Estimation and strong approximation of Hidden Markov Models

László Gerencsér[1] and Gábor Molnár-Sáska[2]

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

Abstract. We give novel conditions for the existence of the limit of the normalized log-likelihood function for a finite-state continuous read-out Hidden Markov Model. Our results complements those of [8]. The result is then applied to derive a strong approximation result for the parameter estimates of the Hidden Markov Model.

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 [11]. 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 [8], [7], [9], [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 [6] using a random-transformation representation for Markov-processes. The advantage of this approach is that, under suitable conditions a more precise characterization of the estimation error-process can be obtained. The purpose of this paper is to give conditions which can be used in practice, and compare our conditions with that of [8].

2 Hidden Markov Models

We consider Hidden Markov Models with a finite state space $\mathcal{X}$ and a possibly continuous observation or read-out space $\mathcal{Y}$. The latter is assumed to be a measurable set of $\mathbf{R}^d$ as in [8] or [9]. 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 given (X_n), i.e. for any integer $n \geq 0$, and for any $i_0, \dots i_n \in \mathcal{X}$

$$P(Y_n \in dy_n, \ldots, Y_0 \in dy_0 | X_n = i_n, \ldots, X_0 = i_0) = \prod_{k=0}^{n} P(Y_k \in dy_k | X_k = i_k),$$

and the conditional distributions, considered as functions of y and x, are independent of k.

Let Q^* be the transition matrix of the unobserved Markov process (X_n), i.e. $Q^*_{ij} = P(X_{n+1} = j | X_n = i)$, and let the read-outs have a density with respect to a fixed nonnegative, σ-finite measure λ on $\mathbf{R}^d$

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

and denote

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

Let the predictive filter be defined by

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

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

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

where π is the normalizing operator, to ensure that p^*_{n+1} is a probability vector: 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 densities $b^{*i}(y)$ of the observation sequence (Y_n) are possibly unknown. For this reason we consider the Baum-equation with running parameters Q, $b^i(y)$

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

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

In the sequel we write (2) shortly as

$$p^*_{n+1} = f(Y_n, p^*_n). \tag{3}$$

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

Let the parameter of the model be denoted by θ^*. Usually θ^* is assumed to be composed of the elements of the unknown transition probability matrix and the parameters of the densities of the read-outs.

3 Exponential forgetting

A key property of the Baum-equation is its exponential stability with respect to the initial condition. This has been established in [8], see Theorem 2.2. Recall that a stochastic matrix Q is primitive if there exists an $r \geq 1$ such that Q^r is positive.

Theorem 1. *If the stochastic matrices Q, Q^* are primitives and the read-out densities are positive, i.e. $b^i(y), b^{*i}(y) > 0$ then for any two initializations q, q'*

$$\|p_m(q) - p_m(q')\|_{TV} \leq C(\omega)(1-\delta)^m \|q - q'\|_{TV}$$

almost surely under the true probability measure with $0 < \delta < 1$, where $C(\omega)$ is a finite random variable and $\| \ \|_{TV}$ denotes the total variation norm.

Theorem 2. *If the matrix Q is positive in Theorem 1, then the "almost surely" can be replaced "for all ω" by $C(\omega) = C$, i.e.*

$$\|p_m(q) - p_m(q')\|_{TV} \leq C(1-\delta)^m \|q - q'\|_{TV}.$$

We say that the prediction filter generated by the Baum-equation is uniformly exponentially stable, i.e. any initial condition for the prediction filter is forgotten exponentially fast.

