COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 7, No. 2, pp. 133-152, 2007

002

AN IMPROVED BOUND FOR THE EXPONENTIAL STABILITY OF PREDICTIVE FILTERS OF HIDDEN MARKOV MODELS*

LÁSZLÓ GERENCSÉR†, GYÖRGY MICHALETZKY‡, AND GÁBOR MOLNÁR-SÁSKA§

Abstract. We consider hidden Markov processes in discrete time with a finite state space $\mathcal{X}$ and a general observation or read-out space $\mathcal{Y}$, which is assumed to be a Polish space. It is well-known that in the statistical analysis of HMMs the so-called predictive filter plays a fundamental role. A useful result establishing the exponential stability of the predictive filter with respect to perturbations of its initial condition was given in [15] in the case, when the assumed transition probability matrix was primitive. The main technical result of the present paper is the extension of the cited result by showing that the random constant and the deterministic positive exponent showing up in the inequality stating exponential stability can be chosen so that for any prescribed $s \geq 1$ the s-th exponential moment of the random constant is finite. An application of this result to the estimation of HMMs with primitive transition probabilities will be also briefly presented.

Key words: hidden Markov models, predictive filters, random mappings, Doeblin-condition, risk processes, L-mixing.

1. Introduction. Hidden Markov Models have become a basic tool for modeling stochastic systems with a wide range of applications in such diverse areas as telecommunication, [20], speech recognition [12], financial mathematics [7] and protein research [22]. A good introduction to HMMs, and stochastic systems in general is given in [23]. For a survey of recent results on HMMs see [8].

We consider hidden Markov processes in discrete time with a finite state space $\mathcal{X}$ and a general observation or read-out space $\mathcal{Y}$, which is assumed to be a Polish space. We will identify $\mathcal{X}$ with $\{1, \ldots, N\}$. The motivation of our investigations is the problem of estimating the unknown true transition probability matrix Q^* and the unknown true read-out probability densities $b^{*x}(y)$ from a sequence of observations $Y_0 = y_0, \ldots, Y_n = y_n$.

The first basic results for finite state-space $\mathcal{X}$ and finite read-out space $\mathcal{Y}$ are due to Baum and Petrie, see [2]. Strong consistency of the maximum-likelihood estimator for finite-state and binary read-out HMMs has been established by Arapostathis and

*Dedicated to Tyrone Duncan in honor of his 65th birthday.

†Corresponding author, MTA SZTAKI (Computer and Automation Institute, Hungarian Academy of Sciences), 13-17 Kende u., Budapest 1111, Hungary. E-mail: gerencser@sztaki.hu, Tel: (36-1)-279-6138, (36-1)-279-6190, Fax: (36-1)-466-7503

‡Eötvös Loránd University, Pázmány Péter sétány 1/C, Budapest, 1117, Hungary. E-mail: michgy@ludens.elte.hu. MTA SZTAKI (Computer and Automation Institute, Hungarian Academy of Sciences), 13-17 Kende u., Budapest 1111, Hungary.

§ Morgan Stanley Hungary Analytics Ltd, Budapest, 15 Deák Ferenc u., Budapest 1052, Hungary. E-mail: gabor.molnar-saska@morganstanley.com, Tel: (36-1)-880-45-11

Marcus in [1]. The extension of these results to continuous read-out space requires new insights.

Taking any feasible assumed values Q and $b^x(y)$ the computation of the conditional likelihood, under the condition that the probability distribution of the initial state X_0 is q, requires the computation of the so-called predictive filter p_n, defined as

$$p_{n+1}^j = p_{n+1}^j(q) = P(X_{n+1} = j | Y_n, \dots, Y_0; q).$$

It is known that the filter process satisfies the so-called Baum-equation, see [2], which is a discrete time version of the celebrated Duncan-Mortensen-Zakai equation, see [5, 18, 24].

The standard, first step in proving consistency of the maximum likelihood method would be to show the validity of the strong law of large numbers for the log-likelihood function. An equivalent problem is to show the validity of the strong law of large numbers for a function of an extended Markov chain (X_n, Y_n, p_n). This has been investigated in the literature using basically three different methods. Leroux [16] used the subadditive ergodic theorem, LeGland and Mevel used the theory of geometric ergodicity for Markov chains, see [15] and [14], finally in [11] and [10] we used the theory of L-mixing processes. The advantage of the latter approach over previous approaches that it enables us to adapt a variety of techniques developed for the statistical theory of linear stochastic systems for HMM-s. In particular the rate of the moments of the estimation error can be established.

A major objective of this paper is to lay down the foundations for an extension of the results of [10], a preliminary version of which has been published in [11], from positive to *primitive* Q-s, i.e when $Q^r > 0$ for some r.

One of the key tools used [10] is the exponential stability of the Baum equation with respect to its initial condition q, established by Arapostathis and Marcus, [1] for binary read-outs and by LeGland and Mevel, [15], for continuous read-outs.

A slight modification of the methods of [15], using simple inequalities for projective products in a different way, gives an improved estimate for *positive* Q-s, stated here as Proposition 2.1. It states, in short, that the predictive filter forgets its initial condition with exponential rate, and the upper bound showing up in this statement is independent of the observation sequence.

In contrast to this, for the case when Q is primitive, the relevant upper bounds of [15] depend inherently on the observation sequence. The best we can get from Theorem 2.1 of [15] is that there exists an $\alpha > 0$ and a non-negative random variable $C(\omega)$ such that for any two initial distributions q, q' we have

$$\|p_n(q) - p_n(q')\|_{TV} \le C(\omega) e^{-\alpha n} \|q - q'\|_{TV}. \tag{1}$$

The more accurate characterization of the random constant $C(\omega)$ plays a key role in extending the results of [10] from positive to primitive Q-s. The main technical contribution of the paper is Theorem 5.1 stating that for any $s > 1$ there exists a small $\alpha > 0$ and a non-negative random variable $C(\omega)$, such that the above inequality holds and $\mathrm{E}C^s(\omega) < \infty$.

