Bernoulli **2**(3), 1996, 199–228

Inference in hidden Markov models I: Local asymptotic normality in the stationary case

PETER J. BICKEL[1]* and YA'ACOV RITOV[2]

[1]*Department of Statistics, University of California, Berkeley, CA 94270, USA*
[2]*Department of Statistics, Hebrew University, Jerusalem 91905, Israel*

Following up on work by Baum and Petrie published 30 years ago, we study likelihood-based methods in hidden Markov models, where the hiding mechanism can lead to continuous observations and is itself governed by a parametric model. We show that procedures essentially equivalent to maximum likelihood estimates are asymptotically normal as expected and consistent estimates of their variance can be constructed, so that the usual inferential procedures are asymptotically valid.

Keywords: geometric ergodicity; hidden Markov models; local asymptotic normality; maximum likelihood

1. Introduction and basic results

Hidden Markov models, that is stochastic point functions of finite Markov chains, have become important in a number of areas of application. These include, first and foremost, speech recognition (for an introduction and survey, see Rabiner 1989); the study of excitation periods in ion channels (for a survey, see Ball and Rice 1992), and models for heterogenous DNA sequences (Churchill 1992). The main focus of these efforts have been algorithms for the fitting of these models and, in particular (see Rabiner 1989), the implementation of likelihood-based methods. It is, in fact, not obvious that the likelihood can be computed in linear time, but that is the case. There has been comparatively little work on the study of the inferential properties of likelihood methods in these models. The notable exceptions to this are the papers of Baum and Petrie (1966), Petrie (1969) and, most recently, Leroux (1989; 1992). Concurrently with our work, Rydén (1994a; 1994b) has also pursued likelihood-based procedures in hidden Markov models.

Specifically, Baum and Petrie (1966) showed that, when observing a deterministic finite point function of a finite Markov chain, maximum likelihood estimates of the parameters of the model governing the chain are consistent and asymptotically normal. Leroux formulated hidden Markov models in the generality we shall present and established consistency of maximum likelihood estimates of both the parameters of the Markov chain and the conditional distribution of the observations given the Markov chain. Unlike the

* To whom correspondence should be addressed.

1350-7265

Baum–Petrie techniques, which were used for establishing both consistency and asymptotic normality, Leroux's approach, based on results of Furstenberg and Kesten (1960) and Kingman's (1976) subadditive ergodic theorem, appears incapable of giving results beyond consistency. On the other hand, we shall show, by adding a few essential ideas to the penetrating analysis of Baum and Petrie, that the log-likelihood for hidden Markov models obeys the local asymptotic normality (LAN) conditions of LeCam (see LeCam and Yang 1990, for instance). Hence, efficient analogues of maximum likelihood estimates can be constructed, and the information bound giving their asymptotic variance estimated. We shall also indicate how our results need to be strengthened to yield asymptotic efficiency of maximum likelihood estimates, when they are consistent. Consistency of maximum likelihood estimates can also be established with our methods but under conditions slightly stronger than those of Leroux (1992).

The paper is constructed as follows. In the rest of this section we formally introduce the models we consider, state our main assumptions and results, and further discuss the strengths and weaknesses of these as well as extensions and further questions, some of which we intend to pursue. In Section 2 we give without proof some lemmas needed to establish our main theorem, discuss the heuristic behind them, and give a proof of the theorem based on these lemmas. Finally, in Section 3 we state more lemmas and give the proofs of all the lemmas which may not immediately be derived from the work of Baum and Petrie or others.

Formally we assume that observations $(Y_1, \dots Y_n) \in \mathcal{Y}^n$, for some space $\mathcal{Y}$, are distributed according to $P_\vartheta^{(n)}, \vartheta \in \Theta$, where Θ is an open subset of R^p and described as follows:

(i) (Hidden chain.) We are given (but do not observe) a stationary ergodic Markov chain $X_1, \dots, X_n, \dots$ with states $\{1, \dots, K\}$, stationary initial probability $\pi_\vartheta(i), 1 \le i \le K$, and transition probability matrix $\|\alpha_\vartheta(i,j)\|_{K\times K}$.

(ii) (Y_i is a function of the present X_i and an external randomization only.) Given $X_1, \dots, X_n$, the Y_i are conditionally independent, and given X_i, Y_i is independent of $X_j, j \neq i$.

(iii) (Stationarity.) The conditional distribution of Y_i given X_i does not depend on i.

(iv) The conditional distribution of Y_i given $X_i = a$ are dominated by ν, a σ-finite measure for all i, a, ϑ. The conditional density is denoted by $g_\vartheta(\,\cdot\,|a)$.

We may then write the density of $(Y_1, \dots, Y_n)$ with respect to product measure $\nu^{(n)}$ as

$$g_\vartheta(y_1, \dots, y_n) = \sum_{(x_1,\dots,x_n)} f_\vartheta(x_1, \dots, x_n, y_1, \dots, y_n), \tag{1.1}$$

where

$$f_\vartheta(x_1, \dots, x_n, y_1, \dots, y_n) = \pi_\vartheta(x_1) \prod_{j=1}^{n-1} \alpha_\vartheta(x_j, x_{j+1}) \prod_{i=1}^{n} g_\vartheta(y_i|x_i) \tag{1.2}$$

is the joint density of $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ with respect to (counting measure)$^{(n)} \times \nu^{(\mathrm{n})}$. We denote the joint distribution of $(X_i, Y_i), 1 \le i < \infty$, by P_ϑ, a probability on $(\Omega, \mathcal{A})$, where Ω is the space of x, y sequences and $\mathcal{A}$ is the Borel σ-field.

This model, more or less given in Leroux (1989), is more general than it appears to be at first sight. It includes all situations where $Y_i = h(X_{i-j}; 1 \le j \le t, \epsilon_i, \vartheta)$, $1 \le i \le n$, where the ϵ_i are i.i.d. and independent of the Xs and t is fixed, since we can always take $(X_{1+i}, \ldots, X_{t+i})$, $i \ge 0$, as our hidden chain. We will need the following assumptions.

Assumption 1. *For all $\vartheta, a, b, \alpha_\vartheta(a,b) \ge \gamma(\vartheta) > 0$.*

Assumption 2. *For all a, b, the map $\vartheta \to \alpha_\vartheta(a,b)$ has three continuous derivatives. Hence so has $\vartheta \to \pi_\vartheta(a)$.*

Note that Assumptions 1 and 2 imply that for all ϑ_0 there exist $\delta > 0, \gamma(\vartheta_0) > 0$, such that

$$\inf\{\alpha_\vartheta(a,b) : |\vartheta - \vartheta_0| \le \delta\} \ge \gamma(\vartheta_0) \tag{1.3}$$

$$\inf\{\pi_\vartheta(a) : |\vartheta - \vartheta_0| \le \delta\} \ge \gamma(\vartheta_0). \tag{1.4}$$

Assumption 3. *The maps $\vartheta \to \nabla \log g_\vartheta(y|a)$ have three derivatives for all y, a. Further, for all ϑ_0 there exist $\delta > 0, \lambda > 0$, such that if*

$$q_{\vartheta_0}(y,\delta) \equiv \sup\{|\nabla \log g_\vartheta(y|a)| : a, |\vartheta - \vartheta_0| \le \delta\},$$

then

$$\mathrm{E}_{\vartheta_0} \exp[\lambda q_{\vartheta_0}(Y_1,\delta)] < \infty. \tag{1.5}$$

Assumption 4. *For all ϑ_0 there exist $\delta > 0, r > 32$, such that if*

$$\rho_{\vartheta_0}(y) = \sup\left\{\frac{g_\vartheta(y|a)}{g_\vartheta(y|b)} : a, b, |\vartheta - \vartheta_0| < \delta\right\},$$

then

$$\mathrm{E}_{\vartheta_0}\rho^r_{\vartheta_0}(Y_1) < \infty. \tag{1.6}$$

Assumption 5. *Let $\vartheta = (\vartheta_1, \ldots, \vartheta_p)$ and*

$$q_{\vartheta_0 j}(y,\delta) = \sup\left\{\left|\frac{\partial^j}{\partial\vartheta_{i_1}\ldots\partial\vartheta_{i_j}} \log g_\vartheta(y|a)\right|\right\},$$

where the supremum is taken over $\{1 \le i_l \le p, l = 1, \ldots, j, 1 \le a \le K, |\vartheta - \vartheta_0| \le \delta\}$. Assume, for all ϑ_0, some $\delta > 0, j = 2, 3$,

$$\mathrm{E}_{\vartheta_0}\{(q_{\vartheta_0 j}(Y_1,\delta))^{2+\delta}\} < \infty. \tag{1.7}$$

Let (X_i, Y_i), $-\infty < i < \infty$, be the two-sided stationary sequence defined by our model, and

$$W(Y_1, Y_0, Y_{-1}, \ldots) \equiv \sum_{m=-\infty}^{1} W_m(Y_1, Y_0, \ldots), \tag{1.8}$$

where

$$W_m(Y_1, Y_), \ldots) \equiv \mathrm{E}_{\vartheta_0}\{\nabla \log g(Y_m|X_m)|Y_1, Y_0, \ldots\} - \mathrm{E}_{\vartheta_0}\{\nabla \log g(Y_m|X_m)|Y_0, Y_{-1}, \ldots\}$$
$$+ \mathrm{E}_{\vartheta_0}\{\nabla \log \alpha(X_m, X_{m+1})|Y_1, Y_0, \ldots\} \tag{1.9}$$
$$- \mathrm{E}_{\vartheta_0}\{\nabla \log \alpha(X_m, X_{m+1})|Y_0, Y_{-1}, \ldots\}.$$

We show in Lemma 3.5 that, under Assumptions 1–4, $W \in L_2(P_{\vartheta_0})$, and we can then define

$$I(\vartheta_0) \equiv \mathrm{E}_{\vartheta_0}\{WW^{\mathrm{T}}\}. \tag{1.10}$$

Fix ϑ_0 and let $\mathcal{L}_0, P_0, E_0$ be law, probability and expectation under ϑ_0. Let $\delta_n \equiv n^{-1/2}$, $\vartheta_n \equiv \vartheta_0 + \tau\delta_n$, and

$$L_n(\tau) \equiv \frac{g_{\vartheta_n}}{g_{\vartheta_0}}(Y_1, \ldots, Y_n). \tag{1.11}$$

Our main goal is to establish the following theorem:

Theorem 1.1. *Suppose Assumptions 1–5 hold. Then there exist Δ_n, random p-vectors, such that if $|\tau_n| = O(1)$, then*

$$\log L_n(\tau_n) = \tau_n^{\mathrm{T}}\Delta_n - \tfrac{1}{2}\tau_n^{\mathrm{T}} J_n \tau_n + R_n(\tau_n), \tag{1.12}$$

where

$$E_0\Delta_n = 0, \tag{1.13}$$

$$E_0\Delta_n\Delta_n^{\mathrm{T}} \to I(\vartheta_0), \tag{1.14}$$

$$J_n \to I(\vartheta_0), \tag{1.15}$$

$$\Delta_n \xrightarrow{\mathcal{L}_0} \mathcal{N}(0, I(\vartheta_0)), \tag{1.16}$$

$P_0(|R_n(\tau_n)| < n^{-\gamma/2}/e_n) < \max\{e_n, n^{-1}\}$ for any $e_n \to 0$ and $\gamma < 2(1 - 16/r)/5$ for r satisfying (1.6), and $I(\vartheta_0)$ given in (1.10).

