STATISTICAL ANALYSIS OF HIDDEN MARKOV MODELS

OUTLINE OF PHD THESIS

Molnár-Sáska Gábor

Supervisor:
László Gerencsér

2005.

Institute of Mathematics, Technical University of Budapest

and

Computer and Automation Research Institute
of the Hungarian Academy of Sciences

In the subsequent sections we give an overview of the results of the PhD thesis "Statistical Analysis of Hidden Markov Models". We start with a short introduction of the considered topics. Then we describe the investigated model and the results.

1 Introduction

A Hidden Markov Model (HMM) is a discrete-time finite-state homogenous Markov chain observed through a discrete-time memoryless invariant channel. The channel is characterized by a finite set of transition densities indexed by the states of the Markov chain. These densities may be members of any parametric family such as Gaussian, Poisson, etc. The initial distribution of the Markov chain, the transition matrix, and the densities of the channel may depend on some parameter that characterizes the HMM.

Hidden Markov Models have become a basic tool for modelling stochastic systems with a wide range of applications in such diverse areas as nano-tecnology [31], telecommunication [52], speech recognition [30], switching systems [16, 20], financial mathematics [13] and protein research [53]. A good introduction to HMM-s with recent results is given in [15].

The estimation of the dynamics of a Hidden Markov Model is a basic problem in applications. The first fundamental result is due to Baum and Petrie for finite state Markov chains with finite-range read-outs [5]. 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 Araposthatis and Marcus in [1]. An important technical tool, the exponential forgetting of the predictive filter has also been established. Strong consistency of the maximum-likelihood estimator for continuous read-out space has been first proven by Leroux in [41] 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 [40] and [39] 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 [9].

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 numbers. 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.

However, it is known from the statistical theory of linear stochastic systems that these classical results of statistics 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 [28], see also [26]. In fact, the performance analysis of adaptive predictors and controllers has lead prompted research in deriving strong approximation results for estimators of linear stochastic systems. For off-line estimators the strongest result on such a strong approximation is given in [24].

A main technical tool for deriving these results is the concept of L-mixing processes, developed in [23], a generalization of what is known as exponentially stable process, introduced by Caines and Rissanen in [48] and Ljung [42]. This is a concept which, in its motivation, strongly exploits the stability and the linear algebraic structure of the underlying stochastic system.

A simple, but important observation is that using a random mapping representation of HMM-s (which goes back to Borkar [8], see also [33]), 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 Theorem 3.1 below.

2 Preliminaries

2.1 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, equipped with their respective Borel-fields.

Definition 2.1 *The pair (X_n, Y_n) is a Hidden Markov process if (X_n) is a homogenous Markov process with state space $\mathcal{X}$ and the observation sequence (Y_n) is conditionally independent and identically distributed given the σ-field generated by the process (X_n).*

To illustrate the basic concepts let the state space of the Hidden Markov Model be finite now, i.e. $|\mathcal{X}| = N$. The results for general compact state space are discussed in Section 4.2.

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 dissertation 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.

The 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.

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

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

Writing $p_{n+1} = (p^1_{n+1}, \dots, p^N_{n+1})^T$, we know from [5] that the filter process satisfies the Baum-equation

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

with initial condition $p_0 = q$, where $Q \in \mathbb{R}^{N \times N}$ is a stochastic matrix, $B(y) = \mathrm{diag}(b^i(y))$ is a collection of conditional probabilities, and $q \in \mathbb{R}^N$ 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 $p_n(q)$.

From the statistical point of view it is crucial whether the Baum equation is exponentially stable, 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 [40] for continuous read-outs under appropriate conditions.

Proposition 2.1 *Assume that $Q > 0$ and $b^x(y) > 0$ for all x, y. Let q, q' be any two initializations. Then for some $0 < \delta < 1$,*

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

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

Let D be a non-empty, open subset of $\mathbb{R}^r$. Consider the following estimation problem: let $Q(\theta)$ and $b(\theta)$ be parameterized by $\theta \in D$, and let

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

Usually the entries of Q are included in θ.

A standard step in proving consistency of the maximum likelihood estimator is to show that

$$\lim_{N \to \infty} \frac{1}{N} \log p(y_0, \dots y_N, \theta) \tag{4}$$

exists almost surely (uniformly in θ), see [42].

The limit of (4) was investigated in various setup in the literature, see [4], [41], [20], [34], [9], [10].

2.2 L-mixing processes

In this section an overview of L-mixing processes is presented. The concept of L-mixing introduced by László Gerencsér [23] seemed to be a very powerful tool in the analysis of linear stochastic systems. Establishing a connection between HMM-s and linear stochastic systems this technique became the main technical tool analyzing Hidden Markov Models in this thesis. Here we give the definition of L-mixing:

Let a probability space $(\Omega, \mathcal{F}, P)$ be given. Consider an $\mathbb{R}^m$-valued stochastic process (X_n), $n \geq 0$ defined on $(\Omega, \mathcal{F}, P)$. From now on we do not make explicit reference to $(\Omega, \mathcal{F}, P)$ any more.

Definition 2.2 *We say that the stochastic process (X_n), $n \geq 0$ is M-bounded if for all $1 \leq q < \infty$*

$$M_q(x) = \sup_{n \geq 0} E^{\frac{1}{q}} |X_n|^q < \infty.$$

If (X_n) is M-bounded we shall also write $X_n = O_M(1)$. Similarly if c_n is a positive sequence we write $X_n = O_M(c_n)$ if $X_n / c_n = O_M(1)$.

