ADAPTIVE ENCODING AND PREDICTION OF HIDDEN MARKOV PROCESSES

L. Gerencsér, G. Molnár-Sáska

MTA SZTAKI,
Computer and Automation Institute, Hungarian Academy of Sciences,
13-17 Kende u., Budapest 1111, Hungary,
Tel: (36-1)-279-6138, (36-1)-279-6217, Fax: (36-1)-4667503
gerencser@sztaki.hu, molnar@math.bme.hu

Keywords: Hidden Markov Models, maximum-likelihood estimation, adaptive encoding, adaptive prediction, stochastic complexity

Abstract

The purpose of this paper is to provide explicit results on the almost sure asymptotic performance of adaptive encoding and prediction procedures for finite-state Hidden Markov Models. In addition, Rissanen's tail condition [14] will be verified, from which a lower bound for the mean-performance of universal encoding procedures will be derived. The results of this paper are based on [10].

1 Introduction

Hidden Markov Models have become a basic tool for modeling stochastic systems with a wide range of applicability. For a general introduction see [16]. The estimation of the dynamics of a Hidden Markov Model is a basic problem in applications. A key element in the statistical analysis of HMM-s is a strong law of large numbers for the log-likelihood function, see [11], [12], [3]. An alternative tool that has been widely used in linear system identification is theory of L-mixing processes. The relevance of this theory is established in [10] using a random-transformation representation for Markov-processes (see [9]). The advantage of this approach is that, under suitable conditions a more precise characterization of the estimation error-process can be obtained, which, in turn, is crucial for the analysis of the performance of adaptive prediction, see [6].

The purpose of this paper is to provide explicit results on the almost sure asymptotic performance of adaptive encoding and prediction procedures for finite-state Hidden Markov Models. In addition, Rissanen's tail condition [14] will be verified, from which a lower bound for the mean-performance of universal encoding procedures will be derived.

2 Hidden Markov Models

We consider Hidden Markov Models with a general state space $\mathcal{X}$ and a general observation or read-out space $\mathcal{Y}$. Both are assumed to be Polish spaces, i.e. they are complete, separable metric spaces.

Definition 2.1 *The pair (X_n, Y_n) is a Hidden Markov process if (X_n) is a homogenous Markov chain, with state space $\mathcal{X}$ and the observations (Y_n) are conditionally independent and identically distributed given (X_n).*

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

$$P(Y_n = y_n, \dots Y_0 = y_0 | X_n = x_n, \dots X_0 = x_0) =$$

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

In this case we will use the following notations

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

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

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

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

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

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

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

$$p^*_{n+1} = \pi(Q^{*T} B^*(Y_n) p^*_n), \tag{2}$$

where π is the normalizing operator: for $x \geq 0,\ x \neq 0$ set $\pi(x)^i = x^i / \sum_j x^j$, see [1]. Here $p^{*j}_0 = P(X_0 = j)$.

In practice, the transition probability matrix Q^* and the initial probability distribution p^*_0 of the unobserved Markov chain (X_n) and the conditional probabilities $b^{*i}(y)$ of the observation sequence (Y_n) are possibly unknown. For this reason we consider the Baum-equation in a more general sense

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

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

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

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

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

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

where $i = 1, \ldots, N$, then the conditional probability defined under 1 will satisfy the Baum-equation. In the rest of the section we deal with continuous read-out, which includes the finite case in a natural manner.

We will take an arbitrary probability vector q as initial condition, and the solution of the Baum equation will be denoted by $p_n(q)$.

A key property of the Baum equation is its exponential stability with respect to the initial condition. This has been established in [11] for continuous read-outs. Here we state the result for HMM-s with a positive transition probability matrix:

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

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(1-\delta)^n \|q - q'\|_{TV}, \qquad (4)$$

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

If Q is only primitive, i.e. $Q^r > 0$ with some positive integer $r > 1$, then (4) holds with a random C.

Next we are going to introduce the notion of Doeblin-condition (see [2]):

Definition 2.2 *If there exists an integer $m \geq 1$ such that $P^m(x, A) \geq \delta\nu(A)$ is valid for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ with some probability measure ν, then we say that the Doeblin-condition is satisfied.*

Now let (X_n, Y_n) be a Hidden Markov process and assume that the state space $\mathcal{X}$ and the observed space $\mathcal{Y}$ are Polish.

Lemma 2.1 *Assume that the Doeblin condition holds for the Markov chain (X_n). Then the Doeblin condition holds for (X_n, Y_n) as well.*

3 Markov chains and L-mixing processes

Now we are going to introduce a class of processes called L-mixing processes which have been used extensively in the statistical analysis of linear stochastic systems, see [5].

