The Annals of Applied Probability
2016, Vol. 26, No. 4, 2357–2383
DOI: 10.1214/15-AAP1149

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

BY RANDAL DOUC*, FRANÇOIS ROUEFF† AND TEPMONY SIM†,[1]

Télécom SudParis and Télécom ParisTech†*

This paper deals with a parametrized family of partially observed bivariate Markov chains. We establish that, under very mild assumptions, the limit of the normalized log-likelihood function is maximized when the parameters belong to the equivalence class of the true parameter, which is a key feature for obtaining the consistency of the maximum likelihood estimators (MLEs) 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 (HMMs) and observation-driven time series models. In contrast with previous approaches, the identifiability is addressed by relying on the uniqueness 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, and (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 (HMMs). Nevertheless, Problem 2.1 can be solved independently, and with Step 1 above, this directly yields that the MLE is consistent in a weakened sense, namely, that the estimated parameter converges to the set of all the parameters associated to the same distribution as the one of the observed sample. This consistency result is referred to as *equivalence-class consistency*, as introduced by [23]. In this

Received November 2014; revised September 2015.

[1]Supported by Institut Carnot.

MSC2010 subject classifications. Primary 60J05, 62F12; secondary 62M05, 62M10.

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

contribution, our goal is to provide a general approach to solve Problem 2.1 in the general framework of partially observed Markov models. These include many classes of models of interest; see, for instance, [27] or [16]. The novel aspect of this work is that the result mainly relies on the uniqueness 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 [12, 13, 23]) or as a basic assumption (see [4, 21]). 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. In this example and throughout the paper, we denote by $\mathbb{R}_+ = [0, \infty)$, $\mathbb{R}_- = (-\infty, 0]$, $\mathbb{R}_+^* = (0, \infty)$ and $\mathbb{R}_-^* = (-\infty, 0)$, the sets of nonnegative, nonpositive, (strictly) positive and (strictly) negative real numbers, respectively. Similarly, we use the notation $\mathbb{Z}_+$, $\mathbb{Z}_-$, $\mathbb{Z}_+^*$ and $\mathbb{Z}_-^*$ for the corresponding subsets of integers. Also, $a^+ = \max(a, 0)$ denotes the nonnegative part of a.

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 an HMM on $\mathsf{X} \times \mathsf{Y}$ by the following recursions:

$$
\begin{aligned}
X_k &= (X_{k-1} + U_k - m)^+, \\
Y_k &= a X_k + V_k,
\end{aligned}
\tag{1.1}
$$

where $(m, a) \in \mathbb{R}_+^* \times \mathbb{R}$, and the sequence $((U_k, V_k))_{k \in \mathbb{Z}_+}$ is independent and identically distributed (i.i.d.) and is independent from X_0. This Markov model $(X_k)_{k \in \mathbb{Z}_+}$ was proposed by [29] and further considered by [20] as an example of polynomially ergodic Markov chain, under specific assumptions made on U_k's. Namely, if 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 $(X_k)_{k \in \mathbb{Z}_+}$ is not geometrically ergodic (see Lemma 4 below). In such a situation, the exponential separation of measures condition introduced in [12] seems difficult to check. We will show, nevertheless, in Proposition 2, that under some mild conditions the chain $(X_k)_{k \in \mathbb{Z}_+}$ is ergodic and the equivalence-class consistency holds.

Observation-driven time series models were introduced by [7] and later considered, among others, by [8, 15, 17, 25, 28] 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 function and its derivatives are easy to compute. The consistency of the MLE can however 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 [30] 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 example 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 (Example 2) was introduced in [31]. 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(1, 1)] model is defined by

$$\begin{aligned} &X_{k+1} = \omega + aX_k + bY_k, \\ &Y_{k+1} | X_{0:k+1}, Y_{0:k} \sim \mathcal{NB}\left(r, \frac{X_{k+1}}{1 + X_{k+1}}\right), \end{aligned} \tag{1.2}$$

where X_k takes values in $\mathsf{X} = \mathbb{R}_+$, Y_k takes values in $\mathbb{Z}_+$ and $\theta = (\omega, a, b, r) \in (\mathbb{R}_+^*)^4$ is an unknown parameter. In (1.2), $\mathcal{NB}(r, p)$ denotes the negative binomial distribution with parameters $r > 0$ and $p \in (0, 1)$, whose probability function is $\frac{\Gamma(k+r)}{k!\Gamma(r)} p^r (1-p)^k$ for all $k \in \mathbb{Z}_+$, where Γ stands for the Gamma function.

The second example, Example 3, proposed by [19] and [1], is a 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. The normal mixture GARCH [NM(d)-GARCH(1, 1)] model is defined by:

$$\begin{aligned} &\mathbf{X}_{k+1} = \boldsymbol{\omega} + \mathbf{A}\mathbf{X}_k + Y_k^2 \mathbf{b}, \\ &Y_{k+1} | \mathbf{X}_{0:k+1}, Y_{0:k} \sim G^\theta(\mathbf{X}_{k+1}; \cdot), \\ &G^\theta(\mathbf{x}; \mathrm{d}y) = \left(\sum_{\ell=1}^d \gamma_\ell \frac{\mathrm{e}^{-y^2/2x_\ell}}{(2\pi x_\ell)^{1/2}} \right) \mathrm{d}y, \\ &\qquad \mathbf{x} = (x_i)_{1 \le i \le d} \in (\mathbb{R}_+^*)^d, y \in \mathbb{R}, \end{aligned} \tag{1.3}$$

where d is a positive integer; $\mathbf{X}_k = [X_{1,k} \cdots X_{d,k}]^T$ takes values in $\mathsf{X} = \mathbb{R}_+^d$; $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_d]^T$ is 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 nonnegative entries, respectively; $\mathbf{A}$ is a $d \times d$ matrix parameter with nonnegative entries; and $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b})$.

The paper is organized as follows. Section 2 is dedicated to the main result (Theorem 1) which shows that the argmax of the limiting criterion reduces to the equivalence class of the true parameter, as defined in [23]. The general setting is introduced in Section 2.1. The theorem is stated and proved in Section 2.2. In Section 2.3, we focus on the kernel involved in the assumptions, and explain how it can be obtained explicitly. Our general assumptions are then shown to hold for two important classes of partially observed Markov models:

– First, the HMMs 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 Examples 2 and 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 $((X_k, Y_k))_{k\in\mathbb{Z}_+}$ 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 will simply write $\mathbb{P}^\theta_{(x,y)}$.