We use a variety of tools to derive the main technical result. In particular, the (weak) realization of Markov processes by stochastic dynamic systems, following [13], [4], and [3] will be extensively used, together with a related coupling argument for Markov processes satisfying Doeblin's condition. A basic result in the theory of risk processes to estimate the supremum of cumulative net payments, see e.g. [19, 21], will also be used. Finally, the application of the main technical result of the paper will be given in the context of L-mixing processes, see [9]. This class of processes has been used extensively in the statistical analysis of linear stochastic systems, and the relevant techniques have been successfully adapted to hidden Markov models in [10].

It is hoped that the results and the techniques of the present paper can be applied also in the context of adaptive control of HMM-s, see [6].

2. Hidden Markov Models. We consider Hidden Markov Models (HMM) or hidden Markov processes with a finite state space $\mathcal{X}$, $|\mathcal{X}| = N$, and a general observation or read-out space $\mathcal{Y}$, which is assumed to be a Polish space, i.e. a complete, separable metric space, or a measurable subset of it. We will identify $\mathcal{X}$ with $\{1, \dots, N\}$. Often $\mathcal{Y}$ is a measurable subset of an Euclidean space.

DEFINITION 2.1. *The stochastic process $(X_n, Y_n), n \geq 0$ taking its values in $\mathcal{X} \times \mathcal{Y}$ is a homogeneous, stationary hidden Markov process if the following conditions are satisfied:*

i. the state process (X_n) is a homogenous, stationary Markov process with state space $\mathcal{X}$,

ii. the observations or readouts Y_k have a conditional density given $X_k = x$ for all k:

$$P(Y_k \in y + dy | X_k = x) = b^{*x}(y)\lambda(dy),$$

where λ is a fixed nonnegative, σ-finite measure on $\mathcal{Y}$. It is thus assumed that these read-out probabilities are independent of k.

iii. we have for any positive n

$$\begin{aligned} &P(Y_n \in y_n + dy_n, \dots Y_0 \in y_0 + dy_0 | X_n = x_n, \dots X_0 = x_0) \\ &= \prod_{k=0}^{n} P(Y_k \in y_k + dy_k | X_k = x_k). \end{aligned}$$

In particular, the random variables $(Y_n, \dots, Y_0)$ are conditionally independent

and identically distributed given $(X_n, \dots, X_0)$. It is easily seen that if (X_n, Y_n) is a hidden Markov process, then $Z_n = (X_n, Y_n)$ is a Markov process. In the present paper the shorthand name "hidden Markov process" will be used for a homogeneous, stationary hidden Markov process.

REMARK 2.1. *We can relax the conditions above for two sided processes by requiring only that the past and future of* (X_n, Y_n) *for* $n < 0$ *and* $n \geq 0$ *are conditionally independent given* X_0*, see [23].*

In the notation b^{*x} the upper index * indicates that we take the true values of the corresponding, possibly unknown read-out densities, as opposed to some assumed value that shows up in the estimation problem.

If $\mathcal{Y}$ is finite, or enumerable, then we use the notation

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

and then the conditional independence can be written as

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

A classical example of a hidden Markov process is a mixture process, where (X_n) is an unobserved i.i.d. sequence, thus the sequence (Y_n) is a random mixture of i.i.d. processes. A more sophisticated example is as follows:

EXAMPLE 2.1. *Gaussian read-outs. In this case the observations are of the form*

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

where $\{\epsilon_n, n \geq 0\}$ *is a Gaussian i.i.d. sequence independent of* (X_n)*, and* $h, \sigma : \mathcal{X} \to \mathbb{R}$ *are arbitrary mappings.*

The read-out probabilities will be collected into the vector

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

We will also use the diagonal matrix with the diagonal elements $(b^{*i}(y)), i = 1, \dots, N$, denoted by

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

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

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

For notational convenience we write $Q^* > 0$ if all the elements of the transition probability matrix are strictly positive. A standing assumption throughout the paper

is that Q^* is primitive, i.e. $(Q^*)^r > 0$ with some positive integer $r > 1$. (The smallest r satisfying $(Q^*)^r > 0$ is called the index of primitivity).

We consider the problem of estimating the unknown transition probability matrix Q^* and the unknown read-out probability densities $b^{*x}(y)$ from a sequence of observations $Y_0 = y_0, \dots, Y_n = y_n$. We do not consider the problem of estimating the unknown initial probability distribution p_0^* of X_0.

To motivate our investigations consider a parametric family of transition probability matrices $Q(\theta)$ where $\theta \in D \subset \mathbb{R}^r$, and a parametric family of read-out probability densities $b^x(y;\theta)$ parameterized by the very same parameter θ. The parametrization of Q is often trivial: we simply take $N-1$ entries of each row as coordinates of the parameter vector. The set D can be arbitrary at this point, except that there should be a true parameter value $\theta^* \in D$ such that

$$b^{*x}(y) = b^x(y;\theta^*) \quad \text{and} \quad Q^* = Q(\theta^*).$$

Let $\theta \in D$ be any parameter-value, and let us write

$$b^x(y) = b^x(y;\theta) \quad \text{and} \quad Q = Q(\theta).$$

Let $(y_0, \dots y_n)$ be a sequence of observed values of $(Y_0, \dots Y_n)$. Write the conditional log-likelihood function, with condition $P(X_0 = i) = q_i$, where $q = (q_1, \dots, q_N)$ is a probability vector, (formally: $q \in \mathcal{P}$) as

$$\log p(y_0, \dots y_n; \theta, q) = \sum_{k=1}^{n-1} \log p(y_k | y_{k-1}, \dots y_0; \theta, q) + \log p(y_0; \theta, q). \tag{2}$$

Expressing the k-th term as

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

it is seen that the predictive filter, defined as

$$p_{n+1}^j = p_{n+1}^j(q) = p_{n+1}^j(\theta, q) = P(X_{n+1} = j | y_n, \dots, y_0; \theta, q),$$

is a basic entity in the analysis of the log-likelihood function. Write $p_{n+1} = (p_{n+1}^1, \dots, p_{n+1}^N)^T$.

