STATISTICAL ANALYSIS OF HIDDEN MARKOV MODELS

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

Ezen értekezés bírálatai és a védésről készült jegyzőkönyv a későbbiekben a Budapesti Műszaki és Gazdaságtudományi Egyetem Természettudományi Karának Dékáni Hivatalában elérhető.

Alulírott Molnár-Sáska Gábor kijelentem, hogy ezt a doktori értekezést magam készítettem és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

.....
Molnár-Sáska Gábor

Preface

The present thesis contains results of the research that the author carried out with his supervisor László Gerencsér between January 2001 and May 2005 while he was a PhD student at the Mathematics Doctoral School of the Technical University of Budapest (Applied mathematics programme, head of school: József Fritz) and junior research fellow at the Computer and Automation Institute of the Hungarian Academy. The grants, as well as the hospitality of the above institutions are gratefully acknowledged.

I owe a debt of gratitude to my supervisor, László Gerencsér for his constant care and guidance. I also thank Domokos Szász and Marianna Bolla for their help.

Abstract

In the present thesis a new approach for the statistical analysis of Hidden Markov Models (HMM-s), in particular for the analysis of the maximum-likelihood estimate, is laid down. Useful connection between the estimation theory of HMM-s and linear stochastic systems is established via the theory of L-mixing processes.

Our analysis is applicable to HMM-s with a general state-space and read-out space, assuming that the state process satisfies the Doeblin's condition. The key technical results give conditions for the functions of the input-output process of a non-linear stochastic systems to be L-mixing. This is then applied to HMM-s extended by the filter process. Several applications are presented: we state a strong approximation theorem for finite state HMM-s, we give an on-line estimation procedure, and we deal with the fixed-gain estimation of HMM-s and apply the results for change detection.

Contents

Chapter 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 nanotecnology [31], quantized Gaussian linear regression [17, 18], 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]. An extension of HMM-s allowing dynamic memory is presented in [49].

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]. Strong consistency for the maximum-likelihood estimate for continous-time HMM-s with finite state-space and Gaussian read-out has been established by Moore and Elliott using martingale-theory in [14]. Adaptive control of HMM-s has been considered in Duncan et al. [11].

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 processes, 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 3.2.1 below.

Although the state space and the read-out space of a general HMM may have no algebraic structure, the filter process is known to be generated by a non-linear algebraic recursion, known as the Baum-equation, with the observation process as the input process. Uniform exponential stability of this non-linear dynamic system has been investigated in several papers, see e.g. [2], [40]. This stability property will play a major role in establishing L-mixing of the extended HMM.

The structure of the thesis is the following: Chapter 2 contains the definitions and an overview of known related results. In Chapter 3 the key technical tools are given 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 the results of Chapter 3 to HMM-s will be given in Chapter 4. HMM-s with finite state-space and general read-out space, under Doeblin-condition for the state-process will be given in Section 4.1. To conclude Section 4.1 we compare our conditions with those of [40] that ensure geometric ergodicity of the extended process. The results of Section 4.1 are extended to HMM-s with general compact state space in Section 4.2.

In further chapters applications of our results in the statistical analysis of HMM-s are presented. In Chapter 5 the recursive estimation of HMMs is investigated. In Chapter 6 we state a strong approximation theorem for finite state HMM-s, inspired by [24]. This fine characterization of the estimator process is not of purely academic interest: it plays a key role in the analysis of the effect of statistical uncertainty and in certain problems of stochastic complexity, see e.g. [26].

In Chapter 7 we follow the same route as in Chapter 6, but this time for Hidden Markov Models with fixed gain or forgetting rate λ. We also establish an explicit formula for the error term. In Chapter 8, using the above representation of the error term, we investigate the effect of parameter uncertainty on the performance of an adaptive encoding procedure. Using

this result and ideas from the theory of stochastic complexity, a change point detection method for HMM-s is developed.

Chapter 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. Throughout the dissertation capital letters with lower index n, such as (X_n), will denote discrete-time processes on the positive axis, i.e. $n \in \mathbb{N}$, if not otherwise stated.

Definition 2.1.1 *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).*

Example 2.1.2 *Assume that the observations are of the form*

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

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

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.

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

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

In this case we will use the following notation:

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

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

$$P(Y_n \in dy | X_n = x) = b^{*x}(y)\lambda(dy), \tag{2.1}$$

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

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

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

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

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

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

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

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

$$p^*_{n+1} = \pi(Q^{*T} B^*(Y_n) p^*_n), \tag{2.2}$$

both in discrete and continuous read-out cases, where π is the normalizing operator: for $x \in \mathbb{R}^N$, $x \geq 0$, $x \neq 0$ set $\pi(x)^i = x^i / \sum_{j=1}^N x^j$. Here $p_0^{*j} = P(X_0 = j)$.

In practice, the transition probability matrix Q^* and the initial probability distribution p_0^* of the unobserved Markov chain (X_n) as well as the conditional probabilities $b^{*i}(y)$ of the observation sequence (Y_n) are possibly unknown. For this reason we consider the Baum-equation in a more general sense:

$$p_{n+1} = \pi(Q^T B(Y_n) p_n), \tag{2.3}$$

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

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(1-\delta)^n \|q - q'\|_{TV}, \tag{2.4}$$

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

That is, the filter forgets its initial condition with an exponential rate. An essential feature of the result is that $\|q - q'\|_{TV}$ shows up in the upper bound, see [2]. We note that Proposition 2.1.3 is a purely linear algebraic statement,

i.e. there is no need for probability. We also note that the total variation norm is not required in this result as the vectors $p_n(q) \in \mathbb{R}^N$ are in a finite dimensional space. We will need the total variation norm when the state space is not finite, see Section 4.2.

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (2.4) holds with a random C, see [40].

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.

For the log-likelihood function we have

$$\log p(y_0, \dots y_n, \theta) = \sum_{k=1}^{n-1} \log p(y_k | y_{k-1}, \dots y_0, \theta) + \log p(y_0, \theta). \tag{2.5}$$

The k-th term in (2.5) for $k \geq 1$ can be written as

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

Now write

$$g(y, p, \theta) = \log \sum_i b^i(y, \theta) p^i(\theta), \tag{2.6}$$

then we have

$$\log p(y_N, \dots, y_0, \theta) = \sum_{k=1}^{N} g(y_k, p_k, \theta) + \log p(y_0, \theta). \tag{2.7}$$

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

$$\lim_{N \to \infty} \frac{1}{N} \log p(y_0, \dots y_N, \theta) \tag{2.8}$$

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

The limit of (2.8) was investigated in various setup in the literature. An overview is presented in the next section.

2.2 Entropy Ergodic Theorems

We review ergodic theorems for the sample entropy and relative entropy densities of HMM-s. The fundamental ergodic theorem for the sample entropy of a stationary ergodic finite state process, not necessarily an HMM, is given by the Shannon-Breiman-McMillan theorem. Let (Y_n) denote such a process and let P_Y denote its distribution. Let $p(y_1, \dots y_n)$ denote the n-dimensional probability mass function induced by P_Y. The theorem states that

$$\lim_{n\to\infty} -\frac{1}{n}\log p(Y_1, \dots Y_n) = H(Y) \qquad P_Y - \text{a.s.},$$

where

$$H(Y) = \lim_{n\to\infty} \frac{1}{n} E_{P_Y}(-\log p(Y_1, \dots Y_n))$$

is the entropy rate of (Y_n)

This theorem implies for a finite state finite read-out HMM:

$$\lim_{n\to\infty} -\frac{1}{n}\log p(Y_1, \dots, Y_n; \theta^*) = H(\theta^*) \qquad P_{\theta^*}\text{-a.s.}$$

The Shannon-Breiman-McMillan theorem has been generalized by Barron to non-discrete processes , see [4]. Let (Y_n) be a stochastic process on a probability space $(\Omega, \mathcal{B}, P)$. Suppose that the joint distribution P_n for $(Y_1, \dots, Y_n)$ has a probability density function $p_n(y_1, \dots, y_n)$ with respect to a σ-finite measure M_n. Let $p(Y_{n+1}|Y_1, \dots, Y_n)$ denote the conditional density for $n \geq 1$. Then we have

Proposition 2.2.1 *(Barron 1985, [4]) If (Y_n) is a stationary ergodic process and there exists an integer m such that for all $n \geq m$*

$$E \log p(Y_{n+1}|Y_1, \dots Y_n) > -\infty,$$

then the sequence of relative entropy densities $\frac{1}{n}\log p_n(Y_1, \dots, Y_n)$ converges almost surely to the relative entropy rate, i.e.

$$\lim_{n\to\infty} \frac{1}{n}\log p_n(Y_1, \dots Y_n) = D,$$

where

$$D = \lim_{n\to\infty} E \log p(Y_{n+1}|Y_1, \dots Y_n).$$

Let P_θ denote a distribution of the misspecified Hidden Markov Model and let $p(y_1, \ldots, y_n, \theta)$ denote the induced n-dimensional density. A central question in estimation problems is the ergodic theorem for $\log p(Y_1, \ldots, Y_n, \theta)$, i.e. the existence of the limit

$$\lim_{n\to\infty} \frac{1}{n} \log p(Y_1, \ldots, Y_n, \theta), \tag{2.9}$$

where (Y_n) is a stationary ergodic HMM with distribution P_{θ^*}. Leroux proved the existence of a limit (2.9) for a stationary ergodic general HMM, see [41].

Proposition 2.2.2 *(Leroux 1992, [41]) Assume that the Markov chain (X_n) is irreducible and aperiodic and observation conditional densities satisfy*

$$E_{\theta^*}(|\log b(Y_1, \theta_j(\theta^*))|) < \infty \qquad \textit{for } j = 1, \ldots N.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log p(Y_1, \ldots, Y_n, \theta) = H(\theta, \theta^*) \qquad P_{\theta^*} - a.s.,$$

where

$$H(\theta, \theta^*) = \lim_{n\to\infty} \frac{1}{n} E_{\theta^*}(\log p(Y_1, \ldots, Y_n, \theta) < \infty.$$

The theorem is proved using Kingman's ergodic theorem for subadditive processes (see [41]).

Similar ergodic theorems for relative entropy densities of several extensions of standard HMM-s were recently proved under suitable conditions. Francq and Roussignol [20] studied stationary ergodic switching autoregressive processes with finite-state Markov regime defined by

$$Z_n = g(Z_{n-1}, X_n, \theta) + h(V_n, S_n, \theta), \tag{2.10}$$

where Z_n is a sequence of r-dimensional random vectors, X_n is a finite-state Markov chain, V_n is a sequence of i.i.d. k-dimensional random vectors independent of X_n, $g(\cdot, \cdot, \cdot)$ and $h(\cdot, \cdot, \cdot)$ are measurable functions from $\mathbb{R}^r \times \mathcal{X} \times \Theta$ to $\mathbb{R}^r$ and from $\mathbb{R}^k \times \mathcal{X} \times \Theta$ to $\mathbb{R}^r$, respectively. They proved an ergodic theorem for the normalized conditional log-likelihood

$$\frac{1}{n} \log p(Z_1, \ldots, Z_n | z_0, \theta)$$

by expressing the conditional density as a product of random matrices and then applying the Furstenberg and Kesten ergodic theorem, see [21]. The sequence converges almost surely to the upper Lyapunov exponent of the sequence of auxiliary random matrices. The standard HMM is a special case of (2.10) which corresponds to $g(\cdot,\cdot,\cdot) \equiv 0$.

Krishnamurty and Rydén studied stationary ergodic switching autoregressive processes with finite-state Markov regime described by

$$Y_n = g(Y_{n-r}, \ldots, Y_{n-1}, X_n, W_n, \theta),$$

where (Y_n) is a scalar process, g is an arbitrary measurable function, and (W_n) is a scalar i.i.d. process. They arrived at a similar ergodic theorem for the normalized conditional log-likelihood using also Kingman's ergodic theorem following Leroux, see [34].

For misspecified HMM-s, defined in Section 2.1, the predictive filter (p_n) is not a Markov chain under P_{θ^*}, but the triplet (state, observation, wrong predictive filter) is a Markov chain. Let $Z_n = (X_n, Y_n, p_n)$ denote this extended Markov chain. LeGland and Mevel proved geometric ergodicity of the Markov chain Z_n and showed existence of a unique invariant distribution under suitable conditions. In particular, this property implies an ergodic theorem for finite-state general HMMs similar to (2.9).

Douc and Matias [9] extended this approach to a general HMM with a compact state space that is not necessarily finite. They developed an ergodic theorem for an HMM with arbitrary, not necessarily stationary initial state density.

Douc, Moulines and Rydén [10] studied general forms of switching autoregressive processes with a compact state space that is not necessarily finite. They proved an ergodic theorem similar to (2.9) for almost sure and L_1 convergence of the normalized conditional log-likelihood of the observation sequence. They relied on uniform exponential forgetting of the initial distribution of the inhomogeneous Markov chain representing the states given the observation sequence.

2.3 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.

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 2.3.1 *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 $X_n/c_n = O_M(1)$.

This definition extends to parameter-dependent stochastic processes $(X_n(\theta))$, $n \geq 0$. Let $D \subset \mathbb{R}^p$ be an open domain. A parameter-dependent stochastic process $(X_n(\theta))$, $n \geq 0$ is a sequence of measurable mappings for $n \geq 0$ from $(\Omega \times D, \mathcal{F} \bigotimes \mathcal{B}(D))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Here $\mathcal{B}(D)$ denotes the σ-field of Borel-sets of D. For each fixed n, $(X_n(\theta))$ can be considered as a random field over D. In this case we require

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

We say that a sequence of random variables X_n tends to a random variable X in the M-sense if for all $q \geq 1$ we have

$$\lim_{n \to \infty} E^{\frac{1}{q}} |X_n - X|^q = 0.$$

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. A standard example is

$$\mathcal{F}_n = \sigma\{e_i : i \leq n\} \qquad \mathcal{F}_n^+ = \sigma\{e_i : i > n\}, \tag{2.11}$$

where (e_i), $i \geq 0$ is an independent sequence of random variables.

Definition 2.3.2 *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

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

The definition extends to parameter-dependent stochastic processes. We say that a stochastic process $(X_n(\theta))$, $n \geq 0$ is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ uniformly in θ, if it is $\mathcal{F}_n \bigotimes \mathcal{B}(D)$-adapted, M-bounded and and for $1 \leq q < \infty$ and $\tau \in \mathbb{Z}^+$

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

is such that

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

In subsequent discussions we often speak about L-mixing processes without making explicit reference to $(\mathcal{F}_n, \mathcal{F}_n^+)$, provided that this does not lead to ambiguity. A basic example of L-mixing processes is obtained as follows: let (e_n), $n \geq 0$, $e_n \in \mathbb{R}^k$, be an M-bounded, independent sequence of random variables and define a vector-valued process (y_n) by

$$x_{n+1} = Ax_n + Be_n \qquad y_n = Cx_n$$

with $A \in \mathbb{R}^{n \times n}$ stable, $B \in \mathbb{R}^{n \times k}$, $C \in \mathbb{R}^{p \times n}$ and $x_0 = 0$. It is easy to see that the process (y_n), $n \geq 0$ is L-mixing with respect to the $(\mathcal{F}_n, \mathcal{F}_n^+)$ defined in (2.11).

To verify that a given process (X_n) is L-mixing, the definition requires the computation of $E(X_n | \mathcal{F}_{n-\tau}^+)$. However, a much simpler method is to find just any $\mathcal{F}_{n-\tau}^+$-measurable random variable, which approximates X_n with reasonable accuracy, and then use the following lemma (see Lemma 2.1 in [23]).

Lemma 2.3.3 *Let ξ, η be two random variable with finite moments of all orders. Let $\mathcal{F}' \subset \mathcal{F}$ be some σ-field and let η be an $\mathcal{F}'$-measurable random variable. Then for all $1 \leq q < \infty$ we have*

$$E^{\frac{1}{q}}|\xi - E(\xi|\mathcal{F}')|^q \leq 2E^{\frac{1}{q}}|\xi - \eta|^q.$$

The lemma implies that for $\tau \leq \tau'$ we have

$$E^{\frac{1}{q}}|X_n - E(X_n|\mathcal{F}^+_{n-\tau'})|^q \leq 2E^{\frac{1}{q}}|X_n - E(X_n|\mathcal{F}^+_{n-\tau})|^q.$$

It follows that, although $\gamma_q(\tau)$ is in general not monotone decreasing in τ, we have for $1 \leq q < \infty$, $\tau \leq \tau'$

$$\gamma_q(\tau') \leq 2\gamma_q(\tau).$$

A fundamental technical tool in estimation theory is the following moment inequality given in [23] (Theorem 1.1).