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

$$K^\theta\big((x, y); \mathrm{d}x' \,\mathrm{d}y'\big) = Q^\theta\big((x, y); \mathrm{d}x'\big)G^\theta\big((x, y, x'); \mathrm{d}y'\big), \tag{2.1}$$

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 now consider the following general setting.

DEFINITION 1. We say that the Markov model $(K^\theta)_{\theta\in\Theta}$ of the form (2.1) is partially dominated if there exists a σ-finite measure ν on Y such that for all $(x, y), (x', y') \in \mathsf{X} \times \mathsf{Y}$,

$$G^\theta\big((x, y, x'); \mathrm{d}y'\big) = g^\theta\big((x, y, x'); y'\big)\nu\big(\mathrm{d}y'\big), \tag{2.2}$$

where the conditional density function g^θ moreover satisfies

$$g^\theta\big((x, y, x'); y'\big) > 0, \qquad \text{for all } (x, y), \big(x', y'\big) \in \mathsf{X} \times \mathsf{Y}. \tag{2.3}$$

It follows from (2.2) that, for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $A \in \mathcal{X}$ and $B \in \mathcal{Y}$,

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

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

$$\kappa^\theta\langle y, y'\rangle(x; \mathrm{d}x') := Q^\theta\big((x, y); \mathrm{d}x'\big)g^\theta\big((x, y, x'); y'\big). \tag{2.4}$$

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 $(x, y, y') \in \mathsf{X} \times \mathsf{Y} \times \mathsf{Y}$,

$$\kappa^\theta\langle y, y'\rangle(x; \mathsf{X}) = \int_{\mathsf{X}} Q^\theta\big((x, y); \mathrm{d}x'\big)g^\theta\big((x, y, x'); y'\big) > 0, \tag{2.5}$$

where the positiveness follows from the fact that $Q^\theta((x, y); \cdot)$ is a probability on $(\mathsf{X}, \mathcal{X})$ and Condition (2.3).

In well-specified models, it is assumed that the observations $Y_{1:n}$ are generated from a process $((X_k, Y_k))_{k\in\mathbb{Z}_+}$, which follows the distribution $\mathbb{P}^{\theta_\star}_{\xi_\star}$ associated to an unknown parameter $\theta_\star \in \Theta$ and an unknown initial distribution $\xi_\star$ (usually, $\xi_\star$ is such that, under $\mathbb{P}^{\theta_\star}_{\xi_\star}$, $(Y_k)_{k\in\mathbb{Z}_+}$ is a stationary sequence). To form a consistent estimate of $\theta_\star$ on the basis of the observations $Y_{1:n}$ only, that is, without access to the hidden process $(X_k)_{k\in\mathbb{Z}_+}$, 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 $\mathsf{X} \times \mathsf{Y}$, that is,

$$\begin{aligned} L_{\xi,n}(\theta) &:= \ln \int \prod_{k=1}^n Q^\theta\big((x_{k-1}, y_{k-1}); \mathrm{d}x_k\big)g^\theta\big((x_{k-1}, y_{k-1}, x_k); y_k\big)\xi(\mathrm{d}x_0\,\mathrm{d}y_0) \\ &= \ln \int \kappa^\theta\langle y_0, y_1\rangle\kappa^\theta\langle y_1, y_2\rangle \cdots \kappa^\theta\langle y_{n-1}, y_n\rangle(x_0; \mathsf{X})\xi(\mathrm{d}x_0\,\mathrm{d}y_0). \end{aligned}$$

This corresponds to the log of the conditional density of $Y_{1:n}$ given (X_0, Y_0) with the latter integrated according to ξ. In practice, ξ is often taken as a Dirac mass at (x, y) with x arbitrarily chosen and y equal to the observation Y_0 when it is available. In this context, a classical way (see, e.g., [23]) 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

$$\Theta_\star := \operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\big[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\big]. \tag{2.6}$$

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 $n^{-1}L_{\xi,n}(\theta)$ 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 (K-1) and (K-2), 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 (2.6) 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 a set of general assumptions which ensures that $\Theta_\star$ is exactly the set of parameters θ such that $\tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta_\star}$. Then this result guarantees that the estimator converges to the set of parameters compatible with the true stationary distribution of the observations. If moreover the model $(\tilde{\mathbb{P}}^\theta)_{\theta\in\Theta}$ is identifiable, then this set reduces to $\{\theta_\star\}$ and consistency of $\hat{\theta}_{\xi,n}$ directly follows.

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

(K-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 (K-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 $\mathsf{Y}^{\mathbb{Z}}$.

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. Moreover, for all $\theta,\theta' \in \Theta$, we write $\theta \sim \theta'$ if and only if $\tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta'}$. 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 [23] as an alternative to the classical identifiability condition.

2.2. *Main result.* Assumption (K-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-2) 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 $A \in \mathcal{X}$,

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

REMARK 2. 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.

REMARK 3. Let us give some insight about the formula appearing in (K-2) and explain why it is important to consider the cases $\theta = \theta'$ and $\theta \neq \theta'$ separately. Since X is a Borel space, [22], Theorem 6.3, applies and the conditional distribution of X_0 given $Y_{-\infty:0}$ under $\mathbb{P}^\theta$ defines a probability kernel denoted by Φ^θ. We prove in Section A.1 of the Appendix that this kernel satisfies, for all $A \in \mathcal{X}$,

$$\frac{\int_{\mathsf{X}} \Phi^{\theta}(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, Y_1\rangle(x_0; A)}{\int_{\mathsf{X}} \Phi^{\theta}(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, Y_1\rangle(x_0; \mathsf{X})} = \Phi^{\theta}(Y_{-\infty:1}; A), \qquad \tilde{\mathbb{P}}^{\theta}\text{-a.s.} \tag{2.7}$$

Assumption (K-2) asserts that the kernel $\Phi^{\theta,\theta'}$ satisfies a similar identity $\tilde{\mathbb{P}}^{\theta'}$-a.s. for $\theta' \neq \theta$. It is not necessary at this stage to precise how $\Phi^{\theta,\theta'}$ shows up. This is done in Section 2.3.

REMARK 4. The denominator in the ratio displayed in (K-2) can be written as $p^{\theta,\theta'}(Y_1|Y_{-\infty:0})$, where, for all $y \in \mathsf{Y}$ and $y_{-\infty:0} \in \mathsf{Y}^{\mathbb{Z}_-}$,

$$p^{\theta,\theta'}(y|y_{-\infty:0}) := \int_{\mathsf{X}} \Phi^{\theta,\theta'}(y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle y_0, y\rangle(x_0; \mathsf{X}) \tag{2.8}$$

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, [22], 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 $B \in \mathcal{Y}^{\otimes n}$,

$$\tilde{\mathbb{P}}^\theta(Y_{1:n} \in B|Y_{-\infty:0}) = \int_B p_n^\theta(y_{1:n}|Y_{-\infty:0})\nu(\mathrm{d}y_1)\cdots\nu(\mathrm{d}y_n), \qquad \tilde{\mathbb{P}}^\theta\text{-a.s.} \tag{2.9}$$

Let us now state the main result.

