The Annals of Statistics
1998, Vol. 26, No. 4, 1614–1635

ASYMPTOTIC NORMALITY OF THE MAXIMUM-LIKELIHOOD ESTIMATOR FOR GENERAL HIDDEN MARKOV MODELS

BY PETER J. BICKEL,[1] YA'ACOV RITOV[1] AND TOBIAS RYDÉN[2]

University of California, Hebrew University and Lund University

Hidden Markov models (HMMs) have during the last decade become a widespread tool for modeling sequences of dependent random variables. Inference for such models is usually based on the maximum-likelihood estimator (MLE), and consistency of the MLE for general HMMs was recently proved by Leroux. In this paper we show that under mild conditions the MLE is also asymptotically normal and prove that the observed information matrix is a consistent estimator of the Fisher information.

1. Introduction. A hidden Markov model (HMM) is a discrete-time stochastic process $\{(X_k, Y_k)\}$ such that (i) $\{X_k\}$ is a finite-state Markov chain, and (ii) given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables with the conditional distribution of Y_n depending on $\{X_k\}$ only through X_n. The Markov chain $\{X_k\}$ is sometimes called the *regime*. The name HMM is motivated by the assumption that $\{X_k\}$ is not observable, so that inference and so on has to be based on $\{Y_k\}$ alone. HMMs have during the last decade become widespread for modeling sequences of weakly dependent random variables, with applications in areas such as speech processing [Rabiner (1989)], neurophysiology [Fredkin and Rice (1992)] and biology [Leroux and Puterman (1992)]. See also the monograph by MacDonald and Zucchini (1997). Commonly, the conditional distributions of Y_n given X_n belong to a single parametric family, such as the normal or Poisson families, so that X_n selects the parameter used to generate Y_n. The distribution of Y_n, that is, the marginal distribution of $\{Y_k\}$, will then be a finite mixture from the parametric family. Mixtures are frequently used in i.i.d. settings to increase the dispersion governed by a specific parametric family, and this effect is obviously found in the marginal distribution of an HMM as well. In addition, $\{Y_k\}$ is dependent. HMMs can thus be viewed as an extension of Markov chains, but also as an extension of mixture models.

Inference for HMMs was first considered by Baum and Petrie, who treated the case when $\{Y_k\}$ takes values in a finite set. In Baum and Petrie (1966), results on consistency and asymptotic normality of the maximum-likelihood estimator (MLE) are given, and the conditions for consistency are weakened in Petrie (1969). In the latter paper the identifiability problem is also discussed,

Received January 1997; revised October 1997.

[1]Supported in part by NSF Grant DMS-91-15577 and by US–Israel Bi-National Science Foundation Grant 90-00031/2.

[2]Supported by the Swedish Natural Science Research Council Contract M-AA/MA 10538-303.

AMS 1991 *subject classification*. Primary 62M09.

Key words and phrases. Hidden Markov model, incomplete data, missing data, asymptotic normality.

that is, under what conditions there are no other parameters that induce the same law for $\{Y_k\}$ as the true parameter does. For general HMMs, Lindgren (1978) constructed consistent and asymptotically normal estimators of the parameters determining the conditional densities of Y_n given X_n, but he did not consider estimation of the transition probabilities. Later, Leroux (1992) proved consistency of the MLE for general HMMs under mild conditions, and local asymptotic normality (LAN) has been proved by Bickel and Ritov (1996).

The topic of the present paper is asymptotic normality of the MLE. Although Bickel and Ritov (1996) prove that an estimator similar to the MLE is asymptotically normal and achieves the information bound, their result falls short of proving that the likelihood function has a second derivative and that the MLE itself is asymptotically normal. Asymptotic normality of the MLE can be inferred from their paper, but an extra argument is needed; see Ritov (1996). In this paper we show that the curvature of the likelihood function is, asymptotically, equal to the information bound and hence the MLE is asymptotically normal. We also work with conditions that are weaker than those in Bickel and Ritov (1996).

