Convergence Rates of the Maximum Likelihood Estimator of Hidden Markov Models

Laurent Mevel
IRISA / INRIA
Campus de Beaulieu
35042 Rennes Cédex, France
e–mail : lmevel@irisa.fr

Lorenzo Finesso
Inst. of Systems Science and Biomed. Eng.
LADSEB-CNR
Corso Stati Uniti, 4, 35127 Padova, Italy
e–mail : finesso@ladseb.pd.cnr.it

Abstract

In this paper we derive the almost sure rate of convergence of the Maximum Likelihood estimator of the parameters of a Hidden Markov Model with continuous observations and finite state space. The analysis is based on the geometric ergodicity properties of the prediction filter and its derivatives. As an example of application of these results we prove that, also in this context, the likelihood ratio is a consistent statistic for model selection.

1 Introduction

In this paper we consider the problem of the identification of a partially observed finite state Markov chain (or Hidden Markov Model, HMM), with observations in $\mathbf{R}^d$. Maximum Likelihood (ML) is the most popular approach to parameter estimation for this class of models. The asymptotic properties of the ML estimator (MLE) have already been investigated under a variety of conditions. Under the assumption of stationarity, Leroux [6] has proved the almost sure consistency of the MLE, and Bickel, Ritov and Ryden [10] have proved its asymptotic normality. More general results, encompassing both previous ones, have been given by Mevel in [4], where a new technique for the study of the convergence of HMM's is developed. The new technique is based on geometric ergodicity properties of the prediction filter and its derivatives [3], derived via results for products of random matrices [2]. The main advantage is that convergence results for Hidden Markov Models can now be reduced to the analysis of a Markov process, with a properly defined state space. In this paper we apply the new technique to derive the almost sure rate of convergence of the MLE and give an example of application in the context of the model selection problem. As will be mentioned in the paper, these results extend easily to conditional least squares estimators.

2 Statistical model

The statistical model is a parametric class of HMM's defined as follows. Let $\{X_n\,,\, n \geq 0\}$ and $\{Y_n\,,\, n \geq 0\}$ be two sequences, defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, with values in the finite set $S = \{1, \cdots, N\}$ and $\mathbf{R}^d$ respectively. On the space $(\Omega, \mathcal{F})$ we consider a family $(\mathbf{P}^\theta\,,\, \theta \in \Theta)$ of probability measures, with Θ compact subset of $\mathbf{R}^p$, such that under $\mathbf{P}^\theta$:

- The unobserved state sequence $\{X_n\,,\, n \geq 0\}$ is a Markov chain with transition probability matrix $Q_\theta = (q_\theta^{i,j})$, i.e. for any $i, j \in S$

  $$q_\theta^{i,j} = \mathbf{P}^\theta[X_{n+1} = j \mid X_n = i]\ ,$$

  and initial probability distribution $\pi_0 = (\pi_0^i)$ independent of $\theta \in \Theta$, and possibly different of the *true* initial probability distribution $\pi_\bullet = (\pi_\bullet^i)$ of X_0, i.e. for any $i \in S$

  $$\pi_0^i \stackrel{\triangle}{=} \mathbf{P}^\theta[X_0 = i] \neq \mathbf{P}[X_0 = i] = \pi_\bullet^i\ .$$

- For any $n \geq 0$, and $i \in S$, the conditional probability distribution of the observation Y_n given $\{X_n = i\}$ is absolutely continuous with respect to a unique σ–finite measure λ, positive on $\mathbf{R}^d$, i.e. for any $i \in S$,

  $$\mathbf{P}^\theta[Y_n \in dy \mid X_n = i] = b_\theta^i(y)\ \lambda(dy)\ ,$$

  with $b_\theta^i(y)$ a λ–a.e. positive density .

- The observations $\{Y_n\,,\, n \geq 0\}$ are mutually independent given the sequence of states of the Markov chain.

For future reference, for any $y \in \mathbf{R}^d$, define

$$b_\theta(y) \stackrel{\triangle}{=} [\,b_\theta^1(y), \cdots, b_\theta^N(y)\,]^*\ ,$$

$$B_\theta(y) \stackrel{\triangle}{=} \mathrm{diag}[b_\theta^1(y), \cdots, b_\theta^N(y)]\ .$$

Here and in the sequel $*$ denotes the transpose operator.