Note that (1.12) is just local asymptotic normality in the sense of Le Cam. In order to implement this result for inferential purposes we can proceed more or less as in Le Cam and Yang (1990, pp. 57–65). We need the following assumption:

Assumption 6. *The parameter ϑ is identifiable in the sense that if for some $\vartheta, \vartheta' \in \Theta$, $P_\vartheta^{(n)} = P_{\vartheta'}^{(n)}$ for all n, then $\vartheta = \vartheta'$.*

Lemma 1.1. *If Assumptions 1–6 hold, then there exists an estimate $\{\tilde{\vartheta}_n(Y_1, \ldots, Y_n)\}_{n \geq 1}$ which is $\sqrt{n}$ consistent. That is, for each $\vartheta_0, \tilde{\vartheta}_n - \vartheta_0 = O_{P_0}(\delta_n)$.*

Let the $\mathcal{G}_n$ grid denote the set of all $(\pm j_1, \ldots, \pm j_p)\delta_n n^{-\gamma/2p}$, where the j_i are integers and γ is as in Theorem 1.1. If Lemma 1.1 holds we can and shall, without loss of generality, suppose that $\tilde{\vartheta}_n$ takes on values in the $\mathcal{G}_n$ grid only. Let

$$\tilde{\vartheta}_n = \text{local maximizer of } g_\vartheta(Y_i, \ldots, Y_n) \text{ on } \mathcal{G}_n \tag{1.17}$$

closest to $\tilde{\vartheta}_n$ among the points of the δ_n grid, and, for given ϵ_n, define the matrix $\hat{I}_n$ by,

$$\hat{I}_{nab} = -\epsilon_n^{-2} \log \left\{ \frac{g_{\hat{\vartheta}_n(a,b)} g_{\hat{\vartheta}_n}}{g_{\hat{\vartheta}_n(a,0)} g_{\hat{\vartheta}_n(0,b)}} (Y_1, \ldots, Y_n) \right\} \tag{1.18}$$

$$\hat{\vartheta}_n(a,b) = \hat{\vartheta}_n + \epsilon_n \delta_n (e_a + e_b) \tag{1.19}$$

where $e_1, \ldots, e_p$ are the standard basis vectors and $e_0 = 0$. Thus, $\hat{\vartheta}_n$ is a grid version of the closest root of the likelihood equation to $\tilde{\vartheta}_n$ and $-\hat{I}_n$ is a second difference grid evaluated version of the Hessian at $\hat{\vartheta}_n$. Then we have the following corollary:

Corollary 1.1. *If Assumptions 1–6 hold, $\hat{\vartheta}_n$ is as in (1.17) and $I(\vartheta_0)$ is non-singular, then*

$$n^{1/2}(\hat{\vartheta}_n - \vartheta_0) \xrightarrow{\mathcal{L}_0} \mathcal{N}(0, I^{-1}(\vartheta_0)) \tag{1.20}$$

$$\hat{I}_n \xrightarrow{P_0} I(\vartheta_0). \tag{1.21}$$

We are now able to construct asymptotically efficient estimates, tests, etc., by pretending that $\hat{\vartheta}_n$ is approximately $\mathcal{N}(\vartheta, \delta_n^2 \hat{I}^{-1})$. This result does not give what one would ideally like:

(a) that the maximum likelihood estimator (MLE) $\hat{\vartheta}_n^*$ is asymptotically $\mathcal{N}(\vartheta_0, \delta_n^2 I^{-1}(\vartheta_0))$;
(b) that the Hessian of the log-likelihood at $\hat{\vartheta}_n^*, n^{-1} \|(\partial^2/\partial\vartheta_n \partial\vartheta_b) \log g_{\hat{\vartheta}_n^*}(Y_1, \ldots, Y_n)\|$ converges in probability to $-I(\vartheta_0)$.

Part (a) requires $\sqrt{n}$-consistency of the MLE and uniform (permitting τ_n to be data determined) LAN, while (b) requires consistency of the MLE and some sort of uniform convergence of the Hessian. These are open problems.

Discussion of assumptions

Evidently using f_ϑ and Bayes's rule we can construct maps from $\mathcal{Y}^n$ to {probabilities on $(\Omega, \mathcal{A})$}, $(y_1, \ldots, y_n) \to P_\vartheta(\cdot \mid y_1, \ldots, y_n)$ such that $P_\vartheta(\cdot \mid Y_1, \ldots, Y_n)$ is a regular conditional probability on Ω given $(Y_1, \ldots, Y_n)$. The key property in Baum and Petrie (1966) and our analysis is that $(X_1, X_2, \ldots)$ are an inhomogeneous Markov chain under $P_\vartheta(\cdot \mid y_1, y_2, \ldots)$. Assumptions 1, 2 and 4 guarantee that, with probability 1, this chain has strong geometric ergodicity properties which, among other things, guarantee the existence of $I(\vartheta_0)$ in (1.10). Assumptions 1 and 2 can easily be relaxed by specifying that only some power of the transition matrix needs to have all entries positive. Assumption 4 is clearly not very demanding. Assumption 3 intersects with Assumptions 1, 2 and 4, guaranteeing the validity of appropriate Taylor expansions. It is evidently a much stronger moment condition than is required for valid Taylor expansions in the i.i.d. case. However, we do not presently see how it can be relaxed. It evidently holds for Gaussian location and scale families, for instance, as does Assumption 5, which is essentially a standard condition of the Cramér type.

Extensions

Two extensions worth considering are:

(a) dropping the requirement that the state space of X be finite;
(b) the case where the hidden process is a Markov random field.

Extension (a) includes most nonlinear ARMA processes which have been proposed (see Priestley 1988, Tong 1991). Let $\ldots, \epsilon_{-1}, \epsilon_0, \epsilon_1, \ldots$ be an i.i.d. sequence of random variables with distribution from a parametric family, $\{F_\vartheta\}$, and

$$Y_j = h(\epsilon_j, \epsilon_{j-1}, \ldots, \vartheta), \qquad 1 \le j \le n. \tag{1.22}$$

Since $X_j = \{\epsilon_{j-k} : k \ge 0\}$ is a Markov chain on R^∞ this falls under case (a). For a discussion of Edgeworth expansions of smooth statistics in such models see Götze and Hipp (1992).

Estimation of parameters in hidden Markov fields by *ad hoc* methods has been considered by Frigessi and Piccioni (1990) and others. Likelihoods, even for directly observed fields, are only computable by simulation, but extension of our approach replacing likelihoods of the hidden process by pseudo-likelihoods may be valuable. See Qian and Titterington (1991).

We intend to pursue special cases of both extensions. It also appears that extensions to continuous-time situations where observations are not simply point functions of the hidden process may also be possible and interesting. A simple example discussed in Daley and Vere Jones (1989), and pursued by Rydén (1994b), is that of Cox processes driven by a finite-state continuous Markov process.

2. Proof of Theorem 1.1

We begin with an outline of our proof of Theorem 1.1. Details are given at the end of the section after the statement of some lemmas. Let $\mathbf{Y}_{a,b} = (Y_a, \ldots, Y_b)$ and $\mathbf{X}_{a,b}$ be the corresponding X block. Also define $\mathbf{Y}_m^{(k)} = \mathbf{Y}_{mk+1,mk+k}$ and $\mathbf{X}_n^{(k)}$ be the corresponding X block where $0 \le m \le N = n/k - 1$. To simplify the notation, we assume that n is a multiple of k. W argue in II below that if k does not divide n we can neglect the resulting end effect. For convenience we use the subscript τ in the following to stand for $\vartheta_n = \vartheta_0 + \tau_n \delta_n$, where $\{\tau_n\}$ is a bounded sequence. Let $\ell_\tau(\mathbf{Y}_m^{(k)} | X_{mk+1})$ denote the conditional likelihood of $\mathbf{Y}_m^{(k)}$ given X_{mk+1}, and let

$$L_{\tau m} \equiv \frac{\sum_{a=1}^{K} P_\tau[X_{mk+1} = a | \mathbf{Y}_{1,mk}] \ell_\tau(\mathbf{Y}_m^{(k)} | a)}{\sum_{a=1}^{K} P_0[X_{mk+1} = a | \mathbf{Y}_{1,mk}] \ell_0(\mathbf{Y}_m^{(k)} | a)} \tag{2.1}$$

denote the likelihood ratio of $\mathbf{Y}_m^{(k)}$ given $\mathbf{Y}_{1,mk}$. Also, let

$$L_{\tau m}^{(d)} \equiv \frac{\sum_{a=1}^{K} P_\tau[X_{mk+1}=a|\mathbf{Y}_{mk-d,mk}]\ell_\tau(\mathbf{Y}_m^{(k)}|a)}{\sum_{a=1}^{K} P_0[X_{mk+1}=a|\mathbf{Y}_{mk-d,mk}]\ell_0(\mathbf{Y}_m^{(k)}|a)}, \tag{2.2}$$

denote the likelihood ratio of $\mathbf{Y}_m^{(k)}$ given $\mathbf{Y}_{mk-d,mk}$, and

$$L_{\tau m}^{*} \equiv \frac{\ell_\tau(\mathbf{Y}_m^{(k)}|X_{mk+1})}{\ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1})} \tag{2.3}$$

the likelihood ratio of $\mathbf{Y}_m^{(k)}$ given X_{mk+1}.

I. Write

$$\log L_n(\tau) = \sum_{m=1}^{N} \log L_{\tau m} + \log \frac{g_{\vartheta_n}}{g_{\vartheta_0}}(Y_1, \ldots, Y_k) \tag{2.4}$$

and

$$\sum_{m=1}^{N} \log L_{\tau m} = \sum_{m=1}^{N} \log L_{\tau m}^{*} + \sum_{m=1}^{N} \log\left(1 + \frac{(L_{\tau m} - L_{\tau m}^{*})}{L_{\tau m}^{*}}\right). \tag{2.5}$$

Taylor expanding, we get

$$\begin{aligned}\sum_{m=1}^{N} \log\left(1 + \frac{(L_{\tau m} - L_{\tau m}^{*})}{L_{\tau m}^{*}}\right) &= \sum_{m=1}^{N}(L_{\tau m} - L_{\tau m}^{*}) - \sum_{m=1}^{N} \frac{(L_{\tau m} - L_{\tau m}^{*})}{L_{\tau m}^{*}}(L_{\tau m}^{*} - 1) \\ &\quad - \frac{1}{2}(1 + R_n) \sum_{m=1}^{N} \frac{(L_{\tau m} - L_{\tau m}^{*})^2}{(L_{\tau m}^{*})^2}. \end{aligned} \tag{2.6}$$

II. We expect $|L_{\tau m} - 1| = O_{P_0}(k/n)^{1/2}$. We shall establish this and, in so doing, also show that if $n = Nk + r, 0 < r < k$, then the difference between $\log L_n(\tau)$ and $\log L_{nk}(\tau)$ is $o_{P_0}(1)$. Further, $X_1, X_2, \ldots$ remains a Markov chain given the Ys. Although the chain is not stationary, it satisfies a strong mixing condition. Thus, we expect that the knowledge of Ys and Xs in the distant past adds very little information to the present and $|L_{\tau m} - L_{\tau m}^{*}| = o_{P_0}((k/n)^{1/2})$ so that we can and do show that the last two terms of (2.6) are negligible. The second term in (2.4) is also negligible. This uses arguments based on the Baum and Petrie (1966) results which are stated under our conditions in Lemmas 3.1–3.4.

III. We write the first term as

$$\sum_{m=1}^{N}(L_{\tau m} - L_{\tau m}^{*}) = \sum_{m=1}^{N}(L_{\tau m}^{(d)} - L_{\tau m}^{*}) + \sum_{m=1}^{N}(L_{\tau m} - L_{\tau m}^{(d)}). \tag{2.7}$$

We show that the second term is negligible for $d \to \infty, d = o(k)$ using Baum and Petrie again, and that the first term is negligible using uniform mixing and the Ibragimov–Linnik lemma (Lemma 3.7 below).

IV. We Taylor expand $\sum_{m=1}^{N} \log L_{tm}^*$ in τ and apply uniform mixing to show it has the LAN structure.

V. Finally, we evaluate $I(\vartheta_0)$ necessarily by a different starting formula than that of Baum and Petries, but again rely on their results to dispose of possible long-range dependence.

The proof of Theorem 1.1 is based on the following lemma whose proofs are given in the next section.

We adopt the following notation. We say

$$A_n = O_{b_n}(a_n) \tag{2.8}$$

if and only if there exists some $M_0, c(\cdot) \searrow 0$, such that for all $M > M_0$ and $n > n(M)$

$$P_0[|A_n| \geq Ma_n] \leq c(M)b_n.$$

In particular, $O_0(a_n) \equiv O(a_n)$ and $O_1(a_n) \equiv O_{P_0}(a_n)$.

Lmma 2.1. *If Assumptions 1–4 hold,* $r > 16, k = n^{4\epsilon+(5/4)\gamma}, \epsilon > 2/r, 4\epsilon + \gamma < 1/2, \gamma > 0$*, then for any* $|\tau| < M$*,*

$$\sum_{m=1}^{N} \log(L_{\tau m}/L_{\tau m}^*) = O_{e_n}(n^{-\gamma/2}/e_n) \tag{2.9}$$

for any $e_n \to 0, ne_n \to \infty$.

Lemma 2.2. *If Assumptions 1–5 hold,* $r > 32, k = n^{4\epsilon+\gamma}, 4\epsilon + \gamma < 1/4$*, then*

$$\begin{aligned} \mathrm{E}_0 \sup_{|\tau|<M} \Bigg\{ \sum_{m=1}^{N} \Bigg| & \log L_{\tau m}^* - \delta_n \tau^{\mathrm{T}} \nabla \log \ell_0(\mathbf{Y}_m^{(k)} | X_{mk+1}) \\ & - \frac{1}{2n} \tau^{\mathrm{T}} \left\| \frac{\partial^2}{\partial\vartheta_i \partial\vartheta_j} \log \ell_0(\mathbf{Y}_m^{(k)} | X_{mk+1}) \right\| \tau \Bigg| \Bigg\} = O(k^2/n^{1/2}), \end{aligned} \tag{2.10}$$

where $\|a_{ij}\|$ *is the matrix with entries* a_{ij}.