Let $(\mathcal{F}_n)$, $n \geq 0$ be a family of monotone increasing σ-fields and $(\mathcal{F}_n^+)$, $n \geq 0$ be a monotone decreasing family of σ-fields. We assume that for all $n \geq 0$, $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent.

Definition 2.3 *A stochastic process (X_n), $n \geq 0$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$, if it is $\mathcal{F}_n$-adapted, M-bounded, and for $1 \leq q < \infty$ and $\tau \in \mathbb{Z}^+$*

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

is such that

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

3 Exponentially stable systems

3.1 Representation of Markov processes

In Section 3.1 we give an overview of the representation of Markov chains following Borkar, see [8], then give some useful statement on the relationship between this representation and the Doeblin condition, see [7]. Using these techniques we prove the following lemma:

Lemma 3.1 *Assume that the Doeblin-condition holds for the Markov chain (X_n). Then the Doeblin-condition holds for (X_n, Y_n) as well.*

3.2 Markov chains and L-mixing processes

Consider an input-output system as follows: Let the input process be a Markov chain which satisfies the Doeblin condition and the output process is generated through a bounded measurable function. Then the Doeblin condition is not satisfied for the output process. The following theorem states that the output process is L-mixing.

Theorem 3.1 *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. Furthermore let $g : \mathcal{X} \longrightarrow \mathbb{R}$ be a bounded, measurable function. Then the process*

$$U_n = g(X_n)$$

is L-mixing.

3.3 Exponentially stable random mappings I.

Now we formulate a general concept of exponential stability motivated by Proposition 2.1. Let $\mathcal{X}$ be an arbitrary abstract measurable space, 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

$$z_{n+1} = f(x_n, z_n), \quad z_0 = \xi. \tag{5}$$

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

Definition 3.1 *The mapping f is uniformly exponentially stable if for every sequence (x_n) $n \geq 0$, $x_n \in \mathcal{X}$*

$$\|z_n(\xi) - z_n(\xi')\| \leq C(1-\varrho)^n \|\xi - \xi'\|, \tag{6}$$

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 [40].

Define the process (Z_n) by

$$Z_{n+1} = f(X_n, Z_n), \quad Z_0 = \xi, \tag{7}$$

where (X_n) is a Markov chain satisfying the Doeblin condition. Let us denote its invariant distribution by π. To prove M-boundedness of (Z_n) we impose the following conditions:

Condition 3.1 *Let the distribution of X_0 be π_0. Assume*

$$\frac{d\pi_0}{d\pi} \leq C_1. \tag{8}$$

Condition 3.2 *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

$$\int_{\mathcal{X}} \|f(x,\xi)\|^q d\pi(x) \leq K_1(\xi) < \infty, \tag{9}$$

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

Lemma 3.2 *Let the mapping $f(x,z)$ be uniformly exponentially stable, and let Condition 3.1 and 3.2 hold. Then the process (Z_n) defined by (7) with any fixed constant $Z_0 = \xi$ is M-bounded.*

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 3.3 *$g(x,z)$ 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 3.2 *Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition, and (Z_n) is defined by (7) with a uniformly exponentially stable mapping f and an arbitrary constant initial condition ξ. Assume that Conditions 3.1 and 3.2 hold. Furthermore let $g(x,z)$ be a bounded function satisfying Condition 3.3. Then*

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

is an L-mixing process.

In some applications Condition 3.3 is too strong. Hence we should weaken this condition with the following one:

Condition 3.4 *$g(x,z)$ is a measurable function on $\mathcal{X}\times\mathcal{Z}$ such that it is Lipschitz-continuous in z for every x with an x-dependent Lipschitz constant $L(x)$ such that all the moments of $L(x)$ exists with respect to the stationary distribution of the Markov chain (X_n), i.e. for all $q \geq 1$*

$$\int_{\mathcal{X}} |L(x)|^q d\pi(x) < L_q^q < \infty.$$

Replacing Condition 3.3 with Condition 3.4 Theorem 3.2 is also valid.

3.4 Exponentially stable random mappings II.

In this section we consider an extension of Theorem 3.2 for *unbounded* function g. We need the following conditions.

Condition 3.5 *Assume that for all $q \geq 1$*

$$\int_{\mathcal{X}} \sup_{z\in\mathcal{Z}} \|g(x,z)\|^q d\pi(x) \leq M_q < \infty. \tag{10}$$

We generalize Theorem 3.2 to unbounded function g as follows.

Theorem 3.3 *Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition, and let (Z_n) be defined by (7) with a uniformly exponentially stable mapping f and an arbitrary constant initial condition $Z_0 = \xi$. Assume that Conditions 3.1 and 3.2 hold. Furthermore assume that Condition 3.3, 3.5 is satisfied for the function $g(x,z)$. Then*

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

is an L-mixing process.

Theorem 3.3 can also be weakened replacing Condition 3.3 with Condition 3.4.

3.5 Exponentially stable random mappings III.

For strong approximation results we will need the L-mixing property of the derivative process $\frac{\partial}{\partial\theta}\log p(y_n|y_{n-1},\dots y_0,\theta)$. Since the conditions of Theorem 3.3 are not satisfied for this derivative process we need an extension of Theorem 3.3. We change Condition 3.3 to the following technical condition:

Condition 3.6 *Let $g(x,z)$ be a measurable function on $\mathcal{X}\times\mathcal{Z}$ such that for every x with an x-dependent Lipschitz constant $L(x)$ we have*

$$|g(x,z_1)-g(x,z_2)|\le L(x)\|z_1-z_2\|(\|z_1\|+\|z_2\|).$$

Furthermore assume that

$$\left(\int_{\mathcal{X}}|L(x)|^q d\pi(x)\right)^{1/q}<L_q<\infty$$

for all $q\ge 1$, where $\pi(x)$ is the stationary distribution of the Markov chain (X_n).

