The maximizing set of the asymptotic normalized log-likelihood for partially observed Markov chains

Randal Douc, François Roueff, Tepmony Sim

THE MAXIMIZING SET OF THE ASYMPTOTIC NORMALIZED LOG-LIKELIHOOD FOR PARTIALLY OBSERVED MARKOV CHAINS

RANDAL DOUC[1], FRANÇOIS ROUEFF[2], AND TEPMONY SIM[2]

Abstract. This paper deals with a parametrized family of partially observed bivariate Markov chains. We establish, under very mild assumptions, that the limit of the normalized log-likelihood is maximized when the parameter belongs to the equivalence class of the true parameter, which is a key feature for obtaining the consistency of the Maximum Likelihood Estimators (MLE) in well-specified models. This result is obtained in the general framework of partially dominated models. We examine two specific cases of interest, namely, Hidden Markov models and Observation-Driven times series. In contrast with previous approaches, the identifiability is addressed by relying on the unicity of the invariant distribution of the Markov chain associated to the complete data, regardless its rate of convergence to the equilibrium.

1. Introduction

Maximum likelihood estimation is a widespread method for identifying a parametric model of a time series from a sample of observations. Under a well-specified model assumption, it is of prime interest to show the consistency of the estimator, that is, its convergence to the true parameter, say $\theta_\star$, as the sample size goes to infinity. The proof generally involves two important steps: 1) the maximum likelihood estimator (MLE) converges to the maximizing set $\Theta_\star$ of the asymptotic normalized log-likelihood 2) the maximizing set indeed reduces to the true parameter. The second step is usually referred to as solving the *identifiability* problem but it can actually be split in two sub-problems: 2.1) show that any parameter in $\Theta_\star$ yields the same distribution for the observations as for the true parameter and 2.2) show that for a sufficiently large sample size, the set of such parameters reduces to $\theta_\star$. Problem 2.2 can be difficult to solve, see [2, 18] and the references therein for recent advances in the case of hidden Markov models (HMM). Nevertheless Problem 2.1 can be solved independently, and, among

Date: November 5, 2014.

2010 *Mathematics Subject Classification.* Primary 60J05, 62F12; Secondary 62M05, 62M10.

Key words and phrases. consistency, ergodicity, hidden Markov models, maximum likelihood, observation-driven models, time series of counts.

with Step 1 above, it directly yields that the maximum likelihood estimator is consistent in a weakened sense, namely, that the estimated parameter converges to a set of parameters, all of them corresponding to the same distribution of the observed sample. This consistency result is referred to as *equivalence-class consistency*, as introduced by [24]. In this contribution our goal is to provide a general approach to solve Problem 2.1 in the general framework of partially observed Markov models, which include many classes of models of interest, see e.g. [28] or [16]. The novel aspect of this work is that the result mainly relies on the unicity of the invariant distribution of the Markov chain associated to the complete data, regardless its rate of convergence to the equilibrium. We then detail how this approach applies in the context of two important subclasses of partially observed Markov models, namely, the class of HMMs and the class of observation-driven time series models.

In the context of HMMs, the consistency of the MLE is of primary importance, either as a subject of study (see [11, 12, 24]) or as a basic assumption (see [4, 22]). The characterization of the maximizing set $\Theta_\star$ of the asymptotic log-likelihood (and thus the equivalence-class consistency of the MLE) remains a delicate question for HMMs. As an illustration, we consider the following example.

Example 1. Set $\mathsf{X} = \mathbb{R}^+$, $\mathcal{X} = \mathcal{B}(\mathbb{R}^+)$, $\mathsf{Y} = \mathbb{R}$ and $\mathcal{Y} = \mathcal{B}(\mathbb{R})$ and define a Hidden Markov model on $\mathsf{X} \times \mathsf{Y}$ by the following recursions:

$$\begin{aligned} X_k &= (X_{k-1} + U_k - m)^+ \ , \\ Y_k &= aX_k + V_k \ , \end{aligned} \tag{1}$$

where $(m, a) \in (0, \infty) \times \mathbb{R}$ and where the sequence $\{(U_k, V_k), k \in \mathbb{N}\}$ is i.i.d and independent from X_0.

The Markov chain $\{X_n\}_{n\in\mathbb{N}}$ was proposed by [30] and further considered by [21] as an example of polynomially ergodic Markov chain, under specific assumptions on the U_k's. Namely, if the U_k's are centered and $\mathbb{E}[\mathrm{e}^{\lambda U_k^+}] = \infty$ for any $\lambda > 0$, it can be shown that the chain $\{X_k\}$ is not geometrically ergodic (see Lemma 13 below). In such a situation, the exponential separation of measures condition introduced in [11] seems difficult to check. We will show, nevertheless, in Proposition 14, that under some mild conditions the chain $\{X_k\}$ is ergodic and the equivalence-class consistency holds.

Observation-driven time series models were introduced by [7] and later considered, among others, by [29], [8], [17], [26], [15] and [10]. The celebrated GARCH(1,1) model introduced by [5], is an observation-driven model as well as most of the models derived from this one, see [6] for a list of some of them. This class of models has the nice feature that the (conditional) likelihood and its derivatives are easy to compute. The consistency of the MLE, however, can be cumbersome and is often derived using computations specific to the studied model. When the observed variable is discrete, general consistency results have been obtained only recently in [9] or [10] (see

also in [20] for the existence of stationary and ergodic solutions to some observation-driven time series models). However, in these contributions, the way of proving that the maximizing set $\Theta_\star$ reduces to $\{\theta_\star\}$ requires checking specific conditions in each given examples and seems difficult to assert in a more general context, for instance when the distribution of the observations given the hidden variable also depends on an unknown parameter. Let us describe two such examples. The first one was introduced in [31]. Up to our knowledge the consistency of the MLE has not been treated for this model.

Example 2. The Negative Binomial Integer-Valued GARCH (NBIN-GARCH) model defined by

$$\begin{aligned} &X_{k+1} = \omega + aX_k + bY_k \ , \\ &Y_{k+1}|X_{0:k+1}, Y_{0:k} \sim \mathcal{NB}\left(r, \frac{1}{1+X_{k+1}}\right) \ , \end{aligned} \tag{2}$$

where X_k takes values in $\mathsf{X} = \mathbb{R}_+$, Y_k takes values in $\mathbb{Z}_+$ and $\theta = (\omega, a, b, r) \in (0,\infty)^4$ is an unknown parameter. In (2), $\mathcal{NB}(r,p)$ denotes the negative binomial distribution with parameters $r > 0$ and $p \in (0,1)$, which has probability $\frac{\Gamma(k+r)}{k!\Gamma(r)} p^r (1-p)^k$ for all $k \geq 0$, where Γ stands for the Gamma function.

The second example, proposed by [19] and [1], is another natural extension of GARCH processes, where the usual Gaussian conditional distribution of the observations given the hidden volatility variable is replaced by a mixture of Gaussian distributions given a hidden vector volatility variable. Up to our knowledge, the usual consistency proof of the MLE for the GARCH cannot be directly adapted to this model.

Example 3 (Normal Mixture GARCH Model)**.** The normal mixture GARCH model (NM(d)-GARCH(1,1)) is defined by

$$\begin{aligned} &\mathbf{X}_{k+1} = \boldsymbol{\omega} + \mathbf{A}\mathbf{X}_k + Y_k^2 \mathbf{b} \ , \\ &Y_{k+1}|\mathbf{X}_{0:k+1}, Y_{0:k} \sim G^\theta(\mathbf{X}_{k+1}; \cdot) \ , \\ &G^\theta(\mathbf{x}; \mathrm{d}y) = \left(\sum_{\ell=1}^d \gamma_\ell \frac{\mathrm{e}^{-y^2/2x_\ell}}{(2\pi x_\ell)^{1/2}} \right) \mathrm{d}y \ , \quad \mathbf{x} \in (0,\infty)^d, \ y \in \mathbb{R} \ , \end{aligned} \tag{3}$$

where d is a positive integer; $\mathbf{X}_k = [X_{1,k} \dots X_{d,k}]^T$ takes values in $\mathsf{X} = \mathbb{R}_+^d$; $\boldsymbol{\gamma} = [\gamma_1 \dots \gamma_d]^T$ a d-dimensional vector of mixture coefficients belonging to the d-dimensional simplex $\mathsf{P}_d = \{\boldsymbol{\gamma} \in \mathbb{R}_+^d \ : \ \sum_{\ell=1}^d \gamma_\ell = 1\}$; $\boldsymbol{\omega}$, $\mathbf{b}$ are d-dimensional vector parameters with positive and non-negative entries, respectively, $\mathbf{A}$ is a $d \times d$ matrix parameter with non-negative entries and $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b})$.

The paper is organized as follows. The main result (Theorem 3) shows that the argmax of the limiting criterion is reduced to the equivalence class of the true parameter. It is stated and proved in Section 2. We then focus

on the kernel involved in the assumptions, and we show that in practice, it can be obtained as a backward limit. Two important classes of partially observed Markov models are then considered.

- First, the hidden Markov models described in Section 3, for which the equivalence-class consistency of the MLE is derived under simplified assumptions. The polynomially ergodic HMM of Example 1 is treated as an application of this result.
- Second, the observation-driven time series models described in Section 4. The obtained results apply to the models of Example 2 and Example 3, where the generating process of the observations may also depend on the parameter.

The technical proofs are gathered in the Appendix.

2. A general approach to identifiability

2.1. **General setting and notation: partially dominated and partially observed Markov models.** Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two Borel spaces, that is, measurable spaces that are isomorphic to a Borel subset of $[0,1]$ and let Θ be a set of parameters. Consider a statistical model determined by a class of Markov kernels $\{K^\theta,\ \theta \in \Theta\}$ on $(\mathsf{X} \times \mathsf{Y}) \times (\mathcal{X} \otimes \mathcal{Y})$. Throughout the paper, we denote by $\mathbb{P}^\theta_\xi$ the probability (and by $\mathbb{E}^\theta_\xi$ the corresponding expectation) induced on $(\mathsf{X} \times \mathsf{Y})^{\mathbb{Z}_+}$ by a Markov chain $\{(X_k, Y_k)\}_{k\geq 0}$ with transition kernel K^θ and initial distribution ξ on $\mathsf{X} \times \mathsf{Y}$. In the case where ξ is a Dirac mass at (x, y) we shall simply write $\mathbb{P}^\theta_{(x,y)}$.

For partially observed Markov chains, that is, when only a sample $Y_{1:n} = (Y_1, \ldots, Y_n) \in \mathsf{Y}^n$ of the second component is observed, it is convenient to write K^θ as

$$K^\theta((x,y); \mathrm{d}x'\mathrm{d}y') = Q^\theta((x,y); \mathrm{d}x') G^\theta((x,y,x'); \mathrm{d}y') \ , \tag{4}$$

where Q^θ and G^θ are probability kernels on $(\mathsf{X}\times\mathsf{Y})\times\mathcal{X}$ and on $(\mathsf{X}\times\mathsf{Y}\times\mathsf{X})\times\mathcal{Y}$, respectively.

We consider the following general setting.