It is known that the filter process satisfies the so-called Baum-equation

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

with initial condition $p_0 = q$, where π is the normalizing operator: for $x \geq 0,\ x \neq 0$ set $\pi(x)^i = x^i / \sum_j x^j$, see [2].

Obviously, the Baum-equation makes sense for any fixed pair of read-out density $b^x(y)$ and transition probability matrix Q.

With this notation the k-th term in (2) can be written as

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

Introducing the function of two independent variables y, p with $p = (p^1,\dots,p^N)$, parameterized by θ,

$$g(y,p\ ;\theta) = \log \sum_i b^i(y;\theta)p^i, \tag{4}$$

we can finally write, with the substitution $p_k = p_k(\theta,q)$,

$$\log p(y_0,\dots,y_n;\theta,q) = \sum_{k=1}^n g(y_k,p_k\ ;\theta) + \log p(y_0;\theta,q). \tag{5}$$

The standard, first step in proving consistency of the maximum likelihood method would be to show that

$$\lim_{n\to\infty} \frac{1}{n} \log p(Y_0,\dots Y_n;\theta,q) = W(\theta) \tag{6}$$

exists almost surely, and the limit does not depend on q. The existence of the limit in (6) has been investigated in the literature using basically three different methods. Leroux [16] used the subadditive ergodic theorem, LeGland and Mevel, [15], used the theory of geometric ergodicity for Markov chains, (a technique that has been extended in [17]), finally in [11] and [10] we used the theory of L-mixing processes. The advantage of the latter approach over previous approaches that it enables us to adapt a variety of techniques developed for the statistical theory of linear stochastic systems for HMM-s. Here we cite only one basic result of [10], the proof of which is fairly technical, but its essential ingredients are given already in [11].

THEOREM 2.1. *Consider a hidden Markov process (X_n, Y_n) as specified in Section 2 and the associated Baum-equation (3) with $y_n = Y_n$. Assume that the "true" transition probability matrix Q^* is primitive, say $Q^{*r} > 0$, while "the assumed" $Q > 0$, and that for all $x \in \mathcal{X}, y \in \mathcal{Y}$ we have $b^x(y) > 0$. Furthermore assume that for all $s \geq 1$ and for all $i, j \in \mathcal{X}$*

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

Let $q \in \mathcal{P}$ be any initial distribution and let $p_n = p_n(q)$ denote the solution of the Baum-equation (3) with $y_n = Y_n$. Then the process $g(Y_n, p_n)$ is L-mixing, the limit

$$\lim_{n\to\infty} \mathrm{E} g(Y_n, p_n) = W$$

exists, and is independent of the initial value q, and finally we have

$$\frac{1}{n}\sum_{k=1}^{n}(g(Y_k,p_k)-W)=O_M(N^{-1/2}).$$

It follows that

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

almost surely for any initial value q.

As said in the introduction a major objective of this paper is to lay down the foundations for an extension of the above result for *primitive* Q-s. A key tool in establishing Theorem 2.1 is the exponential stability of the Baum equation with respect to its initial condition q, proven by LeGland and Mevel, [15] for continuous read-outs. (A similar result for hidden Markov processes with binary read-outs were given by Arapostathis and Marcus, [1]). First we state the result of [15] for hidden Markov models for *positive* Q-s. A slight modification of the methods of [15], using simple inequalities for projective products in a different way, gives an improved estimate, stated here as Proposition 2.1. It states, in short, that the predictive filter forgets its initial condition with exponential rate, and the upper bound showing up in this statement is independent of the observation sequence. Let $\mathcal{P}(\epsilon)\subset\mathcal{P}$ be the set of $\mathbf{q}$-s such that $q_i\geq\epsilon>0$ for all i.

PROPOSITION 2.1. *Consider the Baum-equation (3) and assume that $Q>0$ and for all $x\in\mathcal{X}, y\in\mathcal{Y}$ we have $b^x(y)>0$. Then there exists a constant δ with $0<\delta<1$ depending only on Q, and for any $\epsilon>0$ there exists a constant $C>0$ depending only on Q and ϵ, such that for any $\mathbf{q}$, $\mathbf{q}'\in\mathcal{P}(\epsilon)$, for all $n\geq 0$, and for any sequence of observations $(y_0,\dots,y_{n-1})$ we have*

$$\|p_n(q)-p_n(q')\|_{TV}\leq C(1-\delta)^n\|q-q'\|_{TV}, \tag{8}$$

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

REMARK 2.2. *An essential feature of the result is that the constant term is independent of the observation sequence. Thus the standard condition $b^x(y)>0$ for all $x\in\mathcal{X}, y\in\mathcal{Y}$ can in fact be removed by a simple limiting argument. Thus important examples, such as readouts obtained by quantization are also covered by the theorem.*

3. Realizations of Markov processes. A Markov process can be defined either in terms of its transition probability kernel and its initial distribution, or by an explicit stochastic dynamic state-space system with some initial value. This duality will be explored in this section.

Consider a Polish space $\mathcal{X}$, that is the candidate for a state space, and let $\mathcal{U}$ let be another Polish space in which a noise process will take its values. Let $(U_n), n\geq 1$ be

a sequence of $\mathcal{U}$-valued i.i.d. random variables on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Let f be a Borel measurable deterministic function $f: \mathcal{X} \times \mathcal{U} \longrightarrow \mathcal{X}$. Then the sequence (X_n) defined by

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

is a homogeneous Markov chain, where $\xi \in \mathcal{X}$ is an arbitrary initialization such that ξ is independent of $(U_n), n \geq 0$. Surprisingly, the converse result also holds. The key step is what is called the realization of the Markov transition probability kernel by a stochastic dynamic state-space system given in the following theorem.