Lemma 2.3. *Under Assumptions 1–4*

$$\lim_{k\to\infty} \frac{1}{k} \mathrm{E}_0 \Big\{ \Big(\nabla \log \ell_0(\mathbf{Y}_0^{(k)} | X_1)\Big) \Big(\nabla \log \ell_0(\mathbf{Y}_0^{(k)} | X_1)\Big)^{\mathrm{T}} \Big\} = I(\vartheta_0) \tag{2.11}$$

where $I(\vartheta_0)$ *is defined as in (1.10).*

Lemma 2.4. *Under Assumptions 1–4, if $k = o(n)$, then*

$$\frac{1}{n}\sum_{m=1}^{N} \mathrm{E}_0\{\nabla\nabla^{\mathrm{T}} \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1})|X_{mk+1}\} \xrightarrow{P_0} I(\vartheta_0) \tag{2.12}$$

$$\frac{1}{n}\sum_{m=1}^{N} \mathrm{E}_0^{1/2}\{|\nabla \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1})|^4|X_{mk+1}\} = O_{P_0}(1) \tag{2.13}$$

$$\max_m P_0[|\delta_n \nabla \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1})| \geq \epsilon|X_{mk+1}] = o_{P_0}(1) \tag{2.14}$$

where $\nabla\nabla^T h \equiv (\nabla h)(\nabla h)^T$.

Lemma 2.5. *Under Assumptions 1–4,*

$$\frac{1}{n}\sum_{m=1}^{N} \left\| \frac{\partial^2}{\partial\vartheta_a \partial\vartheta_b} \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1}) \right\| \xrightarrow{P_0} -I(\vartheta_0). \tag{2.15}$$

Proof of Theorem 1.1. From Lemma 2.1 we see that if $\tau \equiv \tau_n$ we can replace the left-hand side of (2.5) by $\Sigma_{m=1}^{N} \log L_{\tau_n m}^* + O_{e_n}(n^{-2\gamma/5}/e_n)$ if $k = n^{4\epsilon+\gamma}$, $\epsilon > 2/r, 4\epsilon + \gamma < 1/4$.

Lemma 2.2 now guarantees that

$$\begin{aligned} \sum_{m=1}^{N} \log L_{\tau_n m}^* &= \delta_n \tau_n^{\mathrm{T}} \sum_{m=1}^{N} \nabla \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1}) \\ &\quad + \frac{1}{2n}\sum_{m=1}^{N} \tau_n^{\mathrm{T}} \left\| \frac{\partial^2 \log}{\partial\vartheta_i \partial\vartheta_j} \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1}) \right\| \tau_n \\ &\quad + O_{e_n}(n^{-1/2+8\epsilon+2\gamma}/e_n). \end{aligned} \tag{a}$$

Let

$$\xi_{mn} = \delta_n \tau_n^{\mathrm{T}} \nabla \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1}), \qquad 1 \leq m \leq N. \tag{b}$$

We claim that this is a triangular sequence of martingale summands with respect to the σ-fields $\mathcal{F}_{mn} = \sigma(\mathbf{X}_{1,(m+1)k+1}, \mathbf{Y}_{1,(m+1)k})$, $1 \leq m \leq N$. This follows from the Markov property which gives

$$\mathrm{E}_0\left\{\frac{\ell_0}{\ell_0}(\mathbf{Y}_m^{(k)}|X_{mk+1}|\mathcal{F}_{(m-1)n}\right\} = \mathrm{E}_0\left\{\frac{\ell_\vartheta}{\ell_0}(\mathbf{Y}_m^{(k)}|X_{mk+1})|X_{mk+1}\right\} \equiv 1 \tag{c}$$

and the usual interchange of differentiation and integration. Further, $I(\vartheta_0)$ is well defined and by (2.12),

$$\sum_{m=1}^{N} \mathrm{E}_0(\xi_{mn}^2|\mathcal{F}_{(m-1)n}) \xrightarrow{P_0} \tau_n^{\mathrm{T}} I(\vartheta_0)\tau_n, \tag{d}$$

and by Lemma 2.4,

$$\sum_{m=1}^{N} \mathrm{E}_0(\xi_{mn}^2 1(|\xi_{mn}| \geq \epsilon)|\mathcal{F}_{(m-1)n}) \tag{e}$$

$$\leq \left[\sum_{m=1}^{N} \mathrm{E}_0^{1/2}(\xi_{mn}^4|\mathcal{F}_{(m-1)n})\right] \max_{1\leq m\leq N} P_0^{1/2}[|\xi_{mn}| \geq \epsilon|\mathcal{F}_{(m-1)n}] = o_{P_0}(1).$$

The central limit theorem for triangular arrays of martingale summands (see Hall and Heyde 1980, for example) establishes that

$$\delta_n \tau^{\mathrm{T}} \sum_{m=1}^{N} \nabla \log \ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1}) \xrightarrow{\mathcal{L}_0} \mathcal{N}(0, \tau^{\mathrm{T}} I(\vartheta_0)\tau). \tag{f}$$

Finally, Lemma 2.5 establishes that the last term in (a) tends to $-\frac{1}{2}\tau^{\mathrm{T}} I(\vartheta_0)\tau$. The theorem is proved. □

Proof of Lemma 1.1. We construct a minimum distance estimator. The proof is based on Le Cam (1956). The construction is simple under the assumption that, for some $k < \infty$, the map $\vartheta \to P_\vartheta^{(k)}$ is one–one and Θ-compact. In that case it is possible to construct $\sqrt{n}$-consistent estimates by considering $P_n^{(k)}$, the empirical distribution of the vectors $\{Y_{a+b} : 0 \leq b \leq k-1\}$, for $1 \leq a \leq n-k+1$. See Rydén (1995) for a proof that $k = 2K$ under somewhat different conditions than ours, and Rydén (1994a) for the construction of the $\sqrt{n}$-consistent estimator. In general, let $\Theta = \cup_{j=1}^{\infty}\Theta_j$ with $\Theta_{j+1} \supset \Theta_j, j \geq 1$ compact sets, and define $T_{njk} = \{t \in \Theta_j : n^{-1/4} d_K(P_{\tilde{t}}^{(k)}, P_n^{(k)}) = \min_{\vartheta\in\Theta_j} d_K(P_\vartheta^{(k)}, P_n^{(k)})\}$, where $d_K(\cdot, \cdot)$ is the Kolmogorov distance. Then let $\tilde{\vartheta} \in T_n$, where $T_n = T_{njk}$ with T_{njk} non-empty and radius less than $n^{-1/4}$ and minimal $j + k$. □

Proof of Corollary 1.1. The corollary follows in a standard fashion by the methods of Le Cam (1986) and Le Cam and Yang (1990). Let $\mathcal{G}_{M_n} = \mathcal{G}_n \cap \{\vartheta : |\vartheta - \vartheta_0| < Mn^{-1/2}\}$. Note that there are $O(n^{\gamma/2})$ points in $\mathcal{G}_{M_n}$. Write $R_n = R_n(\tau)$ for the remainder term in (1.12). It follows from Theorem 1.1 that

$$P_0\left(\sup_{\tau n^{-1/2}\in\mathcal{G}_{M_n}} |L_n(\tau) - \tau\Delta_n + \tfrac{1}{2}\tau^{\mathrm{T}} J_n \tau| > \epsilon\right) \tag{a}$$

$$\leq O(n^{\gamma/2}) \sup_{\tau n^{-1/2}\in\mathcal{G}_{M_n}} P_0(|R_n(\tau)| > \epsilon) \tag{b}$$

$$\xrightarrow{P_0} 0.$$

Hence $\hat{\vartheta}_n$ is within distance $O_{P_0}(n^{(1+\gamma)/2})$ of

$$n^{-1/2} \arg\max\{\tau^{\mathrm{T}}\Delta_n - \tfrac{1}{2}\tau^{\mathrm{T}} J_n \tau\} = n^{-1/2} J_n^{-1} \Delta_n, \tag{c}$$

which proves Corollary 1.1. □

3. Further lemmas and proofs

We begin with four lemmas which are straightforward extensions of key results of Baum and Petrie (1966) (Lemma 2.1, Lemma 2.2 and Corollary 2.3) valid under Assumptions 1–3 and hence the proofs are omitted. They contain the essential information that knowledge of ys and xs in the distant past adds very little information to the present. Lemma 3.1 guarantees strong mixing conditions.

Let
$$\mu_0(y) = (1 + (K-1)\gamma^{-2}(\vartheta_0)\rho_{\vartheta_0}(y))^{-1}.$$
In what follows we write $P_\vartheta(A|B, y_1, \ldots, y_n)$ if $P_\vartheta(A|B, Y_1, \ldots, Y_n)$ is a version of the regular conditional probability of A given B, $Y_1, \ldots, Y_n$, and $P\vartheta(A|B, y_1, \ldots, y_n)$ is defined for all ϑ, A, B and $y_1, \ldots, y_n$. This is easily done if we can define densities $g_\vartheta(y|x)$ valid for all ϑ, y and x.

Lemma 3.1. *For $|\vartheta - \vartheta_0| \le \delta$ and all ϑ_0,*
$$P_\vartheta[X_{i+1} = b | X_i = a, y_1, \ldots, y_n] \ge \mu_0(y_{i+1}) > 0. \tag{3.1}$$

Lemma 3.2. *If C_t is an event depending only on $X_i, Y_i, i \ge t$, then for all $|\vartheta - \vartheta_0| \le \delta, \vartheta_0, d \ge 2$,*
$$\begin{aligned} &|P_\vartheta[C_t|y_{t-1}, \ldots, y_{t-d+1}] - P_\vartheta[C_t|y_{t-1}, \ldots, y_{t-d}]| \\ &\le \prod_{j=t-d+1}^{t-1} (1 - 2\mu_0(y_j)) \le \exp\left\{-2 \sum_{j=t-d+1}^{t-1} \mu_0(y_j)\right\}. \end{aligned}$$

Lemma 3.3. *Let C_t be as above, let*
$$M_d^+(\vartheta) = \max_a P_\vartheta[C_t | y_1, \ldots, y_n, X_{t-d} = a],$$
and define $M_d^-(\vartheta)$ as the corresponding minimum. Then, for all $\vartheta_0, |\vartheta - \vartheta_0| \le \delta$,
$$|M_d^+(\vartheta) - M_d^-(\vartheta)| \le \prod_{j=t-d+1}^{t-1} (1 - 2\mu_0(y_j)). \tag{3.2}$$

Lemma 3.4. *If assumptions 1 and 2 hold, then for all $\vartheta_0, |\vartheta - \vartheta_0| \le \delta, y_1, \ldots y_\ell, a, b$,*
$$P_\vartheta[X_{\ell+1} = a | y_1, \ldots, y_\ell, X_1 = b] \ge \gamma(\vartheta_0). \tag{3.3}$$

The following two lemmas are of general utility in missing-data models.

Lemma 3.5. *If $P \gg Q, \mathrm{e}^\Lambda \equiv \mathrm{d}Q/dP, T \in L_I(Q)$, and $\mathbf{B}$ is a sub-σ-field, then*
$$\mathrm{E}_P|\mathrm{E}_Q(T|\mathbf{B})| \le \mathrm{E}_P^{1/r}\{|T|^r\}\mathrm{E}_P^{1/s}\{\mathrm{e}^{s\Lambda}\}\mathrm{E}_P^{1/t}\{\mathrm{e}^{-t\Lambda}\}, \tag{3.4}$$
where $1/r + 1/s + 1/t = 1$.

