arXiv:0912.4480v2 [math.ST] 9 Mar 2011

The Annals of Statistics
2011, Vol. 39, No. 1, 474–513
DOI: 10.1214/10-AOS834
© Institute of Mathematical Statistics, 2011

CONSISTENCY OF THE MAXIMUM LIKELIHOOD ESTIMATOR FOR GENERAL HIDDEN MARKOV MODELS[1]

By Randal Douc, Eric Moulines, Jimmy Olsson and Ramon van Handel

Télécom SudParis, Télécom ParisTech, Lund University and Princeton University

Consider a parametrized family of general hidden Markov models, where both the observed and unobserved components take values in a complete separable metric space. We prove that the maximum likelihood estimator (MLE) of the parameter is strongly consistent under a rather minimal set of assumptions. As special cases of our main result, we obtain consistency in a large class of nonlinear state space models, as well as general results on linear Gaussian state space models and finite state models.

A novel aspect of our approach is an information-theoretic technique for proving identifiability, which does not require an explicit representation for the relative entropy rate. Our method of proof could therefore form a foundation for the investigation of MLE consistency in more general dependent and non-Markovian time series. Also of independent interest is a general concentration inequality for V-uniformly ergodic Markov chains.

1. Introduction. A hidden Markov model (HMM) is a bivariate stochastic process $(X_k, Y_k)_{k\geq 0}$, where $(X_k)_{k\geq 0}$ is a Markov chain (often referred to as the state sequence) in a state space X and, conditionally on $(X_k)_{k\geq 0}$, $(Y_k)_{k\geq 0}$ is a sequence of independent random variables in a state space Y such that the conditional distribution of Y_k given the state sequence depends on X_k only. The key feature of HMM is that the state sequence $(X_k)_{k\geq 0}$ is not observable, so that statistical inference has to be carried out by means of the observations $(Y_k)_{k\geq 0}$ only. Such problems are far from straightforward

Received December 2009; revised April 2010.
[1]Supported in part by the Grant ANR-07-ROBO-0002-04.
AMS 2000 subject classifications. Primary 60F10, 62B10, 62F12, 62M09; secondary 60J05, 62M05, 62M10, 94A17.
Key words and phrases. Hidden Markov models, maximum likelihood estimation, strong consistency, V-uniform ergodicity, concentration inequalities, state space models.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Statistics*, 2011, Vol. 39, No. 1, 474–513. This reprint differs from the original in pagination and typographic detail.

due to the fact that the observation process $(Y_k)_{k\geq 0}$ is generally a dependent, non-Markovian time series [despite that the bivariate process $(X_k, Y_k)_{k\geq 0}$ is itself Markovian]. HMM appear in a large variety of scientific disciplines including financial econometrics [17, 25], biology [7], speech recognition [19], neurophysiology [11], etc., and the statistical inference for such models is therefore of significant practical importance [6].

In this paper, we will consider a parametrized family of HMM with parameter space Θ. For each parameter $\theta \in \Theta$, the dynamics of the HMM is specified by the transition kernel Q_θ of the Markov process $(X_k)_{k\geq 0}$, and by the conditional distribution G_θ of the observation Y_k given the signal X_k. For example, the state and observation sequences may be generated according to a nonlinear dynamical system (which defines implicitly Q_θ and G_θ) of the form

$$X_k = a_\theta(X_{k-1}, W_k),$$
$$Y_k = b_\theta(X_k, V_k),$$

where a_θ and b_θ are (nonlinear) functions and $(W_k)_{k\geq 1}$, $(V_k)_{k\geq 0}$ are independent sequences of i.i.d. random variables which are independent of X_0.

Throughout the paper, we fix a distinguished element $\theta^\star \in \Theta$. We will always presume that the kernel $Q_{\theta^\star}$ possesses a unique invariant probability measure $\pi_{\theta^\star}$, and we denote by $\bar{\mathbb{P}}_{\theta^\star}$ and $\bar{\mathbb{E}}_{\theta^\star}$ the law and associated expectation of the stationary HMM with parameter $\theta^\star$ (we refer to Section 2.1 for detailed definitions of these quantities). In the setting of this paper, we have access to a single observation path of the process $(Y_k)_{k\geq 0}$ sampled from the distribution $\bar{\mathbb{P}}_{\theta^\star}$. Thus, $\theta^\star$ is interpreted as the *true* parameter value, which is not known a priori. Our basic problem is to form a consistent estimate of $\theta^\star$ on the basis of the observations $(Y_k)_{k\geq 0}$ only, that is, without access to the hidden process $(X_k)_{k\geq 0}$. This will be accomplished by means of the maximum likelihood method.

The maximum likelihood estimator (MLE) is one of the backbones of statistics, and common wisdom has it that the MLE should be, except in "atypical" cases, consistent in the sense that it converges to the true parameter value as the number of observations tends to infinity. The purpose of this paper is to show that this is indeed the case for HMM under a rather minimal set of assumptions. Our main result substantially generalizes previously known consistency results for HMM, and can be applied to many models of practical interest.

1.1. *Previous work.* The study of asymptotic properties of the MLE in HMM was initiated in the seminal work of Baum and Petrie [3, 28] in the 1960s. In these papers, the state space X and the observation space Y were both presumed to be finite sets. More than two decades later, Leroux [23]

proved consistency for the case that X is a finite set and Y is a general state space. The consistency of the MLE in more general HMM has subsequently been investigated in a series of contributions [8, 9, 14, 21, 22] using a variety of methods. However, all these results require very restrictive assumptions on the underlying model, such as uniform positivity of the transition densities, which are rarely satisfied in applications (particularly in the case of a noncompact state space X). A general consistency result for HMM has hitherto remained lacking.

Though the consistency results above differ in the details of their proofs, all proofs have a common thread which serves also as the starting point for this paper. Let us therefore recall the basic approach for proving consistency of the MLE. Denote by $p^\nu(y_0^n;\theta)$ the likelihood of the observations Y_0^n for the HMM with parameter $\theta\in\Theta$ and initial measure $X_0\sim\nu$. The first step of the proof aims to establish that for any $\theta\in\Theta$, there is a constant $H(\theta^\star,\theta)$ such that

$$\lim_{n\to\infty} n^{-1}\log p^\nu(Y_0^n;\theta)=\lim_{n\to\infty} n^{-1}\bar{\mathbb{E}}_{\theta^\star}[\log p^\nu(Y_0^n;\theta)]=H(\theta^\star,\theta),\qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.}$$

For $\theta=\theta^\star$, this convergence follows from the generalized Shannon–Breiman–McMillan theorem [2], but for $\theta\neq\theta^\star$ the existence of the limit is far from obvious. Now set $K(\theta^\star,\theta)=H(\theta^\star,\theta^\star)-H(\theta^\star,\theta)$. Then $K(\theta^\star,\theta)\geq 0$ is the relative entropy rate between the observation laws of the parameters $\theta^\star$ and θ, respectively. The second step of the proof aims to establish identifiability, that is, that $K(\theta^\star,\theta)$ is minimized only at those parameters θ that are equivalent to $\theta^\star$ (in the sense that they give rise to the same stationary observation law). Finally, the third step of the proof aims to prove that the maximizer of the likelihood $\theta\mapsto p^\nu(Y_0^n;\theta)$ converges $\bar{\mathbb{P}}_{\theta^\star}$-a.s. to the maximizer of $H(\theta^\star,\theta)$, that is, to the minimizer of $K(\theta^\star,\theta)$. Together, these three steps imply consistency.

Let us note that one could write the likelihood as

$$n^{-1}\log p^\nu(Y_0^n;\theta)=\frac{1}{n}\sum_{k=0}^{n}\log p^\nu(Y_k|Y_0^{k-1};\theta),$$

where $p^\nu(Y_k|Y_0^{k-1};\theta)$ denotes the conditional density of Y_k given Y_0^{k-1} under the parameter θ (i.e., the one-step predictor). If the limit of $p^\nu(Y_1|Y_{-n}^0;\theta)$ as $n\to\infty$ can be shown to exist $\bar{\mathbb{P}}_{\theta^\star}$-a.s., existence of the relative entropy rate follows from the ergodic theorem and yields the explicit representation $H(\theta^\star,\theta)=\bar{\mathbb{E}}_{\theta^\star}[\log p^\nu(Y_1|Y_{-\infty}^0;\theta)]$. Such an approach was used in [3, 9]. Alternatively, the predictive distribution $p^\nu(Y_k|Y_0^{k-1};\theta)$ can be expressed in terms of a measure-valued Markov chain (the prediction filter), so that existence of the relative entropy rate, as well as an explicit representation for $H(\theta^\star,\theta)$, follows from the ergodic theorem for Markov chains if the prediction filter can be shown to be ergodic. This approach was used in [8, 21, 22].

In [23], the existence of the relative entropy rate is established by means of Kingman's subadditive ergodic theorem (the same approach is used indirectly in [28], which invokes the Furstenberg–Kesten theory of random matrix products). After some additional work, an explicit representation of $H(\theta^\star, \theta)$ is again obtained. However, as noted in [23], page 136, the latter is surprisingly difficult, as Kingman's ergodic theorem does not directly yield a representation of the limit as an expectation.

Though the proofs use different techniques, all the results above rely heavily on the explicit representation of $H(\theta^\star, \theta)$ in order to establish identifiability. This has proven to be one of the main difficulties in developing consistency results for more general HMM. For example, an attempt in [14] to generalize the approach of [23] failed to establish such a representation, and therefore to establish consistency except in a special example. Once identifiability has been established, standard techniques (such as Wald's method) can be used to show convergence of the maximizer of the likelihood, completing the proof.

For completeness, we note that a recent attempt [12] to prove consistency of the MLE for general HMM contains very serious problems in the proof [18] (not addressed in [13]), and therefore fails to establish the claimed results.

1.2. *Approach of this paper.* In this paper, we prove consistency of the MLE for general HMM under rather mild assumptions. Though our proof follows broadly the general approach described above, our approach differs from previous work in two key aspects. First, we note that it is not necessary to establish existence of the relative entropy rate. Indeed, rather than attempting to prove the existence of a limiting contrast function

$$\lim_{n\to\infty} n^{-1} \log p^\nu(Y_0^n; \theta) = H(\theta^\star, \theta), \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.},$$

which must then shown to be identifiable in the sense that $H(\theta^\star, \theta) < H(\theta^\star, \theta^\star)$ for parameters θ not equivalent to $\theta^\star$, it suffices to show directly that

$$\limsup_{n\to\infty} n^{-1} \log p^\nu(Y_0^n; \theta) < H(\theta^\star, \theta^\star), \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.}$$

[note that the existence of $H(\theta^\star, \theta^\star)$ is guaranteed by the Shannon–Breiman–McMillan theorem, and therefore poses little difficulty in the proof]. This simple observation implies that it suffices to obtain a convenient upper bound for $p^\nu(Y_0^n; \theta)$, which we accomplish by introducing the assumption that some iterate Q_θ^l of the transition kernel of the state sequence possesses a bounded density with respect to a σ-finite reference measure λ.

Second, and perhaps more importantly, we avoid entirely the need to obtain an explicit representation for the limiting contrast function $H(\theta^\star, \theta)$ which played a key role in all previous work. Instead, we develop in Section

4.2 a surprisingly powerful information-theoretic device which may be used to prove identifiability in a very general setting (see [26] for related ideas), and is not specific to HMM. This technique yields the following: in order to establish that the normalized relative entropy is bounded away from zero, that is,

$$\liminf_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}\left[n^{-1}\log\frac{\bar{p}(Y_0^n;\theta^\star)}{p^\nu(Y_0^n;\theta)}\right] > 0,$$

it suffices to show that there is a sequence of sets $(A_k)_{k\geq 0}$ such that

$$\liminf_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^n \in A_n) > 0, \qquad \limsup_{n\to\infty} n^{-1}\log \mathbb{P}_\theta^\nu(Y_0^n \in A_n) < 0$$

[here $\mathbb{P}_\theta^\nu$ is the law of the HMM with parameter θ and initial measure ν, while $\bar{p}(y_0^n;\theta^\star)$ denotes the likelihood of Y_0^n under $\bar{\mathbb{P}}_{\theta^\star}$]. It is rather straightforward to find such a sequence of sets, provided the law of the observations $(Y_k)_{k\geq 0}$ is ergodic under $\bar{\mathbb{P}}_{\theta^\star}$ and satisfies an elementary large deviations property under $\mathbb{P}_\theta^\nu$. These properties are readily established in a very general setting. In particular, we will show (Section 5) that any geometrically ergodic state sequence gives rise to the requisite large deviations property, so that our main result can be applied immediately to a large class of models of practical interest. (Let us note, however, that ergodicity of $\mathbb{P}_\theta^\nu$ is not necessary; see Section 3.2.)

Of course, there are some complications. Rather than investigating the likelihood function $p^\nu(Y_0^n;\theta)$ directly, the proof of our main result relies in an essential manner on the asymptotics of the process $p^\lambda(Y_0^n;\theta)$ where λ is the reference measure defined above. The latter process plays a special role in our proofs due to the fact that it satisfies a certain submultiplicativity property; this allows us to upper bound $n^{-1}\log p^\nu(Y_0^n;\theta)$ by a time average, which possesses an almost sure limit by Birkhoff's ergodic theorem (see the proof of Theorem 1 below for further details). As λ is typically only σ-finite, however, it is not immediately obvious that the problem is well-posed. Nonetheless, we will see that these complications can be resolved, provided that the HMM is sufficiently "observable" so that the improper likelihood $p^\lambda(Y_0^n;\theta)$ is well defined for sufficiently large n (under mild integrability conditions). As is demonstrated by the examples in Section 3, this is the case in a wide variety of applications.

Finally, let us note that the techniques used in the proof of our main result appear to be quite general. Though we have restricted our attention in this paper to the case of HMM, these techniques could form the foundation for consistency proofs in other dependent and non-Markovian time series models (such as, e.g., the autoregressive setting of [9]), which share many of the difficulties of statistical inference in hidden Markov models. Other asymptotic properties of the MLE, such as asymptotic normality, merit further investigation.

1.3. *Organization of the paper.* The remainder of the paper is organized as follows. In Section 2, we first introduce the setting and notations that are used throughout the paper. Then, we state our main assumptions and results. In Section 3, our main result is used to establish consistency in three general classes of models: linear-Gaussian state space models, finite state models, and nonlinear state space models of the vector ARCH type (this includes the stochastic volatility model and many other models of interest in time series analysis and financial econometrics). Section 4 is devoted to the proof of our main result. Finally, Section 5 is devoted to the proof of the fact that geometrically ergodic models satisfy the large deviations property needed for identifiability. In particular, we prove in Section 5.2 general Azuma–Hoeffding type concentration inequality for V-uniformly ergodic Markov chains, which is of independent interest.

2. Assumptions and main results.

2.1. *Canonical setup and notation.* We fix the following spaces throughout:

- X is a Polish space endowed with its Borel σ-field $\mathcal{X}$.
- Y is a Polish space endowed with its Borel σ-field $\mathcal{Y}$.
- Θ is a compact metric space endowed with its Borel σ-field $\mathcal{H}$.

X is the state space of the hidden Markov process, Y is the state space of the observations, and Θ is the parameter space of our model. We furthermore assume that Θ is endowed with a given equivalence relation[2] $\sim$, and denote the equivalence class of $\theta \in \Theta$ as $[\theta] \stackrel{\text{def}}{=} \{\theta' \in \Theta : \theta' \sim \theta\}$.

Our model is defined as follows: we are given a transition kernel $Q : \Theta \times \mathsf{X} \times \mathcal{X} \to [0,1]$, a positive σ-finite measure μ on $(\mathsf{Y}, \mathcal{Y})$, and a measurable function $g : \Theta \times \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ such that $\int g_\theta(x,y)\mu(dy) = 1$ for all θ, x. For each $\theta \in \Theta$, we can define the transition kernel T_θ on (X, Y) as

$$T_\theta[(x,y), C] \stackrel{\text{def}}{=} \int \mathbb{1}_C(x', y') g_\theta(x', y') \mu(dy') Q_\theta(x, dx').$$

[2]This is meant here in the broad sense, that is, $\sim$ is a binary relation on Θ indicating which elements $\theta \in \Theta$ should be viewed as "equivalent." We do not require $\sim$ to be transitive.

It should be emphasized that in the setting of this paper, the equivalence relation $\sim$ is presumed to be given as part of the model specification, rather than being defined in terms of the model: the statistician may choose up to which equivalence she wishes to estimate the true parameter generating the observations. One assumption of our main result [assumption (A6) below] then requires that parameters θ, θ' that are not equivalent, denoted $\theta \not\sim \theta'$, give rise to observation laws that are distinguishable in a suitable sense. In many cases, there is a natural equivalence relation which ensures that this is the case; see Section 2.3 below.

We will work on the measurable space $(\Omega, \mathcal{F})$ where $\Omega = (\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}$, $\mathcal{F} = (\mathcal{X} \otimes \mathcal{Y})^{\otimes \mathbb{N}}$, and the canonical coordinate process is denoted as $(X_k, Y_k)_{k \geq 0}$. For each $\theta \in \Theta$ and probability measure ν on $(\mathsf{X}, \mathcal{X})$, we define $\mathbb{P}_\theta^\nu$ to be the probability measure on $(\Omega, \mathcal{F})$ such that $(X_k, Y_k)_{k \geq 0}$ is a time homogeneous Markov process with initial measure $\mathbb{P}_\theta^\nu((X_0, Y_0) \in C) = \int \mathbb{1}_C(x, y) g_\theta(x, y) \times \mu(dy)\nu(dx)$ and transition kernel T_θ. Denote as $\mathbb{E}_\theta^\nu$ the expectation with respect to $\mathbb{P}_\theta^\nu$, and denote as $\mathbb{P}_\theta^{\nu, Y}$ the marginal of the probability measure $\mathbb{P}_\theta^\nu$ on $(\mathsf{Y}^{\mathbb{N}}, \mathcal{Y}^{\otimes \mathbb{N}})$.

Throughout the paper, we fix a distinguished element $\theta^\star \in \Theta$. We will always presume that the kernel $Q_{\theta^\star}$ possesses a unique invariant probability measure $\pi_{\theta^\star}$ on $(\mathsf{X}, \mathcal{X})$ [this follows from assumption (A1) below]. For ease of notation, we will write $\bar{\mathbb{P}}_{\theta^\star}, \bar{\mathbb{E}}_{\theta^\star}, \bar{\mathbb{P}}_{\theta^\star}^Y$ instead of $\mathbb{P}_{\theta^\star}^{\pi_{\theta^\star}}, \mathbb{E}_{\theta^\star}^{\pi_{\theta^\star}}, \mathbb{P}_{\theta^\star}^{\pi_{\theta^\star}, Y}$. Though the kernel, Q_θ need not be uniquely ergodic for $\theta \neq \theta^\star$ in our main result, we will obtain easily verifiable assumptions in a setting which implies that all Q_θ possess a unique invariant probability measure. When this is the case, we will denote as π_θ this invariant measure and we define $\bar{\mathbb{P}}_\theta, \bar{\mathbb{E}}_\theta, \bar{\mathbb{P}}_\theta^Y$ as above.