Example 2.1 [Conditionally Gaussian observations] Assume that for any $\theta \in \Theta$, $n \geq 0$, the observations are of the form

$$Y_n = h_\theta(X_n) + V_n^\theta \ ,$$

where $\{V_n^\theta \ , \ n \geq 0\}$ is a Gaussian white noise sequence, under $\mathbf{P}^\theta$, with identity covariance matrix. The mapping h_θ from S to $\mathbf{R}^d$ is equivalently defined $h_\theta = (h_\theta^i)$ where $h_\theta^i \in \mathbf{R}^d$ for all $i \in S$. In this case, the mutual independence condition is satisfied.

The following set of regularity assumptions on the transition probability matrix and the observation densities will be required :

Assumption A : For the *true* value $\alpha \in \Theta$, the transition probability matrix $Q_\alpha = (q_\alpha^{ij})$ is positive.

Hence, it exists $\varepsilon > 0$, possibly unknown, such that $q_\alpha^{i,j} \geq \varepsilon$, for all $i, j \in S$. Define

$$\Theta_\varepsilon = \{\theta \in \Theta, \, d(\theta, \partial\Theta) \geq \varepsilon\} \ .$$

Assumption A' : The mapping $\theta \to Q_\theta$, and its first four derivatives are Lipschitz continuous.

Assumption B : For the *true* value α of the parameter, for any $k, l = 1, \cdots, p$, and any $i \in S$

$$\int_{\mathbf{R}^d} \partial_k \, b_\alpha^i(y) \, \lambda(dy) = 0, \quad \int_{\mathbf{R}^d} \partial_{k,l}^2 \, b_\alpha^i(y) \, \lambda(dy) = 0 \ ,$$

Assumption B' : For any $y \in \mathbf{R}^d$, the mapping $\theta \to b_\theta(y)$ is four times differentiable.

Remark 2.2 Consider the model of Example 2.1, then Assumption B is satisfied.

Define now some quantities that will be needed later.

Definition 2.3 *For any $y \in \mathbf{R}^d$, and $\theta \in \Theta$, define*

$$\delta_\theta^{(s)}(y) = \frac{\max\limits_{i \in S} \max\limits_{k_0, \cdots, k_s = 1, \cdots, p} |\partial_{k_0, \cdots, k_s}^s \, b_\theta^i(y)|}{\min\limits_{i \in S} b_\theta^i(y)} \ .$$

For any $p \geq 1$, $s \geq 0$, let

$$\Delta_p^{(s)} = \max_{i \in S} \int_{\mathbf{R}^d} \max_{\theta \in \Theta} [\delta_\theta^{(s)}(y)]^p \, b_\alpha^i(y) \, \lambda(dy)$$

$$\Gamma_p = \max_{i \in S} \max_{\theta \in \Theta} \int_{\mathbf{R}^d} [\max_{j \in S} |\log b_\theta^j(y)|]^p b_\alpha^i(y) \, \lambda(dy) \ .$$

3 Consistency of the MLE

We collect here some results of [4], on the consistency of the MLE, which will be used later. For all $n \geq 1$, let $p_n^\bullet = (p_n^i)$ denote the *prediction filter*, i.e. the conditional probability distribution under $\mathbf{P}$, characterized by $(Q_\alpha, b_\alpha, \pi_\bullet)$, the *true* measure, of the state X_n given the observations $(Y_0, \cdots, Y_{n-1})$:

$$p_n^i = \mathbf{P}[X_n = i \mid Y_0, \cdots, Y_{n-1}] \ .$$

The random sequence $\{p_n^\bullet \ , \ n \geq 0\}$ takes values in the set $\mathcal{P}(S)$ of probability distributions over the finite set S, and satisfies the forward Baum equation

$$p_{n+1}^\bullet = \frac{Q_\alpha^* B_\alpha(Y_n) \, p_n^\bullet}{b_\alpha^*(Y_n) \, p_n^\bullet} \ , \tag{1}$$

for all $n \geq 0$, with $p_0^\bullet$ the initial probability distribution of X_0. We also define, for any $\theta \in \Theta$, and $n \geq 0$, the *approximate prediction filter*,

$$p_{n+1}^\theta = \frac{Q_\theta^* B_\theta(Y_n) \, p_n^\theta}{b_\theta^*(Y_n) \, p_n^\theta} \stackrel{\triangle}{=} f_\theta[Y_n, p_n^\theta] \ , \tag{2}$$

where $Q_\theta = (q_\theta^{i,j})$ is the stochastic matrix generated by θ. The first and second derivatives of the approximate prediction filter (2) w.r.t. θ are defined next.

