general conditions in Bhatt et al. [4].

Thus, a wrong initial condition for the prediction filter is rapidly forgotten, so that we could use any initial condition with the same effect. On the other hand, we expect that two different transition probability matrices and different read-out probabilities will produce significantly different observation sequences, so that we could estimate the unknown transition probability matrix and the unknown read-out probabilities by accumulating observations.

III. Representation of Markov processes

Let the state space of a Markov chain be a Polish space $\mathcal{X}$, and let $\mathcal{M} : \mathcal{X} \longrightarrow \mathcal{X}$ be the space of Borel-measurable mappings. Assume that $\mathcal{M}$ is equipped with a σ-algebra of its subsets such that the sets $\{f : \ fx \in G\}$ for any fixed $x \in \mathcal{X}$ and $G \in \mathcal{B}(\mathcal{X})$, where $\mathcal{B}(\mathcal{X})$ is the σ-algebra of Borel sets in $\mathcal{X}$, are measurable. Let $\mathbf{Q}$ be a probability measure on this σ-algebra. Let (T_n) be a sequence of i.i.d. mappings according to $\mathbf{Q}$. Let X_0 be independent of (T_n). Then the process (X_n) defined by $X_0 \in \mathcal{X}$, $X_{n+1} = T_{n+1}X_n$ is Markov. A converse result is given in the following proposition:

Proposition III.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 measure $\mathbf{Q}$ on $\mathcal{M}$, satisfying

$$P(x, G) = \mathbf{Q}\{T : Tx \in G\}.$$

For the proof see [27]. 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 III.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 III.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.

Proof: We outline the proof of [5]. First let us assume that there exists a representation (T_n). In this case $P(x, A) = \mathbf{Q}(T_1 x \in A) \geq \mathbf{Q}(T_1 x \in A | T_1 \in \Gamma_c)\mathbf{Q}(T_1 \in \Gamma_c) \geq \nu(A)\delta$, where ν is the probability measure.

On the other hand assume that the Doeblin-condition is valid. In this case we choose a random element ξ in $\mathcal{X}$ with distribution ν and then define $Tx = \xi$ for all x with probability δ and $Tx = \overline{T}x$ with probability $1 - \delta$, where $\overline{T}$ is received from a representation of a Markov chain with kernel function

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

■

Proposition III.2: Assume that the Doeblin-condition holds with $m = 1$ for a Markov chain (X_n). Then there exists an invariant distribution π, and

$$|P^n(x, A) - \pi(A)| \leq (1 - \delta)^n \quad \text{for} \quad \forall A \in \mathcal{B}(\mathcal{X}). \qquad (4)$$

Proof: For the sake of easy reference see [5] Let (T_n) be the representation of the process. In this case due to Lemma III.1 the limit $\lim_n T_0 \circ \cdots \circ T_{-n}\eta$ exists with probability 1, because $\mathbf{Q}(T_k \in \Gamma_c) \geq \delta > 0$, and so with probability 1 there exists k such that $T_k \in \Gamma_c$, and after using a constant mapping the process $T_0 \dots T_{-n}\eta$ does not depend on n. Further the limit is independent from $\eta \in \mathcal{X}$.

Let $\lim_n T_0 \circ \cdots \circ T_{-n}\eta = X_0^*$. In this case

$$X_0^* = \lim_n T_0 \circ \cdots \circ T_{-n}\eta = T_0 \circ T_{-1} \circ \cdots \circ T_{-k}\eta,$$

where k is such that $T_{-k} \in \Gamma_c$. Therefore

$$T_1 X_0^* = T_1 \circ T_0 \circ \dots T_{-k}\eta = \lim_n T_1 \circ T_0 \circ \cdots \circ T_{-n}\eta$$

We obtained that the distribution of X_0^* is invariant. So

$$|P^n(x, A) - \pi(A)| = |P(X_n \in A) - \pi(Y_n \in A)| =$$

$$= |E(\chi_A(X_n) - \chi_A(Y_n))| \leq P(X_n \neq Y_n),$$

where $X_n = T_n \dots T_1 X_0$ and $Y_n = T_n \dots T_1 X_0^*$.

Otherwise $P(X_n \neq Y_n) \leq \mathbf{Q}(T_k \notin \Gamma_c, k \leq n) \leq (1 - \delta)^n$, so the statement is proved. ■

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.

Lemma III.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.

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

Let $P(x, G)$ be the read-out transition kernel of the original Markov chain (X_n), where $x \in \mathcal{X}$ and $G \in \mathcal{B}(\mathcal{Y})$. By Proposition III.1 there is a probability measure $\mathbf{Q}'$ on the space of Borel mappings $\mathcal{X} \longrightarrow \mathcal{Y}$ such that if U is a random mapping selected according to $\mathbf{Q}'$ then $P(x, G) = \mathbf{Q}'\{U : Ux \in G\}$.

With the notation $Y_n = U_n X_n$ we get

$$Y_{n+1} = U_{n+1}T_{n+1}X_n.$$

It is easy to see that the random mapping $\binom{T}{UT}$ is a representation for $\binom{X}{Y}$. Obviously if $T_n \in \Gamma_c(\mathcal{X} \to \mathcal{X})$, then $U_n T_n \in \Gamma_c(\mathcal{X} \to \mathcal{Y})$, and thus

$$\mathbf{Q} \times \mathbf{Q}'\{\begin{pmatrix} T \\ UT \end{pmatrix} \in \Gamma_c\{\mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}\}\} \geq \delta,$$

and taking into account Lemma III.1, the lemma follows. ■

Remark III.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 III.2 and Lemma III.2 also valid if the Doeblin-condition holds for $m \geq 1$.

IV. Markov chains and L-mixing processes

Now we are going to introduce a class of processes called L-mixing processes which have been used extensively in the statistical analysis of linear stochastic systems, see [18]. First of all we need the definition of M-boundedness.

Definition IV.1: A stochastic process (u_n) $(n \geq 0)$ taking its values in an Euclidean space is M-bounded if for all $q \geq 1$

$$M_q(u) = \sup_{n \geq 0} E^{1/q} \|u_n\|^q < \infty. \tag{5}$$

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 all n.

Definition IV.2: A stochastic process (X_n) taking its values in a finite-dimensional Euclidean space is L-mixing with respect to $((\mathcal{F}_n), (\mathcal{F}_n^+))$, if it is M-bounded and with

$$\gamma_q(\tau) = \sup_{n \geq \tau} E^{1/q} \|X_n - E(X_n | \mathcal{F}_{n-\tau}^+)\|^q \tag{6}$$

we have

$$\Gamma_q = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty. \tag{7}$$

The definition of L-mixing extends to parameter dependent processes in a natural way. Let $\theta \in D \subset \mathbb{R}^p$, where D is a compact domain.