Under the measure $\mathbb{P}_\theta^\nu$, the process $(X_k, Y_k)_{k \geq 0}$ is a *hidden Markov model*. The hidden process $(X_k)_{k \geq 0}$ is a Markov chain in its own right with initial measure ν and transition kernel Q_θ, while the observations $(Y_k)_{k \geq 0}$ are conditionally independent given the hidden process with common observation kernel $G_\theta(x, dy) = g_\theta(x, y)\mu(dy)$. In the setting of this paper, we have access to a single observation path of the process $(Y_k)_{k \geq 0}$ sampled from the distribution $\bar{\mathbb{P}}_{\theta^\star}$. Thus, $\theta^\star$ is interpreted as the *true* parameter value, which is not known a priori. Our basic problem is to obtain a consistent estimate of $\theta^\star$ (up to equivalence, i.e., we aim to identify the equivalence class $[\theta^\star]$ of the true parameter) on the basis of the observations $(Y_k)_{k \geq 0}$ only, without access to the hidden process $(X_k)_{k \geq 0}$. This will be accomplished by the maximum likelihood method.

Define for any positive σ-finite measure ρ on $(\mathsf{X}, \mathcal{X})$

$$p^\rho(dx_{t+1}, y_s^t; \theta) \stackrel{\text{def}}{=} \int \rho(dx_s) \prod_{u=s}^{t} g_\theta(x_u, y_u) Q_\theta(x_u, dx_{u+1}),$$

$$p^\rho(y_s^t; \theta) \stackrel{\text{def}}{=} \int p^\rho(dx_{t+1}, y_s^t; \theta),$$

with the conventions $\prod_{u=v}^{w} a_u = 1$ if $v > w$ and for any sequence $(a_s)_{s \in \mathbb{Z}}$ and any integers $s \leq t$, $a_s^t \stackrel{\text{def}}{=} (a_s, \ldots, a_t)$. For ease of notation, we will write $p^x(y_s^t; \theta) \stackrel{\text{def}}{=} p^{\delta_x}(y_s^t; \theta)$ for $x \in \mathsf{X}$, and we write $\bar{p}(y_s^t; \theta) \stackrel{\text{def}}{=} p^{\pi_\theta}(y_s^t; \theta)$. Note that $p^\rho(dx_{t+1}, y_s^t; \theta)$ is a positive but not necessarily σ-finite measure. However, if ρ is a probability measure, then $p^\rho(dx_{t+1}, y_s^t; \theta)$ is a finite measure and $p^\rho(y_s^t; \theta) < \infty$.

If ν is a probability measure, then $p^\nu(y_0^n;\theta)$ is the likelihood of the observation sequence y_0^n under the law $\mathbb{P}_\theta^\nu$. The maximum likelihood method forms an estimate of $\theta^\star$ by maximizing $\theta \mapsto p^\nu(y_0^n;\theta)$, and we aim to establish consistency of this estimator. However, as the state space X is not compact, it will turn out to be essential to consider also $p^\lambda(y_0^n;\theta)$ for a positive σ-finite measure λ.

We conclude this section with some miscellaneous notation. For any function f, we denote as $|f|_\infty$ its supremum norm [e.g., $|g_\theta|_\infty \stackrel{\text{def}}{=} \sup_{(x,y)\in\mathsf{X}\times\mathsf{Y}} g_\theta(x, y)$]. As we will frequently integrate with respect to the measure μ, we will use the abridged notation dy instead of $\mu(dy)$, and we write $dy_s^t \stackrel{\text{def}}{=} \prod_{i=s}^t dy_i$. For any integer m and $\theta \in \Theta$, we denote by Q_θ^m the mth iterate of the kernel Q_θ. For any pair of probability measures $\mathbb{P}, \mathbb{Q}$ and function $V \geq 1$, we define the norm

$$\|\mathbb{P} - \mathbb{Q}\|_V \stackrel{\text{def}}{=} \sup_{f\,:\,|f|\leq V} \left| \int f \, d\mathbb{P} - \int f \, d\mathbb{Q} \right|.$$

Finally, the relative entropy (or Kullback–Leibler divergence) is defined as

$$\mathrm{KL}(\mathbb{P}||\mathbb{Q}) \stackrel{\text{def}}{=} \begin{cases} \displaystyle\int \log(d\mathbb{P}/d\mathbb{Q})\, d\mathbb{P}, & \text{if } \mathbb{P} \ll \mathbb{Q}, \\ \infty, & \text{otherwise,} \end{cases}$$

for any pair of probability measures $\mathbb{P}$ and $\mathbb{Q}$.

REMARK 1. Throughout the paper, we will encounter partial suprema of measurable functions [e.g., $y_0^n \mapsto \sup_{\theta\in\mathcal{U}} \bar{p}(y_0^n;\theta)$ for some measurable set $\mathcal{U} \in \mathcal{H}$]. As the supremum is taken over an uncountable set, such functions are not necessarily Borel-measurable. However, as all our state spaces are Polish, such functions are always guaranteed to be universally measurable ([4], Proposition 7.47). Similarly, a Borel-measurable (approximate) maximum likelihood estimator need not exist, but the Polish assumption ensures the existence of universally measurable maximum likelihood estimators ([4], Proposition 7.50). All probabilities and expectations can therefore be unambiguously extended to such quantities, which we will implicitly assume to be the case in the sequel.

2.2. *The consistency theorem.* Our main result establishes consistency of the MLE under assumptions (A1)–(A6) below, which hold in a large class of models. Various examples will be treated in Section 3 below.

(A1) The Markov kernel $Q_{\theta^\star}$ is positive Harris recurrent.

(A2) $\bar{\mathbb{E}}_{\theta^\star}[\sup_{x\in\mathsf{X}}(\log g_{\theta^\star}(x, Y_0))^+] < \infty$, $\bar{\mathbb{E}}_{\theta^\star}[|\log \int g_{\theta^\star}(x, Y_0)\pi_{\theta^\star}(dx)|] < \infty$.

Assumptions (A1), (A2) ensure the existence of the entropy rate for $\theta^\star$.

(A3) There is an integer $l \geq 1$, a measurable function $q : \Theta \times \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$, and a σ-finite measure λ on $(\mathsf{X}, \mathcal{X})$ such that $|q_\theta|_\infty < \infty$ and

$$Q_\theta^l(x, A) = \int \mathbb{1}_A(x') q_\theta(x, x') \lambda(dx')$$

for all $\theta \not\sim \theta^\star$, $x \in \mathsf{X}$, $A \in \mathcal{X}$.

Assumption (A3) states that an iterate of the transition kernel Q_θ possesses a density with respect to a σ-finite measure λ. This property will allow us to establish the asymptotics of the likelihood of $\mathbb{P}_\theta^\nu$ in terms of the improper likelihood $p^\lambda(\cdot; \theta)$. The measure λ plays a central role throughout the paper.

(A4) For every $\theta \not\sim \theta^\star$, there is a neighborhood $\mathcal{U}_\theta$ of θ such that

$$\sup_{\theta' \in \mathcal{U}_\theta} |q_{\theta'}|_\infty < \infty, \qquad \bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta' \in \mathcal{U}_\theta} \sup_{x \in \mathsf{X}} (\log g_{\theta'}(x, Y_0))^+\Big] < \infty,$$

and there is an integer r_θ such that

$$\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta' \in \mathcal{U}_\theta} (\log p^\lambda(Y_0^{r_\theta}; \theta'))^+\Big] < \infty.$$

(A5) For any $\theta \not\sim \theta^\star$ and $n \geq r_\theta$, the function $\theta' \mapsto p^\lambda(Y_0^n; \theta')$ is upper-semicontinuous at θ, $\bar{\mathbb{P}}_{\theta^\star}$-a.s.

Assumptions (A4) and (A5) are similar in spirit to the classical Wald conditions in the case of i.i.d. observations. However, an important difference with the classical case is that (A4) applies to $p^\lambda(y_0^{r_\theta}; \theta)$, which is not a probability density (as λ is typically only σ-finite). Assumption (A4) implies in particular that $p^\lambda(y_0^{r_\theta}; \theta)$ is $\bar{\mathbb{P}}_{\theta^\star}$-a.s. finite. When λ is σ-finite, this requires, in essence, that the observations contain some information on the range of values taken by the hidden process.

Finally, the key assumption (A6) below gives identifiability of the model. In principle, what is needed is that $\mathbb{P}_\theta^{\lambda, Y}$ is distinguishable from $\bar{\mathbb{P}}_{\theta^\star}^Y$ in a suitable sense. However, as λ may be σ-finite, $\mathbb{P}_\theta^{\lambda, Y}$ is not well defined. As a replacement, we will consider the probability measure $\tilde{\mathbb{P}}_\theta^\lambda$ defined by

$$\tilde{\mathbb{P}}_\theta^\lambda(Y_0^n \in A) = \int \mathbb{1}_A(y_0^n) \frac{p^\lambda(y_0^n; \theta)}{p^\lambda(y_0^{r_\theta}; \theta)} \bar{p}(y_0^{r_\theta}; \theta^\star)\, dy_0^n \tag{1}$$

for all $n \geq r_\theta$ and $A \in \mathcal{Y}^{\otimes(n+1)}$ (note that the definition of $\tilde{\mathbb{P}}_\theta^\lambda$ depends implicitly on $\theta^\star$ as well as on θ; the former dependence is suppressed for notational simplicity). Lemma 11 shows that $\tilde{\mathbb{P}}_\theta^\lambda$ is well defined, provided that (A4) holds and $p^\lambda(Y_0^{r_\theta}; \theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. The law $\tilde{\mathbb{P}}_\theta^\lambda$ is in essence a normalized version of $\mathbb{P}_\theta^{\lambda, Y}$, and (A6) should be interpreted in this spirit.

(A6) For every $\theta \not\sim \theta^\star$ such that $p^\lambda(Y_0^{r_\theta};\theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s., we have

$$\liminf_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^n \in A_n) > 0, \qquad \limsup_{n\to\infty} n^{-1} \log \tilde{\mathbb{P}}_\theta^\lambda(Y_0^n \in A_n) < 0$$

for some sequence of sets $A_n \in \mathcal{Y}^{\otimes(n+1)}$.

Despite that this assumption looks nontrivial, we will obtain sufficient conditions in Section 2.3 which are satisfied in a large class of models.

Having introduced the necessary assumptions, we now turn to the statement of our main result. Let $\ell_{\nu,n} : \theta \mapsto \log p^\nu(Y_0^n;\theta)$ be the log-likelihood function associated with the initial probability measure ν and the observations Y_0^n. An approximate *maximum likelihood estimator* $(\hat{\theta}_{\nu,n})_{n\geq 0}$ is defined as a sequence of (universally) measurable functions $\hat{\theta}_{\nu,n}$ of Y_0^n such that

$$n^{-1}\ell_{\nu,n}(\hat{\theta}_{\nu,n}) \geq \sup_{\theta\in\Theta} n^{-1}\ell_{\nu,n}(\theta) - o_{\text{a.s.}}(1),$$

where $o_{\text{a.s.}}(1)$ denotes a stochastic process that converges to zero $\bar{\mathbb{P}}_{\theta^\star}$-a.s. as $n \to \infty$ [if the supremum of $\ell_{\nu,n}$ is attained, we may choose $\hat{\theta}_{\nu,n} = \arg\max_{\theta\in\Theta} \ell_{\nu,n}(\theta)$]. The main result of the paper consists in obtaining the consistency of $\hat{\theta}_{\nu,n}$.

THEOREM 1. *Assume* (A1)–(A6)*, and let* ν *be a fixed initial probability measure. Suppose that one of the following assumptions hold:*

1. $\nu \sim \pi_{\theta^\star}$*; or*
2. $g_{\theta^\star}(x,y) > 0$ *for all* x, y*, and* $Q_{\theta^\star}$ *is aperiodic; or*
3. $g_{\theta^\star}(x,y) > 0$ *for all* x, y*, and* ν *has mass in each periodic class of* $Q_{\theta^\star}$*.*

Then $\hat{\theta}_{\nu,n} \overset{n\to\infty}{\longrightarrow} [\theta^\star]$*,* $\bar{\mathbb{P}}_{\theta^\star}$*-a.s.*

The proof of this theorem is given in Section 4.

REMARK 2. The assumptions 1–3 in Theorem 1 impose different requirements on the initial measure ν used for the maximum likelihood procedure. When the true parameter is aperiodic and has nondegenerate observations, consistency holds for any choice of ν. On the other hand, in the case of degenerate observations, it is evident that we cannot expect consistency to hold in general without imposing an absolute continuity assumption of the form $\nu \sim \pi_{\theta^\star}$. The intermediate case, where the observations are nondegenerate but the signal may be periodic, is not entirely obvious. An illuminating counterexample, which shows that the MLE can be inconsistent for a choice of ν that does not satisfy the requisite assumption in this case, is given in Remark 12 below.

REMARK 3. In Theorem 1, we have assumed that the data is generated by the stationary measure $\bar{\mathbb{P}}_{\theta^\star}$. However, it follows directly from Lemma 7 below that, under the assumptions of Theorem 1, we also have $\hat{\theta}_{\nu,n} \overset{n\to\infty}{\longrightarrow} [\theta^\star]$ $\mathbb{P}^\rho_{\theta^\star}$-a.s. for any initial measure ρ that satisfies the same assumptions as ν in Theorem 1. Hence, the initial measure of the underlying chain is largely irrelevant, both for the consistency of the estimator and in the definition of the log-likelihood function $\ell_{\nu,n}(\theta)$ that is used to compute the estimator.

REMARK 4. The assumptions of Theorem 1 can be weakened somewhat. For example, the σ-finite measure λ can be allowed to depend on θ, or one may consider maximum likelihood estimates of the form $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \ell_{\nu_\theta,n}(\theta)$ where the initial measure ν used to compute the likelihood depends on θ (the latter does not affect the asymptotics of the MLE, but may improve finite sample properties in certain cases). Such generalizations are straightforward and require only minor adjustments in the proofs. In order not to further complicate our notation, we leave these modifications to the reader.

REMARK 5. As was pointed out to us by a referee, assumptions (A2) and (A4) depend on the choice of the observation reference measure μ, even though the maximum likelihood estimator itself is independent of the choice of reference measure. It is therefore possible that the assumptions of Theorem 1 are not satisfied for a given reference measure μ, but that consistency of the MLE can be established nonetheless by making a suitable change of reference measure.

2.3. *Geometric ergodicity implies identifiability.* Most of the assumptions of Theorem 1 can be verified in a straightforward manner. The exception is the identifiability assumption (A6), which appears to be nontrivial. Nonetheless, we will show that this assumption holds in a large class of models: it is already sufficient (beside a mild technical assumption) that the transition kernel Q_θ is *geometrically ergodic*, a property that holds in many applications. Moreover, there is a well-established theory of geometric ergodicity for Markov chains [27] which provides a powerful set of tools to verify this assumption. Consequently, our main theorem is directly applicable in many cases of practical interest.

REMARK 6. Before we state a precise result, it is illuminating to understand the basic idea behind the proof of assumption (A6). Assume that Q_θ is ergodic and that $\bar{\mathbb{P}}^Y_\theta \neq \bar{\mathbb{P}}^Y_{\theta^\star}$. Then there is an $s < \infty$ and a bounded function $h : \mathsf{Y}^{s+1} \to \mathbb{R}$ such that $\bar{\mathbb{E}}_\theta[h(Y_0^s)] = 0$ and $\bar{\mathbb{E}}_{\theta^\star}[h(Y_0^s)] = 1$. Define

$$A_n = \left\{ y_1^n : \frac{1}{n-s} \sum_{i=1}^{n-s} h(y_i^{i+s}) > \frac{1}{2} \right\}$$

for $n > s$. By the ergodic theorem, $\bar{\mathbb{P}}^Y_{\theta^\star}(A_n) \to 1$ and $\tilde{\mathbb{P}}^\lambda_\theta(A_n) \to 0$ as $n \to \infty$. To prove (A6), one must show that the convergence $\tilde{\mathbb{P}}^\lambda_\theta(A_n) \to 0$ happens at an exponential rate, that is, one must establish a type of large deviations property. Therefore, the key thing to prove is that geometrically ergodic Markov chains possess such a large deviations property. This will be done in Section 5.

Let us begin by recalling the appropriate notion of geometric ergodicity (the definition of the norm $\|\cdot\|_V$ was given in Section 2.1 above).

DEFINITION 1. Let $V_\theta : \mathsf{X} \to [1, \infty)$ be given. The transition kernel Q_θ is called V_θ-*uniformly ergodic* if it possesses an invariant probability measure π_θ and

$$\|Q^m_\theta(x, \cdot) - \pi_\theta\|_{V_\theta} \le R_\theta \alpha_\theta^{-m} V_\theta(x) \qquad \text{for every } x \in \mathsf{X}, m \in \mathbb{N},$$

for some constants $R_\theta < \infty$ and $\alpha_\theta > 1$.

For equivalent definitions and extensive discussion, see [27], Chapter 16. We can now formulate a practical sufficient condition for assumption (A6).

(A6′) For every $\theta \not\sim \theta^\star$ such that $p^\lambda(Y_0^{r_\theta}; \theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s., there exists a function $V_\theta \ge 1$ such that Q_θ is V_θ-uniformly ergodic, $\bar{\mathbb{P}}^Y_\theta \ne \bar{\mathbb{P}}^Y_{\theta^\star}$, and

$$\bar{\mathbb{P}}_{\theta^\star}\left(\int V_\theta(x_{r_\theta+1}) p^\lambda(dx_{r_\theta+1}, Y_0^{r_\theta}; \theta) < \infty\right) > 0. \tag{2}$$

Note, in particular, that (2) holds if (A4) holds and $|V_\theta|_\infty < \infty$ [in this case, (A6′) implies that the transition kernel Q_θ is *uniformly ergodic*]. In the setting where (A6′) holds, it is most natural to consider the equivalence relation $\sim$ defined by setting $\theta \sim \theta'$ if and only if $\bar{\mathbb{P}}^Y_\theta = \bar{\mathbb{P}}^Y_{\theta'}$ (i.e., two parameters are equivalent precisely when they give rise to the same stationary observation laws).