Definition 1. We say that the Markov model $\{K^\theta,\ \theta \in \Theta\}$ of the form (4) is partially dominated if there exists a σ-finite measure ν on Y such that for all $(x, y), (x', y') \in \mathsf{X} \times \mathsf{Y}$,

$$G^\theta((x,y,x'); \mathrm{d}y') = g^\theta((x,y,x'); y')\nu(\mathrm{d}y') \ , \tag{5}$$

where the conditional density function g^θ moreover satisfies

$$g^\theta((x,y,x'); y') > 0 \quad \text{for all} \quad (x,y), (x',y') \in \mathsf{X} \times \mathsf{Y} \ . \tag{6}$$

It follows from (5) that, for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$ $A \in \mathcal{X}$ and $B \in \mathcal{Y}$,

$$K^\theta\left((x,y); A \times B\right) = \int_B \kappa^\theta\langle y, y'\rangle\left(x; A\right) \nu(\mathrm{d}y') \ ,$$

where, for all $y, y' \in \mathsf{Y}$, $\kappa^\theta\langle y, y'\rangle$ is a kernel defined on $(\mathsf{X}, \mathcal{X})$ by

$$\kappa^\theta\langle y, y'\rangle \left(x; \mathrm{d}x'\right) = Q^\theta((x, y); \mathrm{d}x') g^\theta((x, y, x'); y') \; . \tag{7}$$

Remark 1. Note that, in general, the kernel $\kappa^\theta\langle y, y'\rangle$ is unnormalized since $\kappa^\theta\langle y, y'\rangle (x; \mathsf{X})$ may be different from one. Moreover, we have for all $(x, y, y') \in \mathsf{X} \times \mathsf{Y} \times \mathsf{Y}$,

$$\kappa^\theta\langle y, y'\rangle (x; \mathsf{X}) = \int_{\mathsf{X}} Q^\theta((x, y); \mathrm{d}x') g^\theta((x, y, x'); y') > 0 \; , \tag{8}$$

where the positiveness follows from the fact that $Q^\theta((x, y); \cdot)$ is a probability on $(\mathsf{X}, \mathcal{X})$ and Condition (6).

In well-specified models, it is assumed that the observations $Y_{1:n}$ are generated from a process $\{(X_k, Y_k)\}_{k\geq 0}$, which follows the distribution $\mathbb{P}_{\xi_\star}^{\theta_\star}$ associated to an unknown parameter $\theta_\star \in \Theta$ and an unknown initial distribution $\xi_\star$ (usually, $\xi_\star$ is such that, under $\mathbb{P}_{\xi_\star}^{\theta_\star}$, $\{Y_k\}_{k\geq 0}$ is a stationary sequence). To form a consistent estimate of $\theta_\star$ on the basis of the observations $\{Y_k\}_{k\geq 1}$ only, i.e., without access to the hidden process $\{X_k\}_{k\geq 0}$, we define the maximum likelihood estimator (MLE) $\hat{\theta}_{\xi,n}$ by

$$\hat{\theta}_{\xi,n} \in \operatorname*{argmax}_{\theta\in\Theta} L_{\xi,n}(\theta) \; ,$$

where $L_{\xi,n}(\theta)$ is the (conditional) log-likelihood function of the observations under parameter θ with some arbitrary initial distribution ξ on $\mathsf{X}\times\mathsf{Y}$. In this context, a classical way (see for example [24]) to prove the consistency of a maximum likelihood type estimator $\hat{\theta}_{\xi,n}$ may be decomposed in the following steps. The first step is to show that $\hat{\theta}_{\xi,n}$ is, with probability tending to one, in a neighborhood of the set

$$\Theta_\star := \operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star} \left[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\right] \; . \tag{9}$$

This formula involves two quantities that have not yet been defined since they may require additional assumptions: first, the expectation $\tilde{\mathbb{E}}^\theta$, which corresponds to the distribution $\tilde{\mathbb{P}}^\theta$ of a sequence $\{Y_k\}_{k\in\mathbb{Z}}$ in accordance with the kernel K^θ, and second, the density $p^{\theta,\theta_\star}(\cdot|\cdot)$, which shows up when taking the limit, under $\tilde{\mathbb{P}}^{\theta_\star}$, of the $\tilde{\mathbb{P}}^\theta$-conditional density of Y_1 given its m-order past, as m goes to infinity. In many cases, such quantities appear naturally because the model is ergodic and the normalized log-likelihood $L_{\xi,n}(\theta)$ can be approximated by

$$\frac{1}{n}\sum_{k=1}^{n} \ln p^{\theta,\theta_\star}(Y_k|Y_{-\infty:k-1}) \; .$$

We will provide below some general assumptions, Assumptions (A-1) and (K-1), that yield precise definitions of $\tilde{\mathbb{P}}^\theta$ and $p^{\theta,\theta'}(\cdot|\cdot)$.

The second step consists in proving that the set $\Theta_\star$ in (9) is related to the true parameter $\theta_\star$ in an exploitable way. Ideally, one could have $\Theta_\star = \{\theta_\star\}$,

which would yield the consistency of $\hat{\theta}_{\xi,n}$ for estimating $\theta_\star$. In this work, our first objective is to provide general assumptions which ensure that $\Theta_\star$ is exactly the set of parameters θ such that $\tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta_\star}$. Then, if the model $\{\tilde{\mathbb{P}}^\theta\}_{\theta\in\Theta}$ is identifiable, consistency of $\hat{\theta}_{\xi,n}$ directly follows. On the other hand, this result is also of interest in the case where identifiability is not true (for instance in models involving mixtures) or is difficult to check, since it guarantees that the estimator converges to the set of parameters compatible with the true stationary distribution of the observations.

To conclude with our general setting, we state the main assumption on the model and some subsequent notation and definitions used throughout the paper.

(A-1) For all $\theta \in \Theta$, the transition kernel K^θ admits a unique invariant probability π^θ.

We now introduce some important notation used throughout the paper.

Definition 2. Under Assumption (A-1), we denote by π_1^θ and π_2^θ the marginal distributions of π^θ on X and Y, respectively and by $\mathbb{P}^\theta$ and $\tilde{\mathbb{P}}^\theta$ the probability distributions defined respectively as follows.

a) $\mathbb{P}^\theta$ denotes the extension of $\mathbb{P}^\theta_{\pi^\theta}$ on the whole line $(\mathsf{X}\times\mathsf{Y})^{\mathbb{Z}}$.
b) $\tilde{\mathbb{P}}^\theta$ is the corresponding projection on the component $\mathsf{Y}^{\mathbb{Z}}$.

Moreover, for all $\theta, \theta' \in \Theta$, we write $\theta \sim \theta'$ if and only if $\tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta'}$, where $\tilde{\mathbb{P}}^\theta$ and $\tilde{\mathbb{P}}^{\theta'}$ are the shift-invariant distributions defined above on $\mathsf{Y}^{\mathbb{Z}}$. This defines an equivalence relation on the parameter set Θ and the corresponding equivalence class of θ is denoted by $[\theta] := \{\theta' \in \Theta \ : \ \theta \sim \theta'\}$.

The equivalence relationship $\sim$ was introduced by [24] as an alternative to the classical identifiability condition. The probability distributions $\mathbb{P}^\theta$ and $\tilde{\mathbb{P}}^\theta$ are more formally defined by setting, for all $m \in \mathbb{Z}$ and $B \in \mathcal{Y}^{\otimes(m+\mathbb{Z}_+^*)}$,

$$\tilde{\mathbb{P}}^\theta\left(\mathsf{Y}^{m+\mathbb{Z}_-}\times B\right) = \mathbb{P}^\theta\left(\mathsf{X}^{\mathbb{Z}}\times\left(\mathsf{Y}^{m+\mathbb{Z}_-}\times B\right)\right) = \mathbb{P}^\theta_{\pi^\theta}\left(\mathsf{X}^{m+\mathbb{Z}_+^*}\times B\right)\ , \tag{10}$$

or equivalently, using the canonical functions Y_k, $k \in \mathbb{Z}$,

$$\tilde{\mathbb{P}}^\theta\left(Y_{m+1:\infty}\in B\right) = \mathbb{P}^\theta\left(Y_{m+1:\infty}\in B\right) = \mathbb{P}^\theta_{\pi^\theta}\left(Y_{m+1:\infty}\in B\right)\ . \tag{11}$$

Here and in what follows, we use the same notation Y_k both for the canonical projection defined on $\mathsf{Y}^{\mathbb{Z}}$ and for the original one defined on $(\mathsf{X}\times\mathsf{Y})^{\mathbb{Z}_+}$. We also use the symbols $\mathbb{E}^\theta$ and $\tilde{\mathbb{E}}^\theta$ to denote the expectations corresponding to $\mathbb{P}^\theta$ and $\tilde{\mathbb{P}}^\theta$, respectively.

2.2. **Main result.** Assumption (A-1) is supposed to hold all along this section and $\mathbb{P}^\theta$, $\tilde{\mathbb{P}}^\theta$ and $\sim$ are given in Definition 2. Our main result is stated under the following general assumption.

(K-1) For all $\theta \neq \theta'$ in Θ, there exists a probability kernel $\Phi^{\theta,\theta'}$ on $\mathsf{Y}^{\mathbb{Z}_-} \times \mathcal{X}$ such that for all $A \in \mathcal{X}$,

$$\frac{\int_{\mathsf{X}} \Phi^{\theta,\theta'}(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^{\theta}\langle Y_0, Y_1\rangle (x_0; A)}{\int_{\mathsf{X}} \Phi^{\theta,\theta'}(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^{\theta}\langle Y_0, Y_1\rangle (x_0; \mathsf{X})} = \Phi^{\theta,\theta'}(Y_{-\infty:1}; A), \quad \tilde{\mathbb{P}}^{\theta'}\text{-a.s.}$$

Note that from Remark 1, the denominator in the left-hand side of the last displayed equation is strictly positive, which ensures that the ratio is well defined. Moreover, this denominator can be written as $p^{\theta,\theta'}(Y_1|Y_{-\infty:0})$, where, for all $y \in \mathsf{Y}$ and $y_{-\infty:0} \in \mathsf{Y}^{\mathbb{Z}_-}$,

$$p^{\theta,\theta'}(y|y_{-\infty:0}) = \int_{\mathsf{X}} \Phi^{\theta,\theta'}(y_{-\infty:0}; \mathrm{d}x_0)\kappa^{\theta}\langle y_0, y\rangle (x_0; \mathsf{X}) \tag{12}$$

is a conditional density with respect to the measure ν, since, for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $\int \kappa^{\theta}\langle y, y'\rangle (x; \mathsf{X})\, \nu(\mathrm{d}y') = 1$.

Since Y is a Borel space, [23, Theorem 6.3] applies and the conditional distribution of $Y_{1:n}$ given $Y_{-\infty:0}$ defines a probability kernel. Since $\tilde{\mathbb{P}}^{\theta}(Y_{1:n} \in \cdot)$ is dominated by $\nu^{\otimes n}$, this in turns defines a conditional density with respect to $\nu^{\otimes n}$ which we denote by $p_n^{\theta}(\cdot|\cdot)$, so that, for all $B \in \mathcal{Y}^{\otimes n}$,

$$\tilde{\mathbb{P}}^{\theta}\left(Y_{1:n} \in B \,|\, Y_{-\infty:0}\right) = \int_B p_n^{\theta}(y_{1:n}|Y_{-\infty:0})\, \nu(\mathrm{d}y_1) \ldots \nu(\mathrm{d}y_n), \quad \tilde{\mathbb{P}}^{\theta}\text{-a.s.} \tag{13}$$

Let us now state the main result.

Theorem 3. *Assume that (A-1) holds and define $\mathbb{P}^{\theta}$, $\tilde{\mathbb{P}}^{\theta}$ and $[\theta]$ as in Definition 2. Suppose that Assumption (K-1) holds. For all $\theta, \theta' \in \Theta$, define $p^{\theta,\theta'}(Y_1|Y_{-\infty:0})$ by (12) if $\theta \neq \theta'$ and by $p^{\theta,\theta}(Y_1|Y_{-\infty:0}) = p_1^{\theta}(Y_1|Y_{-\infty:0})$ as in (13) otherwise. Then for all $\theta_\star \in \Theta$, we have*

$$\operatorname*{argmax}_{\theta \in \Theta} \tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\right] = [\theta_\star]\,. \tag{14}$$

Before proving Theorem 3, we first extend the definition of the conditional density on Y in (12) to a conditional density on Y^n.

