Asymptotics of the maximum likelihood estimator for general hidden Markov models

RANDAL DOUC¹ and CATHERINE MATIAS²

¹Dpt. TSI/CNRS URA 820, École Nationale Supérieure des Télécommunications, Paris, France. E-mail: douc@tsi.enst.fr
²CNRS UMR C 8628. Laboratoire de Probabilités, Statistique et Modélisation. Bât. 425. Université Paris Sud, 91405 Orsay Cedex, France. E-mail: Catherine.Matias@math.u-psud.fr

In this paper, we consider the consistency and asymptotic normality of the maximum likelihood estimator for a possibly non-stationary hidden Markov model where the hidden state space is a separable and compact space not necessarily finite, and both the transition kernel of the hidden chain and the conditional distribution of the observations depend on a parameter \( \theta \). For identifiable models, consistency and asymptotic normality of the maximum likelihood estimator are shown to follow from exponential memorylessness properties of the state prediction filter and geometric ergodicity of suitably extended Markov chains.

Keywords: asymptotic normality; consistency; geometric ergodicity; hidden Markov models; identifiability; maximum likelihood estimation

1. Introduction

Hidden Markov models (HMMs) form a wide class of discrete-time stochastic processes, used in different areas such as speech recognition (Juang and Rabiner 1991), neurophysiology (Fredkin and Rice 1987), biology (Churchill 1989), econometrics (Chib et al. 1998) and time-series analysis (De Jong and Shephard 1995; Chan and Ledolter 1995; see also MacDonald and Zucchini 1997 and the references therein).

Most work on maximum likelihood estimation in such models has focused on iterative numerical methods, suitable for approximating the maximum likelihood estimator. For finite hidden state-space models, the pioneering contribution is due to Baum et al. (1970) who presented an early and non-trivial application of the expectation–maximization principle (Dempster et al. 1977), known as the ‘forward–backward’ procedure. The more challenging issue of hidden Markov models with continuous state space was much studied throughout the 1990s, mostly using simulation-based approaches made possible by recent advances in Markov chain Monte Carlo methods (Chib et al. 1998; Durbin and Koopman 1997; De Jong and Shephard 1995; Chan and Ledolter 1995).

By contrast, the statistical issues regarding the asymptotic properties of the maximum likelihood estimator for hidden Markov models have until recently been largely ignored.
Baum and Petrie (1966) showed the consistency and asymptotic normality of the maximum likelihood estimator in the particular case where both the observed and the latent variables take only finitely many values. These results were recently extended in a series of papers by Leroux (1992), Bickel and Ritov (1996), Bickel et al. (1998) and Bakry et al. (1997). Bickel et al. generalized the method followed by Baum and Petrie (1966) to the case where the hidden Markov process $X_n$ takes a finite number of values, the observations belonging to a general topological space. Around the same time, LeGland and Mevel (2000a; 2000b) independently developed a different technique to prove the consistency and asymptotic normality of the maximum likelihood estimator (Mevel 1997) for hidden Markov models with finite hidden state space. The work of LeGland and Mevel (2000a; 2000b) is based on the remark that the likelihood can be expressed as an additive function of an extended Markov chain. The key to the proof of LeGland and Mevel (2000b) consists in showing that, under appropriate conditions, this extended chain is geometrically ergodic, in a sense given below. The proof applies even when the hidden Markov chain is non-stationary, that is, the initial distribution of the chain does not correspond to its invariant distribution, while Baum and Petrie (1966), Leroux (1992) and Bickel et al. (1998) assume stationarity (see, however, Bakry et al. 1997, for an extension to the non-stationary case, based on a coupling technique).

In this paper, we study maximum likelihood estimation when the hidden Markov chain takes values in a topological space, assumed to be compact. The main result is the consistency of the maximum likelihood estimator, under the usual condition of identifiability of the parameters. We also identify the limit of the likelihood process as the Kullback–Leibler divergence. In addition, we prove the asymptotic normality of the maximum likelihood estimator in both the stationary and the non-stationary case. To the best of our knowledge, the only contribution dealing with the maximum likelihood for hidden Markov models taking values in a general topological space (for the definition of Markov chains in such spaces, see Meyn and Tweedie 1993) is Jensen and Petersen (1999). The authors show the asymptotic normality (but not the consistency) of the maximum likelihood estimator when the hidden Markov chain is stationary, the technique being based on an extension of Bickel et al. (1998). The technique used in our contribution to prove the consistency and the asymptotic normality of the maximum likelihood estimator for general hidden Markov models is completely different and is essentially based on the technique developed by LeGland and Mevel (2000a; 2000b) in a finite state space. A distinctive advantage of this technique is that it makes it possible to study in a common framework the stationary and the non-stationary case. The main difficulty encountered when trying to use the ‘extended chain method’ for hidden Markov models with a continuous state space is that some components of the extended chain lie in an infinite-dimensional functional space. This is in sharp contrast to the case of a finite hidden state space in which the predictive density can always be identified with an element of $\mathbb{R}^M$, where $M$ is the number of different possible values for $X_n$. Instead of the Riesz representation theorem (Mevel 1997), we investigate another approach based on the asymptotic tightness of a sequence of measures to prove the existence of an invariant measure for the extended chain (Sunyach 1975). It is then possible to extend the result of Leroux (1992) under the basic assumption that the model is identifiable (in the sense that different parameter values lead to different
The maximum likelihood estimator for general hidden Markov models

2. Consistency of the maximum likelihood estimator

Let \( \{X_n\}_{n=0}^{\infty} \) be a Markov chain on a compact set \( K \subset \mathcal{X} \), where \( \mathcal{X} \) is a separable state space, equipped with a metrizable topology, and \( \mathcal{B}(K) \) is the associated Borel \( \sigma \)-field. We denote by \( \{Q_0(x, \cdot), x \in K, \mathcal{B} \in \mathcal{B}(K)\} \) the Markov transition kernel (see Meyn and Tweedie 1993) of the chain. We also let \( \{Y_n\}_{n=0}^{\infty} \) be a sequence of random variables taking values in a space \( \mathcal{Y} \) which is separable with respect to a topology that is metrizable by a complete metric. Conditional on \( \{X_n\}_{n=0}^{\infty} \), the \( Y_n \) are independent, with a conditional density \( g_0(y | X_n) \) dependent on \( X_n \) only. Here the density is with respect to a \( \sigma \)-finite measure \( \nu \) on the Borel \( \sigma \)-field \( \mathcal{B}(\mathcal{Y}) \). \( \mathcal{X} \) and \( \mathcal{Y} \) are usually subsets of \( \mathbb{R}^s \) and \( \mathbb{R}^t \) respectively, but they may also be higher-dimensional spaces. Moreover, both \( Q_0 \) and \( g_0 \) depend on a parameter \( \theta \) in \( \Theta \), where \( \Theta \) is a compact subset of \( \mathbb{R}^p \). \( \Theta \) is equipped with a norm denoted by \( \| \cdot \| \). The true parameter value will be denoted \( \theta^* \) and is assumed to be in the interior of \( \Theta \).

Assume that each transition kernel \( Q_0 \) has a density \( q_0 \) with respect to the same \( \sigma \)-finite dominating measure \( \gamma \) on \( \mathcal{X} \). For notational simplicity, it is assumed that the initial distribution of \( \{X_0\} \) under \( \theta^* \) has a density with respect to \( \gamma \) denoted by \( \pi^* \). We stress that \( \pi^* \) is not necessarily the invariant distribution of the chain. Denote by \( P^\theta \) the probability distribution of \( \{Y_n\}_{n=0}^{\infty} \) induced by the parameter \( (\theta^*, \pi^*) \), and let \( E^Y \) be the associated expectation. The convergence results that will be proved throughout this paper hold under \( P^\theta \). In the following, for \( m \leq n \), denote by \( Y_n^m \) the family of random variables \( (Y_n, \ldots, Y_m) \). The \( L_p(K, \mathcal{B}(K), \gamma) \)-norm will be denoted \( \| \cdot \|_p \), and \( C, C', \ldots \) will denote unspecified finite constants which may take different values upon each appearance. Moreover, for any measurable function \( f \) on \( (K, \mathcal{B}(K), \gamma) \), write \( \text{ess sup}(f) = \inf\{M \geq 0, \gamma\{M < |f|\} = 0\} \) and, if \( f \) is non-negative, \( \text{ess inf}(f) = \sup\{M \geq 0, \gamma\{M > f\} = 0\} \) (with obvious conventions if those sets are empty).

As outlined in Section 1, the central idea consists in writing the log-likelihood of the observations as an additive function of a Markov chain, comprising the observations \( \{Y_n\} \), the current state \( X_n \) of the initial Markov chain and the predictive density of the state, often referred to as the filter in the HMM literature.

distributions for the stationarized version of the observed process; see the definition below), a point covered by neither Mevel (1997) nor Jensen and Petersen (1999). We stress that the technique developed in Leroux (1992) to prove consistency, which is based on the Kingman subadditive theorem, cannot be immediately adapted to continuous state space. For technical reasons detailed below, we restrict our attention to cases where the hidden state space can be assumed to be compact.

The rest of the paper is organized as follows. In section 2, we first prove the geometric ergodicity of the extended chain. As a corollary, pointwise convergence of the normalized log-likelihood is obtained and its limit is identified with the generalized Kullback–Leibler divergence. The final part of Section 2 deals with the consistency of the maximum likelihood estimator. Section 3 is devoted to the asymptotic normality of the maximum likelihood estimator, using the technique of Section 2 applied to a further extended Markov chain, encompassing the gradient and the Hessian of the prediction filter.
We define a family of probability density functions (pdfs) \( \{ f_{\theta,n}^\xi \}_{n \geq 0} \) over \((K, \mathcal{B}(K), \gamma)\), by the recurrence relations
\[
f_{\theta,0}^\xi(v) = \xi(v),
\]
\[
f_{\theta,n+1}(v) = \Delta \Phi_1(Y_n, f_{\theta,n}^\xi; \theta) = \frac{\int u \, q_\theta(u, v) g_\theta(Y_n|u) f_{\theta,n}^\xi(u) d\gamma(u)}{\int u \, g_\theta(Y_n|u) f_{\theta,n}^\xi(u) d\gamma(u)},
\]
for all \( v \in K \), where \( \xi \) is an initial pdf. Note that \( f_{\theta,n+1}^\xi \) is a deterministic function of \((Y_n, f_{\theta,n}^\xi)\). If \( \xi = \pi^* \), it is easily checked that \( f_{\theta,n}^\pi^* \) is the predictive density distribution, that is, the density of the conditional distribution of \( X_n \) given \( Y_{n-1} \). Otherwise, \( f_{\theta,n}^\xi \) is usually called the prediction filter.

We consider the following family of contrast functions:
\[
l_n(\theta, \xi) = \sum_{m=0}^{n-1} \log \left( \int_K g_\theta(Y_n|x) f_{\theta,m}^\xi(x) d\gamma(x) \right).
\]
When \( \xi \) is fixed and equal to \( \pi^* \), then \( l_n(\theta, \pi^*) \) is the log-likelihood function for the parameter \( \theta \). We show below that the choice of the initial predictive density function \( \xi \) does not affect the limiting value of \((1/n) l_n(\theta, \xi)\). This parameter can thus be chosen by the user. Moreover, the parameter \( \theta^* \) is the only one that can be consistently estimated. As usual, for ergodic non-stationary chains, the initial distribution cannot be consistently estimated. We define the state space
\[
E = K \times Y \times S^+, \quad S^+ = \{ f \in L_1(K, \mathcal{B}(K), \gamma); \ f \geq 0, \| f \|_1 = 1 \},
\]
In the following, \( S^+ \) is equipped with the topology induced by the \( L_1 \)-norm, and \( \mathcal{B}(S^+) \) denotes the corresponding Borel \( \sigma \)-field. We denote by \( \mathcal{B}(E) \) the Borel \( \sigma \)-field on the space \( E \) itself, induced by the product topology. On the state space \( E \), define the extended Markov chain \( Z_n \equiv (X_n, Y_n, F_n, F_n^*) \) by an initial law \( \lambda \), and the following transition kernel: for all \((x, y, f, f^*)\) in \( E \) and \( \mathcal{A}_K \times \mathcal{A}_Y \times \mathcal{A}_F \times \mathcal{A}_F^* \) in \( \mathcal{B}(K) \times \mathcal{B}(Y) \times \mathcal{B}(S^+) \times \mathcal{B}(S^+) \),
\[
\Pi_\theta((x, y, f, f^*)) = P(X_{n+1} \in \mathcal{A}_K, Y_{n+1} \in \mathcal{A}_Y, F_{n+1} \in \mathcal{A}_F, F_n^* \in \mathcal{A}_F^*|X_n = x, Y_n = y, F_n = f, F_n^* = f^*)
\]
\[
= \int_{\mathcal{A}_K \times \mathcal{A}_Y} q_\theta^*(x, x') g_\theta^*(y'|x') d\gamma(x') d\nu(y') (\Phi_1(y, f; \theta))_{\mathcal{A}_F} (\Phi_1(y, f^*; \theta^*)_{\mathcal{A}_F^*}),
\]
where \( \Phi_1(y, f; \theta) \) is defined in (1).

In what follows, we will often consider the particular initial distribution \( \lambda(\xi) \) defined for each pdf \( \xi \) on \((K, \mathcal{B}(K), \gamma)\) by
\[
\lambda(\xi)(\mathcal{A}_K \times \mathcal{A}_Y \times \mathcal{A}_F \times \mathcal{A}_F^*) = \int_{\mathcal{A}_K \times \mathcal{A}_Y} \pi^*(x) g_\theta^*(y|x) d\gamma(x) d\nu(y) (\Phi_1(\xi, \cdot, \cdot, \cdot, \cdot))_{\mathcal{A}_F}(\Phi_1(\xi, \cdot, \cdot, \cdot, \cdot)_{\mathcal{A}_F^*}(\pi^*). \]
This initial distribution generates a Markov chain \( \{ Z_n \} \) where the first coordinate corresponds
to the Markov chain \( \{X_n\} \), the second to the observations \( \{Y_n\} \), the third to the prediction filter \( \{f_{\theta,n}^*\} \) and the fourth to the predictive density of \( X_n \) given \( Y_{n-1}^0 \), that is, \( \{f_{\theta,n}^*\} \).

Finally, we denote by \( P_{\theta,\lambda} \) the probability distribution of the chain \( \{Z_n\} \) induced by the initial distribution \( \lambda \) and the transition kernel \( \Pi_{\theta} \) (\( P_{\theta,\lambda} \) is classically defined on \( \bigotimes_{n=1}^{\infty} \mathcal{B}(E_n) \), where \( (E_n)_{n \geq 0} \) is a sequence of copies of \( E \)).

Using this definition, and assuming that \( \{Z_n\} \) has the initial distribution \( \lambda(\xi) \), the normalized log-likelihood may be expressed as

\[
\frac{1}{n} I_n(\theta, \xi) = \frac{1}{n} \sum_{m=0}^{n-1} h_\theta(Z_m) \quad P_{\theta,\lambda(\xi)} \text{ almost surely,}
\]

where

\[
h_\theta(x, y, f, f^*) \triangleq \log \left( \int_y g_\theta(y|u)f(u)dy(u) \right).
\]

Equation (5) shows that the normalized log-likelihood of the observations for the parameter \( \theta \) can be written as an additive functional of \( Z_n \). Thus, following ideas of LeGland and Mevel (2000a; 2000b), geometric ergodicity of the extended Markov chain will be the key ingredient in proving the consistency of the maximum likelihood estimator.

It is clear from (2) that the log-likelihood is indeed a function of the first three components of \( Z_n \), which themselves form a Markovian process. Extending the chain \( Z_n \) so as to include the prediction density corresponding to the actual value \( \theta_1^* \) of the parameter is, however, needed to obtain the consistency result of Section 2.4 below (for a closely related idea, see Leroux 1992).

The following definitions and assumptions are stated for future reference. For \( y \) in \( \mathcal{Y} \), define

\[
\delta(y) \triangleq \sup_{\theta \in \Theta} \delta_\theta(y) = \sup_{\theta \in \Theta} \left[ \frac{\text{ess sup}_x \ g_\theta(y|x)}{\text{ess inf}_x \ g_\theta(y|x)} \right],
\]

\[
\epsilon \triangleq \inf_{\theta \in \Theta} \frac{\text{ess inf}_{x,x'} q_\theta(x, x')}{\text{ess sup}_{x,x'} q_\theta(x, x')},
\]

Assumption 1. \( 0 < \epsilon < 1 \).