THEOREM 2. *Assume* (A1), (A4) *and* (A6′). *Then* (A6) *holds.*

The proof of this theorem is given in Section 5.1.

A different sufficient condition for assumption (A6), which does not rely on geometric ergodicity of the underlying model, is the following assumption (A6″). We will use this assumption in Section 3.2 to show that when X is finite set, the identifiability assumption holds even for nonergodic signals.

(A6″) For every $\theta \not\sim \theta^\star$ and initial probability measure ν, we have

$$\liminf_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^n \in A_n) > 0, \qquad \limsup_{n\to\infty} n^{-1} \log \mathbb{P}^\nu_\theta(Y_0^n \in A_n) < 0$$

for some sequence of sets $A_n \in \mathcal{Y}^{\otimes(n+1)}$.

PROPOSITION 3. *Assume* (A4) *and* (A6″). *Then* (A6) *holds.*

The proof of this proposition is given in Section 5.1.

3. Examples. In this section, we develop three classes of examples. In Section 3.1 we consider linear Gaussian state space models. In Section 3.2, we consider the classic case where the signal state space is a finite set. Finally, in Section 3.3, we develop a general class of nonlinear state space models. In all these examples, we will find that the assumptions of Theorem 1 are satisfied in a rather general setting.

3.1. *Gaussian linear state space models.* Gaussian linear state space models form an important class of HMM. In this setting, let $\mathsf{X} = \mathbb{R}^d$ and $\mathsf{Y} = \mathbb{R}^p$ for some integers d, p, and let Θ be a compact parameter space. The transition kernel T_θ of the model is specified by the state space dynamics

$$X_{k+1} = A_\theta X_k + R_\theta U_k, \tag{3}$$

$$Y_k = B_\theta X_k + S_\theta V_k, \tag{4}$$

where $\{(U_k, V_k)\}_{k\geq 0}$ is an i.i.d. sequence of Gaussian vectors with zero mean and identity covariance matrix, independent of X_0. Here U_k is q-dimensional, V_k is p-dimensional, and the matrices $A_\theta, R_\theta, B_\theta, S_\theta$ have the appropriate dimensions.

For each $\theta \in \Theta$ and any integer $r \geq 1$, define

$$\mathcal{O}_{\theta,r} \stackrel{\text{def}}{=} \begin{bmatrix} B_\theta \\ B_\theta A_\theta \\ B_\theta A_\theta^2 \\ \vdots \\ B_\theta A_\theta^{r-1} \end{bmatrix} \quad \text{and} \quad \mathcal{C}_{\theta,r} \stackrel{\text{def}}{=} [\, R_\theta A_\theta R_\theta \cdots A_\theta^{r-1} R_\theta \,].$$

It is assumed in the sequel that for any $\theta \in \Theta$, the following hold:

(L1) The pair $[A_\theta, B_\theta]$ is observable and the pair $[A_\theta, R_\theta]$ is controllable, that is, the observability matrix $\mathcal{O}_{\theta,d}$ and controllability matrix $\mathcal{C}_{\theta,d}$ are full rank.

(L2) The state transition matrix A_θ is discrete-time Hurwitz, that is, its eigenvalues all lie in the open unit disc in $\mathbb{C}$.

(L3) The measurement noise covariance matrix S_θ is full rank.

(L4) The functions $\theta \mapsto A_\theta$, $\theta \mapsto R_\theta$, $\theta \mapsto B_\theta$ and $\theta \mapsto S_\theta$ are continuous on Θ.

We show below that the Markov kernel Q_θ is ergodic for every $\theta \in \Theta$. We can therefore define without ambiguity the equivalence relation $\sim$ on Θ as

follows: $\theta \sim \theta'$ iff $\bar{\mathbb{P}}_\theta^Y = \bar{\mathbb{P}}_{\theta'}^Y$. We now proceed to verify the assumptions of Theorem 1.

The fact that A_θ is Hurwitz guarantees that the state equation is stable. Together with the controllability assumption, this implies that Q_θ is V_θ-uniformly ergodic with $V_\theta(x) \asymp |x|^2$ as $|x| \to \infty$ ([15], pages 929 and 930). In particular, $Q_{\theta^\star}$ is $V_{\theta^\star}$-uniformly ergodic, which implies (A1).

By the assumption that S_θ is full rank, and choosing the reference measure μ to be the Lebesgue measure on Y, we find that $g_\theta(x,y)$ is a Gaussian density for each $x \in \mathsf{X}$ with covariance matrix $S_\theta S_\theta^T$. We therefore have $|g_{\theta^\star}|_\infty = (2\pi)^{-p/2} \det{}^{-1/2}(S_{\theta^\star} S_{\theta^\star}^T) < \infty$, so that $\bar{\mathbb{E}}_{\theta^\star}[\sup_{x\in\mathsf{X}}(\log g_{\theta^\star}(x, Y_0))^+] < \infty$. On the other hand, as the stationary distribution π_θ is Gaussian, the function $y \mapsto \int g_{\theta^\star}(x,y)\pi_{\theta^\star}(dx)$ is a Gaussian density with respect to μ. Therefore, is easily seen that $\bar{\mathbb{E}}_{\theta^\star}[|\log \int g_{\theta^\star}(x, Y_0)\pi_{\theta^\star}(dx)|] < \infty$, and we have established (A2).

The dimension q of the state noise vector U_k is in many situations smaller than the dimension d of the state vector X_k and hence $R_\theta R_\theta^T$ may be rank deficient. However, note that $Q_\theta^d(x, dx')$ is a Gaussian distribution with covariance matrix $\mathcal{C}_{\theta,d}\mathcal{C}_{\theta,d}^T$ for each $x \in \mathsf{X}$. Therefore, the controllability of the pair $[A_\theta, R_\theta]$ nonetheless guarantees that $Q_\theta^d(x, dx')$ has a density with respect to the Lebesgue measure λ on X. Thus, (A3) is satisfied with $l = d$.

To proceed, we obtain an explicit expression for $p^\lambda(y_0^r; \theta)$.

LEMMA 4. *For $r \geq d$, we have*

$$p^\lambda(y_0^{r-1};\theta) = (2\pi)^{(d-pr)/2} \det{}^{-1/2}(\mathcal{O}_{\theta,r}^T \Gamma_{\theta,r}^{-1} \mathcal{O}_{\theta,r}) \det{}^{-1/2}(\Gamma_{\theta,r}) \times \exp(-\tfrac{1}{2}\mathbf{y}_r^T H_{\theta,r} \mathbf{y}_r). \tag{5}$$

Here we defined the matrix $\Gamma_{\theta,r} \stackrel{def}{=} \mathcal{H}_{\theta,r}\mathcal{H}_{\theta,r}^T + \mathcal{S}_{\theta,r}\mathcal{S}_{\theta,r}^T$ with

$$\mathcal{H}_{\theta,r} \stackrel{def}{=} \begin{pmatrix} 0 & 0 & \cdots & 0 \\ B_\theta R_\theta & 0 & & 0 \\ B_\theta A_\theta R_\theta & B_\theta R_\theta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ B_\theta A_\theta^{r-2} R_\theta & B_\theta A_\theta^{r-3} R_\theta & \cdots & B_\theta R_\theta \end{pmatrix}$$

and where $\mathcal{S}_{\theta,r}$ is the $pr \times pr$ block diagonal matrix with diagonal blocks equal to S_θ, $\mathbf{y}_r = [y_0, \ldots, y_{r-1}]^T$, and $H_{\theta,r}$ is the matrix defined by

$$H_{\theta,r} \stackrel{def}{=} \Gamma_{\theta,r}^{-1} - \Gamma_{\theta,r}^{-1}\mathcal{O}_{\theta,r}(\mathcal{O}_{\theta,r}^T\Gamma_{\theta,r}^{-1}\mathcal{O}_{\theta,r})^{-1}\mathcal{O}_{\theta,r}^T\Gamma_{\theta,r}^{-1}.$$

PROOF. Define the vectors $\mathbf{Y}_r = [Y_0^T, \ldots, Y_{r-1}^T]^T$, $\mathbf{U}_{r-1} = [U_0^T, \ldots, U_{r-2}^T]^T$ and $\mathbf{V}_r = [V_0^T, \ldots, V_{r-1}^T]^T$. It follows from elementary algebra that

$$\mathbf{Y}_r = \mathcal{O}_{\theta,r} X_0 + \mathcal{H}_{\theta,r}\mathbf{U}_{r-1} + \mathcal{S}_{\theta,r}\mathbf{V}_r$$

for any integer $r \geq 1$. Note that, as $\mathbf{U}_{r-1}$ and $\mathbf{V}_r$ are independent, the covariance matrix of the vector $\mathcal{H}_{\theta,r}\mathbf{U}_{r-1} + \mathcal{S}_{\theta,r}\mathbf{V}_r$ is given by $\Gamma_{\theta,r}$. It follows that

$$p^x(y_0^{r-1};\theta) = (2\pi)^{-pr/2} \det{}^{-1/2}(\Gamma_{\theta,r}) \exp(-\tfrac{1}{2}(\mathbf{y}_r - \mathcal{O}_{\theta,r}x)^T \Gamma_{\theta,r}^{-1}(\mathbf{y}_r - \mathcal{O}_{\theta,r}x)),$$

where we have used that $\Gamma_{\theta,r}$ is positive definite (this follows directly from the assumption that S_θ is full rank). Now let $\tilde{\Pi}_{\theta,r} \stackrel{\text{def}}{=} \tilde{\mathcal{O}}_{\theta,r}(\tilde{\mathcal{O}}_{\theta,r}^T \tilde{\mathcal{O}}_{\theta,r})^{-1}\tilde{\mathcal{O}}_{\theta,r}^T$ be the orthogonal projector on the range of $\tilde{\mathcal{O}}_{\theta,r} \stackrel{\text{def}}{=} \Gamma_{\theta,r}^{-1/2}\mathcal{O}_{\theta,r}$ ($\tilde{\Pi}_{\theta,r}$ is well defined for $r \geq d$ as the pair $[A_\theta, B_\theta]$ is observable, so that $\tilde{\mathcal{O}}_{\theta,r}$ is full rank). Clearly,

$$\begin{aligned}(\mathbf{y}_r - \mathcal{O}_{\theta,r}x)^T \Gamma_{\theta,r}^{-1}(\mathbf{y}_r - \mathcal{O}_{\theta,r}x) &= \|\tilde{\Pi}_{\theta,r}\Gamma_{\theta,r}^{-1/2}\mathbf{y}_r - \Gamma_{\theta,r}^{-1/2}\mathcal{O}_{\theta,r}x\|^2 \\ &\quad + \|(1-\tilde{\Pi}_{\theta,r})\Gamma_{\theta,r}^{-1/2}\mathbf{y}_r\|^2.\end{aligned}$$

The result now follows from

$$\int \exp\left(-\frac{1}{2}\|\tilde{\Pi}_{\theta,r}\Gamma_{\theta,r}^{-1/2}\mathbf{y}_r - \Gamma_{\theta,r}^{-1/2}\mathcal{O}_{\theta,r}x\|^2\right) dx = (2\pi)^{d/2} \det{}^{-1/2}(\mathcal{O}_{\theta,r}^T \Gamma_{\theta,r}^{-1}\mathcal{O}_{\theta,r})$$

(which is immediately seen to be finite due to the fact that $\tilde{\mathcal{O}}_{\theta,r}$ has full rank), and from the identity $H_{\theta,r} = \Gamma_{\theta,r}^{-1/2}(1-\tilde{\Pi}_{\theta,r})\Gamma_{\theta,r}^{-1/2}$. $\square$

REMARK 7. As is evident from the proof, the observability assumption is key in order to guarantee that $p^\lambda(y_0^{r-1};\theta)$ is finite (albeit only for r sufficiently large). Intuitively, observability guarantees that we can estimate X_0 from Y_0^{d-1} "in every direction," so that the likelihood $p^x(y_0^{r-1};\theta)$ becomes small as $|x| \to \infty$. This is needed in order to ensure that $p^x(y_0^{r-1};\theta)$ is integrable with respect to the σ-finite measure λ. It should also be noted that for any $r \geq d$ the matrix $H_{\theta,r}$ is rank-deficient, showing that (5) is not the density of a finite measure.

Now note that, by our assumptions, the functions $\theta \mapsto \det{}^{-1/2}(\mathcal{O}_{\theta,d}^T \Gamma_{\theta,d}^{-1}\mathcal{O}_{\theta,d})$, $\theta \mapsto \det{}^{-1/2}(\Gamma_{\theta,d})$, and $\theta \mapsto H_{\theta,r}$ are continuous on Θ for any $r \geq d$. Thus, $\theta \mapsto p^\lambda(y_0^{r-1};\theta)$ is continuous for every $r \geq d$, and it is easily established that $\bar{\mathbb{E}}_{\theta^\star}[\sup_{\theta' \in \mathcal{U}_\theta}(\log p^\lambda(Y_0^{r_\theta};\theta'))^+] < \infty$ if we choose $r_\theta = d-1$ and a sufficiently small neighborhood $\mathcal{U}_\theta$. Moreover, note that $|g_\theta|_\infty = (2\pi)^{-p/2} \det{}^{-1/2}(S_\theta S_\theta^T)$ and $|q_\theta|_\infty = (2\pi)^{-d/2} \det{}^{-1/2}(\mathcal{C}_{\theta,d}\mathcal{C}_{\theta,d}^T)$. Therefore, by the continuity of S_θ and $\mathcal{C}_{\theta,d}$, we have $\sup_{\theta' \in \mathcal{U}_\theta} |q_{\theta'}|_\infty < \infty$ and $\bar{\mathbb{E}}_{\theta^\star}[\sup_{\theta' \in \mathcal{U}_\theta} \sup_{x \in \mathsf{X}}(\log g_{\theta'}(x, Y_0))^+] < \infty$ for a sufficiently small neighborhood $\mathcal{U}_\theta$. Thus, we have verified (A4) and (A5).

It remains to establish assumption (A6). We established above that Q_θ is V_θ-uniformly ergodic with $V_\theta(x) \asymp |x|^2$ as $|x| \to \infty$. Moreover, $\theta \not\sim \theta^\star$ implies $\bar{\mathbb{P}}_\theta^Y \neq \bar{\mathbb{P}}_{\theta^\star}^Y$ by definition. Therefore, (A6′) would be established if

$$\int |x_d|^2 p^\lambda(dx_d, Y_0^{d-1}; \theta) < \infty, \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.}$$

But note that

$$\int p^\lambda(dx_d, Y_0^{d-1}; \theta) = p^\lambda(Y_0^{d-1}; \theta) < \infty, \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.},$$

so that $p^\lambda(dx_d, Y_0^{d-1}; \theta)$ is a finite measure. Moreover, as $(Y_0^{d-1}, X_d) = MX_0 + \xi$ for a matrix M and a Gaussian vector ξ, it is easily seen that $p^\lambda(dx_d, Y_0^{d-1}; \theta)$ must be a random Gaussian measure. As Gaussian measures have finite moments, we have established (A6′). Therefore, (A6) follows from Theorem 2.

Having verified (A1)–(A6), we can apply Theorem 1. As $g_{\theta^\star}(x, y) > 0$ for all x, y, and as $Q_{\theta^\star}$ is $V_{\theta^\star}$-uniformly ergodic (hence certainly aperiodic), we find that the MLE is consistent for any initial measure ν.

3.2. *Finite state models.* One of the most widely used classes of HMM is obtained when the signal is a finite state Markov chain. In this setting, let $\mathsf{X} = \{1, \ldots, d\}$ for some integer d, let Y be any Polish space, and let Θ be a compact metric space. For each parameter $\theta \in \Theta$, the signal transition kernel Q_θ is determined by the corresponding transition probability matrix $\mathcal{Q}_\theta$, while the observation density g_θ is given as in the general setting of this paper.

It is assumed in the sequel that:

(F1) The stochastic matrix $\mathcal{Q}_{\theta^\star}$ is irreducible.
(F2) $\bar{\mathbb{E}}_{\theta^\star}[|\log g_{\theta^\star}(x, Y_0)|] < \infty$ for every $x \in \mathsf{X}$.
(F3) For every $\theta \in \Theta$, there is a neighborhood $\mathcal{U}_\theta$ of θ such that

$$\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta' \in \mathcal{U}_\theta} (\log g_{\theta'}(x, Y_0))^+\Big] < \infty \qquad \text{for all } x \in \mathsf{X}.$$

(F4) $\theta \mapsto \mathcal{Q}_\theta$ and $\theta \mapsto g_\theta(x, y)$ are continuous for any $x \in \mathsf{X}$, $y \in \mathsf{Y}$.

Following [23], we introduce the equivalence relation on Θ as follows: we write $\theta \sim \theta'$ iff there exist invariant distributions π, π' for $\mathcal{Q}_\theta, \mathcal{Q}_{\theta'}$, respectively, such that $\mathbb{P}_\theta^{\pi, Y} = \mathbb{P}_{\theta'}^{\pi', Y}$. In words, two parameters are equivalent whenever they give rise to the same stationary observation laws for some choice of invariant measures for the underlying signal process. The latter statement is not vacuous as we have not required that $\mathcal{Q}_\theta$ is ergodic for $\theta \neq \theta^\star$, that is, there may be multiple invariant measures for $\mathcal{Q}_\theta$. The possibility that $\mathcal{Q}_\theta$ is not aperiodic or even ergodic is the chief complication in this example, as

the easily verified V-uniform ergodicity assumption (A6′) need not hold. We will show nonetheless that assumption (A6″) is satisfied, so that Theorem 1 can be applied.

LEMMA 5. *Let $C \subseteq \mathsf{X}$ be an ergodic class of $\mathcal{Q}_\theta$, and denote by π_C the unique $\mathcal{Q}_\theta$-invariant measure supported in C. Fix $s \geq 0$, and let $f : \mathsf{Y}^{s+1} \to \mathbb{R}$ be such that $|f|_\infty < \infty$. Then there exists a constant K such that*

$$\mathbb{P}_\theta^\nu\left(\left|\sum_{i=1}^n \{f(Y_i^{i+s}) - \mathbb{P}_\theta^{\pi_C}[f(Y_0^s)]\}\right| \geq t\right) \leq K \exp\left[-\frac{t^2}{Kn}\right]$$

for any probability measure ν supported in C and any $t > 0$, $n \geq 1$.

PROOF. The proof is identical to that of Theorem 14, provided we replace the application of Theorem 17 by a trivial modification of the result of [16]. □

REMARK 8. As stated, the result of [16] would require that the restriction of $\mathcal{Q}_\theta$ to C is aperiodic. However, aperiodicity is only used in the proof to ensure the existence of a solution to the Poisson equation, and it is well known that the latter holds also in the periodic case. Therefore, a trivial modification of the proof in [16] allows us to apply the result without additional assumptions.