Theorem 2.3.4 *(Gerencsér 1989, [23]) Let (X_n), $n \geq 0$ be a real-valued L-mixing process with $EX_n = 0$ for all n and let (f_n) be a deterministic sequence. Then we have for all $1 \leq m < \infty$*

$$E^{\frac{1}{2m}}|\sum_{n=0}^{N} f_n X_n|^{2m} \leq C_m\big(\sum_{n=0}^{N} f_n^2\big)^{1/2} M_{2m}^{1/2}(x)\Gamma_{2m}^{1/2}(x),$$

where C_m depends only on m. We can take $C_m = 4(2m-1)^{1/2}$.

Two applications of this theorem are given below. In the first we take $f_n = 1$ for all n. In the second the process (X_n) is subject to exponential smoothing.

Theorem 2.3.5 *(Gerencsér 1989, [23]) Let (X_n), $n \geq 0$ be a real-valued L-mixing process with $EX_n = 0$ for all n. Then we have for all $1 \leq m < \infty$*

$$E^{\frac{1}{2m}}|\frac{1}{N}\sum_{n=0}^{N} X_n|^{2m} \leq C_m N^{1/2} M_{2m}^{1/2}(x)\Gamma_{2m}^{1/2}(x),$$

where $C_m = 4(2m-1)^{\frac{1}{2}}$. In short, $N^{-1}\sum_{n=0}^{N} X_n = O_M(N^{1/2})$.

Theorem 2.3.6 *(Gerencsér 1989, [23]) Let (X_n), $n \geq 0$ be an L-mixing process with $EX_n = 0$ for all n. Then for any $0 < \lambda < 1$ and for all $1 \leq m < \infty$ we have*

$$E^{\frac{1}{2m}} |\sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda X_n|^{2m} \leq C_m \lambda^{1/2} M_{2m}^{1/2}(x) \Gamma_{2m}^{1/2}(x),$$

where $C_m = 4(2m-1)^{1/2}$. In short, $\sum_{n=1}^{N} (1-\lambda)^{N-n} \lambda X_n = O(\lambda^{1/2})$.

An important technical tool is an inequality that provides an upper bound for the maximal value of random fields. Let $(X_n(\theta))$ be a random field defined for $\theta \in D \subset \mathbb{R}^p$. Let $\alpha > 0$, and define another random field $(\triangle X_n / \triangle^\alpha \theta)$ by

$$(\triangle X_n / \triangle^\alpha \theta)(\theta, \theta + h) = |X_n(\theta + h) - X_n(\theta)| / |h|^\alpha,$$

for $n \geq 0$, $\theta \neq \theta + h \in D$.

Definition 2.3.7 *The random field $(X_n(\theta))$ is M-Hölder-continuous in θ with exponent α, if the process $(\triangle X_n / \triangle^\alpha \theta)$ is M-bounded, i.e. if for all $1 \leq q < \infty$ we have*

$$M_q(\triangle X_n / \triangle^\alpha \theta) = \sup_{n \geq 0, \theta \neq \theta + h \in D} E^{\frac{1}{q}} |X_n(\theta + h) - X_n(\theta)|^q / |h|^\alpha < \infty.$$

If $\alpha = 1$ then we say that $X_n(\theta)$ is M-Lipschitz-continuous.

Let $(X_n(\theta))$ be a measurable, separable, M-bounded random field that is M-Hölder-continuous in θ with exponent α for $\theta \in D$. By Kolmogorov's theorem [32] the realizations of $(X_n(\theta))$ are continuous in θ with probability 1, hence for $D_0 \subset D$ being a compact domain, we can define for almost all ω

$$X_n^* = \max_{\theta \in D_0} |X_n(\theta)|.$$

An upper bound could be given for the moments of the process (X_n^*), see Theorem 3.4 in [23].

Theorem 2.3.8 *(Gerencsér 1989, [23]) Assume that $(X_n(\theta))$ is a measurable, separable, M-bounded random field, which is M-Hölder-continuous with exponent α for $\theta \in D \subset \mathbb{R}^p$. Then we have for all $q \geq 1$ and $r > p/\alpha$*

$$E^{\frac{1}{q}} (X_n^*)^q \leq C(M_{qr}(X) + M_{qr}(\triangle X / \triangle^\alpha \theta)),$$

where C depends only on α, p, q, r and D, D_0.

A useful application of the above result is obtained by combining it with Theorem 2.3.5 to get the following uniform version of Theorem 2.3.5, see Theorem 1.2 in [23].

Theorem 2.3.9 *(Gerencsér 1989, [23]) Let $(X_n(\theta))$, $n \geq 0$ be a zero-mean, measurable and separable stochastic process. Assume that X_n and $\triangle X_n/\triangle^\alpha\theta$ are L-mixing uniformly in θ for $\theta \in D \subset \mathbb{R}^p$ and let D_0 be as above. Then we have for all $m \geq 1$ and $r > p/\alpha$*

$$E^{\frac{1}{2m}} \max_{\theta\in D_0} |\frac{1}{N}\sum_{n=1}^{N} X_n(\theta)|^{2m} \leq CN^{-1/2}\big(M'_{2mr}(x)\Gamma'_{2mr}(x)\big)^{1/2},$$

where

$$M'_{2m}(x) = M_{2m}(x) + M_{2m}(\triangle x/\triangle^\alpha\theta) \text{ and } \Gamma'_{2m}(x) = \Gamma_{2m}(x) + \Gamma_{2m}(\triangle x/\triangle^\alpha\theta),$$

and C depends only on α, p, m, r and the domains D, D_0. In short we can write

$$\max_{\theta\in D_0} |N^{-1}\sum_{n=1}^{N} X_n(\theta)| = O_M(N^{-1/2}).$$

From here a uniform law of large numbers can easily be derived using Markov's inequality and a Borel-Cantelli argument:

Theorem 2.3.10 *(Gerencsér 1989, [23]) Let $(X_n(\theta))$, $n \geq 0$, D_0 be as above. Then we have almost surely*

$$\lim_{N\to\infty} \max_{\theta\in D_0} |\frac{1}{N}\sum_{n=1}^{N} X_n(\theta)| = 0.$$

Combining Theorem 2.3.6 and 2.3.8 we have a similar result for the case when X_n is subject to exponential smoothing:

Theorem 2.3.11 *(Gerencsér 1989, [23]) Let $X_n(\theta)$ be an L-mixing process uniformly in $\theta \in D$ such that $EX_n(\theta) = 0$ for all $n \geq 0$, $\theta \in D$ and assume that $(\triangle X_n/\triangle\theta)$ is also L-mixing, uniformly in $\theta, \theta + h \in D$. Let $0 < \lambda < 1$, then we get*

$$\sup_{\theta\in D^*} \left|\sum_{n=1}^{n} (1-\lambda)^{N-n}\lambda X_n(\theta)\right| = O_M(\lambda^{1/2}).$$

Chapter 3

Exponentially stable systems

3.1 Representation of Markov processes

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

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

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

A converse result is given in the following proposition:

Proposition 3.1.1 *Let (X_n) be a Markov process on a Polish space $\mathcal{X}$ with transition probabilities $P(x, G)$, $x \in \mathcal{X}$, $G \in \mathcal{B}(\mathcal{X})$. Then there exists a Borel measurable function $f : \mathcal{X} \times [0,1] \longrightarrow \mathcal{X}$ such that, with U being uniform in $[0,1]$ over some probability space $(\Omega, \mathcal{F}, \mathcal{Q})$, for all $x \in \mathcal{X}$ and $G \in \mathcal{B}(\mathcal{X})$ we have*

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

For the proof, see [33]. In the sequel we will denote the random mapping $f(\cdot, U_{n-1})$ by T_n, i.e. for $x \in \mathcal{X}$

$$T_n x = f(x, U_{n-1}). \tag{3.1}$$

The process defined by $X_{n+1} = T_{n+1} X_n$, $X_0 = x$ is Markov, if X_0 is independent of (T_n), $n \geq 1$.

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 the subsequent analysis.

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

Definition 3.1.2 *Let (X_n) be a Markov chain with state space $\mathcal{X}$. If there exists an integer $m \geq 1$, $\delta > 0$ and some probability measure ν on $\mathcal{B}(\mathcal{X})$ such that*

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

is valid for all $x \in \mathcal{X}$ and $A \subset \mathcal{B}(\mathcal{X})$, then we say that the Doeblin-condition is satisfied.

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

Lemma 3.1.3 *(Bhattacharya-Waymire 1999, [7]) 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. 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 $\nu(\cdot) = \mathbf{Q}(T_1 x \in \cdot | T_1 \in \Gamma_c)$ 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 obtained from a representation of a Markov chain with kernel function

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

■

Proposition 3.1.4 *(Bhattacharya-Waymire 1999, [7]) Assume that the Doeblin-condition holds with $m = 1$ for a Markov chain (X_n). Then there exists*

an invariant distribution π, and

$$|P^n(x,A) - \pi(A)| \leq (1-\delta)^n \quad \textit{for} \quad \forall A \in \mathcal{B}(\mathcal{X}). \tag{3.2}$$

Proof. Let (T_n) be the representation of the process. Consider the two-sided extension $(T_n)_{n=-\infty}^{\infty}$. Due to Lemma 3.1.3 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 any longer. Furthermore, the limit is independent from $\eta \in \mathcal{X}$.

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

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

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

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

Thus we obtained that the distribution π of X_0^* is invariant. So

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

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

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

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

■

Now let (X_n, Y_n) be a Hidden Markov process and assume that both the state space $\mathcal{X}$ and the observed space $\mathcal{Y}$ are Polish. The following lemma is the first new result of the thesis.

Lemma 3.1.5 *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 3.1.3. 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 3.1.1 there is a Borel measurable function $g : \mathcal{X} \times [0,1] \to \mathcal{Y}$ such that, with V being uniform in $[0,1]$ over some probability space $(\Omega, \mathcal{F}, \mathcal{Q}')$, for all $x \in \mathcal{X}$ and $G \in \mathcal{B}(\mathcal{Y})$ we have $P(x, G) = \mathbf{Q}'\{g(x, V) \in G\}$. Consider a sequence of i.i.d uniformly distributed random variables (V_n) and let us denote the random mapping $g(\cdot, V_{n-1})$ by U_n. Thus we have $Y_n = U_n X_n$ and

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

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

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

and taking into account Lemma 3.1.3, the lemma follows.

■

Remark 3.1.6 *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 3.1.4 and Lemma 3.1.5 also valid if the Doeblin-condition holds for $m \geq 1$.

In the subsequent statements we always consider the case $m = 1$ for simplicity.

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. Indeed, the output process in not necessarily a Markov chain. In the following theorem we prove that the output process is L-mixing.

Theorem 3.2.1 *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

$$X_{n,m}^+ = T_n \dots T_{m+1} X^*, \tag{3.3}$$

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 2.3.3 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. ■

3.3 Exponentially stable random mappings I.

Now we formulate a general concept of exponential stability motivated by Proposition 2.1.3. 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

$$z_{n+1} = f(x_n, z_n), \quad z_0 = \xi. \tag{3.4}$$

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

Definition 3.3.1 *The mapping f is uniformly exponentially stable if for every sequence $(x_n)_{n\geq 0}$, $x_n \in \mathcal{X}$*

$$\|z_n(\xi) - z_n(\xi')\| \leq C(1-\varrho)^n \|\xi - \xi'\|, \tag{3.5}$$

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]. Let $z(n, m, \xi)$ denote the solution of (3.4) initialized at $z_m = \xi$ with $m \leq n$. Let us consider an arbitrary discrete sequence defined by recursion of the form

$$\overline{z}_{n+1} = \overline{f}_n(\overline{z}_n) \tag{3.6}$$

with the same starting point $\overline{z}_0 = \xi$. Extending a simple analytical lemma given in [22] from continuous to discrete time we get

Lemma 3.3.2 *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 3.3.3 *For the solution of (3.4) 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 3.3.2 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

$$Z_{n+1} = f(X_n, Z_n), \quad Z_0 = \xi, \tag{3.7}$$

where (X_n) is a Markov chain satisfying the Doeblin condition. Due to Proposition 3.1.4 an invariant distribution of (X_n) exists. Let us denote it by π . To prove M-boundedness of (Z_n) we impose following conditions:

Condition 3.3.4 *Let the distribution of X_0 be π_0. Assume*

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

Condition 3.3.5 *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,

$$\int_{\mathcal{X}} \|f(x,\xi)\|^q d\pi(x) \le K_1(\xi) < \infty, \tag{3.8}$$

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

Lemma 3.3.6 *Assume Condition 3.3.4. Then we have*

$$\frac{d\pi_n}{d\pi} \le C_1 \quad \textit{for all} \quad n. \tag{3.9}$$

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

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

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

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

Lemma 3.3.7 *Assume Condition 3.3.4 and 3.3.5. Then we have*

$$E\|f(X_n,\xi)\|^q \le K_1(\xi) C_1. \tag{3.10}$$

Proof. We have

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

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

due to Lemma 3.3.6 and Condition 3.3.5. ■

Lemma 3.3.8 *Let the mapping $f(x,z)$ be uniformly exponentially stable, and let Condition 3.3.4 and 3.3.5 hold. Then the process (Z_n) defined by (3.7) with any fixed constant $Z_0 = \xi$ is M-bounded.*

Proof. Using Lemma 3.3.3 and the exponential stability of f we have

$$\|Z_n\| \leq \|\xi\| + \sum_{m=0}^{n-1} C(1-\varrho)^{n-m-1}\|f(X_m,\xi)-\xi\|. \tag{3.11}$$

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 3.3.9 *$g(x,z)$ 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 3.3.10 *Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition with $m = 1$, and (Z_n) is defined by (3.7) 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 3.3.4 and 3.3.5 hold. Furthermore let $g(x,z)$ be a bounded function satisfying Condition 3.3.9 Then*

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

is an L-mixing process.

Remark 3.3.11 *Theorem 3.3.10 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 3.2.1, except that the distribution of X^* be stationary (independent of T_i) and

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

Let an approximation of (Z_n) be defined recursively by

$$Z_{k+1,m}^+ = f(X_{k,m}^+, Z_{k,m}^+), \tag{3.12}$$

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 $(m, m']$:

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

Due to the Doeblin-condition

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

Now consider the event

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

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

$$Z_{k+1,m}^+ = f(X_k, Z_{k,m}^+) \quad \text{for} \quad m' < k \leq n, \tag{3.13}$$

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

The process (Z_k) considered for $m' \leq k \leq n$ satisfies

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

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

$$\|Z_{n,m}^+ - Z_n\| \leq C(1-\varrho)^{[\frac{\tau}{2}]} \|Z_{m',m}^+ - Z_{m'}\|. \tag{3.14}$$

Obviously for $q \geq 1$

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

$$E^{\frac{1}{q}}(\chi_B\|g(X_n,Z_n)-g(X^+_{n,m},Z^+_{n,m})\|^q)$$
$$+E^{\frac{1}{q}}(\chi_{B^C}\|g(X_n,Z_n)-g(X_n,Z^+_{n,m})\|^q). \tag{3.15}$$

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

$$E^{\frac{1}{q}}(\chi_B\|g(X_n,Z_n)-g(X^+_{n,m},Z^+_{n,m})\|^q)\leq$$
$$E^{\frac{1}{q}}(\chi_B(2K)^q)=2KP^{\frac{1}{q}}(B), \tag{3.16}$$

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

Consider the second term of the expression (3.15).

$$\begin{aligned}E^{\frac{1}{q}}(\chi_{B^C}\|g(X_n,Z_n)-g(X_n,Z^+_{n,m})\|^q) &\leq\\ E^{\frac{1}{q}}\|g(X_n,Z_n)-g(X_n,Z^+_{n,m})\|^q &\leq\\ E^{\frac{1}{q}}(L\|Z_n-Z^+_{n,m}\|)^q &\leq\\ E^{\frac{1}{q}}(LC(1-\varrho)^{[\frac{\tau}{2}]}\|Z^+_{m',m}-Z_{m'}\|)^q &=\\ LC(1-\varrho)^{[\frac{\tau}{2}]}E^{\frac{1}{q}}\|Z^+_{m',m}-Z_{m'}\|^q &\end{aligned} \tag{3.17}$$

The second inequality is due to the Lipschitz-continuity of g, and the third inequality follows from (3.14). Using the Minkowski inequality, Condition 3.3.5 and Lemma 3.3.8 (the distribution of X^* is stationary) we have that $Z_{m'}$ and $Z^+_{m',m}$ are M-bounded

$$E^{\frac{1}{q}}\|Z^+_{m',m}-Z_{m'}\|^q\leq E^{\frac{1}{q}}\|Z^+_{m',m}\|^q+E^{\frac{1}{q}}\|Z_{m'}\|^q\leq S, \tag{3.18}$$

and so

$$E^{\frac{1}{q}}\|g(X^+_n,Z^+_{m,n})-g(X_n,Z_n)\|^q\leq 2K(1-\delta)^{\frac{[\tau/2]}{q}}+K'(1-\varrho)^{[\tau/2]}, \tag{3.19}$$

where $K'=LCS$.