Furthermore we weaken Condition 3.5.

Condition 3.7 *Assume that for all $q\ge 1$*

$$\int_{\mathcal{X}}\sup_{z\in\mathcal{Z}}\left(\frac{\|g(x,z)\|}{\|z\|+1}\right)^q d\pi(x)\le M_q<\infty. \tag{11}$$

Then we have

Theorem 3.4 *Consider the process (X_n,Z_n), where (X_n) satisfies the Doeblin-condition, and let (Z_n) be defined by (7) with a uniformly exponentially stable mapping f and an arbitrary constant initial condition $Z_0=\xi$. Assume that Conditions 3.1 and 3.2 hold. Furthermore assume that Conditions 3.6, 3.7 are satisfied for the function $g(x,z)$. Then*

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

is an L-mixing process.

3.6 On-line estimation

In this section we lay down the foundation of the analysis of the convergence of recursive estimation in Hidden Markov Models. For this purpose we investigate Markov processes generated by exponentially stable mappings. First we present the general scheme of Benveniste, Metivier and Priouret, see [6] introduced for investigating stochastic approximation algorithms, then verify the assumptions of [6] for our model class.

3.6.1 The BMP scheme

In this section we present the basics of the theory of recursive estimation developed by Benveniste, Metivier and Priouret, BMP henceforth (see Chapter 2, Part II. of [6]).

Let a family of transition probabilities $\{\Pi_\theta,\ \theta \in D \subset R^d\}$ on $\mathcal{U}$ be given, where $\mathcal{U}$ is a Polish space. Let us denote the metric by d. Note that in [6] $\mathcal{U}$ is R^n, but the results can be generalized for complete separable metric space. Let D be an open set. Assume that for any $\theta \in D$ there exists a unique invariant probability measure, say μ_θ. Let $(U_n(\theta))$ be a Markov-chain such that its initial state $U_0(\theta)$ has distribution μ_θ. Let $H(\theta, u)$ be a mapping from $R^d \times \mathcal{U}$ to R^d. Then the basic estimation problem of the BMP-theory is to solve the equation

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

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

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

$$\theta_{n+1} = \theta_n + \frac{1}{n} H(\theta_n, U_n), \tag{12}$$

where U_n is the time-varying process defined by

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

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

Theorem 13, p. 236 of [6] yields the following convergence result.

Theorem 3.5 *(Benveniste-Métivier-Priouret 1990, [6]) Assume that Conditions* **A1 - A6** *are satisfied, and ϵ is sufficiently small. Let $\theta \in \mathrm{int} D_0$, $U_m = u \in \mathcal{U}$, and consider the stopped process $\theta_n^\circ = \theta_{n \wedge \tau \wedge \sigma}$. Then for any $0 < \lambda < 1$ there exist constants B and s such that for all $m \geq 0$ we have $\lim \theta_n^\circ = \theta^*$ with probability at least*

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

For Conditions **A1 - A6** see [6] or Section 3.6.1 of the dissertation.

3.6.2 Application for exponentially stable nonlinear systems

In this subsection conditions (**A1**)-(**A3**) are verified for exponentially stable nonlinear systems. Let $\mathcal{X}$ be a Polish space and $\mathcal{Z}$ be a closed subset of a separable Banach space. Let us denote the metric on $\mathcal{X}$ by $d_{\mathcal{X}}$.

Consider an exponentially stable random mapping f, see Definition 3.1, and define the process (Z_n) by

$$Z_{n+1} = f(X_n, Z_n, \theta), \quad Z_0 = \xi, \tag{13}$$

where (X_n) is a Markov chain which satisfies the Doeblin condition. Let $U_n = (X_n, Z_n) \in \mathcal{X} \times \mathcal{Z} = \mathcal{U}$. Define the metric on $\mathcal{U}$ by

$$d(u, u') = \|z - z'\| + d_{\mathcal{X}}(x, x'), \tag{14}$$

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

$$V(u) = \|z\|. \tag{15}$$

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

Condition 3.8 *Let the distribution of X_1 be π_1. Assume*

$$\frac{d\pi_1}{d\pi} \leq C_1. \tag{16}$$

Condition 3.9 *Assume for all $\xi \in \mathcal{Z}$ and for $p \geq 1$*

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

or equivalently

$$\int_{\mathcal{X}} \|f(x, \xi)\|^p d\pi(x) \leq K_1(1 + \|\xi\|^p). \tag{17}$$

Note that Condition 3.8 is a modified version of Condition 3.1. As in assumptions (**A1**)-(**A3**) the initialization is always a fixed value and we need it for each initialization, Condition 3.1 is not realistic. Condition 3.9 is a special case of Condition 3.2.

Theorem 3.6 *Consider a process $U_n = (X_n, Z_n)$ defined by (7), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.8 and 3.9 are satisfied. Then assumption (**A1**) holds, i.e. there exists positive constant K such that for all $n \geq 0$, $u \in \mathcal{U}$ and $\theta \in Q$:*

$$E_{u,\theta}(|V(U_n)|^{p+1}) \leq K(1 + |V(u)|^{p+1}).$$

Since we have not used the metric property in Theorem 3.6 $\mathcal{X}$ can be any measurable abstract space. Furthermore, we have used the Doeblin property only for the existence of a stationary distribution of the Markov chain (X_n).