Before we proceed, we need to introduce some notation. We let $\{X_k\}_{k=1}^{\infty}$ be a stationary Markov chain on $\{1, \ldots, K\}$ with transition probabilities $\alpha(a, b) = P(X_{k+1} = b \mid X_k = a)$. We also let $\{Y_k\}$ be an $\mathscr{Y}$-valued sequence such that given $\{X_k\}$, $\{Y_k\}$ is a sequence of conditionally independent random variables, Y_n having (conditional) density $g(y|X_n)$ with respect to some σ-finite measure ν on $\mathscr{Y}$. Usually $\mathscr{Y}$ is a subset of $\mathbb{R}^q$ for some q, but it may also be a higher dimensional space. Moreover, both $\{\alpha(a, b)\}$ and $\{g(\cdot|a)\}$ depend on a parameter ϑ, that is $\alpha(a, b) = \alpha_\vartheta(a, b)$ and $g(\cdot|a) = g_\vartheta(\cdot|a)$, where ϑ is to be estimated from a realization of $\{Y_k\}$. The set to which ϑ belongs is denoted by Θ, and we assume $\Theta \subseteq \mathbb{R}^d$. Note that the stationary distribution of $\{X_k\}$, denoted by $\{\pi(a)\}_{a=1}^{K}$, does also depend on ϑ.

The most common set-up is that where ϑ contains the transition probabilities themselves, together with some parameters characterizing the g's. In particular, it is often the case that $g_\vartheta(y|a) = f(y; \phi(a))$ for some parametric family $f(y; \phi)$. We refer to this situation as the "usual parametrization." We now give a few examples of HMMs.

EXAMPLE 1 (Mixture of normal distributions). Let $K = 2$, $\vartheta = (\alpha(1, 2), \alpha(2, 1), \mu(1), \mu(2), \sigma^2)$ and $g_\vartheta(y|a) = \sigma^{-1}\varphi((y - \mu(a))/\sigma)$, where $\varphi(\cdot)$ is the standard normal density. Hence, $\mathscr{Y} = \mathbb{R}$ and ν is Lebesgue measure. The distribution of Y_n is a mixture of two normal distributions with different means but equal variances. This model has been used to model electric current through channels in ion membranes; see Guttorp [(1995), page 109], for a short description and Fredkin and Rice (1992) for a fuller treatment.

EXAMPLE 2 (Mixture of Poisson distributions). Let $K = 2$, $\vartheta = (\alpha(1, 2), \alpha(2, 1), \mu(1), \mu(2))$, and let $g_\vartheta(y|a)$ be the Poisson density with mean $\mu(a)$. Hence, $\mathscr{Y} = \{0, 1, 2, \ldots\}$ and ν is counting measure. The distribution of Y_n is a mixture of two Poisson distributions. Albert (1991) proposed this HMM

as a model for series of daily counts of epileptic seizures in one patient [see also Le, Leroux and Puterman (1992) and MacDonald and Zucchini (1997), page 146], Leroux and Puterman (1992) used it for modeling fetal lamb movements.

EXAMPLE 3 (Markov-modulated Poisson process). Let $\{X(t)\}$ be a continuous-time Markov chain on $\{1, \ldots, K\}$ with intensity matrix $Q = \{q(i, j)\}$, let $\lambda(1), \ldots, \lambda(K)$ be nonnegative numbers and let $\{N(t)\}$ be a doubly stochastic Poisson process (or Cox process) with random intensity function $\{\lambda(X(t))\}$; that is, given $\{\lambda(X(t))\}$, $\{N(t)\}$ has conditionally independent increments and $N(t+s) - N(t)$ has a Poisson distribution with mean $\int_t^{t+s} \lambda(X(u))\, du$. Such processes are called Markov-modulated Poisson processes, and they have been proposed for modeling traffic streams in complex telecommunication networks. See, for example, Heffes and Lucantoni (1986). The parameters of the model are the q's and the λ's. To make the connection to discrete-time HMMs, let $T_0 = 0$, let T_k be the time of the kth event in $\{N(t)\}$, $Y_k = T_k - T_{k-1}$ and $X_k = X(T_k)$. Then $\{(X_k, Y_k)\}$ is an HMM, except that given $\{X_k\}$, the distribution of Y_n depends on both X_{n-1} and X_n. Replacing $\{X_k\}$ by $\{X'_k\} = \{(X_{k-1}, X_k)\}$ takes us back to the standard set-up, however.