Now we are going to apply Lemma 2.3.3 and obtain

$$\gamma_q(\tau)\leq 2(2K(1-\delta)^{\frac{[\tau/2]}{q}}+K'(1-\varrho)^{[\tau/2]}). \tag{3.20}$$

Thus

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

$$\sum_{\tau=0}^{\infty}(4K(1-\delta)^{\frac{[\tau/2]}{q}}+2K'(1-\varrho)^{[\tau/2]})<\infty, \tag{3.21}$$

hence the claim of the theorem follows. ■

In some applications Condition 3.3.9 is too strong. Hence we should weaken this condition as follows:

Condition 3.3.12 *$g(x,z)$ 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 $q\geq 1$*

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

Theorem 3.3.13 *Consider the process (X_n,Z_n), where (X_n) satisfies the Doeblin-condition with $m=1$, and (Z_n) is defined by (3.7) 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 3.3.4 and 3.3.5 hold. Furthermore let $g(x,z)$ be a bounded function satisfying Condition 3.3.12 Then*

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

is an L-mixing process.

Proof. There is only one place, (3.17), where we have used the Lipschitz continuity, Condition 3.3.9, of $g(x,z)$. Let us replace (3.17) with the following train of thought:

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

$$E^{\frac{1}{2q}}|L(X_n)|^{2q}E^{\frac{1}{2q}}\|Z_n-Z_{n,m}^+\|)^{2q}$$

by Condition 3.3.12 and the Hölder inequality. Using Lemma 3.3.6 we have

$$\int_{\mathcal{X}}|L(x)|^{2q}d\pi_n(x)\leq\int_{\mathcal{X}}|L(x)|^{2q}C_1 d\pi(x)$$

Thus we have

$$E^{\frac{1}{q}}(\chi_{B^C}\|g(X_n,Z_n)-g(X_n,Z^+_{n,m})\|^q)\le$$
$$C_1^{\frac{1}{2q}}L_{2q}C(1-\varrho)^{[\frac{\tau}{2}]}E^{\frac{1}{2q}}(\|Z^+_{m',m}-Z_{m'}\|)^{2q}$$

and can continue the proof of Theorem 3.3.10 from (3.17). ■

3.4 Exponentially stable random mappings II.

Considering that our motivation is to prove that $g(y_k,p_k)$ in (2.7) is an L-mixing process observe that g is not necessarily bounded due to the logarithm function, see (2.6). Thus, in this section we consider an extension of Theorem 3.3.10 for *unbounded* function g.

We need the following conditions for the function g.

Condition 3.4.1 *Assume that for all* $q\ge 1$

$$\int\limits_{\mathcal{X}}\sup_{z\in\mathcal{Z}}\|g(x,z)\|^q d\pi(x)\le M_q<\infty. \tag{3.22}$$

Lemma 3.4.2 *Conditions 3.3.4 and 3.4.1 imply that the process* $g(X_n,Z_n)$ *is* M*-bounded, i.e. for all* $q\ge 1$

$$E\|g(X_n,Z_n)\|^q<\infty. \tag{3.23}$$

Proof. Let us denote the distribution of (X_n,Z_n) by μ_n.

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

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

$$\int\limits_{\mathcal{X}}\sup_{z\in\mathcal{Z}}\|g(x,z)\|^q C_1 d\pi(x)\le M_qC_1. \tag{3.24}$$

■

We are going to generalize Theorem 3.3.10 to unbounded function g.

Theorem 3.4.3 *Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition with $m = 1$, and let (Z_n) be defined by (3.7) with a uniformly exponentially stable mapping f and an arbitrary constant initial condition $Z_0 = \xi$. Assume that X_0 is independent of $\{T_n\}$, $n \geq 1$, and Conditions 3.3.4 and 3.3.5 hold. Furthermore assume that Condition 3.3.9, 3.4.1 is satisfied for the function $g(x, z)$. Then*

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

is an L-mixing process.

Proof. The proof is analogous to the proof of Theorem 3.3.10. Consider the expression (3.15). The estimation of the second part is the same. Consider the first term. By the Hölder inequality we get

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

$$(E^{\frac{1}{2}}(\chi_B)^2 E^{\frac{1}{2}} \|g(X_n, Z_n) - g(X^+_{n,m}, Z^+_{n,m})\|^{2q})^{\frac{1}{q}}. \tag{3.25}$$

Due to the Minkowski inequality we have

$$E^{\frac{1}{2q}}(\|g(X_n, Z_n) - g(X^+_n, Z^+_{n,m})\|)^{2q} \leq$$

$$E^{\frac{1}{2q}} \|g(X_n, Z_n)\|^{2q} + E^{\frac{1}{2q}} \|g(X^+_n, Z^+_{n,m})\|^{2q}. \tag{3.26}$$

Both $E^{\frac{1}{2q}} \|g(X_n, Z_n)\|^{2q}$ and $E^{\frac{1}{2q}} \|g(X^+_n, Z^+_{n,m})\|^{2q}$ can be majorized by Lemma 3.4.2. (note that (X^+_n) is a stationary process, which implies that the conditions of Lemma 3.4.2 are satisfied). Thus the right hand side of (3.25) can be majorized by

$$P^{\frac{1}{2q}}(B)(2M_{2q}C_1)^{\frac{1}{2q}} \leq 2K_1 P^{\frac{1}{2q}}(B). \tag{3.27}$$

Let us turn back to the proof of Theorem 3.3.10. In the present case inequality (3.19) is replaced by

$$E^{\frac{1}{q}} \|g(X^+_n, Z^+_{m,n}) - g(X_n, Z_n)\|^q \leq 2K_1(1-\delta)^{\frac{[\tau/2]}{2q}} + K'(1-\varrho)^{[\tau/2]}. \tag{3.28}$$

Then we can continue the proof as in Theorem 3.3.10. ■

Remark 3.4.4 *Theorem 3.4.3 holds if the Doeblin-condition holds with* $m > 1$.

Following Theorem 3.3.13 we can weaken Condition 3.3.9 to 3.3.12. Then we get the following statement.

Corollary 3.4.5 *Consider the process* (X_n, Z_n), *where* (X_n) *satisfies the Doeblin-condition with* $m = 1$, *and let* (Z_n) *be defined by (3.7) with a uniformly exponentially stable mapping* f *and an arbitrary constant initial condition* $Z_0 = \xi$. *Assume that* X_0 *is independent of* $\{T_n\}$, $n \geq 1$, *and Conditions 3.3.4 and 3.3.5 hold. Furthermore assume that Condition 3.3.12, 3.4.1 is satisfied for the function* $g(x, z)$. *Then*

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

is an L-mixing process.

In the following we give some remarks on Condition 3.4.1. We start with a lemma on the existence of a stationary distribution for the process (X_n, Z_n).

Lemma 3.4.6 *Assume that the Doeblin-condition holds with* $m \geq 1$ *for the Markov process* (X_n), f *is uniformly exponentially stable mapping and Condition 3.3.5 holds. Then the process* (X_n, Z_n) *has a stationary distribution.*

Proof. Define X_{-n} as the limit

$$X_{-n} = \lim_{k \to \infty} T_{-n} \circ \cdots \circ T_{-n-k}\eta, \tag{3.29}$$

with any fixed η, similar to Proposition 3.1.4. It has been shown in the proof of Proposition 3.1.4 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 set

$$Z_0^* = \lim_n f_{X_{-1}} \circ \cdots \circ f_{X_{-n}}\xi. \tag{3.30}$$

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

$$\|f_{x_{-1}} \circ \cdots \circ f_{x_{-n}}\xi - f_{x_{-1}} \circ \cdots \circ f_{x_{-m}}\xi\|, \tag{3.31}$$

where $n < m$. Using notations like in Lemma 3.3.2 with $\varphi = z(-n-1, -m-1, \xi)$ we have

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

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

$$C(1-\varrho)^n\|\xi - \varphi\|, \tag{3.32}$$

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

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

$$C(1-\varrho)^n\big(\|\xi\| + E^{\frac{1}{q}}\|Z(-n-1, -m-1, \xi)\|^q\big), \tag{3.33}$$

and by Lemma 3.3.8 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

$$X_0 = \lim_n T_0 \circ T_{-1} \circ \cdots \circ T_{-n}\eta, \tag{3.34}$$

$$Z_0^* = \lim_n f_{X_{-1}} \circ \cdots \circ f_{X_{-n}}\xi. \tag{3.35}$$

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 3.1.4, and

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

$$\lim_n f_{X_0} \circ \cdots \circ f_{X_{-n}}\xi, \tag{3.36}$$

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

According to Lemma 3.4.6 under the conditions of Theorem 3.4.3 a stationary distribution of the process (X_n, Z_n) exists. Let this stationary distribution be denoted by μ and for an arbitrary initialization let the distribution of (X_n, Z_n) be μ_n. If we replace Condition 3.4.1 with the following conditions then Lemma 3.4.2 still holds true: the Radon-Nikodym derivative of μ_0 w.r.t. μ is bounded, say

$$\frac{d\mu_0}{d\mu} \leq K. \tag{3.37}$$

and

$$\int\limits_{\mathcal{X}\times\mathcal{Z}} \|g(x,z)\|^q d\mu(x,z) \leq M'_q. \tag{3.38}$$

(3.37) implies Condition 3.3.4 and we have (see Lemma 3.3.6)

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

thus indeed

$$E\|g(X_n, Z_n)\|^q \leq KM'_q. \tag{3.39}$$

Condition 3.4.1 is motivated by Legland and Mevel [40]. This condition is easier to use when we wish to analyze the log-likelihood function as it will be seen in Chapter 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 3.4.3 are not satisfied for this derivative process we need an extension of Theorem 3.4.3. Consider now the process $V_n = g(X_n, Z_n)$, where $g : \mathcal{X}\times\mathcal{Z} \to \mathbb{R}$ is a measurable function and (X_n, Z_n) is defined as in (3.7). We change Condition 3.3.9 to the following technical condition:

Condition 3.5.1 *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*

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

Furthermore assume that

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

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

Furthermore we weaken Condition 3.4.1.

Condition 3.5.2 *Assume that for all* $q \geq 1$

$$\int\limits_{\mathcal{X}} \sup_{z \in \mathcal{Z}} \left(\frac{\|g(x,z)\|}{\|z\|+1} \right)^q d\pi(x) \leq M_q < \infty. \tag{3.40}$$

A version of Lemma 3.5.3 is the following.

Lemma 3.5.3 *Conditions 3.3.4,3.3.5 and 3.5.2 imply that the process* $g(X_n, Z_n)$ *is* M*-bounded, i.e. for all* $q \geq 1$

$$E\|g(X_n, Z_n)\|^q < \infty. \tag{3.41}$$

Proof.

$$E\|g(X_n, Z_n)\|^q = \int\limits_{\mathcal{X}\times\mathcal{Z}} \left(\frac{\|g(x,z)\|}{\|z\|+1} \right)^q (\|z\|+1)^q d\mu_n(x,z) \leq$$

$$\left(\int\limits_{\mathcal{X}\times\mathcal{Z}} \left(\frac{\|g(x,z)\|}{1+\|z\|} \right)^{2q} d\mu_n(x,z) \right)^{1/2} \left(\int\limits_{\mathcal{X}\times\mathcal{Z}} (1+\|z\|)^{2q} d\mu_n(x,z) \right)^{1/2} \leq$$

$$\left(\int\limits_{\mathcal{X}\times\mathcal{Z}} \left(\sup_{z\in\mathcal{Z}} \frac{\|g(x,z)\|}{1+\|z\|} \right)^{2q} d\mu_n(x,z) \right)^{1/2} E^{1/2}(1+\|Z_n\|)^{2q} =$$

$$\left(\int\limits_{\mathcal{X}} \left(\sup_{z\in\mathcal{Z}} \frac{\|g(x,z)\|}{1+\|z\|} \right)^{2q} d\pi_n(x) \right)^{1/2} E^{1/2}(1+\|Z_n\|)^{2q} \leq$$

$$\leq M_{2q}^q C_1 C_2. \tag{3.42}$$

We have used here that Z_n is M-bounded by Lemma 3.3.8, i.e. $E^{1/2}(1+\|Z_n\|)^{2q} \leq C_2$ and $\frac{d\pi_n}{d\pi} \leq C_1$ by Lemma 3.3.6.

■

Theorem 3.5.4 *Consider the process (X_n, Z_n), where (X_n) satisfies the Doeblin-condition with $m = 1$, and let (Z_n) be defined by (3.7) with a uniformly exponentially stable mapping f and an arbitrary constant initial condition $Z_0 = \xi$. Assume that X_0 is independent of $\{T_n\}$, $n \geq 1$, and Conditions 3.3.4 and 3.3.5 hold. Furthermore assume that Conditions 3.5.1, 3.5.2 are satisfied for the function $g(x, z)$. Then*

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

is an L-mixing process.

Proof. The process $V_n = g(X_n, Z_n)$ is M-bounded by Lemma 3.5.3. For the L-mixing property we follow the same route as in the proof of Theorem 3.3.10. Let us repeat the proof up to the inequality (3.15).

Consider the second term of the right hand side of (3.15).

$$\begin{aligned}
E^{\frac{1}{q}}(\chi_{B^C}\|g(X_n, Z_n) - g(X_n, Z^+_{n,m})\|^q) &\leq \\
E^{\frac{1}{q}}\|g(X_n, Z_n) - g(X_n, Z^+_{n,m})\|^q &\leq \\
E^{\frac{1}{q}}(L\|Z_n - Z^+_{n,m}\|(\|Z_n\| + \|Z^+_{n,m}\|))^q &\leq \\
\left(E^{\frac{1}{3q}}(L)^{3q}\right)\left(E^{\frac{1}{3q}}\|Z_n - Z^+_{n,m}\|^{3q}\right)\left(E^{\frac{1}{3q}}(\|Z_n\| + \|Z^+_{n,m}\|)^{3q}\right) &\leq \\
L_{3q}E^{\frac{1}{3q}}(C(1-\varrho)^{[\frac{\tau}{2}]}\|Z^+_{m',m} - Z_{m'}\|)^{3q}\left(E^{\frac{1}{3q}}\|Z_n\|^{3q} + E^{\frac{1}{3q}}\|Z^+_{n,m}\|^{3q}\right) &\leq \\
\overline{C}_q(1-\varrho)^{[\frac{\tau}{2}]}\left(E^{\frac{1}{3q}}\|Z^+_{m',m}\|^{3q} + E^{\frac{1}{3q}}\|Z_{m'}\|)^{3q}\right)\left(E^{\frac{1}{3q}}\|Z_n\|^{3q} + E^{\frac{1}{3q}}\|Z^+_{n,m}\|^{3q}\right)&
\end{aligned}$$

The second inequality is due to Condition 3.5.1, the third inequality is due to the Hölder inequality, and the fourth inequality follows from the exponential stability of f.

Using that $Z^+_{m',m}$, $Z_{m'}$, Z_n and $Z^+_{n,m}$ are M-bounded, see Lemma 3.5.3, we have

$$E^{\frac{1}{q}}(\chi_{B^C}\|g(X_n, Z_n) - g(X_n, Z^+_{n,m})\|^q) \leq K'(1-\varrho)^{[\tau/2]}. \tag{3.43}$$

Consider now the first term of the right hand side of (3.15).