For \( s > 0 \), define \( \Delta_s \triangleq \text{ess sup}_x \int \text{sup}_\theta[\delta_\theta(y)]^s \int g_\theta(y|x) \nu(y) \).

Assumption 2. \( \Delta_1 \) is finite.

For \( s > 0 \), define \( \Gamma_s \triangleq \text{ess sup}_x \int \text{sup}_\theta[k_\theta(y)]^s \int g_\theta(y|x) \nu(y) \), where \( k_\theta(y) \triangleq \text{ess sup}_x \log g_\theta(y|x) \).

Assumption 3. There exists some \( s > 1 \) such that \( \Gamma_s \) is finite.

If Assumption 1 holds, then \( \inf_\theta \text{ess inf}_{x,x'} q_\theta(x, x') > 0 \) (note that \( \text{ess sup}_{x,x'} q_\theta(x, x') \geq \gamma(K)^{-1} \)). Doeblin’s condition is then satisfied for the Markov chain associated...
with \( Q_\theta \) (see Meyn and Tweedie 1993, p. 391). Moreover, the kernel \( Q_\theta \) is strongly aperiodic (see Meyn and Tweedie 1993, p. 118). Thus, by Theorems 16.0.2 and 16.2.4 in Meyn and Tweedie (1993, p. 384), there exists some real number \( 0 < \rho_0 < 1 \) (uniform in \( \theta \)), such that, for all \( \theta \) in \( \Theta \), there exists a probability measure on \( \mathcal{B}(K) \), having density \( \alpha_\theta \) with respect to \( \gamma \), such that, for all \( x \) in \( K \),

\[
\| Q_\theta^n(x, \cdot) - \alpha_\theta(\cdot) \|_{TV} \leq \rho_0^n,
\]

where \( \| \cdot \|_{TV} \) is the total variation norm, and \( Q_\theta^n \) is the kernel \( Q_\theta \) iterated \( n \) times. Then \( \alpha_\theta(\cdot) \) is the unique invariant probability measure of the chain. Thus, for all \( \theta \) in \( \Theta \), the Markov chain associated with the kernel \( Q_\theta \) is uniformly ergodic.

**Remark 1.** Note that we may relax \( \epsilon \in ]0, 1[ \) (Assumption 1) and derive the same results under the following:

**Assumption 1’.** There exists \( m \) in \( \mathbb{N}^* \) such that for all \( \theta \) in \( \Theta \), \( Q_\theta \) is aperiodic and

\[
0 < \inf_{\theta} \text{ess inf}_{x, x'} q^n_\theta(x, x') \leq 1,
\]

where \( q^n_\theta \) denotes the density of the kernel \( Q_\theta \) iterated \( m \) times.

In that case, all the properties concerning the Markov chain associated with \( Q_\theta \) are derived from properties concerning the \( m \)-skeleton (see the definition in Meyn and Tweedie 1993, p. 68).

### 2.1. Exponential memorylessness of the prediction filter

Ergodic properties for \( \{Y_n\} \) are inherited from those of \( \{X_n\} \) (see, for instance, Leroux 1992; Celeux et al. 1993). The point of this subsection is to obtain those of the extended chain \( Z_n = (X_n, Y_n, F_n, F_n^+) \) from the exponential memorylessness of the prediction filter. Iterating the recurrence relation (1) \( n - m + 1 \) times backwards \( (m \leq n) \) yields \( f_{\theta,m+1}^n = \Phi_{n-m+1}(Y_n^m, f_{\theta,m}^n; \theta) \), where

\[
\Phi_{n-m+1}(Y_n^m, f; \theta)(v) \triangleq \\
\int_{u_m} q_\theta(u_m, v) g_\theta(Y_n|u_n) \cdots \\
\int_{u_m} \int_{u_m} g_\theta(Y_n|u_n) \cdots q_\theta(u_m, u_{m+1}) g_\theta(Y_m|u_m) f(u_m) dy(u_n) \cdots dy(u_m).
\]

Using this expression, we may show the following exponential memorylessness inequality, which is a non-trivial adaptation from Mevel (1997, Proposition 2.2, part A).

**Proposition 1.** Suppose that Assumption 1 holds. Then for all \( f, f' \) in \( S^+ \),
The proof is given in Appendix A. The upper-bound of the right-hand side only depends on the initial conditions \((Y_m, f, f')\). Note that Assumption 1 implies that \(\epsilon_\theta\) is bounded from below by a strictly positive \(\epsilon\). Hence, \(1 - \epsilon_\theta\) is bounded from above by \(1 - \epsilon\) and the geometrical rate of the right-hand side of (10) can be bounded uniformly for \(\theta\) in \(\Theta\), provided that \(\delta(Y_m)\) is finite.

2.2. Geometric ergodicity of the extended Markov chain

In this section, it is shown that the exponential memorylessness of the prediction filter implies the geometric ergodicity (in a sense to be defined) of the extended chain \(\{Z_n\}\).

**Definition 1.** \(\text{Lip}(E)\) is the set of real-valued measurable functions \(h\) on \((E, \mathcal{B}(E))\) such that, for all \((x, y)\) in \(K \times \mathcal{Y}\), there exist \(\text{lip}(h, x, y)\) and \(k(h, x, y)\) such that, for all \(f_1, f_1^*, f_2, f_2^*\) in \(S^*\),

\[
|h(x, y, f_1, f_1^*) - h(x, y, f_2, f_2^*)| \leq \text{lip}(h, x, y)(\|f_1 - f_2\|_1 + \|f_1^* - f_2^*\|_1),
\]

and

\[
|h(x, y, f_1, f_1^*)| \leq k(h, x, y)
\]

and

\[
\text{lip}(h) \triangleq \text{ess sup}_x \int \text{lip}(h, x, y) g_{\theta^0}(y|x) d\nu(y) < \infty,
\]

\[
k(h) \triangleq \text{ess sup}_x \int k(h, x, y) g_{\theta^0}(y|x) d\nu(y) < \infty.
\]

These functions are of Lipschitz type with respect to the two last components and are bounded independently of the two last components by a function which satisfies an integrability condition. We will prove that, for all \(h\) in \(\text{Lip}(E)\) verifying some moment condition (see (15)), a law of large numbers holds for \(\{h(Z_n)\}_{n \geq 0}\).

**Proposition 2.** Suppose that Assumptions 1 and 2 hold, and define

\[
\rho \triangleq \max(\rho_0^{1/2}, (1 - \epsilon)^{1/2}),
\]

where \(\epsilon\) and \(\rho_0\) are as defined in (8) and (9). There exists a constant \(C > 0\) such that, for all \(\theta\) in \(\Theta\), for all \(h\) in \(\text{Lip}(E)\), for all \(z, z'\) in \(E\), and for all \(n \geq 1\),

\[
|\Pi_{\theta}^n h(z) - \Pi_{\theta}^n h(z')| \leq C[\text{lip}(h) + k(h)] \rho^n,
\]

where \(\Pi_{\theta}^n\) denotes the \(n\)th iterate of \(\Pi_{\theta}\) and \(\Pi_{\theta}^n h(z) = \int \Pi_{\theta}^n(z, dw) h(w)\).

The proof of this proposition is given in Appendix B. As a consequence, the extended
Markov chain satisfies a geometric ergodicity property, stated in the following corollaries, also proved in Appendix B.

**Corollary 1.** Suppose that Assumptions 1 and 2 hold. Then there exists a constant $C > 0$ such that, for all $\theta$ in $\Theta$ and for all $h$ in Lip($E$), there exists $\Lambda_\theta(h) < \infty$ such that, for all $z$ in $E$ and for all $n \geq 1$,

$$|\Pi_\theta^n h(z) - \Lambda_\theta(h)| \leq C[\text{lip}(h) + k(h)] \frac{\rho^n}{1 - \rho}. \quad (12)$$

This inequality may be used in the following way: for any initial probability measure $\lambda$ on $(E, \mathcal{B}(E))$, and for any function $h$ in Lip($E$),

$$\Lambda_\theta(h) = \lim_{n \to \infty} E_{\theta, \lambda}(h(Z_n)). \quad (13)$$

In addition, for all $h$ in Lip($E$), there exists a unique solution $V$ of the Poisson equation

$$[I - \Pi_\theta]V(z) = h(z) - \Lambda_\theta(h).$$

**Corollary 2.** Suppose that Assumption 1 and 2 hold. The kernel $\Pi_\theta$ admits a unique invariant probability measure $\lambda_\theta$ on $(E, \mathcal{B}(E))$, and, for any function $h$ in Lip($E$),

$$\Lambda_\theta(h) = \int h(z)\lambda_\theta(dz). \quad (14)$$

The type of convergence in Corollary 1 does not coincide with the usual definition of uniform geometric ergodicity (see Meyn and Tweedie 1993, Theorem 16.0.2, p. 384) since our result applies here only to certain Lipschitz functions on $E$, whereas in the standard literature about Markov chains the ergodicity is expressed in terms of exponential decrease to zero of the total variation norm, implying that (12) applies to all measurable bounded functions.

Note that it is easy to show that $\Lambda_\theta(h)$ is a linear form. Nevertheless, it is not straightforward to prove that there exists a measure $\lambda_\theta$ on $E$ representing $\Lambda_\theta(h)$. This is because the Riesz representation theorem does not apply. To prove the existence of an invariant probability measure, we use a technique based on asymptotic tightness of an appropriately defined sequence of probability measures, exploiting the basic ideas of Sunyach (1975).

**Remark 2.** We will often make use of the following property. Let $E_n$ be a sequence of copies of the space $E$, let $\mathcal{A}$ belong to $\bigvee_{n \geq 0} \mathcal{B}(E_n)$ and let $\Upsilon$ a real-valued function on $\prod_{n \geq 0} E_n$, measurable with respect to $\bigvee_{n \geq 0} \mathcal{B}(E_n)$. Note that, due to the initial distribution $\lambda(\xi)$, $(F_n, F^*_n)$ is $P_{\theta, \lambda(\xi)}$ almost surely equal to $(f^\theta_{\theta,n}, f^{\theta^*}_{\theta,n})$ for all $n$ in $\mathbb{N}$. This implies that

$$\Upsilon(\{Z_n\}_{n \geq 0}) = \Upsilon(\{(X_n, Y_n, f^\theta_{\theta,n}, f^{\theta^*}_{\theta,n})\}_{n \geq 0}) \quad P_{\theta, \lambda(\xi)}\text{-a.s.}$$

Thus,

$$P_{\theta, \lambda(\xi)}(\Upsilon(\{Z_n\}_{n \geq 0}) \in \mathcal{A}) = P^*[\Upsilon(\{(X_n, Y_n, f^\theta_{\theta,n}, f^{\theta^*}_{\theta,n})\}_{n \geq 0}) \in \mathcal{A}]$$
since the marginal distribution of \( P_{\theta, \lambda(\zeta)} \) on events only depending on \( \{(X_n, Y_n)\}_{n \geq 0} \) is equal to \( P^* \).

**Proposition 3.** Suppose that Assumptions 1 and 2 hold. For any \( \theta \) in \( \Theta \), for any pdf \( \zeta \) on \((K, \mathcal{B}(K), \gamma)\), and for any function \( h \) in \( \text{Lip}(E) \) such that there exists \( s > 1 \) satisfying

\[
\text{ess sup}_x \int k^x(h, x, y) \gamma(y|x) \text{d}y(y) < \infty, \quad (15)
\]

the following convergence results hold:

(i) \( \lim_{n \to \infty} (1/n) \sum_{m=0}^{n-1} h(Z_m) = \Lambda_\theta(h) \quad P_{\theta, \lambda(\zeta)} - \text{a.s.} \);

(ii) \( \lim_{n \to \infty} (1/n) \sum_{m=0}^{n-1} h(X_m, Y_m, f_{\theta, m}^\zeta, f_{\theta, m}^\gamma) = \Lambda_\theta(h) \quad P^* - \text{a.s.} \)

The proof of this proposition (see Appendix C) is based on the ergodicity of \( \{Z_n\} \) and, more precisely, on the existence of a solution to the Poisson equation stated in Corollary 1 using a classical martingale decomposition method (the proof is analogous to that of Meyn and Tweedie 1993, Theorem 17.4.3). Note that the statement (ii) is a direct consequence of (i) using Remark 2.

### 2.3. Pointwise convergence

The following proposition ensures that, under some regularity conditions, the normalized log-likelihood of the observations converges almost everywhere to a finite function of \( \theta \), and identifies the limit function. It is a straightforward application of Proposition 3.

**Proposition 4.** Suppose that Assumptions 1–3 hold. Then, for any \( \theta \) in \( \Theta \), there exists a finite value \( l(\theta) \) such that, for any initial pdf \( \zeta \) on \((K, \mathcal{B}(K), \gamma)\),

\[
\lim_{n \to \infty} \frac{1}{n} l_n(\theta, \zeta) = l(\theta) \quad P^* - \text{a.s.}
\]

Moreover,

\[
l(\theta) = \int \log \left[ \int g_{\theta}(y)f(u) \text{d}y(u) \right] \lambda_\theta(\text{d}x, \text{d}y, \text{d}f, \text{d}f^*) \quad (16)
\]

Note that the limiting value \( l(\theta) \) does not depend on the initial distribution \( \lambda(\zeta) \). This means that the limiting value of \( l_n(\theta, \zeta)/n \) does not depend on the initial distribution of the chain \( \{X_n\} \) (which does not necessarily coincide with the stationary distribution of the chain, in contrast with Bickel et al. 1998 and Jensen and Petersen 1999) or on the initial value of the prediction filter \( \zeta \), which may be chosen arbitrarily. The limiting value \( l(\theta) \) is the integral of \( h_\theta \) with respect to the invariant measure of \( \{Z_n\} \).
2.4. Consistency of the maximum likelihood estimator

In order to prove the consistency of the maximum likelihood estimator, we need to check that the limit of the normalized log-likelihood is maximized only at \( \theta^* \), the true value of the parameter; that is, \( l(\theta) \leq l(\theta^*) \) with equality if and only if \( \theta = \theta^* \). Define

\[
g_{\theta,n+1}(y_{n}^{0}|X_{0}) \overset{\Delta}{=} g_{\theta}(y_{0}|X_{0}) q_{\theta}(X_{0}, u_{1}) g_{\theta}(y_{1}|u_{1}) \cdots q_{\theta}(u_{n-1}, u_{n}) g_{\theta}(y_{n}|u_{n}) d\gamma(u_{1}) \cdots d\gamma(u_{n}).
\]

(17)

the density under the parameter \( \theta \) of \( \{Y_{n}^{0}\} \) conditionally on \( X_{0} \) with respect to the measure \( \nu^\otimes(n+1) \).

We make the following assumption:

**Assumption 4.** The equality

\[
\int_{K \times \mathcal{A}_0 \times \cdots \times \mathcal{A}_p} g_{\theta,p+1}(y_{0}, \ldots, y_{p}|u_{0}) d\alpha_{\theta}(u_{0}) d\nu(y_{0}) \cdots d\nu(y_{p})
= \int_{K \times \mathcal{A}_0 \times \cdots \times \mathcal{A}_p} g_{\theta',p+1}(y_{0}, \ldots, y_{p}|u_{0}) d\alpha_{\theta'}(u_{0}) d\nu(y_{0}) \cdots d\nu(y_{p})
\]

(18)

holds for all non-negative \( p \) and all Borel sets \( \mathcal{A}_i \) in \( \mathcal{B}(\mathcal{F}) \), \( 0 \leq i \leq p \), if and only if \( \theta = \theta' \).

(Recall that \( \alpha_{\theta} \) is the invariant measure for the kernel \( q_{\theta} \).) This condition is equivalent to the following:

**Assumption 4'.** The equality

\[
\int q_{\theta}(u, u_{1}) \cdots q_{\theta}(u_{p-1}, u_{p}) g_{\theta}(y_{0}|u) \cdots g_{\theta}(y_{p}|u_{p}) d\alpha_{\theta}(u) d\gamma(u_{1}) \cdots d\gamma(u_{p})
= \int q_{\theta}(u, u_{1}) \cdots q_{\theta}(u_{p-1}, u_{p}) g_{\theta}(y_{0}|u) \cdots g_{\theta}(y_{p}|u_{p}) d\alpha_{\theta}(u) d\gamma(u_{1}) \cdots d\gamma(u_{p})
\]

almost everywhere with respect to the measure \( \nu^\otimes(p+1) \), for all non-negative \( p \), if and only if \( \theta = \theta' \).