Proof of Lemma 3.5. Note that

$$\mathrm{E}_Q(T|\mathbf{B}) = \frac{\mathrm{E}_P(T\mathrm{e}^{\Lambda}|\mathbf{B})}{\mathrm{E}_P(\mathrm{e}^{\Lambda}|\mathbf{B})}. \tag{a}$$

So (3.4) is bounded by

$$\mathrm{E}_P|\mathrm{E}_P(T\mathrm{e}^{\Lambda}|\mathbf{B})\mathrm{E}_P(\mathrm{e}^{-\Lambda}|\mathbf{B})| \le \mathrm{E}_P\{|T|\mathrm{e}^{\Lambda}\mathrm{E}_P(\mathrm{e}^{-\Lambda}|\mathbf{B})\} \le \mathrm{E}_P^{1/r}\{|T|^r\}\mathrm{E}_P^{1/s}\{\mathrm{e}^{s\Lambda}\}\mathrm{E}_P^{1/t}\{\mathrm{e}^{-t\Lambda}\}. \tag{b}$$

$\square$

Lemma 3.6. *Let $\vartheta \to U_\vartheta, \vartheta \in R$, be continuously differentiable, where $U_\vartheta(\cdot)$ is a stochastic process on $(\Omega, \mathcal{A})$, $\mathbf{B}$ is a sub-field of $\mathcal{A}$. Then, if $P_\vartheta \ll \nu$ and $\ell_\vartheta \equiv \mathrm{d}P_\vartheta/\mathrm{d}\nu$, suppose*

(i) $\vartheta \to \dfrac{\partial}{\partial\vartheta}\log \ell_\vartheta$

(ii) $\vartheta \to \mathrm{E}_\vartheta\left|\dfrac{\partial U_\vartheta}{\partial\vartheta}\right|$

(iii) $\vartheta \to \mathrm{E}_\vartheta[U_\vartheta^2]$

(iv) $\vartheta \to \mathrm{E}_\vartheta\left(\dfrac{\partial}{\partial\vartheta}\log \ell_\vartheta\right)^2$

are all continuous. Then,

$$\frac{\partial}{\partial\vartheta}\mathrm{E}_\vartheta(U_\vartheta|\mathbf{B}) = \mathrm{E}_\vartheta\left(\frac{\partial U_\vartheta}{\partial\vartheta}\middle|\mathbf{B}\right) + \mathrm{cov}_\vartheta\left\{\left(U_\vartheta, \frac{\partial}{\partial\vartheta}\log \ell_\vartheta\right)\middle|\mathbf{B}\right\}. \tag{3.5}$$

Proof of Lemma 3.6. Write $\Lambda(\vartheta, \vartheta+\Delta) = \log(\ell_{\vartheta+\Delta}/\ell_\vartheta)$,

$$\mathrm{E}_{\vartheta+\Delta}(U_{\vartheta+\Delta}|\mathbf{B}) = \frac{\mathrm{E}_\vartheta(U_{\vartheta+\Delta}\mathrm{e}^{\Lambda(\vartheta,\vartheta+\Delta)}|\mathbf{B})}{\mathrm{E}_\vartheta(\mathrm{e}^{\Lambda(\vartheta,\vartheta+\Delta)}|\mathbf{B})}. \tag{a}$$

Then

$$\frac{\partial}{\partial\vartheta}\mathrm{E}_\vartheta(U_\vartheta|\mathbf{B}) = \frac{\partial}{\partial\Delta}\mathrm{E}_\vartheta(U_{\vartheta+\Delta}\mathrm{e}^{\Lambda(\vartheta,\vartheta+\Delta)}|\mathbf{B})|_{\Delta=0} - \mathrm{E}_\vartheta(U_\vartheta|\mathbf{B})\frac{\partial}{\partial\Delta}\mathrm{E}_\vartheta(\mathrm{e}^{\Lambda(\vartheta,\vartheta+\Delta)}|\mathbf{B})|_{\Delta=0}, \tag{b}$$

provided the right-hand side exists. Interchange of integration and differentiation may be justified under our condition by a delicate but standard argument we do not reproduce. We get that the right-hand side of (b) is

$$\mathrm{E}_\vartheta\left(\frac{\partial U_\vartheta}{\partial\vartheta}\middle|\mathbf{B}\right) + \mathrm{E}_\vartheta\left(U_\vartheta\frac{\partial}{\partial\vartheta}\log \ell_\vartheta\middle|\mathbf{B}\right) - \mathrm{E}_\vartheta(U_\vartheta|\mathbf{B})\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\log \ell_\vartheta\middle|\mathbf{B}\right), \tag{c}$$

and (3.5) follows. $\square$

We also need a basic lemma (Theorem 17.2.2) from Ibragimov and Linnik (1971 p. 307), which we quote for completeness.

Lemma 3.7. *If* ξ, η *have joint distribution* P *with marginals* P_1, P_2 *such that* $\|P - (P_1 \times P_2)\| \leq \alpha$, *where* $\|\cdot\|$ *is variational distance, and, for some* $\delta > 0$, *and* $\mathrm{E}|\xi|^{2+\delta} \leq c_1, \mathrm{E}|\eta|^{2+\delta} \leq c_2$, *then*

$$|\mathrm{E}(\xi\eta) - \mathrm{E}(\xi)\mathrm{E}(\eta)| \leq c\alpha^{1-\beta}, \tag{3.6}$$

where $\beta = 2/(2+\delta)$ *and* $c = 4 + 3c_1^{\beta/2}c_2^{1-\beta/2} + 3c_1^{1-\beta/2}c_2^{\beta/2}$.

Here are the additional lemmas we need to carry out I–V from Section 2. Let

$$\alpha_{\tau,i,m}(a,b) \equiv P_\tau[X_{i+1} = b | X_i = a, Y_1, \ldots, Y_m]. \tag{3.7}$$

Lemma 3.8. *In our model, if* $1 \leq i \leq m-1$,

$$\frac{\alpha_{\tau,i,m}(a,b)}{\alpha_{0,i,m}(a,b)} = \frac{\mathrm{E}_0\left\{\dfrac{f_\tau}{f_0}(\mathbf{X}_{1,m}, \mathbf{Y}_{1,M}) | X_i = a, X_{i+1} = b, \mathbf{Y}_{1,m}\right\}}{\mathrm{E}_0\left\{\dfrac{f_\tau}{f_0}(\mathbf{X}_{1,m}, \mathbf{Y}_{1,n}) | X_i = 1, \mathbf{Y}_{1,m}\right\}}. \tag{3.8}$$

Proof of Lemma 3.8. Note that

$$P_\tau[X_{i+1} = b, X_i = a | \mathbf{Y}_{1.m}] = \mathrm{E}_0 \frac{\left\{\dfrac{f_\tau}{f_0}(\mathbf{X}_{1,m}, \mathbf{Y}_{1,m}) 1(X_{i+1} = b, X_i = a) | \mathbf{Y}_{1,m})\right\}}{\mathrm{E}_0\left\{\dfrac{f_\tau}{f_0}(\mathbf{X}_{1.m}, \mathbf{Y}_{1,m}) | \mathbf{Y}_{1.m}\right\}} \tag{a}$$

$$P_\tau[X_i = a | \mathbf{Y}_{1.m}] = \frac{\mathrm{E}_0\left\{\dfrac{f_\tau}{f_0}(\mathbf{X}_{1,m}, \mathbf{Y}_{1,m}) 1(X_i = a) | \mathbf{Y}_{1,m}\right\}}{\mathrm{E}_0\left\{\dfrac{f_\tau}{f_0}(\mathbf{X}_{1.m}, \mathbf{Y}_{1,m}) | \mathbf{Y}_{1,m}\right\}} \tag{b}$$

$$\mathrm{E}_0\left\{\frac{f_\tau}{f_0}(\mathbf{X}_{1,m}, \mathbf{Y}_{1,m}) 1(X_i = a) | \mathbf{Y}_{1.m}\right\} = \mathrm{E}_0\left\{\frac{f_\tau}{f_0}(\mathbf{X}_{1,m}, \mathbf{Y}_{1,m}) | X_i = 1, \mathbf{Y}_{1,m}\right\} P_0[X_i = a | \mathbf{Y}_{1,m}]. \tag{c}$$

Substitute (a), (b) on the left-hand side of (3.8) and simplify using (c) and an analogous expression for the numerator in (a) to get the right-hand side. □

Let

$$S_n \equiv \{(a, b, i, m, \tau) : m - i \leq d_n, 1 \leq m \leq n, |\tau| \leq M\}$$

and

$$\mathrm{E}_{0m}(\cdot) \equiv \mathrm{E}_0(\cdot | \mathbf{Y}_{1,m}), P_{\tau m}(\cdot) \equiv P_\tau(\cdot | \mathbf{Y}_{1,m}), \quad \text{etc.}$$

Lemma 3.9. *Suppose Assumptions 1, 3 and 4 hold and*

$$d_n = o(n^{1/2}/\log n). \tag{3.9}$$

Then

$$P_0\left[\inf_{S_n} \mathrm{E}_{0m}\left\{\frac{f_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{i,m}) | X_i = a, X_{i+1} = b\right\} \geq \frac{1}{2}\right] = 1 - o(n^{-1}). \tag{3.10}$$

Proof of Lemma 3.9. From (1.2), if $|\tau| \leq M$,

$$\frac{f_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{i,m}) \geq \left(\inf_{c,d}\frac{\alpha_\tau}{\alpha_0}(c, d)\right)^{m-i+1} \inf_c \frac{\pi_\tau}{\pi_0}(c) \exp\left\{-M\delta_n \sum_{j=i}^m q_0(Y_j, M\delta_n)\right\}. \tag{a}$$

By Assumptions 1 and 2, if $|\tau| \leq M$ then the first two terms are larger than $(1 - r\tau)^{m-i+1}$ for a fixed $r = r(M) < \infty$, so that

$$\begin{aligned}&\inf_{S_n} \mathrm{E}_{0m}\left\{\frac{F_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{j,m}) | X_i = a, X_{i+1} = b\right\}\\ &\geq (1 + o(1)) \exp\{(-(d_n + 1)M\delta_n \max_{1\leq j\leq n} q_0(Y_j, M\delta_n)\}.\end{aligned} \tag{b}$$

But by (3.9) and Assumption 3, for some $\lambda > 0$,

$$\begin{aligned}&P_0\left[\max_{1\leq j\leq n} q_0(Y_j, M\delta_n) \geq (\log 2)/Md_n\delta_n\right]\\ &\leq nP_0[q_0(Y_1, M\delta_n) \geq (\log 2)/Md_n\delta_n]\\ &\leq n \exp\{-\lambda(\log 2/M)c_n \log n\}\mathrm{E}_0 \mathrm{e}^{\lambda q_0(Y_1, M\delta_n)},\end{aligned} \tag{c}$$

where $c_n \to \infty$ and (3.10) follows. □

Lemma 3.10. *Suppose Assumptions 1–4 hold and $\epsilon > 2/r$. Suppose $d_n \to \infty$, $d_n = o(n^{1/2}/(log\, n)^2)$. Then*

$$\sup_{S_n}\left|\frac{\alpha_{\tau,i,m}}{\alpha_{0,i,m}}(a, b) - 1\right| = O_{1/n}(n^{-1/2+\epsilon}). \tag{3.11}$$

Proof of Lemma 3.10. By Lemmas 3.8 and 3.9 it is enough to show that

$$\begin{aligned}&\sup_{S_n}\left\{\left|\mathrm{E}_{0m}\left(\frac{f_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{i,m}) | X_i = a, X_{i+1} = b\right)\right.\right.\\ &\left.\left.\quad - \mathrm{E}_{0m}\left(\frac{f_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{i,m}) | X_i = a, X_{i+1} = c\right)\right|\right\}\\ &= O_{n^{-1}}(n^{-1/2+\epsilon}).\end{aligned} \tag{a}$$

Consider the three Markov chains $X'_{i+1}, \ldots, X'_m; X''_{i+1}, \ldots, X''_m; X'''_{i+1}, \ldots, X'''_m$. Here $\{X'_j\}$ and $\{X''_j\}$ are independent, both with transition probabilities $\alpha_{o,j,m}$ from j to $j+1, i \le j \le m$, with $Y_{i,m}$ held fixed, $X'_i = X''_{i+1} = a, X'_{i+1} = b, X''_{i+1} = c$. Also $X'''_\ell = X''_\ell 1(\ell \le T) + X'_\ell 1(\ell > T)$, where $T = \min\{\ell : X'_\ell = X''_\ell, i < \ell \le m\} \wedge m$. Note that

$$\{X''_\ell : i \le \ell \le T\} \quad \text{and} \quad \{X'''_\ell : i \le \ell \le T\} \text{ have the same distribution.} \tag{b}$$

Further, if E_{0m}, P_{0m} now refer to probabilities on the space on which the data and the X'_j, X''_j, X'''_j are defined.

$$\left|\mathrm{E}_{0m}\left(\frac{f_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{i,m}) | X_i = a, X_{i+1} = b\right) - \mathrm{E}_{0m}\left(\frac{f_\tau}{f_0}(\mathbf{X}_{i,m}, \mathbf{Y}_{i,m}) | X_i = a, X_{i+1} = c\right)\right| \tag{c}$$

$$\begin{aligned} &= \left|\mathrm{E}_{0m}\left(\frac{f_\tau}{f_0}(\mathbf{X}'_{i,m}, \mathbf{Y}_{i,m}) - \frac{f_\tau}{f_0}(\mathbf{X}''_{i,m}, \mathbf{Y}_{i,,})\right)\right| \\ &= \left|\mathrm{E}_{0m}\left[\left(\frac{f_\tau}{f_0}(\mathbf{X}'_{i,T}, \mathbf{Y}_{i,T}) - \frac{f_\tau}{f_0}(\mathbf{X}'''_{i,T}, \mathbf{Y}_{i,T})\right)\right.\right. \\ &\quad \left.\left. \cdot \frac{\pi_0}{\pi_\tau}(X'_{T+1}) \frac{f_\tau}{f_0}(X'_{T+1}, \ldots, X'_m, Y_{T+1}, \ldots, Y_m) \frac{\alpha_\tau}{\alpha_0}(X'_T, X'_{T+1})\right]\right|. \end{aligned} \tag{d}$$

By Assumptions 1 and 2, for $|\tau| \le M$, d_n as above, there exists $c(M) < \infty$ such that, if $A_n \equiv \max\{q_0(Y_j, M\delta_n) : 1 \le j \le n\}$,

$$\exp\{-\delta_n(T-i)(MA_n + c)\} \le \frac{f_\tau}{f_0}(\mathbf{X}'_{i,T}, \mathbf{Y}_{i,T}) \le \exp\{\delta_n(T-i)(MA_n + c)\}. \tag{e}$$