THEOREM 1. *Assume that* (K-1) *holds and define* $\mathbb{P}^\theta$, $\tilde{\mathbb{P}}^\theta$ *and* $[\theta]$ *as in Definition* 2. *Suppose that Assumption* (K-2) *holds. For all* $\theta, \theta' \in \Theta$, *define* $p^{\theta,\theta'}(Y_1|Y_{-\infty:0})$ *by* (2.8) *if* $\theta \neq \theta'$ *and by* $p^{\theta,\theta}(Y_1|Y_{-\infty:0}) = p_1^\theta(Y_1|Y_{-\infty:0})$ *as in* (2.9) *otherwise. Then for all* $\theta_\star \in \Theta$, *we have*

$$\operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\big[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\big] = [\theta_\star]. \tag{2.10}$$

Before proving Theorem 1, we first extend the definition of the conditional density on Y in (2.8) to a conditional density on Y^n.

DEFINITION 3. 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

$$\begin{aligned} &p_n^{\theta,\theta'}(y_{1:n}|y_{-\infty:0}) \\ &\quad := \int_{\mathsf{X}^n} \Phi^{\theta,\theta'}(y_{-\infty:0}; \mathrm{d}x_0) \prod_{k=0}^{n-1} \kappa^\theta \langle y_k, y_{k+1}\rangle(x_k; \mathrm{d}x_{k+1}). \end{aligned} \tag{2.11}$$

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-2) 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 1. *Suppose that Assumption* (K-2) *holds and let* $p_n^{\theta,\theta'}(\cdot|\cdot)$ *be as defined in Definition* 3. *Then for all* $\theta, \theta' \in \Theta$ *and* $n \geq 2$, *we have*

$$p_n^{\theta,\theta'}(Y_{1:n}|Y_{-\infty:0}) = p_1^{\theta,\theta'}(Y_n|Y_{-\infty:n-1})\, p_{n-1}^{\theta,\theta'}(Y_{1:n-1}|Y_{-\infty:0}), \quad \tilde{\mathbb{P}}^{\theta'}\text{-a.s.} \tag{2.12}$$

The proof of this lemma is postponed to Section A.2 in the Appendix. We now have all the tools for proving the main result.

PROOF OF THEOREM 1. Within this proof section, we will 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.

For all $\theta \in \Theta$, we have by conditioning on $Y_{-\infty:0}$ and by using (2.9),

$$\begin{aligned} &\tilde{\mathbb{E}}^{\theta_\star}\big[\ln p^{\theta_\star}(Y_1|Y_{-\infty:0})\big] - \tilde{\mathbb{E}}^{\theta_\star}\big[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\big] \\ &\quad = \tilde{\mathbb{E}}^{\theta_\star}\left[\tilde{\mathbb{E}}^{\theta_\star}\left[\ln \frac{p^{\theta_\star}(Y_1|Y_{-\infty:0})}{p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})}\middle| Y_{-\infty:0}\right]\right] \\ &\quad = \tilde{\mathbb{E}}^{\theta_\star}\big[\mathrm{KL}\big(p_1^{\theta_\star}(\cdot|Y_{-\infty:0}) \,\|\, p_1^{\theta,\theta_\star}(\cdot|Y_{-\infty:0})\big)\big], \end{aligned} \tag{2.13}$$

where $\mathrm{KL}(p\|q)$ denotes the Kullback–Leibler divergence between the densities p and q. The nonnegativity of the Kullback–Leibler divergence shows that $\theta_\star$ belongs to the maximizing set on the left-hand side of (2.10). This implies

$$\operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\big[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\big] \supseteq [\theta_\star], \tag{2.14}$$

where we have used the following lemma.

LEMMA 2. *Assume that* (K-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* $\theta_\star \in \Theta$,

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

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

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

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

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

which completes the proof. □

The proof of the reverse inclusion of (2.14) 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 (2.13), we have

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

Consequently,

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

Applying Lemma 1 and using that $\tilde{\mathbb{P}}^{\theta_\star}$ is shift-invariant, this relation propagates to all $n \geq 2$, so that

$$p^{\theta_\star}(Y_{1:n}|Y_{-\infty:0}) = p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0}), \qquad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{2.15}$$

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

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

where the last equality follows from (2.9). Using Definition 3 and Tonelli's theorem, we obtain

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

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

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

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

$$\tilde{\mathbb{P}}^{\theta_\star}\big(\mathsf{Y}^{\mathbb{Z}_-}\times B\big) = \mathbb{P}^\theta_{\pi^{\theta,\theta_\star}}\big(\mathsf{X}^{\mathbb{Z}_+}\times(\mathsf{Y}\times B)\big). \tag{2.16}$$

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 $\pi^\theta=\pi^{\theta,\theta_\star}$, which by (K-1) is equivalent to showing 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 $(A,B)\in\mathcal{X}\times\mathcal{Y}$,

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

where in the last equality we have used (2.15). Using (K-2), we then get

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

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

2.3. *Construction of the kernel* $\Phi^{\theta,\theta'}$ *as a backward limit.* Again, all along this section, Assumption (K-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 (K-1), Theorem 1 fundamentally relies on Assumption (K-2). 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 HMMs, 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 4. 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 $A \in \mathcal{X}$,

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

We will 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 derive the desired $\Phi^{\theta,\theta'}$ we 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-3) below, which requires the following definition. For all $\theta \in \Theta$ and for all nonnegative measurable functions f defined on X, we set

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

We can now state the assumption as follows:

(K-3) 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 $f \in \mathcal{F}$,

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

The next lemma shows that, provided that $\mathcal{F}$ is rich enough, Assumption (K-3) can be directly used to obtain Assumption (K-2). 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 3. *Suppose that Assumption* (K-3) *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-2).