$$\partial_k p_{n+1}^\theta = \Phi_\theta[Y_n, p_n^\theta] \partial_k p_n^\theta + u_\theta^k[Y_n, p_n^\theta] \ ,$$

$$\partial_{k,l}^2 \, p_{n+1}^\theta = \Phi_\theta[Y_n, p_n^\theta] \partial_{k,l}^2 \, p_n^\theta + U_\theta^{k,l}[Y_n, p_n^\theta, \partial \, p_n^\theta] \ ,$$

We also denote $\partial \, p_n^\theta = (\partial_k \, p_n^\theta)$ and $\partial^2 \, p_n^\theta = (\partial_{k,l}^2 \, p_n^\theta)$.

Remark 3.1 The terms $u_\theta^k[y, p]$ and $U_\theta^{k,l}[Y, p, w]$ are assumed to be Lipschitz continuous w.r.t. (p, w).

Let

$$Z_n^\theta \stackrel{\triangle}{=} \{X_n, Y_n, p_n^\theta, \partial \, p_n^\theta, \partial^2 \, p_n^\theta\} \ .$$

The process $\{Z_n^\theta \ , \ n \geq 0\}$ is an extended Markov chain with values in $T \stackrel{\triangle}{=} S \times \mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p \times \Sigma^{p \times p}$, where $\Sigma = \{w \in \mathbf{R}^N : e^* w = 0\}$ and $e = (1, \cdots, 1)^*$.

Definition 3.2 *Let L denote the set of functions $g = (g_{k,l}^i)$ defined on T such that for any $i \in S$, any $k, l = 1, \cdots, p$, and any $y \in \mathbf{R}^d$, the partial mapping $(p, w, W) \mapsto g_{k,l}^i(y, p, w, W)$ is locally Lipschitz continuous, i.e. if $u = (w, W)$,*

$$|g_{k,l}^i(y, p, u) - g_{k,l}^i(y, p', u')| \leq$$
$$\leq \mathrm{Lip}(g^i, y) \, [\, \|u - u'\| (1 + \|u\| + \|u'\|)$$
$$+ \|p - p'\| [\, (1 + \|u\| + \|u'\|)^2 \,]$$

such that

$$|g_{k,l}^i(y, p, u)| \leq \mathrm{K}(g^i, y) \, [\, (1 + \|w\|)^2 + \|W\| \,] \ ,$$

and such that

$$\mathrm{Lip}(g) = \max_{i \in S} \int \mathrm{Lip}(g^i, y) \, b_\alpha^i(y) \, \lambda(dy) < \infty \ ,$$

$$\mathrm{K}(g) = \max_{i \in S} \int \mathrm{K}(g^i, y) \, b_\alpha^i(y) \, \lambda(dy) < \infty \ .$$

Following LeGland and Mevel [3], the paper will be based on the following geometric ergodicity result :

Proposition 3.3 *Under Assumptions A, A', B, B', if* $\Delta_6^{(0)}$, $\Delta_4^{(1)}$, *and* $\Delta_2^{(2)}$ *are finite, then, under the true probability measure* $\mathbf{P}$, $\{Z_n^\theta,\, n \geq 0\}$ *has a unique invariant probability distribution* $\mu_\theta = (\mu_\theta^i)$ *on* T. *In particular, there exists some finite constants* C, $0 < \rho < 1$, *such that, for any function* $g = (g_{k,l}^i)$ *in* L, *any* $\theta \in \Theta_\varepsilon$, *any* $z \in T$, *and any* $k, l = 1, \cdots, p$, *we prove the geometric convergence of the* n*–th transition probability matrix/kernel of* Z_n^θ *under* $\mathbf{P}$

$$|\Pi_\theta^n\, g_{k,l}(z) - \lambda| \leq C[\,\mathrm{Lip}(g) + \mathrm{K}(g)\,]\frac{\rho^n}{1-\rho}\ ,$$