PROPOSITION 3.1. *Let $P(x, G)$, $x \in \mathcal{X}$, $G \in \mathcal{B}(\mathcal{X})$ be a transition probability kernel of a homogeneous Markov process (X_n) on a Polish space $\mathcal{X}$. 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 $(\Omega, \mathcal{F}, \mathcal{P})$, for all $x \in \mathcal{X}$ and $G \in \mathcal{B}(\mathcal{X})$ we have*

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

For the proof see [13]. It should be noted that the realization above is by no way unique.

Now if we consider a sequence of independent, $[0,1]$-uniform random variables $(U_n), n \geq 1$ on $(\Omega, \mathcal{F}, \mathcal{P})$, and a random variable $\xi' \in \mathcal{X}$ such that ξ' and $\xi = X_0$ have the same distribution, and ξ' is independent of (U_n), then the sequence (X'_n) defined over $(\Omega, \mathcal{F}, \mathcal{P})$ by

$$X'_n = f(X'_{n-1}, U_n), \quad X'_0 = \xi'$$

is a Markov-process having the same finite dimensional joint distributions as (X_n). The process (X'_n) will be called a (weak) realization of (X_n).

Assuming a general, $\mathcal{U}$-valued random variable U over $(\Omega, \mathcal{F}, \mathcal{P})$, denote the random Borel mapping $f(\cdot, U)$ of $\mathcal{X}$ into itself by T, i.e. for $x \in \mathcal{X}$ write

$$Tx = f(x, U).$$

Now if $(U_n), n \geq 1$ is an i.i.d. sequence of $\mathcal{U}$-valued random variables over $(\Omega, \mathcal{F}, \mathcal{P})$ with the same distribution as U, then we will write for $x \in \mathcal{X}, n \geq 1$

$$T_n x = f(x, U_n).$$

Thus the homogeneous Markov-process process (X'_n) defined above can be defined as

$$X'_n = T_n X'_{n-1}, \quad X'_0 = \xi'.$$

This formalism plays a key role in subsequent analysis.

In the special case when Tx does not depend on x we get that $(X'_n), n \geq 1$ is an i.i.d. sequence. Conversely, if $(X_n), n \geq 1$ is an i.i.d. process then it has a random mapping realization such that Tx is independent of x. The set of constant mappings $\mathcal{X} \to \mathcal{X}$ will be denoted by Γ_c:

$$\Gamma_c = \{R : \mathcal{X} \to \mathcal{X}, Rx \equiv z \ \text{ for some } \ z \in \mathcal{X}\}.$$

Another important tool that we will use is the Doeblin-condition, see [3]:

DEFINITION 3.1. *Let (X_n) be a homogeneous Markov process on a measurable space $\mathcal{X}$. We say that the Doeblin-condition is satisfied for (X_n) if there exists a probability measure ν on $\mathcal{X}$, a $\delta > 0$ and an integer $m \geq 1$ such that*

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

is valid for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$.

For the sake of convenience we will mostly restrict ourself to the case $m = 1$. It follows that, using the notation $\overline{P}(x, A) = (P(x, A) - \delta\nu(A))/(1 - \delta)$, we have

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

where $\overline{P}(x, A)$ is a probability transition kernel. This means that (X_n) is stochastically equivalent to a mixture of two Markov processes, one of which, with weight δ, is in fact an i.i.d. process. Thus (X_n) can be realized as follows: take an i.i.d. sequence of Bernoulli random variables, say (δ_n), such that

$$\mathcal{P}(\delta_n = 1) = \delta,$$

let (T^1_n) be a sequence of i.i.d. mappings of $\mathcal{X}$ into itself such that

$$T^1_n(\omega) \in \Gamma_c \quad \text{and} \quad \mathcal{P}(T^1_n x \in A) = \nu(A),$$

for all $\omega \in \Omega$ and $x \in \mathcal{X}$, and let (T^0_n) be a sequence of i.i.d. Borel mappings of $\mathcal{X}$ into itself realizing $\overline{P}(x, A)$. Let X'_0 has the same distribution as X_0, and assume that X'_0, (δ_n), (T^1_n) and (T^0_n) are mutually independent. Then it is easily seen that the random mappings

$$T_n = T^{\delta_n}_n$$

together with X'_0 define a realization of (X_n). From this argument we get with some extra work for the case $m \geq 1$, the following known result, see [3]:

LEMMA 3.1. *Let (X_n) be a Markov chain on a Polish space. The Doeblin-condition is valid for (X_n) with $m \geq 1$ if and only if there exists an i.i.d. random mapping realization (T_n) of its transition probability kernel such that $\mathcal{P}(T_m \dots T_1 \in \Gamma_c) \geq \delta$.*

Thus the Doeblin condition with $m \geq 1$ implies that the Markov process (X_n) has a realization such that it completely forgets its initial condition in m steps with probability at least δ. Alternatively, if two copies of the process are generated by the above mechanism choosing two different initial values, then the two processes will be coupled within at most m steps with probability at least δ.

REMARK 3.1. *An alternative realization can be obtained if we keep a record of the number of* 1*-s and* 0*-s. Define*

$$t_n^1 = |\{k : 1 \leq k \leq n,\ \delta_k = 1\}|, \tag{9}$$

and define t_n^0 *similarly. Then it is easily seen that the random mappings*

$$T_n = T_{t_n^{\delta_n}}^{\delta_n}$$

define a realization of (X_n)*. This realization will be used later in this paper.*

Finally we need the following simple observation, see [10]

LEMMA 3.2. *Let* (X_n, Y_n) *be a hidden Markov process with values in* $\mathcal{X} \times \mathcal{Y}$*, where* $\mathcal{X}$ *and* $\mathcal{Y}$ *are Polish spaces. Assume that the Doeblin-condition holds for* (X_n) *with some* $m \geq 1$*. Then the Doeblin-condition holds also for* (X_n, Y_n) *with the same* m*.*

4. Markov chains and L-mixing processes. In this short section we summarize the basic definitions related to L-mixing, as developed in [9]. L-mixing has been used extensively in the statistical analysis of linear stochastic systems, see [9]. However it is a concept which, in its motivation, strongly exploits the stability and the linear algebraic structure of the underlying stochastic system. At first sight it is not clear at all how this concept could be extended to hidden Markov processes. First of all we need the definition of M-boundedness.