By the Hölder inequality we get

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

$$(E^{\frac{1}{2}}(\chi_B)^2 E^{\frac{1}{2}}\|g(X_n,Z_n)-g(X^+_{n,m},Z^+_{n,m})\|^{2q})^{\frac{1}{q}}. \tag{3.44}$$

Due to Minkowski inequality we have

$$E^{\frac{1}{2q}}(\|g(X_n,Z_n)-g(X^+_n,Z^+_{n,m})\|)^{2q} \leq$$

$$E^{\frac{1}{2q}}\|g(X_n,Z_n)\|^{2q}+E^{\frac{1}{2q}}\|g(X^+_n,Z^+_{n,m})\|^{2q} \tag{3.45}$$

and by Lemma 3.5.3 the right hand side of (3.44) is majorized by

$$P^{\frac{1}{2q}}(B)(2M^q_{2q}C_1C_2)^{\frac{1}{2q}} \leq 2KP^{\frac{1}{2q}}(B).$$

Thus we have

$$E^{\frac{1}{q}}(\chi_B\|g(X_n,Z_n)-g(X^+_{n,m},Z^+_{n,m})\|^q) \leq 2KP^{\frac{1}{2q}}(B). \tag{3.46}$$

Adding (3.43) and (3.46) we get

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

$$2K(1-\delta)^{\frac{[\tau/2]}{2q}}+K'(1-\varrho)^{[\tau/2]}. \tag{3.47}$$

The end of the proof is similar to the proof of Theorem 3.3.10.

■

3.6 On-line estimation

In this section we lay 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 Prioret, 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

$$\theta_{n+1} = \theta_n + \frac{1}{n} H(\theta_n, U_n), \tag{3.48}$$

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 $\mathcal{X}$.

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

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

and

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

Introduce the class of functions

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The associated ODE is then given by

$$\dot{\theta}_s = h(\theta_s). \tag{3.49}$$

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

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

$$\begin{aligned}
&(i) \quad \widetilde{U}(\theta^*) = 0,\ \widetilde{U}(\theta) > 0 \text{ for all } \theta \in D\backslash\{\theta^*\} \\
&(ii) \quad \widetilde{U}'(\theta)h(\theta) < 0 \text{ for all } \theta \in D\backslash\{\theta^*\} \\
&(iii) \ \widetilde{U}(\theta) \to \infty \text{ if } \theta \to \partial D \text{ or } |\theta| \to \infty.
\end{aligned}$$

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

Theorem 3.6.1 *(Benveniste-Métivier-Priouret 1990, [6]) Assume that Conditions A1 - A6 are satisfied, and ϵ is sufficiently small. Let $\theta \in \text{int}D_0$, $U_m = u \in \mathcal{U}$, 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*

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

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 3.3.1, and define the process (Z_n) by

$$Z_{n+1} = f(X_n, Z_n, \theta), \quad Z_0 = \xi, \tag{3.50}$$

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

$$X_{n+1} = T_n X_n, \tag{3.51}$$

where (T_n) is a sequence of i.i.d. random mappings, see (3.1). Let $U_n = (X_n, Z_n) \in \mathcal{X} \times \mathcal{Z} = \mathcal{U}$. Define the metric on $\mathcal{U}$ by

$$d(u, u') = \|z - z'\| + d_{\mathcal{X}}(x, x'), \tag{3.52}$$

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

$$V(u) = \|z\|. \tag{3.53}$$

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

Verification of BMP conditions

By Proposition 3.1.4 a stationary distribution of X_n exists. Let us denote it by π. For assumption (**A1**) we need two conditions: the first one ensures that there are no states in "large distances", the second one is (**A1**) for one-step when X_0 has an invariant distribution.

Condition 3.6.2 *Let the distribution of X_1 be π_1. Assume*

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

Condition 3.6.3 *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

$$\int_{\mathcal{X}} \|f(x, \xi)\|^p d\pi(x) \leq K_1(1 + \|\xi\|^p). \tag{3.54}$$

Note that Condition 3.6.2 is a modified version of Condition 3.3.4. As in assumptions (**A1**)-(**A3**) the initialization is always a fixed value and we need it for each initialization, Condition 3.3.4 is not realistic. Condition 3.6.3 is a special case of Condition 3.3.5.

Theorem 3.6.4 *Consider a process $U_n = (X_n, Z_n)$ defined by (3.7), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.6.2 and 3.6.3 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 $\theta \in Q$:*

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

Proof. Similar to Lemma 3.3.6 we have that Condition 3.6.2 implies that

$$\frac{d\pi_n}{d\pi} \leq C_1 \quad \text{for all} \quad n. \tag{3.55}$$

Repeating the arguments of Lemma 3.3.7 we have that

$$E\|f(X_n, \xi)\|^q \leq K_1(1 + \|\xi\|^p)C_1, \tag{3.56}$$

and similarly to Lemma 3.3.8 we have that

$$E\|Z_n\|^p \leq K(1 + \|\xi\|^p). \tag{3.57}$$

By the definition of the function V, see (3.53), we get the statement from (3.57). ■

Since we have not used the metric property in Theorem 3.6.4 $\mathcal{X}$ 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 3.6.5 *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 3.6.6 *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 3.6.7 *Consider a process $U_n = (X_n, Z_n)$ defined by (3.7), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.6.2, 3.6.3, 3.6.5 and 3.6.6 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 the proof of Theorem 3.6.7 we need a lemma first.

Lemma 3.6.8 *Consider a process $U_n = (X_n, Z_n)$ defined by (3.7), where f is an exponentially stable mapping. Assume that Conditions 3.6.5 and 3.6.6 are satisfied. Then we have*

$$Ed(U_n, U_n') \leq Kd(u_0, u_0'),$$

where K is independent of n.

Proof. By definition

$$d(u_n, u_n') = \|z_n - z_n'\| + d_{\mathcal{X}}(x_n, x_n').$$

To estimate $\|z_n - z_n'\|$ we use the idea of Lemma 3.3.2, but this time z_n and z_n' are not generated with the same sequence (x_n). To highlight the generating sequence (x_n) we introduce the following notations. For $k \geq i$ let

$$z_k^i = f(i, k, z_i', x_1^{n-1})$$

be the sequence starting from z_i' at step i with the generating process (x_n). Note that $z_i^i = z_i'$. With this notation we have

$$z_n - z_n' = (z_n - z_n^0) + \sum_{i=0}^{n-1}(z_n^i - z_n^{i+1}),$$

i.e.

$$\|z_n - z_n'\| \leq \|z_n - z_n^0\| + \sum_{i=1}^{n} \|z_n^{i-1} - z_n^i\|. \tag{3.58}$$

By the exponential stability of f we have

$$\|z_n^{i-1} - z_n^i\| \leq C\rho^{n-i}\|z_i^i - z_i^{i-1}\|, \tag{3.59}$$

and

$$\|z_n - z_n^0\| \leq C\rho^n\|z_0 - z_0'\|. \tag{3.60}$$

Furthermore,

$$\|z_i^i - z_i^{i-1}\| = \|f(z_{i-1}', x_{i-1}') - f(z_{i-1}', x_{i-1})\| \leq$$

$$Ld_{\mathcal{X}}(x_{i-1}, x_{i-1}') \tag{3.61}$$

by Condition 3.6.5. Using (3.59), (3.60) and (3.61) inequality (3.58) implies that

$$\|z_n - z_n'\| \leq C\rho^n\|z_0 - z_0'\| + \sum_{i=1}^{n} C\rho^{n-i}Ld_{\mathcal{X}}(x_{i-1}, x_{i-1}').$$

Taking the expectation of both sides and considering Condition 3.6.6 we get the lemma. ■

Let us turn to the proof of Theorem 3.6.7.

Proof. (Theorem 3.6.7) For $g \in Li(p)$ we have

$$|g(u_n) - g(u_n')| \leq \|\Delta g\|_p d(u_n, u_n')(1 + |V(u_n)|^p + |V(u_n')|^p). \tag{3.62}$$

Let $A = \{\omega : \ T_k(\omega) \in \Gamma_c \text{ for } k \leq n/2\}$. From Lemma 3.1.3 we have $P(A) = 1 - (1-\delta)^{n/2}$. On A we have $x_k = x_k'$ for all $n/2 \leq k \leq n$. Thus from the definition of d and the exponential stability of the mapping f we have on the set A

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

$$C\rho^{n/2}d(u_{n/2}, u_{n/2}').$$

Taking the expectation of both sides of (3.62) and considering that $Ed(U_{n/2}, U_{n/2}') \leq d(u_0, u_0')$ (see Lemma 3.6.8 and Theorem 3.6.4) we have

$$E\chi_A|g(U_n) - g(U_n')| \leq \|\Delta g\|_p C\rho^{n/2}d(u, u')(1 + |V(u)|^p + |V(u')|^p). \tag{3.63}$$

Consider now the complement of A. We have $P(A^c) = (1-\delta)^{n/2}$. Taking the expectation of (3.62) on the set A^c and using Lemma 3.6.8 we have

$$E\chi_{A^c}|g(U_n)-g(U'_n)| \leq (1-\delta)^{n/2}\|\Delta g\|_p d(u,u')(1+|V(u)|^p+|V(u')|^p) \quad (3.64)$$

Adding (3.63) and (3.64) we finish the proof.

■

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 3.6.9 *Consider a process $U_n = (X_n, Z_n)$ defined by (3.7), 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 3.6.2 and 3.6.3 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 $\theta, \theta' \in Q$:*

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

We start with a very important lemma which states that if the exponentially stable mapping f is smooth in the parameter θ then the derivative process $\partial z_n/\partial\theta$ is also an exponentially stable process.

Lemma 3.6.10 *Let f be a uniformly exponentially stable mapping smooth in θ. Then the derivative process $w_n = \frac{\partial z_n}{\partial\theta}$ is also exponentially stable, i.e. we have*

$$\|w_n(\eta) - w_n(\eta')\| \leq C(1-\varrho)^n\|\eta-\eta'\|, \quad (3.65)$$

where $\eta = \frac{\partial\xi}{\partial\theta}$ and $\eta' = \frac{\partial\xi'}{\partial\theta}$.

Proof. Let the derivative of z_n with respect to the initial condition ξ be v_n, i.e. $v_n = \frac{\partial z_n}{\partial\xi}$. Then we have

$$v_n = f_z(z_{n-1}, x_{n-1}, \theta)v_{n-1}. \quad (3.66)$$

Note that x_n and θ do not depend on the initialization ξ.

Define $v'_n = \frac{\partial w_n}{\partial \eta}$. For the derivative of w_n with respect to the initialization we have

$$v'_n = f_z(z_{n-1}, x_{n-1}, \theta)v'_{n-1} \tag{3.67}$$

We used that x_n and θ do not depend on the initialization η.

Comparing (3.66) and (3.67) we have that if the filter process is exponentially stable then the same property holds for its derivative. ■

Proof. (Theorem 3.6.9) Fix $\omega \in \Omega$. Consider the derivative of $g(x_n, z_n)$ with respect to the parameter θ:

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

Since $g \in Li(p)$ we have that

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

and by Lemma 3.6.10 we have

$$\|\frac{\partial z_n}{\partial \theta}\| < K,$$

for a fix $K > 0$ (independent of the sequence (x_n)). Here we have used that for a fix ω the sequence (x_n) is fixed. Thus we have for a fix ω

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

Taking the expectation of both sides and using Theorem 3.6.4 we get the proof. ■

We conclude this section with the following theorem.

Theorem 3.6.11 *Consider a process $U_n = (X_n, Z_n)$ defined by (3.7), where f is an exponentially stable mapping and X_n is a Markov chain satisfying the Doeblin condition. Assume that Conditions 3.6.2, 3.6.3, 3.6.5 and 3.6.6 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 3.6.1 holds for the algorithm (3.48).

We apply Theorem 3.6.11 for Hidden Markov Models in Chapter 5

Chapter 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. Indeed, $Q^* > 0$ implies the Doeblin condition for the Markov chain (X_n) and if the Doeblin condition is satisfied for (X_n) then it is also satisfied for the pair (X_n, Y_n). Note that if the Doeblin condition is satisfied for a Markov chain then an invariant distribution exists for the process, see Proposition 3.1.4.

Let the invariant distribution of (X_n) be ν and the invariant distribution of (X_n, Y_n) be π. Note that (X_n, Y_n) corresponds to (X_n) and (p_n) corresponds to (Z_n) in Chapter 3.

$$\pi(\{i\}, dy) = \nu_i b^{*i}(y)\lambda(dy). \tag{4.1}$$

The logarithm of the likelihood function is

$$\sum_{k=1}^{n-1} \log p(y_k|y_{k-1}, \dots y_0, \theta) + \log p(y_0, \theta), \tag{4.2}$$

where D is a domain and $\theta \in D$ parameterizes the transition matrix Q and the conditional read-out densities $b^i(y)$. Usually the entries of Q are included in θ. The k-th term in (4.2) for $k \geq 1$ can be written as

$$\log \sum_{i=1}^{N} b^i(y_k, \theta) P(i|y_{k-1}, \dots, y_0, \theta) = \log \sum_{i=1}^{N} b^i(y_k, \theta) p_k^i(\theta).$$

Now define g as

$$g(y, p) = \log \sum_{i=1}^{N} b^i(y) p^i, \tag{4.3}$$

then we have

$$\log p(y_n, \dots, y_0, \theta) = \sum_{k=1}^{n} g(y_k, p_k) + \log p(y_0, \theta). \tag{4.4}$$

Although the problem is thought of as a parametric one, to simplify the notations we will drop the parameter θ in this chapter. Instead, the true value of the corresponding unknown quantity is indicated by $*$ and the running value is denoted by letters without $*$.

The parameter dependence will be used from Chapter 6 on.

4.1 Estimation of Hidden Markov Models

A central question in estimation problems is proving the ergodic theorem for (2.9), see Chapter 2, which is equivalent to the existence of the limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} g(y_k, p_k). \tag{4.5}$$

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

With the notation $p_n^i = P(X_n = i|Y_{n-1}, \dots, Y_0)$ 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 4.1.1 *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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{4.6}$$

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

$$\int |\log b^j(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{4.7}$$

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 3.4.3. The exponential stability of f follows from Proposition 2.1.3. As p_n is a probability vector Condition 3.3.5 is trivially satisfied.

We prove that Condition 3.4.1 is satisfied. Let $[x]_- = \max\{-x, 0\}$ and $[x]_+ = \max\{x, 0\}$. On one hand

$$\sum_{j=1}^{N} b^j(y)p^j \geq \min_i b^i(y),$$

leads to

$$[\log \sum_{j=1}^{N} b^j(y)p^j]_- \leq [\log \min_i b^i(y)]_-,$$

or

$$[g(y,p)]_- \leq \max_i [\log b^i(y)]_- \leq \max_i |\log b^i(y)|. \tag{4.8}$$

On the other hand the inequality

$$\sum_{j=1}^{N} b^j(y)p^j \leq \max_i b^i(y),$$

leads to

$$[\log \sum_{j=1}^{N} b^j(y)p^j]_+ \leq [\log \max_i b^i(y)]_+,$$

or

$$[g(y,p)]_+ \leq \max_i [\log b^i(y)]_+ \leq \max_i |\log b^i(y)|. \tag{4.9}$$

Since the right hand sides in (4.8) and (4.9) are independent of p we get

$$\sup_p |g(y,p)| \leq \max_i |\log b^i(y)|. \tag{4.10}$$

Combining (4.7) and (4.10) we get that for all $i \in \mathcal{X}$

$$\int \left(\sup_p |g(y,p)|^q \right) b^{*i}(y)\lambda(dy) < \infty. \tag{4.11}$$

Since

$$\int \sup_p |g(y,p)|^q d\pi = \sum_{i=1}^{N} \nu_i \int \left(\sup_p |g(y,p)|^q \right) b^{*i}(y)\lambda(dy), \tag{4.12}$$

the finiteness of the left hand side follows.

Now, only Condition 3.3.9 remained to be checked, i.e. that $g(y,p) = \log \sum_i b^i(y)p^i$ is Lipschitz-continuous in p with Lipschitz constant independent of y. For an arbitrary fixed $y \in \mathcal{Y}$ we have

$$\| \frac{\partial g(y,p)}{\partial p} \| = \| \frac{1}{\sum_{j=1}^{N} b^j(y)p^j} (b^1(y), \dots b^N(y))^T \| \leq \tag{4.13}$$

$$\frac{\sqrt{N} \max_i b^i(y)}{\sum_{j=1}^{N} b^j(y)p^j} \leq \sqrt{N} \max_i \frac{1}{p^i} = \sqrt{N} (\min_i p^i)^{-1}. \tag{4.14}$$

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

$$\varepsilon = \min_{i,j} q_{ij} > 0. \tag{4.15}$$

Due to the Baum-equation (2.3) 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 (4.15)

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

Thus

$$p_{n+1} \geq \frac{\varepsilon \mathbf{1}\mathbf{1}^T B(y_n)p_n}{\mathbf{1}^T B(y_n)p_n} = \varepsilon \mathbf{1} \tag{4.16}$$

and we get

$$\|\frac{\partial g(y,p)}{\partial p}\| \leq \frac{\sqrt{N}}{\varepsilon}. \tag{4.17}$$

Hence the function $g(y,p)$ is Lipschitz continuous and thus Theorem 3.4.3 implies that $g(Y_n, p_n)$ is an L-mixing process. ■

Remark 4.1.2 *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, (4.6) 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 (4.5) consider the following lemma.