where the constant λ *is defined*

$$\lambda = \sum_{i \in S} \int g_{k,l}^i(y,p,w,W)\, \mu_\theta^i(dy,dp,dw,dW)\ ,$$

and there exists a unique solution V_θ, *defined on* T, *of the* Poisson equation

$$[\, I - \Pi_\theta\,]\, V_\theta(z) = g_{k,l}(z) - \lambda\ .$$

By definition, the log–likelihood function (suitably normalized) based on observations $(Y_0, \cdots, Y_n)$ is

$$\ell_n(\theta) = \frac{1}{n+1}\ \log p^\theta[Y_n, \cdots, Y_0]\ ,$$

Notice that the log–likelihood function can be expressed a sum of terms depending on the observations and the prediction filter: for any $\theta \in \Theta$

$$\ell_n(\theta) = \frac{1}{n+1}\ \sum_{k=0}^{n} \log[\, b_\theta^*(Y_k)\, p_k^\theta\,].$$

The following strong law of large numbers holds :

Proposition 3.4 *Under Assumption A, if* $\Delta_1^{(0)}$ *and* Γ_1 *are finite, then for any* $\theta \in \Theta$ *there exists a finite constant* $\ell(\theta)$ *such that*

$$\ell_n(\theta) \longrightarrow \ell(\theta)\ , \qquad \mathbf{P}\text{–}a.s.$$

as $n \to \infty$, *where*

$$\ell(\theta) = \int_{\mathcal{P}(S)\times\mathbf{R}^d} \log[\, b_\theta^*(y)\, p\,]\, \nu_\theta(dy,dp)\ ,$$

and ν_θ *denotes the marginal of* μ_θ *on* $\mathbf{R}^d \times \mathcal{P}(S)$.

Define the maximum likelihood estimator as

$$\widehat{\theta}_n \in \operatorname*{argmax}_{\theta \in \Theta_\varepsilon} \ell_n(\theta)\ ,$$

its properties have been investigated, via the Poisson equation approach, in Mevel [4]. Let

$$M(\alpha) \stackrel{\triangle}{=} \{\theta \in \Theta_\varepsilon\ : \ell(\theta) = \ell(\alpha)\}\ ,$$

be the set of global maxima of the function $\ell(\cdot)$.

The main consistency result for the MLE is

Theorem 3.5 *[4] Under Assumptions A, A', B', if* $\Delta_2^{(0)}$, $\Delta_1^{(1)}$, Γ_1, *and* Γ_2 *are finite, then any MLE sequence* $\widehat{\theta}_n$ *converges* $\mathbf{P}$*–almost surely to* $M(\alpha)$, *as* $n \to \infty$.

To further the investigation of the asymptotic properties of the MLE, we need to introduce some form of identifiability for the class of HMM's as follows. We say that a point θ in Θ_ε is identifiable modulo permutations (i.m.p.) if the only points θ' in Θ_ε with P_θ equivalent to $P_{\theta'}$ are obtained by simultaneous permutations of the rows and columns of the Q_θ matrix and of the elements of the family $b_\theta(\cdot)$ (corresponding to a renaming of the states of the Markov chain X_n).

Remark 3.6 A generic set (i.e. open, dense and of full measure) of i.m.p. points has been described in [7, page 99] in the case of discrete observations. For continuous observations, see Leroux [6].

4 Higher order asymptotics

We collect here some results of [4], on the normality of the score function and of the estimation error, which will be used later to derive the rate of convergence.

4.1 Asymptotics of the score function

For any $(y, p, w = (w_k))$ in $\mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p$, any $\theta \in \Theta$, and any $k = 1, \cdots, p$, define the *score* function by

$$H_\theta^k(y,p,w) = \frac{b_\theta^*(y)\, w_k}{b_\theta^*(y)p}\ +\ \frac{\partial_k b_\theta^*(y)\, p}{b_\theta^*(y)\, p}\ .$$

Define, for any $n \geq 1$, any $\theta \in \Theta$, and any $k = 1, \cdots, p$,

$$h^k(\theta) = \sum_{i \in S} \int H_\theta^k(y,p,w)\ \lambda_\theta^i(dy,dp,dw)\ ,$$

where $\lambda_\theta = (\lambda_\theta^i)$ denotes the marginal of μ_θ on $S \times \mathbf{R}^d \times \mathcal{P}(S) \times \Sigma^p$. Also define, for $k, l = 1, \cdots, p$,

$$I_\alpha^{k,l} = \sum_{i \in S} \int [H_\alpha^k(y,p,w)]^*\, [H_\alpha^l(y,p,w)]\, \lambda_\alpha^i(dy,dp,dw)\ .$$

Introduce the following assumptions

Assumption I : I_α is an invertible matrix.