Note that when the state space \( K \) is finite, Assumption 4 is implied by Condition 2 in Leroux (1992). In fact, Assumption 4 is the minimal assumption guaranteeing identifiability for the model.

**Remark 3.** In many models of interest, the parameter itself is identifiable only up to a permutation of states. It is most often possible to reparametrize the model to avoid this type of degeneracy. When it is not possible, one should define a slightly extended notion of identifiability, based on an equivalence relation, for example \( \theta \equiv \theta' \) if and only if \( \theta \) is equal...
to $\theta'$ up to a permutation of its components. The results presented below can be straightforwardly adapted in such a situation.

**Theorem 1.** Under Assumptions 1–4, $l(\theta) \equiv l(\theta^*)$ and $l(\theta) = l(\theta^*)$ if and only if $\theta = \theta^*$.

The key point in the proof is that the difference $l(\theta^*) - l(\theta)$ can be written as the expected difference between two functions under the same measure. This is possible since the fourth coordinate of $Z_n$ is the prediction filter associated with the true value of the parameter $\theta^*$. This is the main motivation for extending $(X_n, Y_n, F_n)$ to $(X_n, Y_n, F_n, F_n^*)$.

**Proof.** The first part of the proof follows an idea from Leroux (1992, Lemma 6), but the conclusion has to be adapted to cope with general state space. We first express $l(\theta)$ in a more tractable way.

Using Corollary 1,

$$2l(\theta) = \lim_n E_{\theta, \lambda(\theta)}(h_{\theta}(Z_n) + h_{\theta}(Z_{n+1})).$$

By direct application of the recurrence relation between $f_{\theta, n}$ and $f_{\theta, n+1}$, we have

$$h_{\theta}(Z_n) + h_{\theta}(Z_{n+1}) = \log \left[ g_{\theta, 2}(Y_n, Y_{n+1}|u)f_{\theta, n}(u)d\gamma(u) \right] \quad \text{P}_{\theta, \lambda(\theta)} \text{-a.s.}$$

where $g_{\theta, 2}$ is defined by (17). Then, using Remark 2,

$$2l(\theta) = \lim_n E^*(H_{\theta}(Y_n, Y_{n+1}, f_{\theta, n}^*)(|Y_{n-1}^0))$$

Let $G_\theta(x, y, f, f^*) \triangleq \int H_{\theta}(y_0, y_1, f)g_{\theta, 2}(y_0, y_1|u)f^*(u)d\nu(y_0)d\nu(y_1)d\gamma(u)$. It is easy to check that $G_\theta$ is in Lip$(E)$. Thus, by Corollary 1,

$$\Lambda_\theta(G_\theta) = \lim_n E_{\theta, \lambda(\theta)}(G_{\theta}(Z_n)).$$

Using the equality

$$E^*[H_{\theta}(Y_n, Y_{n+1}, f_{\theta, n}^*)(|Y_{n-1}^0)] = E^*[H_{\theta}(Y_n, Y_{n+1}, f_{\theta, n}^*, X_n, Y_{n-1}^0)|Y_{n-1}^0]$$

$$= G_{\theta}(X_n, Y_n, f_{\theta, n}, f_{\theta, n}^*)$$

combined with (19) and (20), yields that

$$2l(\theta) = \Lambda_\theta(G_{\theta})$$

$$= \log \left[ \int g_{\theta, 2}(y_0, y_1|u_0)f(u_0)d\gamma(u_0) \right]$$

$$\times g_{\theta, 2}(y_0, y_1|u)f^*(u)d\nu(y_0)d\nu(y_1)d\gamma(u)\lambda_\theta(K, \gamma, \gamma, df, df^*).$$
Define the function $h^*(x, y, f, f^*) = \log(\int [g_{\theta^*}(y|u)] f^*(u) dy(u))$. Replacing the function $h_0$ by $h^*$, it is easily checked in the same way that

$$2l(\theta^*) = \int \log \left[ \int g_{\theta^*, 2}(y_0, y_1|u_0) f^*(u_0) dy(u_0) \right] \times g_{\theta^*, 2}(y_0, y_1|u)f^*(u) dv(y_0) dv(y_1) dy(u) \lambda_0(K, \mathcal{Y}, S^+, df^*).$$

Hence,

$$2(l(\theta^*) - l(\theta)) = \int \left[ \int g_{\theta^*, 2}(y_0, y_1|u)f^*(u) dy(u) \right] \times \log \left[ \int g_{\theta^*, 2}(y_0, y_1|u)f^*(u) dy(u) \right] dv(y_0) dv(y_1) \lambda_0(K, \mathcal{Y}, df, df^*)$$

which is the integral of a Kullback–Leibler divergence and is thus non-negative. We now show that the equality $l(\theta) = l(\theta^*)$ implies that $\theta = \theta^*$ under the assumption of identifiability (Assumption 4). If $l(\theta) = l(\theta^*)$ the Kullback–Leibler divergence is null, and thus

$$\int g_{\theta^*, 2}(y_0, y_1|u)f^*(u) dy(u) = \int g_{\theta, 2}(y_0, y_1|u)f(u) dy(u)$$

almost everywhere with respect to the measure $dv(y_0) dv(y_1) \lambda_0(K, \mathcal{Y}, df, df^*)$. This technique can easily be generalized by computing the quantities $(m + 1)_r(\theta, \zeta)$ for each integer $m$, such that $l(\theta) = l(\theta^*)$ implies almost everywhere with respect to the measure $\lambda_0(K, \mathcal{Y}, df, df^*)$, that, for all $m \geq 1$,

$$\int g_{\theta^*, m+1}(y_0, \ldots, y_m|u)f^*(u) dy(u) = \int g_{\theta, m+1}(y_0, \ldots, y_m|u)f(u) dy(u)$$

$v(dy_0) \otimes \ldots \otimes v(dy_m)$-a.e. (22)

Now, let us prove that assumption 4 holds. Let $(f_0, f_0^*)$ be in $S^+ \times S^+$ such that property (22) is satisfied. Let $p \leq m$ and $h$ be any bounded, continuous function on $\mathcal{Y}_{p+1}$. Multiplying both sides of the equation by $h(y_{m-p}, \ldots, y_m)$, and then taking the integral with respect to $dv(y_0) \ldots dv(y_m)$, yields

$$\int F_\theta(u_{m-p}) q_{\theta^*, m-p}^*(u_0, u_{m-p}) f_0(u_0) dy(u_0) dy(u_{m-p}) = \int F_\theta(u_{m-p}) q_{\theta^*, m-p}^*(u_0, u_{m-p}) f_0^*(u_0) dy(u_0) dy(u_{m-p})$$

where, for any $\beta$,

$$F_\beta(u) \triangleq \int h(y_{m-p}, \ldots, y_m) g_{\beta, p+1}(y_{m-p}, \ldots, y_m|u) dv(y_{m-p}) \ldots dv(y_m),$$

which is continuous and bounded by $\text{ess sup}_{u \in \mathcal{Y}_{p+1}} |h(u)|$ on $K$. Now, using (9), we have, for any $u$ in $K$, 

...
The maximum likelihood estimator for general hidden Markov models

\[ \left| F_\theta(u')q_\theta^{m-p}(u, u')d\gamma(u') - F_\theta(u')d\alpha_\theta(u') \right| \leq \rho_0^{m-p}\times \text{ess sup}_{u \in \mathcal{Y}^{p+1}} |h(u)|. \]

This implies, by dominated convergence,

\[ \lim_{m \to \infty} \int F_\theta(u')q_\theta^{m-p}(u, u')f_0(u)d\gamma(u)d\gamma(u') = \int F_\theta(u')d\alpha_\theta(u'). \]

This result also holds with the parameter \( \theta^* \) instead of \( \theta \). Now, combining with (23) yields

\[ \int F_\theta(u')d\alpha_\theta(u') = \int F_{\theta^*}(u')d\alpha_{\theta^*}(u'). \]

Replacing \( F_\theta \) and \( F_{\theta^*} \) by their definition in (24) and knowing that \( h \) is any bounded continuous function in \( \mathcal{Y}^{p+1} \),

\[ \int g_{\theta, p+1}(y_0, \ldots, y_p|u_0)d\alpha_\theta(u_0) = \int g_{\theta^*, p+1}(y_0, \ldots, y_p|u_0)d\alpha_{\theta^*}(u_0) \]

almost everywhere with respect to the measure \( d\nu(y_0) \cdots d\nu(y_p) \), for any \( p \) in \( \mathbb{N} \). Thus \( \theta = \theta^* \) by Assumption 4. This completes the proof. \( \square \)

Let \( \hat{\theta}_n(\zeta) \triangleq \arg \max_\theta l_n(\theta, \zeta) \) be the maximum likelihood estimator for a fixed initial pdf \( \zeta \) (the existence of which is guaranteed under assumptions stated below). It has been proved in Proposition 4 and Theorem 1 that the normalized log-likelihood for a fixed \( \theta \) converges \( P^\times\)-a.s. to \( l(\theta) \), which is maximum at the point \( \theta^* \). Under additional assumptions on the regularity of some functions with respect to \( \theta \), uniformity (in a sense to be defined) in the convergence of the normalized log-likelihood may be shown, and thus, consistency can be stated, using Wald’s (1949) consistency proof for the \( M \)-estimator.

Define for all \( y \) in \( \mathcal{Y} \) and for \( \eta > 0 \), the expression \( \omega^\zeta(y, \eta) \triangleq \sup_{\theta-\theta' \leq \eta} \text{ess sup}_x |g_\theta(y|x) - g_{\theta'}(y|x)|. \) Let \( \eta > 0 \) and \( \theta, \theta' \in \Theta \) such that \( \|\theta - \theta'\| < \eta \). Then, for all pdfs \( \zeta \) on \( (K, \mathcal{B}(K), \gamma) \),

\[ l_n(\theta, \zeta) - l_n(\theta', \zeta) \]

\[ = \sum_{m=0}^{n-1} \log \left( 1 + \frac{[g_{\theta}(Y_m|u)f_{\theta, m}^{\zeta}(u) - g_{\theta}(Y_m|u)f_{\theta', m}^{\zeta}(u)]d\gamma(u)}{g_{\theta}(Y_m|u)f_{\theta', m}^{\zeta}(u)d\gamma(u)} \right) \]

\[ \leq \sum_{m=0}^{n-1} \left( \frac{|g_{\theta}(Y_m|u) - g_{\theta}(Y_m|u)|f_{\theta, m}^{\zeta}(u)d\gamma(u)}{g_{\theta}(Y_m|u)f_{\theta', m}^{\zeta}(u)d\gamma(u)} + \frac{g_{\theta}(Y_m|u)f_{\theta, m}^{\zeta}(u) - f_{\theta, m}^{\zeta}(u)d\gamma(u)}{g_{\theta}(Y_m|u)f_{\theta', m}^{\zeta}(u)d\gamma(u)} \right), \]

\[ |l_n(\theta, \zeta) - l_n(\theta', \zeta)| \leq 2 \sum_{m=0}^{n-1} \left( \frac{\omega^\zeta(Y_m, \eta)}{\inf_\theta \text{ess sup}_x g_\theta(Y_m|x)} + \delta(Y_m)\|f_{\theta, m}^{\zeta} - f_{\theta', m}^{\zeta}\|_1 \right). \]
\[ \delta'(y, \eta) \triangleq \left[ \frac{\omega^g(y, \eta)}{\inf_{\theta} \text{ess inf}_x g_{\theta}(y|x)} \right] \quad (26) \]

\[ \Delta'(\eta) \triangleq \text{ess sup}_x \int [\delta'(y, \eta)]^t g_{\theta^*}(y|x) d\nu(y), \quad (27) \]

\[ (\Delta\Delta')_{s,t}(\eta) \triangleq \text{ess sup}_x \int [\delta(y)]^s [\delta'(y, \eta)]^t g_{\theta^*}(y|x) d\nu(y), \quad (28) \]

\[ \omega^g(\eta) \triangleq \sup_{\|\theta - \theta^*\| \leq \eta} \int \text{ess sup}_u |q_{\theta}(u, v) - q_{\theta^*}(u, v)| d\gamma(v). \quad (29) \]

**Assumption 5.** For all \( y \) in \( \mathcal{Y} \),

\[ \lim_{\eta \to 0} \omega^g(\eta) = 0, \]

\[ \lim_{\eta \to 0} \omega^S(y, \eta) = 0. \]

**Assumption 6.** \( \Delta_2 \) is finite and, for some positive \( \eta_0 \), \( \Delta_1(\eta_0) \) and \( (\Delta\Delta')_{1,1}(\eta_0) \) are finite.

We prove in Appendix C the following lemma:

**Lemma 1.** Under Assumption 1,

\[ \sup_{\zeta \in \mathbb{S}^+} \|f_{\hat{\theta},m} - f_{\hat{\theta}',m}\|_1 \leq 2 \epsilon^{-1} \sum_{k=0}^{m-1} \delta(Y_{k+1}) \rho^{m-k} [\omega^g(\eta) + 2 \delta'(Y_k, \eta)] + [\omega^g(\eta) + 2 \delta'(Y_m, \eta)]. \quad (30) \]

Combining (25) with this lemma, we prove under Assumption 5 that \( \theta \mapsto l_n(\theta, \zeta) \) is continuous. The continuity of \( \theta \mapsto l(\theta) \) is then straightforward: the family of functions \( (\Pi_n^\theta h_{\theta}(z))_{n \geq 0} \) converges uniformly in \( \theta \) to \( l(\theta) \) as we have

\[ |\Pi_n^\theta h_{\theta}(z) - \Lambda_\theta(h_{\theta})| \leq C[\Lambda_1 + \Gamma_1] \frac{\rho^n}{1 - \rho}. \]

Moreover,

\[ \Pi_n^\theta h_{\theta}(z) = \int \log \left( \int g_{\theta}(y_n|u) f_{\hat{\theta},n}(u) d\gamma(u) \right) q_{\theta^*}(x_1) \cdots q_{\theta^*}(x_{n-1}, x_n) \]

\[ \times g_{\theta^*}(y_1|x_1) \cdots g_{\theta^*}(y_n|x_n) d\gamma(x_1) \cdots d\gamma(x_n) d\gamma(y_1) \cdots d\gamma(y_n) \]

The functions \( \theta \mapsto \log(\int g_{\theta}(y_n|u) f_{\hat{\theta},n}(u) d\gamma(u)) \) are continuous (by the same argument as for \( l_n(\theta, \zeta) \)) and uniformly integrable as \( \Gamma_s \) is finite for some \( s > 1 \). Then we obtain that \( \theta \mapsto \Pi_n^\theta h_{\theta}(z) \) is continuous, and so is \( \theta \mapsto l(\theta) \).
The maximum likelihood estimator for general hidden Markov models 395

Now write \( \omega_n^l(\eta) \hat{=} \sup \{(1/n)I_n(\theta, \zeta) - I_n(\theta', \zeta), \|\theta - \theta'\| < \eta\} \). Combining (25), (30) and Assumptions 5 and 6, we obtain that

\[
\lim \limsup_n E^* \left[ \omega_n^l \left( \frac{1}{k} \right) \right] = 0.
\]

Thus, using that

\[
P^* \left[ \omega_n^l \left( \frac{1}{k} \right) \geq \epsilon_k \right] \leq \frac{1}{\epsilon_k} E^* \left[ \omega_n^l \left( \frac{1}{k} \right) \right], \tag{31}
\]

it is easy to choose \( \epsilon_k \) decreasing to 0 such that

\[
\lim \limsup_n P^* \left[ \omega_n^l \left( \frac{1}{k} \right) \geq \epsilon_k \right] = 0.
\]

Then a straightforward adaptation of Dacunha-Castelle and Duflot (1986, Theorem 3.2.8, p. 93), shows that the maximum likelihood estimator is consistent.

**Theorem 2.** Suppose that Assumptions 1–6 hold. Then for all initial pdf \( \zeta \),

\[
\hat{\theta}_n(\zeta) \to \theta^* \quad \text{in } P^* \text{-probability).
\]

**2.5. Examples**

Let us give some illustrations.