4 The log-likelihood function

For the consistency of the maximum-likelihood estimation we need to show the existence of the limit $\log p(y_{n-1}, \dots y_0, \theta)/n$ as n goes to infinity. We can write $\log p(y_{n-1}, \dots y_0, \theta)$ as

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

Here the k-th term is $\log \sum_i b^i(y_k) P(i | y_{k-1}, \dots, y_0, \theta) = \log \sum_i b^i(y_k) p_k^i(\theta)$. Let

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

Thus we get

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

Definition 1. *(see [10]) We say that the Markov process (X_n) is geometrically ergodic if there exists a stationary distribution π and there exists an x-dependent constant M_x and an x-independent constant $0 < \rho < 1$ for all initialization $x \in \mathcal{X}$ such that*

$$\|P^n(x, \cdot) - \pi\|_{TV} \leq M_x \rho^n.$$

If the transition probability matrix Q^* is primitive then the process (X_n, Y_n) is geometrically ergodic (see [8], pp. 70) under the true probability measure, with a unique invariant probability distribution π on $\mathcal{X} \times \mathcal{Y}$, and for any $i \in \mathcal{X}$

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

where ν denotes the stationary distribution on $\mathcal{X}$. In addition the geometrically ergodicity of (X_n, Y_n, p_n) is proved (see Theorem 4 below).

To describe the results of [6] we are going to introduce the notion of Doeblin-condition, see [2]:

Definition 2. *Given a Markov chain (X_n) with state space $\mathcal{X}$. We say that the Doeblin-condition is satisfied if there exists an integer $m \geq 1$ such that*

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

is valid for all $x \in \mathcal{X}$ and $A \subset \mathcal{B}(\mathcal{X})$, where $\mathcal{B}(\mathcal{X})$ denotes the Borel measurable sets, with $\delta > 0$ and some probability measure ν.

Lemma 1. *(see [6]) Assume that the Doeblin-condition holds for the Markov process (X_n), and the mapping f defined in (3) is uniformly exponentially stable. Then the process (X_n, Y_n, p_n) has a stationary distribution.*

Theorem 3. *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 $\mathbf{R}^d$. Let the running value of the transition probability matrix Q and the running value of the conditional read-out densities be positive, i.e. $Q > 0$ and $b^i(y) > 0$, respectively. Assume that the process (X_n, Y_n) satisfies the Doeblin-condition. Let the initialization of the process (X_n, Y_n) be random, where the Radon-Nikodym derivate of the initialized distribution π_0 w.r.t the stationary distribution π is bounded, i.e.*

$$\frac{d\pi_0}{d\pi} \leq K. \tag{6}$$

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

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

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

For the definition of L-mixing see the Appendix. The law of large numbers is valid for L-mixing processes, see [4], which implies the existence of the limit in (5).

The corresponding result in [8] is the following, given as Theorem 3.5 of [8], which also implies the existence of the limit in (5).

Theorem 4. *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 $\mathbf{R}^d$. Let the true and the running value of the conditional read-out densities be positive, i.e. $b^i(y), b^{*i}(y) > 0$, respectively. Let the true and the* *running transition probability matrices Q^* and Q be primitive, and for all $i, j \in \mathcal{X}$*

$$\int\limits_{\mathbf{R}^d} \frac{\max\limits_{j\in\mathcal{X}} b^j(y)}{\min\limits_{j\in\mathcal{X}} b^j(y)} b^{*i}(y)\lambda(dy) < \infty, \tag{8}$$

and

$$\int\limits_{\mathbf{R}^d} |\log b^j(y)| b^{*i}(y)\lambda(dy) < \infty \tag{9}$$

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

Geometric ergodicity

We compare the results of Theorem 3 and Theorem 4. In Theorem 4 only the weaker primitivity condition is assumed for the transition probability matrix and condition (7) is relaxed to condition (9) where the exponent q becomes 1. On the other hand in Theorem 3 condition (8), which is a type of Lipschitz-condition is not needed. To see that (8) is a type of Lipscitz-condition consider the following inequality for an arbitrary fix $y \in \mathcal{Y}$:

$$\|\frac{\partial g(y,p)}{\partial p}\| = \|\frac{1}{\sum\limits_j b^j(y)p^j}(b^1(y), \dots b^N(y))^T\| \le \tag{10}$$

$$\frac{\sqrt{N}\max\limits_i b^i(y)}{\sum\limits_j b^j(y)p^j} \le \sqrt{N}\frac{\max\limits_i b^i(y)}{\min\limits_i b^i(y)}. \tag{11}$$

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 (6) is not a strong condition.