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

Condition 3.10 *Assume that f is Lipschitz continuous in x, i.e.*

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

Condition 3.11 *Assume that for the process (X_n) we have*

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

Theorem 3.7 *Consider a process $U_n = (X_n, Z_n)$ defined by (7), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.8, 3.9, 3.10 and 3.11 are satisfied. Then assumption (**A2**) holds, i.e. there exist positive constants K, p and $0 < \rho < 1$ such that for all $g \in Li(p)$, $\theta \in Q$, $n \geq 0$ and $u, u' \in \mathcal{U}$:*

$$|\Pi_\theta^n g(u) - \Pi_\theta^n g(u')| \leq K\rho^n \|\Delta g\|_{V_p} (1 + |V(u)|^p + |V(u')|^p) d(u, u')$$

For assumption (**A3**) we need the smoothness of f with respect to the parameter θ. Assume that $f : \mathcal{X} \times \mathcal{Z} \times \Theta \to \mathcal{Z}$ is a Borel-measurable function, differentiable in θ and for any fix θ the function $f(\cdot, \cdot, \theta)$ is exponentially stable.

Theorem 3.8 *Consider a process $U_n = (X_n, Z_n)$ defined by (7), where f is an exponentially stable mapping which is smooth is θ, and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.8 and 3.9 are satisfied. Then assumption (**A3**) holds, i.e. there exist positive constants K, p such that for all $g \in Li(p)$, $u \in \mathcal{U}$, $n \geq 0$ and $\theta, \theta' \in Q$:*

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

We conclude this section with the following theorem.

Theorem 3.9 *Consider a process $U_n = (X_n, Z_n)$ defined by (7), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.8, 3.9, 3.10 and 3.11 are satisfied. Then assumptions (**A1**)-(**A3**) hold.*

Thus we get that if assumption (**A5**) is satisfied for a function H, and we have a Lyapunov function satisfying (**A6**) then convergence result Theorem 3.5 holds for the algorithm (12).

We apply Theorem 3.9 for Hidden Markov Models in Chapter 5.

4 Application to Hidden Markov Models

This chapter demonstrates the relevance of the previous results for the 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 possibly 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. In practice $\mathcal{Y}$ is usually a measurable subset of $\mathbb{R}^d$. Although the results of this chapter are valid for a general read-out space, we will assume that $\mathcal{Y}$ is a measurable subset of $\mathbb{R}^d$ and λ is the Lebesgue-measure. Assume that the transition probability matrix and the conditional read-out densities are positive, i.e. $Q^* > 0$ and $b^{*i}(y) > 0$ 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 3.3. Then

$$\pi(\{i\}, dy) = \nu_i b^{*i}(y)\lambda(dy). \tag{18}$$

4.1 Estimation of Hidden Markov Models

A central question in estimation problems is proving the ergodic theorem

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g(y_k, p_k), \tag{19}$$

where

$$g(y,p) = \log \sum_i b^i(y) p^i. \tag{20}$$

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 $\mathbb{R}^d$. 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 derivative of the initial distribution π_0 w.r.t the stationary distribution π is bounded, i.e.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{21}$$

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

$$\int |\log b^j(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{22}$$

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

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 (21) 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 $b_0^i(y)$.

For the asymptotic properties of (19) we have

Theorem 4.2 *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 $\mathbb{R}^d$. 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 derivative of the initial distribution π_0 w.r.t the stationary distribution π is bounded, i.e.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{23}$$

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

$$\int |\log b^j(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{24}$$

Then the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g(Y_k, p_k)$$

exists almost surely.

Consider now a finite state-finite read-out HMM. This case follows from Theorem 4.1, but the integrability condition (22) is simplified due to the discrete measure.

Theorem 4.3 *Consider a Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Assume that $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) > 0$ 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.*

Finally, we compare our results with those of Legland and Mevel, see [40] or Proposition 4.1.6 of the dissertation. We give an example where Theorem 4.2 is applicable, but conditions of Proposition 4.1.6 are not satisfied.

4.2 Extension to general state space

We extend the results of Section 4.1 for a general compact state space. Let (X_n) be a Markov chain on a compact set $K \subset \mathcal{X}$, where $\mathcal{X}$ is a Polish space, and $\mathcal{B}(K)$ is the associated Borel σ-field. Let us fix a σ-finite dominating measure on $\mathcal{X}$. Let $Q^*(x, A)$ ($x \in K$, $A \in \mathcal{B}(K)$) be the Markov transition kernel of the chain, see [44]. The observations (Y_n) are conditionally independent and identically distributed given (X_n) with conditional densities $b^{*x_n}(y)$, see (1), where the read-out space $\mathcal{Y}$ is assumed to be a Polish space. Let the initial distribution of (X_n) be P_0^*.

Assume that the densities $b^x(y)$ are with respect to the same σ-finite measure λ and the transition kernel Q has a density q with respect to the σ-finite dominating measure μ on $\mathcal{X}$. Furthermore, it is assumed that the initial distribution of (X_n) has a density p_0 with respect to μ.