Definition IV.3: A stochastic process $(u_n(\theta))$ $(n \geq 0)$ taking its values in an Euclidean space is uniformly M-bounded if for all $q \geq 1$

$$\sup_{\theta \in D} M_q(u(\theta)) = \sup_{n \geq 0, \theta \in D} E^{1/q} \|u_n(\theta)\|^q < \infty. \tag{8}$$

Definition IV.4: A stochastic process $(X_n(\theta))$ taking its values in a finite-dimensional Euclidean space is L-mixing uniformly in θ with respect to $((\mathcal{F}_n), (\mathcal{F}_n^+))$, if it is uniformly M-bounded and for every $q \geq 1$ we have with

$$\gamma_q(\tau) = \sup_{n \geq \tau, \theta \in D} E^{1/q} \|X_n(\theta) - E(X_n(\theta) | \mathcal{F}_{n-\tau}^+)\|^q \tag{9}$$

$$\Gamma_q = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty. \tag{10}$$

The following lemma is useful in checking if a process is L-mixing, see e.g. [18].

Lemma IV.1: Let X be a random variable with $E\|X\|^q < \infty$ for all $q \geq 1$, and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let η be a $\mathcal{G}$ measurable random variable. Then we have

$$E^{1/q} \|X - E(X|\mathcal{G})\|^q \leq 2E^{1/q} \|X - \eta\|^q. \tag{11}$$

The first new result of this paper is given in the following proposition:

Proposition IV.1: Let (X_n) be a Markov chain with state space $\mathcal{X}$, where $\mathcal{X}$ is a Polish space, and assume that the Doeblin condition is valid for $m = 1$. Furthermore let $g : \mathcal{X} \longrightarrow \mathbb{R}$ be a bounded, measurable function. Then the process

$$U_n = g(X_n)$$

is L-mixing.

Proof: Let

$$\mathcal{F}_n = \sigma\{X_0, T_k : k \leq n\},$$

$$\mathcal{F}_n^+ = \sigma\{T_k : k \geq n+1\}.$$

Let $n \geq m$ and $n - m = \tau$. To approximate the process $g(X_n)$, first we approximate X_n by $X_{n,m}^+$, where

$$X_{n,m}^+ = T_n \dots T_{m+1} X^*, \tag{12}$$

and X^* is a constant. Obviously $X_{n,m}^+$ is $\mathcal{F}_m^+$ measurable. Furthermore

$$P(X_n \neq X_{n,m}^+) \leq \mathbf{Q}(T_k \text{ is not constant for } m+1 \leq k \leq n) \leq$$

$$(1-\delta)^{n-m}.$$

So

$$E^{1/q} \|g(X_n) - g(X_{n,m}^+)\|^q \leq 2KP^{1/q}(X_n \neq X_{n,m}^+) \leq$$

$$2K(1-\delta)^{\frac{n-m}{q}},$$

where K is an upper bound for $|g|$. Due to Lemma IV.1 we have

$$\gamma_q(\tau, U) \leq 4K(1-\delta)^{\frac{\tau}{q}},$$

and thus

$$\Gamma_q(U) \leq 4K \frac{1}{1-(1-\delta)^{\frac{1}{q}}},$$

and the statement is proved. ■

V. Exponentially stable random mappings I.

Now we formulate a general concept of exponential stability motivated by Proposition II.1. Let $\mathcal{X}$ be an arbitrary abstract set, and let $\mathcal{Z}$ be a closed subset of a Banach space (e.g. $\mathcal{Z} \subset L_1(\mathbb{R})$ can be the set of density functions). 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

$$z_{n+1} = f(x_n, z_n), \quad z_0 = \xi. \tag{13}$$

Let the solution be denoted by $z_n(\xi)$.

Definition V.1: The mapping f is uniformly exponentially stable if for every sequence $(x_n)_{n \geq 0}$, $x_n \in \mathcal{X}$

$$\|z_n(\xi) - z_n(\xi')\| \leq C(1-\varrho)^n \|\xi - \xi'\|, \tag{14}$$

where $C > 0, 1 > \varrho > 0$ are independent of the sequence (x_n). Under reasonable technical conditions this condition is satisfied for the Baum-equation and its derivatives, see [29]. Let $z(n, m, \xi)$ denote the solution of (13) initialized at $z_m = \xi$ with $m \leq n$. Let us consider an arbitrary discrete sequence defined by recursion of the form

$$\overline{z}_{n+1} = \overline{f}_n(\overline{z}_n) \tag{15}$$

with the same starting point $\overline{z}_0 = \xi$. Extending a simple analytical lemma given in [17] from continuous to discrete time we get

Lemma V.1: For the sequence (z_n) and $(\overline{z}_n)$ we have

$$z_n - \overline{z}_n =$$

$$\sum_{m=0}^{n-1} (z(n, m+1, f(x_m, \overline{z}_m)) - z(n, m+1, \overline{f}_m(\overline{z}_m))).$$

Proof: Due to the definition of z_n and $\overline{z}_n$ we have

$$z_n = z(n, 1, f(x_0, \overline{z}_0)) \quad \text{and} \quad \overline{z}_n = z(n, n, \overline{f}_{n-1}(\overline{z}_{n-1}))$$

Using

$$z(n, m+1, \overline{f}_m(\overline{z}_m)) = z(n, m+2, f(x_{m+1}, \overline{z}_{m+1})),$$

for $m = 0, \ldots, n-2$, we obtain the statement of the lemma. ■

A trivial corollary is the following key lemma:

Lemma V.2: For the solution of (13) we have

$$z_n = \xi + \sum_{m=0}^{n-1} (z(n, m+1, f(x_m, \xi)) - z(n, m+1, \xi)).$$

Proof: Let $\overline{f}$ be the constant mapping, so that $\overline{z}_n \equiv \xi$. Due to Lemma V.1 we have

$$z_n = \xi + \sum_{m=0}^{n-1} (z(n, m+1, f(x_m, \xi)) - z(n, m+1, \xi))$$

■

Define the process (Z_n) by

$$Z_{n+1} = f(X_n, Z_n), \quad Z_0 = \xi, \tag{16}$$

where (X_n) is a Markov chain with unique invariant distribution π . To prove M-boundedness of (Z_n) we impose following conditions:

Condition V.1: Let the distribution of X_0 be π_0. Assume

$$\frac{d\pi_0}{d\pi} \leq C_1.$$

Condition V.2: Assume for all $\xi \in \mathcal{Z}$ and for any $q \geq 1$

$$E_\pi \|Z_1(\xi)\|^q \leq K_1(\xi) < \infty,$$

or equivalently

$$\int_{\mathcal{X}} \|f(x, \xi)\|^q d\pi(x) \leq K_1(\xi) < \infty, \tag{17}$$

where π is the unique stationary distribution of (X_n) and $K_1(\cdot)$ is a measurable function.