Definition 3.1 *A stochastic process (X_n) ($n \geq 0$) taking its values in an Euclidean space is M-bounded if for all $q \geq 1$*

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

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

Definition 3.2 *A stochastic process (X_n) taking its values in a finite-dimensional Euclidean space is L-mixing, if it is M-bounded and with*

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

we have

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

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

Proposition 3.1 *Let (X_n) be a Markov chain with state space $\mathcal{X}$, where $\mathcal{X}$ is a Polish space, and assume that the Doeblin condition is valid for $m = 1$. Furthermore let $g : \mathcal{X} \longrightarrow \mathbb{R}$ be a bounded, measurable function. Then $g(X_n)$ is an L-mixing process.*

4 Estimation of Hidden Markov Models

This section gives a brief outline of the maximum likelihood estimation of Hidden Markov Models. Consider a Hidden Markov Process (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is continuous, a measurable subset of $\mathbb{R}^d$. Assume that the transition probability matrix and the conditional read-out densities are positive, i.e. $Q^* > 0$ and $b^{*i} > 0$ for all i, y. Then the process (X_n, Y_n) satisfies the Doeblin-condition.

Let the invariant distribution of $\mathcal{X}$ be ν and the invariant distribution of $\mathcal{X} \times \mathcal{Y}$ be π. Then

$$\pi^i(dy) = \nu_i b^{*i}(y)\lambda(dy), \qquad (5)$$

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

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

The logarithm of the likelihood function is

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

Here the k-th term for $k \geq 1$ can be written as

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

Now write

$$g(y, p) = \log \sum_i b^i(y) p^i, \tag{6}$$

then we have

$$\log p(y_N, \ldots, y_0, \theta) = \sum_{k=1}^N g(y_k, p_k) + \log p(y_0, \theta). \tag{7}$$

It is easy to see that the Doeblin condition is not satisfied for the process (X_n, Y_n, p_n), thus Proposition 3.1 is not applicable directly. For this reason we look for a different characterization of (X_n, Y_n, p_n).

Theorem 4.1 *Consider a Hidden Markov Model (X_n, Y_n), where the state space $\mathcal{X}$ is finite and the observation space $\mathcal{Y}$ is continuous, a measurable subset of $\mathbb{R}^d$. Let $Q, Q^* > 0$ and $b^i(y), b^{*i}(y) > 0$ for all i, y. Let the initialization of the process (X_n, Y_n) be random, where the Radon-Nikodym derivate of the initial distribution π_0 w.r.t the stationary distribution π is bounded, i.e.*

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

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

$$\int |\log b^j(y)|^q b^{*i}(y) \lambda(dy) < \infty. \tag{9}$$

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

Remark 4.1 *Since the positivity of Q implies that the stationary distribution of (X_n) is strictly positive in every state and the densities of the read-outs are strictly positive Condition (8) is not a strong condition. For example for the random initialization we can take a uniform distribution on $\mathcal{X}$ and an arbitrary set of λ a.e. positive density functions $b_0^i(y)$.*

To analyze the asymptotic properties of the right hand side of (7) Theorem 4.1 seems to be relevant. Under the conditions of Theorem 4.1 $g(y, p)$ is an L-mixing process and the law of large numbers is valid for such processes, see [5]. This implies the existence of the limit of (7).

Consider now a *finite state-finite read-out* HMM. This case follows from Theorem 4.1, but the integrability condition (9) is simplified due to the discrete measure.

Theorem 4.2 *Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Assume that the process (X_n, Y_n) satisfies the Doeblin condition. Let the running value of the transition probability matrix Q be positive and $b^i(y) \geq \delta > 0$ for all i, y. Then with a random initialization on $\mathcal{X} \times \mathcal{Y}$ we have that $g(Y_n, p_n)$ is an L-mixing process.*

Consider a finite state-finite read-out HMM, parameterized by θ, where $|\mathcal{X}| = N$ and $|\mathcal{Y}| = M$ and θ containing the elements of the transition probability matrix and the read-out probabilities. Thus θ is an $N^2 + NM - 2N$ dimensional vector with coordinates between 0 and 1. Furthermore let the ML estimate of the true parameter θ^* be denoted by $\hat{\theta}_N$. Due to [11] the gradient process $\partial p_n(\theta)/\partial\theta$ is also exponentially stable, thus the process $\partial g(Y_n, p_n(\theta))/\partial\theta$ is an L-mixing process, see [10]. Similarly it can be shown that $\partial^2 g(Y_n, p_n(\theta)/\partial\theta^2$ is also an L-mixing process. The arguments of [6] yield the following result.