Definition 4. For every positive integer n and $\theta \neq \theta' \in \Theta$, define the function $p_n^{\theta,\theta'}(\cdot|\cdot)$ on $\mathsf{Y}^n \times \mathsf{Y}^{\mathbb{Z}_-}$ by

$$p_n^{\theta,\theta'}(y_{1:n}|y_{-\infty:0}) = \int_{\mathsf{X}^n} \Phi^{\theta,\theta'}(y_{-\infty:0}; \mathrm{d}x_0) \prod_{k=0}^{n-1} \kappa^{\theta}\langle y_k, y_{k+1}\rangle (x_k; \mathrm{d}x_{k+1})\,. \tag{15}$$

Again, it is easy to check that each $p_n^{\theta,\theta'}(\,\cdot\,|y_{-\infty:0})$ is indeed a density on Y^n. Assumption (K-1) ensures that, these density functions moreover satisfy the successive conditional formula, as for conditional densities, provided that we restrict ourselves to sequences in a set of $\tilde{\mathbb{P}}^{\theta'}$-probability one, as stated in the following lemma.

Lemma 5. *Suppose that Assumption (K-1) holds and let $p_n^{\theta,\theta'}(\cdot|\cdot)$ be as in Definition 4. Then for all $\theta, \theta' \in \Theta$ and $n \geq 2$, we have*

$$p_n^{\theta,\theta'}(Y_{1:n}|Y_{-\infty:0}) = p_1^{\theta,\theta'}(Y_n|Y_{-\infty:n-1}) p_{n-1}^{\theta,\theta'}(Y_{1:n-1}|Y_{-\infty:0}), \quad \tilde{\mathbb{P}}^{\theta'}\text{-a.s.} \tag{16}$$

The proof of this lemma is postponed to Section A.1 in Appendix A. We have now all the tools for proving the main result.

Proof of Theorem 3. Within this proof section, we shall drop the subscript n and respectively write $p^{\theta,\theta'}(y_{1:n}|y_{-\infty:0})$ and $p^{\theta}(y_{1:n}|y_{-\infty:0})$ instead of $p_n^{\theta,\theta'}(y_{1:n}|y_{-\infty:0})$ and $p_n^{\theta}(y_{1:n}|y_{-\infty:0})$ when no ambiguity occurs.

Let us denote

$$\Theta_\star := \operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\right] . \tag{17}$$

For all $\theta \in \Theta$, we have by conditioning on $Y_{-\infty:0}$ and by using (13),

$$\begin{aligned}
&\tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta_\star}(Y_1|Y_{-\infty:0})\right] - \tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\right] \\
&\quad = \tilde{\mathbb{E}}^{\theta_\star}\left[\tilde{\mathbb{E}}^{\theta_\star}\left[\ln \frac{p^{\theta_\star}(Y_1|Y_{-\infty:0})}{p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})}\middle| Y_{-\infty:0}\right]\right] \\
&\quad = \tilde{\mathbb{E}}^{\theta_\star}\left[\mathrm{KL}\left(p_1^{\theta_\star}(\,\cdot\,|Y_{-\infty:0})\|p_1^{\theta,\theta_\star}(\,\cdot\,|Y_{-\infty:0})\right)\right] . \qquad (18)
\end{aligned}$$

where $\mathrm{KL}(p\|q)$ denotes the Kullback-Leibler divergence between the densities p and q. The non-negativity of the Kullback-Leibler divergence shows that $\theta_\star$ belongs to the maximizing set on the left-hand set of (14). This implies

$$\operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\right] \supseteq [\theta_\star] , \tag{19}$$

where we used the following lemma.

Lemma 6. *Assume that (A-1) holds and define $\tilde{\mathbb{E}}^{\theta}$ and $[\theta]$ as in Definition 2. Suppose that for all $\theta \in \Theta$, $G(\theta)$ is a $\sigma(Y_{-\infty:\infty})$-measurable random variable such that, for all $\theta_\star \in \Theta$,*

$$\sup_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta)\right] = \tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta_\star)\right] .$$

Then, for all $\theta_\star \in \Theta$ and all $\theta' \in [\theta_\star]$, we have

$$\tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta')\right] = \sup_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta)\right] .$$

Proof. Take $\theta_\star \in \Theta$ and $\theta' \in [\theta_\star]$. Then we have, for all $\theta \in \Theta$, $\tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta)\right] = \tilde{\mathbb{E}}^{\theta'}\left[G(\theta)\right]$ and it follows that

$$\tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta')\right] = \tilde{\mathbb{E}}^{\theta'}\left[G(\theta')\right] = \sup_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta'}\left[G(\theta)\right] = \sup_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\left[G(\theta)\right] ,$$

which concludes the proof. □

The proof of the reverse inclusion of (19) is more tricky. Let us take $\theta \in \Theta_\star$ such that $\theta \neq \theta_\star$ and show that it implies $\theta \sim \theta_\star$. By (18), we have

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\mathrm{KL}\left(p_1^{\theta_\star}(\,\cdot\,|Y_{-\infty:0})\|p_1^{\theta,\theta_\star}(\,\cdot\,|Y_{-\infty:0})\right)\right] = 0\,.$$

Hence we have

$$p^{\theta_\star}(Y_1|Y_{-\infty:0}) = p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0}), \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

Applying Lemma 5 and using that $\tilde{\mathbb{P}}^{\theta_\star}$ is shift-invariant, this relation propagates to all $n \geq 2$, so that

$$p^{\theta_\star}\left(Y_{1:n}|Y_{-\infty:0}\right) = p^{\theta,\theta_\star}\left(Y_{1:n}|Y_{-\infty:0}\right), \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{20}$$

For any measurable function $H : \mathsf{Y}^n \to \mathbb{R}_+$, we get that

$$\begin{aligned}\tilde{\mathbb{E}}^{\theta_\star}\left[H(Y_{1:n})\right] &= \tilde{\mathbb{E}}^{\theta_\star}\left\{\tilde{\mathbb{E}}^{\theta_\star}\left[H(Y_{1:n})\frac{p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0})}{p^{\theta_\star}(Y_{1:n}|Y_{-\infty:0})}\middle| Y_{-\infty:0}\right]\right\} \\ &= \tilde{\mathbb{E}}^{\theta_\star}\left[\int H(y_{1:n})p^{\theta,\theta_\star}(y_{1:n}|Y_{-\infty:0})\nu^{\otimes n}(\mathrm{d}y_{1:n})\right]\,,\end{aligned}$$

where the last equality follows from (13). Using Definition 4 and Tonelli's theorem, we get

$$\begin{aligned}\tilde{\mathbb{E}}^{\theta_\star}\left[H(Y_{1:n})\right] &= \tilde{\mathbb{E}}^{\theta_\star}\int H(y_{1:n})\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0};\mathrm{d}x_0)\kappa^\theta\langle Y_0,y_1\rangle\left(x_0;\mathrm{d}x_1\right)\times \\ &\qquad \prod_{k=1}^{n-1}\kappa^\theta\langle y_k,y_{k+1}\rangle\left(x_k;\mathrm{d}x_{k+1}\right)\nu^{\otimes n}(\mathrm{d}y_{1:n})\,, \\ &= \tilde{\mathbb{E}}^{\theta_\star}\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0};\mathrm{d}x_0)\int H(y_{1:n})\kappa^\theta\langle Y_0,y_1\rangle\left(x_0;\mathrm{d}x_1\right)\times \\ &\qquad \prod_{k=1}^{n-1}\kappa^\theta\langle y_k,y_{k+1}\rangle\left(x_k;\mathrm{d}x_{k+1}\right))\nu^{\otimes n}(\mathrm{d}y_{1:n})\,, \\ &= \tilde{\mathbb{E}}^{\theta_\star}\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0};\mathrm{d}x_0)\mathbb{E}^\theta_{(x_0,Y_0)}\left[H(Y_{1:n})\right]\,, \\ &= \mathbb{E}^\theta_{\pi^{\theta,\theta_\star}}\left[H(Y_{1:n})\right]\,,\end{aligned}$$

where $\pi^{\theta,\theta_\star}$ is a probability on $\mathsf{X}\times\mathsf{Y}$ defined by

$$\pi^{\theta,\theta_\star}(A\times B) = \tilde{\mathbb{E}}^{\theta_\star}\left[\Phi^{\theta,\theta_\star}(Y_{-\infty:0};A)\mathbb{1}_B(Y_0)\right]\,,$$

for all $(A,B)\in\mathcal{X}\times\mathcal{Y}$. Consequently, for all $B\in\mathcal{Y}^{\otimes\mathbb{Z}_+^*}$,

$$\tilde{\mathbb{P}}^{\theta_\star}(\mathsf{Y}^{\mathbb{Z}_-}\times B) = \mathbb{P}^\theta_{\pi^{\theta,\theta_\star}}(\mathsf{X}^{\mathbb{Z}_+^*}\times B)\,. \tag{21}$$

If we had $\pi^\theta = \pi^{\theta,\theta_\star}$, then we could conclude that the two shift-invariant distributions $\tilde{\mathbb{P}}^{\theta_\star}$ and $\tilde{\mathbb{P}}^\theta$ are the same and thus $\theta\sim\theta_\star$. Therefore, to complete the proof, it only remains to show that $\pi^\theta = \pi^{\theta,\theta_\star}$, or by (A-1), equivalently, that $\pi^{\theta,\theta_\star}$ is an invariant distribution for K^θ.

Let us now prove this latter fact. Using that $\tilde{\mathbb{P}}^{\theta_\star}$ is shift-invariant and then conditioning on $Y_{-\infty:0}$, we have, for any $(A, B) \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned}\pi^{\theta,\theta_\star}(A \times B) &= \tilde{\mathbb{E}}^{\theta_\star}\left[\Phi^{\theta,\theta_\star}(Y_{-\infty:1}; A)\mathbb{1}_B(Y_1)\right] , \\ &= \tilde{\mathbb{E}}^{\theta_\star}\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0}, y_1; A)\mathbb{1}_B(y_1)\, p^{\theta_\star}(y_1|Y_{-\infty:0})\, \nu(\mathrm{d}y_1) , \\ &= \tilde{\mathbb{E}}^{\theta_\star}\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0}, y_1; A)\mathbb{1}_B(y_1)\, p^{\theta,\theta_\star}(y_1|Y_{-\infty:0})\, \nu(\mathrm{d}y_1) ,\end{aligned}$$

where, in the last equality, we used again (20). Then, using (K-1), we get

$$\begin{aligned}&\pi^{\theta,\theta_\star}(A \times B) \\ &= \tilde{\mathbb{E}}^{\theta_\star}\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle\,(x_0; \mathrm{d}x_1)\,\mathbb{1}_A(x_1)\mathbb{1}_B(y_1)\nu(\mathrm{d}y_1) , \\ &= \tilde{\mathbb{E}}^{\theta_\star}\int \Phi^{\theta,\theta_\star}(Y_{-\infty:0}; \mathrm{d}x_0)K^\theta((x_0, Y_0); A \times B) , \\ &= \pi^{\theta,\theta_\star}K^\theta(A \times B) .\end{aligned}$$

Thus, $\pi^{\theta,\theta_\star}$ is an invariant distribution for K^θ, which concludes the proof. □

2.3. **Construction of the kernel $\Phi^{\theta,\theta'}$ as a backward limit.** Again, all along this section, Assumption (A-1) is supposed to hold and the symbols $\mathbb{P}^\theta$ and $\tilde{\mathbb{P}}^\theta$ refer to the probabilities introduced in Definition 2. In addition to Assumption (A-1), Theorem 3 fundamentally relies on Assumption (K-1). These assumptions ensure the existence of the probability kernel $\Phi^{\theta,\theta'}$ that yields the definition of $p_1^{\theta,\theta'}(\cdot|\cdot)$. We now explain how the kernel $\Phi^{\theta,\theta'}$ may arise as a limit under $\mathbb{P}^{\theta'}$ of explicit kernels derived from K^θ. It will generally apply to observation-driven models, treated in Section 4, but also in the more classical case of hidden Markov models, as explained in Section 3. A natural approach is to define the kernel $\Phi^{\theta,\theta'}$ as the weak limit of the following ones.