Lemma V.3: Assume Condition V.1. Then we have

$$\frac{d\pi_n}{d\pi} \leq C_1 \quad \text{for all} \quad n. \tag{18}$$

Proof: For an arbitrary set $A \subset \mathcal{X}$

$$\pi_n(A) = \int_{\mathcal{X}} \chi_A d\pi_n = \int_{\mathcal{X}} P^n(x, A) d\pi_0 \leq$$

$$\leq \int_{\mathcal{X}} P^n(x, A) C_1 d\pi = C_1 \pi(A),$$

since π is the stationary distribution, so the lemma is proved. ■

Lemma V.4: Assume Condition V.1 and V.2. Then we have

$$E\|f(X_n, \xi)\|^q \leq K_1(\xi) C_1. \tag{19}$$

Proof: We have

$$E\|f(X_n, \xi)\|^q = \int_{\mathcal{X}} \|f(x, \xi)\|^q d\pi_n \leq$$

$$\int_{\mathcal{X}} \|f(x, \xi)\|^q C_1 d\pi \leq K_1(\xi) C_1,$$

due to Lemma V.3 and Condition V.2. ■

Lemma V.5: Let the mapping $f(x, z)$ be uniformly exponentially stable, and let Condition V.1 and V.2 hold. Then the process (Z_n) defined by (16) with any fixed constant $Z_0 = \xi$ is M-bounded.

Proof: Using Lemma V.2 and the exponential stability of f we have

$$\|Z_n\| \leq \|\xi\| + \sum_{m=0}^{n-1} C(1-\varrho)^{n-m-1} \|f(X_m, \xi) - \xi\|. \tag{20}$$

Since $q \geq 1$ and $f(X_m, \xi)$ is M-bounded, we have

$$E^{\frac{1}{q}} \|Z_n\|^q \leq$$

$$\|\xi\| + \sum_{m=0}^{n-1} C(1-\varrho)^{n-m-1} (E^{\frac{1}{q}} \|f(X_m, \xi)\|^q + \|\xi\|) \leq$$

$$\|\xi\| + C((K_1(\xi) C_1)^{\frac{1}{q}} + \|\xi\|) \frac{1}{\varrho},$$

so the lemma is proved. ■

Consider now processes of the form $V_n = g(X_n, Z_n)$, where g is a measurable function. We need the following technical condition:

Condition V.3: $g(x, z)$ is a measurable function on $\mathcal{X} \times \mathcal{Z}$ such that it is Lipschitz-continuous in z for every x with an x-independent Lipschitz constant L.

Theorem V.1: Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition with $m = 1$, and (Z_n) is defined by (16) with a uniformly exponentially stable mapping f and an arbitrary constant initial condition ξ. Assume that X_0 is independent of $\{T_n\}$, $n \geq 1$, and Conditions V.1 and V.2

hold. Furthermore let $g(x,z)$ be a bounded function satisfying Condition V.3 Then

$$V_n = g(X_n, Z_n)$$

is an L-mixing process.

Remark V.1: Theorem V.1 is valid also if the Doeblin-condition for (X_n) with $m > 1$ is assumed.

Proof: The process $V_n = g(X_n, Z_n)$ is obviously M-bounded. Now let $n \geq m$, $\tau = n - m$, $\mathcal{F}_n, \mathcal{F}_n^+$, and $X_{n,m}^+$ be the same as in the proof of Proposition IV.1, except that the distribution of X^* be stationary and

$$\mathcal{F}_n^+ = \sigma\{X^*, T_i : \ i \geq n+1\}.$$

Let an approximation of (Z_n) be defined recursively by

$$Z_{k+1,m}^+ = f(X_{k,m}^+, Z_{k,m}^+), \tag{21}$$

where $Z_{m,m}^+ = z^*$ is a constant.

Obviously, $Z_{n,m}^+$ is $\mathcal{F}_m^+$-measurable. Let $m' = n - [\frac{\tau}{2}]$ and let B denote the event that no coupling occurs in the interval $(m, m']$:

$$B = \{\omega : \text{ for } \quad m < k \leq m' \quad T_k \notin \Gamma_c\}.$$

Due to the Doeblin-condition

$$P(B) \leq (1-\delta)^{m'-m} = (1-\delta)^{[\frac{\tau}{2}]}.$$

Now consider the event

$$B^C = \{\omega : \exists k, \quad m < k \leq m' \quad T_k \in \Gamma_c\}.$$

On B^C we have $X_{k,m}^+ = X_k$ for all $k \geq m'$. Consider the following process:

$$Z_{k+1,m}^+ = f(X_k, Z_{k,m}^+) \quad \text{for} \quad m' < k \leq n, \tag{22}$$

with starting point at time $m' \quad Z_{m',m}^+$.

The process (Z_k) considered for $m' \leq k \leq n$ satisfies $Z_{k+1} = f(X_k, Z_k) \quad$ with starting point at time m' $Z_{m'}$.

On the set B^C by the exponential stability of f we have

$$\|Z_{n,m}^+ - Z_n\| \leq C(1-\varrho)^{[\frac{\tau}{2}]} \|Z_{m',m}^+ - Z_{m'}\|. \tag{23}$$

Hence for $q \geq 1$

$$E^{\frac{1}{q}} \|g(X_n, Z_n) - g(X_{n,m}^+, Z_{n,m}^+)\|^q \leq$$

$$E^{\frac{1}{q}} (\chi_B \|g(X_n, Z_n) - g(X_{n,m}^+, Z_{n,m}^+)\|^q)$$

$$+ E^{\frac{1}{q}} (\chi_{B^C} \|g(X_n, Z_n) - g(X_n, Z_{n,m}^+)\|^q) \tag{24}$$

As $g(x,z)$ is bounded, the first term on the right hand side can be bounded from above trivially

$$E^{\frac{1}{q}} (\chi_B \|g(X_n, Z_n) - g(X_{n,m}^+, Z_{n,m}^+)\|^q) \leq$$

$$E^{\frac{1}{q}} (\chi_B (2K)^q) = 2K P^{\frac{1}{q}}(B), \tag{25}$$

where $\|g(x,z)\| \leq K$.