DEFINITION 4.1. *A stochastic process* (X_n) *(*$n \geq 0$*) taking its values in an Euclidean space is* M*-bounded if for all* $q \geq 1$

$$M_q(X) = \sup_{n \geq 0} E^{1/q} \|X_n\|^q < \infty. \tag{10}$$

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

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

DEFINITION 4.2. *A stochastic process* (X_n) *taking its values in a finite-dimensional Euclidean space is* L*-mixing with respect to* $(\mathcal{F}_n, \mathcal{F}_n^+)$*, if it is* M*-bounded and with*

$$\gamma_q(\tau) = \sup_{n \geq \tau} E^{1/q} \|X_n - E(X_n | \mathcal{F}_{n-\tau}^+)\|^q \tag{11}$$

we have

$$\Gamma_q = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty. \tag{12}$$

To compute $E(X_n|\mathcal{F}^+_{n-\tau})$ may be hard, but in most cases the following lemma is useful, see e.g. [9]:

LEMMA 4.1. *Let X be a random variable with $E\|X\|^q < \infty$ for all $q \geq 1$, and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let η be a $\mathcal{G}$ measurable random variable. Then we have*

$$E^{1/q}\|X - E(X|\mathcal{G})\|^q \leq 2E^{1/q}\|X - \eta\|^q. \tag{13}$$

5. Hidden Markov Models with primitive transition probability matrix. The main purpose of this section is to provide an improvement of the technical result of [15] on the exponential stability of the predictive filter for HMM-s when Q and Q^* are primitive, with possibly different indices of primitivity. We will take a single, common r such that both $Q^{*r} > 0$ and $Q^r > 0$ are satisfied. The result of this section is the key technical contribution of the present paper. Its application in extending Theorem 2.1 from positive Q-s to primitive Q-s will be shortly described at the end of the section.

First we restate Theorem 2.1 of [15] below. Let $b^x(y)$ denote the read-out density let $Q(x, x')$ denote the transition probability matrix used in the Baum-equation (3), and let

$$\delta(y) = \frac{\max_x b^x(y)}{\min_x b^x(y)} \tag{14}$$

and

$$\epsilon = \min_{x,x'}{}^{+} Q(x, x'), \tag{15}$$

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

PROPOSITION 5.1. *Consider the Baum-equation (3) and assume that the transition probability matrix Q is primitive, say $Q^r > 0$ with some r. Furthermore assume that for all $x \in \mathcal{X}, y \in \mathcal{Y}$ we have $b^x(y) > 0$. Let q, $q' \in \mathcal{P}$ be any two initializations. Then for any sequence of observations $(y_0, \ldots, y_{n-1}) \in \mathcal{Y}^n$ with $n \geq r$ we have*

$$\begin{aligned} &\|p_n(q) - p_n(q')\|_{TV} \\ &\leq \epsilon^{-r}\delta(y_0)\ldots\delta(y_{r-1}) \prod_{k=1}^{\lfloor n/r \rfloor} (1 - \epsilon^r \delta^{-1}(y_{kr-r+1})\ldots\delta^{-1}(y_{kr-1}))\|q - q'\|_{TV}, \end{aligned}$$

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

Note that if $n = qr + s$ with $0 \le s < r$, then $\lfloor n/r \rfloor \cdot r = qr$, and thus the last few observations are not used in computing the bound given in the right hand side for $p_n(q)$. Furthermore note that the read-outs at times kr do not show up in the product $\prod_{k=1}^{\lfloor n/r \rfloor}$. The inclusion of these terms would give a tighter upper bound, which may not be valid.

Setting $y_k = Y_k$ and

$$\xi_k = \log(1 - \epsilon^r \delta^{-1}(Y_{kr-r+1}) \dots \delta^{-1}(Y_{kr-1})) \tag{16}$$

we get that

$$\log \|p_n(q) - p_n(q')\| \le -r \log \epsilon + \log \delta(Y_0) + \dots + \log \delta(Y_{r-1}) + \left(\sum_{k=1}^{\lfloor n/r \rfloor} \xi_k \right) + \log \|q - q'\|.$$

Note that in the special case when the transition probability matrix is positive, i.e. $r = 1$, the product in the definition of ξ_k is empty, therefore $\xi_k = \log(1 - \epsilon^r)$ identically, but the term $\delta(y_0)$ is still present, contrary to what we had in Proposition 2.1. This is due to the differences in the methods that are used for the positive and the primitive case, respectively.