The joint density of $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ [with respect to (counting measure)$^n \times \nu^n$] is given by

$$p_\vartheta(x_1, \ldots, x_n, y_1, \ldots, y_n) = \pi_\vartheta(x_1) \prod_{k=1}^{n-1} \alpha_\vartheta(x_k, x_{k+1}) \prod_{k=1}^{n} g_\vartheta(y_k | x_k),$$

and the joint density of $(Y_1, \ldots, Y_n)$ (with respect to ν^n) is

$$p_\vartheta(y_1, \ldots, y_n) = \sum_{x_1=1}^{K} \cdots \sum_{x_n=1}^{K} p_\vartheta(x_1, \ldots, x_n, y_1, \ldots, y_n); \tag{1}$$

here, as well in the sequel, p is used as a generic symbol for densities. Looking at (1), one might think that the complexity for computing $p_\vartheta(y_1, \ldots, y_n)$ is exponential in n. Fortunately, we can compute the likelihood much faster by introducing the matrix $G_\vartheta(y) = \operatorname{diag}\{g_\vartheta(y|a)\}$ and noting that

$$p_\vartheta(y_1, \ldots, y_n) = \pi_\vartheta \left\{ \prod_{k=1}^{n} G_\vartheta(y_k) A_\vartheta \right\} \mathbf{1}, \tag{2}$$

where $A_\vartheta = \{\alpha_\vartheta(a, b)\}$ and $\mathbf{1}$ is a $K \times 1$-vector of ones. The computational complexity of (2) is only linear in n. A further useful observation is that conditional on the Y's, $\{X_k\}$ is still a Markov chain, although nonhomogeneous. It mixes geometrically fast, however, and this is the key to our analysis below.

The MLE, denoted by $\widehat{\vartheta}_n$, maximizes $p_\vartheta(Y_1, \ldots, Y_n)$ over the parameter set Θ. In many cases we may renumber the state space of $\{X_k\}$ and the g's, leaving the likelihood unchanged, and the MLE is then not unique. In particular we may do so if the usual parametrization is employed. This ambiguity is obviously not a big concern, though.

In practice, the MLE is often computed using the EM (expectation-maximization) algorithm; $\{X_k\}$ then play the role as missing data. In the context of HMMs, the EM algorithm was formulated by Baum and co-workers; see, for example, Baum, Petrie, Soules and Weiss (1970). A recent general reference is the monograph by McLachlan and Krishnan (1997). For HMMs with the usual parametrization, the M-step, in which the parameters are updated, is always explicit in the transition probabilities; that is, the new α's are obtained without a numerical search. If the parametric family $f(y;\phi)$ is an exponential family, the M-step is often explicit in the ϕ's as well. The E-step, in which conditional expectations are evaluated, is computationally more demanding. In most cases it is carried out using the so-called forward–backward algorithm, the complexity of which is linear in n; we refer to Rabiner (1989) and Leroux and Puterman (1992) for details. The major drawback of the EM algorithm is its rate of convergence, which is only linear in the vicinity of the MLE. Various modifications of the basic algorithm have been suggested to improve on this; see, for example, Jamshidian and Jennrich (1997), Meng and van Dyk (1997) and references therein. Little has been published on which of these modifications perform well for HMMs, however.