Consider the second term of the expression (24).

$$\begin{aligned} E^{\frac{1}{q}} (\chi_{B^C} \|g(X_n, Z_n) - g(X_n, Z_{n,m}^+)\|^q) &\leq \\ E^{\frac{1}{q}} \|g(X_n, Z_n) - g(X_n, Z_{n,m}^+)\|^q &\leq \\ E^{\frac{1}{q}} (L \|Z_n - Z_{n,m}^+\|)^q &\leq \\ E^{\frac{1}{q}} (LC(1-\varrho)^{[\frac{\tau}{2}]} \|Z_{m',m}^+ - Z_{m'}\|)^q &= \end{aligned}$$

$$LC(1-\varrho)^{[\frac{\tau}{2}]} E^{\frac{1}{q}} \|Z_{m',m}^+ - Z_{m'}\|^q \tag{26}$$

The second inequality is due to the Lipschitz-continuity of g, and the third inequality follows from the exponential stability of f. Using the Minkowski inequality, Condition V.2 and Lemma V.5 (the distribution of X^* is stationary) we have that $Z_{m'}$ and $Z_{m',m}^+$ are M-bounded

$$E^{\frac{1}{q}} \|Z_{m',m}^+ - Z_{m'}\|^q \leq E^{\frac{1}{q}} \|Z_{m',m}^+\|^q + E^{\frac{1}{q}} \|Z_{m'}\|^q \leq S, \tag{27}$$

and so

$$E^{\frac{1}{q}} \|g(X_n^+, Z_{m,n}^+) - g(X_n, Z_n)\|^q \leq$$

$$2K(1-\delta)^{\frac{[\tau/2]}{q}} + K'(1-\varrho)^{[\tau/2]}, \tag{28}$$

where $K' = LCS$.

Now we are going to apply Lemma IV.1 and obtain

$$\gamma_q(\tau) \leq 2(2K(1-\delta)^{\frac{[\tau/2]}{q}} + K'(1-\varrho)^{[\tau/2]}). \tag{29}$$

Thus

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

$$\sum_{\tau=0}^{\infty} (4K(1-\delta)^{\frac{[\tau/2]}{q}} + 2K'(1-\varrho)^{[\tau/2]}) < \infty, \tag{30}$$

hence the claim of the theorem follows. ■

VI. Exponentially stable random mappings II.

In this section we consider an extension of Theorem V.1 for *unbounded* g. For this we will need to prove the existence of a stationary distribution for the process (X_n, Z_n). For not necessarily Markovian, but strictly stationary process (X_n) and (Z_n) defined by (16), Has'minskii's criteria gives a necessary and sufficient condition for the existence of stationary distribution on (X_n, Z_n) as follows:

Proposition VI.1: Let $\mathcal{X}, \mathcal{Y}$ be Polish spaces, and let (X_n) be a stationary process. Consider the recursion (16), where f is a continuous function. The process (X_n, Z_n) has a stationary distribution if and only if there exists an initial value ξ, such that

$$\frac{1}{N} \sum_{n=1}^{N} P(|Z_n| > c) \longrightarrow 0$$

uniformly in N, when $c \to \infty$.

A continuous time version of this theorem has been stated and proved in Thereom 2.1 of Chapter II of [24]. The proof of the discrete time proposition given above follows the same path. In our case it seems to be difficult to check the condition of Has'minskii, hence we follow a different path.

In our case the process (X_n, Z_n) is Markov. A representation of this Markov process is

$$(x, z) \longrightarrow (Tx, f(x, z)). \tag{31}$$

Lemma VI.1: Assume that the Doeblin-condition holds with $m \geq 1$ for the Markov process (X_n), f is uniformly exponentially stable mapping and Condition V.2 holds. Then the process (X_n, Z_n) has a stationary distribution.

Proof: Define X_{-n} as the limit

$$X_{-n} = \lim_{k\to\infty} T_{-n} \circ \cdots \circ T_{-n-k}\eta, \tag{32}$$

with any fixed η. It has been shown in Proposition III.2 that the limit is well-defined. It is easy to see that the process (X_{-n}) is stationary. Denote the mapping $f(x_n, \cdot) : \mathcal{Z} \longrightarrow \mathcal{Z}$ by f_{x_n} and let be

$$Z_0^* = \lim_n f_{X_{-1}} \circ \cdots \circ f_{X_{-n}}\xi. \tag{33}$$

We prove that the limit exists. Take a realization of (X_{-n}) denoted by (x_{-n}). Consider the difference

$$\|f_{x_{-1}} \circ \cdots \circ f_{x_{-n}}\xi - f_{x_{-1}} \circ \cdots \circ f_{x_{-m}}\xi\|, \tag{34}$$

where $n < m$. Using notations like in Lemma V.1 with $\varphi = z(-n-1, -m-1, \xi)$ we have

$$\|f_{x_{-1}} \circ \cdots \circ f_{x_{-n}}\xi - f_{x_{-1}} \circ \cdots \circ f_{x_{-m}}\xi\| =$$

$$\|f_{x_{-1}} \circ \cdots \circ f_{x_{-n}}\xi - f_{x_{-1}} \circ \cdots \circ f_{x_{-n}}\varphi\| \leq$$

$$C(1-\varrho)^n\|\xi - \varphi\|, \tag{35}$$

where the last inequality is due to the exponential stability of f. Thus

$$E^{\frac{1}{q}}\|f_{X_{-1}} \circ \cdots \circ f_{X_{-n}}\xi - f_{X_{-1}} \circ \cdots \circ f_{X_{-m}}\xi\|^q \leq$$

$$C(1-\varrho)^n\big(\|\xi\| + E^{\frac{1}{q}}\|Z(-n-1, -m-1, \xi)\|^q\big), \tag{36}$$

and by Lemma V.5 the sequence $f_{x_{-1}} \circ \cdots \circ f_{x_{-n}}\xi$ is Cauchy in L_q-norm, hence it converges. Thus Z_0^* is well-defined when convergence is interpreted in L_q-norm for any $q \geq 1$. Consider now the pair

$$X_0 = \lim_n T_0 \circ T_{-1} \circ \cdots \circ T_{-n}\eta, \tag{37}$$

$$Z_0^* = \lim_n f_{X_{-1}} \circ \cdots \circ f_{X_{-n}}\xi. \tag{38}$$

We prove that the distribution of (X_0, Z_0^*) is invariant, i.e. it is the same as the distribution of $(T_1X_0, f_{X_0}Z_0^*)$. Let $\overline{X}_1 = T_1X_0$ and $\overline{Z}_1 = f_{X_0}Z_0^*$ As

$$\overline{X}_1 = T_1 \lim_n T_0 \circ \cdots \circ T_{-n}\eta = T_1 \circ T_0 \circ T_{-1} \circ \cdots \circ T_{-k}\eta,$$

where k is such that $T_{-k} \in \Gamma_c$. Therefore

$$\overline{X}_1 = \lim_n T_1 \circ T_0 \circ \cdots \circ T_{-n}\eta$$

as in Proposition III.2, and

$$\overline{Z}_1 = f_{X_0}Z_0^* = f_{X_0} \circ \lim_n f_{X_1} \circ \cdots \circ f_{X_{-n}}\xi =$$

$$\lim_n f_{X_0} \circ \cdots \circ f_{X_{-n}}\xi, \tag{39}$$

since f_{X_0} is continuous in z. Thus the distribution of (X_0, Z_0^*) is the same as the distribution of $(\overline{X}_1, \overline{Z}_1)$, so the statement is proved. ■

Let the distribution of the process (X_0^*, Z_0^*) be denoted by μ, and for an arbitrary initialization let the distribution of (X_n, Z_n) be μ_n. To generalize the results of Theorem V.1 to unbounded g-s we need conditions, generalizing Condition V.2.