Consider the predictive density function, i.e. the density of the conditional distribution of X_n given $(Y_i)_{i=0}^{n-1}$. Using the Baum-equation, see (2), we have the following recursion for the density of the predictive filter:

$$p_{n+1}(x) = \frac{\int_u q(u,x) b^u(Y_n) p_n(u) d\mu(u)}{\int_u b^u(Y_n) p_n(u) d\mu(u)}.$$

In this section we use the following notations: for any measurable function f on the space $(K, \mathbf{B}(K), \mu)$ define

$$\operatorname{ess\,sup}(f) = \inf\{M \geq 0 : \mu(\{M < |f|\}) = 0\}$$

and if f is non-negative,

$$\operatorname{ess\,inf}(f) = \sup\{M \geq 0 : \mu(\{M > |f|\}) = 0\}.$$

For $y \in \mathcal{Y}$ define

$$\delta(y) = \frac{\operatorname{ess\,sup}_x b^x(y)}{\operatorname{ess\,inf}_x b^x(y)} \tag{25}$$

$$\epsilon = \frac{\operatorname{ess\,inf}_{x,x'} q(x,x')}{\operatorname{ess\,sup}_{x,x'} q(x,x')}. \tag{26}$$

The following statement, which is an adaptation of Proposition 2.1, shows the exponential memorylessness of the predictive density function, see [9].

Proposition 4.1 *(Douc-Matias 2001, [9]) Suppose that $0 < \epsilon$. Let p_0' and p_0'' be any two initial density functions of X_0 with respect to the measure μ. Then*

$$\|p_n(p_0') - p_n(p_0'')\|_{L_1} \leq C(1-\epsilon)^n \|p_0' - p_0''\|_{L_1}. \tag{27}$$

4.2.1 Estimation of HMMs: continuous state space

Assume that the Markov chain (X_n) has an invariant distribution ν. This implies that the density of the invariant distribution of the pair (X_n, Y_n) is

$$\pi(x,y) = b^x(y)\nu(x).$$

The logarithm of the likelihood function is

$$\sum_{k=1}^{n-1} \log\left(\int_K b^x(Y_k) p_k \mu(dx)\right),$$

and define the function g as

$$g(y,p) = \log\left(\int_K b^x(y) p(x) \mu(dx)\right). \tag{28}$$

The following theorem is a modified version of Theorem 4.1.

Theorem 4.4 *Consider a Hidden Markov Model* (X_n, Y_n)*, where the state space* $K \subset \mathcal{X}$ *is a compact subset of a Polish space* $\mathcal{X}$ *and the observation space* $\mathcal{Y}$ *is a measurable subset of* $\mathbb{R}^d$*. Assume that* $\epsilon > 0$ *in (26). Furthermore, assume that the Doeblin condition is satisfied for the Markov chain* (X_n)*. Let the initialization of the process* (X_n, Y_n) *be random such that the Radon-Nikodym derivative of the initial distribution* π_0 *w.r.t the stationary distribution* π *is bounded, i.e.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{29}$$

Assume that for all $q \geq 1$

$$\operatorname*{ess\,sup}_x \int |\log \operatorname*{ess\,sup}_{x'} b^{x'}(y)|^q b^{*x}(y)\lambda(dy) < \infty. \tag{30}$$

and

$$\operatorname*{ess\,sup}_x \int |\delta(y)|^q \, b^{*x}(y)\lambda(dy) < \infty \tag{31}$$

Then the process $g(Y_n, p_n)$ *is L-mixing and the limit*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g(Y_k, p_k)$$

exists almost surely.

5 Recursive Estimation of Hidden Markov Models

In this paragraph we consider Hidden Markov Models with finite state-space and finite read-out space. Consider the following estimation problem: let Q and b be parameterized by $\theta \in D$, where D is a compact subset of $\mathbb{R}^r$ and let

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

Consider the parameter-dependent Baum-equation

$$\mathbf{p}_{n+1}(\theta) = \frac{Q^T(\theta)B(y_n,\theta)\mathbf{p}_n(\theta)}{\mathbf{b}(y_n,\theta)^T\mathbf{p}_n(\theta)} = \Phi_1(y_n, \mathbf{p}_n, \theta), \tag{32}$$

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

$$W_{n+1} = Q^T \left(I - \frac{B(y_n)\mathbf{p}_n\mathbf{e}^T}{\mathbf{b}^T(y_n)\mathbf{p}_n} \right) \frac{B(y_n)W_n}{\mathbf{b}^T(y_n)\mathbf{p}_n} + F, \tag{33}$$

where

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

$W_n = \frac{\partial \mathbf{p}_n}{\partial \theta}$ and $\beta(y_n) = \frac{\partial B(y_n)}{\partial \theta}$.

In a compact form

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

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

Let the score function be

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

Using that

$$\log p(y_n | y_{n-1}, \ldots, y_0, \theta) = \log \mathbf{b}^T(y) \mathbf{p}_n,$$

we get

$$\varphi_n = \frac{\beta(y_n)\mathbf{p}_n + W_n \mathbf{b}(y_n)}{\mathbf{b}(y_n)^T \mathbf{p}_n}. \tag{34}$$

Let

$$H(\theta, u) = H(\theta, x, y, \mathbf{p}, W) = \frac{\beta(y, \theta)\mathbf{p} + W\mathbf{b}(y, \theta)}{\mathbf{b}(y, \theta)^T \mathbf{p}}, \tag{35}$$

and consider the following adaptive algorithm.

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \gamma_{n+1} H(\overline{\theta}_n, x_n, y_n, \overline{\mathbf{p}}_n, \overline{W}_n), \tag{36}$$

$$\overline{\mathbf{p}}_{n+1} = \Phi_1(y_n, \overline{\mathbf{p}}_n, \overline{\theta}_n), \tag{37}$$

$$\overline{W}_{n+1} = \Phi_2(y_n, \overline{\mathbf{p}}_n, \overline{W}_n, \overline{\theta}_n). \tag{38}$$

For the convergence of this algorithm we use the approach of Benveniste, Metivier and Priouret, see Section 3.6.1 and [6]. We verify the conditions of Theorem 3.9.