Alternatively, one may maximize (2) with respect to ϑ directly, using any standard numerical optimization scheme. The downhill simplex algorithm [see for example Press, Flannery, Teukolsky and Vetterling (1989)], is particularly attractive since it does not require any derivatives of the objective function, and derivatives of (2) are time-consuming to compute.

Whatever optimization algorithm is used, one always faces the problem that the likelihood surface of an HMM in general is multimodal. Any algorithm, including EM, may thus converge towards a local maximum or even a saddle point. Today there are no methods guaranteed to find the MLE, but the best advice available is to start the optimization algorithm from several different, possibly random, points in Θ.

2. Further notation and assumptions. The true parameter is denoted by ϑ_0. We deliberately replace the subindex ϑ_0 by '0' in notation like P_{ϑ_0} (becoming P_0) and so on. The $\mathbb{L}_q(P_0)$-norm will be denoted $||\cdot||_q$; that is, $||\cdot||_q = \{E_0|\cdot|^q\}^{1/q}$. Sometimes $Y_m, \ldots, Y_n$ will be abbreviated $\mathbf{Y}_m^n$, with an entirely similar notation for the X-process. The symbol D denotes differentiation with respect to ϑ, with D forming the gradient and D^2 forming the Hessian. Occasionally we will use a dot instead of D and two dots instead of D^2. Finally, C denotes a generic constant, finite and nonnegative, whose value may change from one expression to another.

The following assumptions will be referred to in the sequel.

(A1) The transition probability matrix $\{\alpha_0(a,b)\}$ is ergodic, that is, irreducible and aperiodic.

(A2) For all a and b, the maps $\vartheta \mapsto \alpha_\vartheta(a,b)$ and $\vartheta \mapsto \pi_\vartheta(a)$ have two continuous derivatives in some neighborhood $|\vartheta - \vartheta_0| < \delta$ of ϑ_0. For all a

and $y \in \mathscr{Y}$, the map $\vartheta \mapsto g_\vartheta(y|a)$ has two continuous derivatives in the same neighborhood.

(A3) Write $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$. There exists a $\delta > 0$ such that (i) for all $1 \le i \le d$ and all a,

$$E_0\left[\sup_{|\vartheta-\vartheta_0|<\delta}\left|\frac{\partial}{\partial\vartheta_i}\log g_\vartheta(Y_1|a)\right|^2\right] < \infty;$$

(ii) for all $1 \le i,\ j \le d$ and all a,

$$E_0\left[\sup_{|\vartheta-\vartheta_0|<\delta}\left|\frac{\partial^2}{\partial\vartheta_i\partial\vartheta_j}\log g_\vartheta(Y_1|a)\right|\right] < \infty;$$

(iii) for $j = 1, 2$, all $1 \le i_l \le d$, $l = 1, \ldots, j$, and all a,

$$\int \sup_{|\vartheta-\vartheta_0|<\delta}\left|\frac{\partial^j}{\partial\vartheta_{i_1}\cdots\partial\vartheta_{i_j}} g_\vartheta(y|a)\right| \nu(dy) < \infty.$$

(A4) There exists a $\delta > 0$ such that with

$$\rho_0(y) = \sup_{|\vartheta-\vartheta_0|<\delta} \max_{1\le a,\, b\le K} \frac{g_\vartheta(y|a)}{g_\vartheta(y|b)},$$

$P_0(\rho_0(Y_1) = \infty \mid X_1 = a) < 1$ for all a.

(A5) ϑ_0 is an interior point of Θ.

(A6) The maximum-likelihood estimator is strongly consistent.

Without loss of generality, we assume that the δ's in (A2)–(A4) agree.