Lemma 4.1.3 *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. 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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{4.18}$$

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

$$\int |\log b^j(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{4.19}$$

Then the limit

$$\lim_{n\to\infty} Eg(Y_n, p_n)$$

exists.

Proof. Let us go back to the proof of Lemma 3.4.6. Identify (X_n, Y_n) with (X_n) and (p_n) with (Z_n) in Lemma 3.4.6. Furthermore let us identify the initialization of the true process (X_n, Y_n) with η and the initialization of the predictive filter with ξ. By the proof of Lemma 3.4.6 we have that (X_n, Y_n, p_n) converges in law to the stationary distribution. Thus it is enough to prove that the sequence $g(Y_n, p_n)$ is uniformly bounded in L_q $(q > 1)$ norm.

Using the fact that

$$\min_j b^j(y_n) \le \mathbf{b}^T(y_n)p_n \le \max_j b^j(y_n),$$

we have that

$$g(Y_n, p_n) \le \max_j |\log b^j(Y_n)|.$$

Let us denote the distribution of (X_n, Y_n) by π_n. Considering condition (4.18) and Lemma 3.3.6 we have that

$$E|\log b^j(Y_n)|^q \le K \max_i \int |\log b^j(y)|^q b^{*i}(y)\lambda(dy).$$

Thus condition (4.19) implies the uniform boundedness of $g(Y_n, p_n)$ in L_q norm. ■

Theorem 4.1.4 *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.*

$$\frac{d\pi_0}{d\pi} \le K. \tag{4.20}$$

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

$$\int |\log b^j(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{4.21}$$

Then the limit

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

exists almost surely.

Proof. Under the conditions of Theorem 4.1.1 $g(Y_n, p_n)$ is an L-mixing process. Normalizing this process we have that

$$g(Y_n, p_n) - Eg(Y_n, p_n)$$

is also L-mixing. According to Theorem 2.3.5 the law of large numbers is valid for this process. Combining this with the results of Lemma 4.1.3, we have that

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

also exists almost surely. ■

Consider now a finite state-finite read-out HMM. This case follows from Theorem 4.1.1, but the integrability condition (4.7) is simplified due to the discrete measure.

Theorem 4.1.5 *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, [40]. For easier reference we restate the results of [40] collecting the relevant conditions.

Proposition 4.1.6 *(Legland-Mevel 2000, [40]) 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 $i \in \mathcal{X}$

$$\int \frac{\max_{j\in\mathcal{X}} b^j(y)}{\min_{j\in\mathcal{X}} b^j(y)} b^{*i}(y)\lambda(dy) < \infty, \tag{4.22}$$

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

$$\int |\log b^j(y)| b^{*i}(y)\lambda(dy) < \infty. \tag{4.23}$$

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

Geometric ergodicity also implies the existence of limit in (4.5).

Remark 4.1.7 *Inequality (4.22) is a Lipschitz condition in the mean in the following sense. Due to (4.13) 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 (4.22) states that the Lipschitz constant $L(y)$ is bounded in average.

Now we demonstrate that our result applies in certain cases where Proposition 4.1.6 does not.

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 $\sigma_1 \le \cdots \le \sigma_N$. Denote the true parameter by (m_i^*, σ_i^*). Since $\log b^i(y)$ is quadratic in y, (4.7) is satisfied

as all moments of the normal distribution exist. Hence Theorem 4.1.4 is applicable, and the limit of the log-likelihood function (4.5) exists.

On the other hand, Condition (4.22) of Proposition 4.1.6 may not be satisfied if $\sigma_1 < \sigma_N$. Indeed, for large y-s the integrand of (4.22) 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, and this expression 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

$$(\sigma_i^*)^2 > \frac{(\sigma_1\sigma_N)^2}{(\sigma_N)^2 - (\sigma_1)^2}. \tag{4.24}$$

4.2 Extension to general state space

In Section 4.1 we investigated the case when the state space $\mathcal{X}$ is finite. We consider now 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 (2.1), 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 (2.3), 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 will use the following notation: for any measurable function f on $(K, \mathbf{B}(K), \mu)$ 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

$$\delta(y) = \frac{\operatorname{ess\,sup}_x b^x(y)}{\operatorname{ess\,inf}_x b^x(y)} \tag{4.25}$$

$$\epsilon = \frac{\operatorname{ess\,inf}_{x,x'} q(x,x')}{\operatorname{ess\,sup}_{x,x'} q(x,x')}. \tag{4.26}$$

The following statement, which is an adaptation of Proposition 2.1.3, shows the exponential memorylessness of the predictive density function, see [9].

Proposition 4.2.1 *(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*

$$\|p_n(p_0') - p_n(p_0'')\|_{L_1} \leq C(1-\epsilon)^n \|p_0' - p_0''\|_{L_1}. \tag{4.27}$$

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

$$g(y,p) = \log\left(\int_K b^x(y) p(x) \mu(dx)\right), \tag{4.28}$$

similarly to (4.3).

The following theorem is a modified version of Theorem 4.1.1.

Theorem 4.2.2 *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 (4.26). 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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{4.29}$$

Assume that for all $q \geq 1$

$$\operatorname*{ess\,sup}_x \int |\log \operatorname*{ess\,sup}_{x'} b^{x'}(y)|^q b^{*x}(y)\lambda(dy) < \infty. \tag{4.30}$$

and

$$\operatorname*{ess\,sup}_x \int |\delta(y)|^q \, b^{*x}(y)\lambda(dy) < \infty \tag{4.31}$$

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

Remark 4.2.3 *By Lemma 3.1.5 and Proposition 3.1.4 the Doeblin condition for the Markov chain implies the existence of an invariant distribution for the pair (X_n, Y_n).*

Proof. (Theorem 4.2.2) Identify (X_n, Y_n) with (X_n) and (p_n) with (Z_n) in Corollary 3.4.5. The exponential stability of f follows from Proposition 4.2.1. As p_n is a conditional density function Condition 3.3.5 is trivially satisfied.

We prove that Condition 3.4.1 is satisfied. For this we should check whether

$$\int \sup_p \left| \log \left(\int_K b^x(y)p(x)\mu(dx) \right) \right|^q b^{*x}(y)p^*(x)\lambda(dy)\mu(dx) < \infty \tag{4.32}$$

is true for all $q \geq 1$. Using that

$$\operatorname*{ess\,inf}_{x'} b^{x'}(y) < \int_K b^x(y)p(x)\mu(dx) < \operatorname*{ess\,sup}_{x'} b^{x'}(y),$$

it is enough to show that both

$$\int \left|\log\left(\operatorname*{ess\,sup}_{x'} b^{x'}(y)\right)\right|^q b^{*x}(y)p^*(x)\lambda(dy)\mu(dx)$$

and

$$\int \left|\log\left(\operatorname*{ess\,inf}_{x'} b^{x'}(y)\right)\right|^q b^{*x}(y)p^*(x)\lambda(dy)\mu(dx)$$

are finite.

$$\int \left|\log\left(\operatorname*{ess\,sup}_{x'} b^{x'}(y)\right)\right|^q b^{*x}(y)p^*(x)\lambda(dy)\mu(dx) <$$

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

by condition (4.30) and using the definition of $\delta(y)$ in (4.25) and the fact that $|a-b|^q \le 2^q(|a|^q + |b|^q)$ we have

$$\int \left|\log\left(\operatorname*{ess\,inf}_{x'} b^{x'}(y)\right)\right|^q b^{*x}(y)p^*(x)\lambda(dy)\mu(dx) <$$

$$\int 2^q \left(\left|\log(\operatorname*{ess\,sup}_{x'} b^{x'}(y))\right|^q + |\log\delta(y)|^q\right) b^{*x}(y)p^*(x)\lambda(dy)\mu(dx) <$$

$$2^q \operatorname*{ess\,sup}_{x} \int |\log \operatorname*{ess\,sup}_{x'} b^{x'}(y)|^q b^{*x}(y)\lambda(dy) +$$

$$2^q \operatorname*{ess\,sup}_{x} \int |\log\delta(y)|^q\, b^{*x}(y)\lambda(dy) < \infty$$

by condition (4.30) and (4.31). For the second term we have used that $\delta(y) \ge 1$, thus $|\log\delta(y)|^q \le |\delta(y)|^q$. Thus we have that (4.32) holds indeed.

To finish the proof we have to check the Lipschitz continuity of $g(y,p)$ in p for all y (see Condition 3.3.12). Consider the definition of g in (4.28)

$$|g(y,p_1) - g(y,p_2)| =$$

$$\left|\log\left(\int_K b^x(y)p_1(x)\mu(dx)\right) - \log\left(\int_K b^x(y)p_2(x)\mu(dx)\right)\right| =$$

$$\left|\log\frac{\int_K b^x(y)p_1(x)\mu(dx)}{\int_K b^x(y)p_2(x)\mu(dx)}\right| \tag{4.33}$$

As $|\log A| = |\log 1/A|$ for $A > 0$ assume that the numerator is greater then the denominator.

Using the fact that $\log x \leq x - 1$ for $x > 1$ we can estimate (4.33) from above by

$$\left| \frac{\left(\int_K b^x(y)p_1(x)\mu(dx)\right) - \left(\int_K b^x(y)p_2(x)\mu(dx)\right)}{\int_K b^x(y)p_2(x)\mu(dx)} \right| \leq$$

$$\frac{\operatorname{ess\,sup}_{x'} b^{x'}(y) \left(\int_K |p_1(x) - p_2(x)|\mu(dx)\right)}{\operatorname{ess\,inf}_{x'} b^{x'}(y)} \leq$$

$$\delta(y)\|p_1 - p_2\|_{L_1},$$

i.e. the function $g(y,p)$ is Lipschitz-continuous in p for all y and all the moments of the Lipshitz constant exists by (4.31).

Thus the conditions of Corollary 3.4.5 are satisfied and the process $g(Y_n, p_n)$ is L-mixing.

■

Let us turn to the analyze of the asymptotic properties of (4.5). The following lemma is similar to Lemma 4.1.3.

Lemma 4.2.4 *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 (4.26). 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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{4.34}$$

Assume that for all $q \geq 1$

$$\operatorname{ess\,sup}_x \int |\log \operatorname{ess\,sup}_{x'} b^{x'}(y)|^q b^{*x}(y)\lambda(dy) < \infty \tag{4.35}$$

and

$$\operatorname{ess\,sup}_x \int |\log \delta(y)|^q \, b^{*x}(y)\lambda(dy) < \infty \tag{4.36}$$

Then the limit

$$\lim_{n\to\infty} Eg(Y_n, p_n)$$

exists.

Proof. We follow the arguments of Lemma 4.1.3. Identify (X_n, Y_n) with (X_n) and (p_n) with (Z_n) in Lemma 3.4.6. Furthermore let us identify the initialization of the true process (X_n, Y_n) with η and the initialization of the predictive filter with ξ. By the proof of Lemma 3.4.6 we have that (X_n, Y_n, p_n) converges in law to the stationary distribution. Thus it is enough to prove that the sequence $g(Y_n, p_n)$ is uniformly bounded in L_q $(q > 1)$ norm.

Note that

$$\operatorname{ess\,inf}_x b^x(y_n) \le \int b^x(y_n) p_n(x)\mu(dx) \le \operatorname{ess\,sup}_x b^x(y_n)$$

and

$$|\log \operatorname{ess\,inf}_x b^x(y_n)| \le |\log \operatorname{ess\,sup}_x b^x(y_n)| + |\log \delta(y_n)|.$$

Let us denote the distribution of (X_n, Y_n) by π_n. Considering condition (4.34) and Lemma 3.3.6 we have that

$$E|\log \operatorname{ess\,sup}_x b^x(y_n)|^q \le K \operatorname{ess\,sup}_x \int |\log \operatorname{ess\,sup}_{x'} b^{x'}(y)|^q b^{*x}(y)\lambda(dy).$$

and

$$E|\log \delta(y_n)|^q \le K \operatorname{ess\,sup}_x \int |\log \delta(y)|^q b^{*x}(y)\lambda(dy)$$

Thus conditions (4.35) and (4.36) imply the uniform boundedness of $g(Y_n, p_n)$ in L_q norm.

■

Theorem 4.2.5 *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 (4.26). Furthermore assume that the Doeblin condition is satisfied for the Markov chain (X_n). Assume that for all $q \ge 1$*

$$\operatorname{ess\,sup}_x \int |\log \operatorname{ess\,sup}_{x'} b^{x'}(y)|^q b^{*x}(y)\lambda(dy) < \infty. \tag{4.37}$$

and

$$\operatorname{ess\,sup}_x \int |\delta(y)|^q \, b^{*x}(y)\lambda(dy) < \infty \tag{4.38}$$

Then the limit

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

exists almost surely.

At the end of this section we compare our results with those of Douc and Matias, [9].

Proposition 4.2.6 *(Douc-Matias 2001, [9]) Consider a Hidden Markov Process (X_n, Y_n), where the state space $\mathcal{X}$ is compact and the observation space $\mathcal{Y}$ is continuous. Assume that $0 < \epsilon < 1$,*

$$\operatorname{ess\,sup}_x \int \operatorname{ess\,sup}_{x'} |\log b^{x'}(y)|^q b^{*x}(y)\lambda(dy) < \infty. \tag{4.39}$$

for some $q > 0$ and

$$\operatorname{ess\,sup}_x \int |\delta(y)| \, b^{*x}(y)\lambda(dy) < \infty \tag{4.40}$$

Then the limit $\frac{1}{n} \sum_{k=1}^{n} g(Y_k, p_k)$ exists almost surely.

The proof is based on the geometric ergodicity of the process $g(Y_n, p_n)$.

Chapter 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

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

In this case θ is often the parameter of the model parameterizing the transition matrix Q and the conditional read-out probabilities $b^i(y)$. Usually the entries of Q are included in θ.

Consider the parameter-dependent Baum-equation

$$\mathbf{p}_{n+1}(\theta) = \frac{Q^T(\theta)B(y_n,\theta)\mathbf{p}_n(\theta)}{\mathbf{b}(y_n,\theta)^T\mathbf{p}_n(\theta)} = \Phi_1(y_n,\mathbf{p}_n,\theta), \tag{5.1}$$

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

$$W_{n+1} = Q^T\left(I - \frac{B(y_n)\mathbf{p}_n\mathbf{e}^T}{\mathbf{b}^T(y_n)\mathbf{p}_n}\right)\frac{B(y_n)W_n}{\mathbf{b}^T(y_n)\mathbf{p}_n} + F, \tag{5.2}$$

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

$$W_{n+1} = \Phi_2(y_n, \mathbf{p}_n, W_n, \theta). \tag{5.3}$$

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

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

see (4.3), we get

$$\varphi_n = \frac{\beta(y_n)\mathbf{p}_n + W_n \mathbf{b}(y_n)}{\mathbf{b}(y_n)^T \mathbf{p}_n}. \tag{5.4}$$

Let

$$H(\theta, u) = H(\theta, x, y, \mathbf{p}, W) = \frac{\beta(y, \theta)\mathbf{p} + W\mathbf{b}(y, \theta)}{\mathbf{b}(y, \theta)^T \mathbf{p}}, \tag{5.5}$$

and consider the following adaptive algorithm.

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+1} H(\overline{\theta}_n, x_n, y_n, \overline{\mathbf{p}}_n, \overline{W}_n), \tag{5.6}$$

$$\overline{\mathbf{p}}_{n+1} = \Phi_1(y_n, \overline{\mathbf{p}}_n, \overline{\theta}_n), \tag{5.7}$$

$$\overline{W}_{n+1} = \Phi_2(y_n, \overline{\mathbf{p}}_n, \overline{W}_n, \overline{\theta}_n). \tag{5.8}$$

For the convergence of this algorithm we use the approach of Benveniste, Metivier and Priouret, see Theorem 3.6.1 and [6]. In the following we verify the conditions of Theorem 3.6.11

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.

Theorem 5.0.7 *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**) of Section 3.6.1 are satisfied.*

Proof. Identify X_n of Theorem 3.6.11 with (X_n, Y_n) and Z_n of Theorem 3.6.11 with (p_n, W_n). Then the mapping f of Theorem 3.6.11 is identified with the pair (Φ_1, Φ_2). Exponentially stability of the pair (Φ_1, Φ_2) is implied by Proposition 2.1.3 and Lemma 3.6.10. The Doeblin condition and Condition 3.6.2 is satisfied for the process (X_n, Y_n) since $Q^* > 0$ and $b^{*x}(y) > 0$. Conditions 3.6.3 and 3.6.5 are trivially satisfied for finite state space and finite read-out space if $Q(\theta) > 0$ and $b^x(y) > 0$ for all x, y. Condition 3.6.6 is automatically satisfied for finite systems. ■

Let us investigate assumption (**A5**).