**Example 1** Markov chain in additive noise.

Define \( Y_t = h(X_t) + B_t \), where \( \{B_t\} \) is a sequence of independent and identically distributed centred Gaussian random variables with variance \( \sigma^2 \), and \( \{X_t\} \) is a Markov chain on a compact set in \( \mathbb{R}^d \), independent of \( \{B_t\} \), and with transition \( Q_\alpha \). The parameter here is \( \theta = (\alpha, \sigma^2) \). Let \( \phi_t \) be the log-Laplace function of the marginal distributions of \( \{Y_t\} \),

\[
\phi_t(u) = \log(E(e^{uY_t})) = \log E(e^{uh(X_t)}) + \log \left( e^{u^2\sigma^2/2} \right),
\]

so that

\[
\frac{2}{u^2} \phi_t(u) = \sigma^2 + \frac{2}{u^2} \log(E(e^{uh(x)})) \to \sigma^2.
\]

\( \sigma^2 \) is thus identifiable. If different parameters \( \alpha \) lead to different stationary distributions for the process \( \{h(X_t)\} \), then Assumption 4 holds. Such models are used in the following applications:

- Underwater acoustics (see Quinn 1995), where \( h(x) = A \cos(2\pi x) \), and \( A \) is a constant.
- Speech processing, count data analysis, and so on (see Pagano 1974), where \( h(x) = x \) and \( \{X_t\} \) is an autoregressive process of first order, with bounded innovations, or more generally an ARMA process. In this last case, just note that Theorems 1 and 2 may be
generalized when Assumption 1 holds only for a power of the transition $Q_a$ — that is, when Assumption 1′ holds.

**Example 2** Markov chain in multiplicative noise.
Define $Y_t = B_t \cdot h(X_t)$, where $\{B_t\}$ is a process of independent and identically distributed, centred Gaussian random variables with variance $\sigma^2$, $\{X_t\}$ is a Markov chain on a compact subset of $\mathbb{R}$, independent of $\{B_t\}$, and $h$ is any measurable function. Such models are used in a number of different areas (see Chib et al. 1998).

### 3. Asymptotic normality of the maximum likelihood estimator

The asymptotic normality of the maximum likelihood estimator involves a second-order Taylor expansion of the log-likelihood. We shall use similar arguments to those for the extended chain $Z_n$, but we will now consider Markov chains involving both the gradient and the Hessian of the prediction filter. These chains have an exponential lack of memory of the initial condition, and we prove that they are uniformly ergodic. Denote by $\partial_k$ the first-order partial derivative with respect to the component $\theta_k$ of $\theta$, and by $\nabla$, $\nabla^2$ the Jacobian and Hessian operators with respect to the parameter $\theta$. Define, for all $y$ in $\mathcal{Y}$,

$$
\delta'(y) = \sup_{\theta, \theta'} \max_{1 \leq k \leq p} \left[ \text{ess sup}_x |\partial_k g_\theta(y|x)| \right]
$$

and, for all $s$, $t$ in $\mathbb{N}$,

$$
\Delta_s' = \text{ess sup}_x \int [\delta'(y)']^T g_{\theta,s}(y|x) dv(y),
$$

$$
(\Delta \Delta')_{s,t} = \text{ess sup}_x \int [\delta'(y)][\delta'(y)]^T g_{\theta,s}(y|x) dv(y).
$$

**Assumption 7.** (i) For any $(x, x', y)$ in $K \times K \times \mathcal{Y}$, the functions $\theta \mapsto g_\theta(y|x)$ and $\theta \mapsto q_\theta(x, x')$ are differentiable.

(ii) $\sup_{\theta} \text{ess sup}_{u, v} q_\theta(u, v) < \infty$, $\max_{1 \leq k \leq p} \sup_{\theta} \text{ess sup}_{u, v} |\partial_k q_\theta(u, v)| < \infty$. For some $s > 1$, $\Delta_{s, s}$, $\Delta_{s, t}$, $(\Delta \Delta')_{s, t}$, and $(\Delta \Delta')_{s, t}$, are finite.

The first step consists in differentiating the recurrence relation between $f_{\theta,n}^\beta$ and $f_{\theta,n+1}^\beta$ (which is possible under Assumption 7). Then $\{\nabla f_{\theta,n}^\beta\}$ appears as a functional autoregressive chain. For all $1 \leq k \leq p$,

$$
\partial_k f_{\theta,n+1}^\beta(v) = \int u a_\theta(Y_n, f_{\theta,n}^\beta(u, v)) \partial_k f_{\theta,n}^\beta(u) dv(u) + U_{0,k}(Y_n, f_{\theta,n}^\beta(v)),
$$

where
We denote by \( f \) responding expectation. The exponential ergodicity of the extended chain \( \hat{\Theta} \).

\[
\text{Denote by } U \text{ the vector function } (U_{\theta,k})_{1 \leq k \leq p}.
\]

We consider the space \( \mathcal{E}' = K \times \mathcal{Y} \times S^+ \times \Sigma \), where

\[
\Sigma \triangleq \left\{ \sigma = (\sigma_k)_{1 \leq k \leq p}; \ \sigma_k \in L_1(K, \mathcal{B}(K), \gamma); \quad \int \sigma_k(u)d\gamma(u) = 0 \right\}.
\]

The space \( \Sigma \) is equipped with the topology induced by the \( L_1 \)-norm, and denote by \( \mathcal{B}(\mathcal{E}') \) the Borel \( \sigma \)-field induced by the product topology on \( \mathcal{E}' \). For all \((y, f, \sigma) \in \mathcal{Y} \times S^+ \times \Sigma, \theta \in \Theta, \) and \( v \in K \), write

\[
\Psi_\theta(y, f, \sigma; \theta)(v) \triangleq \int a_\theta(y, f)(u, v)\sigma(u)d\gamma(u) + U_\theta(y, f)(v).
\]

(Note that \( \Psi_\theta = (\Psi_\theta^k)_{1 \leq k \leq p} \) is a vector in \( \mathbb{R}^p \).)

We define an extended Markov chain \( \hat{Z}_n = (X_n, Y_n, F_n, G_n) \) on \((\mathcal{E}', \mathcal{B}(\mathcal{E}'))\) by its initial law \( \hat{\lambda} \) and the transition kernel \( \hat{\Pi}_0 \) given, for all \((x, y, f, \sigma) \) in \( \mathcal{E}' \) and \( \mathcal{A}_K \times \mathcal{A}_\mathcal{Y} \times \mathcal{A}_f \times \mathcal{A}_\pm \) in \( \mathcal{B}(K) \times \mathcal{B}(\mathcal{Y}) \times \mathcal{B}(S^+) \times \mathcal{B}(\Sigma) \), by

\[
\hat{\Pi}_0((x, y, f, \sigma); \mathcal{A}_K \times \mathcal{A}_\mathcal{Y} \times \mathcal{A}_f \times \mathcal{A}_\pm) = P(X_{n+1} \in \mathcal{A}_K, Y_{n+1} \in \mathcal{A}_\mathcal{Y}, F_{n+1} \in \mathcal{A}_f, G_{n+1} \in \mathcal{A}_\pm|X_n = x, Y_n = y, F_n = f, G_n = \sigma)
\]

\[
= \int_{\mathcal{A}_K \times \mathcal{A}_\mathcal{Y}} q_{\theta^s}(x, x') g_{\theta^s}(y'|x')d\gamma(x')d\nu(y')^{\mathcal{A}_f}((\Phi_1(y, f; \theta))^{\mathcal{A}_\mathcal{Y}}(\Psi_\theta(y, f, \sigma; \theta)).
\]

In what follows, we will often consider the particular initial distribution \( \hat{\lambda}(\xi) \) defined for each pdf \( \xi \) on \((K, \mathcal{B}(K), \gamma)\) by

\[
\hat{\lambda}(\xi)(\mathcal{A}_K \times \mathcal{A}_\mathcal{Y} \times \mathcal{A}_f \times \mathcal{A}_\pm) = \int_{\mathcal{A}_K \times \mathcal{A}_\mathcal{Y}} \pi^s(x) g_{\theta^s}(y|x)d\gamma(x)d\nu(y)^{\mathcal{A}_f}(\xi)^{\mathcal{A}_\mathcal{Y}}(\theta^s) = (36).
\]

We denote by \( \hat{\Pi}_{\theta,\xi}(\xi) \) the law of \( \{\hat{Z}_n\} \) induced by \( \hat{\lambda}(\xi) \), and \( \hat{\Pi}_{\theta} \), by \( \hat{E}_{\theta,\hat{\lambda}(\xi)} \) the corresponding expectation. The exponential ergodicity of the extended chain \( \{\hat{Z}_n\} \) (similar to the result obtained in Proposition 1) and convergence of normalized sums of some functions of \( \{\hat{Z}_n\} \), are stated in Appendix D.

These properties, combined with the results of Propositions 6 and 7, ensure the convergence of the gradient of the normalized log-likelihood.
Assumption 8. For any \((x, x', y)\) in \(K \times K \times \mathcal{Y}\), the functions \(\theta \mapsto g_\theta(y|x)\) and \(\theta \mapsto q_\theta(x, x')\) are twice differentiable. We also assume that

\[
\sup_{\theta} \max_{1 \leq k, l \leq p} \text{ess sup}_{u, v} |\partial^2_{k, l} q_\theta(u, v)| < \infty.
\]

Define for all \(y \in \mathcal{Y}\), and for all \(s, t, r \in \mathbb{N}\),

\[
\delta''(y) = \sup_{\theta, \theta'} \max_{1 \leq k, l \leq p} \left[ \frac{\text{ess sup}_y |\partial^2_{k, l} g_\theta(y|x)|}{\text{ess sup}_y g_\theta(y|x)} \right],
\]

\[
\Delta_s^\triangleq \text{ess sup}_x \int [\delta''(y)]^r g_{\theta r}(y|x) dv(y),
\]

\[
(\Delta\delta''_{s,t})^\triangleq \text{ess sup}_x \int [\delta''(y)]^r [\delta''(y)]^t g_{\theta t}(y|x) dv(y),
\]

\[
(\Delta\Delta\delta''_{s,t,r})^\triangleq \text{ess sup}_x \int [\delta''(y)]^r [\delta''(y)]^t [\delta''(y)]^s g_{\theta s}(y|x) dv(y).
\]

Assumption 9. There exists some \(s > 1\) such that \(\Delta''_s\) is finite. Moreover, the quantities \(\Delta_{11}, (\Delta\delta''_{11,1}), (\Delta\delta''_{10,2}), (\Delta\delta''_{9,3}), (\Delta\delta''_{8,4}), (\Delta\delta''_{7,1}), (\Delta\delta''_{7,1,1})\) and \((\Delta\Delta\delta''_{6,2,1})\) are supposed to be finite.

Proposition 5. Under Assumptions 1 and 7–9 and the assumption that the maximum likelihood estimator is consistent (see Theorem 2), there exists a matrix \(J_{\theta^*}\) such that

\[
n^{-1/2} \nabla l_n(\theta^*, \xi) \rightarrow N(0, J_{\theta^*}) \text{ weakly under } P^*.
\]

Moreover, \(J_{\theta^*} = \lim_{n \rightarrow \infty} (1/n) E^* (\nabla l_n(\theta^*, \xi)' \nabla l_n(\theta^*, \xi))\).

This result extends Theorem 3.1 in Jensen and Petersen (1999) to the non-stationary case, and its proof is given in Section D.1. The second-order Taylor expansion of the log-likelihood also involves its Hessian matrix. Differentiating (35), we obtain \(\nabla^2 f_{\theta, n}^*\) as a functional autoregressive chain:

\[
\nabla^2 f_{\theta, n+1}^* \overset{\Delta}{=} \alpha_1 (Y_n, f_{\theta, n}^*, \nabla f_{\theta, n}^*, \nabla^2 f_{\theta, n}^*, \theta).
\]

As we have done with \(Z_n\) and \(\hat{Z}_n\) we now consider \(\hat{Z}_n = (X_n, Y_n, F_n, G_n, H_n)\), an extended Markov chain on a space \(E''\) defined by \(E'' = K \times \mathcal{Y} \times S^* \times \Sigma \times \Sigma_2\), where \(\Sigma_2 = \{\tau = (\tau_{k,l}); \tau_{k,l} \in L_1(K, \mathcal{B}(K), \gamma) \text{ and } \forall v \in K, \tau(v) \text{ is symmetric and non-negative}\}\).

(The matrix \(M = (M_{i,j})_{1 \leq i, j \leq n}\) is said to be non-negative if for all \((a_i)_{1 \leq i \leq n}\) in \(\mathbb{R}_+\), we have \(\sum_{1 \leq i, j \leq n} a_i M_{i,j} a_j \geq 0\), with equality if and only if \(a_i = 0\) for all \(1 \leq i \leq n\).) The space \(\Sigma_2\) is equipped with the topology induced by the \(L_1\)-norm and \(\mathcal{B}(E'')\) will denote the Borel \(\sigma\)-field on \(E''\) induced by the product topology. The Markov chain \(\hat{Z}_n\) is defined by its initial law \(\hat{\lambda}\).
and its transition kernel $\hat{\Pi}_0$: for all $(x, y, f, \sigma, \tau)$ in $\mathbb{E}^n$ and $\mathcal{H}_0 \times \mathcal{H}_y \times \mathcal{S} \times \mathcal{S}_e \times \mathcal{S}_e^s$, in $\mathcal{B}(K) \times \mathcal{B}(Y) \times \mathcal{B}(S) \times \mathcal{B}(\Sigma) \times \mathcal{B}(\Sigma_2)$,

$$\hat{\Pi}_0(x, y, f, \sigma, \tau) = P[\{(X_{n+1}, Y_{n+1}, F_{n+1}, G_{n+1}, H_{n+1}) \in \mathcal{H}_0 \times \mathcal{H}_y \times \mathcal{S} \times \mathcal{S}_e \times \mathcal{S}_e^s\} | (X_n, Y_n, F_n, G_n, H_n) = (x, y, f, \sigma, \tau)]$$

$$= \int_{\mathcal{H}_0 \times \mathcal{H}_y} q_{\theta^*}(x, x') g_{\theta^*}(y'|x') dy'(x') dv(y')$$

$$\times 1_{\mathcal{H}_f}(\Phi_1(y, f; \theta)) 1_{\mathcal{H}_g}(\Psi_1(y, f, \sigma; \theta)) 1_{\mathcal{H}_s}(a_1(y, f, \sigma, \tau; \theta)).$$

We will consider the particular initial distribution $\hat{\lambda}(\zeta)$ defined for each pdf $\zeta$ on $(K, \mathcal{B}(K), \gamma)$ by

$$\hat{\lambda}(\zeta)(\mathcal{H}_0 \times \mathcal{H}_y \times \mathcal{S} \times \mathcal{S}_e \times \mathcal{S}_e^s)$$

$$= \int_{\mathcal{H}_0 \times \mathcal{H}_y} \pi^*(x) g_{\theta^*}(y|x) dy' d\gamma(x) dx(y) 1_{\mathcal{H}_f}(\Phi_1(\zeta, 0)) 1_{\mathcal{H}_g}(\Psi_1(\zeta, 0)) 1_{\mathcal{H}_s}(0).$$

We denote by $\hat{P}_{\theta, \hat{\lambda}(\zeta)}$ the law of $\{Z_n\}_n$ induced by $\hat{\lambda}(\zeta)$ and $\hat{\Pi}_0$, and by $\hat{E}_{\theta, \hat{\lambda}(\zeta)}$ the corresponding expectation. In Section D.3, we prove the exponential lack of memory of the initial condition for this chain, its ergodicity, and the convergence of normalized sums of some functions of this chain. Then we are able to prove the convergence of the Hessian of the normalized log-likelihood.

**Proposition 6.** Under Assumptions 1 and 7–9, there exists a matrix $I_0$ such that

$$\frac{1}{n} \nabla^2 l_n(\theta, \zeta) \to -I_0 \quad \text{P*-a.s.}$$

Moreover, $I_{\theta^*}$ is the Fisher information matrix, that is,

$$I_{\theta^*} = -\lim_{n \to \infty} \frac{1}{n} \mathbb{E}^*(\nabla^2 l_n(\theta^*, \zeta)) = \lim_{n \to \infty} \frac{1}{n} \mathbb{E}^*(\nabla l_n(\theta^*, \zeta) \nabla l_n(\theta^*, \zeta)) = J_{\theta^*},$$

where superscript $T$ denotes matrix transposition.