Consider a Hidden Markov Model with finite state space and finite read-out space.

Assume that $Q(\theta)$ and $b(\theta)$ are smooth functions of the parameter, i.e. the second derivatives exist and are continuous.

Theorem 5.1 *Consider a Hidden Markov Model with finite state space and finite read-out space. Assume that $Q^* > 0$, $b^{*x}(y) > 0$, and $Q(\theta) > 0$, $b^x(y, \theta) > 0$ for all x, y and $\theta \in D$, where D is a compact subset of $\mathbb{R}^d$. Then assumptions (**A1**)-(**A3**) and (**A5**) of Section 3.6.1 are satisfied.*

Note that if the state space and the read-out space are finite then assumption (**A4**) is trivially satisfied.

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

$$h(\theta) = \lim_{n \to \infty} E \frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \ldots Y_0, \theta) \tag{39}$$

This limit exists, see Theorem 6.2, and assume that the following identifiability condition is satisfied, see also Condition 6.1:

Condition 5.1 *The equation $h(\theta) = 0$ has exactly one solution in D, namely θ^*.*

Condition 5.1 implies assumption (**A6**) in a small domain. Thus we conclude with the following theorem as an application of Theorem 3.5.

Theorem 5.2 *Consider a Hidden Markov Model with finite state space and finite read-out space. Assume that $Q^* > 0$, $b^{*x}(y) > 0$, and $Q(\theta) > 0$, $b^x(y, \theta) > 0$ for all θ, x, y. Assume Condition 5.1. Then the algorithm defined by (36), (37), (38) converges to the true value θ^* with probability arbitrary close to 1.*

6 Strong Estimation of Hidden Markov Models

6.1 Parametrization of the Model

In this chapter the rate of convergence of the parameter is investigated. Let $G \subset \mathbb{R}^r$ be an open set, $D \subset G$ be a compact set, and $D^* \subset \text{int}D$ be another compact set, where $\text{int}D$ 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 chapter are valid for a general read-out space, we will always assume that $\mathcal{Y}$ is a measurable subset of $\mathbb{R}^d$ and λ is the Lebesgue-measure, similarly to Chapter 4. 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.

6.2 L-mixing property of the derivative process

For strong approximation theorems we will need that the derivative processes

$$\frac{\partial^k}{\partial \theta^k} \log p(y_n, y_{n-1}, \ldots, y_0, \theta),$$

where $k = 1, 2, 3$, are L-mixing.

For $y \in \mathcal{Y}$ define

$$\delta(y) = \frac{\max\limits_x b^x(y)}{\min\limits_x b^x(y)} \tag{40}$$

and

$$\delta'(y) = \frac{\max_x \|\partial b^x(y)/\partial\theta\|}{\min_x b^x(y)} \tag{41}$$

Theorem 6.1 *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 $\mathbb{R}^d$. Let $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) > 0$ for all i, y. Assume that $Q(\theta)$ and $b^i(y, \theta)$ are continuously differentiable functions in the parameter θ. 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.*

$$\frac{d\pi_0}{d\pi} \le K. \tag{42}$$

Assume that

$$\int |\delta(y)|^q b^{*i}(y)\lambda(dy) < \infty, \tag{43}$$

$$\int |\delta(y)'|^q b^{*i}(y)\lambda(dy) < \infty. \tag{44}$$

Then

$$\frac{\partial}{\partial\theta} \log p(y_n | p_{n-1}, \ldots p_0, \theta)$$

is L-mixing.

In applications we need that the limit of the expectation of the derivative process exists, see (39) or (49).

Theorem 6.2 *Under the conditions of Theorem 6.1 we have that the limit*

$$\lim_{n\to\infty} E\frac{\partial}{\partial\theta} \log p(y_n | y_{n-1}, \ldots y_0, \theta)$$

exists.

We prove similar theorems as Theorem 6.1 for the second and third derivatives.

6.3 Characterization theorem for the error

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 $\mathbb{R}^d$. Let $Q(\theta), Q^* > 0$ 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.

$$\frac{d\pi_0}{d\pi} \le K. \tag{45}$$

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

$$\int |\log b^j(y,\theta)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{46}$$

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 to 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

$$\frac{\partial}{\partial\theta} L_N(\theta,\theta^*) = L_{\theta N}(\theta,\theta^*) = 0 \tag{47}$$

Let us introduce the asymptotic cost function

$$W(\theta,\theta^*) = \lim_{n\to\infty} E_{\theta^*} \log p(Y_n|Y_{n-1},\ldots,Y_0,\theta). \tag{48}$$

Assume that the function $W(\theta,\theta^*)$ is smooth in the interior of D, i.e. the third derivative exists. Under the conditions of Theorem 6.2 we have

$$W_\theta(\theta,\theta^*) = \lim_{n\to\infty} E_{\theta^*} \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1},\ldots,Y_0,\theta), \tag{49}$$

and for the Fisher-information matrix we have

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

$$\lim_{n\to\infty} E_{\theta^*}\left((\frac{\partial}{\partial\theta}\log p(y_n|y_{n-1},\ldots,y_0,\theta^*))^T (\frac{\partial}{\partial\theta}\log p(y_n|y_{n-1},\ldots,y_0,\theta^*))\right).$$

Remark 6.1 *Note that $W_\theta(\theta^*,\theta^*) = 0$.*

Consider the following identifiability condition:

Condition 6.1 *The equation*

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

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

We prove a characterization theorem for the error term of the off-line ML estimation following the arguments of [24].