LEMMA 6. *In the present setting, assumption* (A6″) *holds.*

PROOF. Let $\theta \not\sim \theta^\star$. We can partition $\mathsf{X} = E_1 \cup \cdots \cup E_p \cup T$ into the $p \leq d$ ergodic classes $E_1, \ldots, E_p$ and the set of transient states T of the stochastic matrix $\mathcal{Q}_\theta$. Denote as π_θ^i the unique invariant measure of $\mathcal{Q}_\theta$ that is supported in E_i. Then we can find an integer $s \geq 1$ and bounded function $h : \mathsf{Y}^{s+1} \to \mathbb{R}$ such that $\mathbb{E}_\theta^{\pi_\theta^i}[h(Y_0^s)] \leq 0$ for all $i = 1, \ldots, p$ and such that $\bar{\mathbb{E}}_{\theta^\star}[h(Y_0^s)] = 1$.

Define for $n > 2s$ the set $A_n \in \mathcal{Y}^{\otimes(n+1)}$ as

$$A_n \stackrel{\text{def}}{=} \left\{ y_0^n \in \mathsf{Y}^{n+1} : \frac{1}{\lfloor n/2 \rfloor - s} \sum_{i=\lceil n/2 \rceil + 1}^{n-s} h(y_i^{i+s}) \geq \frac{1}{2} \right\}.$$

As Y_0^∞ is stationary and ergodic under $\bar{\mathbb{P}}_{\theta^\star}$ (because $\mathcal{Q}_{\theta^\star}$ is irreducible), we have

$$\lim_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^n \in A_n) = \lim_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}\left[\frac{1}{\lfloor n/2 \rfloor - s} \sum_{i=1}^{\lfloor n/2 \rfloor - s} h(Y_i^{i+s}) \geq \frac{1}{2}\right] = 1$$

by Birkhoff's ergodic theorem. On the other hand, for any initial probability measure ν, we can estimate as follows: for some constant $K > 0$,

$$\begin{aligned}\mathbb{P}_\theta^\nu(Y_0^n \in A_n) &= \mathbb{P}_\theta^\nu(Y_0^n \in A_n, X_{\lceil n/2 \rceil} \in T) \\ &\quad + \sum_{j=1}^{p} \mathbb{P}_\theta^\nu(Y_0^n \in A_n | X_{\lceil n/2 \rceil} \in E_j) \mathbb{P}_\theta^\nu(X_{\lceil n/2 \rceil} \in E_j) \\ &\le \mathbb{P}_\theta^\nu(X_{\lceil n/2 \rceil} \in T) \\ &\quad + \max_{j=1,\ldots,p} \sup_{\operatorname{supp} \mu \subseteq E_j} \mathbb{P}_\theta^\mu \left[\frac{1}{\lfloor n/2 \rfloor - s} \sum_{i=1}^{\lfloor n/2 \rfloor - s} h(Y_i^{i+s}) \ge \frac{1}{2} \right] \\ &\le K \exp\left[-\frac{n}{K} \right].\end{aligned}$$

The latter inequality follows from the fact that the population in the transient states decays exponentially, while we may apply Lemma 5 to obtain an exponential bound for every ergodic class E_i. We therefore find that

$$\limsup_{n\to\infty} n^{-1} \log \mathbb{P}_\theta^\nu(Y_0^n \in A_n) \le -\frac{1}{K} < 0,$$

completing the proof of assumption (A6″). □

Let us now check the assumptions of Theorem 1. (A1) follows directly from the assumption that $\mathcal{Q}_{\theta^\star}$ is irreducible. To establish (A2), note that

$$\bar{\mathbb{E}}_{\theta^\star}\left[\sup_{x \in \mathsf{X}} (\log g_{\theta^\star}(x, Y_0))^+ \right] \le \sum_{x \in \mathsf{X}} \bar{\mathbb{E}}_{\theta^\star}[|\log g_{\theta^\star}(x, Y_0)|] < \infty,$$

while we can estimate

$$\begin{aligned}&\bar{\mathbb{E}}_{\theta^\star}\left[\left| \log \int g_{\theta^\star}(x, Y_0) \pi_{\theta^\star}(dx) \right| \right] \\ &\quad \le \bar{\mathbb{E}}_{\theta^\star}\left[\sup_{x \in \mathsf{X}} (\log g_{\theta^\star}(x, Y_0))^+ \right] + \bar{\mathbb{E}}_{\theta^\star}\left[\sup_{x \in \mathsf{X}} (\log g_{\theta^\star}(x, Y_0))^- \right] \\ &\quad \le \sum_{x \in \mathsf{X}} \bar{\mathbb{E}}_{\theta^\star}[|\log g_{\theta^\star}(x, Y_0)|] < \infty.\end{aligned}$$

Assumption (A3) holds trivially for $l = 1$ and with λ the counting measure on X [note that $|q_\theta|_\infty \le 1$ for all θ, as $q_\theta(x, x')$ is simply the transition probability from x to x']. To establish (A4), note that $\sup_{\theta \in \Theta} |q_\theta|_\infty < \infty$, while

$$\bar{\mathbb{E}}_{\theta^\star}\left[\sup_{\theta' \in \mathcal{U}_\theta} \sup_{x \in \mathsf{X}} (\log g_{\theta'}(x, Y_0))^+ \right] \le \sum_{x \in \mathsf{X}} \bar{\mathbb{E}}_{\theta^\star}\left[\sup_{\theta' \in \mathcal{U}_\theta} (\log g_{\theta'}(x, Y_0))^+ \right] < \infty$$

by our assumptions. Moreover, as

$$\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in\mathcal{U}_\theta}(\log p^\lambda(Y_0^0;\theta'))^+\Big] \le \bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in\mathcal{U}_\theta}\sup_{x\in\mathsf{X}}(\log d + \log g_{\theta'}(x,Y_0))^+\Big] < \infty,$$

we have shown that (A4) holds with $r_\theta = 0$ for all θ. Next, we note that the continuity of $\theta \mapsto \mathcal{Q}_\theta$ and $\theta \mapsto g_\theta(x,y)$ yield immediately that $\theta \mapsto p^\lambda(y_0^n;\theta)$ is a continuous function for every $n \ge 0$ and $y_0^n \in \mathsf{Y}^{n+1}$, establishing (A5). Finally, Lemma 6 and Proposition 3 establish (A6).

Having verified (A1)–(A6), we can apply Theorem 1. Note that as $\mathcal{Q}_{\theta^\star}$ is irreducible, $\pi_{\theta^\star}$ charges every point of X. Therefore, by Theorem 1, the MLE is consistent provided that ν charges every point of X (so that $\nu \sim \pi_{\theta^\star}$).

REMARK 9. The result obtained in this section as a special case of Theorem 1 is almost identical to the result of Leroux [23]. The main difference in [23] is that there the parameter space Θ may be noncompact, provided the parametrization of the model vanishes at infinity. This setting reduces directly to the compact case by compactifying the parameter space Θ, so that this does not constitute a major generalization from the technical point of view.

However, it should be noted that one cannot immediately apply Theorem 1 to the compactified model. The problem is that the new parameters "at infinity" are typically sub-probabilities rather than true probability measures, while we have assumed in this paper that every parameter $\theta \in \Theta$ corresponds to a probability measure on the space of observation paths. Theorem 1 can certainly be generalized to allow for sub-probabilities without significant technical complications. We have chosen to concentrate on the compact setting, however, in order to keep the notation and results of the paper as clean as possible.

3.3. *Nonlinear state space models.* In this section, we consider a class of nonlinear state space models. Let $\mathsf{X} = \mathbb{R}^d$, $\mathsf{Y} = \mathbb{R}^\ell$, and let Θ be a compact metric space. For each $\theta \in \Theta$, the Markov kernel Q_θ of the hidden process $(X_k)_{k\ge0}$ is defined through the nonlinear recursion

$$X_k = G_\theta(X_{k-1}) + \Sigma_\theta(X_{k-1})\zeta_k.$$

Here $(\zeta_k)_{k\ge1}$ is an i.i.d. sequence of d-dimensional random vectors which are assumed to possess a density ρ_ζ with respect to the Lebesgue measure λ on $\mathbb{R}^d$, and $G_\theta : \mathbb{R}^d \to \mathbb{R}^d$, $\Sigma_\theta : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are given (measurable) functions. The model for the hidden chain $(X_k)_{k\ge0}$ is sometimes known as a vector ARCH model, and covers many models of interest in time series analysis and financial econometrics (including the AR model, the ARCH model, threshold ARCH, etc.). We let the reference measure μ be the Lebesgue measure on

$\mathbb{R}^\ell$, and define the observed process $(Y_k)_{k\geq 0}$ by means of a given observation density $g_\theta(x,y)$.

For any positive matrix B, denote by $\lambda_{\min}(B)$ its minimal eigenvalue. For any bounded set $\mathcal{A}\subset\mathbb{R}^{d\times d}$, define $\mathcal{A}_m \stackrel{\text{def}}{=} \{A_1A_2\cdots A_m : A_i\in\mathcal{A}, i=1,\ldots,m\}$. Denote by $\rho(\mathcal{A})$ the joint spectral radius of the set of matrices $\mathcal{A}$, defined as

$$\rho(\mathcal{A}) \stackrel{\text{def}}{=} \limsup_{m\to\infty}\Big(\sup_{A\in\mathcal{A}_m}\|A\|\Big)^{1/m}.$$

Here $\|\cdot\|$ is any matrix norm [it is elementary that $\rho(\mathcal{A})$ does not depend on the choice of the norm]. We now introduce the basic assumptions of this section.

(NL1) The random variables ζ_k have mean zero and identity covariance matrix. Moreover, $\rho_\zeta(x)>0$ for all $x\in\mathbb{R}^d$, and $|\rho_\zeta|_\infty<\infty$.

(NL2) For each $\theta\in\Theta$, the function Σ_θ is bounded on compact sets, $\Sigma_\theta(x)=o(|x|)$ as $|x|\to\infty$, and $0<\inf_{\theta'\in\mathcal{U}_\theta}\inf_{x\in\mathbb{R}^d}\lambda_{\min}[\Sigma_{\theta'}(x)\Sigma_{\theta'}^T(x)]$ for a sufficiently small neighborhood $\mathcal{U}_\theta$ of θ.

(NL3) For each $\theta\in\Theta$, the drift function G_θ has the form

$$G_\theta(x)=A_\theta(x)x+h_\theta(x)$$

for some measurable functions $A_\theta:\mathbb{R}^d\to\mathbb{R}^{d\times d}$ and $h_\theta:\mathbb{R}^d\to\mathbb{R}^d$. Moreover, we assume that G_θ is bounded on compact sets, $h_\theta(x)=o(|x|)$ as $|x|\to\infty$, and that there exists $R_\theta>0$ such that the set of matrices $\mathcal{A}_\theta \stackrel{\text{def}}{=} \{A_\theta(x): x\in\mathbb{R}^d, |x|\geq R_\theta\}$ is bounded and $\rho(\mathcal{A}_\theta)<1$.

(NL4) For each $\theta\in\Theta$, there is a neighborhood $\mathcal{U}_\theta$ of θ such that

$$\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in\mathcal{U}_\theta}\sup_{x\in\mathsf{X}}(\log g_{\theta'}(x,Y_0))^+\Big]<\infty,$$

$$\bar{\mathbb{E}}_{\theta^\star}\left[\sup_{\theta'\in\mathcal{U}_\theta}\left(\log\int g_{\theta'}(x,Y_0)\lambda(dx)\right)^+\right]<\infty.$$

Moreover, $\bar{\mathbb{P}}_{\theta^\star}(\int|x|g_\theta(x,Y_0)\lambda(dx)<\infty)>0$ for each $\theta\in\Theta$, and

$$\bar{\mathbb{E}}_{\theta^\star}\left[\int(\log g_{\theta^\star}(x,Y_0))^-\pi_{\theta^\star}(dx)\right]<\infty.$$

(NL5) The functions $\theta\mapsto g_\theta(x,y)$, $\theta\mapsto G_\theta(x)$, $\theta\mapsto\Sigma_\theta(x)$ and $x\mapsto\rho_\zeta(x)$ are continuous on Θ for every x,y. Moreover, for each $\theta\in\Theta$, the function $\theta'\mapsto\int g_{\theta'}(x,Y_0)\lambda(dx)$ is positive and continuous at θ, $\bar{\mathbb{P}}_{\theta^\star}$-a.s.

REMARK 10. We have made no attempt at generality here: for the sake of example, we have chosen a set of conditions under which the assumptions

of Theorem 1 are easily verified. Of course, the applicability of Theorem 1 extends far beyond the simple assumptions imposed in this section.

Nonetheless, even the present assumptions already cover a broad class of nonlinear models. Consider, for example, the stochastic volatility model [17]

$$\begin{cases} X_{k+1} = \phi_\theta X_k + \sigma_\theta \zeta_k, \\ Y_k = \beta_\theta \exp(X_k/2)\varepsilon_k, \end{cases} \tag{6}$$

where (ζ_k, ε_k) are i.i.d. Gaussian random variables in $\mathbb{R}^2$ with zero mean and identity covariance matrix, $\beta_\theta > 0$, $\sigma_\theta > 0$ and $|\phi_\theta| < 1$ for every $\theta \in \Theta$, and the functions $\theta \mapsto \phi_\theta$, $\theta \mapsto \sigma_\theta$ and $\theta \mapsto \beta_\theta$ are continuous. Assumptions (NL1)–(NL3) are readily seen to hold. The observation likelihood g_θ is given by

$$g_\theta(x, y) = (2\pi\beta_\theta^2)^{-1/2} \exp[-\exp(-x)y^2/2\beta_\theta^2 - x/2].$$

We can compute

$$\sup_{x \in \mathsf{X}} g_\theta(x, y) = \frac{1}{\sqrt{2\pi e}} \frac{1}{|y|}, \qquad \int g_\theta(x, y)\lambda(dx) = \frac{1}{|y|}.$$

As the stationary distribution $\pi_{\theta^\star}$ is Gaussian, it is easily seen that the law of Y_0 under $\bar{\mathbb{P}}_{\theta^\star}$ has a bounded density with respect to the Lebesgue measure μ on Y. As $\int (\log(1/|y|))^+ \mu(dy) < \infty$, the first equation display of (NL4) follows. To prove that $\bar{\mathbb{P}}_{\theta^\star}(\int |x| g_\theta(x, Y_0)\lambda(dx) < \infty) > 0$, it suffices to note that $x \mapsto g_\theta(x, y)$ has exponentially decaying tails for all $|y| > 0$. The remaining part of (NL4) follows easily using that $\pi_{\theta^\star}$ is Gaussian and $\bar{\mathbb{E}}_{\theta^\star}(Y_0^2) < \infty$. Finally, (NL5) now follows immediately, and we have verified that the assumptions of this section hold for the stochastic volatility model. Similar considerations apply in a variety of nonlinear models commonly used in financial econometrics.

We show below that the Markov kernel Q_θ is ergodic for every $\theta \in \Theta$. We can therefore define without ambiguity the equivalence relation $\sim$ on Θ as follows: $\theta \sim \theta'$ iff $\bar{\mathbb{P}}_\theta^Y = \bar{\mathbb{P}}_{\theta'}^Y$. We now proceed to verify the assumptions of Theorem 1.

It is shown in [24], Theorem 2, that under conditions (NL1)–(NL3), the Markov kernel Q_θ is V-uniformly ergodic for each $\theta \in \Theta$ with $V(x) = 1 + |x|$. In particular, assumption (A1) holds. The first part of (A2) follows directly from (NL4). To prove the second part, we first note that Q_θ has a transition density

$$q_\theta(x, x') = |\det[\Sigma_\theta(x)]|^{-1} \rho_\zeta(\Sigma_\theta^{-1}(x)\{x' - G_\theta(x)\})$$

with respect to the Lebesgue measure λ on X. This evidently gives

$$|q_\theta|_\infty = \sup_{x \in \mathsf{X}} |\det[\Sigma_\theta(x)]|^{-1} |\rho_\zeta|_\infty < \infty$$

by (NL1) and (NL2), which implies in particular that $\pi_{\theta^\star}$ has a bounded density with respect to λ. Therefore

$$\begin{aligned}&\bar{\mathbb{E}}_{\theta^\star}\left[\left(\log\int g_{\theta^\star}(x,Y_0)\pi_{\theta^\star}(dx)\right)^+\right]\\&\quad\le|q_{\theta^\star}|_\infty\bar{\mathbb{E}}_{\theta^\star}\left[\left(\log\int g_{\theta^\star}(x,Y_0)\lambda(dx)\right)^+\right]<\infty\end{aligned}$$

by (NL4). On the other hand, as $x\mapsto(\log x)^-$ is convex, we have

$$\bar{\mathbb{E}}_{\theta^\star}\left[\left(\log\int g_{\theta^\star}(x,Y_0)\pi_{\theta^\star}(dx)\right)^-\right]\le\bar{\mathbb{E}}_{\theta^\star}\left[\int(\log g_{\theta^\star}(x,Y_0))^-\pi_{\theta^\star}(dx)\right]<\infty$$

by Jensen's inequality and (NL4). Therefore, (A2) is established. We have already shown that Q_θ possesses a bounded density, so (A3) holds with $l=1$. Assumption (A4) with $r_\theta=0$ follows directly from (NL4) and (NL1), (NL2).