REMARK. If (A1) holds, $\{X_k\}$ is ergodic under P_0. This implies that $\{Y_k\}$ is ergodic as well; see Leroux [(1992), page 130]. (A2) and (A3) are essentially regularity conditions of "Cramér type," that we cannot expect to weaken considerably. (A4) fails to hold if there are two g_0's with disjoint support; let, for example, the g's be location shifts of Beta densities. Heuristically, the result is a gain of information, however, rather than a loss, and it is possible that our results could be refined to include also this case.

In (A6) we assume that $\widehat{\vartheta}_n \to \vartheta_0$, P_0-a.s. as $n \to \infty$ (up to a possible permutation of states). Consistency of the MLE is discussed by Leroux (1992), and the conditions needed to ensure (A6) are essentially the following: (i) (A1); (ii) for all a and b, the map $\vartheta \mapsto \alpha_\vartheta(a, b)$ is continuous on Θ; (iii) for all a and $y \in \mathscr{Y}$, the map $\vartheta \mapsto g_\vartheta(y|a)$ is continuous on Θ; (iv) Θ is compact (this assumption can be relaxed somewhat; see Leroux's paper); (v) for each a, $E_0|\log g_0(Y_1|a)| < \infty$; (vi) For each a and ϑ there is a $\delta > 0$ such that $E_0[\sup_{|\vartheta'-\vartheta|<\delta}(\log g_{\vartheta'}(Y_1|a))^+] < \infty$; (vii) for each ϑ such that the laws P_ϑ and P_0 agree, $\vartheta = \vartheta_0$ (up to a possible permutation of states).

Obviously, conditions (ii), (iii) and (vi) are global, whereas conditions (A2)–(A4) are all local. Condition (vii) holds, for example, if the HMM has the usual

parametrization, finite mixtures of the parametric family $\{f(y;\phi)\}$ are identifiable and the ϕ_0's are distinct. Families of which finite mixtures are identifiable include the normal distribution, the Poisson distribution and the exponential distribution.

EXAMPLE 1 (Continued). We may define Θ by $\alpha(1,2), \alpha(2,1) \in [0,1]$, $\mu(a) \in [-1/\varepsilon, 1/\varepsilon]$, and $\sigma^2 \in [\varepsilon, 1/\varepsilon]$ for some small $\varepsilon > 0$. Conditions (A2)–(A4) are then all satisfied, as are the conditions for consistency listed above provided $\alpha_0(1,2), \alpha_0(2,1) \in (0,1)$ [implying (A1)].

EXAMPLE 2 (Continued). We define Θ by $\alpha(1,2), \alpha(2,1) \in [0,1]$ and $\mu(a) \in [0, 1/\varepsilon]$ for some small $\varepsilon > 0$. Then (A2)–(A4) and the consistency conditions are satisfied provided (A1) also holds.

EXAMPLE 3 (Continued). Define Θ by Q having off-diagonal elements bounded by $1/\varepsilon$ and $\lambda(a) \in [0, 1/\varepsilon]$ for some small $\varepsilon > 0$. Then (A2)–(A4) and the consistency conditions are satisfied provided (A1) also holds; it does if Q_0 is irreducible and all $\lambda_0(a) > 0$. Parameter estimation and consistency of the MLE are further discussed in Rydén (1994).

3. Main results. To prove asymptotic normality of the MLE, we need two lemmas which themselves are of considerable interest. These lemmas involve the loglikelihood, denoted by $L_n(\vartheta) = \log p_\vartheta(Y_1, \ldots, Y_n)$, and the Fisher information matrix for $\{Y_k\}$, denoted by $\mathscr{I}_0$. Intuitively, $\mathscr{I}_0$ may be thought of as the limiting covariance matrix of either $n^{-1/2}\dot{L}_n(\vartheta_0)$ or $D \log p_{\vartheta_0}(Y_n | Y_{n-1}, \ldots, Y_1)$. In Section 4 we show that both of these definitions are valid.

The first lemma is a central limit theorem for the score function at ϑ_0.