Now since $\xi_k < 0$ for all $\omega \in \Omega$ we have $\mathrm{E}\xi_k < 0$. Let $\alpha > 0$ be a small positive number such that $\mathrm{E}\xi_k + \alpha < 0$ still holds. Note that ξ_k is an integrable function of the homogenous Markov chain $V_k \in \mathcal{X}^r \times \mathcal{Y}^r$ defined by

$$V_k = (X_{(k-1)r}, Y_{(k-1)r} \dots X_{kr-1}, Y_{kr-1}).$$

Assume now that the transition probability matrix Q^* is primitive, say $Q^r > 0$. Then it will be shown that (V_k) satisfies the Doeblin-condition with $m = 2$ (see Lemma 5.1 below). Thus the strong law of large numbers is applicable and we get that

$$\sigma^* = \sup_n \sum_{k=1}^{n} (\xi_k + \alpha) \tag{17}$$

is finite almost surely. It follows that

$$\sum_{k=1}^{n} \xi_k \le \sigma^* - \alpha n, \tag{18}$$

for all $n \ge 1$, and exponentiating this we get that

$$\|p_n(q) - p_n(q')\| \le \epsilon^{-r} \delta(y_0) \dots \delta(y_{r-1})\ e^{\sigma^*} e^{-\alpha n}\ \|q - q'\|.$$

Thus the exponential stability of the predictive filter follows, with the random constant term

$$C(\omega) = \epsilon^{-r}\delta(y_0)\dots\delta(y_{r-1})\ e^{\sigma^*}. \tag{19}$$

The more accurate characterization of this random constant is a key element in the statistical theory of Hidden Markov Models. The main technical result of this section with full proof is the following theorem, which is a significant improvement of Theorem 2.2 of [15]:

THEOREM 5.1. *Consider the hidden Markov process (X_n, Y_n) specified in Section 2, and the associated Baum equation (3) with $y_n = Y_n$. Assume that the transition probability matrices Q^* and Q are primitive, with possibly different indices of primitivity, and that for all $x \in \mathcal{X}, y \in \mathcal{Y}$ we have $b^x(y) > 0$. Furthermore, assume that for all $s \geq 1$ and all $i \in \mathcal{X}$*

$$\int |\delta(y)|^s b^{*i}(y)\lambda(dy) < \infty. \tag{20}$$

Then for any $s > 1$ there exists a small $\alpha > 0$, and a nonnegative random variable $C(\omega)$, with $\mathrm{E}C^s(\omega) < \infty$, so that for any two initial distributions q, q' we have

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(\omega)e^{-\alpha n}\|q - q'\|_{TV}. \tag{21}$$

This result provides the basis for the extension Theorem 2.1, from positive Q-s to primitive Q-s, to be stated at the end of the section. For the proof of the above theorem we need the following general result:

THEOREM 5.2. *Let (X_n) be a Markov chain on a Polish space $\mathcal{X}$ satisfying the Doeblin-condition with some $m \geq 1$. Assume that g is a real-valued measurable function over $\mathcal{X}$ such that $g(x) < 0$ for all $x \in \mathcal{X}$. Then for any $s > 1$ there exist an $\alpha > 0$ such that with*

$$\sigma^* = \sup_n \sum_{k=1}^{n} (g(X_k) + \alpha) \tag{22}$$

we have $\mathrm{E}(e^{s\sigma^*}) < \infty$.

REMARK 5.1. *The condition on (X_n) can be modified by assuming that there exists an atom $x^* \in \mathcal{X}$ such that for the recurrence time*

$$\tau = \min\{n > 0 : X_n = x^* | X_0 = x^*\}$$

we have with some $s' > 0$

$$\mathrm{E}(e^{s'\tau}) < \infty.$$

The proof of this remark is based on the application of Theorem 5.3 below to a sequence of i.i.d. copies of τ.

First we prove Theorem 5.2 in the special case when the Markov process is an i.i.d. sequence. We restate this for the sake of convenience in a slightly stronger form:

THEOREM 5.3. *Let (ξ_k) be a sequence of i.i.d random variables such that $\xi_k \le 0$, and $\mathcal{P}(\xi_k < 0) > 0$. Then for any $s > 1$ there exist an $\alpha > 0$ such that with*

$$\sigma^* = \sup_n \sum_{k=1}^{n} (\xi_k + \alpha) \tag{23}$$

we have $\mathrm{E}(e^{s\sigma^*}) < \infty$.

To prepare the proof we need the following comment. Let $\eta_k = \xi_k + \alpha$ and let

$$g(s) = \mathrm{E}(e^{s\eta_k})$$

denote the common moment generating function of η. Then for all $n \ge 1$ we have

$$\mathrm{E}e^{s\sigma^*} \ge \mathrm{E}e^{s\sum_{k=1}^{n}\eta_k} = g(s)^n.$$

Thus if $g(s) > 1$, then $\mathrm{E}e^{s\sigma^*} = \infty$. A surprising and beautiful result of risk theory is that on the other hand if $g(s) < 1$, then $\mathrm{E}e^{s\sigma^*} < \infty$! For details see [19, 21].

REMARK 5.2. *To see the relevance of risk theory let $\overline{\xi}_k$ denote the liability of an insurance company at time k, and let $\overline{\alpha}$ denote the fee collected at time k. Then $Y_k = \overline{\xi}_k + \overline{\alpha}$ denotes the net payment in the k-th period. The fees are established so that $\mathrm{E}Y_k > 0$. To determine the reserve fund of the insurance company that is needed to avoid bankruptcy it is crucial to estimate the expression $\inf_n \sum_{k=1}^{n} Y_k$, which is mathematically equivalent to estimating σ^* in (23).*

To complete the proof of Theorem 5.3 fix any $s > 0$ and note that

$$g(s) = \mathrm{E}(e^{s\eta_k}) = \mathrm{E}(e^{s(\xi_k+\alpha)}) = \mathrm{E}(e^{s\xi_k}) \cdot e^{s\alpha}.$$

Since $\mathrm{E}(e^{s\xi_k}) < 1$ for any $s > 0$ we can choose $\alpha > 0$ so small that $g(s) < 1$, and thus the proposition follows.

Let us turn to the proof of Theorem 5.2.

Proof. We give the proof for the case $m = 1$. The case $m \ge 1$ can be handled similarly. Let (T_n) be an i.i.d. sequence of random Borel mappings of $\mathcal{X}$ into itself giving a realization of the Markov process (X_n). By Lemma 3.1 we have $\mathcal{P}(T_n \in \Gamma_c) > \delta$. Let us define the following sequence of stopping times: $\tau_0 = 0$ and for $k \ge 1$

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

Equivalently:

$$\tau_k = \min_n \{n > \tau_{k-1} : \delta_n = 1\}.$$

Using the realization of (X_n) given in Remark 3.1 it is easily seen that

$$\xi_k = \sum_{j=\tau_k}^{\tau_{k+1}-1} g(X_j)$$

is a sequence of i.i.d. random variables.