Theorem 5.0.8 *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 assumption (**A5**) of Section 3.6.1 is satisfied.*

Proof. Noting that by the condition $b^x(y, \theta) > 0$ we have that $b^x(y, \theta) > \epsilon$, since the read-out space is finite and D is a compact domain. Thus we have

$$b^T(y, \theta)p > \epsilon,$$

and using the definition of H, see (5.5), assumption (**A5**) follows by the smoothness of $b(y, \theta)$ and $Q(\theta)$. ■

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

$$h(\theta) = \lim_{n\to\infty} E \frac{\partial}{\partial\theta} \log p(Y_n | Y_{n-1}, \dots Y_0, \theta) \tag{5.9}$$

This limit exists, see Theorem 6.2.3, and assume that the following identifiability condition is satisfied, see also Condition 6.3.2:

Condition 5.0.9 *The equation*

$$h(\theta) = 0$$

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

Note that $h(\theta)$ is identified with $W_\theta(\theta, \theta^*)$ in (6.26).

Condition 5.0.9 implies assumption (**A6**) in a small domain. Thus we conclude with the following theorem as an application of Theorem 3.6.1.

Theorem 5.0.10 *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 5.0.9. Then the algorithm defined by (5.6), (5.7), (5.8) converges to the true value* θ^* *with probability arbitrary close to 1.*

Chapter 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 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. We only prove our statement for the first derivative, i.e. when $k = 1$, for $k = 2, 3$ the proofs are very similar. Throughout this section we will assume that $Q(\theta)$ and $b^x(y, \theta)$ are smooth functions in the parameter $\theta \in G$.

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

$$\delta(y) = \frac{\max\limits_x b^x(y)}{\min\limits_x b^x(y)} \tag{6.1}$$

and

$$\delta'(y) = \frac{\max\limits_x \|\partial b^x(y)/\partial \theta\|}{\min\limits_x b^x(y)} \tag{6.2}$$

Theorem 6.2.1 *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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{6.3}$$

Assume that

$$\int |\delta(y)|^q b^{*i}(y) \lambda(dy) < \infty, \tag{6.4}$$

$$\int |\delta(y)'|^q b^{*i}(y) \lambda(dy) < \infty. \tag{6.5}$$

Then

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

is L-mixing.

Proof. To simplify the notations we drop the dependence on the parameter θ. Using the notations of Chapter 5 we have

$$\frac{\partial}{\partial \theta} \log p(y_n | p_{n-1}, \ldots p_0) = \frac{\partial}{\partial \theta} \log b^T(y_n) p_n = \frac{\beta(y_n) p_n + W_n b(y_n)}{b^T(y_n) p_n},$$

where $W_n = \frac{\partial p_n}{\partial \theta}$ and $\beta(y_n) = \frac{\partial b(y_n)}{\partial \theta}$, see (5.4).

Identify (X_n, Y_n) with (X_n) and (p_n, W_n) with (Z_n) in Theorem 3.5.4. According to (5.1) and (5.3) let f be (Φ_1, Φ_2). Finally, let us define g as

$$g(x_n, y_n, p_n, W_n) = \frac{\beta(y_n)p_n + W_n b(y_n)}{b^T(y_n)p_n}.$$

Thus we should check the conditions of Theorem 3.5.4. The exponential stability of f follows from Proposition 2.1.3 and Lemma 3.6.10.

We prove Condition 3.5.2. For this consider the following lemma.

Lemma 6.2.2

$$\|\frac{\beta(y)p + Wb(y)}{b^T(y)p}\| \leq \delta(y)\|W\| + \delta'(y)$$

Proof. To simplify the expressions here we give the proof when $\dim \Theta = 1$:

$$\|\frac{\beta(y)p}{b^T(y)p}\| \leq \frac{\max_x(\partial b^x(y)/\partial\theta)}{\min_x b^x(y)} = \delta'(y), \tag{6.6}$$

since p is a probability vector. On the other hand

$$\left\|\frac{Wb(y)}{b^T(y)p}\right\| \leq \frac{\max_x b^x(y)\|W\|}{\min_x b^x(y)} = \delta(y)\|W\|. \tag{6.7}$$

■

Lemma 6.2.2 and conditions (6.4), (6.5) imply Condition 3.5.2.

Let us turn to Condition 3.5.1. To prove that this condition is satisfied we should consider the difference

$$\left\|\frac{\beta(y)p_1 + W_1 b(y)}{b^T(y)p_1} - \frac{\beta(y)p_2 + W_2 b(y)}{b^T(y)p_2}\right\|,$$

where p_1, p_2 are probability vectors and W_1, W_2 are matrices. To simplify the expressions here we consider the case when $\dim \Theta = 1$. In this case $\beta(y)$ and W are row vectors. We have

$$\left\|\frac{\beta(y)p_1 + W_1 b(y)}{b^T(y)p_1} - \frac{\beta(y)p_2 + W_2 b(y)}{b^T(y)p_2}\right\| \leq$$

$$\left\| \frac{\beta(y)p_1}{b^T(y)p_1} - \frac{\beta(y)p_2}{b^T(y)p_2} \right\| + \left\| \frac{W_1 b(y)}{b^T(y)p_1} - \frac{W_2 b(y)}{b^T(y)p_2} \right\|.$$

Consider the first term:

$$\left\| \frac{\beta(y)p_1}{b^T(y)p_1} - \frac{\beta(y)p_2}{b^T(y)p_2} \right\| \leq \left\| \frac{\beta(y)(p_1 - p_2)}{b^T p_1} + \frac{\beta(y)p_2(b^T p_2 - b^T p_1)}{b^T p_1 b^T p_2} \right\| \leq$$

$$\delta'(y)\|p_1 - p_2\| + \delta(y)\delta'(y)\|p_1 - p_2\|.$$

Let us consider the second term.

$$\left\| \frac{W_1 b(y)}{b^T(y)p_1} - \frac{W_2 b(y)}{b^T(y)p_2} \right\| = \left\| \frac{b^T(y)(W_1 - W_2)}{b^T p_1} - \frac{b^T(y)(p_1 - p_2)}{b^T(y)p_1} \frac{b^T(y)W_2}{b^T(y)p_2} \right\| \leq$$

$$\delta(y)\|W_1 - W_2\| + \delta(y)^2\|p_1 - p_2\|\|W_2\|,$$

by (6.7). Thus we have that

$$\left\| \frac{\beta(y)p_1 + W_1 b(y)}{b^T(y)p_1} - \frac{\beta(y)p_2 + W_2 b(y)}{b^T(y)p_2} \right\| \leq (\|p_1 - p_2\| + \|W_1 - W_2\|) *$$

$$(\delta(y) + \delta^2(y) + \delta'(y) + \delta(y)\delta'(y))(\|p_1\| + \|p_2\| + \|W_1\| + \|W_2\|)$$

and by (6.4) and (6.5) Condition 3.5.1 is satisfied.

To finish the proof we should check that Condition 3.3.5 is valid. Since p_1 is a probability vector, it is enough to prove the validity of this condition for W_1. Consider

$$W_1 = Q^T \left(I - \frac{B(y)\mathbf{p}\mathbf{e}^T}{\mathbf{b}^T(y)\mathbf{p}} \right) \frac{B(y)W}{\mathbf{b}^T(y)\mathbf{p}} + F, \tag{6.8}$$

where

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

see (5.2). Here p, W are arbitrary initializations. Similar to the previous proofs we have

$$\|W_1\| \leq \|Q\|\delta(y)(\|W\| + 1) + \delta(y)\|Q_\theta\| + \|Q\|\delta'(y)(1 + \delta(y)). \tag{6.9}$$

Due to conditions (6.4) and (6.5) the moments of W_1 exist. ■

In applications we need that the limit of the expectation of the derivative process exists, see (5.9) or (6.26).

Theorem 6.2.3 *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 $Q(\theta)$ 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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{6.10}$$

Assume that

$$\int |\delta(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{6.11}$$

and

$$\int |\delta(y)'|^q b^{*i}(y)\lambda(dy) < \infty. \tag{6.12}$$

Then the limit

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

exists.

Proof. We follow the arguments of Lemma 4.1.3. Identify (X_n, Y_n) with (X_n) and (p_n, W_n) with (Z_n) in Lemma 3.4.6. Note that by Lemma 3.6.10 the process (p_n, W_n) is exponentially stable. Furthermore let us identify the initialization of the true process (X_n, Y_n) with η and the initialization of the process (p_n, W_n) with ξ. By the proof of Lemma 3.4.6 we have that (X_n, Y_n, p_n, W_n) converges in law to the stationary distribution. Thus it is enough to prove that

$$\frac{\beta(Y_n)p_n + W_n b(Y_n)}{b^T(Y_n)p_n}$$

is uniformly bounded in L_q $(q > 1)$ norm.

From (6.6) and (6.7) we have that

$$\frac{\beta(Y_n)p_n + W_n b(Y_n)}{b^T(Y_n)p_n} \leq \delta'(Y_n) + \delta(Y_n)\|W_n\|.$$

From Lemma 3.3.8 we have the M-boundedness of W_n (the conditions of the lemma are satisfied, see (6.10) and (6.9)) with an arbitrary initialization.

Furthermore by condition (6.10) and Lemma 3.3.6 we have that

$$E|\delta'(Y_n)|^q \leq K \int |\delta(y)|^q b^{*i}(y)\lambda(dy)$$

and

$$E|\delta'(Y_n)|^q \leq K \int |\delta(y)'|^q b^{*i}(y)\lambda(dy).$$

Thus using the Hölder inequality and conditions (6.11), (6.12) we have the uniform boundedness of $\frac{\beta(Y_n)p_n + W_n b(Y_n)}{b^T(Y_n)p_n}$ in L_q norm. ■

Let us turn to the second and the third derivatives. Define

$$\delta_2(y) = \frac{\max\limits_x b^x(y)}{(\min\limits_x b^x(y))^2}$$

$$\delta_2'(y) = \frac{\max\limits_x \|\partial b^x(y)/\partial\theta\|}{(\min\limits_x b^x(y))^2}$$

$$\delta_2''(y) = \frac{\max\limits_x \|\partial^2 b^x(y)/\partial\theta^2\|}{(\min\limits_x b^x(y))^2}$$

Theorem 6.2.4 *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 two times 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.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{6.13}$$

Assume that

$$\int |\delta_2(y)|^q b^{*i}(y)\lambda(dy) < \infty, \tag{6.14}$$

$$\int |\delta_2'(y)|^q b^{*i}(y)\lambda(dy) < \infty, \tag{6.15}$$

$$\int |\delta_2''(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{6.16}$$

Then

$$\frac{\partial^2}{\partial\theta^2}\log p(y_n|p_{n-1},\dots p_0,\theta)$$

is L-mixing and the limit

$$\lim_{n\to\infty} E\frac{\partial^2}{\partial\theta^2}\log p(y_n|y_{n-1},\dots y_0,\theta)$$

exists.

Define

$$\delta_3(y) = \frac{\max\limits_x b^x(y)}{(\min\limits_x b^x(y))^4}$$

$$\delta_3'(y) = \frac{\max\limits_x \|\partial b^x(y)/\partial\theta\|}{(\min\limits_x b^x(y))^4}$$

$$\delta_3''(y) = \frac{\max\limits_x \|\partial^2 b^x(y)/\partial\theta^2\|}{(\min\limits_x b^x(y))^4}$$

$$\delta_3'''(y) = \frac{\max\limits_x \|\partial^3 b^x(y)/\partial\theta^3\|}{(\min\limits_x b^x(y))^4}$$

Theorem 6.2.5 *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 three times 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.*

$$\frac{d\pi_0}{d\pi} \le K. \tag{6.17}$$

Assume that

$$\int |\delta_3(y)|^q b^{*i}(y)\lambda(dy) < \infty, \tag{6.18}$$

$$\int |\delta_3'(y)|^q b^{*i}(y)\lambda(dy) < \infty, \tag{6.19}$$

$$\int |\delta_3''(y)|^q b^{*i}(y)\lambda(dy) < \infty, \tag{6.20}$$

$$\int |\delta_3'''(y)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{6.21}$$

Then

$$\frac{\partial^3}{\partial\theta^3}\log p(y_n|p_{n-1},\dots p_0,\theta)$$

is L-mixing and the limit

$$\lim_{n\to\infty} E\frac{\partial^3}{\partial\theta^3}\log p(y_n|y_{n-1},\dots y_0,\theta)$$

exists.

6.3 Characterization theorem for the error

In this section 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 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 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.

$$\frac{d\pi_0}{d\pi} \le K. \tag{6.22}$$

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

$$\int |\log b^j(y,\theta)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{6.23}$$

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

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

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

$$\frac{\partial}{\partial\theta} L_N(\theta, \theta^*) = L_{\theta N}(\theta, \theta^*) = 0 \tag{6.24}$$

More exactly $\widehat{\theta}_N$ is a random vector such that $\widehat{\theta}_N \in D$ for all ω and if the equation (6.24) has a unique solution in D, then $\widehat{\theta}_N$ is equal to this solution. By the measurable selection theorem such a random variable does exist.

Let us introduce the asymptotic cost function

$$W(\theta, \theta^*) = \lim_{n\to\infty} E_{\theta^*} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta). \tag{6.25}$$

In Lemma 4.1.3 we have proved that this limit exists for all $\theta \in D$. Assume that the function $W(\theta, \theta^*)$ is smooth in the interior of D, i.e. the third derivative exists. Under the conditions of Theorem 6.2.3 and 6.2.4 we have

$$W_\theta(\theta, \theta^*) = \lim_{n\to\infty} E_{\theta^*} \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta), \tag{6.26}$$

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

Consider the following identifiability condition:

Condition 6.3.2 *The equation*

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

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

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

Theorem 6.3.3 *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 4.1.1, 6.2.1, 6.2.4, 6.2.5 are satisfied. Let $\hat{\theta}_N$ be the ML estimate of θ^*. Furthermore assume that the identifiability condition 6.3.2 is satisfied. Then*

$$\hat{\theta}_N - \theta^* = -(I^*)^{-1} \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \dots, Y_0, \theta^*) + O_M(N^{-1}), \quad (6.27)$$

where I^* is the Fisher-information matrix.

For the proof we need several lemmas.

Lemma 6.3.4 *The process*

$$u_n(\theta, \theta^*) = \frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \dots Y_0, \theta) - E \frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \dots Y_0, \theta)$$

is uniformly L-mixing in (θ, θ^*).

Proof. For a fix θ the process $u_n(\theta, \theta^*)$ is L-mixing due to Theorem 6.2.1 and Theorem 6.2.3. Considering that in Proposition 2.1.3 in the right hand side of (2.4) C depends on the parameter θ continuously and by the smoothness conditions on $Q(\theta)$ and $b^i(\theta)$ we have that in the proof of the L-mixing property in Theorem 3.5.4 the left hand side of (3.47) is a continuous function of θ. Since D is a compact domain this implies the uniform L-mixing property. ■

Similarly to Lemma 6.3.4 Theorem 6.2.4 and 6.2.5 imply the following lemmas.

Lemma 6.3.5 *The process*

$$u_{\theta n}(\theta, \theta^*) = \frac{\partial^2}{\partial \theta^2} \log p(Y_n | Y_{n-1}, \dots Y_0, \theta) - E \frac{\partial^2}{\partial \theta^2} \log p(Y_n | Y_{n-1}, \dots Y_0, \theta)$$

is uniformly L-mixing in (θ, θ^*).

Lemma 6.3.6 *The process*

$$u_{\theta n}(\theta,\theta^*)=\frac{\partial^3}{\partial\theta^3}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)-E\frac{\partial^3}{\partial\theta^3}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)$$

is uniformly L-mixing in (θ,θ^*).

Lemma 6.3.7 *Assume* $W_\theta(\theta,\theta^*)=0$ *has a single solution* $\theta=\theta^*$ *in* D *(that is, assume the identifiability condition 6.3.2). Then for any* $d>0$ *the equation (6.24) has a unique solution in* D *such that it is also in the sphere* $\{\theta:\ |\theta-\theta^*|<d\}$ *with probability at least* $1-O(N^{-s})$ *for any* $s>0$ *where the constant in the error term* $O(N^{-s})=CN^{-s}$ *depends only on* d *and* s.