Write, for all $\eta > 0$ and $y$ in $\mathcal{Y}$,
\[
\omega^\alpha(\eta) \triangleq \sup_{\|\theta - \theta^*\| < \eta} \max_{1 \leq k \leq p} \int \text{ess sup}_{u} |\partial_k q_\theta(u, v) - \partial_k q_{\theta^*}(u, v)| d\gamma(v),
\]
\[
\omega^\varphi(\eta) \triangleq \sup_{\|\theta - \theta^*\| < \eta} \max_{1 \leq k \leq p} \int \text{ess sup}_{u} |\partial^2_{k,i} q_\theta(u, v) - \partial^2_{k,i} q_{\theta^*}(u, v)| d\gamma(v),
\]
\[
\omega^\varphi(y, \eta) \triangleq \sup_{\|\theta - \theta^*\| < \eta} \max_{1 \leq k \leq p} \text{ess sup}_{u} |\partial^2_{k,i} g_\theta(y|u) - \partial^2_{k,i} g_{\theta^*}(y|u)|,
\]
\[
\Delta^m(\eta) \triangleq \int \frac{\omega^\varphi(y, \eta)}{\inf_{\theta} \text{ess sup}_{u} g_\theta(y|x)} g_{\theta^*}(y|x) d\gamma(y).
\]

**Assumption 10.** \(\lim_{\eta \to 0} \omega^\alpha(\eta) = \lim_{\eta \to 0} \omega^\varphi(\eta) = 0\). Furthermore, there exists some \(\eta_0 > 0\) such that \(\Delta^m(\eta_0)\) is finite.

We finally prove the following proposition.

**Proposition 7.** Under Assumptions 1, 5–6 and 9–10, we have, for any \(k, l \in \{1, \ldots, p\}\),
\[
\lim_{\eta \to 0} E^* \left( \sup_{\|\theta - \theta^*\| < \eta} \frac{1}{n} |\partial^2_{k,i} l_n(\theta, \zeta) - \partial^2_{k,i} l_n(\theta^*, \zeta)| \right) = 0.
\]

**Remark 4.** This proposition ensures that \(\nabla [\lim_{n \to \infty} (1/n) \nabla l_n(\theta, \zeta)]_{\theta = \theta^*} = -I_{\theta^*}\) and that \(\nabla [\Delta(\theta)]_{\theta = \theta^*} = 0\).

These properties are sufficient to obtain the asymptotic normality of the maximum likelihood estimator using the following assumption.

**Assumption 11.** \(I_{\theta^*}\) is an invertible matrix.

**Theorem 3.** Under Assumptions 7–11 and the assumption that \(\hat{\theta}_n(\zeta)\) is consistent,
\[
n^{1/2}(\hat{\theta}_n - \theta^*) \to N(0, I_{\theta^*}^{-1}) \text{ weakly under } P^*.
\]

**Remark 5.** Combining Propositions 6 and 7 with Theorem 3, we obtain a confidence interval for \(\theta^*\). Moreover, we also have the local asymptotic normality conditions for our model, and the maximum likelihood estimator is locally asymptotically minimax (see Bickel et al. 1993).

**Appendix A. Exponential memorylessness of the prediction filter**

**Proof of Proposition 1.** Only the main steps of the proof will be described here. It is adapted from Mevel (1997) and uses results from Seneta (1981). The original idea comes from work
The maximum likelihood estimator for general hidden Markov models

by Arapostathis and Marcus (1990). Proofs which are not easily deduced from the finite state-space case are described in full.

Step 1. Let $p$ be a transition density function on $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \gamma)$. Define

$$\tau_1(p) \triangleq \frac{1}{2} \text{ess sup}_{x,x'} \int_y \left| p(x, v) - p(x', v) \right| \gamma(dy(v))$$

and, for all $u, v$ in $\mathbf{K}$, for all $f$ measurable and non-negative, write, for $n \geq m$,

$$A^{0}_{n,m}(u, v) \triangleq \int_u q_\theta(u, v) g_\theta(Y_n|u_n) \cdots \int_{u_{m+1}} q_\theta(u_{m+1}, u_{m+2}) g_\theta(Y_{m+1}|u_{m+1})$$

$$\times q_\theta(u, u_{m+1}) g_\theta(Y_{m+1}|u) d\gamma(u_{m+1}) \cdots d\gamma(u_n),$$

$$N(A_{n,m}^\theta)(u) \triangleq \int_v A_{n,m}^\theta(u, v) d\gamma(v),$$

$$N(A_{n,m}^\theta f)(u) = \int_{u,v} A_{n,m}^\theta(u, v) f(u) d\gamma(u) d\gamma(v).$$

Lemma 2.

$$\|\Phi_{n-m+1}(Y^m_n; \theta) - \Phi_{n-m+1}(Y^m_n, f'; \theta)\|_1 \leq \tau_1 \left( \frac{A_{n,m}^\theta}{N(A_{n,m}^\theta)} \right) \left\| \frac{N(A_{n,m}^\theta f - N(A_{n,m}^\theta f')}{N(A_{n,m}^\theta f')} \right\|_1 .$$

Proof. With the previous definitions, we may write

$$\|\Phi_{n-m+1}(Y^m_n, f; \theta) - \Phi_{n-m+1}(Y^m_n, f'; \theta)\|_1$$

$$= \int_v \left\| \int u \left[ \frac{N(A_{n,m}^\theta)(u)}{N(A_{n,m}^\theta)} f(u) - \frac{N(A_{n,m}^\theta)(u)}{N(A_{n,m}^\theta f')} f'(u) \right] d\gamma(u) \right\| d\gamma(v).$$

We then use the following general result. Let $\Psi$ be a measurable function such that $\int_u \Psi(u) d\gamma(u) = 0$. Then we have $\int \Psi^+ = \int \Psi^- = \frac{1}{2} \int |\Psi|$, where $\Psi^+ = \max(\Psi, 0)$, $\Psi^- = \max(-\Psi, 0)$. We may also write

$$\Psi(x) = \int a(s, x) d\gamma(s) - \int a(x, t) d\gamma(t), \quad \text{where} \quad a(s, t) = \frac{\Psi^-(s) \Psi^+(t)}{\int \Psi^+} .$$

Note that $\int a(x, t) d\gamma(t) = \frac{1}{2} \int |\Psi(x)| d\gamma(x)$. Applying this result to the function

$$\Psi(u) = \frac{N(A_{n,m}^\theta)(u)}{N(A_{n,m}^\theta f)} f(u) - \frac{N(A_{n,m}^\theta)(u)}{N(A_{n,m}^\theta f')} f'(u)$$

(with mean zero), we obtain
Lemma 4. Let \( \Psi(u) \) be a measurable and non-negative function. Then

\[
\int_{v} \left| \int_{u} \frac{A_{n,m}^{\Theta}(u, v)}{N(A_{n,m}^{\Theta}N)} \Psi(u) \right| dy(v)
\]

\[
= \int_{v} \left| \int_{u} \frac{A_{n,m}^{\Theta}(u, v)}{N(A_{n,m}^{\Theta}N)} \left( \int_{s} a(s, u)dy(s) - \int_{t} a(u, t)dy(t) \right) dy(u) \right| dy(v)
\]

\[
= \int_{v} \left| \int_{u} \frac{A_{n,m}^{\Theta}(x, v)a(y, x)}{N(A_{n,m}^{\Theta}N)} dy(x)dy(y) - \int_{u} \frac{A_{n,m}^{\Theta}(y, v)a(y, x)}{N(A_{n,m}^{\Theta}N)} dy(x)dy(y) \right| dy(v)
\]

\[
\leq \int_{v} \left| \frac{A_{n,m}^{\Theta}(x, v)}{N(A_{n,m}^{\Theta}N)} - \frac{A_{n,m}^{\Theta}(y, v)}{N(A_{n,m}^{\Theta}N)} \right| a(y, x)dy(x)dy(y)dy(v)
\]

\[
\leq \frac{1}{2} \left( \text{ess sup}_{x,y} \int_{v} \left| \frac{A_{n,m}^{\Theta}(x, v)}{N(A_{n,m}^{\Theta}N)} - \frac{A_{n,m}^{\Theta}(y, v)}{N(A_{n,m}^{\Theta}N)} \right| dy(v) \right)
\]

\[
\times \int_{u} \left| \frac{N(A_{n,m}^{\Theta}(u))}{N(A_{n,m}^{\Theta}f)} f(u) - \frac{N(A_{n,m}^{\Theta}(u))}{N(A_{n,m}^{\Theta}f')} f'(u) \right| dy(u)
\]

\[
= \tau_1 \left( \frac{N(A_{n,m}^{\Theta})}{N(A_{n,m}^{\Theta}f)} \right) \left\| \frac{N(A_{n,m}^{\Theta})}{N(A_{n,m}^{\Theta}f') f} - \frac{N(A_{n,m}^{\Theta})}{N(A_{n,m}^{\Theta}f') f'} \right\|_1,
\]

which concludes the proof of Lemma 2. \( \square \)

Step 2. This consists in bounding from above the last term of the right-hand side of the previous inequality.

Lemma 3.

\[
\left\| \frac{N(A_{n,m}^{\Theta})}{N(A_{n,m}^{\Theta}f)} f - \frac{N(A_{n,m}^{\Theta})}{N(A_{n,m}^{\Theta}f')} f' \right\|_1 \leq 2\omega^0 \| f - f' \|_1, \quad \text{where } \omega^0 = \sup \frac{N(A_{n,m}^{\Theta}f)}{N(A_{n,m}^{\Theta}g)}.
\]

The supremum is taken over the set of measurable and non-negative functions \( f \) and \( g \) such that \( \int f \, dy = \int g \, dy = 1 \).

Step 3. We prove the following result:

Lemma 4. \( \omega^0 \leq \epsilon^{-1} \delta_0(Y_m) \).

The proofs of Lemmas 3 and 4 are straightforward adaptations from Mevel (1997, Propositions A.1 and A.2), and are omitted here for brevity.

Then we conclude that
\[ \| \Phi_{n-m+1}(Y_n^m, f; \theta) - \Phi_{n-m+1}(Y_n^m, f'; \theta) \|_1 \leq 2 \epsilon_p^{-1} \delta_p(Y_n) \times \tau_1 \left( \frac{A_{n,m}^0}{N(A_{n,m}^0)} \right) \times \| f - f' \|_1. \]

Let us now examine the behaviour of \( \tau_1(A_{n,m}^0/N(A_{n,m}^0)) \). We wish to proceed by analogy between \( A_{n,m}^0 \) and a product of random matrices (which appears in finite state-space models): this technique will bring out the term \( (1 - \epsilon_p)^m \). We obtain an upper bound for \( \tau_1 \) by using another coefficient, denoted by \( \tau_B \) (Birkhoff’s coefficient), which does not depend on the normalization \( N(A_{n,m}^0) \).

**Step 4.** All these results are generalizations of those obtained by Seneta (1981, Theorems 3.12 and 3.13, pp. 106–110) and the proofs will be omitted. Let us define Birkhoff’s coefficient by

\[ \tau_B(p) = \sup \frac{d(p^* f, p^* g)}{d(f, g)}, \]

where

\[ d(f, g) = \text{ess sup} \log \frac{f(x)g(x')}{g(x)f(x')}, \]

\[ p^* f(u) = \int p^*(v, u)f(u)dy(v) = \int p(u, v)f(u)dy(u). \]

The first supremum is taken over the set of measurable and non-negative functions \( f \) and \( g \) such that \( f \neq \lambda g \ (\lambda \in \mathbb{R}) \), and \( p^* \) is the transpose density function associated with \( p \) (i.e. \( p^*(x, y) = p(y, x) \)). With this definition, \( \tau_B \) does not depend on the normalization, and is, moreover, submultiplicative. We now give another expression for \( \tau_B \), and then the upper bound of \( \tau_1 \) by \( \tau_B \).

**Lemma 5.**

\[ \tau_B(p) = \frac{1 - \alpha(p)^{1/2}}{1 + \alpha(p)^{1/2}} \geq \tau_1(p), \quad \text{where} \quad \alpha(p) = \text{ess inf} p(x, z) p(y, t) / p(x, t) p(y, z) = \alpha(p^*). \]

We now apply these results with \( p = A_{n,m}^0/(N(A_{n,m}^0)) \). We have

\[ \tau_1 \left( \frac{A_{n,m}^0}{N(A_{n,m}^0)} \right) \leq \tau_B \left( \frac{A_{n,m}^0}{N(A_{n,m}^0)} \right). \]

It follows from the definition that \( \alpha(A_{n,m}^0/N(A_{n,m}^0)) = \alpha(A_{n,m}^0) \), and then \( \tau_B(A_{n,m}^0/N(A_{n,m}^0)) = \tau_B(A_{n,m}^0) \) by application of Lemma 5. Since \( \tau_B \) is submultiplicative,

\[ \tau_B(A_{n,m}^0) \leq \prod_{k=m}^n \tau_B[q_\theta(\cdot, \cdot) g_\theta(Y_k|\cdot)]. \]

Finally, again using Lemma 5, \( \tau_B[q_\theta(\cdot, \cdot) g_\theta(Y_k|\cdot)] = \tau_B(q_\theta) \), and, under Assumption 1,
Then, using \( \text{Lip}(E) \),

\[
\alpha(q_\theta) = \text{ess inf}_{x,z,t} q_\theta(x, z)q_\theta(y, t) \quad \geq \quad \epsilon_\theta^2.
\]

Thus \( \tau_B(q_\theta) \leq 1 - \epsilon_\theta \), which finally implies that \( \tau_B(A_{n,m}^\theta) \leq (1 - \epsilon_\theta)^{n-m+1} \).

Appendix B. Geometric ergodicity of the extended chain

\textbf{Proof of Proposition 2.} Let \( h \) be in \( \text{Lip}(E) \), and \( z = (x, y, f, f^*) \), \( z' = (x', y', f', (f^*)') \) be in \( E \). In order to obtain an upper bound for \( |\Pi_\theta^n h(z) - \Pi_\theta^n h(z')| \), use the decomposition

\[
\Pi_\theta^n h(z) - \Pi_\theta^n h(z') = [\Pi_\theta^n h(x, y, f, f^*) - \Pi_\theta^n h(x, y', f', (f^*)')]
+ [\Pi_\theta^n h(x, y', f', (f^*)') - \Pi_\theta^n h(x', y', f', (f^*)')].
\]

Denote by \( A \) and \( B \) the first and second terms on the right-hand side of this equation, respectively. \( A \) and \( B \) will be treated separately.

For \( i \leq j \), write \( y^i_j = (y_i, \ldots, y_j) \) and, with an abuse of notation, \( d\gamma(y^i_j) = dy(x_i) \cdots dy(x_j) \) and \( d\nu(y^i_j) = d\nu(y_i) \cdots d\nu(y_j) \). Write also \( \{y^i_j, y\} \triangleq (y_i, \ldots, y_j, y) \). Then

\[
A = \int_{y^1_{n-1}} \left[ h(x_n, y_n, \Phi_n(y^1_{n-1}, y), f, \theta), \Phi_n(y^1_{n-1}, y, f^*, \theta^*) \right]
- h(x_n, y_n, \Phi_n(y^1_{n-1}, y'), f', \theta), \Phi_n(y^1_{n-1}, y', f^*, \theta^*) \right]
\times q^\theta(x_1) \cdots q^\theta(x_n) g^\theta(y_1|x_1) \cdots g^\theta(y_n|x_n) d\gamma(x_1) d\nu(y^1_n).
\]

Then, using \( \text{Lip}(E) \) as in Definition 1, and the fact that, for all \( \theta \) in \( \Theta \), \( \Phi_n(y^1_{n-1}, y, f, \theta) = \Phi_{n-1}(y^1_{n-1}, \Phi_1(y, f; \theta); \theta) \), we have

\[
|A| \leq \int_{y^1_{n-1}} \text{lip}(h, x_n, y_n)[||\Phi_{n-1}(y^1_{n-1}, \Phi_1(y, f; \theta); \theta) - \Phi_{n-1}(y^1_{n-1}, \Phi_1(y', f'; \theta); \theta)||_1
+ ||\Phi_{n-1}(y^1_{n-1}, \Phi_1(y, f^*; \theta^*); \theta^*) - \Phi_{n-1}(y^1_{n-1}, \Phi_1(y', (f^*)'; \theta^*); \theta^*)||_1]
\times q^\theta(x_1) \cdots q^\theta(x_n) g^\theta(y_1|x_1) \cdots g^\theta(y_n|x_n) d\gamma(x_1) d\nu(y^1_n).
\]

Exponential lack of memory of the prediction filter yields

\[
||\Phi_{n-1}(y^1_{n-1}, \Phi_1(y, f; \theta); \theta) - \Phi_{n-1}(y^1_{n-1}, \Phi_1(y', f'; \theta); \theta)||_1
\leq 2\epsilon_\theta^{-1} \delta_\theta(y_1)(1 - \epsilon_\theta)^{n-1} ||\Phi_1(y, f; \theta) - \Phi_1(y', f'; \theta)||_1
\leq 4\epsilon_\theta^{-1} \delta_\theta(y_1)(1 - \epsilon_\theta)^{n-1}.
\]

The same inequality holds at the point \( \theta^* \). Integrating the above expression and using Assumption 1 yields
The maximum likelihood estimator for general hidden Markov models

\[ |A| \leq 4 \text{lip}(h)[\epsilon_t^{-1} \Delta_t (1 - \epsilon_t)^{n-1} + \epsilon_t^{\frac{3}{2}} \Delta_t (1 - \epsilon_t)^{n-1}] \]
\[ \leq 8 \text{lip}(h) \epsilon_t^{-1} \Delta_t (1 - \epsilon)^{n-1}. \]  

(37)

**Remark 6.** We may prove along the same lines that if \( h \) is in \( \text{Lip}(E) \) then, for any integer \( n \), \( \Pi_0^n h \) also belongs to \( \text{Lip}(E) \), with \( \text{lip}(\Pi_0^n h, x, y) \leq \text{lip}(h) \delta_0(y) \) and \( k(\Pi_0^n h, x, y) \leq k(h) \).