Theorem 6.3 *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 $\mathbb{R}^d$. Let $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) > 0$ for all i, y. Assume that conditions of Theorem 4.1, 6.1 are satisfied. Let $\hat{\theta}_N$ be the ML estimate of θ^*. Furthermore assume that the identifiability condition 6.1 is satisfied. Then*

$$\hat{\theta}_N - \theta^* = -(I^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\ldots,Y_0,\theta^*) + O_M(N^{-1}), \tag{50}$$

where I^* 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 $\hat{\theta}_N - \theta^*$.

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

Theorem 6.4 *Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let $Q(\theta), Q^* > 0$ 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 6.1 is satisfied. Then*

$$\hat{\theta}_N - \theta^* = -(I^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\ldots,Y_0,\theta^*) + O_M(N^{-1}), \tag{51}$$

where I^* is the Fisher-information matrix.

7 Estimation with forgetting

If the dynamics changes slowly in time, then we should adapt to the actual system. But then the estimation procedure must be modified: instead of cumulating past data we must gradually forget them. Forgetting past data is technically realized by using exponential forgetting in the off-line case.

To estimate the unknown parameter we use the modified maximum-likelihood method: let $\widehat{\theta}_N(\lambda)$ be the estimator of θ^* obtained by minimizing

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\log p(y_n|y_{n-1},\ldots,y_0;\theta), \tag{52}$$

with $0 < \lambda < 1$. Here λ is the so-called forgetting factor: small value of λ means slow forgetting.

Let

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

We shall refer to this as the cost function associated with the modified ML estimation of the parameter. The right hand side depends on θ^* through the sequence (Y_n).

It is easy to see that the cost function can be computed recursively as follows:

$$L_N^\lambda(\theta,\theta^*) = (1-\lambda)L_{N-1}^\lambda(\theta,\theta^*) + \lambda \log p(Y_N|Y_{N-1},\dots,Y_0,\theta),$$

i.e. the correction term corresponding to the latest observation enters the cost function always with the same fixed weight. This representation of the cost function justifies the terminology "fixed gain estimation".

The modified ML estimation $\widehat{\theta}_N(\lambda)$ of θ^* is defined as the solution of the equation

$$\frac{\partial}{\partial\theta}L_N^\lambda(\theta,\theta^*) = L_{\theta N}^\lambda(\theta,\theta^*) = 0 \tag{53}$$

Combining Theorem 4.1 and the results of Section 6.2 with the techniques of [25] we have a version of Theorem 6.3:

Theorem 7.1 *Under the conditions of Theorem 6.3 we have*

$$\hat{\theta}_N(\lambda) - \theta^* = -I(\theta^*)^{-1}\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*) + r_N,$$

where $0<\alpha<1$, $r_N = O_M(\lambda) + O_M(\alpha^N)$, *and* $I(\theta^*)$ *is the Fischer-information matrix.*

Theorem 7.1 implies that for the covariance matrix we have

$$E(\widehat{\theta}_{n-1} - \theta^*)(\widehat{\theta}_{n-1} - \theta^*)^T = \frac{\lambda}{2}I(\theta^*)^{-1} + O(\lambda^{3/2}) + o(1). \tag{54}$$

8 Change detection of HMM-s

We consider change-detection problems for Hidden Markov Models following [3]. For this we first note that the negative of the log-likelihood can be interpreted as a codelength, modulo a constant, which is obtained when encoding the data sequence $(y_N,\dots,y_1)$ with a prescribed accuracy, using the assumed joint density $p(y_N,\dots,y_0;\theta)$. This interpretation of the likelihood is a central idea of the theory of stochastic complexity. Thus we interpret

$$C_n(y_n;\theta) \stackrel{\Delta}{=} -\log p(y_n|y_{n-1},\dots,y_0;\theta),$$

as a codelength. A key result in the theory of the stochastic complexity can be extended for the present case (see [26]).

Theorem 8.1 *Under the conditions of Theorem 6.3 we have*

$$\mathrm{E}(C_n(Y_n,\widehat{\theta}_{n-1}(\lambda)) - C_n(Y_n,\theta^*) = \frac{1}{2}r\lambda + O(\lambda^{3/2-c''}) + o(1),$$

with an arbitrary small $c''>0$, *where* $r = \dim\theta$.

The faster the forgetting is i.e. the closer λ is to 1, the more we loose in encoding performance. An easy consequence of Theorem 8.1 is the following.

Proposition 8.1 *Consider two different forgetting factors* $0 < \lambda_1 < \lambda_2 < 1$. *Then we have*

$$\mathrm{E}(C_n(y_n, \widehat{\theta}_{n-1}(\lambda_1)) - C_n(y_n, \widehat{\theta}_{n-1}(\lambda_2))) \simeq \frac{1}{2} r(\lambda_1 - \lambda_2) < 0.$$

Proposition 8.1 has been useful in the design of a new model selection criterion. However the theoretical analysis of the new method is not powerful enough. Thus we need a sample path characterization of the prediction error process. Let the cumulative error be

$$S_N(\lambda) = \sum_{n=1}^{N} (C_n(y_n, \widehat{\theta}_{n-1}(\lambda)) - C_n(y_n, \theta^*))$$

Theorem 8.2 *Under the conditions of Theorem 6.3 we have*

$$\limsup_{N \to \infty} |\frac{1}{N} S_N(\lambda) - \frac{\lambda}{2} r| \leq C\lambda^{3/2}$$

We state a similar easy consequence as above.