Proof. We show first that the probability to have a solution outside the sphere $\{\theta:\ |\theta-\theta^*|<d\}$ is less than $O(N^{-s})$ with any $s>0$. Indeed, the equation $W_\theta(\theta,\theta^*)=0$ has a single solution $\theta=\theta^*$ in D, thus for any $d>0$ we have

$$d'=\inf\{|W_\theta(\theta,\theta^*)|:\theta\in D,\theta^*\in D^*,|\theta-\theta^*|\geq d\}>0$$

since $W_\theta(\theta,\theta^*)$ is continuous in (θ,θ^*) and $D\times D^*$ is compact. Therefore if a solution of (6.24) exists outside the sphere $\{\theta:\ |\theta-\theta^*|<d\}$ then we have for

$$\delta L_{\theta N}=\sup_{\theta\in D,\theta^*\in D^*}|\frac{1}{N}L_{\theta N}(\theta,\theta^*)-W_\theta(\theta,\theta^*)|$$

the inequality $\delta L_{\theta N}>d'$. Due to Lemma 6.3.4 and 6.3.5 the process

$$u_n(\theta,\theta^*)=\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)-E\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)$$

and the process

$$u_{\theta n}(\theta,\theta^*)=\frac{\partial^2}{\partial\theta^2}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)-E\frac{\partial^2}{\partial\theta^2}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)$$

are L-mixing processes uniformly in (θ,θ^*).

Since $Eu_n(\theta,\theta^*)=0$ Theorem 2.3.9 is applicable, i.e.

$$\sup_{\theta\in D,\theta^*\in D^*}\left|\frac{1}{N}L_{\theta N}(\theta,\theta^*)-\frac{1}{N}\sum_{n=1}^{N}E\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)\right|=O_M(N^{-1/2}).$$

Observe that

$$\delta_n = E\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta) - W_\theta(\theta,\theta^*) = O(\alpha^n) \tag{6.28}$$

with some $0 < \alpha < 1$. Indeed if the initial value of the predictive filter process is from a stationary distribution then $\delta_n = 0$. On the other hand the effects of nonstationary initial values decay exponentially, see Theorem 6.2.3 and Lemma 3.4.6. Thus we have

$$\delta L_{\theta N} = O_M(N^{-1/2}),$$

therefore

$$P(\delta L_{\theta N} > d') = O(N^{-s})$$

with any s by Markov's inequality.

Let us now consider the random variable

$$\delta L_{\theta\theta N} = \sup_{\theta\in D,\theta^*\in D_0}\left\|\frac{1}{N}L_{\theta\theta N}(\theta,\theta^*) - W_{\theta\theta}(\theta,\theta^*)\right\|.$$

By the same argument as above we have

$$\delta L_{\theta\theta N} = O_M(N^{-1/2}), \tag{6.29}$$

therefore

$$P(\delta L_{\theta\theta N} > d'') = O(N^{-s})$$

for any $d'' > 0$ and hence for the event

$$A_N = \{\omega :\ \delta L_{\theta N} < d', \delta L_{\theta\theta N} < d''\} \tag{6.30}$$

we have for N big enough

$$P(A_N) > 1 - O(N^{-s}) \tag{6.31}$$

with any $s > 0$. But on A_N the equation (6.24) has a unique solution whenever d' and d'' are sufficiently small. Indeed by Condition 6.3.2 the equation $W_\theta(\theta,\theta^*) = 0$ has a unique solution in D and hence the existence of a unique solution of (6.24) can easily be derived from the following version of the implicit function theorem.

Lemma 6.3.8 *Let $W_\theta(\theta), \delta W_\theta(\theta)$, $\theta \in D \subset \mathbb{R}^p$ be $\mathbb{R}^p$-valued continuously differentiable functions, let for some $\theta^* \in D_0 \subset D$, $W_\theta(\theta^*) = 0$, and let $W_{\theta\theta}(\theta^*)$ be non-singular. Then for any $d > 0$ there exists positive numbers d', d'' such that*

$$|\delta W_\theta(\theta)| < d' \quad \textit{and} \quad \|\delta W_{\theta\theta}(\theta)\| < d''$$

for all $\theta \in D_0$ implies that the equation $W_\theta(\theta) + \delta W_\theta(\theta) = 0$ has exactly one solution in a neighborhood of radius d of θ^*.

■

Lemma 6.3.9 *We have*

$$\hat{\theta}_N - \theta^* = O_M(N^{-1/2}).$$

Proof. Consider the Taylor-series expansion of $L_{\theta N}(\theta, \theta^*)$ around $\theta = \theta^*$ and evaluate the value of the function at $\theta = \hat{\theta}_N$. Then we have

$$L_{\theta N}(\hat{\theta}_N, \theta^*) = L_{\theta N}(\theta^*, \theta^*) + (\hat{\theta}_N - \theta^*) \int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda = 0 \tag{6.32}$$

First we prove that

$$L_{\theta N}(\theta^*, \theta^*) = O_M(N^{1/2}). \tag{6.33}$$

Note that $\frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \dots Y_0, \theta^*)$ is a martingal difference process. Indeed,

$$\int_{\mathcal{Y}} \left(\frac{\partial}{\partial\theta} \log p(y|y_{n-1}, \dots, y_0, \theta^*)\right) p(y|y_{n-1}, \dots y_0, \theta^*) dy =$$

$$\int_{\mathcal{Y}} \frac{\partial}{\partial\theta} p(y|y_{n-1}, \dots y_0, \theta^*) dy = \frac{\partial}{\partial\theta} \int_{\mathcal{Y}} p(y|y_{n-1}, \dots y_0, \theta^*) dy = 0. \tag{6.34}$$

Here we have used that $p(y|y_{n-1}, \dots y_0, \theta)$ is a density function and D is a compact domain, thus the uniform integrability condition for the class $p(y|y_{n-1}, \dots y_0, \theta)$ is satisfied.

For (6.33) we use the Burkholder's inequality for martingales, see Theorem 2.10 in [29]:

$$E^{1/q}|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*)|^q \le$$

$$CE^{1/q}\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*))^2\right)^{q/2}$$

Taking the square of both sides and using the triangle inequality for the $L_{q/2}$ norm of the right hand side we get

$$E^{2/q}|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*)|^q \le$$

$$C^2\frac{1}{N}\sum_{n=1}^{N}E^{2/q}|\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*)|^q.$$

M-boundedness of the process $\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*)$ follows from Theorem 6.2.1, thus we get (6.33).

Let us now investigate the integral. Since the function W is smooth we have for $0\le\lambda\le 1$ on the set A_N (defined in (6.30))

$$\|W_{\theta\theta}(\theta^*+\lambda(\hat\theta_N-\theta^*),\theta^*)-W_{\theta\theta}(\theta^*,\theta^*)\|<C|\hat\theta_N-\theta^*|<Cd \qquad (6.35)$$

Hence if d is sufficiently small then the positive definiteness of $W_{\theta\theta}(\theta^*,\theta^*)$ implies that

$$\int_0^1 W_{\theta\theta}\left((1-\lambda)\theta^*+\lambda\hat\theta_N,\theta^*\right)d\lambda > cI$$

with some positive c. Since on A_N

$$\left\|\frac{1}{N}\int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^*+\lambda\hat\theta_N,\theta^*\right)d\lambda-\int_0^1 W_{\theta\theta}\left((1-\lambda)\theta^*+\lambda\hat\theta_N,\theta^*\right)d\lambda\right\|<d''$$

it follows that if d'' is sufficiently small then denoting the minimal eigenvalue of a matrix by $\lambda_{\min}$ we have

$$\lambda_{\min}\left(\int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda\right) > c > 0,$$

i.e.

$$\left\|\left(\int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda\right)^{-1}\right\| < CN^{-1} \tag{6.36}$$

on A_N with a nonrandom constant C.

Considering equation (6.32) and that the first term is $O_M(N^{1/2})$ (see (6.33)) we have that

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = O_M(N^{-1/2}). \tag{6.37}$$

Using (6.31) and the fact that $|\hat{\theta}_N - \theta^*|$ is bounded we have

$$\chi_{A_N^c}(\hat{\theta}_N - \theta^*) = O_M(N^{-s}) \tag{6.38}$$

with any $s > 0$.

Combining (6.37) and (6.38) we get the lemma. ■

Proof. (Theorem 6.3.3) According to Lemma 6.3.9 the inequality (6.35) can be improved by $O_M(N^{-1/2})$. Thus we get after integration with respect to λ that

$$\left\|\int_0^1 W_{\theta\theta}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda - W_{\theta\theta}(\theta^*, \theta^*)\right\| = O_M(N^{-1/2})$$

On the other hand from (6.29) we have

$$\left\|\frac{1}{N}\int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda - \int_0^1 W_{\theta\theta}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda\right\| =$$

$$O_M(N^{-1/2}).$$

Hence we finally get

$$\left\| \frac{1}{N} \int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda - W_{\theta\theta}(\theta^*, \theta^*) \right\| = O_M(N^{-1/2})$$

Considering that on A_N (6.36) is satisfied and the fact that $W_{\theta\theta}(\theta^*, \theta^*) > 0$ we have

$$\left\| \left(\int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda \right)^{-1} - \frac{1}{N} W_{\theta\theta}^{-1}(\theta^*, \theta^*) \right\| = O_M(N^{-3/2}). \tag{6.39}$$

Consider (6.32) on the set A_N. We have

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = -\chi_{A_N} \left(\int_0^1 L_{\theta\theta N}\left((1-\lambda)\theta^* + \lambda\hat{\theta}_N, \theta^*\right) d\lambda \right)^{-1} L_{\theta N}(\theta^*, \theta^*).$$

Taking into account estimation (6.39) we get that

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = -\chi_{A_N} \left(\frac{1}{N} W_{\theta\theta}^{-1}(\theta^*, \theta^*) + O_M(N^{-3/2}) \right) L_{\theta N}(\theta^*, \theta^*)$$

and (6.33) implies that

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = -\chi_{A_N} \frac{1}{N} W_{\theta\theta}^{-1}(\theta^*, \theta^*) L_{\theta N}(\theta^*, \theta^*) + O_M(N^{-1}).$$

Considering (6.31) and (6.33) we have

$$(1 - \chi_{A_N}) \frac{1}{N} W_{\theta\theta}^{-1}(\theta^*, \theta^*) L_{\theta N}(\theta^*, \theta^*) = O_M(N^{-1/2}),$$

which implies that

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = W_{\theta\theta}^{-1}(\theta^*, \theta^*) \frac{1}{N} L_{\theta N}(\theta^*, \theta^*) + O_M(N^{-1})$$

Combining this with (6.38) and using the definition of $W_{\theta\theta}^{-1}(\theta^*, \theta^*)$ and $L_{\theta N}(\theta^*, \theta^*)$ we get the proof of Theorem 6.3.3.

■

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. For this consider Theorems 6.2.1, 6.2.4, 6.2.5 when $\mathcal{Y}$ is finite. We restate these theorems in this special case as a corollary.

Corollary 6.3.10 *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. Then*

$$\frac{\partial}{\partial\theta}\log p(y_n|p_{n-1},\dots p_0,\theta), \frac{\partial^2}{\partial\theta^2}\log p(y_n|p_{n-1},\dots p_0,\theta)$$

and

$$\frac{\partial^3}{\partial\theta^3}\log p(y_n|p_{n-1},\dots p_0,\theta)$$

are L-mixing processes and the limits

$$\lim_{n\to\infty} E\frac{\partial}{\partial\theta}\log p(y_n|y_{n-1},\dots y_0,\theta), \lim_{n\to\infty} E\frac{\partial^2}{\partial\theta^2}\log p(y_n|y_{n-1},\dots y_0,\theta),$$

and

$$\lim_{n\to\infty} E\frac{\partial^2}{\partial\theta^2}\log p(y_n|y_{n-1},\dots y_0,\theta),$$

exist.

Using Corollary 6.3.10 we conclude this section with a version of Theorem 6.3.3 when the state-space and the read-out space are finite.

Theorem 6.3.11 *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 6.3.2 is satisfied. Then*

$$\hat{\theta}_N - \theta^* = -(I^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots,Y_0,\theta^*) + O_M(N^{-1}), \quad (6.40)$$

where I^* is the Fisher-information matrix.

Chapter 7

Estimation with forgetting

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 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 $\theta \in D$.

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.

$$\frac{d\pi_0}{d\pi} \leq K. \tag{7.1}$$

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

$$\int |\log b^j(y,\theta)|^q b^{*i}(y)\lambda(dy) < \infty. \tag{7.2}$$

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

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda \log p(y_n|y_{n-1},\ldots,y_0;\theta), \tag{7.3}$$

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:

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

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

$$\frac{\partial}{\partial\theta}L_N^\lambda(\theta,\theta^*) = L_{\theta N}^\lambda(\theta,\theta^*) = 0 \tag{7.4}$$

More exactly $\widehat{\theta}_N(\lambda)$ is a random vector such that $\widehat{\theta}_N(\lambda) \in D$ for all ω and if the equation (7.4) has a unique solution in D, then $\widehat{\theta}_N(\lambda)$ is equal to this solution. By the measurable selection theorem such a random variable does exist.

Consider the following notations introduced in Chapter 6: let the asymptotic cost function be

$$W(\theta,\theta^*) = \lim_{n\to\infty} E_{\theta^*} \log p(Y_n|Y_{n-1},\ldots,Y_0,\theta). \tag{7.5}$$

Assume that the function $W(\theta, \theta^*)$ is smooth in the interior of D. We have

$$W_\theta(\theta^*, \theta^*) = \lim_{n\to\infty} E_{\theta^*} \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \dots, Y_0, \theta^*) = 0,$$

and for the Fisher-information matrix we have

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

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

Combining Theorem 4.1.1 and the results of Section 6.2 with the techniques of [25] we have a version of Theorem 6.3.3:

Theorem 7.0.12 *Under the conditions of Theorem 6.3.3 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 7.0.12 implies that for the covariance matrix we have

$$E(\widehat{\theta}_{n-1} - \theta^*)(\widehat{\theta}_{n-1} - \theta^*)^T = \frac{\lambda}{2} I(\theta^*)^{-1} + O(\lambda^{3/2}) + o(1). \tag{7.6}$$

Lemma 7.0.13 *Assume* $W_\theta(\theta, \theta^*) = 0$ *has a single solution* $\theta = \theta^*$ *in* D *(that is, assume the identifiability condition 6.3.2). Then for any* $d > 0$ *and* $s > 0$ *the equation (7.4) has a unique solution in* D *for* $N > c/\lambda$, *where* c *is a deterministic constant, such that it is also in the sphere* $\{\theta : |\theta - \theta^*| < d\}$ *with probability at least* $1 - c'\lambda^s$. *Here the constants depend only on* d *and* s.

Proof. We show first that the probability to have a solution outside the sphere $\{\theta : |\theta - \theta^*| < d\}$ is less than $c'\lambda^s$ with any $s > 0$ for $N > c/\lambda$. Indeed, the equation $W_\theta(\theta, \theta^*) = 0$ has a single solution $\theta = \theta^*$ in D, thus for any $d > 0$ we have

$$d' = \inf\{|W_\theta(\theta, \theta^*)| : \theta \in D, \theta^* \in D^*, |\theta - \theta^*| \geq d\} > 0$$

since $W_\theta(\theta,\theta^*)$ is continuous in (θ,θ^*) and $D\times D^*$ is compact. Therefore if a solution of (7.4) exists outside the sphere $\{\theta:\ |\theta-\theta^*|<d\}$ then we have for

$$\delta L^\lambda_{\theta N}=\sup_{\theta\in D,\theta^*\in D^*}|L^\lambda_{\theta N}(\theta,\theta^*)-W_\theta(\theta,\theta^*)|$$

the inequality $\delta L^\lambda_{\theta N}>d'$.

Due to Lemma 6.3.4 and 6.3.5 the process

$$u_n(\theta,\theta^*)=\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)-E\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)$$

and the process

$$u_{\theta n}(\theta,\theta^*)=\frac{\partial^2}{\partial\theta^2}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)-E\frac{\partial^2}{\partial\theta^2}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)$$

are L-mixing processes uniformly in (θ,θ^*).