To establish (A5), let $\nu_\theta(dx,y)\stackrel{\text{def}}{=}g_\theta(x,y)\lambda(dx)/\int g_\theta(x,y)\lambda(dx)$. By (NL5), $\nu_\theta(dx,Y_0)$ is a probability measure $\bar{\mathbb{P}}_{\theta^\star}$-a.s., and for every $\theta\in\Theta$ the density function $\theta'\mapsto g_{\theta'}(x,Y_0)/\int g_{\theta'}(x,Y_0)\lambda(dx)$ is continuous at θ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. By Scheffé's lemma, this implies that for any $\theta\in\Theta$, the map $\theta'\mapsto\nu_{\theta'}(\cdot,Y_0)$ is continuous at θ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. with respect to the total variation norm $\|\cdot\|_1$. Similarly, as $\theta\mapsto q_\theta(x,x')$ is continuous by (NL5), the map $\theta\mapsto Q_\theta(x,dx')$ is continuous with respect to the total variation norm. Now note that we can write

$$p^\lambda(Y_0^n;\theta)=\left(\int g_\theta(x,Y_0)\lambda(dx)\right)\int p^{x'}(Y_1^n;\theta)Q_\theta(x,dx')\nu_\theta(dx,Y_0).$$

From (NL4), it follows that $x\mapsto\sup_{\theta'\in\mathcal{U}_\theta}g_{\theta'}(x,Y_k)$ is bounded $\bar{\mathbb{P}}_{\theta^\star}$-a.s. for every k. Therefore, $x\mapsto\sup_{\theta'\in\mathcal{U}_\theta}p^x(Y_1^n;\theta')$ is a bounded function $\bar{\mathbb{P}}_{\theta^\star}$-a.s., and by dominated convergence the function $\theta'\mapsto p^x(Y_1^n;\theta')$ is continuous at θ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. for every $\theta\in\Theta$. Therefore, it follows that $\bar{\mathbb{P}}_{\theta^\star}$-a.s.

$$\begin{aligned}&\left|\int p^{x'}(Y_1^n;\theta_n)Q_{\theta_n}(x,dx')\nu_{\theta_n}(dx,Y_0)-\int p^{x'}(Y_1^n;\theta)Q_\theta(x,dx')\nu_\theta(dx,Y_0)\right|\\&\quad\le\int|p^{x'}(Y_1^n;\theta_n)-p^{x'}(Y_1^n;\theta)|Q_\theta(x,dx')\nu_\theta(dx,Y_0)\\&\qquad+\sup_{\theta'\in\mathcal{U}_\theta}|p^{\cdot}(Y_1^n;\theta')|_\infty\|\nu_{\theta_n}(\cdot,Y_0)Q_{\theta_n}-\nu_\theta(\cdot,Y_0)Q_\theta\|_1\\&\quad\stackrel{n\to\infty}{\longrightarrow}0\end{aligned}$$

for any sequence $(\theta_n)_{n\ge0}\subset\mathcal{U}_\theta$, $\theta_n\to\theta$. Here we have used the dominated convergence theorem to conclude convergence of the first term, and the

continuity in total variation established above for the second term. (A5) follows.

It remains to establish assumption (A6). We established above that Q_θ is V-uniformly ergodic with $V(x) = 1 + |x|$. Moreover, $\theta \not\sim \theta^\star$ implies $\bar{\mathbb{P}}_\theta^Y \neq \bar{\mathbb{P}}_{\theta^\star}^Y$ by definition. Therefore, (A6′) would be established if

$$\bar{\mathbb{P}}_{\theta^\star}\left(\int |x'| Q_\theta(x, dx') g_\theta(x, Y_0) \lambda(dx) < \infty\right) > 0.$$

But as Q_θ is V-uniformly ergodic, it follows that $Q_\theta V \leq \alpha_\theta V + K_\theta$ for some positive constants α_θ, K_θ ([27], Theorem 16.0.1). Assumption (A6′) therefore follows from (NL4), and (A6) follows from Theorem 2.

Having verified (A1)–(A6), we can apply Theorem 1. As $g_{\theta^\star}(x, y) > 0$ for all x, y, and as $Q_{\theta^\star}$ is V-uniformly ergodic (hence certainly aperiodic), we find that the MLE is consistent for any initial measure ν.

REMARK 11. The assumption in (NL4) that

$$\bar{\mathbb{E}}_{\theta^\star}\left[\int (\log g_{\theta^\star}(x, Y_0))^- \pi_{\theta^\star}(dx)\right] < \infty$$

is used to verify the second part of (A2). This condition can be replaced by the following assumption: there exists a set $D \in \mathcal{X}$ such that:

(i) $\bar{\mathbb{E}}_{\theta^\star}[(\log \int_D g_{\theta^\star}(x, Y_0)\lambda(dx))^-] < \infty$, and
(ii) $\inf_{x,x' \in D} q_{\theta^\star}(x, x') > 0$.

The latter condition is sometimes easier to check.

To see that the result still holds under this modified condition, note that

$$\bar{\mathbb{E}}_{\theta^\star}\left[\left(\log \int g_{\theta^\star}(x, Y_0)\pi_{\theta^\star}(dx)\right)^+\right] < \infty$$

follows as above. On the other hand,

$$\begin{aligned}\int g_{\theta^\star}(x, Y_0)\pi_{\theta^\star}(dx) &\geq \int_{D\times D} g_{\theta^\star}(x', Y_0) q_{\theta^\star}(x, x')\pi_{\theta^\star}(dx)\lambda(dx') \\ &\geq \pi_{\theta^\star}(D) \inf_{x,x' \in D} q_{\theta^\star}(x, x') \int_D g_{\theta^\star}(x, Y_0)\lambda(dx).\end{aligned}$$

It follows from (i) that $\lambda(D) > 0$, so that $\pi_{\theta^\star}(D) > 0$ also (as $Q_{\theta^\star}$, and therefore $\pi_{\theta^\star}$, has a positive density with respect to λ). It now follows directly that also $\bar{\mathbb{E}}_{\theta^\star}[(\log \int g_{\theta^\star}(x, Y_0)\pi_{\theta^\star}(dx))^-] < \infty$, and the claim is established.

4. Proof of Theorem 1. The proof of Theorem 1 consists of three parts. First, we prove pointwise convergence of the log-likelihood under the true parameter $\theta^\star$ (Section 4.1). Next, we establish identifiability of every $\theta \not\sim \theta^\star$ (Section 4.2). Finally, we put everything together to complete the proof of consistency (Section 4.3).

4.1. *Pointwise convergence of the normalized log-likelihood under $\theta^\star$.* The goal of this section is to show that our hidden Markov model possesses a finite entropy rate and that the asymptotic equipartition property holds. We begin with a simple result, which will be used to reduce to the stationary case.

LEMMA 7. *Assume* (A1). *Then $(Y_k)_{k\geq 0}$ is ergodic under $\bar{\mathbb{P}}_{\theta^\star}$. Moreover, if one of the assumptions 1–3 of Theorem 1 hold, then $\mathbb{P}^{\nu,Y}_{\theta^\star} \sim \bar{\mathbb{P}}^{Y}_{\theta^\star}$.*

PROOF. As $Q_{\theta^\star}$ possesses a unique invariant measure by (A1), the kernel $T_{\theta^\star}$ possesses a unique invariant measure also. This implies that the process $(X_k, Y_k)_{k\geq 0}$ is ergodic under the stationary measure $\bar{\mathbb{P}}_{\theta^\star}$ (as the latter is trivially an extreme point of the set of stationary measures). Therefore, $(Y_k)_{k\geq 0}$ is ergodic also.

If $\nu \sim \pi_{\theta^\star}$, it is easily seen that $\mathbb{P}^{\nu,Y}_{\theta^\star} \sim \bar{\mathbb{P}}^{Y}_{\theta^\star}$ [as $d\mathbb{P}^{\nu}_{\theta^\star}/d\bar{\mathbb{P}}_{\theta^\star} = (d\nu/d\pi_{\theta^\star})(X_0)$]. Otherwise, we argue as follows. Suppose that $Q_{\theta^\star}$ has period d [this is guaranteed to hold for some $d \in \mathbb{N}$ by (A1)]. Then we can partition the signal state space as $\mathsf{X} = C_1 \cup \cdots \cup C_d \cup F$, where $C_1, \ldots, C_d$ are the periodic classes and $\pi(F) = 0$ ([27], Section 5.4.3). Note that $C_1, \ldots, C_d$ are absorbing sets for $Q^d_{\theta^\star}$ where the restriction of $Q^d_{\theta^\star}$ to C_i is positive Harris and aperiodic with the corresponding invariant probability measure $\pi^i_{\theta^\star}$. Moreover, the Harris recurrence assumption guarantees that $\mathbb{P}^x_{\theta^\star}(X_n \notin F \text{ eventually}) = 1$ for all $x \in \mathsf{X}$. Therefore, $\nu Q^{nd}_{\theta^\star}(F) \downarrow 0$ and $\nu Q^{nd}_{\theta^\star}(C_i) \uparrow c^i_\nu$ as $n \to \infty$. It follows from the ergodic theorem for aperiodic Harris recurrent Markov chains that

$$\|\nu Q^n_{\theta^\star} - \pi^\nu_{\theta^\star} Q^n_{\theta^\star}\|_1 \overset{n\to\infty}{\longrightarrow}, \qquad \pi^\nu_{\theta^\star} \overset{\text{def}}{=} \sum_{i=1}^{d} c^i_\nu \pi^i_{\theta^\star}.$$

Using $g_{\theta^\star}(x,y) > 0$ and [31], Lemma 3.7, this implies that $\mathbb{P}^{\nu,Y}_{\theta^\star} \sim \mathbb{P}^{\pi^\nu_{\theta^\star},Y}_{\theta^\star}$. But if ν has mass in each periodic class C_i or if $d = 1$, then $c^i_\nu > 0$ for all $i = 1, \ldots, d$. Thus, $\pi^\nu_{\theta^\star} \sim \pi = \frac{1}{d}\sum_{i=1}^d \pi^i_{\theta^\star}$ which implies $\bar{\mathbb{P}}^{Y}_{\theta^\star} \sim \mathbb{P}^{\pi^\nu_{\theta^\star},Y}_{\theta^\star} \sim \mathbb{P}^{\nu,Y}_{\theta^\star}$. □

We will also need the following lemma.

LEMMA 8. *Assume* (A2). *Then $\bar{\mathbb{E}}_{\theta^\star}[|\log \bar{p}(Y_0^n; \theta^\star)|] < \infty$ for all $n \geq 0$.*

PROOF. We easily obtain the upper bound

$$\bar{\mathbb{E}}_{\theta^\star}[(\log \bar{p}(Y_0^n; \theta^\star))^+] \leq \bar{\mathbb{E}}_{\theta^\star}\left[\sum_{k=0}^{n} \sup_{x\in\mathsf{X}} (\log g_{\theta^\star}(x, Y_k))^+\right] < \infty.$$

On the other hand, we can estimate

$$\begin{aligned}\bar{\mathbb{E}}_{\theta^\star}[\log \bar{p}(Y_0^n;\theta^\star)] &= \bar{\mathbb{E}}_{\theta^\star}\left[\log \frac{\bar{p}(Y_0^n;\theta^\star)}{\prod_{k=0}^n \int g_{\theta^\star}(x,Y_k)\pi_{\theta^\star}(dx)}\right] \\ &\quad + \bar{\mathbb{E}}_{\theta^\star}\left[\sum_{k=0}^n \log \int g_{\theta^\star}(x,Y_k)\pi_{\theta^\star}(dx)\right] \\ &\geq -(n+1)\bar{\mathbb{E}}_{\theta^\star}\left[\left|\log \int g_{\theta^\star}(x,Y_0)\pi_{\theta^\star}(dx)\right|\right] \\ &> -\infty,\end{aligned}$$

where we have used that relative entropy is nonnegative. □

The main result of this section follows. After a reduction to the stationary case by means of the previous lemma, the proof concludes by verifying the assumptions of the generalized Shannon–Breiman–McMillan theorem [2].

THEOREM 9. *Assume* (A1) *and* (A2). *There exists* $-\infty < \ell(\theta^\star) < \infty$ *such that*

$$\ell(\theta^\star) = \lim_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}[n^{-1}\log \bar{p}(Y_0^n;\theta^\star)], \tag{7}$$

and such that

$$\ell(\theta^\star) = \lim_{n\to\infty} n^{-1}\log p^\nu(Y_0^n;\theta^\star), \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.} \tag{8}$$

for any probability measure ν *such that one of the assumptions 1–3 of Theorem 1 is satisfied (in particular, the result holds for* $\nu = \pi_{\theta^\star}$*).*

PROOF. Note that $D_n \stackrel{\text{def}}{=} \bar{\mathbb{E}}_{\theta^\star}[\log \bar{p}(Y_0^{n+1};\theta^\star)] - \bar{\mathbb{E}}_{\theta^\star}[\log \bar{p}(Y_0^n;\theta^\star)]$ is a nondecreasing sequence by [2], page 1292, and Lemma 8. Therefore (7) follows immediately. As Y_0^∞ is stationary and ergodic under $\bar{\mathbb{P}}_{\theta^\star}$ by (A1), we can estimate

$$\begin{aligned}-\infty < D_0 \leq \sup_{n\geq 0} D_n = \ell(\theta^\star) &\stackrel{\text{def}}{=} \lim_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}[n^{-1}\log \bar{p}(Y_0^n;\theta^\star)] \\ &\leq \bar{\mathbb{E}}_{\theta^\star}\left[\sup_{x\in\mathsf{X}} (\log g_{\theta^\star}(x,Y_0))^+\right] < \infty,\end{aligned}$$

where we have used (A2). To proceed, we note that the generalized Shannon–Breiman–McMillan theorem ([2], Theorem 1), implies that (8) holds for $\nu = \pi_{\theta^\star}$. Therefore, to prove (8) for arbitrary ν, it suffices to prove the existence of a random variable C^ν satisfying $\bar{\mathbb{P}}_{\theta^\star}(0 < C^\nu < \infty) = 1$ and

$$\lim_{n\to\infty} \frac{p^\nu(Y_0^n;\theta^\star)}{\bar{p}(Y_0^n;\theta^\star)} = C^\nu, \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.} \tag{9}$$

Let $P_n^\nu \stackrel{\text{def}}{=} \mathbb{P}_{\theta^\star}^\nu(Y_0^n \in \cdot)$ and $\bar{P}_n \stackrel{\text{def}}{=} \bar{\mathbb{P}}_{\theta^\star}(Y_0^n \in \cdot)$. Then $p^\nu(Y_0^n;\theta^\star)/\bar{p}(Y_0^n;\theta^\star) = dP_n^\nu/d\bar{P}_n$, and we find that (9) holds with $0 < C^\nu = d\mathbb{P}_{\theta^\star}^{\nu,Y}/d\bar{\mathbb{P}}_{\theta^\star}^Y < \infty$ provided $\mathbb{P}_{\theta^\star}^{\nu,Y} \sim \bar{\mathbb{P}}_{\theta^\star}^Y$. But the latter was already established in Lemma 7. $\square$

REMARK 12. In the case that $g_{\theta^\star}(x,y) > 0$ but $Q_{\theta^\star}$ is periodic, the assumption in the above theorem that the initial probability measure ν has mass in each periodic class of $Q_{\theta^\star}$ cannot be eliminated, as the following example shows. Let $\mathsf{X} = \mathsf{Y} = \{1,2\}$, and let Q_θ be the Markov chain with transition probability matrix $\mathcal{Q}$ and invariant measure π (independent of θ)

$$\mathcal{Q} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \pi = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

Then Q_θ is positive (Harris) recurrent with period 2. For each $\theta \in \Theta = [0.5, 0.9]$, define the observation density $g_\theta(x,y)$ (with respect to the counting measure)

$$g_\theta(x,y) = \theta \mathbb{1}_{y=x} + (1-\theta)\mathbb{1}_{y\neq x},$$

and let $\theta^\star = 0.7$, for example. Then certainly assumptions (A1) and (A2) are satisfied.

Now consider $\nu = \delta_1$. Then ν only has mass in one of the two periodic classes of $Q_{\theta^\star}$. We can compute the observation likelihood as follows:

$$\begin{aligned}\log p^\nu(Y_0^{2n};\theta) = &\sum_{k=0}^{n} \{\mathbb{1}_{Y_{2k}=1}\log\theta + \mathbb{1}_{Y_{2k}=2}\log(1-\theta)\} \\ &+ \sum_{k=1}^{n} \{\mathbb{1}_{Y_{2k-1}=2}\log\theta + \mathbb{1}_{Y_{2k-1}=1}\log(1-\theta)\}.\end{aligned}$$

A straightforward computation shows that

$$\begin{aligned}&\lim_{n\to\infty} (2n)^{-1}\log p^\nu(Y_0^{2n};\theta) \\ &\quad = \{\theta^\star\log\theta + (1-\theta^\star)\log(1-\theta)\}\mathbb{1}_{X_0=1} \\ &\qquad + \{(1-\theta^\star)\log\theta + \theta^\star\log(1-\theta)\}\mathbb{1}_{X_0=2}, \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.}\end{aligned}$$

Therefore, $\lim_{n\to\infty} n^{-1}\log p^\nu(Y_0^n;\theta^\star)$ is not even nonrandom $\bar{\mathbb{P}}_{\theta^\star}$-a.s., let alone equal to $\ell(\theta^\star)$. Thus we see that Theorem 9 does not hold for such ν. Moreover, we can compute directly in this example that

$$\lim_{n\to\infty} \hat{\theta}_{\nu,n} = \theta^\star \mathbb{1}_{X_0=1} + 0.5\mathbb{1}_{X_0=2}, \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.},$$

so that evidently the maximum likelihood estimator is not consistent when we choose the initial measure ν. This shows that also in Theorem 1 the assumption that ν has mass in each periodic class of $Q_{\theta^\star}$ cannot be eliminated.

4.2. *Identifiability.* In this section, we establish the identifiability of the parameter θ. The key issue in the proof consists in showing that the relative entropy rate between $\bar{p}(\cdot;\theta^\star)$ and $p^\lambda(\cdot;\theta)$ may be zero only if $\theta^\star \sim \theta$. Our proof is based on a very simple and intuitive information-theoretic device, given as Lemma 10 below, which avoids the need for an explicit representation of the asymptotic contrast function as in previous proofs of identifiability.

DEFINITION 2. For each n, let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be probability measures on a measurable space $(\mathsf{Z}_n, \mathcal{Z}_n)$. Then $(\mathbb{Q}_n)$ is *exponentially separated* from $(\mathbb{P}_n)$, denoted as $(\mathbb{Q}_n) \dashv (\mathbb{P}_n)$, if there exists a sequence (A_n) of sets $A_n \in \mathcal{Z}_n$ such that

$$\liminf_{n\to\infty} \mathbb{P}_n(A_n) > 0,$$

$$\limsup_{n\to\infty} n^{-1} \log \mathbb{Q}_n(A_n) < 0.$$

If $\mathbb{P}$ and $\mathbb{Q}$ are probability measures on $(\mathsf{Y}^{\mathbb{N}}, \mathcal{Y}^{\otimes\mathbb{N}})$, then we will write $\mathbb{Q} \dashv \mathbb{P}$ if $(\mathbb{Q}_n) \dashv (\mathbb{P}_n)$ with $\mathbb{Q}_n = \mathbb{Q}(Y_0^n \in \cdot)$ and $\mathbb{P}_n = \mathbb{P}(Y_0^n \in \cdot)$.

LEMMA 10. *If* $(\mathbb{Q}_n) \dashv (\mathbb{P}_n)$, *then* $\liminf_{n\to\infty} n^{-1}\mathrm{KL}(\mathbb{P}_n||\mathbb{Q}_n) > 0$.