PROOF. Let $x \in \mathsf{X}$ be given in Assumption (K-3). From Definition 4, we may write, for all $f \in \mathcal{F}$, setting $x_{-m} = x$,

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

and, similarly,

$$\Phi^{\theta}_{x,m+1}(Y_{-m:1}; f) = \frac{\int f(x_1) \prod_{k=-m}^{0} \kappa^{\theta}\langle Y_k, Y_{k+1}\rangle(x_k; \mathrm{d}x_{k+1})}{\int \prod_{k=-m}^{0} \kappa^{\theta}\langle Y_k, Y_{k+1}\rangle(x_k; \mathrm{d}x_{k+1})}. \tag{2.17}$$

Dividing both numerator and denominator of (2.17) 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 (2.17) can be rewritten as

$$\Phi^{\theta}_{x,m+1}(Y_{-m:1}; f) = \frac{\Phi^{\theta}_{x,m}(Y_{-m:0}; \kappa^{\theta}\langle Y_0, Y_1\rangle(\cdot; f))}{\Phi^{\theta}_{x,m}(Y_{-m:0}; \kappa^{\theta}\langle Y_0, Y_1\rangle(\cdot; \mathbb{1}_{\mathsf{X}}))}. \tag{2.18}$$

Letting $m \to \infty$ and applying Assumption (K-3), then $\tilde{\mathbb{P}}^{\theta'}$-a.s.,

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

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

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 5. Consider a partially observed and partially dominated Markov model given in Definition 1 with Markov kernels $(K^{\theta})_{\theta \in \Theta}$. We will say that this model is a hidden Markov model if the kernel K^{θ} satisfies

$$K^{\theta}\big((x, y); \mathrm{d}x' \, \mathrm{d}y'\big) = Q^{\theta}\big(x; \mathrm{d}x'\big) G^{\theta}\big(x'; \mathrm{d}y'\big). \tag{3.1}$$

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

In (3.1), 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 $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

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

One can directly observe that the unnormalized kernel $\kappa^\theta\langle y, y'\rangle$ defined in (2.4) does no longer depend on y, and in this case, one can write

$$\kappa^\theta\langle y, y'\rangle(x; \mathrm{d}x') = \kappa^\theta\langle y'\rangle(x; \mathrm{d}x') = Q^\theta(x; \mathrm{d}x')g^\theta(x'; y'). \tag{3.2}$$

For any integer $n \geq 1$, $\theta \in \Theta$ and 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 by, for all $x_0 \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\mathbf{L}^\theta\langle y_{0:n-1}\rangle(x_0; A) = \int \cdots \int \left[\prod_{k=0}^{n-1} g^\theta(x_k; y_k) Q^\theta(x_k; \mathrm{d}x_{k+1})\right] \mathbb{1}_A(x_n), \tag{3.3}$$

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 [11] 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 (K-1).

REMARK 5. 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 1) in the context of HMMs by considering the assumptions of [11] 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 so-called *local Doeblin set* (in the particular case where $r = 1$) as introduced in [11], Definition 1.

DEFINITION 6. A set C is local Doeblin with respect to the family of kernels $(Q^\theta)_{\theta \in \Theta}$ if there exist positive constants $\varepsilon_C^-, \varepsilon_C^+$ 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 $x \in C$,

$$\varepsilon_C^- \lambda_C^\theta(A) \le Q^\theta(x; A \cap C) \le \varepsilon_C^+ \lambda_C^\theta(A).$$

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

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

(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 $y \in K$,

$$\sup_{x\in C^c} g^\theta(x; y) \le \eta \sup_{x\in\mathsf{X}} g^\theta(x; y) < \infty. \tag{3.4}$$

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

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

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

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

REMARK 6. Under (D-1), for all $\theta \in \Theta$, the Markov kernel Q^θ is μ-irreducible, so that, using Remark 5, (K-1) reduces to the existence of a stationary distribution for Q^θ.

REMARK 7. Assumptions (D-3), (D-4) and (D-5) and (2.3) in Definition 1 correspond to (A1), (A2) and (A3) in [11], where the blocks are of size $r = 1$.

REMARK 8. Assumption (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 4 writes

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

The distribution $\Phi_{x,n}^\theta(Y_{0:n}; \cdot)$ is usually referred to as the *filter distribution*. Proposition 1 (below) can be derived from [11], Proposition 1. For blocks of size 1, the initial distributions in [11] 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

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^- \inf_{\theta\in\Theta} \int \xi(\mathrm{d}x) g^\theta(x; Y_0) Q^\theta(x; D)\right] < \infty, \tag{3.5}$$

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 (3.5), so the constraint on the initial distribution vanishes in our case.

PROPOSITION 1. *Assume* (D-3) *and* (D-4). *Then the following assertions hold*:

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

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

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

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

where

$$\ell(\theta, \theta_\star) := \tilde{\mathbb{E}}^{\theta_\star}\left[\ln \int \Phi^{\theta,\theta_\star}(Y_{-\infty:0}; \mathrm{d}x_0) \kappa^\theta \langle Y_1\rangle (x_0; \mathsf{X})\right]. \tag{3.6}$$

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

THEOREM 2. *Assume that* (K-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 Assumptions* (D-1)–(D-5) *hold. Then, for any probability measure ξ,*

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

PROOF. According to [11], Theorem 2, $\theta \mapsto \ell(\theta, \theta_\star)$ defined by (3.6) is upper semi-continuous [so that $\Theta_\star := \operatorname{argmax}_{\theta\in\Theta} \ell(\theta, \theta_\star)$ is nonempty] and moreover

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

The proof then follows from Theorem 1, provided that $\ell(\theta, \theta_\star)$ can be expressed as in the statement of Theorem 1 and that (K-2) is satisfied. First note that, for $\theta \neq \theta_\star$, the integral appearing within the logarithm in (3.6) corresponds to $p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})$ with $p^{\theta,\theta_\star}$ as defined in (2.8).

Let $\mathcal{F}$ be a countable separating class of nonnegative bounded functions containing $\mathbb{1}_{\mathsf{X}}$; see [26], Theorem 6.6, Chapter 6, for the existence of such a class.

By Lemma 3, we check (K-2) by showing that (K-3) is satisfied. Condition (D-2) and (3.2) imply that for all bounded functions f, $\mathcal{F}_f^\theta$ is a class of bounded functions, and this in turn implies (K-3) by applying Proposition 1(i) to all x. Thus, (K-2) is satisfied, and for $\theta \neq \theta_\star$, $\ell(\theta, \theta_\star)$ can be expressed as in the statement of Theorem 1. To complete the proof, it only remains to consider the case where $\theta = \theta_\star$ and to show that $\ell(\theta_\star, \theta_\star)$ can be written as

$$\ell(\theta_\star, \theta_\star) = \tilde{\mathbb{E}}^{\theta_\star}[\ln p_1^{\theta_\star}(Y_1 | Y_{-\infty:0})], \tag{3.7}$$

where $p_1^{\theta_\star}(\cdot|\cdot)$ is the conditional density given in (2.9). According to [3], Theorem 1, we have

$$\tilde{\mathbb{E}}^{\theta_\star}[\ln p_1^{\theta_\star}(Y_1 | Y_{-\infty:0})] = \lim_{n\to\infty} n^{-1} \ln \pi_1^{\theta_\star} \mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}, \qquad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{3.8}$$