Since $Eu_n(\theta,\theta^*)=0$ Theorem 2.3.11 is applicable, i.e.

$$\sup_{\theta\in D,\theta^*\in D^*}|L^\lambda_{\theta N}(\theta,\theta^*)-\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda E\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta)|=$$

$$O_M(\lambda^{1/2}) \tag{7.7}$$

Define the error term δ_n as in (6.28). We have that $\delta_n=O(\alpha^n)$, and the error process δ_n through an exponentially smoothing filter results the output process order of magnitude $O((1-\lambda)^N)$ for small λ at time N, i.e. for small λ $(1-\lambda>\alpha)$

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\delta_n=O((1-\lambda)^N).$$

Using the fact that

$$\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda=1-(1-\lambda)^{N+1}$$

we have

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

$$|\sum_{n=1}^{N}(1-\lambda)^{N-n}\lambda\delta_n + (1-\lambda)^{N+1}W_\theta(\theta,\theta^*)| = O((1-\lambda)^N) \tag{7.8}$$

Combining (7.7) and (7.8) we have

$$\delta L^\lambda_{\theta N} = O_M(\lambda^{1/2}) + O((1-\lambda)^N).$$

Here the second term on the right hand side is deterministic. Therefore, with some $c > 0$ that for $N > c/\lambda$ we have that $O((1-\lambda)^N) < d'/2$ and hence $P(\delta L^\lambda_{\theta N} > d') \leq P(O_M(\lambda^{1/2}) > d'/2) = O(\lambda^s)$ with any s by Markov's inequality, thus the proposition at the beginning of the proof follows.

Let us now consider the random variable

$$\delta L^\lambda_{\theta\theta N} = \sup_{\theta\in D,\theta^*\in D_0} \left\| L^\lambda_{\theta\theta N}(\theta,\theta^*) - W_{\theta\theta}(\theta,\theta^*)\right\|.$$

By the same argument as above we have

$$\delta L^\lambda_{\theta\theta N} = O_M(\lambda^{1/2}) + O((1-\lambda)^N) \tag{7.9}$$

and

$$P(\delta L^\lambda_{\theta\theta N} > d'') = O(\lambda^s)$$

for any $d'' > 0$ and $N > c/\lambda$. Hence for the event

$$A^\lambda_N = \{\omega :\ \delta L^\lambda_{\theta N} < d', \delta L^\lambda_{\theta\theta N} < d''\} \tag{7.10}$$

we have with any $s > 0$ and $N > c/\lambda$

$$P(A^\lambda_N) > 1 - O(\lambda^s). \tag{7.11}$$

But on A^λ_N the equation (7.4) has a unique solution whenever d' and d'' are sufficiently small. Indeed by Condition 6.3.2 the equation $W_\theta(\theta,\theta^*) = 0$ has a unique solution in D and hence the existence of a unique solution of (7.4) can easily be derived from the implicit function theorem, see Lemma 6.3.8. ■

Lemma 7.0.14 *We have*

$$\hat{\theta}_N(\lambda) - \theta^* = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N).$$

Proof. Consider the Taylor-series expansion of $L^\lambda_{\theta N}(\theta, \theta^*)$ around $\theta = \theta^*$ and evaluate the value of the function at $\theta = \hat{\theta}_N(\lambda)$. To simplify the notations we drop the dependence on λ, i.e. $\hat{\theta}_N = \hat{\theta}_N(\lambda)$. Then we have

$$L^\lambda_{\theta N}(\hat{\theta}_N, \theta^*) = L^\lambda_{\theta N}(\theta^*, \theta^*) + (\hat{\theta}_N - \theta^*) \int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu = 0 \tag{7.12}$$

First we prove that

$$L^\lambda_{\theta N}(\theta^*, \theta^*) = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N). \tag{7.13}$$

The process $(1-\lambda)^{N-n}\lambda\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1}, \dots Y_0, \theta)$ is a martingale difference process, see (6.34).

For (7.13) we use the Burkholder's inequality for martingales, see Theorem 2.10 in [29]:

$$E^{1/q}|\sum_{n=1}^N (1-\lambda)^{N-n}\lambda\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1}, \dots Y_0, \theta^*)|^q \leq$$

$$CE^{1/q}\left(\sum_{n=1}^N (1-\lambda)^{N-n}\lambda(\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1}, \dots Y_0, \theta^*))^2\right)^{q/2}$$

Taking the square of both sides and using the triangle inequality for the $L_{q/2}$ norm of the right hand side we get

$$E^{2/q}|\sum_{n=1}^N (1-\lambda)^{N-n}\lambda\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1}, \dots Y_0, \theta^*)|^q \leq$$

$$C^2\sum_{n=1}^N (1-\lambda)^{2(N-n)}\lambda^2 E^{2/q}|\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1}, \dots Y_0, \theta^*)|^q =$$

$$O(\lambda) + O((1-\lambda)^{2N}).$$

Here we have used the M-boundedness of the process $\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots Y_0,\theta^*)$ which follows from Theorem 6.2.1. Thus we get (7.13).

Let us now investigate the integral in (7.12). Since the function W is smooth we have for $0\le\mu\le 1$ on the set A_N^λ (defined in (7.10))

$$\|W_{\theta\theta}(\theta^*+\mu(\hat\theta_N-\theta^*),\theta^*)-W_{\theta\theta}(\theta^*,\theta^*)\|<C|\hat\theta_N-\theta^*|<Cd \tag{7.14}$$

Hence if d is sufficiently small then the positive definiteness of $W_{\theta\theta}(\theta^*,\theta^*)$ implies that

$$\int_0^1 W_{\theta\theta}\left((1-\mu)\theta^*+\mu\hat\theta_N,\theta^*\right)d\mu>cI$$

with some positive c. Since on A_N^λ

$$\left\|\int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^*+\mu\hat\theta_N,\theta^*\right)d\mu-\int_0^1 W_{\theta\theta}\left((1-\mu)\theta^*+\mu\hat\theta_N,\theta^*\right)d\mu\right\|<d''$$

it follows that if d'' is sufficiently small then denoting the minimal eigenvalue of a matrix by $\lambda_{\min}$ we have

$$\lambda_{\min}\left(\int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^*+\mu\hat\theta_N,\theta^*\right)d\mu\right)>c>0,$$

i.e.

$$\left\|\left(\int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^*+\mu\hat\theta_N,\theta^*\right)d\mu\right)^{-1}\right\|<C \tag{7.15}$$

on A_N^λ with a nonrandom constant C.

Considering equation (7.12) and that the first term is $O_M(\lambda^{1/2})+O_M((1-\lambda)^N)$ (see (7.13)) we have that

$$\chi_{A_N}(\hat\theta_N-\theta^*)=O_M(\lambda^{1/2})+O_M((1-\lambda)^N). \tag{7.16}$$

Using (7.11) and the fact that $|\hat{\theta}_N - \theta^*|$ is bounded we have

$$\chi_{A_N^c}(\hat{\theta}_N - \theta^*) = O(\lambda^s). \tag{7.17}$$

Combining (7.16) and (7.17) we get the lemma. ■

Proof. (Theorem 7.0.12) According to Lemma 7.0.14 the inequality (7.14) can be improved by $O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$. Thus we get after integration with respect to μ that

$$\left\| \int_0^1 W_{\theta\theta}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu - W_{\theta\theta}(\theta^*, \theta^*) \right\| = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$$

On the other hand from (7.9) we have

$$\left\| \int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu - \int_0^1 W_{\theta\theta}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu \right\| =$$

$$O_M(\lambda^{1/2}) + O_M((1-\lambda)^N).$$

Hence we finally get

$$\left\| \int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu - W_{\theta\theta}(\theta^*, \theta^*) \right\| = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)$$

Considering that on A_N^λ (7.15) is satisfied and the fact that $W_{\theta\theta}(\theta^*, \theta^*) > 0$ we have

$$\left\| \left(\int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu \right)^{-1} - W^{-1}_{\theta\theta}(\theta^*, \theta^*) \right\| =$$

$$O_M(\lambda^{1/2}) + O_M((1-\lambda)^N). \tag{7.18}$$

Consider (7.12) on the set A_N^λ. We have

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = -\chi_{A_N} \left(\int_0^1 L^\lambda_{\theta\theta N}\left((1-\mu)\theta^* + \mu\hat{\theta}_N, \theta^*\right) d\mu \right)^{-1} L^\lambda_{\theta N}(\theta^*, \theta^*).$$

Taking into account estimation (7.18) we have that

$$\chi_{A_N}(\hat{\theta}_N-\theta^*) = -\chi_{A_N}\left(W_{\theta\theta}^{-1}(\theta^*,\theta^*) + O_M(\lambda^{1/2}) + O_M((1-\lambda)^N)\right) L_{\theta N}^{\lambda}(\theta^*,\theta^*).$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ (7.13) implies that

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) =$$

$$-\chi_{A_N}W_{\theta\theta}^{-1}(\theta^*,\theta^*)L_{\theta N}^{\lambda}(\theta^*,\theta^*)+O_M(\lambda)+O((1-\lambda)^N)O_M(\lambda^{1/2})+O_M((1-\lambda)^{2N}) =$$

$$-\chi_{A_N}W_{\theta\theta}^{-1}(\theta^*,\theta^*)L_{\theta N}^{\lambda}(\theta^*,\theta^*) + O_M(\lambda) + O_M((1-\lambda)^{2N}).$$

Considering (7.11) and (7.13) we have

$$(1-\chi_{A_N})W_{\theta\theta}^{-1}(\theta^*,\theta^*)L_{\theta N}^{\lambda}(\theta^*,\theta^*) = O_M(\lambda^{1/2}) + O_M((1-\lambda)^N),$$

which implies that

$$\chi_{A_N}(\hat{\theta}_N - \theta^*) = W_{\theta\theta}^{-1}(\theta^*,\theta^*)L_{\theta N}^{\lambda}(\theta^*,\theta^*) + O_M(\lambda) + O_M((1-\lambda)^{2N})$$

Combining this with (7.17) and using the definition of $W_{\theta\theta}^{-1}(\theta^*,\theta^*)$ and $L_{\theta N}^{\lambda}(\theta^*,\theta^*)$ we complete the proof of Theorem 7.0.12.

■

Chapter 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 $(y_N, \ldots, y_1)$ with a prescribed accuracy, using the assumed joint density $p(y_N, \ldots, y_0; \theta)$. This interpretation of the likelihood is a central idea of the theory of stochastic complexity. Thus we interpret

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

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

Let the score function be

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

We also use lower cases for the random variable $\frac{\partial}{\partial \theta} \log p(Y_n | Y_{n-1}, \ldots, Y_0, \theta)$. In the following let

$$r = \dim \theta.$$

Theorem 8.0.15 *Under the conditions of Theorem 6.3.3 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 $c'' > 0$.

The faster the forgetting is i.e. the closer λ is to 1, the more we loose in encoding performance.

Proof. Consider the second order Taylor-series expansion of the function $\log p(y_n|y_{n-1},\dots,y_0,\widehat{\theta}_{n-1})$ around θ^*. To simplify the expressions we drop λ from the arguments and use the notation $p_n(\theta)$ for the conditional probability $p(y_n|y_{n-1},\dots,y_0,\theta)$ expressing the dependence on the parameter θ. We also use the score function defined above.

$$\log p_n(\widehat{\theta}_{n-1}) - \log p_n(\theta^*) = \varphi_n(\theta^*)(\widehat{\theta}_{n-1}-\theta^*)+$$

$$(\widehat{\theta}_{n-1}-\theta^*)^T\left(\frac{\partial^2}{\partial\theta^2}\log p_n\right)(\theta^*)(\widehat{\theta}_{n-1}-\theta^*)+O_M(\lambda^{3/2})+o_M(1),$$

where term $o_M(1)$ is due to the nonstationary initial condition and we took into account that $(\widehat{\theta}_{n-1}-\theta^*)=O_M(\lambda^{1/2})$.

Now consider the expectation of the above equality. Using that $\varphi_n(\theta^*)$ is a martingale difference, the expectation of the first term is 0. In the second term write $\left(\frac{\partial^2}{\partial\theta^2}\log p_n\right)(\theta^*)$ as

$$\left(\frac{1}{p_n}(\frac{\partial^2 p_n}{\partial\theta^2})(\theta^*)\right)-(\varphi_n(\theta^*))^T(\varphi_n(\theta^*)).$$

The expectation of the first term is 0, so

$$E\left(\log p_n(\widehat{\theta}_{n-1})-\log p_n(\theta^*)\right)=$$

$$E\left((\widehat{\theta}_{n-1}-\theta^*)^T(\varphi_n(\theta^*))^T(\varphi_n(\theta^*))(\widehat{\theta}_{n-1}-\theta^*)\right)+O(\lambda^{3/2})+o(1). \qquad (8.1)$$

Noting that $\varphi_n(\theta^*)$ depends on the past weakly, while $(\widehat{\theta}_{n-1}-\theta^*)$ depends on the past strongly for small λ, we can use the following cutting argument. Choose a positive integer $d=-c\log\lambda$ and consider the following approximations:

$$\varphi_n^+(\theta^*)=E(\varphi_n(\theta^*)|\mathcal{F}_{n-d}^+)$$

and

$$\widehat{\theta}_{n-1}^- - \theta^* = -I(\theta^*)^{-1}\sum_{i=1}^{n-d}(1-\lambda)^{n-i-1}\lambda\varphi_i(\theta^*).$$

It is easy to see that

$$\varphi_n^+(\theta^*) - \varphi_n(\theta^*) = O_M(\alpha^d)$$

with some $0 < \alpha < 1$, thus

$$\varphi_n^+(\theta^*) - \varphi_n(\theta^*) = O_M(\lambda^{c'})$$

with some $c' > 0$ and

$$\widehat{\theta}_{n-1}^- - \widehat{\theta}_{n-1} = O_M(\lambda^{1-c''}),$$

for any $c'' > 0$ (see Theorem 7.0.12 and Theorem 2.3.6). Furthermore $\varphi_n^+(\theta^*)$ and $\widehat{\theta}_{n-1}^- - \theta^*$ are independent. Approximate (8.1) by

$$E\left((\widehat{\theta}_{n-1}^- - \theta^*)^T(\varphi_n^+(\theta^*))^T(\varphi_n^+(\theta^*))(\widehat{\theta}_{n-1}^- - \theta^*)\right).$$

This can be written as

$$\mathrm{Tr}\left(E((\widehat{\theta}_{n-1}^- - \theta^*)(\widehat{\theta}_{n-1}^- - \theta^*)^T)E(\varphi_n^+(\theta^*))^T(\varphi_n^+(\theta^*)))\right).$$

The error of this approximation is $O(\lambda^{3/2-c''})$, for sufficienlty large c.

Combining the above approximation with (7.6) we have

$$E(\widehat{\theta}_{n-1}^- - \theta^*)(\widehat{\theta}_{n-1}^- - \theta^*)^T = \frac{\lambda}{2}I(\theta^*)^{-1} + O(\lambda^{3/2-c''}) + o(1).$$

Furthermore noting that

$$E(\varphi_n(\theta^*))^T(\varphi_n(\theta^*)) = I(\theta^*) + O(\alpha^i),$$

we get

$$E(\varphi_n^+(\theta^*))^T(\varphi_n^+(\theta^*)) = I(\theta^*) + O(\alpha^i) + O(\lambda^{c'}).$$

Thus

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

for all $c'' > 0$. ■

An easy consequence of Theorem 8.0.15 is the following.

Proposition 8.0.16 *Consider two different forgetting factors* $0 < \lambda_1 < \lambda_2 < 1$. *Then we have*

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

Theorem 8.0.15 has been useful in the design of a new model selection criterion. However, for a theoretical analysis of the new method it is not powerful enough. For this purpose 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 8.0.17 *Under the conditions of Theorem 6.3.3 we have*

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

Sketch of the proof: The proof of this theorem is based on the fact that under the conditions of the theorem the estimator error process $\widehat{\theta}_n - \theta^*$ has an L-mixing version ([25]). Then $C(y_n, \widehat{\theta}_n)$ and $C_n(y_n, \widehat{\theta}_{n-1}(\lambda)) - C_n(y_n, \theta^*)$ are also L-mixing. Using the law of large numbers for the latter process we get the statement.

We state a similar easy consequence as above.

Proposition 8.0.18 *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 8.0.18 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].

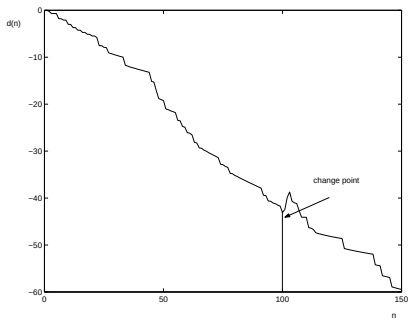

Figure 8.1: We have generated a binary HMM. The change has occured at step 100.

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

Bibliography

[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.