We now turn our attention to \( B \), given by

\[ B = \int_{y_1, \ldots, y_n} h(x_n, y_n, \Phi_n(\{y_{n-1}^1, y'\}, f', \theta), \Phi_n(\{y_{n-1}^1, y'\}, (f^*)'; \theta^*)) \times [q_{\theta^*}(x, x_1) - q_{\theta^*}(x', x_1)]q_{\theta^*}(x_1, x_2) \cdots q_{\theta^*}(x_{n-1}, x_n) \times g_{\theta^*}(y_1|x_1) \cdots g_{\theta^*}(y_n|x_n) dy_1 dy_n. \]

Let \( m = \lfloor n/2 \rfloor \) and \( F, F^* \) be fixed functions in \( S^+ \). Note that

\[ h(x_n, y_n, \Phi_n(\{y_{n-1}^1, y'\}, f'; \theta), \Phi_n(\{y_{n-1}^1, y'\}, (f^*)'; \theta^*)) \]
\[ = h(x_n, y_n, \Phi_{n-m}(y_{n-1}^m, \Phi_{m}(\{y_{m-1}^1, y'\}, f'; \theta), \theta), \]
\[ \Phi_{n-m}(y_{n-1}^m, \Phi_{m}(\{y_{m-1}^1, y'\}, (f^*)'; \theta^*)) \]
\[ - h(x_n, y_n, \Phi_{n-m}(y_{n-1}^m, F; \theta), \Phi_{n-m}(y_{n-1}^m, F^*; \theta^*)) \]
\[ + h(x_n, y_n, \Phi_{n-m}(y_{n-1}^m, F; \theta), \Phi_{n-m}(y_{n-1}^m, F^*; \theta^*)). \]

Insert this form into \( B \) and use the fact that \( h \) is in \( \text{Lip}(E) \) for the first part and take the integral with respect to \( y_{m-1}^1 \) and \( x_{m-1}^1 \) in the second; this yields

\[ |B| \leq \int_{x_1, \ldots, x_n} \text{lip}(h, x_n, y_n) \times (||\Phi_{n-m}(y_{n-1}^m, \Phi_{m}(\{y_{m-1}^1, y'\}, f'; \theta); \theta) - \Phi_{n-m}(y_{n-1}^m, F; \theta)||_1 \]
\[ + \|\Phi_{n-m}(y_{n-1}^m, \Phi_{m}(\{y_{m-1}^1, y'\}, (f^*)'; \theta^*); \theta^*) - \Phi_{n-m}(y_{n-1}^m, F^*; \theta^*)||_1) \times q_{\theta^*}(x, x_1) - q_{\theta^*}(x', x_1)]q_{\theta^*}(x_1, x_2) \cdots q_{\theta^*}(x_{n-1}, x_n) \times g_{\theta^*}(y_1|x_1) \cdots g_{\theta^*}(y_n|x_n) dy_1 dy_n \]
\[ + \int_{y_1, \ldots, y_n} \int_{x_1, \ldots, x_n} |h(x_n, y_n, \Phi_n(\{y_{n-1}^m, F; \theta), \Phi_n(\{y_{n-1}^m, F^*; \theta^*)|) \times q_{\theta^*}(x, x_m) - q_{\theta^*}(x', x_m)]q_{\theta^*}(x_m, x_{m+1}) \cdots q_{\theta^*}(x_{n-1}, x_n) \times g_{\theta^*}(y_m|x_m) \cdots g_{\theta^*}(y_n|x_n) dy_m dy_n. \]

Using Proposition 1 and inequality (9) yields
The existence of Proof of Corollaries 1 and 2. Inequalities (37) and (38) complete the proof.

Proof of Corollaries 1 and 2. The existence of \( \Lambda_\theta(h) \) follows from Proposition 2, using the same ideas as in Sunyach (1975, Theorem 1). Then we can easily check that the function defined on \( E \) by

\[
V(z) = \sum_{n \geq 0} [\Pi_0^n h(z) - \Lambda_\theta(h)]
\]

is a solution to the Poisson equation. As the space \( E \) is not locally compact, we cannot use the Riesz representation theorem to show that \( \Lambda_\theta(h) \) is the expectation of \( h \) under a suitable measure. Let us show that the family of probability measures \( (\Pi_0^n(z, \cdot))_{n \geq 0} \) is asymptotically tight (for definition see, van der Vaart and Wellner 1996, p. 20), using the same basic ideas as in Sunyach (1975, Theorem 1).

Let \( d(\cdot, \cdot) \) be a distance in \( E \). For \( D \) a compact set of \( E \) and \( \delta > 0 \), define

\[
\varphi_{\delta, D}(z) \triangleq \inf(1, (1/\delta)d(z, D)).
\]

It is straightforward to show that \( \varphi_{\delta, D}(z) \) is in \( \text{Lip}(E) \), with \( \text{lip}(\varphi_{\delta, D}) \leq \delta^{-1} \) and \( k(\varphi_{\delta, D}) \leq 1 \). Moreover, for any set \( \mathcal{A} \), write \( \mathcal{A}_\delta \triangleq \{ z \in E; d(z, \mathcal{A}) \leq \delta \} \) for the \( \delta \)-enlargement of \( \mathcal{A} \).

Let \( \epsilon, \delta > 0 \) and \( z \) in \( E \). There exists \( p \) sufficiently large such that \( C[\delta^{-1} + 1] \times \rho^p(1 - \rho)^{-1} < \epsilon/2 \) (where \( C \) is the constant appearing in (12)). Now, choose a compact \( D' \) such that \( \Pi_0^n[z, (D')^c] < \epsilon/2 \) for \( k = 1, \ldots, p \), which is possible since any probability measure in a separable and complete space is tight (the superscript \( c \) denotes the complement in the space \( E \)). Note that we have \( \tilde{\gamma}_{((D')_k)^c} \leq \varphi_{\delta, D'} \leq \tilde{\gamma}_{(D')^c} \), so we obtain, for any \( n > p \),
The maximum likelihood estimator for general hidden Markov models

\[
\Pi_\theta^b[z, ((D')_b)^c] \leq \Pi_\theta^b(\varphi_{\theta, D'})(z) \\
\leq \Pi_\theta^b(\varphi_{\theta, D'})(z) + C \left[ \frac{1}{\delta} + 1 \right] \frac{\rho^p}{1 - \rho} \\
\leq \Pi_\theta^b[z, (D')^c] + \frac{c}{2} \\
\leq \epsilon. \quad (40)
\]

So \((\Pi_\theta^b(z, \cdot))_{n \geq 0}\) is asymptotically tight. Hence, by Prohorov’s theorem (van der Vaart and Wellner 1996, p. 21), there exists a probability measure \(\lambda_\theta\) and a subsequence \((n_i)_{i \geq 0}\) such that \(\Pi_{\theta_i}(z, \cdot)\) converges weakly to \(\lambda_\theta\). This proves that for any continuous and bounded function \(h\) in \(\text{Lip}(E)\), \(\Lambda_\theta(h) = \int h(z) d\lambda_\theta(z)\). Using the result of Corollary 1, we obtain that, for any continuous and bounded function \(h\) in \(\text{Lip}(E)\),

\[
\lambda(\xi) \Pi_{\theta_i}^b h \overset{\Delta}{=} \int \Pi_{\theta_i} h(z) \lambda(z)(dz) = E_{\theta, \lambda}(h(Z_n)) \to_{n \to \infty} \int h(z) d\lambda_\theta(z) = \Lambda_\theta(h). \quad (41)
\]

Then, using the fact that, for any \(h\) in \(\text{Lip}(E)\), \(\Pi_{\theta_i} h\) is still in \(\text{Lip}(E)\) (see Remark 6), \(\lambda_\theta\) is the unique invariant probability measure associated with the kernel \(\Pi_{\theta_i}\).

Now let \(h\) be a function in \(\text{Lip}(E)\) \((h\) is not necessarily continuous and bounded). Define

\[
G_h(x, y, f, f^*) \overset{\Delta}{=} \int h(x_0, y_0, f, f^*) g_{\theta^*}(y_0|x_0) f^*(x_0) d\nu(y_0) d\gamma(x_0) \quad (42)
\]

It can be easily checked that \(G_h\) is a bounded and continuous function in \(\text{Lip}(E)\). Applying the previous result to the function \(G_h\), we obtain

\[
\lim_n \lambda(\xi) \Pi_{\theta_i}^b G_h = \int G_h(z) d\lambda_\theta(z) \\
= \int h(x_0, y_0, f, f^*) g_{\theta^*}(y_0|x_0) f^*(x_0) d\nu(y_0) d\gamma(x_0) \lambda_\theta(K, \gamma, d\nu, d\gamma) \lambda_\theta(K, \gamma, d\nu, d\gamma) \quad (43)
\]

Using Remark 2,

\[
\lambda(\xi) \Pi_{\theta_i}^b h = E_{\theta, \lambda}(h(Z_n)) \\
= E^*[E^*(h(X_n, Y_n, f_{\theta^*, n}^*, f_{\theta^*, n}^* Y_{n-1})^\theta)] \\
= E^*(G_h(X_n, Y_n, f_{\theta^*, n}^*, f_{\theta^*, n}^*)) \\
= E_{\theta, \lambda}(G_h(Z_n)) \\
= \lambda(\xi) \Pi_{\theta_i}^b G_h. \quad (44)
\]

But we already know by (41) that

\[
\lim_n \lambda(\xi) \Pi_{\theta_i}^b h = \Lambda_\theta(h). \quad (45)
\]
Combining with (43), (44) and (45),
\[ \Lambda_\theta(h) = \int h(x_0, y_0, f, f^*) g_{\theta}(y_0|x_0)f^*(x_0)dy(y_0)dy(\gamma(x_0)\lambda_\theta(K, \gamma, df, df^*). \]
But we have
\[ g_{\theta}(y_0|x_0)f^*(x_0)dy(y_0)dy(\gamma(x_0)\lambda_\theta(K, \gamma, df, df^*) = \lambda_\theta(dx_0, dy_0, df, df^*), \]
as these measures are entirely identified by their integrals on Lipschitz and bounded functions on \( E \). Then we obtain, for all \( h \) in \( \text{Lip}(E) \),
\[ \Lambda_\theta(h) = \int h(z)d\lambda_\theta(z). \]

Appendix C. Consistency of the maximum likelihood estimator

Proof of Proposition 3. (i) Denote by \( V \) by the solution of the Poisson equation associated to the function \( h \) (see Corollary 1) and by \( \mathcal{F}_n \) the \( \sigma \)-algebra generated by \( \{Z_0, \ldots, Z_n\} \). It is easy to show, using inequality (12) for \( n > 1 \), that
\[ |V(z)| \leq C_\rho \text{lip}(h) + k(h) \left( \frac{1}{1 - \rho} \right)^2 + |\Lambda_\theta(h)| + k(h, x, y), \]
and thus there exist \( s > 1 \) (cf. Assumption 3) such that \( \Pi_\theta |V|^s \) is bounded, which implies that \( E_\theta(|V(Z_{n+1}) - \Pi_\theta V(Z_n)|^s|\mathcal{F}_n) \) is bounded independently of \( n \). The proof follows from the classical identity (Meyn and Tweedie 1993, 17.4.3)
\[ \frac{1}{n} \sum_{m=0}^{n-1} h(Z_m) \]
\[ = \Lambda_\theta(h) + \frac{1}{n} \sum_{m=1}^{n} [V(Z_m) - \Pi_\theta V(Z_{m-1})] + \frac{1}{n} [\Pi_\theta V(Z_0) - \Pi_\theta V(Z_n) + h(Z_0) - h(Z_n)] \]
by applying the corollary of Chow’s theorem (Hall and Heyde 1980, p. 36).
(ii) This is a direct consequence of (i) by applying Remark 2.

Proof of Proposition 4. It is easy to show that
\[ |h_\theta(x, y, f_1, f_2) - h_\theta(x, y, f_1', f_2')| = \left| \log \left( 1 + \frac{\int_{u \in \Theta} g_{\theta}(y|u)(f_1 - f_1')(u)dy(u)}{\int_{u \in \Theta} g_{\theta}(y|u)f_1'(u)dy(u)} \right) \right| \]
\[ \leq \delta_\theta(y)\|f_1 - f_1'\|_1 \]
(47)
The maximum likelihood estimator for general hidden Markov models

and

\[ |h_\theta(x, y, f_1, f_2)| \leq k_\theta(y), \]  

which proves that \( h \) is in Lip(\( E \)) and verifies equation (15), using Assumptions 2 and 3.

\[ \tag{48} \]

**Proof of Lemma 1.** Let \( \eta > 0 \) and \( \| \theta - \theta' \| \leq \eta \). By Proposition 1,

\[
\| \Phi_{n+1}(Y^0_n, \zeta; \theta) - \Phi_{n+1}(Y^0_n, \zeta; \theta') \|_1
\]

\[
= \| \Phi_n(Y^1_n, \Phi_1(Y_0, \zeta; \theta); \theta) - \Phi_n(Y^1_n, \Phi_1(Y_0, \zeta; \theta'); \theta') \|_1
\]

\[
\leq \| \Phi_n(Y^1_n, \Phi_1(Y_0, \zeta; \theta); \theta) - \Phi_n(Y^1_n, \Phi_1(Y_0, \zeta; \theta'); \theta) \|_1
\]

\[
+ \| \Phi_n(Y^1_n, \Phi_1(Y_0, \zeta; \theta'); \theta) - \Phi_n(Y^1_n, \Phi_1(Y_0, \zeta; \theta'); \theta') \|_1
\]

\[
\leq 2\epsilon_\theta^{-1}\delta_\theta(Y_1)\rho^n\| \Phi_1(Y_0, \zeta; \theta) - \Phi_1(Y_0, \zeta; \theta') \|_1 + \sup_\zeta \| \Phi_n(Y^1_n, f; \theta) - \Phi_n(Y^1_n, f; \theta') \|_1.
\]

\( \tag{49} \)

It is straightforward to show that

\[
\| \Phi_1(Y_0, \zeta; \theta) - \Phi_1(Y_0, \zeta; \theta') \|_1
\]

\[
\leq \left\| \frac{\int (q_\theta(u, \cdot) - q_\theta(u, \cdot))g_\theta(Y_0|u)\zeta(u)dy(u)}{\int g_\theta(Y_0|u)\zeta(u)dy(u)} \right\|_1
\]

\[
+ \left\| \frac{\int q_\theta(u, \cdot)g_\theta(Y_0|u)\zeta(u)dy(u)}{\int g_\theta(Y_0|u)\zeta(u)dy(u)} - \frac{\int q_\theta(u, \cdot)g_\theta(Y_0|u)\zeta(u)dy(u)}{\int g_\theta(Y_0|u)\zeta(u)dy(u)} \right\|_1.
\]