PROOF. A standard property of the relative entropy ([10], Lemma 1.4.3(g)), states that for any pair of probability measures $\mathbb{P}, \mathbb{Q}$ and measurable set A

$$\mathrm{KL}(\mathbb{P}||\mathbb{Q}) \geq \mathbb{P}(A)\log\mathbb{P}(A) - \mathbb{P}(A)\log\mathbb{Q}(A) - 1,$$

where $0\log 0 = 0$ by convention. As $x\log x \geq -e^{-1}$, we have

$$\liminf_{n\to\infty} n^{-1}\mathrm{KL}(\mathbb{P}_n||\mathbb{Q}_n) \geq \Big(\liminf_{n\to\infty} \mathbb{P}(A_n)\Big)\Big(-\limsup_{n\to\infty} n^{-1}\log\mathbb{Q}(A_n)\Big).$$

The result follows directly. □

As a consequence of this result, we obtain positive entropy rates:

$$\mathbb{P}_\theta^{\nu,Y} \dashv \bar{\mathbb{P}}_{\theta^\star}^{Y} \quad\Longrightarrow\quad \liminf_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}\left[n^{-1}\log\frac{\bar{p}(Y_0^n;\theta^\star)}{p^\nu(Y_0^n;\theta)}\right] > 0. \tag{10}$$

This yields identifiability of the asymptotic contrast function in a very simple and natural manner. It turns out that the exponential separation assumption $\mathbb{P}_\theta^{\nu,Y} \dashv \bar{\mathbb{P}}_{\theta^\star}^{Y}$ always holds when the Markov chain $\mathbb{P}_\theta^\nu$ is V-uniformly ergodic and $\nu(V) < \infty$; this is proved in Section 5 below. This observation allows us to establish the consistency of the MLE in a large class of models.

There is an additional complication that arises in our proof of consistency. Rather than (10), the following result turns out to be of crucial importance:

$$\liminf_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}\left[n^{-1}\log\frac{\bar{p}(Y_0^n;\theta^\star)}{p^\lambda(Y_0^n;\theta)}\right] > 0.$$

This result seems almost identical to (10). However, note that the probability measure ν is replaced here by λ, the dominating measure on $(\mathsf{X}, \mathcal{X})$, which may only be σ-finite [$\lambda(\mathsf{X}) = \infty$]. In this case, a direct application of Lemma 10 is not possible since $y_0^n \mapsto p^\lambda(y_0^n; \theta)$ is not a probability density:

$$\int p^\lambda(y_0^n; \theta)\, dy_0^n = \lambda(\mathsf{X}) = \infty.$$

Nevertheless, the following lemma allows us to reduce the proof in the case of an improper initial measure λ to an application of Lemma 10.

LEMMA 11. *Assume* (A4). *For $\theta \not\sim \theta^\star$ such that $p^\lambda(Y_0^{r_\theta}; \theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s., there exists a probability measure $\tilde{\mathbb{P}}_\theta^\lambda$ on $(\mathsf{Y}^{\mathbb{N}}, \mathcal{Y}^{\otimes\mathbb{N}})$ such that*

$$\tilde{\mathbb{P}}_\theta^\lambda(Y_0^n \in A) = \int \mathbb{1}_A(y_0^n) \frac{p^\lambda(y_0^n; \theta)}{p^\lambda(y_0^{r_\theta}; \theta)} \bar{p}(y_0^{r_\theta}; \theta^\star)\, dy_0^n$$

for all $n \geq r_\theta$ and $A \in \mathcal{Y}^{\otimes(n+1)}$.

PROOF. As $\int p^\lambda(y_0^n; \theta)\, dy_{r_\theta+1}^n = p^\lambda(y_0^{r_\theta}; \theta) < \infty$ $\bar{\mathbb{P}}_{\theta^\star}$-a.e. for all $n \geq r_\theta$ by Fubini's theorem and assumption (A4), we can define for $n \geq r_\theta$

$$\tilde{p}^\lambda(y_0^n, \theta) \stackrel{\text{def}}{=} p^\lambda(y_0^n; \theta) \frac{\bar{p}(y_0^{r_\theta}; \theta^\star)}{p^\lambda(y_0^{r_\theta}; \theta)} < \infty, \qquad dy_0^n\text{-a.e.} \tag{11}$$

Note that, by construction, $\{\tilde{p}^\lambda(y_0^n, \theta)\, dy_0^n : n \geq r_\theta\}$ is a consistent family of probability measures. By the extension theorem, we may construct a probability measure $\tilde{\mathbb{P}}_\theta^\lambda$ on $(\mathsf{Y}^{\mathbb{N}}, \mathcal{Y}^{\otimes\mathbb{N}})$ such that $\tilde{\mathbb{P}}_\theta^\lambda(Y_0^n \in A) = \int \mathbb{1}_A(y_0^n) \tilde{p}^\lambda(y_0^n, \theta)\, dy_0^n$. □

THEOREM 12. *Assume* (A2), (A4) *and* (A6). *Then for every $\theta \not\sim \theta^\star$*

$$\liminf_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}\left[n^{-1} \log \frac{\bar{p}(Y_0^n; \theta^\star)}{p^\lambda(Y_0^n; \theta)}\right] > 0. \tag{12}$$

PROOF. Fix $\theta \not\sim \theta^\star$. Let us assume first that $\bar{\mathbb{P}}_{\theta^\star}(p^\lambda(Y_0^{r_\theta}; \theta) = 0) > 0$. As we have $\int p^\lambda(y_0^n; \theta)\, dy_{r_\theta+1}^n = p^\lambda(y_0^{r_\theta}; \theta)$ by Fubini's theorem, it must be the case that $\bar{\mathbb{P}}_{\theta^\star}(p^\lambda(Y_0^n; \theta) = 0) > 0$ for all $n \geq r_\theta$, so that the expression in (12) is clearly equal to $+\infty$. Therefore, in this case, the claim holds trivially.

We may therefore assume that $p^\lambda(Y_0^{r_\theta}; \theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. Let $\tilde{p}^\lambda(y_0^n, \theta)$ be as in the proof of Lemma 11. Note that $\bar{\mathbb{E}}_{\theta^\star}[|\log \bar{p}(Y_0^{r_\theta}; \theta^\star)|] < \infty$ by Lemma 8, while $\bar{\mathbb{E}}_{\theta^\star}[(\log p^\lambda(Y_0^{r_\theta}; \theta))^+] < \infty$ by assumption (A4). Then we find

$$\liminf_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}\left[n^{-1} \log \frac{\bar{p}(Y_0^n; \theta^\star)}{\tilde{p}^\lambda(Y_0^n; \theta)}\right] \leq \liminf_{n\to\infty} \bar{\mathbb{E}}_{\theta^\star}\left[n^{-1} \log \frac{\bar{p}(Y_0^n; \theta^\star)}{p^\lambda(Y_0^n; \theta)}\right].$$

Assumption (A6) gives $\tilde{\mathbb{P}}_\theta^\lambda \dashv \bar{\mathbb{P}}_{\theta^\star}^Y$. Therefore, (12) follows from Lemma 10. □

4.3. *Consistency of the MLE.* Proofs of convergence of the MLE typically require to establish the convergence of the normalized likelihood $n^{-1}\log p^\nu(Y_0^n;\theta)$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. for any parameter θ. The existence of a limit follows from the Shannon–Breiman–McMillan theorem when $\theta=\theta^\star$ (as in Theorem 9), but is far from clear for other θ. In [23], the convergence of $n^{-1}\log p^\nu(Y_0^n;\theta)$ is established using Kingman's subadditive ergodic theorem. This approach fails in the present setting, as $\log p^\nu(Y_0^n;\theta)$ may not be subadditive even up to a constant.

The approach adopted here is inspired by [23]. We note, however, that it is not necessary to prove convergence of $n^{-1}\log p^\nu(Y_0^n;\theta)$ as long as it is asymptotically bounded away from $\ell(\theta^\star)$, the likelihood of the true parameter. It therefore suffices to bound $n^{-1}\log p^\nu(Y_0^n;\theta)$ above by an auxiliary sequence that is bounded away from $\ell(\theta^\star)$. Here the asymptotics of $n^{-1}\log p^\lambda(Y_0^n;\theta)$ come into play.

LEMMA 13. *Assume* (A1)–(A6). *Then, for any $\theta\not\sim\theta^\star$, there exists an integer n_θ and $\eta_\theta>0$ such that $B(\theta,\eta_\theta)\subseteq\mathcal{U}_\theta$ and*

$$\begin{aligned}&\frac{1}{n_\theta+l}\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in B(\theta,\eta_\theta)}\log p^\lambda(Y_0^{n_\theta};\theta')\Big]\\&\quad+\frac{1}{n_\theta+l}\sup_{\theta'\in B(\theta,\eta_\theta)}\log|q_{\theta'}|_\infty\\&\quad+\frac{l-1}{n_\theta+l}\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in B(\theta,\eta_\theta)}\sup_{x\in\mathsf{X}}(\log g_{\theta'}(x,Y_0))^+\Big]<\ell(\theta^\star).\end{aligned}$$

Here $B(\theta,\eta)\subseteq\Theta$ is the ball of radius $\eta>0$ centered at $\theta\in\Theta$.

PROOF. By (7) and Theorem 12, $\limsup_n n^{-1}\bar{\mathbb{E}}_{\theta^\star}[\log p^\lambda(Y_0^n;\theta)]<\ell(\theta^\star)$. Using (A4), this implies that there exists a (nonrandom) integer $n_\theta>r_\theta$ such that

$$\begin{aligned}&\frac{1}{n_\theta+l}\bar{\mathbb{E}}_{\theta^\star}[\log p^\lambda(Y_0^{n_\theta};\theta)]+\frac{1}{n_\theta+l}\sup_{\theta'\in\mathcal{U}_\theta}\log|q_{\theta'}|_\infty\\&\quad+\frac{l-1}{n_\theta+l}\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in\mathcal{U}_\theta}\sup_{x\in\mathsf{X}}(\log g_{\theta'}(x,Y_0))^+\Big]<\ell(\theta^\star).\end{aligned}\tag{13}$$

For any $\eta>0$ such that $B(\theta,\eta)\subseteq\mathcal{U}_\theta$, we have

$$\begin{aligned}&\sup_{\theta'\in B(\theta,\eta)}\log p^\lambda(Y_0^{n_\theta};\theta')\\&\quad\le\sup_{\theta'\in\mathcal{U}_\theta}(\log p^\lambda(Y_0^{r_\theta};\theta'))^+\\&\qquad+\sum_{k=r_\theta+1}^{n_\theta}\sup_{\theta'\in\mathcal{U}_\theta}\sup_{x\in\mathsf{X}}(\log g_{\theta'}(x,Y_k))^+,\end{aligned}$$

where the right-hand side does not depend on η and is integrable. But then

$$\begin{aligned}
&\limsup_{\eta\downarrow 0} \bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in B(\theta,\eta)} \log p^\lambda(Y_0^{n_\theta};\theta')\Big] \\
&\qquad \le \bar{\mathbb{E}}_{\theta^\star}\Big[\limsup_{\eta\downarrow 0} \sup_{\theta'\in B(\theta,\eta)} \log p^\lambda(Y_0^{n_\theta};\theta')\Big] \qquad (14)\\
&\qquad \le \bar{\mathbb{E}}_{\theta^\star}[\log p^\lambda(Y_0^{n_\theta};\theta)],
\end{aligned}$$

by (A5) and Fatou's lemma. Together (14) and (13) complete the proof. □

Proof of Theorem 1. Since, by Theorem 9, $\lim_{n\to\infty} n^{-1}\ell_{\nu,n}(\theta^\star) = \ell(\theta^\star)$, $\bar{\mathbb{P}}_{\theta^\star}$-a.s., it is sufficient to prove that for any closed set $\mathsf{C}\subset\Theta$ such that $\mathsf{C}\cap[\theta^\star]=\varnothing$

$$\limsup_{n\to\infty}\sup_{\theta'\in\mathsf{C}} n^{-1}\ell_{\nu,n}(\theta') < \ell(\theta^\star), \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.}$$

Now note that $\{B(\theta,\eta_\theta):\theta\in\mathsf{C}\}$ is a cover of C, where η_θ are defined in Lemma 13. As Θ is compact, C is also compact and thus admits a finite subcover $\{B(\theta_i,\eta_{\theta_i}):\theta_i\in\mathsf{C}, i=1,\dots,N\}$. It therefore suffices to show that

$$\limsup_{n\to\infty}\sup_{\theta'\in B(\theta,\eta_\theta)\cap\mathsf{C}} n^{-1}\ell_{\nu,n}(\theta') < \ell(\theta^\star), \qquad \bar{\mathbb{P}}_{\theta^\star}\text{-a.s.}$$

for any $\theta\not\sim\theta^\star$. Fix $\theta\not\sim\theta^\star$ and let η_θ and n_θ be as in Lemma 13. Note that

$$p^\nu(y_0^n;\theta') \le p^\nu(y_0^m;\theta')p^\lambda(y_{m+l}^n;\theta')g^*_{\theta'}(y_{m+1}^{m+l-1})|q_{\theta'}|_\infty, \qquad (15)$$

$$p^\lambda(y_j^n;\theta') \le p^\lambda(y_j^m;\theta')p^\lambda(y_{m+l}^n;\theta')g^*_{\theta'}(y_{m+1}^{m+l-1})|q_{\theta'}|_\infty, \qquad (16)$$

for any $j\le m$, $m+l\le n$ and $\theta'\not\sim\theta^\star$, where $g^*_\theta(y_i^j)\stackrel{\text{def}}{=}\prod_{\ell=i}^j\sup_{x\in\mathsf{X}} g_\theta(x,y_\ell)$. We can therefore estimate, for all n sufficiently large,

$$\begin{aligned}
\ell_{\nu,n}(\theta') &\le \frac{1}{n_\theta+l}\sum_{r=1}^{n_\theta+l}\{\ell_{\nu,r-1}(\theta')+\log p^\lambda(Y_{l+r-1}^n;\theta')+\log(g^*_{\theta'}(Y_r^{l+r-2})|q_{\theta'}|_\infty)\}\\
&\le \frac{1}{n_\theta+l}\sum_{r=1}^{n_\theta+l}\sum_{k=1}^{i(n)-1}\{\log p^\lambda(Y_{(n_\theta+l)(k-1)+l+r-1}^{(n_\theta+l)k+r-1};\theta')+\log|q_{\theta'}|_\infty\}\\
&\quad+\frac{1}{n_\theta+l}\sum_{r=1}^{n_\theta+l}\sum_{k=1}^{i(n)-1}\log g^*_{\theta'}(Y_{(n_\theta+l)(k-1)+r}^{(n_\theta+l)(k-1)+l+r-2})\\
&\quad+\frac{1}{n_\theta+l}\sum_{r=1}^{n_\theta+l}\log(g^*_{\theta'}(Y_0^{r-1})g^*_{\theta'}(Y_{(n_\theta+l)(i(n)-1)+r}^n))
\end{aligned}$$

$$= \frac{1}{n_\theta + l} \sum_{r=1}^{(n_\theta+l)(i(n)-1)} \left\{ \log p^\lambda(Y_{l+r-1}^{n_\theta+l+r-1}; \theta') + \sum_{k=0}^{l-2} \sup_{x \in \mathsf{X}} \log g_{\theta'}(x, Y_{k+r}) \right\}$$

$$+ (i(n)-1) \log |q_{\theta'}|_\infty + \frac{1}{n_\theta + l} \sum_{r=1}^{n_\theta+l} \log g_{\theta'}^*(Y_0^{r-1})$$

$$+ \frac{1}{n_\theta + l} \sum_{r=1}^{n_\theta+l} \sum_{k=(n_\theta+l)(i(n)-1)+r}^{n} \sup_{x \in \mathsf{X}} \log g_{\theta'}(x, Y_k),$$

where $i(n) \stackrel{\text{def}}{=} \max\{m \in \mathbb{N} : m(n_\theta + l) \le n\}$. Here we have applied (15) with $m = r - 1$ in the first inequality, while we have repeatedly applied (16) for every $m = (n_\theta + l)k + r - 1$, $k \le i(n) - 2$ in the second inequality, together with the simple estimates $\ell_{\nu,r-1}(\theta') \le \log g_{\theta'}^*(Y_0^{r-1})$ and

$$p^\lambda(Y_{(n_\theta+l)(i(n)-2)+l+r-1}^{n}; \theta')$$
$$\le p^\lambda(Y_{(n_\theta+l)(i(n)-2)+l+r-1}^{(n_\theta+l)(i(n)-1)+r-1}; \theta') g_{\theta'}^*(Y_{(n_\theta+l)(i(n)-1)+r}^{n}).$$

We can now estimate, for all n sufficiently large,

$$\sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} \ell_{\nu,n}(\theta')$$

$$\le \frac{1}{n_\theta + l} \sum_{r=1}^{(n_\theta+l)(i(n)-1)} \sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} \log p^\lambda(Y_{l+r-1}^{n_\theta+l+r-1}; \theta')$$

$$+ \sum_{k=0}^{l-2} \frac{1}{n_\theta + l} \sum_{r=1}^{(n_\theta+l)(i(n)-1)} \sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} \sup_{x \in \mathsf{X}} (\log g_{\theta'}(x, Y_{k+r}))^+$$

$$+ (i(n)-1) \sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} \log |q_{\theta'}|_\infty$$

$$+ \frac{1}{n_\theta + l} \sum_{r=1}^{n_\theta+l} \sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} \log g_{\theta'}^*(Y_0^{r-1})$$

$$+ \sum_{k=n-2(n_\theta+l)+1}^{n} \sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} \sup_{x \in \mathsf{X}} (\log g_{\theta'}(x, Y_k))^+,$$

where we have used that $(n_\theta + l)(i(n) - 1) + r \ge n - 2(n_\theta + l) + 1$ to estimate the last term. But as $i(n)/n \to (n_\theta + l)^{-1}$ as $n \to \infty$, we find that

$$\limsup_{n \to \infty} \sup_{\theta' \in B(\theta,\eta_\theta) \cap \mathsf{C}} n^{-1} \ell_{\nu,n}(\theta') \le \frac{1}{n_\theta + l} \bar{\mathbb{E}}_{\theta^\star} \Big[\sup_{\theta' \in B(\theta,\eta_\theta)} \log p^\lambda(Y_0^{n_\theta}; \theta') \Big]$$

$$+\frac{l-1}{n_\theta+l}\bar{\mathbb{E}}_{\theta^\star}\Big[\sup_{\theta'\in B(\theta,\eta_\theta)}\sup_{x\in\mathsf{X}}(\log g_{\theta'}(x,Y_0))^+\Big]$$

$$+\frac{1}{n_\theta+l}\sup_{\theta'\in B(\theta,\eta_\theta)}\log|q_{\theta'}|_\infty$$

$$<\ell(\theta^\star)$$

by (A4), Birkhoff's ergodic theorem, Lemma 13, and the elementary fact that $\lim_n \frac{1}{n}\sum_{k=n-r+1}^{n}\xi_k = \lim_n \frac{1}{n}\sum_{k=1}^{n}\xi_k - \lim_n \frac{1}{n}\sum_{k=1}^{n-r}\xi_k = 0$ for any stationary ergodic sequence $(\xi_k)_{k\geq 0}$ with $\mathbb{E}(|\xi_1|)<\infty$. This completes the proof. □

5. Exponential separation and V-uniform ergodicity. As is explained in Remark 6, the key step in establishing assumption (A6) is to obtain a type of large deviations property. The following Azuma–Hoeffding type inequality provides what is needed in the V-uniformly ergodic case.