The same holds if $\mathbf{X}'_{i,T}$ is replaced by $\mathbf{X}'''_{i,T}$ and also

$$\frac{f_\tau}{f_0}(\mathbf{X}'_{T+1,m}, \mathbf{Y}_{T+1,m}) \le \exp\{\delta_n d_n(MA_n + c)\}. \tag{f}$$

By Assumption 3 and (c) of the proof of Lemma 3.9,

$$A_n = O_{n^{-1}}((\log n)^2). \tag{g}$$

Then, from (d), (e), (f), and (g), (a) follows if

$$\sup_{S_n}\{\mathrm{E}_{0m}(\mathrm{e}^{(T-i)a_n} - \mathrm{e}^{-(T-i)a_n})\} = O_{n^{-1}}(n^{-1/2+\epsilon}) \tag{h}$$

for

$$a_n = O(\delta_n(\log n)^2). \tag{i}$$

Now,

$$P_{0m}[T > i + t] \le \prod_{j=1+1}^{i+t}(1 - K\mu_0^2(Y_j)), \tag{j}$$

since for $j \geq i$

$$P_{0m}[X'_{j+1} = X''_{j+1} | X'_j = a, X''_j = b] = \sum_c \alpha_{0,j,m}(a,c)\alpha_{0,i,m}(b,c) \geq K\mu_0^2(Y_{j+1}) \tag{k}$$

by Lemma 3.1. But, by Assumption 4

$$P_0\left[\min_{1\leq j\leq n}\{K\mu_0^2(Y_j)\} \leq b_n\right] = P_0\left[\max_{1\leq j\leq n}\{\rho_0(Y_j)\} \geq \frac{\gamma^2((K/b_n)^{1/2}-1)}{K-1}\right] = o(n^{-1}) \tag{l}$$

if

$$b_n = o(n^{-2/r}). \tag{m}$$

Note that for any integer-valued random variable $N \geq 1$

$$\mathrm{E}a^N = a + \sum_{t=1}^{\infty}(a^{t+1} - a^t)P[N > t] \tag{n}$$

From (j), (l), (n), if $b_n = o(n^{-2/r}), b_n n^{\epsilon}/(\log n)^2 \to \infty$, then $a_n = o(b_n)$ and, with probability $1 - o(n^{-1})$,

$$\begin{aligned}
&\max\{\mathrm{E}_{0m}(\mathrm{e}^{(T-i)a_n} - \mathrm{e}^{-(T-i)a_n}) : m - i \leq d_n, 1 \leq m \leq n\} \\
&\leq \mathrm{e}^{a_n} - \mathrm{e}^{-a_n} + \sum_{t=1}^{\infty}(\mathrm{e}^{a_n} - 1)\mathrm{e}^{t(a_n - b_n)} \\
&= \mathrm{e}^{a_n} - \mathrm{e}^{-a_n} + (\mathrm{e}^{a_n} - 1)\mathrm{e}^{a_n - b_n}(1 - \mathrm{e}^{(a_n - b_n)})^{-1} \\
&= O(a_n(b_n - a_n)^{-1}) \\
&= O(a_n b_n^{-1})
\end{aligned}$$

and (a) follows from (h). □

Lemma 3.11. *If Assumptions 1–4 hold, $\epsilon > 2/r$, then*

$$\sup_{S_n} |P_{\tau m}[X_m = a] - P_{0m}[X_m = a]| = O_{1/n}(n^{-1/2+2\epsilon}). \tag{3.12}$$

Proof of Lemma 3.11. For fixed a let $V_{\tau,\ell,m} \in R^K$ be the column vector with coordinates:

$$V_{\tau,\ell,m}(\cdot)P_{\tau m}[X_m = a | X_\ell = \cdot], \qquad \ell \leq m. \tag{a}$$

Then,

$$V_{\tau,\ell,m} = \alpha_{\tau,\ell,m} \ldots \alpha_{\tau,m-1,m} V_{\tau,m,m}. \tag{b}$$

By Lemma 3.3

$$\sup\{|V_{\tau,\ell,m}(b) - V_{\tau,\ell,m}(c)| : b, c, |\tau| < M\} \leq \prod_{j=\ell+1}^{m-1}(1 - 2\mu_0(Y_j)) \leq \mathrm{e}^{-(m-\ell-1)B_n}, \tag{c}$$

where

$$B_n = 2 \min_{1 \le j \le n} \mu_0(Y_j). \tag{d}$$

Then

$$\sup\{|V_{\tau,\ell,m}(b) - V_{\tau,\ell,m}(c)|e^{(m-\ell-1)b_n} : b, c, |\tau|M, \ell \le m\} = O_{n^{-1}}(1) \tag{e}$$

if $b_n = o(n^{-2/r})$, by arguing as in (1) of Lemma 3.10. Therefore, if $c_{\tau,\ell,m} = K^{-1}\Sigma_b V_{\tau,\ell,m}(b)$ then

$$\sup\{\|V_{\tau,\ell,m} - c_{\tau,\ell,m}\mathbf{1}\|e^{b_n(m-\ell-1)} : m, \ell, |\tau| \le M\} = O_{n^{-1}}(1), \tag{f}$$

where $\|\cdot\|$ is the L_∞ on R^k and $\mathbf{1}$ is the vector of 1s. Then from (b),

$$\begin{aligned}\|V_{\tau,\ell,m} - V_{0,\ell,m}\| &= \|\alpha_{\tau,\ell,m}V_{\tau,\ell+1,m} - \alpha_{0,\ell,m}V_{0,\ell+1,m}\| \\ &\le \|(\alpha_{\tau,\ell,m} - \alpha_{0,\ell,m})V_{\tau,\ell+1,m}\| + \|V_{\tau,\ell+1,m} - V_{0,\ell+1,m}\|. \end{aligned} \tag{g}$$

Further, from Lemma 3.10, if $m - \ell = o(n^{1/2}/(\log n)^2), b_n = o(n^{-2/r})$, then

$$\|(\alpha_{\tau,\ell,m} - \alpha_{0,\ell,m})V_{\tau,\ell_{+1,m}}\| \le \rho_n e^{-(m-\ell-1)b_n}, \tag{h}$$

where $\rho_n = O_{n^{-1}}(c_n), c_n = n^{-1/2+\epsilon}$, since

$$(\alpha_{\tau,\ell,m} - \alpha_{0,\ell,m})\mathbf{1} = 0. \tag{i}$$

Iterating (g) and using (h), we get, if $d_n = o(n^{1/2}/(\log n)^2), b_n = o(n^{-2/r})$,

$$\sup\{\|V_{\tau,\ell,m} - V_{0,\ell,m}\| : m - \ell \le d_n, |\tau| \le M\} = O_{n^{-1}}(c_n b_n^{-1}). \tag{j}$$

Finally,

$$\begin{aligned} &|P_{\tau m}[X_m = a] - P_{0m}[X_m = a]| \\ &= \left|\sum_b \{P_{\tau m}[X_\ell - b]V_{\tau,l,m}(b) - P_{0m}[X_\ell = b]V_{0,\ell,m}(b)\}\right| \\ &\le \left|\sum_b (P_{\tau m}[X_\ell = b] - P_{0m}[X_\ell = b])V_{\tau,\ell,m}(b)\right| + \|V_{\tau,\ell,m} - V_{0,\ell,m}\|. \end{aligned} \tag{k}$$

By (f) the first term in (k) is, if $m - \ell \ge d_n$, equal to $O_{n^{-1}}(e^{-d_n b_n})$. If we use (j) and put $b_n = n^{-\epsilon}, d_n = n^\epsilon(\log n)^2$, the lemma follows. □

Lemma 3.12. *Under Assumptions 1–4, if* $k = o(n^{1/2-\gamma})$, *for some* $\gamma > 0$,

$$\sup\left\{\left|\frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|X_{mk+1}) - 1\right| : |\tau| \le M, 1 \le m \le N\right\} = O_{1,n}(n^{-\gamma/2}) \tag{3.13}$$

Proof of Lemma 3.12. Note that for any $p > 1$,

$$\begin{aligned}
&\mathrm{E}_0 \sup\left\{\left|\frac{\ell_\tau}{\ell_0}(\mathbf{Y}_0^{(k)}|X_1) - 1\right|^p : |\tau| \le M\right\} \\
&\quad = \mathrm{E}_0 \sup\left\{\left|\mathrm{E}_0\left[\left(\frac{\pi_0}{\pi_\tau}(X_1)\frac{f_\tau}{f_0}(\mathbf{X}_0^{(k)}, \mathbf{Y}_0^{(k)}) - 1\right)|X_1, \mathbf{Y}_0^{(k)}\right]\right|^p : |\tau| \le M\right\} \qquad \text{(a)} \\
&\quad \le (1 + o(1))\mathrm{E}_0 \sup\left\{\left|\frac{f_\tau}{f_0}(\mathbf{X}_0^{(k)}, \mathbf{Y}_0^{(k)}) - 1\right|^p : |\tau| \le M\right\} + o(1).
\end{aligned}$$

But, for any differentiable function $A(\vartheta)$ with $A(0) = 0$,

$$\sup\left\{\left|\mathrm{e}^{A(\vartheta)} - 1\right|, |\vartheta| \le \delta\right\} \le \delta \sup\{|A'(\vartheta)|\mathrm{e}^{A(\vartheta)} : |\vartheta| \le \delta\} \le \delta M_\delta \mathrm{e}^{\delta M_\delta}, \qquad \text{(b)}$$

where $M_\delta \equiv \sup\{|A'(\vartheta)| : |\vartheta| \le \delta\}$. We conclude that

$$\begin{aligned}
&\mathrm{E}_0 \sup\left\{\left|\frac{f_\tau}{f_0}(\mathbf{X}_0^{(k)}, \mathbf{Y}_0^{(k)}) - 1\right|^p : |\tau| \le M\right\} \\
&\quad \le (M\delta_n)^p \mathrm{E}_0\left\{\left(\sum_{j=1}^k \tilde{q}(Y_j, M\delta_n)\right)^p \exp\left[pM\delta_n \sum_{j=1}^k \tilde{q}(Y_j, M\delta_n)\right]\right\}, \qquad \text{(c)}
\end{aligned}$$

where

$$\tilde{q}(y, \delta) = q(y, \delta) + \sup\{|\nabla \log \alpha_\vartheta(a, b)| : |\vartheta - \vartheta_0| < \delta, a, b\}. \qquad \text{(d)}$$

Bound the right-hand side of (c) by

$$(M\delta_n)^p \mathrm{E}_0^{1/(1+\epsilon)}\left\{\left(\sum_{j=1}^k \tilde{q}(Y_j, M\delta_n)\right)^{p(1+\epsilon)}\right\}\mathrm{E}_0^{\epsilon/(1+\epsilon)}\left\{\exp\left[\frac{p(1+\epsilon)M\delta_n}{\epsilon}\sum_{j=1}^k \tilde{q}(Y_j, M\delta_n)\right]\right\}. \qquad \text{(e)}$$

The second term in (e) is bounded by

$$k^p[\mathrm{E}_0\{\tilde{q}^{p(1+\epsilon)}(Y_1, M\delta_n)\}]^{1/(1+\epsilon)}; \qquad \text{(f)}$$

use Assumption 3 to bound the third by

$$\begin{aligned}
&\left[\max_a \mathrm{E}_0\left\{\exp\left(\frac{p(1+\epsilon)}{\epsilon}M\delta_n q(Y_1, M\delta_n)\right)\middle| X_1 = \alpha\right\}\right]^{k\epsilon/(1+\epsilon)} \qquad \text{(g)} \\
&\quad = \left(1 + \frac{O(1)}{\sqrt{n}}\right)^{k\epsilon/(1+\epsilon)} = 1 + o(1)
\end{aligned}$$

since $k = o(n^{1/2})$ and $\delta_n \to 0$. Therefore,

$$P_0\left[\sup_{1 \le m\tau \le M \le N}\left\{\left|\frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|X_{mk+1}) - 1\right|\right\} \ge n^{-\gamma/2}\right] \le O(1)\frac{n}{k}(k\delta_n)^p n^{p\gamma/2} = o(n^{-1}) \qquad \text{(h)}$$

if $k = O(n^{1/2-\gamma}), p > 2 + 3/\gamma$. □

Lemma 3.13. *Under Assumptions 1–4, if* $k = o(n^{1/2-\gamma})$, *for some* $\gamma > 0$ *and* $\epsilon > 2/r$

$$\sup\left\{\frac{|L_{\tau m} - L^*_{\tau m}|}{L^*_{\tau m}} : 1 \le m \le N, |\tau| \le M\right\} = O_{1/n}(n^{-1/2+2\epsilon}). \tag{3.14}$$