LEMMA 1. *Assume that* (A1)–(A4) *hold. Then* $n^{-1/2}\dot{L}_n(\vartheta_0) \to \mathscr{N}(0, \mathscr{I}_0)$ P_0-*weakly as* $n \to \infty$.

We prove this lemma in Section 4. The second lemma is a law of large numbers for the Hessian of the log likelihood.

LEMMA 2. *Assume that* (A1)–(A4) *hold and let* ϑ_n^* *be any, possibly stochastic, sequence in* Θ *such that* $\vartheta_n^* \to \vartheta_0$, P_0-*a.s. as* $n \to \infty$. *Then* $n^{-1}\ddot{L}_n(\vartheta_n^*) \to -\mathscr{I}_0$ *in* P_0-*probability as* $n \to \infty$.

This result will be proved in Section 5. Note that Lemma 2 shows that if (A1)–(A4) and (A6) hold, the observed information, that is $-n^{-1}\ddot{L}_n(\widehat{\vartheta}_n)$, converges to $\mathscr{I}_0$ in P_0-probability. The main result is now as follows.

THEOREM 1. *Assume that* (A1)–(A6) *hold and that* $\mathscr{I}_0$ *is nonsingular. Then* $n^{1/2}(\widehat{\vartheta}_n - \vartheta_0) \to \mathscr{N}(0, \mathscr{I}_0^{-1})$, P_0-*weakly as* $n \to \infty$.

PROOF. The proof essentially uses the approach introduced by Cramér. For n large enough, $\widehat{\vartheta}_n$ is an interior point of Θ and $|\widehat{\vartheta}_n - \vartheta_0| < \delta$, and we can then make a Taylor expansion of $\dot{L}_n$ about ϑ_0,

$$0 = \dot{L}_n(\widehat{\vartheta}_n) = \dot{L}_n(\vartheta_0) + \ddot{L}(\overline{\vartheta}_n)(\widehat{\vartheta}_n - \vartheta_0),$$

where $\overline{\vartheta}_n$ is a point on the line segment between ϑ_0 and $\widehat{\vartheta}_n$. Rewriting this expression, we obtain

$$n^{1/2}(\widehat{\vartheta}_n - \vartheta_0) = \left[-n^{-1}\ddot{L}_n(\overline{\vartheta}_n)\right]^{-1} n^{-1/2}\dot{L}_n(\vartheta_0).$$

The result now follows from the above lemmas. □

REMARK. Lemmas 1 and 2 also imply LAN of our model. In fact, they even imply uniform LAN, that is, that in the expansion

$$L_n(\vartheta_0 + n^{-1/2}u) - L_n(\vartheta_0) = n^{-1/2}u^T\dot{L}_n(\vartheta_0) + n^{-1}\tfrac{1}{2}u^T\ddot{L}_n(\vartheta_0)u + R_n(u),$$

$R_n(u)$ tends to zero in P_0-probability uniformly over compact subsets of $\mathbb{R}^d$. The superindex T denotes transpose.

Throughout the remainder of the paper, we shall make two assumptions that simplify the notation but do not remove any principal difficulties. The first assumption is that ϑ is one-dimensional, which saves us from using notation like uu^T. At one instance we do use this notation, namely, in the definition of the Fisher information matrix below. Our second assumption concerns the transition probabilities. By (A1), there exists a positive integer r such that all r-step transition probabilities $\alpha_0^{(r)}(a,b) = P_0(X_r = b \mid X_0 = a) > 0$. The assumption we make is that this inequality is satisfied with $r = 1$. We comment on the general case after Lemma 3.