Assumption R : $M(\alpha)$ is a set of isolated points. If in addition identifiability holds, then these points are all equal, up to a permutation of S, hence any MLE sequence $\widehat{\theta}_n$ converges to α.

The geometric ergodicity of the *approximate* prediction filter yields the following

Theorem 4.1 *[4] Under Assumptions A, A', B, B', R, and I, if $\Delta_8^{(0)}$, $\Delta_4^{(1)}$, $\Delta_2^{(2)}$, and $\Delta_1^{(3)}$ are finite, we have that $h(\alpha) = 0$ and*

$$\lim_{n\to\infty} \partial_k \ell_n(\theta) = h^k(\theta) = \partial_k \ell(\theta) \quad \mathbf{P}\text{–}a.s. \tag{3}$$

Moreover the following CLT holds

$$\sqrt{n}\, \partial \ell_n(\alpha) \implies \mathcal{N}(0, I_\alpha)$$

One important point is that this result holds for any derivatives of the likelihood, for any $i = 0, 1, 2$,

$$\sqrt{n}\left(\partial^i \ell_n(\alpha) - \partial^i \ell(\alpha)\right) \Longrightarrow \mathcal{N}(0, C_\alpha^i) \ .$$

4.2 Asymptotic normality of the MLE

Using Theorem 4.1, and Proposition 3.3, we prove

Proposition 4.2 *Under the assumptions of Theorem 4.1, for any $\theta \in \Theta_\varepsilon$,*

$$\partial^2 \ell_n(\theta) \longrightarrow \partial^2 \ell(\theta) \ , \qquad \mathbf{P}\text{–}a.s.,$$

and

$$\partial^2 \ell_n(\alpha) \longrightarrow -\, I_\alpha \ , \qquad \mathbf{P}\text{–}a.s.$$

Under the identifiability assumption R and invertibility of the Fisher matrix (assumption I), the following CLT for the estimation error holds.

Theorem 4.3 *[4] Under the assumptions of Theorem 4.1, $\widehat{\theta}_n$ is asymptotically normal, i.e.*

$$\sqrt{n}\,(\widehat{\theta}_n - \alpha) \Longrightarrow \mathcal{N}\left(0, I_\alpha^{-1}\right)$$

Remark 4.4 Invertibility of the Fisher matrix is not required for the convergence of the MLE. In case the Fisher matrix is not invertible, which can happen when the true parameter generates a primitive matrix Q_α, both Proposition 4.2 and Theorem 4.1 are still valid. It is then possible to prove, that

$$n(\theta_n - \alpha) R_n (\theta_n - \alpha)^* \Rightarrow \chi_q^2$$

where q is the dimension of Θ minus the number of zero eigenvalues of I_F, and R_n stands for the pseudo-inverse of $\partial^2 \ell_n(\theta_n)$.

5 The law of the iterated logarithm

Here and in the following sections, we present the main results of the paper. From now on Assumptions A, A', B, B', R and I will always be in effect, moreover we assume that there exist large enough p_0, p_1, p_2, p_3, p_4 such that

$$\Delta_{p_i}^{(i)} < \infty \ , \ i = 0, 1, 2, 3, 4 \ .$$

This section is devoted to establish that

Lemma 5.1 *For $i = 0, 1, 2$, for n large enough*

$$\partial^i \ell_n(\alpha) = \partial^i \ell(\alpha) + O_{a.s.}\left(\frac{\sqrt{\log\log n}}{\sqrt{n}}\right) \ .$$

Remark 5.1 The proof will be presented for $i = 0$. The extension to higher derivatives follows along the same lines. The result is valid also for values of $\theta \neq \alpha$ and it can be extended to mispecified cases where the true model does not belong to the HMM class, provided identifiability holds. In that case, α should be replaced by $\tilde{\theta}$ wich minimizes

$$\tilde{\theta} = \operatorname*{argmax}_{\theta\in\Theta_\varepsilon} \lim_{n\to\infty} \left(\frac{1}{n} \log \mathbf{P}[Y_1, \cdots, Y_n] - \ell_n(\theta)\right) \ ,$$

where $\mathbf{P}$ is the *true* measure generating the observations.