The first term on the right-hand side of this expression is bounded from above by \( \omega^\theta(\eta) \). The second term on the right is bounded from above by \( 2\delta'(Y_0, \eta) \), by breaking the difference into two terms including \( \int q_\theta(u, \cdot)g_\theta(Y_0|u)\zeta(u)dy(u) / \int g_\theta(Y_0|u)\zeta(u)dy(u) \). Combining these two bounds with (49),

\[
\sup_\zeta \| \Phi_{n+1}(Y^0_n, \zeta; \theta) - \Phi_{n+1}(Y^0_n, \zeta; \theta') \|_1
\]

\[
\leq 2\epsilon_\theta^{-1}\delta_\theta(Y_1)(\omega^\theta(\eta) + 2\delta'(Y_0, \eta))\rho^n\| \theta - \theta' \| + \sup_\zeta \| \Phi_n(Y^1_n, \zeta; \theta) - \Phi_n(Y^1_n, \zeta; \theta') \|_1.
\]

A straightforward recurrence completes the proof. \( \square \)
Appendix D. Asymptotic normality of the maximum likelihood estimator

D.1. Exponential memorylessness of the gradient of the prediction filter and consequences

Iterating equation (35) at order \( m, m \leq n \), yields, for all \( v \) in \( K \),

\[
\nabla f_{\theta, n+1}^\wedge(v) = \Psi_{n-m+1}(Y_n^m, f_{\theta, m}^\wedge, \nabla f_{\theta, m}^\wedge; \theta)(v),
\]

where \( \Psi_n \) is the vector function \((\Psi_n^k)_{1 \leq k \leq p}\) defined by

\[
\Psi_{n-m+1}(Y_n^m, f, \sigma; \theta)(v) = \int_{u_m} a_0(Y_n, f_n)(u_n, v) \cdots \int_{u_m} a_0(Y_m, f)(u_m, u_{m+1}) \sigma_k(u_m) d\gamma(u_m) \cdots d\gamma(u_n)
\]

\[
+ \sum_{l=m}^{n-1} \int_{u_l} a_0(Y_n, f_n)(u_n, v) \cdots \int_{u_{l+1}} a_0(Y_{l+1}, f_{l+1})(u_{l+1}, u_{l+2})
\]

\[
\times U_{\theta, k}(Y_{l+1}, f_{l+1}) d\gamma(u_{l+1}) \cdots d\gamma(u_n) + U_{\theta, k}(Y_n, f_n)(v),
\]

where, for all \( l \leq m + 1 \), we use the notation \( f_l \triangleq \Phi_{l-m}(Y_{l+1}^n, f; \theta) \).

In this subsection, the exponential memorylessness for \( \Psi \) of the initial conditions is proved. It is straightforward to show that, for any \( k \) in \( \{1, \ldots, p\} \), \( U_{\theta, k} \) verifies a Lipschitz condition. For all \( y \) in \( \mathcal{Y} \), for all \( f, f' \) in \( \mathbf{S}^+ \),

\[
||U_{\theta, k}(y, f) - U_{\theta, k}(y, f')||_1 \leq \text{lip}(U_{\theta, k}, y)||f - f'||_1,
\]

\[
||U_{\theta, k}(y, f)||_1 \leq k(U_{\theta, k}, y),
\]

where

\[
\text{lip}(U_{\theta, k}, y) = \delta_\theta(y)(1 + \delta_\theta(y)) \left( \max_{1 \leq k \leq p} \int_{x'} \text{ess sup}_x |\partial_k q_\theta(x, x')| d\gamma(x') \right) + \delta_\theta(y)(1 + \delta_\theta(y))
\]

\[
+ 2\delta_\theta(y)\delta_\theta'(y)(1 + \delta_\theta(y))^2,
\]

\[
k(U_{\theta, k}, y) = \delta_\theta(y) \left( \max_{1 \leq k \leq p} \int_{x'} \text{ess sup}_x |\partial_k q_\theta(x, x')| d\gamma(x') \right) + \delta_\theta(y)(1 + \delta_\theta(y)).
\]

Note that \( \text{lip}(U_{\theta, k}) \triangleq \text{ess sup}_y \int_x \text{lip}(U_{\theta, k}, y) g_{\theta'}(y|x) d\nu(y) < \infty \) provided that \( \Delta_2 \) and \((\Delta'\Delta')_{3,1}\) are finite, and \( k(U_{\theta, k}) \triangleq \text{ess sup}_y \int_x k(U_{\theta, k}, y) g_{\theta'}(y|x) d\nu((y)) < \infty \) provided that \( \Delta_1 \) and \((\Delta'\Delta')_{1,1}\) are finite.

**Proposition 8.** Under Assumptions 1 and 7, and for all \( f, f' \) in \( \mathbf{S}^+ \), \( \sigma, \sigma' \) in \( \mathbf{S} \), \( \theta \) in \( \Theta \), and all \( 1 \leq k \leq p \).
The maximum likelihood estimator for general hidden Markov models

\[
\|\Psi_{n-1}^k(Y_n, f, \sigma; \theta) - \Psi_{n-1}^k(Y_n, f', \sigma'; \theta)\|_1
\leq 6\delta_0(Y_m)\epsilon_0^{-1}(1 - \epsilon_0)^{n-1} \left[ \|\sigma_k - \sigma'_{k}\|_1 + \|f - f'\|_1 (1 + \|\sigma_k\|_1 + \|\sigma'_{k}\|_1) \right]
+ 10\delta_0(Y_m)\epsilon_0^{-1}(1 - \epsilon_0)^{n-1} \|f - f'\|_1
\left[ \text{lip}(U_{\theta, k}, Y_n) + \sum_{l=m}^{n-1} \text{lip}(U_{\theta, k}, Y_l)\delta_0(Y_{l+1})^2 + \sum_{l=m}^{n-1} \delta_0(Y_{l+1})^4 \right].
\]

The proof is a straightforward adaptation from Mevel (1997, Part I, Proposition 4.4) and is omitted here.

Now define a class of Lipschitz functions on \( \mathcal{E}^\prime \):

**Definition 2.** \( \text{Lip}(\mathcal{E}^\prime) \) is the set of vector-valued functions \( h = (h_k)_{1 \leq k \leq p} \) such that each \( h_k \) is a real-valued measurable function on \( \mathcal{E}^\prime \); and for all \( (x, y) \) in \( \mathbb{K} \times \mathcal{Y} \), there exist \( \text{lip}'(h_k, x, y) \) and \( k'(h_k, x, y) \) such that, for all \( f, f' \) in \( \mathbb{S}^\prime \), for all \( \sigma, \sigma' \) in \( \Sigma \),

\[
|h_k(x, y, f, \sigma) - h_k(x, y, f', \sigma')| 
\leq \text{lip}'(h_k, x, y)|\|\sigma_k - \sigma'_{k}\|_1 + \|f - f'\|_1 | \times (1 + \|\sigma_k\|_1 + \|\sigma'_{k}\|_1),
\]

\[
|h_k(x, y, f, \sigma)| \leq k'(h_k, x, y)(1 + \|\sigma_k\|_1)
\]
and

\[
\text{lip}'(h_k) = \text{ess sup}_x \int \text{lip}'(h_k, x, y)g_{\theta^y}(y|x)\,d\nu(y) < \infty,
\]

\[
k'(h_k) = \text{ess sup}_x \int k'(h_k, x, y)g_{\theta^y}(y|x)\,d\nu(y) < \infty.
\]

The following analogue of Proposition 2 can be proved.

**Proposition 9.** Under Assumptions 1 and 7, there exist constants \( C > 0 \) and \( \tilde{\rho} \in ]0, 1[ \) such that, for all \( z, z' \) in \( \mathcal{E}^\prime \), for all \( h \) in \( \text{Lip}(\mathcal{E}^\prime) \), for all \( n \geq 1 \), and for all \( 1 \leq k \leq p \), we have

\[
|\hat{\Pi}_n^0 h_k(z) - \hat{\Pi}_n^0 h_k(z')| \leq
C(k'(h_k) + \text{lip}'(h_k))\tilde{\rho}^n (1 + \|\sigma_k\|_1 + \|\sigma'_{k}\|_1)(\delta_0(y)^2 + \delta_0(y')^2 + k(U_{\theta, k}, y) + k(U_{\theta, k}, y')).
\]

**Corollary 3.** Under Assumptions 1 and 7, there exist constants \( C > 0 \) and \( \tilde{\rho} \in ]0, 1[ \) such that, for all \( h \) in \( \text{Lip}(\mathcal{E}^\prime) \), there exists a family of constants \( (\tilde{\Lambda}_0(h_k))_{1 \leq k \leq p} \), such that for all \( z \) in \( \mathcal{E}^\prime \), for all \( n \geq 1 \), and for all \( 1 \leq k \leq p \),

\[
|\hat{\Pi}_n^0 h_k(z) - \tilde{\Lambda}_0(h_k)| \leq C(\text{lip}'(h_k) + k'(h_k))\tilde{\rho}^n (1 + \|\sigma_k\|_1)(\delta_0(y)^2 + k(U_{\theta, k}, y)).
\]

For any initial probability measure \( \tilde{\lambda} \) on \( (\mathcal{E}^\prime, \mathcal{B}(\mathcal{E}^\prime)) \), and for any function \( h \) in \( \text{Lip}(\mathcal{E}^\prime) \),
\[ \tilde{\Lambda}_\theta(h) = \lim_{n \to \infty} \tilde{E}_{\theta, \lambda}(h(\tilde{Z}_n)), \tag{50} \]

and, for all \( 1 \leq k \leq p \), there exists a solution in \( E' \) to the Poisson equation \((I - \tilde{\Pi}_\theta)V(\cdot) = h_k(\cdot) - \tilde{\Lambda}_\theta(h_k)\).

The proofs are similar to the proofs of Proposition 2 and its corollary, and are omitted here (note that they assume that \( \Delta_1 \) is finite).

We can now prove the convergence of the gradient of the normalized log-likelihood.

**D.2. Proof of Proposition 5**

Differentiating the expression for the log-likelihood, we obtain, for all \( k \) in \( \{1, \ldots, p\} \),

\[
\frac{1}{n} \partial_k l_n(\theta, \xi) = \frac{1}{n} \sum_{m=0}^{n-1} \left[ \frac{\partial_k g_\theta(Y_m|u)f^2_{\theta, m}(u)d\gamma(u) + \int g_\theta(Y_m|u)\partial_k f_{\theta, m}(u)d\gamma(u)}{\int g_\theta(Y_m|u)f_{\theta, m}(u)d\gamma(u)} \right] \quad P^*-a.s.
\]

\[
= \frac{1}{n} \sum_{m=0}^{n-1} j_{\theta, k}(\tilde{Z}_m) \quad P^*_{\theta, \lambda(\xi)}-a.s.,
\]

where

\[
j_{\theta, k}(x, y, f, \sigma) \triangleq \frac{\partial_k g_\theta(y|u)f(u)d\gamma(u) + \int g_\theta(y|u)\sigma_k(u)d\gamma(u)}{\int g_\theta(y|u)f(u)d\gamma(u)}.
\]

We check (using simple calculus) that \( j_\theta \) is in \( \text{Lip}(E') \), assuming that \( \Delta_2 \) and \( (\Delta\Delta')_{1,1} \) are finite. (In fact, \( \text{lip}(j_\theta, x, y) = \delta_\theta(y)(1 + \delta_\theta(y)) + \delta_\theta(y)^2 \) and \( k'(j_\theta, x, y) = \delta_\theta(y) + \delta_\theta'(y) \). Assuming, moreover, that, for some \( s > 1 \), \( \Delta_s \) and \( \Delta_s' \) are finite (in order to obtain a condition like (15)), we prove, just as in Proposition 4, that, for all \( k \) in \( \{1, \ldots, p\} \),

\[
\frac{1}{n} \partial_k l_n(\theta, \xi) \rightarrow \tilde{\Lambda}_\theta(j_{\theta, k}) \quad P^*-a.s.
\]

But the score function \( \tilde{\Lambda}_\theta(j_\theta) \) vanishes at \( \theta^* \): Propositions 6 and 7 show that the convergence of \((1/n)\nabla^2 l_n(\theta, \xi)\) happens \( P^*-a.s. \) and uniformly for \( \theta \) in a neighbourhood of \( \theta^* \), for some subsequence; and then for the whole family, as it is still \( P^*-a.s. \) convergent. Thus, \((1/n)\nabla l_n(\theta, \xi)\) is uniformly convergent, vanishes at \( \tilde{\theta}_n(\xi) \), and \( \tilde{\theta}_n(\xi) \) converges to \( \theta^* \), which proves the result.

Let \( V^*_k \) denote the solution to the Poisson equation \((I - \tilde{\Pi}_{\theta^*})V = j_{\theta^*, k} - \tilde{\Lambda}_{\theta^*}(j_{\theta^*, k})\) (which exists by Corollary 3). We have, \( P^*_{\theta^*, \lambda(\xi)}-a.s. \).
\[
\frac{1}{n^{1/2}} \partial_k l_n(\theta, \zeta) = \frac{1}{n^{1/2}} \sum_{m=0}^{n-1} j_{\theta^*} Z_m \hat{\Lambda}_{\theta^*}(j_{\theta^*}) \\
= \frac{1}{n^{1/2}} \sum_{m=1}^{n} V_k^* Z_m - \hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_{m-1}) \\
+ \frac{1}{n^{1/2}} (\hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_0) - \hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_n) + j_{\theta^*} Z_0 - j_{\theta^*} (\hat{Z}_n)).
\]

The second term converges to zero as the quantities \( \hat{\Pi}_{\theta^*} V_k^* (\cdot) \) and \( j_{\theta^*} \) are bounded. (Recall that an inequality like (46) still holds. In fact, \( V_k^* (z) \leq C(\delta_\theta(y) + k(U_{\theta, k}, y)) + k(j_{\theta^*}, x, y) \).) Thus, for all \( 1 \leq k \leq p \),

\[
\lim_{n \to \infty} \left[ \frac{1}{n^{1/2}} \partial_k l_n(\theta, \zeta) - \frac{1}{n^{1/2}} \sum_{m=1}^{n} (V_k^* Z_m - \hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_{m-1})) \right] = 0 \quad \mathbb{P}_{\theta^*, \hat{\lambda}(\zeta)} \text{-a.s.}
\]

Now, since \( (\hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_n))_n \) is bounded, the convergence holds in \( L_2(\mathbb{P}_{\theta^*, \hat{\lambda}(\zeta)}): \)

\[
\mathbb{E}_{\theta^*, \hat{\lambda}(\zeta)} \left[ \frac{1}{n^{1/2}} \partial_k l_n(\theta, \zeta) - \frac{1}{n^{1/2}} \sum_{m=1}^{n} V_k^* (\hat{Z}_{m-1}) - \hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_{m-1}) \right]^2 \to 0 \quad (51)
\]

Let \( u \) be a fixed vector of \( \mathbb{R}^p \), \( \sum_{m=1}^{n} u^T B_m \Delta \equiv \sum_{m=1}^{n} u^T (V_k^* (\hat{Z}_{m-1}) - \hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_{m-1})) \) is a real-valued martingale adapted to the Borel \( \sigma \)-field \( \{\hat{\mathcal{Z}}_n = \sigma((\hat{Z}_k)_{k \leq n})\}_n \geq 0 \). We shall use Theorem 3.2 in Hall and Heyde (1980). We study

\[
\sigma_n^2 \triangleq \sum_{m=1}^{n} \mathbb{E}_{\theta^*, \hat{\lambda}(\zeta)} (u^T B_m)^2 \\
= \sum_{1 \leq k \leq p} u_k u_l \sum_{m=1}^{n} \mathbb{E}_{\theta^*, \hat{\lambda}(\zeta)} [V_k^* (\hat{Z}_{m-1}) - \hat{\Pi}_{\theta^*} V_k^* (\hat{Z}_{m-1})] (V_l^* (\hat{Z}_{m-1}) - \hat{\Pi}_{\theta^*} V_l^* (\hat{Z}_{m-1})) | \hat{\mathcal{Z}}_{m-1} \\
= \sum_{1 \leq k \leq p} u_k u_l \sum_{m=1}^{n} \mathbb{E}_{\theta^*, \hat{\lambda}(\zeta)} [(V_k^* V_l^*) (\hat{Z}_{m-1}) - \hat{\Pi}_{\theta^*} V_l^* (\hat{Z}_{m-1}) \hat{\Pi}_{\theta^*} V_l^* (\hat{Z}_{m-1})].
\]