4. A central limit theorem for the score function. Since the bivariate process $\{(X_k, Y_k)\}$ is stationary, we may extend it to a doubly infinite stationary sequence $\{(X_k, Y_k)\}_{k=-\infty}^{\infty}$, a feature that we will use frequently. Let $p_\vartheta(Y_1|Y_0, \ldots, Y_{-n})$ denote the conditional density of Y_1 given $Y_0, \ldots, Y_{-n}$. By the very definition of an HMM,

$$p_\vartheta(Y_1|\mathbf{Y}_{-n}^0) = \sum_{a=1}^{K} g_\vartheta(Y_1|a)P_\vartheta(X_1 = a \mid \mathbf{Y}_{-n}^0). \tag{3}$$

By a martingale convergence theorem by Lévy [see, e.g., Shiryayev (1984), page 478], $P_\vartheta(X_1 = a \mid \mathbf{Y}_{-n}^0) \to P_\vartheta(X_1 = a \mid \mathbf{Y}_{-\infty}^0)$ P_ϑ-a.s. as $n \to \infty$. Thus, if we define $p_\vartheta(Y_1|Y_0, Y_{-1}, \ldots)$ in analogy with (3), $p_\vartheta(Y_1|\mathbf{Y}_{-n}^0) \to p_\vartheta(Y_1|\mathbf{Y}_{-\infty}^0)$ P_ϑ-a.s.

Now, by a general identity for models with missing data [see Louis (1982), page 227], valid in our case because the X's take values in a finite set,

$$
\begin{aligned}
&D\log p_\vartheta(Y_1|Y_0,\dots,Y_{-n})\\
&\quad= D\log p_\vartheta(Y_{-n},\dots,Y_1) - D\log p_\vartheta(Y_{-n},\dots,Y_0)\\
&\quad= E_\vartheta[D\log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_1)\mid Y_{-n},\dots,Y_1]\\
&\qquad - E_\vartheta[D\log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_0)\mid Y_{-n},\dots,Y_0];
\end{aligned}
\tag{4}
$$

note that in the second term on the right-hand side, we consider X_1 as missing despite that Y_1 is not observed, a trick that will simplify the following computations slightly. Thus, writing $\lambda_\vartheta(a,b) = D\log\alpha_\vartheta(a,b)$, $\gamma_\vartheta(y|a) = D\log g_\vartheta(y|a)$, and $\tau_\vartheta(a) = D\log\pi_\vartheta(a)$, we have

$$
\begin{aligned}
&D\log p_{\vartheta_0}(Y_1|Y_0,\dots,Y_{-n})\\
&\quad= \sum_{k=-n}^{0}\Big\{E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid\mathbf{Y}_{-n}^1]\\
&\qquad - E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid\mathbf{Y}_{-n}^0]\Big\}\\
&\qquad + E_0[\gamma_0(Y_1|X_1)\mid\mathbf{Y}_{-n}^1] + E_0[\tau_0(X_{-n})\mid\mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n})\mid\mathbf{Y}_{-n}^0].
\end{aligned}
\tag{5}
$$

Define

$$
\begin{aligned}
\eta_1 = &\sum_{k=-\infty}^{0}\Big\{E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid\mathbf{Y}_{-\infty}^1]\\
&\qquad - E_0[\gamma_0(Y_k|X_k)+\lambda_0(X_k,X_{k+1})\mid\mathbf{Y}_{-\infty}^0]\Big\}\\
&+ E_0[\gamma_0(Y_1|X_1)\mid\mathbf{Y}_{-\infty}^1].
\end{aligned}
\tag{6}
$$

The sum in (6) is absolutely convergent in $\mathbb{L}_2(P_0)$, so that the right-hand side of (6) defines a random variable in $\mathbb{L}_2(P_0)$. We do not show this here, but it follows from the proof of Lemma 6 below. Under somewhat stronger conditions, the result $\eta_1\in\mathbb{L}_2(P_0)$ is shown in Lemma 2.3 in Bickel and Ritov (1996). We now define the Fisher information matrix as $\mathscr{I}_0 = E_0[\eta_1\eta_1^T]$. Before proving Lemma 1, we give some additional notation and lemmas.