Condition VI.1: Let the initial value of (Z_n), i.e. Z_0 be a random variable with distribution ρ, independent of (X_n), where the distribution of X_n is stationary, and $E^{\frac{1}{q}}\|Z_0\|^q < \infty$ for $q \geq 1$ and for any such Z_0 let us have $E\|Z_1\|^q < \infty$ or equivalently

$$\int_{\mathcal{X}\times\mathcal{Z}} \|f(x, \xi)\|^q d\pi(x)d\rho(\xi) < \infty. \tag{40}$$

Condition VI.2: For the stationary distribution of (Z_n) we have

$$E^{\frac{1}{q}}\|Z_1\|^q < \infty \quad \text{for all} \quad q \geq 1. \tag{41}$$

A simple variant of Lemma V.5 is the following:

Lemma VI.2: Let the mapping $f(x, z)$ be uniformly exponentially stable, and let Conditions V.1, V.2 and VI.1 hold. Further assume that $E^{\frac{1}{q}}\|Z_0\|^q < \infty$ for all $q \geq 1$. Then the process (Z_n) is M-bounded.

Proof: Following the proof of Lemma V.5, we have

$$E^{\frac{1}{q}}\|Z_n\|^q \leq E^{\frac{1}{q}}\|Z_0\|^q +$$

$$\sum_{m=0}^{n-1} C(1-\varrho)^{n-m-1}(E^{\frac{1}{q}}\|f(X_m, Z_0)\|^q + E^{\frac{1}{q}}\|Z_0\|^q) \leq$$

$$E^{\frac{1}{q}}\|Z_0\|^q + C(E^{\frac{1}{q}}\|f(X_m, Z_0)\|^q + E^{\frac{1}{q}}\|Z_0\|^q)\frac{1}{\varrho}. \tag{42}$$

Due to the condition of the lemma Z_0 is M-bounded, and Lemma V.4 implies the M-boundedness of $f(X_m, Z_0)$ with $Z_0 = \xi$. ■

Due to Condition VI.2 the statement of Lemma VI.2 holds with stationary initialization.

Consider now the process $V_n = g(X_m, Z_n)$, where g is a measurable function. We need the following conditions for the function g.

Condition VI.3: Assume that for all $q \geq 1$

$$\int_{\mathcal{X}} \sup_{z\in\mathcal{Z}} \|g(x, z)\|^q d\pi(x) \leq M_q < \infty. \tag{43}$$

Lemma VI.3: Conditions V.1 and VI.3 imply that the process $g(X_n, Z_n)$ is M-bounded, i.e. for all $q \geq 1$

$$E\|g(X_n, Z_n)\|^q < \infty. \tag{44}$$

Proof:

$$E\|g(X_n, Z_n)\|^q = \int_{\mathcal{X}\times\mathcal{Z}} \|g(x, z)\|^q d\mu_n(x, z) \leq$$

$$\int_{\mathcal{X}\times\mathcal{Z}} \sup_{z\in\mathcal{Z}} \|g(x, z)\|^q d\mu_n(x, z) = \int_{\mathcal{X}} \sup_{z\in\mathcal{Z}} \|g(x, z)\|^q d\pi_n(x) \leq$$

$$\int_{\mathcal{X}} \sup_{z\in\mathcal{Z}} \|g(x,z)\|^q C_1 d\pi(x) \le M_q C_1. \tag{45}$$

■

Remark: If we replace Condition VI.3 with the following conditions then Lemma VI.3 still holds true: the Radon-Nikodym derivative of μ_0 w.r.t. μ is bounded, say

$$\frac{d\mu_0}{d\mu} \le K. \tag{46}$$

and

$$\int_{\mathcal{X}\times\mathcal{Z}} \|g(x,z)\|^q d\mu(x,z) \le M'_q. \tag{47}$$

(46) implies Condition V.1 and with a proof similar to Lemma V.3 we have

$$\frac{d\mu_n}{d\mu} \le K \quad \text{for all} \quad n,$$

thus indeed

$$E\|g(X_n,Z_n)\|^q \le KM'_q. \tag{48}$$

Condition VI.3 is motivated by Legland and Mevel [29] and easier to verify in practice as it will be seen in the next section.

We are going to generalize Theorem V.1 to unbounded g.

Theorem VI.1: Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition with $m = 1$, and let (Z_n) be defined by (16) with a uniformly exponentially stable mapping f. Let Z_0 be a random variable with

$$E^{\frac{1}{q}}\|Z_0\|^q < \infty \tag{49}$$

and let X_0 be independent of (T_n). Let Condition V.1, V.2, VI.1 and VI.2 hold for the process (X_n, Z_n), and assume that Condition V.3, VI.3 is satisfied for the function $g(x,z)$. Then

$$V_n = g(X_n, Z_n)$$

is an L-mixing process.

Remark VI.1: Theorem VI.1 holds if the Doeblin-condition holds with $m > 1$.

In Theorem V.1 the initialization of the process (Z_n), i.e. Z_0 was constant and the approximation of the process constructed with $Z^+_{m,m} = z^*$. However, in Theorem VI.1 g is unbounded, thus the proof of V.1 does not work, see (25). Indeed we initialize the process from a random point and make the approximation using stationary distribution. This is the reason of Conditions VI.1 and VI.2.