Consider the function defined by

\[
z \to (V_k^* V_l^*) (z) \triangleq \mathcal{F}_{k,l}(z).
\]

We check that this function is in \( \text{Lip}(E') \) and verifies condition (15). First,

\[
|\mathcal{F}_{k,l}(z)| \leq C(\delta_{\theta^*}(y) + \delta_{\theta^*}(y))^2
\]

(where \( C \) depends on \( \Delta_2 \) and \( \Delta_2' \), with \( \Delta_2 \) and \( \Delta_2' \) finite for some \( s > 1 \), by assumption. The Lipschitz criterion is verified by studying the quantity \( \mathcal{F}_{k,l}(x, y, f, \sigma) - \mathcal{F}_{k,l}(x, y, f', \sigma') \), where \( z = (x, y, f, \sigma) \) and \( z' = (x, y, f', \sigma') \) are elements of the space \( E' \). We have
As this family is uniformly integrable under Assumption 7, the convergence also happens for its first moment, so that

\[
\frac{1}{n} \sum_{m=1}^{n} \tilde{E}_{\theta^*, \lambda(\zeta)}[(u^T B_m)^2] \rightarrow 1 \quad \text{in } \tilde{P}_{\theta^*, \lambda(\zeta)}\text{-probability}.
\]

We prove now that, for any positive \( \epsilon \),

\[
\frac{1}{n} \sum_{m=1}^{n} \tilde{E}_{\theta^*, \lambda(\zeta)}[(u^T B_m)^2 \|

\| u^T B_m \| \geq \alpha \epsilon \| \zeta_{m-1} \|] \rightarrow 0 \quad \text{in } \tilde{P}_{\theta^*, \lambda(\zeta)}\text{-probability}.
\]

For \( \alpha > 0 \),

\[
\tilde{P}_{\theta^*, \lambda(\zeta)}\left( \frac{1}{n} \sum_{m=1}^{n} \tilde{E}_{\theta^*, \lambda(\zeta)}[(u^T B_m)^2 \|

\| u^T B_m \| \geq \alpha \epsilon \| \zeta_{m-1} \|] \geq \alpha \right)
\]
The maximum likelihood estimator for general hidden Markov models

\[ \mathbb{E}_{\theta^*, \tilde{\lambda}(\xi)} \left( \sum_{n=1}^{\infty} \tilde{E}_{\theta^*, \tilde{\lambda}(\xi)} \left[ (u^T B_m)^2 | u^T B_m | \gtrless \sigma_n \right] \right) \]

\[ \leq \frac{1}{m\alpha} \sum_{m=1}^{n} \tilde{E}_{\theta^*, \tilde{\lambda}(\xi)} \left[ (u^T B_m)^2 | u^T B_m | \gtrless \sigma_n \right] \]

\[ \leq \frac{1}{\alpha} \max_{1 \leq m \leq n} \tilde{E}_{\theta^*, \tilde{\lambda}(\xi)} \left[ (u^T B_m)^2 | u^T B_m | \gtrless \sigma_n \right] \]

\[ \leq \frac{1}{\alpha(\sigma_n)^{2s-2}} \max_{1 \leq m \leq n} \tilde{E}_{\theta^*, \tilde{\lambda}(\xi)} \left[ (u^T B_m)^2 \right]; \]

as \( \left( \tilde{E}_{\theta^*, \tilde{\lambda}(\xi)} \left[ (u^T B_m)^2 \right] \right)_m \) is bounded and \( 1/\sigma_n \) tends to zero, we obtain the result. We conclude that

\[ \frac{1}{\sigma_n} \sum_{m=1}^{n} u^T B_m \rightarrow N(0, 1) \text{ weakly under } P_{\theta^*, \tilde{\lambda}(\xi)}, \]

which means that

\[ \frac{1}{n^{1/2}} \sum_{m=1}^{n} u^T B_m \rightarrow N(0, u^T J_\theta^* u) \text{ weakly under } P_{\theta^*, \tilde{\lambda}(\xi)}, \]

where \( (J_\theta^*), k, l = \tilde{\Lambda}_\theta^* (V_k^* V_l^*) - \Lambda_\theta^* (\tilde{\Pi}_k V_k^* \times \tilde{\Pi}_l V_l^*) \). Moreover,

\[ \frac{1}{n} \sum_{m=1}^{n} \tilde{E}_{\theta^*, \tilde{\lambda}(\xi)} \left( B_m^T B_m \right) \rightarrow J_\theta^*. \]

Then, equation (51) shows that

\[ n^{-1/2} \nabla l_n(\theta^*, \xi) \rightarrow N(0, J_\theta^*) \text{ weakly under } P^* \]

and

\[ \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}^* (\nabla l_n(\theta^*, \xi)^T \nabla l_n(\theta^*, \xi)) = J_\theta^*. \]

\[ \square \]

D.3. Exponential memorylessness of the Hessian of the prediction filter and consequences

The chain \( \{ \tilde{Z}_n \} \) lies in the space \( \mathbf{E}^* = K \times \mathcal{Y} \times S_+ \times \Sigma \times \Sigma_2 \) (where \( \Sigma_2 = \{ \tau = (\tau_{k,l})/\tau_{k,l} : K \rightarrow \mathbb{R} \text{ and for all } \nu \text{ in } K, \tau(\nu) \text{ is symmetric and positive} \} \). We first need to find a recurrence relation between \( \nabla^2 f_{\theta, n+1}^\xi \) and \( \nabla^2 f_{\theta, n}^\xi \). Using recurrence relation (35) between \( \nabla f_{\theta, n+1}^\xi \) and \( \nabla f_{\theta, n}^\xi \), we obtain that \( \{ \nabla^2 f_{\theta, n}^\xi \} \) is a functional autoregressive process

\[ \partial^2_{k,l} f_{\theta, n}^\xi (u) = \int_0^1 a_{\theta}(Y_n, f_{\theta, n}^\xi(u)) \partial^2_{k,l} f_{\theta, n}^\xi (u) d\gamma(u) + T_{\theta, k,l}(Y_n, f_{\theta, n}^\xi, \nabla f_{\theta, n}^\xi)(v), \]

where

\[ \square \]
for as in the preceding sections, we prove an exponential lack of memory of the initial conditions

\[ T_{\theta,k;l}(y, f, \sigma)(v) = \frac{\int \partial^2_{k,l} q_\theta(u, v) g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|w)f(w)dy(w)} + \int \frac{\partial q_\theta(u, v) g_\theta(y|u)\sigma_l(u)}{\int g_\theta(y|w)f(w)dy(w)} \times \left( \partial_l q_\theta(u, v) - \frac{\partial l q_\theta(w, v) g_\theta(y|w)f(w)dy(w)}{\int g_\theta(y|w)f(w)dy(w)} \right) dy(u) + \int \frac{\partial q_\theta(u, v) g_\theta(y|u)\sigma_l(u)}{\int g_\theta(y|w)f(w)dy(w)} \times \left( \partial_k q_\theta(u, v) - \frac{\partial k q_\theta(w, v) g_\theta(y|w)f(w)dy(w)}{\int g_\theta(y|w)f(w)dy(w)} \right) dy(u) + \int \frac{\partial^2_{k,l} g_\theta(y|u)f(u)}{\int g_\theta(y|w)f(w)dy(w)} \left[ q_\theta(u, v) - \frac{\int q_\theta(w, v) g_\theta(y|w)f(w)dy(w)}{\int g_\theta(y|w)f(w)dy(w)} \right] dy(u) \]

Relation (52) can be iterated to obtain $\nabla^2 f_{\theta,n+1} = \alpha_n^{-m+1}(Y_n^m, f_{\theta,n}, \nabla f_{\theta,n}^m, \nabla^2 f_{\theta,n}^m; \theta)$. Just as in the preceding sections, we prove an exponential lack of memory of the initial conditions for $\alpha_n^{-m+1}(\cdot; \theta)$.

**Proposition 10.** Under Assumptions 1 and 9, there exists $C > 0$ such that

$$ \sup_{x_1, \ldots, x_n} \int \alpha_{n-m+1}^{k,l}(y_n^m, f, \sigma, \tau; \theta) - \alpha_{n-m+1}^{k,l}(y_n^m, f', \sigma', \tau'; \theta) \leq C(n-m)^2(1 - \epsilon)^{n-m+1} \left[ \left\| \tau_{k,l} - \tau_{k,l}' \right\|_1 + \left\| \sigma_k - \sigma_k' \right\|_1 + \left\| \sigma_l - \sigma_l' \right\|_1 \right] \times (1 + \left\| \sigma_k \right\|_1 + \left\| \sigma_l \right\|_1 + \left\| \sigma_k' \right\|_1 + \left\| \sigma_l' \right\|_1)$$

Then we define a set of Lipschitz functions on $E^n$. 

Definition 3. Lip(E") is the set of vector-valued functions \( h = (h_{k,l})_{1 \leq k, l \leq p} \) with each \( h_{k,l} \) being a measurable function on \( E" \) such that, for all \( k, l \) in \( \{1 \ldots p\} \), for all \( (x, y) \) in \( K \times \mathcal{F} \), there exist \( \text{lip}^n(h(x, y)) \) and \( k^*(h(x, y)) \) such that, for any \( f, f' \) in \( S^+ \), \( \sigma, \sigma' \) in \( \Sigma \), and \( \tau, \tau' \) in \( \Sigma_2 \),

\[
|h_{k,l}(x, y, f, \sigma, \tau) - h_{k,l}(x, y, f', \sigma', \tau')| \\
\leq \text{lip}^n(h(x, y))[||\tau_{k,l} - \tau_{k,l}'||_1 + (||\sigma_k - \sigma_k'||_1 + ||\sigma_l - \sigma_l'||_1) \\
	imes (1 + ||\sigma_k||_1 + ||\sigma_k'||_1 + ||\sigma_l||_1 + ||\sigma_l'||_1) \\
+ ||f - f'||_1[||\tau_{k,l}||_1 + ||\tau_{k,l}'||_1 + (1 + ||\sigma_k||_1 + ||\sigma_k'||_1 + ||\sigma_l||_1 + ||\sigma_l'||_1)]^2] \\
|h_{k,l}(x, y, f, \sigma, \tau)| \leq k^*(h(x, y))[||\tau_{k,l}||_1 + (1 + ||\sigma_k||_1 + ||\sigma_l||_1)^2] \\
\text{and} \\
\text{lip}^n(h) = \text{ess sup}_x \int \text{lip}^n(h(x, y))g_\theta(y|x)d\nu(y) < \infty, \\
k^*(h) = \text{ess sup}_x \int k^*(h(x, y))g_\theta(y|x)d\nu(y) < \infty.
\]

The transition kernel \( \hat{\Pi} \) satisfies a contraction condition for these functions; as a result, we can prove that the chain is ergodic.

Proposition 11. Under Assumptions 1 and 9, there exist finite constants \( \hat{\rho} \in ]\rho, 1[ \) and \( C > 0 \), and for all \( z, z' \) in \( E" \), \( h \) in Lip(E") and \( n \geq 1 \),

\[
|\hat{\Pi}^n h_{k,l}(z) - \hat{\Pi}^n h_{k,l}(z')| \leq C(k^*(h))\hat{\rho}^n[||\alpha^k_l(y, f, \sigma, \tau; \theta)||_1 \\
+ ||\alpha^k_l(y', f', \sigma', \tau'; \theta)||_1 + (1 + ||\Psi_k^1(y, f, \sigma; \theta)||_1 + ||\Psi_1^1(y, f, \sigma; \theta)||_1) \\
+ ||\Psi_k^1(y', f', \sigma'; \theta)||_1 + ||\Psi_1^1(y', f', \sigma'; \theta)||_1)]^2].
\]

Corollary 4. Under Assumptions 1 and 9, the Markov chain \( \hat{Z}_n \) is geometrically ergodic. There exist finite constants \( \hat{\rho} \in ]\rho, 1[ \) and \( C > 0 \) such that, for all \( h \) in Lip(E") , there exists a finite constant \( \hat{\Lambda}_\theta(h) \), such that, for all \( z \) in \( E" \) and for all \( n \geq 1 \),

\[
|\hat{\Pi}^n h_{k,l}(z) - \hat{\Lambda}_\theta(h)| \leq C \times (\text{lip}^n(h) + k^*(h))\hat{\rho}^n[||\alpha^k_l(y, f, \sigma, \tau; \theta)||_1 + (1 + ||\Psi_k^1(y, f, \sigma; \theta)||_1 + ||\Psi_1^1(y, f, \sigma; \theta)||_1) \\
+ ||\Psi_k^1(y', f', \sigma'; \theta)||_1 + ||\Psi_1^1(y', f', \sigma'; \theta)||_1)]^2].
\] (53)

For any initial probability measure \( \hat{\lambda} \) on \( (E", \mathcal{B}(E")) \) and for any function \( h \) in Lip(E"),

\[
\hat{\Lambda}_\theta(h) = \lim_{n \to \infty} \hat{\mathbb{E}}_{\hat{\lambda}}(h(\hat{Z}_n)),
\]

and there exists a solution in \( E" \) to the Poisson equation

\[
(I - \hat{\Pi}_0)\hat{\Lambda} = h - \hat{\Lambda}_\theta(h).
\] (54)
Proofs of Proposition 11 and Corollary 4 follow the same lines as for Proposition 9 and Corollary 3.

**Proof of Proposition 6.** Differentiating the expression for the gradient of the log-likelihood, we obtain, for all \( k, l \) in \( \{1, \ldots, p\} \),

\[
\frac{1}{n} \partial_{k,l}^2 l_n(\theta, \zeta) = \frac{1}{n} \sum_{m=0}^{n-1} i_{\theta,kl}(\hat{Z}_m) \quad \hat{P}_{\theta,kl(\zeta)} \text{ a.s.,} \tag{55}
\]

where

\[
i_{\theta,kl}(x, y, f, \sigma, \tau) \triangleq \frac{\int g_\theta(y|u)T_{kl}(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} + \frac{\int \partial_{kl}^2 g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)}
\]

\[
+ \frac{\int \partial_{kl} g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} \sigma(u) \alpha(u) dy(u) + \frac{\int \partial_{kl} g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} \sigma(u) \alpha(u) dy(u)
\]

\[
- \left( \frac{\int \partial_{kl} g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} + \frac{\int g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} \right) \frac{\int g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)}
\]

\[
\times \left( \frac{\int \partial_{kl} g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} + \frac{\int g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)} \right) \frac{\int g_\theta(y|u)f(u)dy(u)}{\int g_\theta(y|u)f(u)dy(u)}
\]

By a long but simple exercise in calculus, we can check that \( i_{\theta,kl} \) is in Lip(\( E'' \)) and verify (15), assuming that \( \Delta_{2,s}, \Delta_{2,t}, \Delta''_{s}(\Delta\Delta'')_{1,1} \) and \( (\Delta\Delta')_{2,2} \) are finite for some \( r, s, t \geq 1 \). Using similar arguments as for Proposition 3, we obtain

\[
\frac{1}{n} \nabla^2 l_n(\theta^*, \zeta) \to \hat{\Lambda}^*_{\theta^*}(i_{\theta^*}) \triangleq -I_{\theta^*}.
\]

Then

\[
I_{\theta^*} = -\lim_{n \to \infty} E^* \left( \frac{1}{n} \nabla^2 l_n(\theta^*, \zeta) \right).
\]

As permutation between differentiation and integration is valid in our case,

\[
J_{\theta^*} = \lim_{n \to \infty} E^* \left( \frac{1}{n} \nabla l_n(\theta^*, \zeta)^T \nabla l_n(\theta^*, \zeta) \right) = -\lim_{n \to \infty} E^* \left( \frac{1}{n} \nabla^2 l_n(\theta^*, \zeta) \right) = I_{\theta^*}.
\]

**Proof of Proposition 7.** This proof is the same as for the quantity \( l_n(\theta, \zeta) - l_n(\theta', \zeta) \), generalizing Lemma 1 to the gradient and the Hessian of the prediction filter.
Acknowledgements

The authors would like to thank Elisabeth Gassiat and Eric Moulines for helpful and fruitful discussions. The authors are also indebted to Olivier Capp for reading and correcting earlier versions of this manuscript.

References


Received September 1999 and revised February 2001