Proof: Recall that

$$\begin{aligned} \ell_n(\alpha) - \ell(\alpha) &= \frac{1}{n}\sum_{k=1}^n [I - \Pi_\alpha] V_k(Z_k(\alpha)) \\ &= \frac{1}{n}\sum_{k=1}^n W_k + Z_n \ , \end{aligned}$$

with Z_n going to zero as $O_{a.s.}(\frac{1}{n})$ and W_k being a martingale increment sequence. Call $S_n = \sum_{k=1}^n W_k$ and $U_n^2 = \sum_{k=1}^n W_k^2$ and $s_n^2 = \sum_{k=1}^n \mathbf{E}[W_k^2]$. Then $\frac{U_n^2}{n}$ is a uniformly integrable random variable. Almost surely, the following limits hold

$$\begin{aligned} \Sigma_\alpha &= \lim_{n\to\infty} \frac{U_n^2}{n} = \lim_{n\to\infty} \mathbf{E}\frac{U_n^2}{n} = \lim_{n\to\infty} \frac{s_n^2}{n} = \\ &= \int [\log[b_\alpha(y)\, p]^2 \nu_\alpha(dy, dp) - [\ell(\alpha)]^2 \end{aligned}$$

Now we check the tail condition of Corollary 4.2 of Hall and Heyde [9].

$$\begin{aligned} \mathbf{E}[|W_k| I_{[|W_k|>\varepsilon\, s_k]} &\leq \mathbf{E}[|W_k|^2]^{1/2} \mathbf{P}[|W_k| > \varepsilon\, s_k]^{1/2} \\ &\leq C_\varepsilon \left(\frac{\mathbf{E}[|W_k|]^\beta}{\varepsilon^\beta s_k^\beta}\right)^{1/2} \leq \frac{C'_\varepsilon}{s_k^{\beta/2}} \end{aligned}$$

Hence, provided $\beta > 2$, since $\lim_{n\to\infty} s_n^2/n = \Sigma_\alpha$,

$$\sum_k \frac{1}{s_k}\mathbf{E}[|W_k| I_{[|W_k|>\varepsilon\, s_k]}] < \infty \ ,$$

verifying the tail assumption. It follows that

$$\lim_{n\to\infty} \frac{S_n}{\sqrt{2s_n^2 \log\log s_n^2}} \in \{-1, 1\} \ ,$$

$$\text{thus } \ell_n(\alpha) - \ell(\alpha) = O_{a.s.}(\frac{\sqrt{\log\log n}}{\sqrt{n}}) \ ,$$

since s_n^2/n has a positive limit when n goes to infinity.

6 Almost sure rate of convergence of the MLE

The following uniform convergence lemma will be needed.

Lemma 6.1

$$\max_{\theta\in\Theta_\varepsilon} |\partial^3 \ell_n(\theta) - \partial^3 \ell(\theta)| \to 0 \ , \qquad \mathbf{P}\text{–}a.s. \ .$$

Proof: Using the same approach of Section 2, we can introduce

$$W_n^\theta \stackrel{\triangle}{=} (X_n, Y_n, p_n^\theta, \cdots, \partial^3 p_n^\theta) \ ,$$

and show that W_n^θ is a geometrically ergodic Markov process. The third derivative can be dealt with the technique developed in [4] since it follows the recursion rule

$$\partial^3 p_{n+1}^\theta = \phi_\theta[Y_n, p_n^\theta]\partial^3 p_n^\theta + R_n^\theta[Y_n, p_n^\theta, \cdots, \partial^2 p_n^\theta] \ ,$$

with $R_n[\cdots]$ Lipschitz-continous w.r.t. the filter derivatives. This recursion is analog to that of the second derivative therefore allowing a relatively easy extension of the method. From these considerations one can prove the first part of the Theorem i.e.

$$\partial^3 \ell_n(\theta) = \frac{1}{n}\sum_{k=1}^n F_\theta[Y_k, p_k^\theta, \cdots, \partial^3 p_k^\theta] \to \partial^3 \ell(\theta) \ , \mathbf{P}\text{–a.s.} \ .$$

The uniform convergence is extended from a result proved in Mevel and Finesso [5].