Proof of Lemma 3.13. By (3.1)

$$\min_a\left\{\frac{\ell_\tau}{\ell_0}(\mathbf{Y}^{(k)}_m|a)B_\tau(\mathbf{Y}_{1,mk})\right\} \le L_{\tau m} \max_a\left\{\frac{\ell_\tau}{\ell_0}(\mathbf{Y}^{(k)}_m|a)B_\tau(\mathbf{Y}_{1,mk})\right\}, \tag{a}$$

where

$$B_\tau(\mathbf{Y}_{1,mk}) = \frac{\sum_a P_\tau[X_{mk+1} = a|\mathbf{Y}_{1,mk}]\ell_0(\mathbf{Y}^{(k)}_m|a)}{\sum_a P_0[X_{mk+1} = a|\mathbf{Y}_{1,mk}]\ell_0(\mathbf{Y}^{(k)}_m|a)}. \tag{b}$$

But

$$|B_\tau(\mathbf{Y}_{1,mk}) - 1| \le \max_a\left|\frac{P_\tau[X_{mk+1} = a|\mathbf{Y}_{1,mk}]}{P_0[X_{mk+1} = a|\mathbf{Y}_{1,mk}]} - 1\right|. \tag{c}$$

It follows from Lemmas 3.11 and 3.4,

$$\sup\{|B_\tau(\mathbf{Y}_{1,mk}) - 1| : |\tau| \le M, 1 \le m \le N\} = O_{n^{-1}}(n^{-1/2+2\epsilon}). \tag{d}$$

On the other hand,

$$\frac{\ell_\tau}{\ell_0}(\mathbf{Y}^{(k)}_m|a) = \frac{\pi_0}{\pi_\tau}\mathrm{E}_0\left\{\frac{f_\tau}{f_0}(\mathbf{X}^{(k)}_m, \mathbf{Y}^{(k)}_m)|X_{mk+1} = a, \mathbf{Y}^{(k)}_m\right\} \tag{e}$$

so that by (a) of the proof of Lemma 3.10, if $k = o(n^{1/2}/(\log n)^2)$, then

$$\sup\left\{\left|\frac{\ell_\tau}{\ell_0}(\mathbf{Y}^{(k)}_m|a) - \frac{\ell_\tau}{\ell_0}(\mathbf{Y}^{(k)}_m|b)\right| : m, |\tau| \le M, a, b\right\} = O_{n^{-1}}(n^{-1/2+\epsilon}). \tag{f}$$

From (a), (d), (f) and Lemma 3.12 we obtain Lemma 3.13. □

Lemma 3.14. *Under Assumptions 1–5,*

$$\mathrm{E}_0\sum_{m=1}^{N}|L^*_{\tau m} - 1| = O\left(\left(\frac{n}{k}\right)^{1/2}\right). \tag{3.15}$$

Proof of Lemma 3.14. Note that

$$\mathrm{E}_0\{|L^*_{\tau 0} - 1||X_1 = a\} = \|\mathcal{L}_\tau(\mathbf{Y}_{1,k}|X_1 = a) - \mathcal{L}_0(\mathbf{Y}_{1,k}|X_1 = a)\|, \tag{a}$$

where $\|\cdot\|$ denotes variational distance. Therefore,

$$\begin{aligned} &\mathrm{E}_0\{|L^*_{\tau 0} - 1\|X_1 = a\} && \text{(b)}\\ &\quad \le \|\mathcal{L}_\tau((\mathbf{X}_{1,k}, \mathbf{Y}_{1,k})|X_1 = a) - \mathcal{L}_0((\mathbf{X}_{1,k}, \mathbf{Y}_{1,k})|X_1 = a)\| && \text{(c)}\\ &\quad \le 2H(\mathcal{L}_{0,a}, \mathcal{L}_{1a})(2 - H^2(\mathcal{L}_{0a}, \mathcal{L}_{1a}))^{1/2}, \end{aligned}$$

where $\mathcal{L}_{0a}, \mathcal{L}_{1a}$ are the laws in (c) and H is Hellinger distance, by a standard inequality (Le Cam 1986, p.47). But

$$1 - H^2(\mathcal{L}_{0a}, \mathcal{L}_{1a}) = \mathrm{E}_0\left\{\left(\frac{f_\tau}{f_0}\right)^{1/2}(\mathbf{X}_{1,k}, \mathbf{Y}_{1,k})|X_1 = a\right\}$$

$$= \mathrm{E}_0\left\{\left(\frac{\pi_\tau}{\pi_0}\right)^{1/2}(a)\prod_1^{k-1}\left(\frac{\alpha_\tau}{\alpha_0}\right)^{1/2}(X_i, X_{i+1})\prod_{i=1}^k \mathrm{E}_0\left[\left(\frac{g_\tau}{g_0}\right)^{1/2}(Y_i|X_i)\right]|x_1 = a\right\}. \tag{d}$$

But

$$\begin{aligned}\prod_{i=1}^k \mathrm{E}_0\left[\left(\frac{g_\tau}{g_0}\right)^{1/2}(Y_i|X_i)\right] &= \prod_{i=1}^k \mathrm{E}_0\{\mathrm{e}^{(1/2)\log(g_\tau/g_0)(Y_i|X_i)}\} && \text{(e)}\\ &\ge \prod_{i=1}^k\left[1 - \frac{\delta_n^2}{2}|\tau|^2 p\,\mathrm{E}_0\{(\tfrac{1}{4}q_0^2(Y_i, M\delta_n) + \tfrac{1}{2}q_{02}(Y_i, M\delta_n))\mathrm{e}^{(|\tau|/2)\delta_n q_0(Y_i, M\delta_n)}\}\right]\\ &\ge 1 - O\left(\frac{k}{n}\right)\end{aligned}$$

by Taylor expansion and Assumptions 3 and 5. Similarly, by Assumptions 1 and 2:

$$\mathrm{E}_0\left\{\left(\frac{\pi_\tau}{\pi_0}\right)^{1/2}(a)\prod_1^{k-1}\left(\frac{\alpha_\tau}{\alpha_0}\right)^{1/2}(X_i, X_{i+1})|X_1 = a\right\} \ge 1 - O\left(\frac{k}{n}\right). \tag{f}$$

Finally, we conclude from (e) and (f):

$$\sum_{a=1}^K H^2(\mathcal{L}_{0a}, \mathcal{L}_{1a})\pi_0(a) = O\left(\frac{k}{n}\right). \tag{g}$$

The lemma is proved by (a), (b) and (g). □

Lemma 3.15. *If $L^{(d)}_{\tau m}$ is given by (2.2), $r > 8, k = o(n^{1/2-\gamma})$ for some $\gamma > 0$, $d = n^\epsilon \log^2 n$, $\epsilon > 2/r$ then*

$$\sup\{|L^{(d)}_{\tau m} - L_{\tau m}| : |\tau| \le M, 1 \le m \le N\} = O_{n^{-1}}(n^{-1-\epsilon}). \tag{3.16}$$

Proof of Lemma 3.15. By Lemma 3.2, if B_n is given by (d) of Lemma 3.11,

$$\begin{aligned}
&\sup_{\substack{1\le m\le N\\ |\tau|\le M}} \{|P_\tau[X_{mk+1}=a|\mathbf{Y}_{mk-d,mk}]-P_\tau[X_{mk+1}=a|\mathbf{Y}_{1,mk}]|\} \qquad \text{(a)}\\
&\le \max_{m\le 1}\left\{\sum_{\ell=1}^{mk-d-1}\prod_{j=\ell+1}^{mk}(1-2\mu_0(Y_j))\right\}\\
&\le \frac{\mathrm{e}^{-(d-1)B_n}}{1-\mathrm{e}^{-B_n}}\\
&= O_{1/n}(n^{-1-2\epsilon})
\end{aligned}$$

by arguing as for (l) of Lemma 3.10. But, by Lemma 3.1,

$$P_0[X_\ell=a|\mathbf{Y}_{1,n}|\ge\min_{a,b}\alpha_{0,\ell-1,n}(a,b)\ge\min\{\mu_0(Y_j):1\le j\le n\} \tag{b}$$

and, hence

$$P_0[\min_{a,\ell}P_0[X_\ell=a|\mathbf{Y}_{1,n}]\ge n^{-\epsilon}]=1-o(n^{-1}). \tag{c}$$

But, arguing as for Lemma 3.13,

$$|L_{\tau m}-L_{\tau m}^{(d)}|\le A_m(\tau)\max_\alpha\frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|a)+A_m(0)L_{\tau m}, \tag{d}$$

where

$$A_m(\tau)\equiv\max_a\left\{\frac{|P_\tau[X_{mk+1}=a|\mathbf{Y}_{1,mk}]-P_\tau[X_{mk+1}=a|\mathbf{Y}_{mk-d,mk}]|}{P_0[X_{mk+1}=a|\mathbf{Y}_{mk-d,mk}]}\right\}. \tag{e}$$

By (a) and (c),

$$\sup\{A_m(\tau):m,|\tau|\le M\}=O_{n^{-1}}(n^{-(1+\epsilon)}) \tag{f}$$

and Lemma 3.15 follows from (d), Lemma 3.12 and Lemma 3.13. □

Lemma 3.16. *Suppose Assumptions 1–4 hold. Let* $d=n^\epsilon(\log n)^2$, $\epsilon>2/r$, *and* $k=n^{4\epsilon+\gamma}$ *for some* $\gamma>0, 4\epsilon+\gamma<1/2$, *so that* $r>16$. *Then*

$$\mathrm{E}_0\left\{\left(\sum_m(L_{\tau m}^{(d)}-L_{\tau m}^*)\right)^2\right\}=O(n^{-\gamma}). \tag{3.17}$$

Proof of Lemma 3.16. For any fixed u, we first bound

$$\mathrm{E}_0\{(L_{\tau m}^{(d)})^u\}\le\mathrm{E}_0\left\{\left(\max_a\frac{P_\tau}{P_0}[X_{mk+1}=a|\mathbf{Y}_{mk-d,mk}]\max_a\frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|a)\right)^u\right\}. \tag{a}$$

Now the first term in (a) is uniformly bounded by Lemma 3.4. The second is bounded by

$$\exp\left\{M\delta_n u \sum_{j=mk+1}^{(m+1)k} q_0(Y_j, M\delta_n)\right\}. \tag{b}$$

Thus, if $k = o(n^{1/2})$, by Assumption 3, for all u, eventually

$$\mathrm{E}_0(L_{\tau m}^{(d)})^u \le C_1(1 + C_2\delta_n)^k \le C_3. \tag{c}$$

Similarly,

$$\mathrm{E}_0(L_{\tau m}^*)^u \le C_4. \tag{d}$$

Now,

$$\begin{aligned}|L_{\tau m}^{(d)} - L_{\tau m}^*| &\le \max_{a,b}\left|\frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|a) - \frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|b)\right| \\ &\quad + \max_a \frac{\ell_\tau}{\ell_0}(\mathbf{Y}_m^{(k)}|a)\max_a\left|\frac{P_\tau}{P_0}[X_{mk+1} = a|\mathbf{Y}_{mk-d,mk}] - 1\right| \\ &= O_{n^{-1}}(n^{-1/2+\epsilon}) + O_{n^{-1}}(n^{-1/2+2\epsilon})\end{aligned} \tag{e}$$

by (f) of Lemma 3.13, Lemma 3.12, Lemma 3.4, Lemma 3.11 and (a) of Lemma 3.15. Let $c_n = cn^{-1/2+2\epsilon}$ for some large enough c. Note that

$$\begin{aligned}\mathrm{E}_0|L_{\tau m}^{(d)} - L_{\tau m}^*|^{2+8\epsilon} &\le c_n^{2+8\epsilon} + \mathrm{E}_0\{|L_{\tau m}^{(d)} - K_{\tau m}^*|^{2+8\epsilon}1(|L_{\tau m}^{(d)} - L_{\tau m}^*| \ge c_n)\} \\ &\le c_n^{2+8\epsilon} + \mathrm{E}_0^{16\epsilon^2}\{|L_{\tau m}^{(d)} - L_{\tau m}^*|^{(2+8\epsilon)/16\epsilon^2}\}P_0^{1-16\epsilon^2}\{|L_{\tau m}^{(d)} - L_{\tau m}^*| < c_n\} \\ &\le 2c_n^{2+8\epsilon}\end{aligned} \tag{f}$$

for large enough n.