THEOREM 14. *Assume that Q_θ is V_θ-uniformly ergodic. Fix $s\geq 0$, and let $f:\mathsf{Y}^{s+1}\to\mathbb{R}$ be such that $|f|_\infty<\infty$. Then there exists a constant K such that*

$$\mathbb{P}_\theta^\nu\left(\left|\sum_{i=1}^{n}\{f(Y_i^{i+s})-\bar{\mathbb{E}}_\theta[f(Y_0^s)]\}\right|\geq t\right)\leq K\nu(V)\exp\left[-\frac{1}{K}\left(\frac{t^2}{n}\wedge t\right)\right]$$

for any probability measure ν and any $t>0$.

We will first use this result in Section 5.1 to prove Theorem 2. In Section 5.2, we will establish a general Azuma–Hoeffding type large deviations inequality for V-uniformly ergodic Markov chains, which forms the basis for the proof of Theorem 14. Finally, Section 5.3 completes the proof of Theorem 14.

5.1. *Proof of Theorem 2.* We begin by proving that exponential separation holds under the V-uniform ergodicity assumption.

PROPOSITION 15. *Assume* (A1) *and* (A6′). *For any $\theta\not\sim\theta^\star$ with $p^\lambda(Y_0^{r_\theta};\theta)>0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. and probability measure ν such that $\nu(V_\theta)<\infty$, we have $\mathbb{P}_\theta^{\nu,Y}\dashv\bar{\mathbb{P}}_{\theta^\star}^{Y}$.*

PROOF. Fix $\theta\not\sim\theta^\star$. As $\bar{\mathbb{P}}_\theta^Y\neq\bar{\mathbb{P}}_{\theta^\star}^Y$ by assumption (A6′), there exists an integer $s\geq 0$ and a bounded measurable function $h:\mathsf{Y}^{s+1}\to\mathbb{R}$ such that $\bar{\mathbb{E}}_\theta[h(Y_0^s)]=0$ and $\bar{\mathbb{E}}_{\theta^\star}[h(Y_0^s)]=1$. Define for $n\geq s$ the set $A_n\in\mathcal{Y}^{\otimes(n+1)}$ as

$$A_n\stackrel{\text{def}}{=}\left\{y_0^n\in\mathsf{Y}^{n+1}:\left|\frac{1}{n-s}\sum_{i=1}^{n-s}h(y_i^{i+s})\right|\geq\frac{1}{2}\right\}.$$

As Y_0^∞ is stationary and ergodic under $\bar{\mathbb{P}}_{\theta^\star}$ by (A1), Birkhoff's ergodic theorem gives $\bar{\mathbb{P}}^Y_{\theta^\star}(A_n) \to 1$ as $n \to \infty$. On the other hand, Theorem 14 shows that $\limsup_{n\to\infty} n^{-1} \log \mathbb{P}^{\nu,Y}_\theta(A_n) < 0$. Thus, we have established $\mathbb{P}^{\nu,Y}_\theta \dashv \bar{\mathbb{P}}^Y_{\theta^\star}$. □

Proposition 15 is not sufficient to establish (A6), however: the problem is that we are interested in the case where ν is not a probability measure, but the σ-finite measure λ. What remains is to reduce this problem to an application of Proposition 15. To this end, we will use the following lemma.

LEMMA 16. *Assume* (A4), *and fix* $\theta \not\sim \theta^\star$ *such that* $p^\lambda(Y_0^{r_\theta};\theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$*-a.s. For any* $B \in \mathcal{Y}^{\otimes(r_\theta+1)}$ *such that* $\bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B) > 0$, *define the probability measure*

$$\lambda_{B,\theta}(A) = \bar{\mathbb{E}}_{\theta^\star}\left(\int \mathbb{1}_A(x_{r_\theta+1}) \frac{p^\lambda(dx_{r_\theta+1}, Y_0^{r_\theta};\theta)}{p^\lambda(Y_0^{r_\theta};\theta)} \Big| Y_0^{r_\theta} \in B\right)$$

on $(\mathsf{X},\mathcal{X})$. *Then we have*

$$\tilde{\mathbb{P}}^\lambda_\theta(Y_0^{r_\theta} \in B, Y_{r_\theta+1}^n \in A) = \bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B)\mathbb{P}_\theta^{\lambda_{B,\theta}}(Y_0^{n-r_\theta-1} \in A)$$

for any set $A \in \mathcal{Y}^{\otimes(n-r_\theta)}$.

PROOF. Note that by assumption (A4), $\tilde{\mathbb{P}}^\lambda_\theta$ is well defined (as shown in Lemma 11) and $0 < p^\lambda(Y_0^{r_\theta};\theta) < \infty$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s. Moreover, as $p^\lambda(y_0^{r_\theta};\theta) = \int p^\lambda(dx_{r_\theta+1}, y_0^{r_\theta};\theta)$, we find that $\lambda_{B,\theta}$ is indeed a probability measure on $(\mathsf{X},\mathcal{X})$.

Let $B \in \mathcal{Y}^{\otimes(r_\theta+1)}$ be such that $\bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B) > 0$. Then for any $n > r_\theta$

$$\begin{aligned}
&\tilde{\mathbb{P}}^\lambda_\theta(Y_0^{r_\theta} \in B, Y_{r_\theta+1}^n \in A)\\
&\quad= \int \mathbb{1}_A(y_{r_\theta+1}^n)\mathbb{1}_B(y_0^{r_\theta})p^\lambda(y_0^n;\theta)\frac{\bar{p}(y_0^{r_\theta};\theta^\star)}{p^\lambda(y_0^{r_\theta};\theta)}\,dy_0^n\\
&\quad= \bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B)\int\left[\int \mathbb{1}_A(y_{r_\theta+1}^n)p^{x_{r_\theta+1}}(y_{r_\theta+1}^n;\theta)\,dy_{r_\theta+1}^n\right]\lambda_{B,\theta}(dx_{r_\theta+1})\\
&\quad= \bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B)\mathbb{P}_\theta^{\lambda_{B,\theta}}(Y_0^{n-r_\theta-1} \in A),
\end{aligned}$$

where we used $p^\lambda(y_0^n;\theta) = \int p^\lambda(dx_{m+1}, y_0^m;\theta)p^{x_{m+1}}(y_{m+1}^n;\theta)$ for $n > m$. □

We can now complete the proof of Theorem 2.

PROOF OF THEOREM 2. Fix $\theta \not\sim \theta^\star$ such that $p^\lambda(Y_0^{r_\theta};\theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s., and define

$$B = \left\{y_0^{r_\theta} : \int V_\theta(x_{r_\theta+1})\frac{p^\lambda(dx_{r_\theta+1}, y_0^{r_\theta};\theta)}{p^\lambda(y_0^{r_\theta};\theta)} \le K\right\}.$$

By (A6′), we can choose K sufficiently large so that $\bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B) > 0$. Consequently $\lambda_{B,\theta}(V_\theta) \le K < \infty$ by construction. As in the proof of Proposition 15, it follows that there exists a sequence of sets $A_n \in \mathcal{Y}^{\otimes(n-r_\theta)}$ such that

$$\lim_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^{n-r_\theta-1} \in A_n) = 1, \qquad \limsup_{n\to\infty} n^{-1} \log \mathbb{P}_\theta^{\lambda_{B,\theta}}(Y_0^{n-r_\theta-1} \in A_n) < 0.$$

Define the sets

$$\tilde{A}_n \stackrel{\text{def}}{=} \{y_0^n : y_0^{r_\theta} \in B, y_{r_\theta+1}^n \in A_n\}.$$

Using the stationarity of $\bar{\mathbb{P}}_{\theta^\star}$ and Lemma 16, it follows that

$$\lim_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^n \in \tilde{A}_n) = \bar{\mathbb{P}}_{\theta^\star}(Y_0^{r_\theta} \in B) > 0, \qquad \limsup_{n\to\infty} n^{-1} \log \tilde{\mathbb{P}}_\theta^{\lambda}(Y_0^n \in \tilde{A}_n) < 0.$$

This establishes (A6). □

Finally, let us prove Proposition 3.

Proof of Proposition 3. Fix $\theta \not\sim \theta^\star$ such that $p^\lambda(Y_0^{r_\theta}; \theta) > 0$ $\bar{\mathbb{P}}_{\theta^\star}$-a.s., and let $B = \mathsf{Y}^{r_\theta+1}$. By (A6″), there exists a sequence of sets $A_n \in \mathcal{Y}^{\otimes(n-r_\theta)}$ such that

$$\liminf_{n\to\infty} \bar{\mathbb{P}}_{\theta^\star}(Y_0^{n-r_\theta-1} \in A_n) > 0, \qquad \limsup_{n\to\infty} n^{-1} \log \mathbb{P}_\theta^{\lambda_{B,\theta}}(Y_0^{n-r_\theta-1} \in A_n) < 0.$$

Assumption (A6) now follows easily from the stationarity of $\bar{\mathbb{P}}_{\theta^\star}$ and Lemma 16. □

5.2. *An Azuma–Hoeffding inequality.* This section is somewhat independent of the remainder of the paper. We will prove a general Azuma–Hoeffding type large deviations inequality for V-uniformly ergodic Markov chains, on which the proof of Theorem 14 will be based (see Section 5.3). The following result may be seen as an extension of the Azuma–Hoeffding inequality obtained in [16] for uniformly ergodic Markov chains, and the proof of our result is similar to the proof of the Bernstein-type inequality in [1], Theorem 6.

Theorem 17. *Let $(X_k)_{k\ge0}$ be a Markov chain in $(\mathsf{X}, \mathcal{X})$ with transition kernel Q and initial measure η under the probability measure $\mathbb{P}^\eta$. Assume that the transition kernel Q is V-uniformly ergodic, and denote by π its unique invariant measure. Then there exists a constant K such that*

$$\mathbb{P}^\eta\left(\left|\sum_{i=1}^n \{f(X_i) - \pi(f)\}\right| \ge t\right) \le K\eta(V) \exp\left[-\frac{1}{K}\left(\frac{t^2}{n|f|_\infty^2} \wedge \frac{t}{|f|_\infty}\right)\right]$$

for any probability measure η, bounded function $f : \mathsf{X} \to \mathbb{R}$, and $t > 0$.

REMARK 13. The exponential bound of Theorem 17 has a Bernstein-type tail, unlike the usual Azuma–Hoeffding bound. However, unlike the Bernstein inequality, the tail behavior is determined only by $|f|_\infty$, and not by the variance of f. We therefore still refer to this inequality as an Azuma–Hoeffding bound. It is shown in [1] by means of a counterexample that V-uniformly ergodic Markov chains do not admit, in general, a Bernstein bound of the type available for independent random variables (the bound in [1] depends on the variance at the cost of an extra logarithmic factor, which precludes its use for our purposes).

Throughout this section, we let $(X_k)_{k\geq 0}$ be as in Theorem 17. For simplicity, we work with a generic constant K which may change from line to line.

Before we turn to the proof of Theorem 17, let us recall some standard facts from the theory of V-uniformly ergodic Markov chains. It is well known ([27], Chapter 16), that V-uniform ergodicity in the sense of Definition 1 implies (and is essentially equivalent to) the following properties:

Minorization condition. There exist a set $C \in \mathcal{X}$, an integer m, a probability measure ν on $(\mathsf{X}, \mathcal{X})$ and a constant $\varepsilon > 0$ such that

$$Q^m(x, A) \geq \varepsilon \nu(A) \qquad \text{for all } x \in C \text{ and all } A \in \mathcal{X}. \tag{17}$$

Foster–Lyapunov drift condition. There exists a measurable function $V : \mathsf{X} \mapsto [1, \infty)$, $\lambda \in [0, 1)$, and $b < \infty$, such that $\sup_{x\in C} V(x) < \infty$ and

$$QV(x) \leq \lambda V(x) + b\mathbb{1}_C(x) \qquad \text{for all } x \in \mathsf{X}. \tag{18}$$

The set C in the minorization condition is referred to as a (ν, m)-*small set* (see [27] for extensive discussion). For future reference, let us note that

$$1 \leq \pi(V) = (1-\lambda)^{-1}\pi(QV - \lambda V) \leq (1-\lambda)^{-1} b\pi(C) < \infty,$$

which shows that $\pi(V) < \infty$ and $\pi(C) > 0$. Moreover,

$$\varepsilon\pi(C)\nu(V) \leq \pi(Q^m V) = \pi(V) < \infty,$$

so that necessarily $\nu(V) < \infty$ also.

The proof of Theorem 17 is based on an embedding of the Markov chain into a wide sense regenerative process ([20], page 360), known as a *splitting construction*. Let us recall how this can be done. We will employ the canonical process $\check{X}_n \stackrel{\text{def}}{=} (\tilde{X}_n, d_n)$ on the enlarged measure space $(\check{\Omega}, \check{\mathcal{F}})$, where $\check{\Omega} = (\mathsf{X} \times \{0, 1\})^{\mathbb{N}}$ and $\check{\mathcal{F}}$ is the corresponding Borel σ-field. In words, $\tilde{X}_n$ takes values in $(\mathsf{X}, \mathcal{X})$ and d_n is a binary random variable. Define the following stopping times:

$$\sigma_0 \stackrel{\text{def}}{=} \inf\{n \geq 0 : \tilde{X}_n \in C\}, \qquad \sigma_{i+1} \stackrel{\text{def}}{=} \inf\{n \geq \sigma_i + m : \tilde{X}_n \in C\}.$$

We now construct a probability measure $\check{\mathbb{P}}^\eta$ on $(\check{\Omega}, \check{\mathcal{F}})$ with the following properties (e.g., by means of the Ionescu–Tulcea theorem):

$$(d_n)_{n\geq 0} \text{ are i.i.d.} \qquad \text{with } \check{\mathbb{P}}^\eta(d_n = 1) = \varepsilon,$$

$$\tilde{X}_0 \text{ is independent from } (d_n)_{n\geq 0} \quad \text{and} \quad \check{\mathbb{P}}^\eta(\tilde{X}_0 \in \cdot) = \eta,$$

$$\check{\mathbb{P}}^\eta(\tilde{X}_{n+1} \in \cdot | \tilde{X}_0^n, d_0^\infty) = Q(\tilde{X}_n, \cdot) \qquad \text{on } \{n < \sigma_0\} \cup \bigcup_{i\geq 0} \{\sigma_i + m \leq n < \sigma_{i+1}\},$$

$$\check{\mathbb{P}}^\eta(\tilde{X}_{\sigma_i+1}^{\sigma_i+m} \in \cdot | \tilde{X}_0^{\sigma_i}, d_0^\infty) = \begin{cases} \displaystyle\int \mathbf{q}^{\tilde{X}_{\sigma_i}, x}(\cdot)\nu(dx), & \text{if } d_{\sigma_i} = 1, \\ \displaystyle\int \mathbf{q}^{\tilde{X}_{\sigma_i}, x}(\cdot) R(\tilde{X}_{\sigma_i}, dx), & \text{if } d_{\sigma_i} = 0. \end{cases}$$

Here we defined the transition kernel $R(x, A) \stackrel{\text{def}}{=} (1-\varepsilon)^{-1}\{Q^m(x, A) - \varepsilon\nu(A)\}$ for $x \in C$, and (using that X is Polish to ensure existence) the regular conditional probability $\mathbf{q}^{X_0, X_m}(A) \stackrel{\text{def}}{=} \mathbb{P}^\eta(X_1^m \in A | X_0, X_m)$.

The process $(\check{X}_n)_{n\geq 0}$ is not necessarily Markov. However, it is easily verified that the law of the process $(\tilde{X}_n)_{n\geq 0}$ under $\check{\mathbb{P}}^\eta$ is the same as the law of $(X_n)_{n\geq 0}$ under $\mathbb{P}^\eta$, so that our original Markov chain is indeed embedded in this construction. Moreover, at every time σ_n such that additionally $d_{\sigma_n} = 1$, we have by construction that $\tilde{X}_{\sigma_n+m}$ is drawn independently from the distribution ν, that is, the process regenerates in m steps. Let us define the regeneration times as

$$\check{\sigma}_0 \stackrel{\text{def}}{=} \inf\{\sigma_i + m : i \geq 0, d_{\sigma_i} = 1\}, \qquad \check{\sigma}_{n+1} \stackrel{\text{def}}{=} \inf\{\sigma_i + m : \sigma_i \geq \check{\sigma}_n, d_{\sigma_i} = 1\}.$$

The regenerations will allow us to split the path of the process into one-dependent blocks, to which we can apply classical large deviations bounds for independent random variables. We formalize this as the following lemma.

LEMMA 18. *Define for $i \geq 0$ the block sums*

$$\xi_i \stackrel{def}{=} \sum_{k=\check{\sigma}_i}^{\check{\sigma}_{i+1}-1} \{f(\tilde{X}_k) - \pi(f)\}.$$

Then $(\xi_i)_{i\geq 0}$ are identically distributed, one-dependent, and $\check{\mathbb{E}}^\eta(\xi_0) = 0$.