On the other hand, applying Proposition 1(ii) yields

$$\ell(\theta_\star, \theta_\star) = \lim_{n\to\infty} n^{-1} \ln \xi \mathbf{L}^{\theta_\star}\langle Y_{0:n-1}\rangle \mathbb{1}_{\mathsf{X}}, \qquad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{3.9}$$

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

$$\frac{\mathrm{d}\xi \mathbf{L}^{\theta_\star}\langle y_0\rangle}{\mathrm{d}\mu}(x_1) = \int \xi(\mathrm{d}x_0) g^{\theta_\star}(x_0; y_0) q^{\theta_\star}(x_0; x_1). \tag{3.10}$$

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

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

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

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

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

$$\frac{\mathrm{d}\pi_1^{\theta_\star} \mathbf{L}^{\theta_\star}\langle y_0\rangle}{\mathrm{d}\mu}(x_1) = \int \pi_1^{\theta_\star}(\mathrm{d}x_0) g^{\theta_\star}(x_0; y_0) q^{\theta_\star}(x_0; x_1). \tag{3.11}$$

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

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

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

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

where $R(x_1, y_0)$ is the ratio between (3.10) and (3.11), which are positive densities with respect to $\mu \otimes \nu$. Since the denominator (3.11) is the density of (X_1, Y_0) under $\mathbb{P}^{\theta_\star}$, we then have

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

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

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

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

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

Combining with (3.8) and (3.9), we finally obtain (3.7), which completes the proof. □

3.3. *A polynomially ergodic example.* As an application of Theorem 2, 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 given in (E-1)] and $b, c > 0$ such that $\mathbb{E}(|V_0|^\beta) < \infty$ and $h(v) \ge b\mathrm{e}^{-c|v|^\beta}$ 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 $(X_k)_{k\in\mathbb{Z}_+}$ 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 $(x, x') \in \mathbb{R}_+^2$,

$$q^\theta(x; x') = r(x' - x + m)\mathbb{1}\{x' > 0\} + \left(\int_{-\infty}^{m-x} r(u)\,\mathrm{d}u\right)\mathbb{1}\{x' = 0\}. \tag{3.12}$$

Moreover, (1.1) implies

$$g^\theta(x; y) = h(y - ax). \tag{3.13}$$

Following [20], we have the following lemma.

LEMMA 4. *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, *defined on* $\mathsf{X} = \mathbb{R}_+$, *satisfies* $\int \pi_1^\theta(\mathrm{d}x) x^\beta < \infty$, *for all* $\beta \in [1, \alpha - 1)$.

The proof of this lemma is postponed to Section A.3 in the Appendix.

PROPOSITION 2. *Consider the HMM of Example* 1 *under Assumptions* (E-1) *and* (E-2). *Then* (K-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* $\theta_\star \in \Theta$,

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

PROOF. To apply Theorem 2, we need to check (K-1) and (D-1)–(D-5). First observe that Assumption (K-1) immediately follows from Remark 5 and Lemma 4, and Assumptions (D-1) and (D-2) directly follow from the positiveness of the density r and the boundedness of the density h, respectively. Now, using (E-1)(a), it can be easily shown 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, $\inf\{r(x' - x + m) : (x, x', m) \in D^2 \times [\underline{m}, \overline{m}]\} > 0$, which by (3.12) implies

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

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

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

By (E-2)(b), there exist positive constants b and c such that $h(v) \geq b\mathrm{e}^{-c|v|^\beta}$. Plugging this into (3.13) yields

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

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

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

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 where

$n=1$ 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 on 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 (3.13) and the continuity of h. Similarly, the continuity of $\theta \mapsto q^{\theta}(x_0;x_1)$ follows from (3.12) 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

$$\int \mu(\mathrm{d}x_1)\sup_{\theta\in\Theta} q^{\theta}(x_0;x_1)\mathbf{L}^{\theta}\langle y_{1:n}\rangle(x_1;\mathsf{X}) < \infty \tag{3.14}$$

holds. Note further that by the expression of $q^{\theta}(x_0;x_1)$ given in (3.12) and the tail assumption (E-1)(b), we obtain 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 that $\mathbf{L}^{\theta}\langle y_{1:n}\rangle(x_1;\mathsf{X}) \le (\sup_{x\in\mathbb{R}} h(x))^n$ yields (3.14). Finally, we have (D-5), and thus Theorem 2 holds under (E-1) and (E-2). $\square$

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

$$K^{\theta}((x,y);\mathrm{d}x'\,\mathrm{d}y') = \delta_{\psi^{\theta}_y(x)}(\mathrm{d}x')G^{\theta}(x';\mathrm{d}y'), \tag{4.1}$$

where δ_a denotes the Dirac mass at point a, G^{θ} is a probability kernel on $\mathsf{X}\times\mathcal{Y}$ and $((x,y)\mapsto \psi^{\theta}_y(x))_{\theta\in\Theta}$ 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 $(\mathsf{X},\mathrm{d}_{\mathsf{X}})$ is a complete separable metric space and $\mathcal{X}$ denotes the associated Borel σ-field.

Note that a Markov chain $((X_k, Y_k))_{k \in \mathbb{Z}_+}$ with probability kernel given by (4.1) can be equivalently defined by the following recursions:

$$X_{k+1} = \psi^\theta_{Y_k}(X_k),\\ Y_{k+1} | X_{0:k+1}, Y_{0:k} \sim G^\theta(X_{k+1}; \cdot). \tag{4.2}$$

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 $\psi^\theta_y(x)$ 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, (2.3) 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 (K-1), we assume that the model is well specified, that is, 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,

$$y_{1:n} \mapsto \prod_{k=1}^n g^\theta\big(\psi^\theta\langle y_{1:k-1}\rangle(x); y_k\big), \tag{4.3}$$

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 $\psi^\theta_{y_1}, \psi^\theta_{y_2}, \dots,$ and $\psi^\theta_{y_p}$,

$$\psi^\theta\langle y_{1:p}\rangle = \psi^\theta_{y_p} \circ \psi^\theta_{y_{p-1}} \circ \cdots \circ \psi^\theta_{y_1}, \tag{4.4}$$

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

$$\hat{\theta}_{x,n} \in \operatorname*{argmax}_{\theta \in \Theta} \mathsf{L}^\theta_{x,n}\langle Y_{1:n}\rangle, \tag{4.5}$$

where

$$\mathsf{L}^\theta_{x,n}\langle y_{1:n}\rangle := n^{-1} \ln\left(\prod_{k=1}^n g^\theta\big(\psi^\theta\langle y_{1:k-1}\rangle(x); y_k\big)\right). \tag{4.6}$$

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, $\hat{\theta}_{x,n}$ belongs 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

$$\lim_{m\to\infty} \psi^\theta \langle Y_{-m:0}\rangle(x) = \psi^{\theta,\theta_\star}\langle Y_{-\infty:0}\rangle, \qquad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{4.7}$$

(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 $x \mapsto \psi_y^\theta(x)$ is continuous on X.