Definition 7. Let n be a positive integer. For all $\theta \in \Theta$ and $x \in \mathsf{X}$, we define the probability kernel $\Phi^\theta_{x,n}$ on $\mathsf{Y}^{n+1} \times \mathcal{X}$ by: for all $y_{0:n} \in \mathsf{Y}^{n+1}$ and $A \in \mathcal{X}$,

$$\Phi^\theta_{x,n}(y_{0:n}; A) = \frac{\displaystyle\int_{\mathsf{X}^{n-1}\times A}\prod_{k=0}^{n-1}\kappa^\theta\langle y_k, y_{k+1}\rangle\,(x_k; \mathrm{d}x_{k+1})}{\displaystyle\int_{\mathsf{X}^n}\prod_{k=0}^{n-1}\kappa^\theta\langle y_k, y_{k+1}\rangle\,(x_k; \mathrm{d}x_{k+1})} \quad \text{with } x_0 = x.$$

We shall drop the subscript n when no ambiguity occurs.

It is worth noting that $\Phi^\theta_{x,n}(Y_{0:n};\cdot)$ is the conditional distribution of X_n given $Y_{1:n}$ under $\mathbb{P}^\theta_{(x,Y_0)}$. To build the desired $\Phi^{\theta,\theta'}$ we shall take, for a well-chosen x, the limit of $\Phi^\theta_{x,n}(y_{0:n};\cdot)$ as $n\to\infty$ for a sequence $y_{0:n}$ corresponding to a path under $\tilde{\mathbb{P}}^{\theta'}$. The precise statement is provided in Assumption (K'-1) below, which requires the following definition. For all $\theta\in\Theta$ and all nonnegative measurable functions f defined on X, we denote

$$\mathcal{F}^\theta_f := \left\{ x\mapsto \kappa^\theta\langle y,y'\rangle(x;f)\,,\ (y,y')\in\mathsf{Y}^2 \right\}\,.$$

We can now state the assumption as follows.

(K'-1) For all $\theta\neq\theta'\in\Theta$, there exist $x\in\mathsf{X}$, a probability kernel $\Phi^{\theta,\theta'}$ on $\mathsf{Y}^{\mathbb{Z}_-}\times\mathcal{X}$ and a countable class $\mathcal{F}$ of $\mathsf{X}\to\mathbb{R}_+$ measurable functions such that for all $f\in\mathcal{F}$,

$$\tilde{\mathbb{P}}^{\theta'}\left\{\forall f'\in\mathcal{F}^\theta_f\cup\{f\},\quad \lim_{m\to\infty}\Phi^\theta_{x,m}(Y_{-m:0};f') = \Phi^{\theta,\theta'}(Y_{-\infty:0};f')<\infty\right\}=1\,.$$

The next lemma shows that, provided that $\mathcal{F}$ is rich enough, Assumption (K'-1) can be used to obtain Assumption (K-1). In what follows we say that a class of $\mathsf{X}\to\mathbb{R}$ functions is separating if, for any two probability measures μ_1 and μ_2 on $(\mathsf{X},\mathcal{X})$, the equality of $\mu_1(f)$ and $\mu_2(f)$ over f in the class implies the equality of the two measures.

Lemma 8. *Suppose that Assumption (K'-1) holds and that $\mathcal{F}$ is a separating class of functions containing $\mathbb{1}_{\mathsf{X}}$. Then the kernel $\Phi^{\theta,\theta'}$ satisfies Assumption (K-1).*

Proof. Let $x\in\mathsf{X}$ be given in Assumption (K'-1). From Definition 7, we may write, for all $f\in\mathcal{F}$, setting $x_{-m}=x$,

$$\Phi^\theta_{x,m}(Y_{-m:0};f) = \frac{\displaystyle\int f(x_0)\prod_{k=-m}^{-1}\kappa^\theta\langle Y_k,Y_{k+1}\rangle(x_k;\mathrm{d}x_{k+1})}{\displaystyle\int\prod_{k=-m}^{-1}\kappa^\theta\langle Y_k,Y_{k+1}\rangle(x_k;\mathrm{d}x_{k+1})}\,,$$

and, similarly,

$$\Phi^\theta_{x,m+1}(Y_{-m:1};f) = \frac{\displaystyle\int f(x_1)\prod_{k=-m}^{0}\kappa^\theta\langle Y_k,Y_{k+1}\rangle(x_k;\mathrm{d}x_{k+1})}{\displaystyle\int\prod_{k=-m}^{0}\kappa^\theta\langle Y_k,Y_{k+1}\rangle(x_k;\mathrm{d}x_{k+1})}\,. \tag{22}$$

Dividing both numerator and denominator of (22) by

$$\int\prod_{k=-m}^{-1}\kappa^\theta\langle Y_k,Y_{k+1}\rangle(x_k;\mathrm{d}x_{k+1})\,,$$

which is strictly positive by Remark 1, then (22) can be rewritten as

$$\Phi^\theta_{x,m+1}(Y_{-m:1}; f) = \frac{\Phi^\theta_{x,m}\left(Y_{-m:0}; \kappa^\theta\langle Y_0, Y_1\rangle\,(\cdot; f)\right)}{\Phi^\theta_{x,m}\left(Y_{-m:0}; \kappa^\theta\langle Y_0, Y_1\rangle\,(\cdot; \mathbb{1}_{\mathsf{X}})\right)} \,. \tag{23}$$

Letting $m \to \infty$ and applying Assumption (K'-1), then $\tilde{\mathbb{P}}^{\theta'}$-a.s.,

$$\begin{aligned}\Phi^{\theta,\theta'}(Y_{-\infty:1}; f) &= \frac{\Phi^{\theta,\theta'}\left(Y_{-\infty:0}; \kappa^\theta\langle Y_0, Y_1\rangle\,(\cdot; f)\right)}{\Phi^{\theta,\theta'}\left(Y_{-\infty:0}; \kappa^\theta\langle Y_0, Y_1\rangle\,(\cdot; \mathbb{1}_{\mathsf{X}})\right)} \,, \\ &= \frac{\displaystyle\int \Phi^{\theta,\theta'}\left(Y_{-\infty:0}; \mathrm{d}x_0\right) \kappa^\theta\langle Y_0, Y_1\rangle\,(x_0; f)}{\displaystyle\int \Phi^{\theta,\theta'}\left(Y_{-\infty:0}; \mathrm{d}x_0\right) \kappa^\theta\langle Y_0, Y_1\rangle\,(x_0; \mathbb{1}_{\mathsf{X}})} \,.\end{aligned}$$

Since $\mathcal{F}$ is a separating class, the proof is concluded. □

3. Application to hidden Markov models

3.1. **Definitions and assumptions.** Hidden Markov models belong to a subclass of partially observed Markov models defined as follows.

Definition 9. Consider a partially observed and partially dominated Markov model given as in Definition 1 with Markov kernels $\{K^\theta, \theta \in \Theta\}$. We will say that it is a hidden Markov model if the kernel K^θ satisfies

$$K^\theta((x,y); \mathrm{d}x'\mathrm{d}y') = Q^\theta(x; \mathrm{d}x') G^\theta(x'; \mathrm{d}y') \,. \tag{24}$$

Moreover, in this context, we always assume that (X, d) is a complete separable metric space and $\mathcal{X}$ denotes the associated Borel σ-field.

In (24), Q^θ and G^θ are transition kernels on $\mathsf{X}\times\mathcal{X}$ and $\mathsf{X}\times\mathcal{Y}$ respectively. Since the model is partially dominated we denote by g^θ the corresponding Radon-Nikodym derivative of $G^\theta(x;\cdot)$ with respect to the dominating measure ν: for all $(x,y) \in \mathsf{X}\times\mathsf{Y}$,

$$\frac{\mathrm{d}G^\theta(x;\cdot)}{\mathrm{d}\nu}(y) = g^\theta(x;y) \,.$$

Then, the unnormalized kernel $\kappa^\theta\langle y, y'\rangle$ defined in (7) does not depend on y and we write in this case

$$\kappa^\theta\langle y, y'\rangle\left(x; \mathrm{d}x'\right) = \kappa^\theta\langle y'\rangle\left(x; \mathrm{d}x'\right) = Q^\theta(x; \mathrm{d}x') g^\theta(x'; y') \,. \tag{25}$$

For any integer $n \geq 1$, $\theta \in \Theta$ and any sequence $y_{0:n-1} \in \mathsf{Y}^n$, consider the unnormalized kernel $\mathbf{L}^\theta\langle y_{0:n-1}\rangle$ on $\mathsf{X}\times\mathcal{X}$ defined for all $x_0 \in \mathsf{X}$ and $A \in \mathcal{X}$, by

$$\mathbf{L}^\theta\langle y_{0:n-1}\rangle(x_0; A) = \int \cdots \int \left[\prod_{k=0}^{n-1} g^\theta(x_k; y_k) Q^\theta(x_k; \mathrm{d}x_{k+1})\right] \mathbb{1}_A(x_n) \,, \tag{26}$$

so that the MLE $\hat{\theta}_{\xi,n}$, associated to the observations $Y_{0:n-1}$ with an arbitrary initial distribution ξ on X is defined by

$$\hat{\theta}_{\xi,n} \in \operatorname*{argmax}_{\theta\in\Theta} \xi \mathbf{L}^{\theta} \langle Y_{0:n-1} \rangle \mathbb{1}_{\mathsf{X}} \, .$$

We now follow the approach taken by [14] in misspecified models and show that in the context of well-specified models, the maximizing set of the asymptotic normalized log-likelihood can be identified by relying neither on the exponential separation of measures, nor on the rates of convergence to the equilibrium, but only on the uniqueness of the invariant probability. We note the following fact which can be used to check (A-1).

Remark 2. In the HMM context, π^{θ} is an invariant distribution of K^{θ} if and only if π_1^{θ} is an invariant distribution of Q^{θ} and $\pi^{\theta}(\mathrm{d}x\mathrm{d}y) = \pi_1^{\theta}(\mathrm{d}x) G^{\theta}(x;\mathrm{d}y)$.

We illustrate the application of the main result (Theorem 3) in the context of Hidden Markov Models by considering the assumptions of [14] in the particular case of blocks of size 1 ($r = 1$). Of course, general assumptions with arbitrary sizes of blocks could also be used but this complicates significantly the expressions and may confine the attention of the reader to unnecessary technicalities. To keep the discussion simple, we only consider blocks of size 1, which already covers many cases of interest.

Before listing the main assumptions, we recall the definition of a local Doeblin set (in the particular case where $r = 1$) as introduced in [14, Definition 1].

Definition 10. A set C is local Doeblin with respect to the family of kernels $\{Q^{\theta}\,,\ \theta \in \Theta\}$ if there exist positive constants ϵ_C^-, ϵ_C^+ and a family of probability measures $\{\lambda_C^{\theta}\}_{\theta\in\Theta}$ such that for any $\theta \in \Theta$, $\lambda_C^{\theta}(C) = 1$ and, for any $A \in \mathcal{X}$, and $x \in C$,

$$\epsilon_C^- \lambda_C^{\theta}(A) \le Q^{\theta}(x; A \cap C) \le \epsilon_C^+ \lambda_C^{\theta}(A) \, .$$

Consider now the following set of assumptions.