We will apply Lemma 3.7, with $\delta = 8\epsilon$. Note that, if $d \le k$, by the geometric ergodicity of the chain under Assumptions 1 and 2, the variational norm distance between the joint distribution of $(L_{\tau m_1}^{(d)} - L_{\tau m_1}^*, L_{\tau m_2}^{(d)} - L_{\tau m_2}^*)$ and the product of the marginals is bounded by $C\xi^{|m_1 - m_2|}$ for some $C < \infty, \xi < 1$ and all m_1, m_2. Hence, using (f) above,

$$\mathrm{E}_0\left\{\left(\sum_{m=1}^{N}(L_{\tau m}^{(d)} - L_{\tau m}^*)\right)^2\right\} = O\left(\frac{n}{k}c_n^2\right) = O(n^{-\gamma}) \tag{g}$$

under our conditions on k, c_n. □

Proof of Lemma 2.1. It is enough to show that all terms on the right-hand side of (2.6) are $O_{e_n}(n^{-\gamma/2}/e_n)$. The first term is equal to

$$\sum_{m=1}^{N}(L_{\tau m} - L_{\tau m}^{(d)}) + \sum_{m=1}^{N}(L_{\tau m}^{(d)} - L_{\tau m}^*) = O_{e_n}(n^{-\gamma/2}\mathrm{e}_n^{-1/2}) \tag{a}$$

By Lemmas 3.15 and 3.16. The second term can be bounded by

$$\sup_{\substack{1 \le m \le N \\ |\tau| \le M}} \left\{ \frac{|L_{\tau m} - L^*_{\tau m}|}{L^*_{\tau m}} \right\} \sum_{m=1}^{N} |L^*_{\tau m} - 1| = O_{e_n}\left(n^{-1/2+2\epsilon} \left(\frac{n}{k}\right)^{1/2} / e_n \right) = O_{e_n}(n^{-\gamma/2}/e_n) \tag{b}$$

by Lemmas 3.13 and 3.14. Finally, the third term is negligible since

$$|R_n| \le \left(1 - \sup \left\{ \frac{|L_{\tau m} - L^*_{\tau m}|}{L^*_{\tau m}} : 1 \le m \le N, |\tau| \le M \right\} \right)^{-2} = O_{n^{-1}}(1) \tag{c}$$

and

$$\sum_{m=1}^{N} \left(\frac{|L_{\tau m} - L^*_{\tau m}|}{L^*_{\tau m}} \right)^2 = O_{1/n}(n^{-\gamma}), \tag{d}$$

both by Lemma 3.13. □

Proof of Lemma 2.2. Expand

$$\begin{aligned} \log L^*_{\tau m} &= \delta_n \tau^{\mathrm{T}} \nabla \log \ell_0(\mathbf{Y}_m^{(k)} | X_{mk+1}) \\ &\quad + \frac{1}{2n} \tau^{\mathrm{T}} \left\| \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \log \ell_0(\mathbf{Y}_m^{(k)} | X_{mk+1}) \right\| \tau \\ &\quad + \delta_n^3 \int_0^1 \frac{(1-\lambda)^2}{2} \sum_{a,b,c} \tau_a \tau_b \tau_c \frac{\partial^3}{\partial \vartheta_a \partial \vartheta_b \partial \vartheta_c} \log \ell_{\lambda \tau}(\mathbf{Y}_m^{(k)} | \mathbf{X}_{mk+1}) \mathrm{d}\lambda. \end{aligned} \tag{a}$$

We use a classical formula based on Lemma 3.6. If $\mathbf{B}$ is generated by X_{mk+1}, $\mathbf{Y}_m^{(k)}$, and we suppress arguments in f_ϑ,

$$\begin{aligned} &\frac{\partial^3}{\partial \vartheta_a \partial \vartheta_b \partial \vartheta_c} \log \ell_\vartheta(\mathbf{Y}_m^{(k)} | X_{mk+1}) \\ &= \mathrm{E}_0 \left\{ \frac{\partial^3}{\partial \vartheta_a \partial \vartheta_b \partial \vartheta_c} \log f_\vartheta | \mathbf{B} \right\} + \mathrm{cov}_\vartheta \left\{ \frac{\partial^2}{\partial \vartheta_a \partial \vartheta_b} \log f_\vartheta, \frac{\partial}{\partial \vartheta_c} \log f_\vartheta | \mathbf{B} \right\} \\ &\quad + \mathrm{cov}_\vartheta \left\{ \frac{\partial^2}{\partial \vartheta_a \partial \vartheta_c} \log f_\vartheta, \frac{\partial}{\partial \vartheta_b} \log f_\vartheta | \mathbf{B} \right\} + \mathrm{cov}_\vartheta \left\{ \frac{\partial^2}{\partial \vartheta_b \partial \vartheta_c} \log f_\vartheta, \frac{\partial}{\partial \vartheta_a} \log f_\vartheta | \mathbf{B} \right\} \\ &\quad - \mathrm{cov}_\vartheta \left\{ \frac{\partial}{\partial \vartheta_a} \log f_\vartheta \frac{\partial}{\pi \vartheta_b} \log f_\vartheta, \frac{\partial}{\partial \vartheta_c} \log f_\vartheta | \mathbf{B} \right\} \\ &\quad - \mathrm{cov}_\vartheta \left\{ \frac{\partial}{\partial \vartheta_a} \log f_\vartheta \frac{\partial}{\partial \vartheta_c} \log f_\vartheta, \frac{\partial}{\partial \vartheta_b} \log f_\vartheta | \mathbf{B} \right\} \\ &\quad - \mathrm{cov} \left\{ \frac{\partial}{\partial \vartheta_b} \log f_\vartheta \frac{\partial}{\partial \vartheta_c} f_\vartheta, \frac{\partial}{\partial \vartheta_a} \log f_\vartheta | \mathbf{B} \right\} - \frac{\partial^3}{\partial \vartheta_a \partial \vartheta_b \partial \vartheta_c} \log \pi_\vartheta(X_{mk+1}). \end{aligned} \tag{b}$$

We see from (b) and Assumptions 1 and 2 that to bound the third term in (a) it suffices to bound, for $|\vartheta - \vartheta_0| \le M\delta_n$, all a, b, c,

$$E_0\left\{E_\vartheta\left[\left|\sum_{j=1}^{k}\frac{\partial^3}{\partial\vartheta_a\partial\vartheta_b\partial\vartheta_c}\log g_\vartheta(Y_j|X_j)\right| | Y_i,\dots,Y_k\right]\right\}, \tag{c}$$

$$E_0\left\{E_\vartheta\left[\left|\sum_{j=1}^{k}\frac{\partial^2}{\partial\vartheta_a\partial\vartheta_b}\log g_\vartheta(Y_j|X_j)\right|\left(1+\left|\sum_{j=1}^{k}\frac{\partial}{\partial\vartheta_c}\log g_\vartheta(Y_j|X_j)\right|\right) | Y_1,\dots,Y_k\right]\right\}, \tag{d}$$

and

$$E_0\left\{E_\vartheta\left[\left|\sum_{j=1}^{k}\frac{\partial}{\partial\vartheta_a}\log g_\vartheta(Y_j|X_j)\right|^3 | Y_1,\dots,Y_k\right]\right\}. \tag{e}$$

We can apply Lemma 3.5 to all of these and use Assumption 3 to conclude that, under Assumption 5, (c)–(e) are uniformly $O(k^3)$. To do so we take r in the lemma as close to 1 as possible and s and t as large as necessary since, by A3, and by arguing as in (b) of Lemma 3.16, $E_0 \exp|t\Lambda| < \infty$ for all $k = o(n^{1/2})$, t. Therefore, the expectation of the remainder in (a) is $O(n^{-3/2}k^3)$. The lemma follows since there are n/k terms like that in the left-hand side of (2.10). □

Lemma 3.17. *Let* $-k \ge -j+2$ *and*

$$S(j,k) \equiv \max_{a,b,c}\left\{\left|P_0[X_{-k}=a|X_{-j+2},\mathbf{Y}_{-j+2,0},X_1=b] - P_0[X_{-k}=a|X_{-j+2},\mathbf{Y}_{-j+2,0},X_1=c]\right|\right\}. \tag{3.18}$$

Then

$$S(j,k) \le 2\gamma^{-1}(\vartheta_0)\prod_{i=-k+1}^{0}(1-2\mu_0(Y_i)). \tag{3.19}$$

Proof of Lemma 3.17.

$$\begin{aligned}&P_0[X_{-k}=a|X_{-j+2},\mathbf{Y}_{-j+2,0},X_1=b]\\&\quad=\frac{P_0[X_1=b\,|\,X_{-j+2},\mathbf{Y}_{-j+2,0},X_{-k}=a]}{P_0[X_1=b\,|\,X_{-j+2},\mathbf{Y}_{-j+2,0}}P_0[X_{-k}=a|\,X_{-j+2},\mathbf{Y}_{-j+2,0}].\end{aligned} \tag{a}$$

Then

$$S(j,k) \le 2\max_{a,b}\left\{\left|\frac{P_0[X_1=b\,|\,X_{-j+2},\mathbf{Y}_{-j+2,0},X_{-k}=a]}{P_0[X_1=b\,|\,X_{-j+2},\mathbf{Y}_{-j+2,0}]}-1\right|\right\}. \tag{b}$$

But

$$P_0[X_1 = b | X_{-j+2}, \mathbf{Y}_{-j+2,0}] = \sum_c P_0[X_1 = b | X_{-k} = c, \mathbf{Y}_{-k+1,0}] P_0[X_{-k} = c | X_{-j+2}, \mathbf{Y}_{-j+2,0}], \tag{c}$$

and hence

$$\begin{aligned} S(j,k) &\le 2 \max_{a,b} \frac{\sum_c |P_0[X_1 = b | X_{-k} = c, \mathbf{Y}_{-k+1,0}] - P_0[X_1 = b | X_{-k} = a, \mathbf{Y}_{-k+1,0}]|}{\min_b P_0[X_1 = b | X_{-j+2}, \mathbf{Y}_{-j+2,0}]} \\ &\le 2\gamma^{-1}(\vartheta_0) K \prod_{j=-k+1}^{0} (1 - 2\mu_0(Y_j)) \qquad \text{(d)} \end{aligned}$$

by Lemmas 3.3 and 3.4. □

Proof of Lemma 2.3. Without loss of generality, take $\vartheta_0 = 0$. Write

$$\ell_\vartheta(Y_1, \ldots, Y_k | X_1) = \prod_{j=1}^{k} \frac{g_{j\vartheta}}{g_{(j-1)\vartheta}}(X_1, \mathbf{Y}_{1,j}), \tag{a}$$

where $g_{j\vartheta}(X_1, \mathbf{Y}_{1,j})$ is the joint density of $(X_1, \mathbf{Y}_{1,j})$ for $j \ge 1$, and $g_{0\vartheta} = \pi_\vartheta(X_1)$. Take $\dim(\vartheta) = 1$. The generalization is trivial. Then

$$\frac{\partial}{\partial\vartheta} \log \ell_\vartheta(\mathbf{Y}_0^{(k)} | X_1) = \sum_{j=1}^{k} \left[\frac{\partial}{\partial\vartheta} \log g_{j\vartheta}(X_1, \mathbf{Y}_{1,j}) - \frac{\partial}{\partial\vartheta} \log g_{(j-1)\vartheta}(X_1, \mathbf{Y}_{1,j-1}) \right].$$

The terms in brackets are of course martingale summands, and we arrive at the identity

$$\begin{aligned} &\mathrm{E}_0 \left\{ \left(\frac{\partial}{\partial\vartheta} \log \ell_0(\mathbf{Y}_0^{(k)} | X_1) \right)^2 \right\} \\ &= \sum_{j=1}^{k} \mathrm{E}_0 \left\{ \left(\frac{\partial}{\partial\vartheta} \log g_{j0}(X_1, \mathbf{Y}_{1,j}) - \frac{\partial}{\partial\vartheta} \log g_{(j-1)0}(X_1, \mathbf{Y}_{1,j-1}) \right)^2 \right\} \qquad \text{(c)} \\ &= \sum_{j=1}^{k} \mathrm{E}_0 \{ U_j^2(X_1, \mathbf{Y}_{i,j}) \}, \quad \text{say} \\ &= \sum_{j=1}^{k} \mathrm{E}_0 \{ U_j^2(X_{-j+2}, \mathbf{Y}_{-j+2,1}) \}, \end{aligned}$$

where $(X_j, Y_j), -\infty < j < \infty$, is the two-sided stationary sequence such that $(X_j, Y_j), j \ge 1$, are distributed according to P_ϑ. We claim that

$$\mathrm{E}_0 \{ U_j^2(X_{-j+2}, \mathbf{Y}_{-j+2,1} \} \to I(\vartheta_0), \tag{d}$$

and that, combined with (c), clearly establishes (2.11). Now, if we use $(b'/b)(\vartheta)$ for $(\partial/\partial\vartheta)\log b(\vartheta)$,

$$U_j(X_{-j+2}, \mathbf{Y}_{-j+2,1}) = \mathrm{E}_0\left\{\sum_{m=-j+2}^{1}\frac{g_0'}{g_0}(Y_m|X_m) + \sum_{m=-j+2}^{0}\frac{\alpha_0'}{\alpha_0}(X_m, X_{m+1})|X_{-j+2}, \mathbf{Y}_{-j+2,1}\right\} \quad \text{(e)}$$
$$- \mathrm{E}_0\left\{\sum_{m=-j+2}^{0}\frac{g_0'}{g_0}(Y_m|X_m) + \sum_{m=-j+2}^{-1}\frac{\alpha_0'}{\alpha_0}(X_m, X_{m+1})|X_{-j+2}, \mathbf{Y}_{-j+2,0}\right\}$$

by the usual formula. Consider the first part of the mth term in the sum in (e),

$$U_{jm}^{(1)} \equiv \mathrm{E}_0\left\{\frac{g_0'}{g_0}(Y_m|X_m)|X_{-j+2}, \mathbf{Y}_{-j+2,1}\right\} - \mathrm{E}\left\{\frac{g_0'}{g_0}(Y_m|X_m)|X_{-j+2}, \mathbf{Y}_{-j+2,0}\right\} \quad \text{(f)}$$
$$= \sum_{a=1}^{K}\frac{g_0'}{g_0}(Y_m|a)\{P_0[X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,1}] - P_0[X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}]\}.$$

Note that, by the (backward) martingale convergence theorem, for fixed $m < 0$,

$$U_{jm}^{(1)} \xrightarrow{P_0} \mathrm{E}_0\left\{\frac{g_0'}{g_0}(Y_m|X_m)|Y_1, Y_0, \dots\right\} - \mathrm{E}_0\left\{\frac{g_0'}{g_0}(Y_m|X_m)|Y_0, Y_{-1}, \dots\right\} \quad \text{(g)}$$

as $j \to \infty$.