Proposition 8.2 *Let* $0 < \lambda_1 < \lambda_2 < 1$ *be two different forgetting factors. Then we have*

$$\limsup_{N \to \infty} |\frac{1}{N} S_N(\lambda_1) - \frac{1}{N} S_N(\lambda_2) - \frac{\lambda_1 - \lambda_2}{2} r| \leq C\lambda_2^{3/2}$$

Assume now that a jump in the parameter occurs at τ: the true value of θ is θ_1 for $n \leq \tau$ and it is θ_2 for $n \geq \tau + 1$, i.e.

$$\theta^* := \begin{cases} \theta_1, & \text{if} \quad n \leq \tau \\ \theta_2, & \text{if} \quad n \geq \tau \end{cases}$$

Let $0 < \lambda_1 < \lambda_2 < 1$. Then from Proposition 8.2 we have for $N \leq \tau$

$$S_N(\lambda_1) - S_N(\lambda_2) \approx \frac{\lambda_1 - \lambda_2}{2} Nr.$$

On the other hand at the time of change the performance of the estimator with faster forgetting, i.e. with λ_2 expected to be better. Hence, consider the following algorithm for detecting the change.

The algorithm: Let $d(N) := S_N(\lambda_1) - S_N(\lambda_2)$ and set

$$d_N^* = \min_{n \leq N} d(n).$$

An alarm is generated if $d(N) - d_N^* > \epsilon$, where $\epsilon > 0$ is a prescribed threshold value. This type of algorithm is called *Hinkley detector* in the literature, see [12].

Publications

- Gerencsér, L., Molnár-Sáska, G,. A New Method for the Analysis of Hidden Markov Model Estimates, Proceedings of the 15th Triennial World Congress of the International Federation of Automatic Control, Barcelona, 2002., T-Fr-M03

- Gerencsér L., Molnár-Sáska G., Michaletzky Gy., Tusnády G., Vágó Zs., New methods for the statistical analysis of Hidden Markov Models, Proceedings of the 41th IEEE Conference on Decision & Control, Las Vegas, 2002., WeP09-6 2272-2277.

- Gerencsér L., Molnár-Sáska G., Adaptive encoding and prediction of Hidden Markov processes, In proceedings of the European Control Conference, ECC2003, Cambridge, 2003.,

- Gerencsér L., Molnár-Sáska G., Estimation error in adaptive prediction of Hidden Markov Processes, In proceeding of the 11th Mediterranean Conference on Control and Automation MED03, Rhodes, 2003.

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

- Gerencsér L., Molnár-Sáska G., Change detection of Hidden Markov Models, Proceedings of the 43th IEEE Conference on Decision & Control, 1754-1758, 2004.

- Gerencsér L., Molnár-Sáska G., Michaletzky Gy., Tusnády G., A new approach for the statistical analysis of Hidden Markov Models, IEEE Transactions on Automatic Control, submitted

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] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.*, 33 (6):697–725, 1997.

[3] J. Baikovicius and L. Gerencsér. Change point detection in a stochastic complexity framework. In *Proc. of the 29-th IEEE CDC*, volume 6, pages 3554–3555, 1990.

[4] A. R. Barron. The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *The Annals of Probability*, 13:1292–1303, 1985.

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

[6] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations.* Springer-Verlag, Berlin, 1990.

[7] R. Bhattacharya and E. C. Waymire. An approach to the existence of unique invariant probabilities for Markov processes. *Limit theorems in probability and statistics, János Bolyai Math. Soc.*, I (Balatonlelle 1999):181–200, 2002.

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

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

[10] R. Douc, É. Moulines, and T. Ryden. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics*, 32:2254–2304, 2004.

[11] 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.

[12] D.V.Hinkley. Inference about the change-point from Cumulative Sum Tests. *Biometrika*, 58 (3):509–523., 1971.

[13] 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.

[14] 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.

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

[16] 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.

[17] L. Finesso, L. Gerencsér, 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.

[18] L. Finesso, L. Gerencsér, 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.

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

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

[21] H. Furstenberg and H. Kesten. Products of random matrices. *Ann. Math. Statist.*, 31:457–469., 1960.

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

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

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

[25] L. Gerencsér. Fixed gain off-line estimators of ARMA parameters. *Journal of Mathematical Systems, Estimation and Control*, 4(2):249–252., 1994.

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

[27] 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.

[28] 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.

[29] P. Hall and C.C. Heyde. *Martingale Limit Theory and Its Application.* Academic Press, 1980.

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

[31] 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.

[32] I. Ibragimov and R. Khasminskii. *Statistical Estimation. Asymptotic Theory.* Springer Verlag, Berlin, 1981.

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

[34] V. Krishnamurthy and T. Rydén. Consistent estimation of linear and nonlinear autoregressive models with Markov regime. *J. Time Ser. Anal.*, 19 (3):291–307., 1998.

[35] 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.

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

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

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

[39] 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.

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

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

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

[43] L. Mevel. *Statistique asymptotique pour les modéles de Markov cachés.* Doctoral Thesis, Université de Rennes, 1997.

[44] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Springer Verlag, London, 1993.

[45] J. Neveu. *Discrete-Parameter Martingales.* North-Holland Publishing Company, 1975.

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

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

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

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

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

[51] E. Seneta. *Non-negative Matrices and Markov Chains.* Springer Verlag, New York, 1981.

[52] 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.

[53] G.E. Tusnady and I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol.*, 283(2):489–506., 1998.