Proof: The proof is analogous to the proof of Theorem V.1. Let the distribution of $X^+_{m,m}$ and $Z^+_{m,m}$ be stationary, which exists due to Lemma VI.1. Consider the expression (24). The estimation of the second part is the same, but Lemma VI.2 is applied to prove the M-boundedness of $Z^+_{m',m}$ and $Z_{m'}$ in (27). Consider the first term. By the Hölder inequality we get

$$E^{\frac{1}{q}}(\chi_B\|g(X_n,Z_n) - g(X^+_{n,m}, Z^+_{n,m})\|^q) \le$$

$$(E^{\frac{1}{r}}(\chi_B)^r E^{\frac{1}{s}}\|g(X_n,Z_n) - g(X^+_{n,m}, Z^+_{n,m})\|^{qs})^{\frac{1}{q}}, \tag{50}$$

where $r, s > 0$ and $\frac{1}{r} + \frac{1}{s} = 1$. Due to Minkowski inequality we have

$$E^{\frac{1}{qs}}(\|g(X_n,Z_n) - g(X^+_n, Z^+_{n,m})\|)^{qs} \le$$

$$E^{\frac{1}{qs}}\|g(X_n,Z_n)\|^{qs} + E^{\frac{1}{qs}}\|g(X^+_n, Z^+_{n,m})\|^{qs} \tag{51}$$

and by Lemma VI.3 the right hand side is majorized by

$$P^{\frac{1}{qr}}(B)(2M_qC_1)^{\frac{1}{sq}} \le 2K_1 P^{\frac{1}{q}}(B) \tag{52}$$

and we can continue the proof of Theorem V.1 using K_1 instead of K. ■

VII. Application to Hidden Markov Models

This section demonstrates the relevance of the previous results for estimation of Hidden Markov Models. Consider a Hidden Markov Process (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is continuous, i.e. let $\mathcal{Y}$ be a general measurable space with a σ-field $\mathcal{B}(\mathcal{Y})$ and a σ-finite measure λ. Assume that the transition probability matrix and the conditional read-out densities are positive, i.e. $Q^* > 0$ and $b^{*i} > 0$ for all i, y. Then the process (X_n, Y_n) satisfies the Doeblin-condition.

Let the invariant distribution of (X_n) be ν and the invariant distribution of (X_n, Y_n) be π following the notations used in Theorem VI.1. Then

$$\pi^i(dy) = \nu_i b^{*i}(y)\lambda(dy), \tag{53}$$

where π^i denotes the components of π.

The logarithm of the likelihood function is

$$\sum_{k=1}^{n-1} \log p(y_k|y_{k-1},\dots y_0,\theta) + \log p(y_0,\theta), \tag{54}$$

where θ is the parameter of the model parameterizing the transition matrix Q and the conditionally read-out densities $b^i(y)$. Usually the entries of Q are part of θ.

The k-th term in (54) for $k \ge 1$ can be written as

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

Now write

$$g(y,p) = \log \sum_i b^i(y) p^i, \tag{55}$$

then we have

$$\log p(y_N,\dots,y_0,\theta) = \sum_{k=1}^{N} g(y_k,p_k) + \log p(y_0,\theta). \tag{56}$$

Let the running value of the transition probability matrix Q and the running value of the conditional read-out densities be also positive, i.e. $Q > 0$, $b^i(y) > 0$, respectively.

With the notation $p^i_n = P(X_n = i|Y_{n-1},\dots,Y_0)$ we have

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

We use capital letters for random variables and lower cases for their realizations, i.e. X is a random variable and x is a realization of X. The only exception is p, where the meaning depends on the context.

Theorem VII.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 continuous, a measurable subset of $\mathbb{R}^d$. Let $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) > 0$ for all i, y. Let the initialization of the process (X_n, Y_n) be random, where the Radon-Nikodym derivate of the initial distribution π_0 w.r.t the stationary distribution π is bounded, i.e.

$$\frac{d\pi_0}{d\pi} \leq K. \tag{57}$$

Assume that for all $i, j \in \mathcal{X}$

$$\int |\log b^j(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{58}$$

Then the process $g(Y_n, p_n)$ is L-mixing.

Proof: Identify (X_n, Y_n) with (X_n) and (p_n) with (Z_n) in Theorem VI.1. The exponential stability of f follows from Proposition II.1. As p_n is a probability vector Condition V.2, VI.1, VI.2 and the momentum condition of the initialization (49) are trivially satisfied.

We prove that Condition VI.3 is satisfied. Let $[x]_- = \max\{-x, 0\}$ and $[x]_+ = \max\{x, 0\}$. On one hand

$$\sum_j b^j(y)p^j \geq \min_i b^i(y),$$

leads to

$$[\log \sum_j b^j(y)p^j]_- \leq [\log \min_i b^i(y)]_-,$$

or

$$[g(y,p)]_- \leq \max_i [\log b^i(y)]_- \leq \max_i |\log b^i(y)|. \tag{59}$$

On the other hand the inequality

$$\sum_j b^j(y)p^j \leq \max_i b^i(y),$$

leads to

$$[\log \sum_j b^j(y)p^j]_+ \leq [\log \max_i b^i(y)]_+,$$

or

$$[g(y,p)]_+ \leq \max_i [\log b^i(y)]_+ \leq \max_i |\log b^i(y)|. \tag{60}$$

Since the right hand sides in (59) and (60) are independent of p we get

$$\sup_p |g(y,p)| \leq \max_i |\log b^i(y)|. \tag{61}$$

Combining (58) and (61) we get that for all $i \in \mathcal{X}$

$$\int \left(\sup_p |g(y,p)|^q\right) b^{*i}(y)\lambda(dy) < \infty. \tag{62}$$

Since

$$\int \sup_p |g(y,p)|^q d\pi = \sum_{i=1}^N \nu_i \int \left(\sup_p |g(y,p)|^q\right) b^{*i}(y)\lambda(dy), \tag{63}$$

the finiteness of the left hand side follows.

Now, only Condition V.3 remained to be checked, i.e. that $g(y,p) = \log \sum_i b^i(y)p^i$ is Lipschitz-continuous in p. For an arbitrary fix $y \in \mathcal{Y}$ we have

$$\|\frac{\partial g(y,p)}{\partial p}\| = \|\frac{1}{\sum_j b^j(y)p^j}(b^1(y), \dots b^N(y))^T\| \leq \tag{64}$$

$$\frac{\sqrt{N}\max_i b^i(y)}{\sum_j b^j(y)p^j} \leq \sqrt{N}\max_i \frac{1}{p^i} = \sqrt{N}(\min_i p^i)^{-1}. \tag{65}$$

It is easy to see that p^i has a positive lower bound. Let

$$\varepsilon = \min_{i,j} q_{ij} > 0. \tag{66}$$

Due to the Baum-equation (2) we have

$$p_{n+1} = \pi(Q^T B(y_n)p_n) = \frac{Q^T B(y_n)p_n}{\mathbf{1}^T Q^T B(y_n)p_n},$$

where $\mathbf{1}^T = (1, \dots, 1)^T$. As Q is a stochastic matrix, $\mathbf{1}^T Q^T B(y_n)p_n = \mathbf{1}^T B(y_n)p_n$, and due to (66)

$$Q^T B(y_n)p_n \geq \varepsilon \mathbf{1}\mathbf{1}^T B(y_n)p_n.$$

Thus

$$p_{n+1} \geq \frac{\varepsilon \mathbf{1}\mathbf{1}^T B(y_n)p_n}{\mathbf{1}^T B(y_n)p_n} = \varepsilon \mathbf{1} \tag{67}$$

and we get

$$\|\frac{\partial g(y,p)}{\partial p}\| \leq \frac{\sqrt{N}}{\varepsilon}. \tag{68}$$

Hence the function $g(y,p)$ is Lipschitz continuous.