Note that

$$P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}\}$$
$$= \sum_b P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}, X_1 = b\}P_0\{X_1, = b|X_{-j+2}, \mathbf{Y}_{-j+2,0}\} \quad \text{(h)}$$

and

$$P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,1}\}$$
$$= \sum_c P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}, X_1 = c\}P_0\{X_1 = c|X_{-j+2}, \mathbf{Y}_{-j+2,1}\}, \quad \text{(i)}$$

so that

$$\max_a |P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}\} - P_0\{X_m = a|X_{j+2}, \mathbf{Y}_{-j+2,1}\}|$$
$$\le \max_{a,b,c}|P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}, X_1 = b\} \quad \text{(j)}$$
$$- P_0\{X_m = a|X_{-j+2}, \mathbf{Y}_{-j+2,0}, X_1 = c\}| = S(j, -m).$$

We conclude, by Lemma 3.17, that

$$|U_{jm}^{(1)}| \le 2\gamma^{-1}(\vartheta_0)\sum_{a=1}^{K}\left|\frac{g_0'}{g_0}(Y_m|a)\right|\prod_{k=m+1}^{0}(1-2\mu_0(Y_k))$$
$$\le 2\gamma^{-1}(\vartheta_0)Kq_0(Y_m, M\delta_n)\exp\left(-2\sum_{k=m+1}^{0}\mu_0(Y_k)\right). \qquad \text{(k)}$$

Now, by (k)

$$\mathrm{E}_0\left(\sum_{m=-j+2}^{-k}U_{jm}^{(1)}\right)^2 \le 4\gamma^{-2}(\vartheta_0)K^2\sum_{m_1=-j+2}^{-k}\sum_{m_2=-j+2}^{-k}\mathrm{E}_0\Bigg\{q_0(Y_{m_1}, M\delta_n)q_0(Y_{m_2}, M\delta_n) \qquad \text{(l)}$$
$$\exp\left[-2\left(\sum_{t=m_1}^{0}\mu_0(Y_t)+\sum_{t=m_2}^{0}\mu_0(Y_t)\right]\right\}.$$

Applying the Hölder inequality to each term and using Assumption 3, we obtain

$$\mathrm{E}_0\left\{\left(\sum_{m=-j+2}^{-k}U_{jm}^{(1)}\right)^2\right\} \le C_\epsilon\sum_{m_1}\sum_{m_2}\mathrm{E}_0^{(1+\epsilon)^{-1}}\exp\left[-2(1+\epsilon)\left(\sum_{t=m_1}^{0}\mu_0(Y_t)+\sum_{t=m_2}^{0}\mu_0(Y_t)\right)\right]. \quad \text{(m)}$$

But, if $m_1 \le m_2$,

$$\mathrm{E}_0\left\{\exp\left[-2(1+\epsilon)\left(\sum_{t=m_1}^{0}\mu_0(Y_t)+\sum_{t=m_2}^{0}\mu_0(Y_t)\right)\right]\right\} \qquad \text{(n)}$$
$$= \mathrm{E}_0\left\{\prod_{t=m_2}^{0}\mathrm{E}_0(\mathrm{e}^{-4(1+\epsilon)\mu_0(Y_t)}|X_t)\prod_{t=m_1}^{m_2-1}\mathrm{E}_0(\mathrm{e}^{-2(1+\epsilon)\mu_0(Y_t)}|X_t)\right\}$$
$$\le \gamma_{4(1+\epsilon)}^{-m_2}\gamma_{2(1+\epsilon)}^{m_2-m_1} \le \gamma_{2(1+\epsilon)}^{|m_1|}$$

where $\gamma_s = \max_a \mathrm{E}_0(\mathrm{e}^{-s\mu_0(Y_1)}|X_1 = a) < 1$ for all $s > 0$. Using the bound from (n) in (m) we obtain, for some $C_\epsilon < \infty, \gamma \equiv \gamma_{2(1+\epsilon)}$,

$$\mathrm{E}_0\left\{\left(\sum_{m=-j+2}^{-k}U_{jm}^{(1)}\right)^2\right\} \le 2C_\epsilon\sum_{m=k}^{j-2}m\gamma^{m(1+\epsilon)^{-1}} \le 2C_\epsilon\gamma^{k(1+\epsilon)^{-1}}(1-\gamma^{(1+\epsilon)^{-1}})^{-1}. \qquad \text{(o)}$$

Thus for any $\delta > 0$ there exists $k = k(\delta)$ such that, for all $j > k+2$,

$$\mathrm{E}_0\left\{\left(\sum_{m=-j+2}^{-k}U_{jm}^{(1)}\right)^2\right\} \le \delta. \qquad \text{(p)}$$

A similar argument shows that for fixed k, some $C < \infty$, all j,

$$E_0\left(\sum_{m=-k}^{0} U_{jm}^{(1)}\right)^4 \le C. \tag{q}$$

By a similar but easier argument, if

$$U_{jm}^{(2)} = E_0\left\{\frac{\alpha'_0}{\alpha_0}(X_m, X_{m+1})|X_{-j+2}, \mathbf{Y}_{-j+2,1}\right\} - E_0\left\{\frac{\alpha'_0}{\alpha_0}(X_m, X_{m+1})|X_{-j+2}, \mathbf{Y}_{-j+2,1}\right\}. \tag{r}$$

then

$$U_{jm}^{(2)} \xrightarrow{P_0} E_0\left\{\frac{\alpha'_0}{\alpha_0}(X_m, X_{m+1})|Y_1, Y_0, \ldots\right\} - E_0\left\{\frac{\alpha'_0}{\alpha_0}(X_m, , X_{m+1})|Y_0, Y_{-1}, \ldots\right\} \tag{s}$$

and (p) and (q) carry over as well. We conclude that (d) follows since in fact, by (g), (p)–(s),

$$U_j(X_{-j+2}, \mathbf{Y}_{-j+2,1}) \xrightarrow{L_2} W(Y_1, Y_0, \ldots). \tag{t}$$

The lemma follows. □

Proof of Lemma 2.4. We begin by proving (2.12). In view of Lemma 2.3 it is enough to show that, for all τ,

$$\mathrm{var}_0\left(\frac{1}{n}\sum_{m=1}^{N} \tau^{\mathrm{T}} E_0\{\nabla\nabla^{\mathrm{T}} \log \ell_0(\mathbf{Y}_n^{(k)}|X_{mk+1})|X_{mk+1}\}\tau^{\mathrm{T}}\right) \to 0. \tag{a}$$

But if we let $h_{k,m}(X_{mk+1})$ denote the mth summand in (a), then Lemma 3.7 and geometric ergodicity of the $\{X_j\}$ guarantee that the expression in (a) is bounded by

$$C E_0 h_{k,1}^2(X_1) N n^{-2}. \tag{b}$$

Also

$$\begin{aligned} E_0 h_{k,1}^2(X_1) &\le M^4 E_0|\nabla \log \ell_0(\mathbf{Y}_i^{(k)}|X_1)|^4 \\ &\le M^4 E_0\left|\sum_{i=1}^{k} \nabla \log g_0(Y_i|X_i) + \sum_{i=1}^{k-1} \nabla \log \alpha_0(X_i, X_{i+1}) + \frac{\pi'_0}{\pi_0}(X_1)\right|^4 \\ &= O(k^2) \end{aligned} \tag{c}$$

by invoking Lemma 3.7 and (e) of Lemma 2.3 again. Thus, $E_0 h_{k,1}^2(X_1) N n^{-2} = O(kn^{-1}) = o(1)$ and (a) and (2.12) follow. To prove (2.13) we take expectations and note that it is enough to show that

$$E_0|\nabla \log \ell_0(\mathbf{Y}_1^{(k)}|X_1)|^4 = O(k^2). \tag{d}$$

But this is just (c). Finally, (2.14) follows from

$$P_0[\delta_n|\nabla \log \ell_0(\mathbf{Y}_1^{(k)}|X_1)| \le \epsilon] \le n^{-2}\epsilon^{-4} E_0|\nabla \log \ell_0(\mathbf{Y}_1^{(k)}|X_1)|^4 = O(k^2 n^{-2}). \qquad \square \tag{e}$$

Proof of Lemma 2.5. By a standard identity valid under our conditions,

$$\left\| \mathrm{E}_0\left(\frac{\partial^2}{\partial\vartheta_a\partial\vartheta_b}\log\ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1})\right)\right\| = -\mathrm{E}_0(\nabla\nabla^{\mathrm{T}}\log\ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1})). \tag{a}$$

Therefore, by Lemma 2.3 and stationarity,

$$\frac{1}{n}\mathrm{E}_0\sum_{m=1}^{N}\left\|\frac{\partial^2}{\partial\vartheta_a\partial\vartheta_b}\log\ell_0(\mathbf{Y}_m^{(k)}|X_{mk+1}\right\| \to -I(\vartheta_0). \tag{b}$$

Now use A5 and argue as in the proof of (2.12) to obtain the lemma. □

Acknowledgements

The research of both authors was partially supported by NSF grant DMS-9115577 and by US/Israel Bi-National Science Foundation grant 90-00031/2.

References

Ball, F. and Rice, J. (1992) Stochastic models for ion channels: introduction and bibliography. *Math. Biosci.*, **112,** 189–206.

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

Churchill, G. (1992) Hidden Markov chains and the analysis of genome structure. *Comput. in Chemistry*, **16,** 107–115.

Daley, D. and Vere Jones, D. (1989) *Introduction to the Theory of Point Processes*. New York: Springer-Verlag.

Frigessi, A. and Piccioni, M. (1990) Parameter estimation for 2-dimensional Ising fields corrupted by noise. *Stochastic Process. Applic.* **34,** 297–311.

Furstenberg, H. and Kesten, H. (1960) Products of random matrices. *Ann. Math. Statist.*, **31,** 457–469.

Götze, F. and Hipp, C. (1992) On the validity of Edgeworth expansions. Technical Report, University of Bielefeld.

Hall, P. and Heyde, C. (1980) *Martingale Limit Theory and Its Applications*. New York: Academic Press.

Ibragimov, I. and Linnik, Y. (1971) *Independent and Stationary Sequences of Random Variables*. Groningen: Walters–Noordhoff.

Kingman, J.F.C. (1976) Subadditive processes. In P.L. Hennequin (ed.), *Ecole d'Eté de Probabilités de Saint-Flour V-1976*, pp. 167–223. Lecture Notes in Math. 539. Berlin: Springer-Verlag.

Le Cam, L. (1956) On the asymptotic theory of estimation and testing hypotheses. In J. Neyman (ed.), *Proc. Third Berkeley Symp. Math. Statist. and Prob.*, **1,** 129–156. University of California Press, Berkeley.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Le Cam, L. and Yang, G. (1990) *Asymptotics in Statistics. Some Basic Concepts*. New York: Springer-Verlag.

Leroux, B. (1989) Maximum likelihood estimation for mixture distributions and hidden Markov models. Doctoral Thesis, University of British Columbia.

Leroux, B. (1992) Maximum likelihood estimation for hidden Markov models. *Stochastic Process. Applic*. **40,** 127–143.

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

Priestley, M.B. (1988) *Nonlinear and Nonstationary Time Series*. London: Academic Press.

Qian, W. and Titterington, D.M. (1991) Estimation of parameters in hidden Markov models. *Philos. Trans. Roy. Soc. London Ser. A*, Physical Science and Engineering, **337,** 407–428.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77,** 257–284.

Rydén, T. (1994a) Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist*. **22,** 1884–1895.

Rydén, T. (1994b) Parameter estimation for Markov modulated Poisson Processes. *Commun. Statist. Stochastic Models*, **10,** 795–829.

Rydén, T. (1995) Estimating the order of hidden Markov models. *Statistics*, **26,** 345–354.

Tong, H. (1991) *Nonlinear Time Series*. Oxford: Oxford University Press.

Received September 1994 and revised October 1995