Theorem 6.1 *Under the previous assumptions, we have*

$$\widehat{\theta}_n - \alpha = O_{a.s.}(\frac{\sqrt{\log\log n}}{\sqrt{n}}) \ .$$

Proof: We only sketch the proof. The ideas behind it are quite well-known, and only their application to the HMM model is new. By Section 4, we have

$$\partial \ell_n(\alpha) = O_{a.s.}(\frac{\sqrt{\log\log n}}{\sqrt{n}}) \ .$$

$$\partial^2 \ell_n(\alpha) = I_\alpha + O_{a.s.}(\frac{\sqrt{\log\log n}}{\sqrt{n}}) \ ,$$

Under Assumption I, $\partial^2 \ell_n(\alpha)$ is positive definite for n large enough, and

$$\widehat{\theta}_n = \alpha + \partial^2 \ell_n(\alpha)^{-1} \left(\partial_n \ell_n(\alpha) + r_n\right) \ .$$

Lemma 6.1 leads to

$$r_n = (\widehat{\theta}_n - \alpha)\partial^3 \ell_n(\tilde{\theta})(\widehat{\theta}_n - \alpha)^* = O_{a.s.}(\|\widehat{\theta}_n - \alpha\|^2) \ .$$

From the previous results, we conclude that

$$\widehat{\theta}_n - \alpha = O_{a.s.}(\frac{\sqrt{\log\log n}}{\sqrt{n}}) \ .$$

Then, by a Taylor expansion procedure, we prove that

Corollary 6.1

$$\begin{aligned} \ell_n(\widehat{\theta}_n) - \ell_n(\alpha) &= O_{a.s.}(\frac{\log\log n}{n}) \\ \ell_n(\widehat{\theta}_n) - \ell(\alpha) &= O_{a.s.}(\frac{\sqrt{\log\log n}}{\sqrt{n}}) \ . \end{aligned}$$

The last result gives us the rate of convergence of the likelihood ratio.

Remark 6.2 In case of multiple maxima, the MLE is a mutivalued function. For n large enough, all these points are isolated and have the same convergence rate.

Remark 6.3 Straight application of the method gives us the rate of convergence for the conditional least squares estimator [4]

$$\theta_n = \operatorname*{argmax}_{\theta\in\Theta_\varepsilon} \frac{1}{2n}\sum_{k=1}^n |Y_k - \mathbf{E}^\theta[Y_k|Y_0, \cdots, Y_{k-1}]|^2 \ .$$

7 Likelihood ratio asymptotics

We consider here the problem of testing between two different models. The test statistic is the likelihood ratio defined as

$$\Delta_n \stackrel{\triangle}{=} \ell_n^1(\theta_n^1) - \ell_n^2(\theta_n^2) \ ,$$

where $\ell_n^j(\cdot)$ is the likelihood corresponding to the j-th model, $j \in \{1, 2\}$, and

$$\theta_n^j = \operatorname*{argmax}_{\theta\in\Theta_\varepsilon^j} \ell_n^j(\theta) \ .$$

Define, for $j \in \{1, 2\}$,

$$\ell^j(\theta) = \lim_{n\to\infty} \ell_n^j(\theta) = \int \log[b_\theta^* p]\nu_j(dy, dp) \ ,$$

where $\nu_j(\cdot,\cdot)$ is the marginal on $\mathbf{R}^d \times \mathcal{P}(S)$ of the invariant measure of $\{Y_n, p_n^\theta\}$, p_n^θ being defined with respect to model j. Define

$$\alpha^j = \operatorname*{argmax}_{\theta \in \Theta^j} \ell^j(\theta) \ ,$$

then, under identifiability assumptions, we have that $\theta_n^j \to \alpha^j$ **P**–a.s., and the following holds