(D-1) There exists a σ-finite measure μ on $(\mathsf{X}, \mathcal{X})$ that dominates $Q^{\theta}(x;\cdot)$ for all $(x,\theta) \in \mathsf{X}\times\Theta$. Moreover, denoting $q^{\theta}(x;x') = \frac{\mathrm{d}Q^{\theta}(x;\cdot)}{\mathrm{d}\mu}(x')$, we have

$$q^{\theta}(x;x') > 0\,, \quad \text{for all } (x,x',\theta) \in \mathsf{X}\times\mathsf{X}\times\Theta \, .$$

(D-2) For all $y \in \mathsf{Y}$, we have $\sup_{\theta\in\Theta}\sup_{x\in\mathsf{X}} g^{\theta}(x;y) < \infty$.

(D-3) (a) For all $\theta_\star \in \Theta$, there exists a set $K \in \mathcal{Y}$ with $\tilde{\mathbb{P}}^{\theta_\star}(Y_0 \in K) > 2/3$ such that for all $\eta > 0$, there exists a local Doeblin set $C \in \mathcal{X}$ with respect to $\{Q^{\theta}\,,\ \theta \in \Theta\}$ satisfying : for all $\theta \in \Theta$ and all $y \in K$,

$$\sup_{x\in C^c} g^{\theta}(x;y) \le \eta \sup_{x\in\mathsf{X}} g^{\theta}(x;y) < \infty \, . \tag{27}$$

(b) For all $\theta_\star \in \Theta$, there exists a set $D \in \mathcal{X}$ satisfying

$$\inf_{\theta\in\Theta}\inf_{x\in D} Q^{\theta}(x;D) > 0 \quad \text{and} \quad \tilde{\mathbb{E}}^{\theta_\star}\left[\ln^- \inf_{\theta\in\Theta}\inf_{x\in D} g^{\theta}(x;Y_0)\right] < \infty \, .$$

(D-4) For all $\theta_\star \in \Theta$, $\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^+ \sup_{\theta\in\Theta}\sup_{x\in\mathsf{X}} g^\theta(x;Y_0)\right] < \infty$.

(D-5) There exists $p \in \mathbb{N}$ such that for any $x \in \mathsf{X}$ and $n \geq p$, the function $\theta \mapsto \mathbf{L}^\theta\langle Y_{0:n}\rangle(x;\mathsf{X})$ is continuous on Θ $\tilde{\mathbb{P}}^{\theta_\star}$-a.s.

Remark 3. Under (D-1), for all $\theta \in \Theta$, the Markov kernel Q^θ is μ-irreducible, so that, using Remark 2, (A-1) reduces to the existence of a stationary distribution for Q^θ.

Remark 4. Assumptions (D-3), (D-4) and (D-5) and (6) in Definition 1 correspond to (A1), (A2) and (A3) in [14], where the blocks are of size $r = 1$.

Remark 5. (D-4) implies (D-2) up to a modification of $g^\theta(x;y)$ on ν-negligible set of $y \in \mathsf{Y}$ for all $x \in \mathsf{X}$. Indeed (D-4) implies that $\sup_\theta \sup_x g^\theta(x;Y_0) < \infty$, $\tilde{\mathbb{P}}^{\theta_\star}$-a.s., and it can be shown that under (D-1), $\pi_2^{\theta_\star} = \pi^{\theta_\star}(\mathsf{X}\times\cdot)$ is equivalent to ν for all $\theta \in \Theta$.

In these models, the kernel $\Phi_{x,n}^\theta$ introduced in Definition 7 writes

$$\Phi_{x,n}^\theta(y_{1:n};A) = \frac{\displaystyle\int_{\mathsf{X}^{n-1}\times A}\prod_{k=0}^{n-1} Q^\theta(x_k;\mathrm{d}x_{k+1})g^\theta(x_{k+1};y_{k+1})}{\displaystyle\int_{\mathsf{X}^n}\prod_{k=0}^{n-1} Q^\theta(x_k;\mathrm{d}x_{k+1})g^\theta(x_{k+1};y_{k+1})} \quad \text{with } x_0 = x\ .$$

The distribution $\Phi_{x,n}^\theta(Y_{0:n};\cdot)$ is usually referred to as the *filter distribution*. Proposition 11 (below) can be derived from [14, Proposition 1]. For blocks of size 1, the initial distributions in [14] are constrained to belong to the set $\mathcal{M}^{\theta_\star}(D)$ of all probability distributions ξ defined on $(\mathsf{X},\mathcal{X})$ such that

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^- \inf_{\theta\in\Theta}\int \xi(\mathrm{d}x)g^\theta(x;Y_0)Q^\theta(x;D)\right] < \infty\ , \tag{28}$$

where $D \in \mathcal{X}$ is the set appearing in (D-3). It turns out that under (D-3)-(b), all probability distributions ξ satisfy (28), so the constraint on the initial distribution vanishes in our case.

Proposition 11. *Assume (D-3) and (D-4). Then the following assertions hold.*

(i) For any $\theta,\theta_\star \in \Theta$, there exists a kernel $\Phi^{\theta,\theta_\star}$ on $\mathsf{Y}^{\mathbb{Z}_-}\times\mathcal{X}$ such that for any $x \in \mathsf{X}$,

$$\tilde{\mathbb{P}}^{\theta_\star}\left\{\text{for all bounded } f,\ \lim_{m\to\infty}\Phi_{x,m}^\theta(Y_{-m:0};f) = \Phi^{\theta,\theta_\star}(Y_{-\infty:0};f)\right\} = 1\ .$$

(ii) For any $\theta,\theta_\star \in \Theta$ and any probability measure ξ,

$$\lim_{n\to\infty} n^{-1}\ln \xi\mathbf{L}^\theta\langle Y_{0:n-1}\rangle\mathbb{1}_{\mathsf{X}} = \ell(\theta,\theta_\star), \quad \mathbb{P}^{\theta_\star}\text{-a.s.}$$

where

$$\ell(\theta,\theta_\star) := \mathbb{E}^{\theta_\star}\left[\ln \int \Phi^{\theta,\theta_\star}(Y_{-\infty:0};\mathrm{d}x_0)\kappa^{\theta}\langle Y_1\rangle\,(x_0;\mathsf{X})\right]\,. \tag{29}$$

3.2. **Equivalence-class consistency.** We can now state the main result on the consistency of the MLE for hidden Markov models.

Theorem 12. *Assume that (A-1) holds and define $\mathbb{P}^\theta$, $\tilde{\mathbb{P}}^\theta$ and the equivalence class $[\theta]$ as in Definition 2. Moreover, suppose that (Θ,Δ) is a compact metric space and that all the assumptions (D-1)–(D-5) hold. Then, for any probability measure ξ,*

$$\lim_{n\to\infty} \Delta(\hat{\theta}_{\xi,n},[\theta_\star]) = 0, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

Proof. According to [14, Theorem 2], $\theta \mapsto \ell(\theta,\theta_\star)$ defined by (29) is upper semi-continuous (so that $\Theta_\star = \mathrm{argmax}_{\theta\in\Theta}\,\ell(\theta,\theta_\star)$ is non-empty) and moreover

$$\lim_{n\to\infty} \Delta(\hat{\theta}_{\xi,n},\Theta_\star) = 0\,, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

The proof then follows from Theorem 3, provided that $\ell(\theta,\theta_\star)$ can be expressed as in the statement of Theorem 3 and that (K-1) is satisfied. First note that, for $\theta \neq \theta_\star$, the quantity in the ln of (29) corresponds to $p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})$ with $p^{\theta,\theta_\star}$ defined as in (12).

Let $\mathcal{F}$ be a countable separating class of nonnegative bounded functions containing $\mathbb{1}_{\mathsf{X}}$, see [27, Theorem 6.6, Chapter 6] for the existence of such a class. By Lemma 8, we check (K-1) by showing that (K'-1) is satisfied. Condition (D-2) and (25) imply that for all bounded function f, $\mathcal{F}_f^\theta$ is a class of bounded functions, and this in turn implies (K'-1) by applying Proposition 11-(i) to all x. Thus, (K-1) is satisfied and, for $\theta \neq \theta_\star$, $\ell(\theta,\theta_\star)$ can be expressed as in the statement of Theorem 3. To complete the proof, it only remains to consider the case $\theta = \theta_\star$ and show that $\ell(\theta_\star,\theta_\star)$ can be written as

$$\ell(\theta_\star,\theta_\star) = \mathbb{E}^{\theta_\star}\left[\ln p_1^{\theta_\star}(Y_1|Y_{-\infty:0})\right]\,, \tag{30}$$

where $p_1^{\theta_\star}(\cdot|\cdot)$ is the conditional density defined as in (13). According to [3, Theorem 1], we have

$$\mathbb{E}^{\theta_\star}\left[\ln p_1^{\theta_\star}(Y_1|Y_{-\infty:0})\right] = \lim_{n\to\infty} n^{-1}\ln \pi_1^{\theta_\star}\mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle\mathbb{1}_{\mathsf{X}}\,, \quad \mathbb{P}^{\theta_\star}\text{-a.s.} \tag{31}$$

On the other hand, applying Proposition 11-(ii) yields

$$\ell(\theta_\star,\theta_\star) = \lim_{n\to\infty} n^{-1}\ln \xi\mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle\mathbb{1}_{\mathsf{X}}, \quad \mathbb{P}^{\theta_\star}\text{-a.s.} \tag{32}$$

Observe that using (D-1), the probability measure $\xi\mathbf{L}^{\theta_\star}\langle y_0\rangle$ admits a density with respect to μ given by

$$\frac{\mathrm{d}\xi\mathbf{L}^{\theta_\star}\langle y_0\rangle}{\mathrm{d}\mu}(x_1) = \int \xi(\mathrm{d}x_0)g^{\theta_\star}(x_0;y_0)\,q^{\theta_\star}(x_0;x_1)\,. \tag{33}$$

We further get that, for all $y_{0:n-1} \in \mathsf{Y}^n$,

$$\xi \mathbf{L}^{\theta_\star}\langle y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}} = \int \frac{\mathrm{d}\xi \mathbf{L}^{\theta_\star}\langle y_0\rangle}{\mathrm{d}\mu}(x_1) \times \left(\delta_{x_1}\mathbf{L}^{\theta_\star}\langle y_{1:n-1}\rangle \mathbb{1}_{\mathsf{X}}\right)\, \mu(\mathrm{d}x_1)\ ,$$

and that, under $\mathbb{P}^{\theta_\star}$, the density of $X_1, Y_{0:n-1}$ with respect to $\mu \otimes \nu^{\otimes n}$ is given by

$$p^{\theta_\star}_{1,n}(x_1, y_{0:n-1}) := \frac{\mathrm{d}\pi_1^{\theta_\star}\mathbf{L}^{\theta_\star}\langle y_0\rangle}{\mathrm{d}\mu}(x_1) \times \left(\delta_{x_1}\mathbf{L}^{\theta_\star}\langle y_{1:n-1}\rangle \mathbb{1}_{\mathsf{X}}\right)\ .$$

Note that we similarly have, for all $y_0 \in \mathsf{Y}$ ans $x_1 \in \mathsf{X}$,

$$\frac{\mathrm{d}\pi_1^{\theta_\star}\mathbf{L}^{\theta_\star}\langle y_0\rangle}{\mathrm{d}\mu}(x_1) = \int \pi_1^{\theta_\star}(\mathrm{d}x_0) g^{\theta_\star}(x_0; y_0)\, q^{\theta_\star}(x_0; x_1)\ . \tag{34}$$