In observation-driven models, the kernel κ^θ defined in (2.4) reads

$$\begin{aligned}\kappa^\theta\langle y, y'\rangle(x; \mathrm{d}x') &= g^\theta(x'; y')\delta_{\psi_y^\theta(x)}(\mathrm{d}x') \\ &= g^\theta\big(\psi_y^\theta(x); y'\big)\delta_{\psi_y^\theta(x)}(\mathrm{d}x'),\end{aligned} \tag{4.8}$$

and the probability kernel $\Phi_{x,n}^\theta$ in Definition 4 reads, for all $x \in \mathsf{X}$ and $y_{0:n} \in \mathsf{Y}^{n+1}$,

$$\Phi_{x,n}^\theta(y_{0:n}; \cdot) = \delta_{\psi^\theta\langle y_{0:n-1}\rangle(x)} \tag{4.9}$$

[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 1.

THEOREM 3. *Assume that* (K-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.* $y_{-\infty:0} \in \mathsf{Y}^{\mathbb{Z}_-}$,

$$p^{\theta,\theta_\star}(y_1|y_{-\infty:0}) = \begin{cases} g^\theta\big(\psi^{\theta,\theta_\star}\langle y_{-\infty:0}\rangle; y_1\big), & \text{if } \theta \neq \theta_\star, \\ p_1^\theta(y_1|y_{-\infty:0}) \text{ as defined by (2.9)}, & \text{otherwise.}\end{cases} \tag{4.10}$$

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

$$\operatorname*{argmax}_{\theta\in\Theta} \tilde{\mathbb{E}}^{\theta_\star}\big[\ln p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0})\big] = [\theta_\star]. \tag{4.11}$$

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

$$\Phi^{\theta,\theta_\star}(y_{-\infty:0}; \cdot) = \delta_{\psi^{\theta,\theta_\star}\langle y_{-\infty:-1}\rangle}, \qquad \text{for all } y_{-\infty:0} \in \mathsf{Y}^{\mathbb{Z}_-}, \tag{4.12}$$

and that for $\theta \neq \theta_\star$, the conditional density $p^{\theta,\theta_\star}$ defined by (2.8) satisfies

$$p^{\theta,\theta_\star}(y|Y_{-\infty:0}) = g^\theta\big(\psi^{\theta,\theta_\star}\langle Y_{-\infty:0}\rangle; y\big), \qquad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{4.13}$$

By Lemma 3, it is sufficient to prove that Assumption (K-3) 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 [26], Theorem 6.6, Chapter 6, there is a countable and separating subclass $\mathcal{F}$ of nonnegative 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 (4.8), we have

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

By (4.9), (C-1) and (4.12), we obtain (K-3) with x chosen as in (C-1).

To conclude, we need to show (4.13). Note that (4.12) together with (4.8) and the usual definition (2.8) of $p^{\theta,\theta_\star}$ yields

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

By Assumption (C-3) and the definition of $\psi^{\theta,\theta_\star}\langle\cdot\rangle$ in (C-1), we get (4.13). $\square$

4.3. *Examples*. In the context of observation-driven time series, easy-to-check conditions are derived in [14] in order to establish the convergence of the MLE $\hat{\theta}_{x,n}$ defined by (4.5) to the maximizing set of the asymptotic normalized log-likelihood. It turns out that the conditions of Theorem 1, [14], also imply the conditions of Theorem 3. More precisely, the assumptions (B-2) and (B-3) of [14], Theorem 1, are stronger than (C-2) and (C-3) used in Theorem 3 above, and it is shown that the assumptions of Theorem 1, [14], imply (C-1) (see the proof of Lemma 2 in Section 6.3 of [14]). Moreover, the conditions of Theorem 1 are shown to be satisfied in the context of Examples 2 and 3 (see [14], Theorems 3 and 4), provided that Θ in (4.5) 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: POSTPONED PROOFS

A.1. Proof of equation (2.7). Let $\theta \in \Theta$. Recall that in Remark 3, Φ^θ is defined as the probability kernel of the conditional distribution of X_0 given $Y_{-\infty:0}$ under $\mathbb{P}^\theta$, that is, for all $A \in \mathcal{X}$,

$$\Phi^\theta(Y_{-\infty:0}; A) = \mathbb{P}^\theta(X_0 \in A | Y_{-\infty:0}), \qquad \tilde{\mathbb{P}}^\theta\text{-a.s.}$$

Conditioning on X_0, Y_0 and using the definition of κ^θ in (2.4), we get that, for all $A \in \mathcal{X}$, $B \in \mathcal{Y}$ and $C \in \mathcal{Y}^{\otimes \mathbb{Z}_-}$,

$$\begin{aligned}
&\mathbb{P}^\theta(X_1 \in A, Y_1 \in B, Y_{-\infty:0} \in C) \\
&\quad = \mathbb{E}^\theta\left[\int_B \kappa^\theta\langle Y_0, y_1\rangle(X_0; A)\mathbb{1}_C(Y_{-\infty:0})\nu(\mathrm{d}y_1)\right] \qquad \text{(A.1)}\\
&\quad = \tilde{\mathbb{E}}^\theta\left[\int_{\mathsf{X}\times B} \Phi^\theta(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle(x_0; A)\mathbb{1}_C(Y_{-\infty:0})\nu(\mathrm{d}y_1)\right].
\end{aligned}$$

Let us denote

$$\hat{\Phi}^\theta(Y_{-\infty:0}, y_1; A) = \frac{\int \Phi^\theta(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle(x_0; A)}{\int \Phi^\theta(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle(x_0; \mathsf{X})},$$

which is always defined since the denominator does not vanish by Remark 1. With this notation, we deduce from (A.1) that

$$\begin{aligned}
&\mathbb{P}^\theta(X_1 \in A, Y_1 \in B, Y_{-\infty:0} \in C) \\
&\quad = \tilde{\mathbb{E}}^\theta\biggl[\int_B \hat{\Phi}^\theta(Y_{-\infty:0}, y_1; A)\biggl(\int \Phi^\theta(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle(x_0; \mathsf{X})\biggr) \\
&\qquad \times \mathbb{1}_C(Y_{-\infty:0})\nu(\mathrm{d}y_1)\biggr].
\end{aligned}$$