Theorem 4.3 *Consider the Hidden Markov Model (X_n, Y_n), where $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let $Q, Q^* > 0$ and $b^i(y) \geq \delta$, $b^{*i}(y) \geq \delta$ for all i, y, where $\delta > 0$. Let $\hat{\theta}_N$ be the ML estimate of θ^*. Then $\hat{\theta}_N - \theta^*$ can be written as*

$$-(I(\theta^*)^{-1} \frac{1}{N} \sum_{n=1}^N \frac{\partial}{\partial\theta} \log p(Y_n|Y_{n-1}, \ldots, Y_0, \theta^*) + r_n, \tag{10}$$

where $r_n = O_M(N^{-1})$, i.e Nr_n is M-bounded, and $I(\theta^*)$ is the Fisher-information matrix.

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

Next we are going to prove that the tail-condition in Rissanen-theorem, see in [14], for the error term of the estimation θ is satisfied.

Theorem 4.4 *Under the condition of Theorem 4.3 we have*

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

where $c > 0$ is an arbitrary constant

Proof: Let

$$J_n = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log p(Y_i|Y_{i-1}, \ldots, Y_0, \theta)_{|\theta=\theta^*}$$

Then (J_n) is a martingale, and $\sup(J_n - J_{n-1})$ is

$$\frac{\partial}{\partial\theta} \log p(y_n|y_{n-1}, \ldots, y_0, \theta)_{|\theta=\theta^*} \leq c \tag{11}$$

with some positive constant $c > 0$. Let the martingale (J_n) be $\mathcal{G}_n$-adapted. Furthermore let (A_n) denote the increasing process associated with the submartingale (J_n^2), making $J_n^2 - A_n$ a martingale. A_n has a form

$$A_n = \sum_{k=1}^n E((J_k - J_{k-1})^2|\mathcal{G}_{k-1}).$$

Then (11) implies that

$$A_n \leq c'n \tag{12}$$

with some positive constant c'. The following very simple lemma is given in [13]:

Lemma 4.1 *Let J_n be a square-integrable martingale such that* $\sup(J_{n+1} - J_n) \le c$ *a.s., where $c > 0$. Then*

$$\exp(\lambda J_n - \mu A_n)$$

is a positive supermartingale, where λ is an arbitrary positive number, and μ is a positive number depending only on λ and c.

As a consequence of the lemma we have that

$$E \exp(\lambda J_n - \mu A_n) \le K.$$

Using (12) we have

$$E \exp(\lambda J_n - \mu c' n) \le K,$$

$$E \exp(\lambda J_n) \le \exp(c'' n), \tag{13}$$

where $e^{c''} = K e^{\mu c'}$.

Consider the process $\widehat{J}_n = \frac{1}{\sqrt{N}} J_n$ for an arbitrary fix N. Using inequality (13) for $\widehat{J}_n$ with the respective increasing associated process $\widehat{A}_n = \frac{A_n}{N}$ and $\widehat{c} = \frac{c}{N} < c$ at $n = N$ we have

$$E \exp(\lambda \frac{1}{\sqrt{N}} J_N) \le c''. \tag{14}$$

The last inequality and Markov's inequality imply that

$$P(\frac{1}{\sqrt{N}} J_N > d \log N) =$$

$$P(\exp(\lambda \frac{1}{\sqrt{N}} J_N) > \exp(\lambda d \log N)) \le \frac{c''}{N^{d\lambda}}.$$

Thus if $d\lambda > 1$ then

$$\sum_{N=1}^{\infty} P(\frac{1}{\sqrt{N}} J_N > d \log N) < \infty. \tag{15}$$

Using Theorem 4.3 we have

$$\frac{1}{\sqrt{N}} J_N = N^{\frac{1}{2}} (-I(\theta^*))(\hat{\theta}_N - \theta^*) + O_M(N^{-\frac{1}{2}}).$$

This implies that one term in the sum is

$$P(N^{\frac{1}{2}} (-I(\theta^*))(\hat{\theta}_N - \theta^*) + O_M(N^{-\frac{1}{2}}) > d \log N) <$$

$$P(N^{\frac{1}{2}} (\hat{\theta}_N - \theta^*) > \frac{d}{2} \log N) + P(O_M(N^{-\frac{1}{2}}) > \frac{d}{2} \log N),$$

and $P(O_M(N^{-\frac{1}{2}}) > \frac{d}{2} \log N) < C N^{-s}$ for all $s > 1$ thus we get that the second term is summable.

Using (15) this implies that

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

thus the tail probabilities are uniformly summable as stated in the theorem.

In the next section we are going to introduce some consequences of this result.