The four previous displays give that, for all $y_{0:n-1} \in \mathsf{Y}^n$,

$$\xi \mathbf{L}^{\theta_\star}\langle y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}} = \int \frac{\int \xi(\mathrm{d}x_0) g^{\theta_\star}(x_0; y_0) q^{\theta_\star}(x_0; x_1)}{\int \pi_1^{\theta_\star}(\mathrm{d}x_0) g^{\theta_\star}(x_0; y_0) q^{\theta_\star}(x_0; x_1)} p^{\theta_\star}_{1,n}(x_1, y_{0:n-1})\, \mu(\mathrm{d}x_1)\ .$$

Dividing by the density of $Y_{0:n-1}$ with respect to $\nu^{\otimes n}$ under $\mathbb{P}^{\theta_\star}$, we get

$$\frac{\xi \mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}}{\pi_1^{\theta_\star}\mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}} = \mathbb{E}^{\theta_\star}\left[R(X_1, Y_0) \,|\, Y_{0:n-1}\right] \quad \mathbb{P}^{\theta_\star}\text{-a.s.}\ ,$$

where $R(x_1, y_0)$ is the ratio between the densities (33) and (34). Since the denominator also is the density of X_1, Y_0 with respect to $\mu \otimes \nu^{\otimes n}$ under $\mathbb{P}^{\theta_\star}$, we have

$$\mathbb{E}^{\theta_\star}[R(X_1, Y_0)] = 1\ ,$$

By Lévy's zero-one law, we thus get that

$$\lim_{n\to\infty} \frac{\xi \mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}}{\pi_1^{\theta_\star}\mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}} = \mathbb{E}^{\theta_\star}\left[R(X_1, Y_0) \,|\, Y_{0:\infty}\right] \quad \mathbb{P}^{\theta_\star}\text{-a.s.}\ ,$$

and since by (D-1) $R(x_1, y_0)$ only takes positive values, this limit is positive. This implies that

$$\lim_{n\to\infty} n^{-1} \ln \frac{\xi \mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}}{\pi_1^{\theta_\star}\mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}} = 0 \quad \mathbb{P}^{\theta_\star}\text{-a.s.}$$

Combining with (31) and (32), we finally obtain (30), which concludes the proof. □

3.3. **A polynomially ergodic example.** As an application of Theorem 12, we consider the HMM model described in Example 1. In addition to the assumptions introduced in Example 1, we assume that U_0 and V_0 are independent and centered and they both admit densities with respect to the Lebesgue measure λ over $\mathbb{R}$, denoted by r and h, respectively, and

(E-1) the density r satisfies:

(a) r is continuous and positive over $\mathbb{R}$,

(b) there exists $\alpha > 2$ such that $r(u)|u|^{\alpha+1}$ is bounded away from ∞ as $|u| \to \infty$ and from 0 as $u \to \infty$,

(E-2) the density h satisfies:

(a) h is continuous and positive over $\mathbb{R}$, and $\lim_{|v|\to\infty} h(v) = 0$,

(b) there exist $\beta \in [1, \alpha - 1)$ (where α is defined in (E-1)) and $b, c > 0$ such that $\mathbb{E}(|V_0|^\beta) < \infty$ and $h(v) \geq b\,\mathrm{e}^{-c|v|^\beta}$ for all $v \in \mathbb{R}$.

For example, a symmetric Pareto distribution with a parameter strictly larger than 2 satisfies (E-1) and provided that $\alpha > 3$, (E-2) holds with a centered Gaussian distribution. The model is parameterized by $\theta = (m, a) \in \Theta := [\underline{m}, \overline{m}] \times [\underline{a}, \overline{a}]$ where $0 < \underline{m} < \overline{m}$ and $\underline{a} < \overline{a}$. In this model, the Markov transition Q^θ of $\{X_n\}_{n\in\mathbb{N}}$ has a transition density q^θ with respect to the dominating measure $\mu(\mathrm{d}x) = \lambda(\mathrm{d}x) + \delta_0(\mathrm{d}x)$, which can be written as follows: for all $(x, x') \in \mathbb{R}_+^2$,

$$q^\theta(x; x') = r(x' - x + m)\mathbb{1}\{x' > 0\} + \left(\int_{-\infty}^{m-x} r(u)\,\mathrm{d}u\right)\mathbb{1}\{x' = 0\}\,. \tag{35}$$

Moreover, (1) implies

$$g^\theta(x; y) = h(y - ax)\,. \tag{36}$$

Following [21], we have:

Lemma 13. *Assume (E-1) and (E-2). For all $\theta \in \Theta$, the Markov kernel Q^θ is not geometrically ergodic. Moreover, Q^θ is polynomially ergodic and its (unique) stationary distribution π_1^θ, which is defined on $\mathsf{X} = \mathbb{R}_+$, satisfies: $\int \pi_1^\theta(\mathrm{d}x)x^\beta < \infty$ for all $\beta \in [1, \alpha - 1)$.*

Proof. The proof of this Lemma is postponed to Section A.2 in Appendix A. □

Proposition 14. *Consider the HMM of Example 1 under Assumptions (E-1)–(E-2). Then (A-1) holds and we define $\mathbb{P}^\theta$, $\tilde{\mathbb{P}}^\theta$ and the equivalence class $[\theta]$ as in Definition 2. Moreover, for any probability measure ξ, the MLE $\hat{\theta}_{\xi,n}$ is equivalence-class consistent, that is, for any $\theta_\star \in \Theta$,*

$$\lim_{n\to\infty} \Delta(\hat{\theta}_{\xi,n}, [\theta_\star]) = 0, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

Proof. To apply Theorem 12, we need to check (A-1) and (D-1)–(D-5). First note that Assumption (A-1) directly follows from Remark 2 and Lemma 13, Assumption (D-1) from the positiveness of r and Assumption (D-2) from the boundedness of the density h. Now, using (E-1)-(a), it can be easily checked that all compact sets are local Doeblin sets and this in turn implies, via $\lim_{|x|\to\infty} h(x) = 0$ that Assumption (D-3)-(a) is satisfied. We now check (D-3)-(b). By (E-1)-(a), we have for all compact sets D, $\inf\left\{r(x' - x + m) : (x, x', m) \in D^2 \times [\underline{m}, \overline{m}]\right\} > 0$, which by (35), implies that

$$\inf_{\theta\in\Theta} \inf_{x\in D} Q^\theta(x; D) > 0\,.$$

To obtain (D-3)-(b), it thus remains to show

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^- \inf_{\theta\in\Theta}\inf_{x\in D} g^\theta(x;Y_0)\right] < \infty\ .$$

By (E-2)-(b), there exist positive constants b, c such that $h(v) \geq b\mathrm{e}^{-c|v|^\beta}$. Plugging this into (36) yields

$$\begin{aligned}\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^- \inf_{\theta\in\Theta}\inf_{x\in D} g^\theta(x;Y_0)\right] &\leq \tilde{\mathbb{E}}^{\theta_\star}\left[|\ln(b)| + c(|Y_0| + \overline{a}\sup_{x\in D}|x|)^\beta\right]\\ &= \mathbb{E}^{\theta_\star}\left[|\ln(b)| + c(|aX_0+V_0| + \overline{a}\sup_{x\in D}|x|)^\beta\right] < \infty\ ,\end{aligned}$$

where the finiteness follows from (E-2)-(b) and Lemma 13. Finally, (D-3) is satisfied. (D-4) is checked by writing

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^+ \sup_{\theta\in\Theta}\sup_{x\in\mathsf{X}} g^\theta(x;Y_0)\right] \leq \ln^+ \sup_{x\in\mathbb{R}} h(x) < \infty\ .$$

To obtain (D-5), we show by induction on n that for all $n \geq 1$, $y_{0:n-1} \in \mathbb{R}^n$ and $x_0 \in \mathbb{R}_+$, the function $\theta \mapsto \mathbf{L}^\theta\langle y_{0:n-1}\rangle(x_0;\mathsf{X})$ is continuous on Θ. The case $n = 1$ is obvious since $\mathbf{L}^\theta\langle y_0\rangle(x_0;\mathsf{X}) = g^\theta(x_0;y_0) = h(y_0 - ax_0)$. We next assume the induction hypothesis with n and note that

$$\mathbf{L}^\theta\langle y_{0:n}\rangle(x_0;\mathsf{X}) = g^\theta(x_0;y_0)\int \mu(\mathrm{d}x_1) q^\theta(x_0;x_1)\mathbf{L}^\theta\langle y_{1:n}\rangle(x_1;\mathsf{X})\ .$$

The continuity of $\theta \mapsto g^\theta(x_0;y_0)$ follows from (36) and the continuity of h. Similarly, the continuity of $\theta \mapsto q^\theta(x_0;x_1)$ follows from (35) and the continuity of r. Moreover, $\theta \mapsto \mathbf{L}^\theta\langle y_{1:n}\rangle(x_1;\mathsf{X})$ is continuous by the induction assumption. The continuity of $\theta \mapsto \int \mu(\mathrm{d}x_1) q^\theta(x_0;x_1)\mathbf{L}^\theta\langle y_{1:n}\rangle(x_1;\mathsf{X})$ then follows from the Lebesgue convergence theorem provided that we show

$$\int \mu(\mathrm{d}x_1)\sup_{\theta\in\Theta}\left\{q^\theta(x_0;x_1)\mathbf{L}^\theta\langle y_{1:n}\rangle(x_1;\mathsf{X})\right\} < \infty\ . \tag{37}$$

Note that by the expression of $q^\theta(x_0;x_1)$ given in (35) and the tail assumption (E-1)-(b), we obtain that, for all $x_0 \in \mathsf{X}$,

$$\int \mu(\mathrm{d}x_1)\sup_{\theta\in\Theta} q^\theta(x_0;x_1) < \infty\ .$$

Combining with $\mathbf{L}^\theta\langle y_{1:n}\rangle(x_1;\mathsf{X}) \leq (\sup_{x\in\mathbb{R}} h(x))^n$ yields (37). Finally, (D-5) holds and Theorem 12 may be applied under (E-1) and (E-2). □

4. Application to observation-driven models

Observation-driven models are a subclass of partially dominated and partially observed Markov models.

We split our study of the observation-driven model into several parts. Specific definitions and notation are introduced in Section 4.1. Then we

provide sufficient conditions that allow to apply our general result Theorem 3 , that is, $\Theta_\star = [\theta_\star]$. This is done in Section 4.2.

4.1. **Definitions and notation.** Observation-driven models are formally defined as follows.

Definition 15. Consider a partially observed and partially dominated Markov model given as in Definition 1 with Markov kernels $\{K^\theta, \theta \in \Theta\}$. We say that it is an observation-driven model if the kernel K^θ satisfies

$$K^\theta((x,y);\mathrm{d}x'\mathrm{d}y') = \delta_{\psi_y^\theta(x)}(\mathrm{d}x')\, G^\theta(x';\mathrm{d}y') \,, \tag{38}$$

where δ_a denotes the Dirac mass at point a, G^θ is a probability kernel on $\mathsf{X}\times\mathcal{Y}$ and $\{(x,y)\mapsto\psi_y^\theta(x)\,,\,\theta\in\Theta\}$ is a family of measurable functions from $(\mathsf{X}\times\mathsf{Y},\mathcal{X}\otimes\mathcal{Y})$ to $(\mathsf{X},\mathcal{X})$. Moreover, in this context, we always assume that (X,d) is a complete separable metric space and $\mathcal{X}$ denotes the associated Borel σ-field.

Note that a Markov chain $\{(X_k,Y_k)\,;\,k\in\mathbb{N}\}$ with probability kernel given by (38) can be equivalently defined by the following recursions

$$\begin{aligned} X_{k+1} &= \psi_{Y_k}^\theta(X_k) \,, \\ Y_{k+1}|X_{0:k+1},Y_{0:k} &\sim G^\theta(X_{k+1};\cdot) \,. \end{aligned} \tag{39}$$

The most celebrated example is the GARCH(1,1) process, where $G^\theta(x;\cdot)$ is a centered (say Gaussian) distribution with variance x and $\psi_y^\theta(x)$ is an affine function of x and y^2.