This can be more compactly written as

$$\begin{aligned}
&\mathbb{P}^\theta(X_1 \in A, Y_1 \in B, Y_{-\infty:0} \in C) \\
&\quad = \tilde{\mathbb{E}}^\theta\biggl[\int \hat{\Phi}^\theta(Y_{-\infty:0}, y_1; A) \qquad\qquad \text{(A.2)}\\
&\qquad \times \mathbb{1}_B(y_1)\mathbb{1}_C(Y_{-\infty:0})\Phi^\theta(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle(x_0; \mathsf{X})\nu(\mathrm{d}y_1)\biggr].
\end{aligned}$$

Observe that (A.1) with $A = \mathsf{X}$ provides a way to write $\tilde{\mathbb{E}}^\theta[g(Y_{-\infty:0}, Y_1)]$ for $g = \mathbb{1}_{C\times B}$ that can be extended to any nonnegative measurable function g defined on $\mathsf{Y}^{\mathbb{Z}_-} \times \mathsf{Y}$ as

$$\begin{aligned}
&\tilde{\mathbb{E}}^\theta[g(Y_{-\infty:0}, Y_1)] \\
&\quad = \tilde{\mathbb{E}}^\theta\biggl[\int g(Y_{-\infty:0}, y_1)\Phi^\theta(Y_{-\infty:0}; \mathrm{d}x_0)\kappa^\theta\langle Y_0, y_1\rangle(x_0; \mathsf{X})\nu(\mathrm{d}y_1)\biggr].
\end{aligned}$$

Now, we observe that the right-hand side of (A.2) can be interpreted as the right-hand side of the previous display with $g(Y_{-\infty:0}, y_1) = \hat{\Phi}^\theta(Y_{-\infty:0}, y_1; A)\mathbb{1}_B(y_1) \times \mathbb{1}_C(Y_{-\infty:0})$. Hence, we conclude that, for all $A \in \mathcal{X}$ and $C \in \mathcal{Y}^{\otimes\mathbb{Z}_-}$,

$$\mathbb{P}^\theta(X_1 \in A, Y_1 \in B, Y_{-\infty:0} \in C) = \mathbb{E}^\theta[\hat{\Phi}^\theta(Y_{-\infty:0}, Y_1; A)\mathbb{1}_B(Y_1)\mathbb{1}_C(Y_{-\infty:0})].$$

Notice that $\hat{\Phi}^\theta(Y_{-\infty:0}, Y_1; A)$ precisely is the probability kernel on $(\mathsf{Y}^{\mathbb{Z}_-} \times \mathsf{Y}) \times \mathcal{X}$ appearing on the left-hand side of (2.7). The last display implies that this probability kernel is the conditional distribution of X_1 given $Y_{-\infty:1}$ under $\mathbb{P}^\theta$, which concludes the proof of (2.7).

A.2. Proof of Lemma 1. First observe that, by induction on n, having (2.12) for all $n \geq 2$ is equivalent to having, for all $n \geq 2$,

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

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

$$p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0}) = p^{\theta,\theta_\star}(Y_{2:n}|Y_{-\infty:1})p^{\theta,\theta_\star}(Y_1|Y_{-\infty:0}), \qquad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{A.3}$$

Thus to conclude the proof, we only need to show that (A.3) holds for all $n \geq 2$. By Definition 3, we have, for all $n \geq 2$ and $y_{-\infty:n} \in \mathsf{Y}^{\mathbb{Z}_-}$,

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

Using (K-2) we now get, for all $n \geq 2$,

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

We conclude (A.3) by observing that, according to Definition 3, the second line of the last display is $p^{\theta,\theta_\star}(Y_{1:n}|Y_{-\infty:0})$.

A.3. Proof of Lemma 4. Let $\beta \in [1, \alpha - 1)$. Since $1 + \beta < \alpha$ and by (E-1)(b), we obtain $\mathbb{E}[(U_0^+)^{1+\beta}] < \infty$. Combining this with $\mathbb{E}[U_0 - m] = -m < 0$, we may apply [20], 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 well defined on $\mathsf{X} = \mathbb{R}_+$. Moreover, [20], Proposition 5.1, also shows that there exist a finite interval $C = [0, x_0]$ and some constants $\varrho, \varrho' \in (0, \infty)$ 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 [24], 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 the kernel Q^θ is not geometrically ergodic for all $\theta \in \Theta$ and this will be done by contradiction.

Now suppose on the contrary that the kernel Q^θ is geometrically ergodic for some $\theta \in \Theta$. Since the singleton $\{0\}$ is an accessible atom (for Q^θ), then there exists some $\rho > 1$ such that

$$\sum_{k=0}^{\infty} \rho^k \left|(Q^\theta)^k(0, \{0\}) - \pi_1^\theta(\{0\})\right| < \infty.$$

Hence, the atom $\{0\}$ is geometrically ergodic as defined in [24], Section 15.1.3. Applying [24], Theorem 15.1.5, then there exists some $\kappa > 1$ 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 the i.i.d. sequence $(U_k)_{k\in\mathbb{Z}_+}$ is linked to $(X_k)_{k\in\mathbb{Z}_+}$ through (1.1), and note that $\mathbb{E}_0[\kappa^{\tau_0}] = \mathbb{E}[\kappa^{\tau(0)}]$, where we have set for all $u \in \mathbb{R}$,

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

Now, 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

$$\liminf_{u\to\infty} \kappa^{-\gamma u}\mathbb{E}\big[\kappa^{\tau(-u+m)}\big] > 0. \tag{A.4}$$

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

$$\begin{aligned}\mathbb{E}\big[\kappa^{\tau(0)}\big] &\geq \mathbb{E}\big[\kappa^{\tau(0)}\mathbb{1}\{U_1 \geq m\}\big] = \mathbb{E}\big[\kappa^{1+\tilde{\tau}(-U_1+m)}\mathbb{1}\{U_1 \geq m\}\big]\\ &= \mathbb{E}\bigg[\int_m^\infty \kappa^{1+\tilde{\tau}(-u+m)} r(u)\,\mathrm{d}u\bigg]\\ &= \kappa\int_m^\infty \mathbb{E}\big[\kappa^{\tau(-u+m)}\big] r(u)\,\mathrm{d}u,\end{aligned}\tag{A.5}$$

where the last equality follows from $\tau \stackrel{d}{=} \tilde{\tau}$. Provided that (A.4) holds, the right-hand side of (A.5) is infinite since $r(u) \gtrsim u^{-\alpha-1}$ as $u \to \infty$ by (E-1)(b).