Thus Theorem VI.1 implies that $g(Y_n, p_n)$ is an L-mixing process. ∎

Remark VII.1: Since the positivity of Q implies that the stationary distribution of (X_n) is strictly positive in every state and the densities of the read-outs are strictly positive Condition (57) is not a strong condition. For example for the random initialization we can take a uniform distribution on $\mathcal{X}$ and an arbitrary set of λ a.e. positive density functions $b_0^i(y)$.

To analyze the asymptotic properties of the right hand side of (56) Theorem VII.1 seems to be relevant. Under the conditions of Theorem VII.1 $g(y,p)$ is an L-mixing process and the law of large numbers is valid for such processes, see [18]. This implies the existence of the limit of (56).

Consider now a finite state-finite read-out HMM. This case follows from Theorem VII.1, but the integrability condition (58) is simplified due to the discrete measure.

Theorem VII.2: Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Assume that the process (X_n, Y_n) satisfies the Doeblin condition. Let the running value of the transition probability matrix Q be positive and $b^i(y) \geq \delta > 0$ for all i, y. Then with a random initialization on $\mathcal{X} \times \mathcal{Y}$ we have that $g(Y_n, p_n)$ is an L-mixing process.

VIII. Estimation of Hidden Markov Models

!!! Add text.

In the sequel we compare the results with those of Legland and Mevel, [29]. For easier reference we restate the results collecting the relevant conditions.

Proposition VIII.1: Consider a Hidden Markov Process (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is continuous. Let the transition probability matrix of the unobserved Markov chain be primitive and the conditional read-out densities be positive, i.e. let there exist a positive integer r such that $Q^{*r} > 0$, and let $b^{*i}(y) > 0$, respectively. For the running parameter assume also that $Q^r > 0$ and $b^i(y) > 0$ for all i. Furthermore assume that for all $i \in \mathcal{X}$

$$\int \frac{\max_{j\in\mathcal{X}} b^j(y)}{\min_{j\in\mathcal{X}} b^j(y)} b^{*i}(y)\lambda(dy) < \infty, \tag{69}$$

and for all $i, j \in \mathcal{X}$

$$\int |\log b^j(y)| b^{*i}(y)\lambda(dy) < \infty. \tag{70}$$

Then the process $g(Y_n, p_n)$ is geometrically ergodic

Geometric ergodicity also implies the existence of limit in (56).

Remark VIII.1: Condition (69) is a Lipschitz condition in the mean in the following sense. Due to (64) for an arbitrary fix $y \in \mathcal{Y}$ the function $\|\partial g(y,p)/\partial p\|$ is bounded uniformly in p

$$\|\frac{\partial g(y,p)}{\partial p}\| \le \sqrt{N}\max_i \frac{b^i(y)}{\sum_j b^j(y)p^j} \le \sqrt{N}\frac{\max_i b^i(y)}{\min_j b^j(y)}$$

since $\sum_j p^j = 1$, thus $L(y) = \sqrt{N}\max_i b^i(y)/\min_j b^j(y)$ is an y-dependent Lipschitz constant. Condition (69) states that the Lipschitz constant $L(y)$ is bounded in average.

Example: Consider an example with finite state space $\mathcal{X}$ and read-out space $\mathbb{R}$. Assume that the process (X_n) satisfies the Doeblin-condition with $m = 1$ and let the running value of the transition probability matrix be positive, i.e. $Q > 0$. Let the read-outs be continuous with normal density functions, i.e.

$$b^i(y) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{(y-m_i)^2}{2\sigma_i}),$$

where (m_i, σ_i)s are the parameters. Assume that $\sigma_1 \le \cdots \le \sigma_N$. Let denote the true parameter by (m_i^*, σ_i^*). Since $\log b^i(y)$ is quadratic in y, (58) is satisfied as the momentums of the normal distribution exist. Hence Theorem VII.1 is applicable, and the limit of the log-likelihood function (56) exists.

On the other hand Condition (69) of Proposition VIII.1 may not be satisfied if $\sigma_1 < \sigma_N$. Indeed, for large y-s the integrand of (69) is

$$C\exp\left(-\frac{(y-m_N)^2}{2\sigma_N^2} + \frac{(y-m_1)^2}{2\sigma_1^2} - \frac{(y-m_i^*)^2}{2(\sigma_i^*)^2}\right),$$

where C is a constant, which is integrable only if

$$-\frac{1}{\sigma_N^2} + \frac{1}{\sigma_1^2} - \frac{1}{(\sigma_i^*)^2} < 0$$

for all i, i.e. if

$$(\sigma_i^*)^2 > \frac{(\sigma_1\sigma_N)^2}{(\sigma_N)^2 - (\sigma_1)^2}, \tag{71}$$

Consider a finite state-finite read-out HMM, parameterized by θ, where $|\mathcal{X}| = N$ and $|\mathcal{Y}| = M$ and θ containing the elements of the transition probability matrix and the read-out probabilities. Thus θ is an $N^2 + NM - 2N$ dimensional vector with coordinates between 0 and 1. Furthermore let the ML estimate of the true parameter θ^* be denoted by $\hat{\theta}_N$. Due to [29] the gradient process $\partial p_n(\theta)/\partial\theta$ is also exponentially stable, thus Theorem VI.1 yields that the process $\partial g(Y_n, p_n(\theta))/\partial\theta$ is an L-mixing process. Similarly it can be shown that $\partial^2 g(Y_n, p_n(\theta)/\partial\theta^2$ is also an L-mixing process.

The arguments of [19] yield the following result.

Theorem VIII.1: Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) \ge \delta > 0$ for all i, y. Let $\hat{\theta}_N$ be the ML estimate of θ^*. Then $\hat{\theta}_N - \theta^*$ can be written as

$$-(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots,Y_0,\theta^*) + r_n, \tag{72}$$

where $r_n = O_M(N^{-1})$, and R^* is the Fisher-information matrix.

Remark VIII.2: For the continuous version of Theorem VIII.1 more conditions are needed, which is out of the scope of this paper: identifiability conditions needed for the read-out densities, see [30], and further smoothness and integrability conditions needed for the gradient process to be L-mixing.
A key point here 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 $\hat{\theta}_N - \theta^*$.