As a special case of Definition 1, for all $x\in\mathsf{X}$, $G^\theta(x;\cdot)$ is dominated by some σ-finite measure ν on $(\mathsf{Y},\mathcal{Y})$ and we denote by $g^\theta(x;\cdot)$ its Radon-Nikodym derivative, $g^\theta(x;y)=\frac{\mathrm{d}G^\theta(x;\cdot)}{\mathrm{d}\nu}(y)$. A dominated parametric observation-driven model is thus defined by the collection $\{(g^\theta,\psi^\theta)\,:\theta\in\Theta\}$. Moreover, (6) may be rewritten in this case: for all $(x,y)\in\mathsf{X}\times\mathsf{Y}$ and for all $\theta\in\Theta$,

$$g^\theta(x;y)>0\,.$$

Under (A-1), we assume that the model is well-specified, i.e., the observation sample $(Y_1,\dots,Y_n)$ is distributed according to $\tilde{\mathbb{P}}^{\theta_\star}$ for some unknown parameter $\theta_\star$. The inference of $\theta_\star$ is based on the conditional likelihood of $(Y_1,\dots,Y_n)$ given $X_1=x$ for an arbitrary $x\in\mathsf{X}$. The corresponding density function with respect to $\nu^{\otimes n}$ is, under parameter θ,

$$y_{1:n}\mapsto\prod_{k=1}^n g^\theta\left(\psi^\theta\langle y_{1:k-1}\rangle(x);y_k\right) \,, \tag{40}$$

where, for any vector $y_{1:p}=(y_1,\dots,y_p)\in\mathsf{Y}^p$, $\psi^\theta\langle y_{1:p}\rangle$ is the $\mathsf{X}\to\mathsf{X}$ function defined as the successive composition of $\psi_{y_1}^\theta$, $\psi_{y_2}^\theta$, ..., and $\psi_{y_p}^\theta$,

$$\psi^\theta\langle y_{1:p}\rangle = \psi_{y_p}^\theta\circ\psi_{y_{p-1}}^\theta\circ\cdots\circ\psi_{y_1}^\theta \,, \tag{41}$$

with the convention $\psi^\theta\langle y_{s:t}\rangle(x) = x$ for $s > t$. Then, the corresponding (conditional) Maximum Likelihood Estimator (MLE) $\hat{\theta}_{x,n}$ of the parameter θ, is defined by

$$\hat{\theta}_{x,n} \in \operatorname*{argmax}_{\theta\in\Theta} \mathsf{L}^\theta_{x,n}\langle Y_{1:n}\rangle \ , \tag{42}$$

where

$$\mathsf{L}^\theta_{x,n}\langle y_{1:n}\rangle := n^{-1} \ln \left(\prod_{k=1}^{n} g^\theta \left(\psi^\theta\langle y_{1:k-1}\rangle(x); y_k \right) \right) \ . \tag{43}$$

We will provide simple conditions for the consistency of $\hat{\theta}_{x,n}$ in the sense that, with probability tending to one, for a well chosen x, $\hat{\theta}_{x,n}$ belong to a neighborhood of the equivalence class $[\theta_\star]$ of $\theta_\star$, as given by Definition 2.

4.2. **Identifiability.** Let us consider the following assumptions.

(C-1) For all $\theta \neq \theta_\star \in \Theta$, there exist $x \in \mathsf{X}$ and a measurable function $\psi^{\theta,\theta_\star}\langle\cdot\rangle$ defined on $\mathsf{Y}^{\mathbb{Z}_-}$ such that

$$\lim_{m\to\infty} \psi^\theta\langle Y_{-m:0}\rangle(x) = \psi^{\theta,\theta_\star}\langle Y_{-\infty:0}\rangle, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{44}$$

(C-2) For all $\theta \in \Theta$ and $y \in \mathsf{Y}$, the function $x \mapsto g^\theta(x; y)$ is continuous on X.

(C-3) For all $\theta \in \Theta$ and $y \in \mathsf{Y}$, the function $x \mapsto \psi^\theta_y(x)$ is continuous on X.

In observation-driven models, the kernel κ^θ defined in (7) reads

$$\begin{aligned} \kappa^\theta\langle y, y'\rangle \left(x; \mathrm{d}x'\right) &= g^\theta(x'; y') \ \delta_{\psi^\theta_y(x)}(\mathrm{d}x') \\ &= g^\theta \left(\psi^\theta_y(x); y'\right) \ \delta_{\psi^\theta_y(x)}(\mathrm{d}x') \ , \end{aligned} \tag{45}$$

and the probability kernel $\Phi^\theta_{x,n}$ in Definition 7 reads: for all $x \in \mathsf{X}$ and $y_{0:n} \in \mathsf{Y}^{n+1}$,

$$\Phi^\theta_{x,n}(y_{0:n}; \cdot) = \delta_{\psi^\theta\langle y_{0:n-1}\rangle(x)} \ , \tag{46}$$

(the Dirac point mass at $\psi^\theta\langle y_{0:n-1}\rangle(x)$). Using these expressions, we get the following result which is a special case of Theorem 3.

Theorem 16. *Assume that (A-1) holds in the observation-driven model setting and define $\mathbb{P}^\theta$, $\tilde{\mathbb{P}}^\theta$ and $[\theta]$ as in Definition 2. Suppose that Assumptions (C-1), (C-2) and (C-3) hold and define $p^{\theta,\theta_\star}(\cdot|\cdot)$ by setting, for $\tilde{\mathbb{P}}^{\theta_\star}$-a.e. $y_{-\infty:0} \in \mathsf{Y}^{\mathbb{Z}_-}$,*

$$p^{\theta,\theta_\star}(y_1 \,|\, y_{-\infty:0}) = \begin{cases} g^\theta \left(\psi^{\theta,\theta_\star}\langle y_{-\infty:0}\rangle; y_1\right) & \text{if } \theta \neq \theta_\star, \\ p^\theta_1(y_1 \,|\, y_{-\infty:0}) \text{ as defined by (13)} & \text{otherwise.} \end{cases} \tag{47}$$

Then, for all $\theta_\star \in \Theta$, we have

$$\operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star} \left[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0}) \right] = [\theta_\star] \ . \tag{48}$$

Proof. We apply Theorem 3. It is thus sufficent to show that (C-1), (C-2) and (C-3) implies (K-1) with

$$\Phi^{\theta,\theta_\star}(y_{-\infty:0};\cdot) = \delta_{\psi^{\theta,\theta_\star}\langle y_{-\infty:-1}\rangle}, \quad \text{for all } y_{-\infty:0} \in \mathsf{Y}^{\mathbb{Z}_-} , \tag{49}$$

and that for $\theta \neq \theta_\star$, the conditional density $p^{\theta,\theta_\star}$ defined by (12) satisfies

$$p^{\theta,\theta_\star}(y|Y_{-\infty:0}) = g^\theta\left(\psi^{\theta,\theta_\star}\langle Y_{-\infty:0}\rangle; y\right) \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{50}$$

By Lemma 8, it is sufficient to prove that Assumption (K'-1) holds for the kernel $\Phi^{\theta,\theta_\star}$ defined above. Denote by $\mathcal{C}(\mathsf{X})$ the set of continuous functions on X, and by $\mathcal{C}_b(\mathsf{X})$ the set of bounded functions in $\mathcal{C}(\mathsf{X})$. By [27, Theorem 6.6, Chapter 6], there is a countable and separating subclass $\mathcal{F}$ of non-negative functions in $\mathcal{C}_b(\mathsf{X})$ such that $\mathbb{1}_\mathsf{X} \in \mathcal{F}$. Now, let us take $\theta, \theta_\star \in \Theta$ and $f \in \mathcal{F}$. Then, by (C-2), (C-3) and (45), we have

$$\mathcal{F}_f^\theta = \left\{ x \mapsto \kappa^\theta\langle y, y'\rangle\,(x; f)\,,\ (y, y') \in \mathsf{Y}^2 \right\} \subset \mathcal{C}(\mathsf{X})\,.$$

By (46), (C-1) and (49), we obtain (K'-1) with x being chosen as in (C-1).

To conclude, we need to show (50). Note that (49) together with (45) and the usual definition (12) of $p^{\theta,\theta_\star}$ yields

$$p^{\theta,\theta_\star}(y|y_{-\infty:0}) = g^\theta\left(\psi_{y_0}^\theta\left(\psi^{\theta,\theta_\star}\langle y_{-\infty:-1}\rangle\right); y\right)\,.$$

By Assumption (C-3) and the definition of $\psi^{\theta,\theta_\star}\langle\cdot\rangle$ in (C-1), we get (50). □

4.3. **Examples.** In the context of observation-driven time series, easy-to-check conditions are derived in [13] in order to establish the convergence of the MLE $\hat{\theta}_{x,n}$ defined by (42) to the maximizing set of the asymptotic normalized log-likelihood. It turns out that the conditions of [13, Theorem 3] also imply the conditions of Theorem 16. More precisely the assumptions (B-2) and (B-3) of [13, Theorem 3] are stronger than (C-2) and (C-3) used in Theorem 16 above, and it is shown that the assumptions of [13, Theorem 3] imply (C-1) (see the proof of Lemma 10 in Appendix A of [13]). Moreover the conditions of Theorem 3 are shown to be satisfied in the context of Examples 2 and 3 (see [13, Theorem 6 and Theorem 7]), provided that Θ in (42) is a compact metric space such that

(1) in the case of Example 2, all $\theta = (\omega, a, b, r) \in \Theta$ satisfy $rb + a < 1$;
(2) in the case of Example 3, all $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b}) \in \Theta$ are such that the spectral radius of $\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T$ is strictly less than 1.

Under these assumptions, we conclude that the MLE is equivalence-class consistent for both examples, which up to our best knowledge, had not been proven so far.

Appendix A. Postponed proofs

A.1. **Proof of Lemma 5.** First observe that, by induction on n, having (16) for all $n \geq 2$ is equivalent to having, for all $n \geq 2$,

$$p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0}) = p^{\theta,\theta_\star}(Y_n|Y_{-\infty:n-1})p^{\theta,\theta_\star}(Y_{n-1}|Y_{-\infty:n-2})\cdots p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0}),\ \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.},$$

which, using that $\tilde{\mathbb{P}}^{\theta_\star}$ is shift-invariant, is in turn equivalent to having that, for all $n \geq 2$,

$$p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0}) = p^{\theta,\theta_\star}(Y_{2:n}|Y_{-\infty:1})p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0}), \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{51}$$

Thus to conclude the proof, we only need to show that (51) holds for all $n \geq 2$. By Definition 4, we have, for all $n \geq 2$, and all $y_{-\infty:n} \in \mathsf{Y}^{\mathbb{Z}_-}$,

$$p^{\theta,\theta_\star}(y_{2:n}|y_{-\infty:1})p^{\theta,\theta_\star}(y_1|y_{-\infty:0}) = \int \Phi^{\theta,\theta_\star}(y_{-\infty:1};\mathrm{d}x_1)p^{\theta,\theta_\star}(y_1|y_{-\infty:0}) \prod_{k=1}^{n-1} \kappa^\theta\langle y_k, y_{k+1}\rangle\,(x_k;\mathrm{d}x_{k+1})\ .$$

Now, using (K-1), we get that, for all $n \geq 2$,

$$p^{\theta,\theta_\star}(Y_{2:n}|Y_{-\infty:1})p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0}) = \int \Phi^{\theta,\theta_\star}(Y_{-\infty:0};\mathrm{d}x_0) \prod_{k=0}^{n-1} \kappa^\theta\langle Y_k, Y_{k+1}\rangle\,(x_k;\mathrm{d}x_{k+1}) \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

We conclude (51) by observing that by Definition 4, the second line of the last display is $p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0})$.