For given n consider $t = t_n^1 = |\{k : 1 \le k \le n,\ \delta_k = 1\}|$, see (9). Then $\tau_t \le n < \tau_{t+1}$. Since $g(x) < 0$ for all $x \in \mathcal{X}$, we have

$$\sum_{k=1}^{n} g(X_k) \le \sum_{k=\tau_1}^{\tau_t - 1} g(X_k) = \sum_{k=1}^{t-1} \xi_k. \tag{24}$$

Using Theorem 5.3 we have that for any given s there exist an $\alpha_1 > 0$ and a random variable σ_1^* such that

$$\sum_{k=1}^{t-1} \xi_k \le \sigma_1^* - \alpha_1(t-1) = \sigma_1^* + \alpha_1 - \alpha_1 t, \tag{25}$$

and $\mathrm{E}e^{2s\sigma_1^*} < \infty$, and thus . Note that α_1 can be chosen so that $\alpha_1 < 1$.

To bound t from below for given n write

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

Since $\chi_{\Gamma_c}(T_k)$ are independent, identically distributed, non-negative, non-zero random variables, Theorem 5.3 implies that for any given s there exist an $\alpha_2 > 0$ and a random variable σ_2^* such that

$$-t = -t_n^1 = \sum_{k=1}^{n} -\chi_{\Gamma_c}(T_k) \le -\alpha_2 n + \sigma_2^*, \tag{26}$$

so that $\mathrm{E}e^{2s\sigma_2^*} < \infty$, yielding a lower bound for t.

From (25) and (26) we get that

$$\sum_{k=1}^{t-1} \xi_k \le \sigma_1^* + \alpha_1 - \alpha_1(\alpha_2 n - \sigma_2^*), \tag{27}$$

and thus, by (24), we have that

$$\sum_{k=1}^{n} g(X_k) \le \sigma_1^* + \alpha_1 + \alpha_1\sigma_2^* - \alpha_1\alpha_2 n.$$

For the constant term $\sigma_1^* + \alpha_1 + \alpha_1\sigma_2^*$ we have, using the Cauchy-Schwarz inequality,

$$\mathrm{E}e^{s(\sigma_1^* + \alpha_1 + \alpha_1\sigma_2^*)} \le e^{s\alpha_1}\mathrm{E}^{1/2}e^{2s\sigma_1^*}\mathrm{E}^{1/2}e^{2s\alpha_1\sigma_2^*} < \infty,$$

since $\alpha_1 < 1$, thus the proof of Theorem 5.2 is complete. □

To complete the proof of Theorem 5.1 we need a simple technical result.

LEMMA 5.1. *Let $(X_n), n \geq 0$ be a Markov process with a primitive transition probability matrix Q, say $Q^r > 0$. Then the process $(V_n), n \geq 1$ with $V_n \in \mathcal{V} = \mathcal{X}^r$ defined by*

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

satisfies the Doeblin-condition with $m = 2$.

Proof. It is easy to see that a Markov-process (V_n) on a finite state-space $\mathcal{V}$ satisfies the Doeblin-condition with some m if and only if there exists a v^* such that we have

$$P^m(u, v^*) > 0$$

for all u. Now consider an arbitrary $i_0 \in \mathcal{X}$ and a sequence $i_1, \dots, i_r$ such that

$$P(X_{n+1} = i_{l+1} | X_n = i_l) > 0$$

for $l = 0, \dots, r-1$ and let

$$v^* = (i_1, \dots, i_r)^T.$$

We prove that v^* is accessible from any $u \in \mathcal{V}$ in two steps with positive probability.

Consider an arbitrary state $u = (k_1, \dots, k_r) \in \mathcal{V}$, and consider the last coordinate $k_r \in \mathcal{X}$. Since Q is primitive with $Q^r > 0$, the state i_0 can be reached from k_r in r steps. Set $j_0 = k_r$. Thus there exists a sequence $j_1, \dots, j_{r-1}, j_r = i_0$ such that

$$P(X_{n+1} = j_{l+1} | X_n = j_l) > 0$$

for $l = 0, \dots, r-1$. Set

$$u_1 = (j_1, \dots, j_r).$$

Then obviously

$$P(V_{n+1} = u_1 | V_n = u) > 0 \quad \text{and} \quad P(V_{n+2} = v^* | V_{n+1} = u_1) > 0$$

and thus Lemma 5.1 follows. □

Proof. (of Theorem 5.1). By Proposition 5.1 for any q, q', any integer $n \geq r$ and any sequence $y_0 = Y_0, \dots y_{n-1} = Y_{n-1} \in \mathcal{Y}$ we have

$$\log \|p_n(q) - p_n(q')\| \leq -r \log \epsilon + \log \delta(Y_0) + \dots + \log \delta(Y_{r-1}) + \sum_{k=1}^{\lfloor n/r \rfloor} \xi_k + \log \|q - q'\|,$$

where

$$\xi_k = \log(1 - \epsilon^r \delta^{-1}(Y_{kr-r+1}) \dots \delta^{-1}(Y_{kr-1})). \tag{28}$$

Consider the term

$$\sum_{k=1}^{\lfloor n/r \rfloor} \xi_k = \sum_{k=1}^{\lfloor n/r \rfloor} \log(1 - \epsilon^r \delta^{-1}(Y_{kr-r+1}) \dots \delta^{-1}(Y_{kr-1})).$$