We now turn to the proof of (A.4). By Markov's inequality, we have for any $\gamma > 0$,

$$\kappa^{-\gamma u}\mathbb{E}\big[\kappa^{\tau(-u+m)}\big] \geq \mathbb{P}\big(\tau(-u+m) > \gamma u\big). \tag{A.6}$$

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

$$\begin{aligned}\Big\{\Big(\inf_{1\leq k\leq\gamma u} M_k\Big) - \gamma u m \geq -u + m\Big\} &\subset \Big\{\inf_{1\leq k\leq\gamma u}(M_k - km) \geq -u + m\Big\}\\ &= \big\{\tau(-u+m) > \gamma u\big\}.\end{aligned}\tag{A.7}$$

Moreover, since $(U_k)_{k\in\mathbb{Z}_+^*}$ is i.i.d. and centered, Doob's maximal inequality implies, for all $\tilde{\gamma} > 0$,

$$\begin{aligned}\mathbb{P}\Big(\inf_{1\leq k\leq\gamma u} M_k < -\tilde{\gamma}\Big) &\leq \mathbb{P}\Big(\sup_{1\leq k\leq\gamma u} |M_k| > \tilde{\gamma}\Big)\\ &\leq \frac{\mathbb{E}[|M_{\lfloor\gamma u\rfloor}|]}{\tilde{\gamma}} \leq \frac{\lfloor\gamma u\rfloor\mathbb{E}[|U_1|]}{\tilde{\gamma}}.\end{aligned}\tag{A.8}$$

Now, pick $\gamma > 0$ sufficiently small so that $\gamma \mathbb{E}[|U_1|]/(1-\gamma m) < 1$. Observe that for this γ, $\tilde{\gamma} = (1-\gamma m)u - m$ is positive for u sufficiently large, so that combining (A.8) with (A.7) and (A.6) yields

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

This shows (A.4) and the proof is complete.

Acknowledgments. The authors are thankful to the anonymous referee for insightful comments and helpful suggestions that led to improving the quality of the paper.

REFERENCES

[1] ALEXANDER, C. and LAZAR, E. (2006). Normal mixture GARCH(1, 1): Applications to exchange rate modelling. *J. Appl. Econometrics* **21** 307–336. MR2224236

[2] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

[3] BARRON, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303. MR0806226

[4] BICKEL, P. J., RITOV, Y. and RYDÉN, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26** 1614–1635. MR1647705

[5] BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. MR0853051

[6] BOLLERSLEV, T. (2008). Glossary to ARCH (GARCH). Technical report, CREATES Research Paper.

[7] COX, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scand. J. Stat.* **8** 93–115. MR0623586

[8] DAVIS, R. A., DUNSMUIR, W. T. M. and STREETT, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika* **90** 777–790. MR2024757

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

[10] DOUC, R., DOUKHAN, P. and MOULINES, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Process. Appl.* **123** 2620–2647. MR3054539

[11] DOUC, R. and MOULINES, E. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *Ann. Statist.* **40** 2697–2732. MR3097617

[12] DOUC, R., MOULINES, E., OLSSON, J. and VAN HANDEL, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.* **39** 474–513. MR2797854

[13] DOUC, R., MOULINES, É. and RYDÉN, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32** 2254–2304. MR2102510

[14] DOUC, R., ROUEFF, F. and SIM, T. (2015). Handy sufficient conditions for the convergence of the maximum likelihood estimator in observation-driven models. *Lith. Math. J.* **55** 367–392. MR3379032

[15] DOUKHAN, P., FOKIANOS, K. and TJØSTHEIM, D. (2012). On weak dependence conditions for Poisson autoregressions. *Statist. Probab. Lett.* **82** 942–948. MR2910041

[16] EPHRAIM, Y. and MARK, B. L. (2012). Bivariate Markov processes and their estimation. *Foundations and Trends in Signal Processing* **6** 1–95.

[17] FOKIANOS, K., RAHBEK, A. and TJØSTHEIM, D. (2009). Poisson autoregression. *J. Amer. Statist. Assoc.* **104** 1430–1439. MR2596998

[18] GASSIAT, E., CLEYNEN, A. and ROBIN, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.* **26** 61–71. MR3439359

[19] HAAS, M., MITTNIK, S. and PAOLELLA, M. S. (2004). Mixed normal conditional heteroskedasticity. *J. Financ. Econom.* **2** 211–250.

[20] JARNER, S. F. and ROBERTS, G. O. (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Probab.* **12** 224–247. MR1890063

[21] JENSEN, J. L. and PETERSEN, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27** 514–535. MR1714719

[22] KALLENBERG, O. (2002). *Foundations of Modern Probability*, 2nd ed. *Probability and Its Applications*. Springer, New York. MR1876169

[23] LEROUX, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.* **40** 127–143. MR1145463

[24] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

[25] NEUMANN, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17** 1268–1284. MR2854772

[26] PARTHASARATHY, K. R. (2005). *Probability Measures on Metric Spaces*. AMS Chelsea Publishing, Providence, RI. Reprint of the 1967 original. MR2169627

[27] PIECZYNSKI, W. (2003). Pairwise Markov chains. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 634–639.

[28] STREETT, S. (2000). Some observation driven models for time series of counts. Ph.D. thesis, Dept. Statistics, Colorado State Univ., Fort Collins, CO.

[29] TUOMINEN, P. and TWEEDIE, R. L. (1994). Subgeometric rates of convergence of f-ergodic Markov chains. *Adv. in Appl. Probab.* **26** 775–798. MR1285459

[30] WOODARD, D. B., MATTESON, D. S. and HENDERSON, S. G. (2011). Stationarity of generalized autoregressive moving average models. *Electron. J. Stat.* **5** 800–828. MR2824817

[31] ZHU, F. (2011). A negative binomial integer-valued GARCH model. *J. Time Series Anal.* **32** 54–67. MR2790672

R. DOUC
INSTITUT MINES-TÉLÉCOM
TÉLÉCOM SUDPARIS
CNRS UMR 5157
9 RUE CHARLES FOURIER
91000 EVRY
FRANCE
E-MAIL: randal.douc@telecom-sudparis.eu

F. ROUEFF
T. SIM
LTCI CNRS
TÉLÉCOM PARISTECH
UNIVERSITÉ PARIS-SACLAY
46 RUE BARRAULT
75634 PARIS CEDEX 13
FRANCE
E-MAIL: roueff@telecom-paristech.fr
tepmonysim@gmail.com