A.2. **Proof of Lemma 13.** Let $\beta \in [1, \alpha - 1)$. By (E-1)-(b) and $1 + \beta < \alpha$, we obtain $\mathbb{E}\left[(U_0^+)^{1+\beta}\right] < \infty$. Combining with $\mathbb{E}(U_0 - m) = -m < 0$, we may apply [21, Proposition 5.1] so that the Markov kernel Q^θ is polynomially ergodic and thus admits a unique stationary distribution π_1^θ, which is defined on $\mathsf{X} = \mathbb{R}_+$. Moreover, [21, Proposition 5.1] also shows that there exists a finite interval $C = [0, x_0]$ and constants $0 < \varrho, \varrho' < \infty$ such that

$$Q^\theta V \leq V - \varrho W + \varrho' \mathbb{1}_C\ ,$$

where $V(x) = (1 + x)^{1+\beta}$ and $W(x) = (1 + x)^\beta$. Applying [25, Theorem 14.0.1] yields $\int \pi_1^\theta(\mathrm{d}x)x^\beta \leq \pi_1^\theta W < \infty$. It remains to show that Q^θ is not geometrically ergodic for all $\theta \in \Theta$.

This will be done by contradiction. Assume first that Q^θ is geometrically ergodic for some $\theta \in \Theta$ and note that the singleton $\{0\}$ is an accessible atom. Then, there exists $\rho > 1$ such that $\sum_{k=0}^\infty \rho^k |(Q^\theta)^k(0, \{0\}) - \pi_1^\theta(\{0\})| < \infty$ so that the atom $\{0\}$ is geometrically ergodic as defined in [25, Section 15.1.3]. Applying [25, Theorem 15.1.5], there exists $\kappa > 1$ such that $\mathbb{E}_0[\kappa^{\tau_0}] < \infty$ where $\tau_0 = \inf\{n \geq 1 : X_n = 0\}$ is the first return time to $\{0\}$.

Recall that $\{U_n\}_{n\in\mathbb{N}}$ denotes the i.i.d sequence, linked to $\{X_n\}_{n\in\mathbb{N}}$ by (1) and note that $\mathbb{E}_0[\kappa^{\tau_0}] = \mathbb{E}[\kappa^{\tau(0)}]$ where we have set for all $u \in \mathbb{R}$,

$$\tau(u) := \inf\left\{ n \geq 1 \,:\, \sum_{k=1}^{n}(U_k - m) < u \right\} .$$

Denote

$$\tilde{\tau}(u) := \inf\left\{ n \geq 1 \,:\, \sum_{k=1}^{n}(U_{k+1} - m) < u \right\} .$$

To arrive at the contradiction, it is finally sufficient to show that for all $\kappa > 1$, $\mathbb{E}[\kappa^{\tau(0)}] = \infty$. Actually, we will show that there exists a constant $\gamma > 0$ such that

$$\liminf_{u\to\infty} \kappa^{-\gamma u}\mathbb{E}[\kappa^{\tau(-u+m)}] > 0 . \tag{52}$$

This will indeed imply $\mathbb{E}[\kappa^{\tau(0)}] = \infty$ by writing

$$\begin{aligned}\mathbb{E}[\kappa^{\tau(0)}] &\geq \mathbb{E}[\kappa^{\tau(0)}\mathbb{1}\{U_1 \geq m\}] = \mathbb{E}[\kappa^{1+\tilde{\tau}(-U_1+m)}\mathbb{1}\{U_1 \geq m\}] \\ &= \mathbb{E}\left[\int_m^\infty \kappa^{1+\tilde{\tau}(-u+m)} r(u)\mathrm{d}u\right] = \kappa\int_m^\infty \mathbb{E}[\kappa^{\tau(-u+m)}]r(u)\mathrm{d}u , \end{aligned} \tag{53}$$

where the last equality follows from $\tau \stackrel{d}{=} \tilde{\tau}$. Provided that (52) holds, the r.h.s. of (53) is infinite since $r(u) \gtrsim u^{-\alpha-1}$ as $u \to \infty$ by (E-1)-(b).

We now turn to the proof of (52). By the Markov inequality,

$$\kappa^{-\gamma u}\mathbb{E}[\kappa^{\tau(-u+m)}] \geq \mathbb{P}(\tau(-u+m) > \gamma u) . \tag{54}$$

Now, let $M_n = \sum_{k=1}^n U_i$ for $n \geq 1$ and note that for all nonnegative u,

$$\begin{aligned}\left\{\left(\inf_{1\leq k\leq \gamma u} M_k\right) - \gamma u m \geq -u + m\right\} &\subset \left\{\inf_{1\leq k\leq\gamma u}(M_k - km) \geq -u + m\right\} \\ &= \{\tau(-u+m) > \gamma u\} . \end{aligned} \tag{55}$$

Moreover, since $\{U_n\}_{n\in\mathbb{N}^*}$ is i.i.d. and centered, the Doob maximal inequality yields: for all $\beta > 0$,

$$\begin{aligned}\mathbb{P}\left(\inf_{1\leq k\leq\gamma u} M_k < -\beta\right) &\leq \mathbb{P}\left(\sup_{1\leq k\leq\gamma u}|M_k| > \beta\right) \\ &\leq \frac{\mathbb{E}\left[|M_{\lfloor\gamma u\rfloor}|\right]}{\beta} \leq \frac{\lfloor\gamma u\rfloor\mathbb{E}\left[|U_1|\right]}{\beta} . \end{aligned} \tag{56}$$

Now, pick $\gamma > 0$ sufficiently small such that $\gamma\mathbb{E}\left[|U_1|\right]/(1-\gamma m) < 1$. This γ being set, note that $\beta = (1-\gamma m)u - m$ is positive for u sufficiently large so that combining (56) with (55) and (54) gives

$$\liminf_{u\to\infty} \kappa^{-\gamma u}\mathbb{E}[\kappa^{\tau(-u+m)}] \geq 1 - \limsup_{u\to\infty} \frac{\lfloor\gamma u\rfloor\mathbb{E}\left[|U_1|\right]}{(1-\gamma m)u - m} = 1 - \frac{\gamma\mathbb{E}\left[|U_1|\right]}{1-\gamma m} > 0 .$$

This shows (52) and the proof is completed.

References

[1] Carol Alexander and Emese Lazar, *Normal mixture garch (1, 1): Applications to exchange rate modelling*, Journal of Applied Econometrics **21** (2006), no. 3, 307–336.

[2] Elizabeth S Allman, Catherine Matias, and John A Rhodes, *Identifiability of parameters in latent structure models with many observed variables*, The Annals of Statistics (2009), 3099–3132.

[3] A. Barron, *The strong ergodic theorem for densities; generalized Shannon-McMillan-Breiman theorem*, Ann. Probab. **13** (1985), 1292–1303.

[4] P. J. Bickel, Y. Ritov, and T. Rydén, *Asymptotic normality of the maximum likelihood estimator for general hidden Markov models*, Ann. Statist. **26** (1998), 1614–1635.

[5] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, J. Econometrics **31** (1986), 307–327.

[6] Tim Bollerslev, *Glossary to arch (garch)*, CREATES Research Paper, 2008.

[7] DR Cox, *Statistical analysis of time-series: some recent developments*, Scand. J. Statist. **8** (1981), no. 2, 93–115.

[8] RA Davis, WTM Dunsmuir, and SB Streett, *Observation-driven models for Poisson counts*, Biometrika **90** (2003DEC), no. 4, 777–790.

[9] R.A. Davis and H. Liu, *Theory and inference for a class of observation-driven models with application to time series of counts*, Preprint, arXiv:1204.3915 (2012).

[10] R. Douc, P. Doukhan, and E. Moulines, *Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator*, Stochastic Processes and Their Applications **123** (2013), no. 7, 2620–2647.

[11] R. Douc, E. Moulines, J. Olsson, and R. van Handel, *Consistency of the maximum likelihood estimator for general hidden Markov models*, Ann. Statist. **39** (2011), no. 1, 474–513. MR2797854

[12] R. Douc, E. Moulines, and T. Rydén, *Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime*, Ann. Statist. **32** (2004), no. 5, 2254–2304.

[13] R. Douc, F. Roueff, and T. Sim, *Practical conditions for the convergence of the maximum likelihood estimator in observation-driven models*, 2014. Submitted.

[14] Randal Douc and Eric Moulines, *Asymptotic properties of the maximum likelihood estimation in misspecified hidden markov models*, Ann. Statist. **40** (2012), no. 5, 2697–2732.

[15] Paul Doukhan, Konstantinos Fokianos, and Dag Tjøstheim, *On weak dependence conditions for Poisson autoregressions*, Statist. Probab. Lett. **82** (2012), no. 5, 942–948.

[16] Yariv Ephraim and Brian L Mark, *Bivariate markov processes and their estimation*, Foundations and Trends in Signal Processing **6** (2012), no. 1, 1–95.

[17] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim, *Poisson autoregression*, J. Am. Statist. Assoc. **104** (2009), no. 488, 1430–1439. With electronic supplementary materials available online. MR2596998 (2011d:62256)

[18] Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin, *Finite state space non parametric hidden markov models are in general identifiable*, Stats and Computing. To appear. (2013).

[19] Markus Haas, Stefan Mittnik, and Marc S Paolella, *Mixed normal conditional heteroskedasticity*, Journal of Financial Econometrics **2** (2004), no. 2, 211–250.

[20] S. G. Henderson, D.S. Matteson, and D.B. Woodard, *Stationarity of generalized autoregressive moving average models*, Electronic Journal of Statistics **5** (2011), 800–828.

[21] Søren F. Jarner and Gareth O. Roberts, *Polynomial convergence rates of Markov chains*, Ann. Appl. Probab. **12** (2002), no. 1, 224–247.

[22] J. L. Jensen and N. V. Petersen, *Asymptotic normality of the maximum likelihood estimator in state space models*, Ann. Statist. **27** (1999), 514–535.
[23] Olav Kallenberg, *Foundations of modern probability*, Second, Probability and its Applications (New York), Springer-Verlag, New York, 2002. MR1876169 (2002m:60002)
[24] B. G. Leroux, *Maximum-likelihood estimation for hidden Markov models*, Stoch. Proc. Appl. **40** (1992), 127–143.
[25] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer, London, 1993.
[26] Michael H. Neumann, *Absolute regularity and ergodicity of Poisson count processes*, Bernoulli **17** (2011NOV), no. 4, 1268–1284.
[27] K. R. Parthasarathy, *Probability measures on metric spaces*, AMS Chelsea Publishing, Providence, RI, 2005. Reprint of the 1967 original. MR2169627 (2006d:60004)
[28] Wojciech Pieczynski, *Pairwise markov chains*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **25** (2003), no. 5, 634–639.
[29] S. Streett, *Some observation driven models for time series of counts*, Ph.D. Thesis, 2000.
[30] P. Tuominen and R. Tweedie, *Subgeometric rates of convergence of f-ergodic Markov Chains.*, Advances in Applied Probability **26** (1994), 775–798.
[31] Fukang Zhu, *A negative binomial integer-valued GARCH model*, J. Time Series Anal. **32** (2011), no. 1, 54–67. MR2790672 (2012f:62187)

[1]Department CITI, CNRS UMR 5157, Telecom Sudparis, Evry. France.
E-mail address, R. Douc: randal.douc@telecom-sudparis.eu

[2]Institut Mines-Telecom, Telecom Paristech, CNRS LTCI, Paris. France.
E-mail address, F. Roueff: roueff@telecom-paristech.fr

E-mail address, T. Sim: sim@telecom-paristech.fr