5 Encoding of finite state Hidden Markov Models

The negative logarithm of the conditional probability

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

can be interpreted as a code length, see [15]. An adaptive encoding procedure is obtained if we set $\theta = \hat{\theta}_{n-1}$. Following [7] we get the following result:

Theorem 5.1 *Let s_n denote the loss in codelength:*

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

Under the conditions of Theorem 4.3 we have

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

where $p = \dim \theta$. Furthermore

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

with probability 1.

This result can be used for model selection for HMM-s, see [8], [4]. Due to the validity of Rissanen's tail condition the following "converse theorem" is also true by virtue of the fundamental theorem of the theory of stochastic complexity (cf. [14]):

Theorem 5.2 *Let $g_n(y_1, \ldots, y_n)$ be an arbitrary sequence of compatible probability distributions. Then*

$$\liminf_{n \to \infty} \frac{1}{\log n} E_\theta(-\log g_n(y_n, \ldots, y_1) + \log p(y_n, \ldots y_1, \theta))$$

is at least $p/2$ except for a set of θ's with Lebesgue-measure 0.

Theorem 5.1 can be extended to performance indexes different from the conditional entropy. Let (y_n) be a binary process taking value 0 or 1. Let e.g. $\hat{y}_n$ be the predictor defined by

$$\hat{y}_n(\theta) = \begin{cases} 1 & \text{if } q_n(\theta) = p(y_n = 1 | y_{n-1}, \ldots, y_1, \theta) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Define $q_n^* = P_{\theta^*}(Y_n = 1 | Y_{n-1}, \ldots, Y_1, \theta^*)$ and similarly $q_n = P_{\theta^*}(Y_n = 1 | Y_{n-1}, \ldots, Y_1, \theta)$. Then the failure probability can be expressed as

$$P_{\theta^*}(\hat{Y}_n(\theta) \neq Y_n) = \int_0^{1/2} (1 - q_n) q_n^* d\varphi_n(q_n(\theta)) +$$

$$\int_{1/2}^{1} (1 - q_n^*) q_n d\varphi_n(q_n(\theta)) = W_n(\theta),$$

where $d\varphi_n(q_n(\theta))$ is the distribution of $q_n(\theta)$ under P_{θ^*}.

Under the condition of Theorem 4.3 $\varphi_n(q_n(\theta))$ can be shown to converge in distribution to $\varphi(q(\theta))$ having an invariant distribution $\varphi(q,\theta)$. Let

$$W(\theta) = \lim_n W_n(\theta).$$

For finite n the function $W_n(\theta)$ is smooth in θ. Assuming that smoothness is inherited by $W(\theta)$ define

$$S^* = \frac{\partial^2}{\partial\theta^2} W(\theta)_{|\theta=\theta^*}.$$

The adaptive predictor of y_n is defined as

$$\hat{y}_n = \hat{y}_n(\hat{\theta}_{n-1}).$$

We have the following result:

Theorem 5.3 *Let the loss in prediction performance be*

$$T_n = P_{\theta^*}(\hat{Y}_n(\hat{\theta}_{n-1}) \neq Y_n) - P_{\theta^*}(\hat{Y}_n(\theta^*) \neq Y_n).$$

Under the conditions of Theorem 5.1 we have

$$E(T_n) = \frac{1}{2n}(\mathrm{Tr} S^* I(\theta^*)^{-1} + o(1)),.$$

Moreover

$$\lim_{N\to\infty} \frac{1}{\log N} \sum_{n=1}^{N} T_n = \mathrm{Tr} S^* I(\theta^*)^{-1}$$

with probability 1.

The invariant distribution of $\varphi(q(\theta))$ in exact form even in the simplest cases is unknown. Thus the theoretical value of $I(\theta^*)$ and S^* is unknown.

6 Acknowledgement

The authors acknowledge the support of the National Research Foundation of Hungary (OTKA) under Grant no. T 032932.

References

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

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

[3] R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernoulli*, 7:381–420, 2001.

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

[5] L. Gerencsér. On a class of mixing processes. *Stochastics*, 26:165–191, 1989.

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

[7] L. Gerencsér. On Rissanen's predictive stochastic complexity for stationary ARMA processes. *Statistical Planning and Inference*, 41:303–325, 1994.

[8] 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.

[9] L. Gerencsér and G. Molnár-Sáska. A new method for the analysis of Hidden Markov Model estimates. In *Proceedings of the 15th Triennial World Congress of the International Federation of Automatic Control, Barcelona*, pages T–Fr–M03, 2002.

[10] 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*, pages WeP09–6 2272–2277., 2002.

[11] F. LeGland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems*, 13:63–93, 2000.

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

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

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

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

[16] J. H. van Schuppen. Lecture notes on stochastic systems. Technical report. Manuscript.