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

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

Also note that ξ_k is a function of

$$V_k = (X_{(k-1)r}, Y_{(k-1)r} \dots X_{kr-1}, Y_{kr-1}),$$

and here $V_k \in \mathcal{V} = \mathcal{X}^r \times \mathcal{Y}^r$ form a homogenous, stationary Markov chain. By Lemma 5.1 the homogeneous, stationary Markov chain

$$Z_k = (X_{(k-1)r}, \dots, X_{kr-1})$$

satisfies the Doeblin-condition with $m = 2$. It is easily seen that (V_k) is a hidden Markov process with state space process (Z_k), therefore Lemma 3.2 implies that it also satisfies the Doeblin-condition with $m = 2$. Thus using Theorem 5.2 with $g(V_k) = \log(1 - \epsilon^r \delta^{-1}(Y_{kr-r+1}) \dots \delta^{-1}(Y_{kr-1}))$ we get that for any $s \geq 1$ there exist an $\alpha > 0$ and a random variable $\widetilde{c}$ such that

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

and $\mathrm{E} e^{s\widetilde{c}} < \infty$. The exponential moments of the term $\log \delta(Y_0) + \dots + \log \delta(Y_{r-1})$ are finite for any s by condition (20), thus the proof of Theorem 5.1 is complete.

□

A major application of the previous result is given in the following theorem, which is an extension of Theorem 2.1. The conditions of this theorem are obtained by merging the conditions of Theorem 2.1, (replacing the condition $Q > 0$ by the condition that Q is primitive), and the conditions of Theorem 5.1. The proof is fairly technical, but very similar to the proof of Theorem 2.1.

THEOREM 5.4. *Consider a hidden Markov process (X_n, Y_n) specified in Section 2 and the associated Baum-equation (3) with $y_n = Y_n$. Assume that the transition probability matrices Q^* and Q are primitive, with possibly different indices of primitivity, and that for all $x \in \mathcal{X}$ we have $b^x(y) > 0$ for λ almost all $y \in \mathcal{Y}$. Furthermore assume that for all $s \geq 1$ and for all $i, j \in \mathcal{X}$*

$$\int |\log b^j(y)|^s \; b^{*i}(y) \lambda(dy) < \infty, \tag{29}$$

and also that for all $s \geq 1$ *and for all* $i \in \mathcal{X}$

$$\int |\delta(y)|^s b^{*i}(y)\lambda(dy) < \infty. \tag{30}$$

Let $q \in \mathcal{P}$ *be any initial distribution and let* $p_n = p_n(q)$ *denote the solution of the Baum-equation (3) with* $y_n = Y_n$. *Then the process* $g(Y_n, p_n)$ *is L-mixing, the limit*

$$\lim_{n\to\infty} \mathrm{E} g(Y_n, p_n) = W$$

exists, and is independent of the initial distribution q, *and finally we have*

$$\frac{1}{n}\sum_{k=1}^{n}(g(Y_k, p_k) - W) = O_M(N^{-1/2}).$$

It follows that

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

almost surely for any initial distribution q.

REMARK 5.3. *Note that if* $b^j(y)$ *is bounded in* y *for all* j *then condition (29) is implied by condition (30). Although the latter is not satisfied for Gaussian read-outs, it can be enforced by a deliberate distortion of the observations by projecting data outside of a given ball onto its boundary.*

REFERENCES

[1] A. ARAPOSTATHIS AND S. I. MARCUS. *Analysis of an Identification Algorithm Arising in the Adaptive Estimation of Markov Chains*. Math. Control Signals Systems, 3(1990), pp. 1–29.

[2] L. E. BAUM AND T. PETRIE. *Statistical inference for probabilistic functions of finite state Markov chains*. Ann. Math. Stat., 37(1966), pp. 1559–1563.

[3] 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.

[4] V. S. BORKAR. *On white noise representations in stochastic realization theory*. SIAM J. Control Optim., 31(1993), pp. 1093–1102.

[5] T. E. DUNCAN. *Probability densities for diffusion processes with applications to nonlinear filtering theory*. Technical report, PhD thesis, Standford, 1967.

[6] 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, pp. 1369–1374, 2002.

[7] R. J. ELLIOTT, W. P. MALCOLM, AND A. TSOI. *HMM Volatility Estimation*. In: Proceedings of the 41th IEEE Conference on Decision & Control, Las Vegas, pp. 398–404, 2002.

[8] Y. EPHRAIM AND N. MERHAV. *Hidden Markov Processes*. IEEE Transactions on Information Theory, 48(2002), pp. 1508–1569.

[9] L. GERENCSÉR. *On a Class of Mixing Processes*. Stochastics, 26(1989), pp. 165–191.

[10] L. Gerencsér, G. Molnár-Sáska, Gy. Michaletzky, and G. Tusnády. *A new approach for the statistical analysis of hidden Markov models*. IEEE Transactions on Automatic Control, under revision.

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

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

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

[14] 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(2000), pp. 41–62.

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

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

[17] L. Mevel and L. Finesso. *Asymptotical statistics of misspecified hidden markov models*. IEEE Transactions on Automatic Control, 49(2004), pp. 1123–1132.

[18] R. E. Mortensen. *Optimal control of continuous time stochastic systems*. Technical report, PhD thesis, University of California, Berkeley, CA, USA, 1966.

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

[20] 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, pp. 4123–4124., 1999.

[21] E. Sparre Andersen. *On the collective theory of risk in the case of contagation between the claims*. In: Transactions of the XV-th International Congress of Actuaries, pp. 219–229, 1957.

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

[23] J. H. van Schuppen. *Mathematical control and system theory of stochastic systems in discrete time*, 2006.
http://homepages.cwi.nl/ schuppen/courses/lecturenotes/controlstocdtlecnotes.pdf, CWI, Amsterdam, The Netherlands.

[24] M. Zakai. *On the optimal filtering of diffusion pro- cesses*. Zeitschrift für Wahrscheinlichkeitstheorie und ver- wandte Gebiete, 11(1969), pp. 230-243.