The main advantage of Theorem 3 introduced in [6] is that, it is potentially useful for deriving strong approximation results, see [5], which are in turn applicable to analyze adaptive predictors.

Consider a Hidden Markov Model which satisfies the conditions of Theorem 3. Let the parameters of the model be the elements of the transition probability matrix and the parameters of the read-out densities, say $b^i(y) = f(y, \theta^i)$, $\theta^i \in \Theta$. We will use the following identifiability condition, see [9]:

Identifiability Condition: The family of mixtures of $f(y, \theta^i)$ is identifiable, i.e.

$$\sum_{j=1}^N \alpha_j f(y, \theta^j) = \sum_{j=1}^N \alpha' f(y, \theta^{'j}) \ \lambda\text{-a.e.}$$

implies

$$\sum_{j=1}^N \alpha_j \delta_{\theta^j} = \sum_{j=1}^N \alpha'_j \delta_{\theta'^j},$$

where δ_θ denotes the distribution function of a point mass at θ.

Theorem 5. *Consider a Hidden Markov Model under the conditions of Theorem 3. Let the parameters of the model be the elements of the transition probability matrix and the parameters of the read-out densities. Assume that the identifiability condition above is satisfied Let $\hat{\theta}_N$ be the ML estimate of θ^*. Then*

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

where R^* is the Fisher-information matrix and $r_N = O_M(N^{-1})$ means that $r_N N$ is M-bounded, see the Appendix.

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

5 Gaussian read-outs

Consider the most commonly used example, the conditionally Gaussian observations. Let the read-out space be **R**. Assume that the process (X_n) satisfies the Doeblin-condition with $m = 1$ and let the true and the running value of the transition probability matrix be positive, i.e. $Q, Q^* > 0$. Let the read-outs be continuous with normal density functions, i.e.

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

where (m_i, σ_i)-s are the parameters. Assume that $\sigma_1 \leq \ldots \leq \sigma_N$. Let denote the true parameter by (m_i^*, σ_i^*). Since $\log b^i(y)$ is quadratic in y, (7) is satisfied as the momentums of the normal distribution exist. Hence Theorem 3 is applicable, and the limit of the log-likelihood function (5) exists and Theorem 5 is applicable.

On the other hand Condition (8) of Theorem 4 may not be satisfied if $\sigma_1 < \sigma_N$. For large y-s the integrand of (8) is

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

where C is a constant, which is integrable only if

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

for all i, i.e. if

$$\sigma_i^* > \frac{(\sigma_1 \sigma_N)^2}{(\sigma_N)^2 - (\sigma_1)^2}, \tag{12}$$

6 Appendix

In this section we summarize some basic definitions of [4].

Definition 3. *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 4. *A stochastic process* (X_n) *taking its values in a finite-dimensional Euclidean space is* L*-mixing, if it is* M*-bounded and with*

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

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

holds.

A simple application of L-mixing processes in the context of Markovian processes is given by the following result, see [6]

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

7 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. Gerencsér. On a class of mixing processes. *Stochastics*, 26:165–191, 1989.

5. L. Gerencsér. On the martingale approximation of the estimation error of ARMA parameters. *Systems & Control Letters*, 15:417–423, 1990.
6. 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.
7. 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.
8. F. LeGland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems*, 13:63–93, 2000.
9. B.G. Leroux. Maximum-likelihood estimation for hidden Markov-models. *Stochastic Processes and their Applications*, 40:127–143, 1992.
10. S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.
11. J. H. van Schuppen. Lecture notes on stochastic systems. Technical report. Manuscript.