In the sequel we are going to present some consequences of this result. The tail-condition in Rissanen-theorem, see in [32], for the error term of the estimation θ is satisfied, see [22].

Theorem VIII.2: Under the condition of Theorem VII.1 we have

$$\sum_{N=1}^{\infty} P(N^{\frac{1}{2}}(\hat{\theta}_N - \theta^*) > c\log N)) < \infty, \tag{73}$$

where $c > 0$ is an arbitrary constant

The negative logarithm of the conditional probability

$$-\log p(y_n|y_{n-1},\dots,y_1,\theta)$$

can be interpreted as a code length, see [33]. An adaptive encoding procedure is obtained if we set $\theta = \hat{\theta}_{n-1}$. Following [20] we get the following result:

Theorem VIII.3: Let s_n be

$$-\log p(y_n|y_{n-1},\dots,y_1,\hat{\theta}_{n-1}) + \log p(y_n|y_{n-1},\dots,y_1,\theta^*).$$

Under the conditions of Theorem VIII.1 we have

$$E_{\theta^*}(s_n) = \frac{1}{2n}p(1+o(1)), \tag{74}$$

where $p = \dim\theta$. Furthermore

$$\lim_{N\to\infty}\frac{1}{\log N}\sum_{n=1}^{N} s_n = \frac{p}{2} \tag{75}$$

with probability 1.

This result can be used for model selection for HMM-s, see [15], [21].

REFERENCES

[1] A. Arapostathis and S.I. Marcus. Analysis of an Identification Algorithm Arising in the Adaptive Estimation of Markov Chains. *Math. Control Signals Systems*, 3:1–29., 1990.

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

[3] J.O. Berger. Some Recent Developments in Bayesian Analysis with Astronomical illustrations. In G. J. Babu and E.D. Feigelson, editors, *Statistical Challenges in modern Astronomy*, pages 15–39. Springer, 1997.

[4] A.G. Bhatt, A. Budhiraja, and R.L. Karandikar. Some properties of the nonlinear filter: Markovity and Ergodicity. In *In Proceedings of the 40th IEEE Conference on Decision and Control*, pages 1699–1704., 2001.

[5] R. Bhattacharya and E. C. Waymire. An approach to the existence of unique invariant probabilities for Markov processes. 1999.

[6] V. S. Borkar. On white noise representations in stochastic realization theory. *SIAM J. Control Optim.*, 31:1093–1102, 1993.

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

[8] T.E. Duncan, B. Pasik-Duncan, and L. Stettner. Some Results on Ergodic and Adaptive Control of Hidden Markov Models. In *Proceedings of the 41th IEEE Conference on Decision & Control, Las Vegas*, pages WeA07–1, 1369–1374, 2002.

[9] R. J. Elliott, W. P. Malcolm, and A. Tsoi. HMM Volatility Estimation. In *Proceedings of the 41th IEEE Conference on Decision & Control, Las Vegas*, pages TuA 12–6, 398–404, 2002.

[10] R.J. Elliott and J.B. Moore. Almost sure parameter estimation and convergence rates for Hidden Markov models. *Systems and Control Letters*, 32:203–207., 1997.

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

[12] X. Feng, K.A. Loparo, Y. Ji, and H.J. Chizeck. Stochastic stability properties of jump linear systems. *IEEE Transactions on Automatic Control*, 37:38–53., 1992.

[13] L. Finesso, L. Gerencsr, and I. Kmecs. Estimation of parameters from quantized noisy observations. In *Proceedings of the European Control Conference, ECC99, Karlsruhe*, pages AM–3, F589, 6p., 1999.

[14] L. Finesso, L. Gerencsr, and I. Kmecs. A randomized EM-algorithm for estimating quantized linear Gaussian regression. In *Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix*, pages 5100–5101., 1999.

[15] L. Finesso, C.C. Liu, and P. Narayan. The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory*, 42:1488–1497, 1996.

[16] C. Francq and M. Roussignol. Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics*, 32:151–173., 1998.

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

[18] L. Gerencsér. On a class of Mixing Processes. *Stochastics*, 26:165–191, 1989.

[19] L. Gerencsér. On the martingale approximation of the estimation error of ARMA parameters. *Systems & Control Letters*, 15:417–423, 1990.

[20] L. Gerencsér. On Rissanen's Predictive Stochastic Complexity for Stationary ARMA Processes. *Statistical Planning and Inference*, 41:303–325, 1994.

[21] L. Gerencsér and J. Baikovicius. A computable criterion for model selection for linear stochastic systems. In L. Keviczky and Cs. Bányász, editors, *Identification and System Parameter Estimation, Selected papers from the 9th IFAC-IFORS Symposium, Budapest*, volume 1, pages 389–394, Pergamon Press,Oxford, 1991.

[22] L. Gerencsér and G. Molnár-Sáska. Adaptive encoding and prediction of Hidden Markov processes. In *European Control Conference*.

[23] L. Gerencsér and J. Rissanen. A prediction bound for Gaussian ARMA processes. *Proc. of the 25th Conference on Decision and Control, Athens*, 3:1487–1490., 1986.

[24] R. Z. Has'minskii. *Stochastic stability of differential equations*. Sijthoff & Noordhoff, 1980.

[25] X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden Markov models for speech recognition*. Edinburgh University Press, 1990.

[26] I.W. Hunter, L.A. Jones, M. Sagar, S.R. Lafontaine, and P.J. Hunter. Opthalmic microsurgical robot and associated virtual environment. *Computers in Biology and Medicine*, 25:173–182., 1995.

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

[28] F. LeGland and L. Mevel. Basic Properties of the Projective Product with Application to Products of Column-Allowable Nonnegative Matrices. *Mathematics of Control, Signals and Systems*, 13:41–62, 2000.

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

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

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

[32] J. Rissanen. Stochastic complexity and predictive modelling. *Annals of Statistics*, 14(3):1080–1100, 1986.

[33] J. Rissanen. *Stochastic complexity in statistical inquiry*. World Scientific Publisher, 1989.

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

[35] J. Rissanen and S. Forchhammer. Partially Hidden Markov Models. *IEEE Trans. on Information Theory*, 42:1253–1256., 1996.

[36] L. Shue, S. Dey, B.D.O. Anderson, and F. De Bruyne. Remarks on Filtering Error due to Quantisation of a 2-state Hidden Markov Model. In *Proceedings of the 40th IEEE Conference on Decision & Control*, pages FrA05, 4123–4124., 1999.

[37] J. H. van Schuppen. Lecture Notes on Stochastic Systems. Technical report. Manuscript.