Theorem 7.1

$$\Delta_n \to \Delta_{1,2} \stackrel{\triangle}{=} \ell^1(\alpha^1) - \ell^2(\alpha^2) \ , \qquad \mathbf{P}\text{–}a.s. \ ,$$

and

$$\sqrt{n^{-1}}\,(\Delta_n - \Delta_{1,2}) \Rightarrow \mathcal{N}(0, \Sigma_{1,2}) \ ,$$

where $\Sigma_{1,2}$ is well defined as

$$\Sigma_{1,2} \stackrel{\triangle}{=} \int \left(log[\frac{b_{\alpha^1}(y)p}{b_{\alpha^2}(y)q}] \right)^2 \nu_{1,2}(dy, dp, dq) - (\Delta_{1,2})^2 \ ,$$

and $\nu_{1,2}$ is the joint density, with marginals $\nu_j(\cdot,\cdot)$.

Proof: The proof is based on both Lemma 6.1 and Lemma 4.1 applied to the likelihood. First use a Taylor expansion to write

$$\begin{aligned} \ell_n^j(\theta_n^j) &= \ell_n^j(\alpha^j) + (\theta_n^j - \alpha^j)\partial\ell_n^j(\alpha^j) \\ &\quad + \frac{1}{2}(\theta_n^j - \alpha^j)^2[\partial^2\ell^j(\alpha^j) + \varepsilon_n^j] \ , \end{aligned}$$

with $\varepsilon_n^j \to 0$, **P**–a.s. by Lemma 6.1. Then, remark that

$$\sqrt{n}\left(\theta_n^j - \alpha^j\right) = O_{a.s.}(\frac{\log\log n}{\sqrt{n}})$$

by Theorem 6.1, and, using Lemma 5.1,

$$\sqrt{n}\left(\theta_n^j - \alpha^j\right)\partial\ell_n^j(\alpha^j) = O_{a.s.}(\frac{\log\log n}{\sqrt{n}}) \ .$$

Finally, by Theorem 4.1 applied to the likelihood,

$$\sqrt{n}(\ell_n^1(\alpha^1) - \ell_n^2(\alpha^2) - \Delta_{1,2}) \Rightarrow \mathcal{N}(0, \Sigma_{1,2}) \ .$$

Following the same lines of Nishii [8], Theorem 1, where the independent case was studied, we can now prove that

Corollary 7.2 *The likelihood ratio test Δ_n is a consistent statistic for deciding between two models.*

8 Conclusions

In this paper we have derived the rate of almost sure convergence of the MLE of general Hidden Markov Models and used the result to prove that the likelihood ratio is a consistent statistic for the model selection problem. The assumption of positivity of the *true* transition probability matrix Q_α can be relaxed to primitivity. The investigation of the mispecified case needs also to be furthered. This will appear in a forthcoming paper as well as all technical details, and other possible model selection applications.

References

[1] F. Le Gland, L. Mevel (1997). Asymptotic behaviour of the MLE in hidden Markov models. *Proceedings of the 4th European Control Conference*, Brussels, Paper FRA–F6, July 1997.

[2] F. Le Gland, L. Mevel (2000). Some properties of the projective product, with applications to products of column–allowable nonnegative matrices. *Mathematics of Control, Signals and Systems*, 13(1):41–62.

[3] F. Le Gland, L. Mevel (2000). Exponential forgetting and geometric ergodicity in HMM's. *Mathematics of Control, Signals and Systems*, 13(1):63–93.

[4] L. Mevel (1997). *Statistique Asymptotique pour les Modèles de Markov Cachés.* Thèse de doctorat, université de Rennes 1, novembre 1997.

[5] L. Mevel and L. Finesso (2000). Bayesian estimation of hidden Markov models. *Proceedings of the 14th Symposium on the Mathematical Theory of Networks and Systems*, Perpignan, July 2000.

[6] B.G. Leroux (1992). Maximum–likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127–143.

[7] T. Petrie (1969). Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115.

[8] R. Nishii (1988). ML principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27(2):392–403.

[9] P. Hall and C.C. Heyde (1980). *Martingale Limit Theory and its Application.* Academic Press, Inc.

[10] P.J. Bickel, Y. Ritov and T. Ryden (1998). Asymptotic normality of the MLE for general HMM. *The Annals of Statistics*, 26(4):1614–1635.