PROOF. First, we note that $\check{\mathbb{P}}^\eta(\check{X}_{\check{\sigma}_i}^{\check{\sigma}_{i+1}-1} \in \cdot | \check{X}_0^{\check{\sigma}_i - m}) = \check{\mathbb{P}}^\nu(\check{X}_0^{\check{\sigma}_0 - 1} \in \cdot)$ for all i. It follows directly that $(\xi_i)_{i\geq 0}$ are identically distributed and one-dependent. Moreover, as $\check{\sigma}_i$ is $\sigma\{X_0^{\check{\sigma}_i - m}\}$-measurable, we find that the inter-regeneration times $(\check{\sigma}_{i+1} - \check{\sigma}_i)_{i\geq 0}$ are independent. Now note that, by the

law of large numbers,

$$\check{\mathbb{E}}^{\eta}(\xi_0) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} \xi_i = \lim_{n\to\infty} \frac{1}{n}\sum_{k=\check{\sigma}_0}^{\check{\sigma}_n-1} \{f(\tilde{X}_k) - \pi(f)\}$$
$$= \lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=0}^{n-1}\{\check{\sigma}_{i+1} - \check{\sigma}_i\}\right)\left(\frac{1}{\check{\sigma}_n - \check{\sigma}_0}\sum_{k=\check{\sigma}_0}^{\check{\sigma}_n-1}\{f(\tilde{X}_k) - \pi(f)\}\right).$$

But $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1}\{\check{\sigma}_{i+1} - \check{\sigma}_i\} = \check{\mathbb{E}}^{\eta}(\check{\sigma}_1 - \check{\sigma}_0) < \infty$ by the law of large numbers and (19) below, while $\lim_{n\to\infty} \frac{1}{\check{\sigma}_n - \check{\sigma}_0}\sum_{k=\check{\sigma}_0}^{\check{\sigma}_n-1}\{f(\tilde{X}_k) - \pi(f)\} = 0$ by the ergodic theorem for Markov chains. This completes the proof. □

In the proof of Theorem 17, we will need that fact that the inter-regeneration times $\check{\sigma}_0$ and $\check{\sigma}_{i+1} - \check{\sigma}_i$ possess exponential moments. We presently establish that this is necessarily the case, adapting the proof of [29], Theorem 2.1.

PROPOSITION 19. *There exists a constant K such that*

$$\check{\mathbb{E}}^{\eta}[\exp(\check{\sigma}_0/K)] \le K\eta(V) \quad \textit{and} \quad \check{\mathbb{E}}^{\eta}[\exp(\{\check{\sigma}_1 - \check{\sigma}_0\}/K)] \le K \tag{19}$$

for every probability measure η.

PROOF. We begin by writing

$$\{\check{\sigma}_0 - m = n\} = \bigcup_{j\ge 0}\{d_{\sigma_0}, \ldots, d_{\sigma_{j-1}} = 0, d_{\sigma_j} = 1, \sigma_j = n\}.$$

Using the independence of d_{σ_j} from $d_0, \ldots, d_{\sigma_{j-1}}, \sigma_j$, we have

$$\check{\mathbb{P}}^{\eta}(\check{\sigma}_0 - m = n) = \sum_{j=0}^{\infty} \varepsilon(1-\varepsilon)^j \check{\mathbb{P}}^{\eta}(\sigma_j = n | d_{\sigma_0}, \ldots, d_{\sigma_{j-1}} = 0).$$

In particular, we can write

$$\check{\mathbb{E}}^{\eta}(e^{\check{\sigma}_0/K}) = e^{m/K}\sum_{j=0}^{\infty} \varepsilon(1-\varepsilon)^j \check{\mathbb{E}}^{\eta}(e^{\sigma_j/K} | d_{\sigma_0}, \ldots, d_{\sigma_{j-1}} = 0).$$

Now note that by construction, we have

$$\check{\mathbb{E}}^{\eta}(e^{\{\sigma_j - \sigma_{j-1} - m\}/K} | \check{X}_0^{\sigma_{j-1}}) = \check{\mathbb{E}}^{R(\tilde{X}_{\sigma_{j-1}}, \cdot)}(e^{\sigma_0/K}) \qquad \text{on } \{d_{\sigma_{j-1}} = 0\}.$$

Define $G(K) \stackrel{\text{def}}{=} \sup_{x\in C} \check{\mathbb{E}}^{R(x,\cdot)}(e^{\sigma_0/K})$. It is now easily established that

$$\check{\mathbb{E}}^{\eta}(e^{\sigma_j/K} | d_{\sigma_0}, \ldots, d_{\sigma_{j-1}} = 0) \le e^{jm/K} G(K)^j \check{\mathbb{E}}^{\eta}(e^{\sigma_0/K}).$$

We can therefore estimate

$$\check{\mathbb{E}}^{\eta}(e^{\check{\sigma}_0/K}) \le \frac{\varepsilon e^{m/K}\check{\mathbb{E}}^{\eta}(e^{\sigma_0/K})}{1-(1-\varepsilon)e^{m/K}G(K)},$$

provided that $(1-\varepsilon)e^{m/K}G(K)<1$.

Now note that it follows from [27], Theorem 15.2.5, that

$$\check{\mathbb{E}}^{x}(e^{\sigma_0/K}) \le K\{\lambda V(x) + b\mathbb{1}_C(x)\} \qquad \text{for all } x \in \mathsf{X}, \tag{20}$$

provided K is chosen sufficiently large. Therefore, it is easily established that $\check{\mathbb{E}}^{\eta}(e^{\sigma_0/K}) \le K\eta(V)$ for K sufficiently large. On the other hand, by Jensen's inequality, $G(K) \le G(\beta)^{\beta/K}$ for $\beta \le K$. As $G(\beta)<\infty$ for some β by (20), we have $G(K)\to 1$ as $K\to\infty$. Thus, $(1-\varepsilon)e^{m/K}G(K)<1$ for K sufficiently large, and we have proved $\check{\mathbb{E}}^{\eta}[\exp(\check{\sigma}_0/K)] \le K\eta(V)$. To complete the proof, is suffices to note that $\check{\mathbb{E}}^{\eta}[\exp(\{\check{\sigma}_1-\check{\sigma}_0\}/K)] = \check{\mathbb{E}}^{\nu}[\exp(\check{\sigma}_0/K)]$ and $\nu(V)<\infty$. $\square$

With these preliminaries out of the way, we now prove Theorem 17.

PROOF THEOREM 17. Define the sequence $(\xi_\ell)_{\ell\ge 0}$ as in Lemma 18. We begin by splitting the sum $S_n \stackrel{\text{def}}{=} \sum_{i=1}^n \{f(X_i)-\pi(f)\}$ into three different terms:

$$\begin{aligned} S_n = &\sum_{j=1}^{\check{\sigma}_0\wedge n-1}\{f(X_j)-\pi(f)\} + \sum_{k=0}^{i(n)-1}\xi_k \\ &+ \sum_{j=l(n)\wedge n}^{n}\{f(X_j)-\pi(f)\}, \end{aligned} \tag{21}$$

where $i(n) \stackrel{\text{def}}{=} \sum_{k=1}^{\infty}\mathbb{1}_{\{\check{\sigma}_k\le n\}}$ and $l(n) \stackrel{\text{def}}{=} \check{\sigma}_{i(n)}$. Using (19), we have for $t>0$

$$\begin{aligned} &\check{\mathbb{P}}^{\eta}\left[\left|\sum_{j=1}^{\check{\sigma}_0\wedge n-1}\{f(X_j)-\pi(f)\}\right| \ge t\right] \\ &\quad\le \check{\mathbb{P}}^{\eta}[\check{\sigma}_0 \ge t/2|f|_\infty] \\ &\quad\le \check{\mathbb{E}}^{\eta}[\exp(\check{\sigma}_0/K)]\exp\left(-\frac{t}{2K|f|_\infty}\right) \\ &\quad\le K\eta(V)\exp\left(-\frac{t}{2K|f|_\infty}\right). \end{aligned} \tag{22}$$

This bounds the first term of (21). To bound the last term of (21), we proceed as in the proof of [1], Lemma 3. First note that, for any $t > 1$,

$$\begin{aligned}\check{\mathbb{P}}^{\eta}[n - l(n) \wedge n + 1 \geq t] &= \check{\mathbb{P}}^{\eta}[l(n) \leq n + 1 - t] \\ &= \sum_{\ell=0}^{n} \check{\mathbb{P}}^{\eta}[\check{\sigma}_{\ell} \leq n + 1 - t, i(n) = \ell] \\ &= \sum_{\ell=0}^{n} \check{\mathbb{P}}^{\eta}[\check{\sigma}_{\ell} \leq n + 1 - t, \check{\sigma}_{\ell+1} > n].\end{aligned}$$

Recall that the inter-regeneration time $\check{\sigma}_{\ell+1} - \check{\sigma}_{\ell}$ is independent from $\check{\sigma}_0, \ldots, \check{\sigma}_{\ell}$, and $(\check{\sigma}_{\ell+1} - \check{\sigma}_{\ell})_{\ell \geq 0}$ are identically distributed (see the proof of Lemma 18). Thus,

$$\begin{aligned}\check{\mathbb{P}}^{\eta}[\check{\sigma}_{\ell} \leq n + 1 - t, \check{\sigma}_{\ell+1} > n] &= \sum_{k=0}^{\lfloor n+1-t \rfloor} \check{\mathbb{P}}^{\eta}[\check{\sigma}_{\ell} = k, \check{\sigma}_{\ell+1} - \check{\sigma}_{\ell} > n - k] \\ &= \sum_{k=0}^{\lfloor n+1-t \rfloor} \check{\mathbb{P}}^{\eta}[\check{\sigma}_{\ell} = k] \check{\mathbb{P}}^{\eta}[\check{\sigma}_1 - \check{\sigma}_0 > n - k].\end{aligned}$$

But as $\check{\sigma}_{\ell} < \check{\sigma}_{\ell+1}$ for all $\ell \geq 0$, we have $\sum_{\ell=0}^{n} \check{\mathbb{P}}^{\check{\eta}}[\check{\sigma}_{\ell} = k] \leq 1$ for all k, so that

$$\begin{aligned}\check{\mathbb{P}}^{\eta}[n - l(n) \wedge n + 1 \geq t] &\leq \sum_{k=0}^{\lfloor n+1-t \rfloor} \check{\mathbb{P}}^{\eta}[\check{\sigma}_1 - \check{\sigma}_0 > n - k] \\ &\leq \sum_{k=\lceil t-1 \rceil}^{\infty} \check{\mathbb{P}}^{\eta}[\check{\sigma}_1 - \check{\sigma}_0 \geq k] \\ &\leq \check{\mathbb{E}}^{\eta}[e^{\{\check{\sigma}_1 - \check{\sigma}_0\}/K}] \sum_{k=\lceil t-1 \rceil}^{\infty} e^{-k/K} \\ &\leq \left(\frac{K e^{1/K}}{1 - e^{-1/K}} \right) e^{-t/K},\end{aligned}$$

where we have used (19). We therefore find that for $t > 2|f|_{\infty}$

$$\begin{aligned}\check{\mathbb{P}}_{\eta}\left[\left| \sum_{j=l(n)\wedge n}^{n} \{f(X_j) - \pi(f)\} \right| \geq t \right] &\leq \check{\mathbb{P}}_{\eta}[n - l(n) \wedge n + 1 \geq t/2|f|_{\infty}] \\ &\leq K \exp\left(-\frac{t}{2K|f|_{\infty}} \right)\end{aligned} \tag{23}$$

(recall that the constant K changes from line to line). But we may clearly choose K sufficiently large that $Ke^{-1/K} \geq 1$, so that (23) holds for any $t > 0$.

It remains to bound the middle term in (21). As $i(n) \leq n$, we can estimate

$$\left|\sum_{k=0}^{i(n)-1} \xi_k\right| \leq \max_{0\leq j\leq \lfloor n/2\rfloor}\left|\sum_{k=0}^{j} \xi_{2k}\right| + \max_{0\leq j\leq \lfloor n/2\rfloor}\left|\sum_{k=0}^{j} \xi_{2k+1}\right|.$$

Both terms on the right-hand side of this expression are identically distributed. We can therefore estimate using Etemadi's inequality ([5], Theorem 22.5),

$$\check{\mathbb{P}}^\eta\left[\left|\sum_{k=0}^{i(n)-1} \xi_k\right| \geq t\right] \leq 8 \max_{0\leq j\leq \lfloor n/2\rfloor} \check{\mathbb{P}}^\eta\left[\left|\sum_{k=0}^{j} \xi_{2k}\right| \geq t/8\right].$$

Note that $|\xi_k| \leq 2|f|_\infty(\check{\sigma}_{k+1} - \check{\sigma}_k)$, so that using (19)

$$(2K|f|_\infty)^2 \check{\mathbb{E}}^\eta\left(e^{|\xi_k|/2K|f|_\infty} - 1 - \frac{|\xi_k|}{2K|f|_\infty}\right) \leq 4K^3|f|_\infty^2.$$

Using Bernstein's inequality ([30], Lemma 2.2.11), we obtain

$$\check{\mathbb{P}}^\eta\left[\left|\sum_{k=0}^{j} \xi_{2k}\right| \geq t/8\right] \leq 2\exp\left(-\frac{1}{K}\frac{t^2}{(j+1)|f|_\infty^2 + t|f|_\infty}\right).$$

We can therefore estimate for $t > 0$

$$\check{\mathbb{P}}^\eta\left[\left|\sum_{k=0}^{i(n)-1} \xi_k\right| \geq t\right] \leq K\exp\left(-\frac{1}{K}\frac{t^2}{n|f|_\infty^2 + t|f|_\infty}\right). \tag{24}$$

The proof is completed by combining (22), (23) and (24). □

5.3. *Proof of Theorem 14.* Assume without loss of generality that $\bar{\mathbb{E}}_\theta[f(Y_0^s)] = 0$. To prove the result, it suffices to bound each term in the decomposition

$$\sum_{i=1}^{n} f(Y_i^{i+s}) = \sum_{j=0}^{s}\left(\sum_{i=1}^{n} \xi_{i,j}\right) + \sum_{i=1}^{n} \mathbb{E}_\theta^\nu(f(Y_i^{i+s})|X_0^{i-1}, Y_0^{i-1}),$$

where we have defined for any $0 \leq j \leq s$ and $i \geq 1$

$$\xi_{i,j} \stackrel{\text{def}}{=} \mathbb{E}_\theta^\nu(f(Y_i^{i+s})|X_0^{i+j}, Y_0^{i+j}) - \mathbb{E}_\theta^\nu(f(Y_i^{i+s})|X_0^{i+j-1}, Y_0^{i+j-1}).$$

By construction, $(\xi_{i,j})_{1\leq i\leq n}$ are martingale increments for each j, and $|\xi_{i,j}|_\infty \leq 2|f|_\infty$. Therefore, by the Azuma–Hoeffding inequality ([32], page 237), we have

$$\mathbb{P}_\theta^\nu\left(\left|\sum_{i=1}^{n} \xi_{i,j}\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{8n|f|_\infty^2}\right)$$

for each $0 \le j \le s$. On the other hand, note that $\mathbb{E}_\theta^\nu(f(Y_i^{i+s})|X_0^{i-1}, Y_0^{i-1}) = F(X_{i-1})$ for all i, where F satisfies $\pi_\theta(F) = 0$ (as we assumed $\bar{\mathbb{E}}_\theta[f(Y_0^s)] = 0$) and $|F|_\infty \le |f|_\infty$. The result therefore follows by applying Theorem 17.

REFERENCES

[1] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034. MR2424985

[2] Barron, A. (1985). The strong ergodic theorem for densities; generalized Shannon–McMillan–Breiman theorem. *Ann. Probab.* **13** 1292–1303. MR0806226

[3] Baum, L. E. and Petrie, T. P. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37** 1554–1563. MR0202264

[4] Bertsekas, D. P. and Shreve, S. E. (1978). *Stochastic Optimal Control: The Discrete Time Case. Mathematics in Science and Engineering* **139**. Academic Press, New York. MR0511544

[5] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York. MR1324786

[6] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer, New York. MR2159833

[7] Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers & Chemistry* **16** 107–115.

[8] Douc, R. and Matias, C. (2001). Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7** 381–420. MR1836737

[9] Douc, R., Moulines, E. and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32** 2254–2304. MR2102510

[10] Dupuis, P. and Ellis, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York. MR1431744

[11] Fredkin, D. and Rice, J. (1987). Correlation functions of a function of a finite-state Markov process with application to channel kinetics. *Math. Biosci.* **87** 161–172. MR0929996

[12] Fuh, C.-D. (2006). Efficient likelihood estimation in state space models. *Ann. Statist.* **34** 2026–2068. MR2283726

[13] Fuh, C.-D. (2010). Reply to "On some problems in the article Efficient Likelihood Estimation in State Space Models" by Cheng-Der Fuh [*Ann. Statist.* **34** (2006) 2026–2068]. *Ann. Statist.* **38** 1282–1285. MR2604694

[14] Genon-Catalot, V. and Laredo, C. (2006). Leroux's method for general hidden Markov models. *Stochastic Process. Appl.* **116** 222–243. MR2197975

[15] Glynn, P. W. and Meyn, S. P. (1996). A Liapounov bound for solutions of the Poisson equation. *Ann. Probab.* **24** 916–931. MR1404536

[16] Glynn, P. W. and Ormoneit, D. (2002). Hoeffding's inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.* **56** 143–146. MR1881167

[17] Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *J. Finance* **42** 281–300.

[18] Jensen, J. L. (2010). On some problems in the article Efficient Likelihood Estimation in State Space Models by Cheng-Der Fuh [*Ann. Statist.* **34** (2006) 2026–2068]. *Ann. Statist.* **38** 1279–1281. MR2604693

[19] Juang, B. and Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics* **33** 251–272. MR1132665

[20] Kalashnikov, V. V. (1994). Regeneration and general Markov chains. *J. Appl. Math. Stochastic Anal.* **7** 357–371. MR1301706

[21] Le Gland, F. and Mevel, L. (2000). Basic properties of the projective product with application to products of column-allowable nonnegative matrices. *Math. Control Signals Systems* **13** 41–62. MR1742139

[22] Le Gland, F. and Mevel, L. (2000). Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems* **13** 63–93. MR1742140

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

[24] Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *J. Time Ser. Anal.* **26** 669–689. MR2188304

[25] Mamon, R. S. and Elliott, R. J. (2007). *Hidden Markov Models in Finance. International Series in Operations Research & Management Science* **104**. Springer, Berlin. MR2407726

[26] Marton, K. and Shields, P. C. (1994). The positive-divergence and blowing-up properties. *Israel J. Math.* **86** 331–348. MR1276142

[27] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

[28] Petrie, T. (1969). Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **40** 97–115. MR0239662

[29] Roberts, G. O. and Tweedie, R. L. (1999). Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Process. Appl.* **80** 211–229. MR1682243

[30] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[31] van Handel, R. (2009). The stability of conditional Markov processes and Markov chains in random environments. *Ann. Probab.* **37** 1876–1925. MR2561436

[32] Williams, D. (1991). *Probability With Martingales*. Cambridge Univ. Press, Cambridge. MR1155402

R. Douc
CITI/Télécom SudParis
9 rue Charles Fourier
91000 Evry
France
E-mail: randal.douc@it-sudparis.eu

E. Moulines
CNRS/LTCI/Télécom ParisTech
46 rue Barrault
75013 Paris
France
E-mail: eric.moulines@telecom-paristech.fr

J. Olsson
Center of Mathematical Sciences
Lund University
Box 118
SE-22100 Lund
Sweden
E-mail: jimmy@maths.lth.se

R. van Handel
Department of Operations Research and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: rvan@princeton.edu