Note that if (A1) and (A2) hold, there exist a $\delta>0$ and a $\sigma_0>0$ such that $\inf\{\alpha_\vartheta(a,b)\colon a,b,|\vartheta-\vartheta_0|<\delta\}\ge\sigma_0$, $\inf\{\alpha^*_\vartheta(a,b)\colon a,b,|\vartheta-\vartheta_0|<\delta\}\ge\sigma_0$ and $\inf\{\pi_\vartheta(a)\colon a,|\vartheta-\vartheta_0|<\delta\}\ge\sigma_0$, where $\alpha^*_\vartheta(a,b)=\pi_\vartheta(b)/\pi_\vartheta(a)\times\alpha_\vartheta(b,a)$ are the transition probabilities of the time-reversed version of $\{X_k\}$ (recall that we assume $r=1$). Without loss of generality, we assume that this δ agrees with the one in (A2)–(A4). Let

$$
\mu_0(y) = \big\{1+(K-1)\sigma_0^{-2}\rho_0(y)\big\}^{-1};
$$

if (A4) holds, $P_0(\mu_0(Y_1) > 0 \mid X_1 = a) > 0$ for all a. For further reference, we cite the following result from Bickel and Ritov (1996); it is their Lemma 3.3.

LEMMA 3. *Let $-n \le l < k \le 0$ and let H_k be an event defined in terms of $X_k, X_{k+1}, \ldots, X_0$ and $Y_k, Y_{k+1}, \ldots, Y_0$ only. Then for all ϑ such that $|\vartheta - \vartheta_0| < \delta$,*

$$\begin{aligned}\max_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_l = a) - \min_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_l = a) \\ \le \prod_{i=l+1}^{k-1} (1 - 2\mu_0(Y_i)) \\ \le \prod_{i=l+1}^{k-1} \exp(-2\mu_0(Y_i)).\end{aligned}$$

REMARK. If $r > 1$, the result corresponding to Lemma 3 (and with an entirely similar proof) reads

$$\begin{aligned}\max_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_{k-qr} = a) - \min_a P_\vartheta(H_k \mid \mathbf{Y}_{-n}^0, X_{k-qr} = a) \\ \le \prod_{i=2}^{q} \exp(-2\mu_0(Y_{k-ir+1}, \ldots, Y_{k-ir+2r-1})),\end{aligned} \tag{7}$$

where now

$$\mu_0(y_1, \ldots, y_{2r-1}) = \frac{1}{1 + (K-1)\sigma_0^{-2} \prod_{i=1}^{2r-1} \rho(y_i)},$$

and with σ_0 defined as above but in terms of the r-step transition probabilities. By deleting every second factor in (7) we obtain a bound with factors containing disjoint blocks of Y's. The proofs below then go through as when $r = 1$, except for some very minor changes caused by the need to work with the Y's in blocks of size r.

LEMMA 4. *Let $-n \le k \le 0$ and define*

$$S_\vartheta(n, k) = \max_{a,b,c} \left| P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^0,\ X_1 = b) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^0,\ X_1 = c) \right|.$$

Then, for any ϑ such that $|\vartheta - \vartheta_0| < \delta$,

$$S_\vartheta(n, k) \le \prod_{i=k+1}^{0} \exp(-2\mu_0(Y_i)).$$

The proof follows from Lemma 3 and the observation that the time-reversed version of $\{(X_k, Y_k)\}$ is an HMM as well.

LEMMA 5. *Let $-m \le -n \le k \le 0$. Then, for any ϑ such that $|\vartheta - \vartheta_0| < \delta$,*

$$\max_a \left|P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^0)\right| \le \prod_{i=k+1}^{0} \exp(-2\mu_0(Y_i)),$$

$$\max_{a,b} \left|P_\vartheta(X_k = a,\ X_{k+1} = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a,\ X_{k+1} = b \mid \mathbf{Y}_{-n}^0)\right| \le \prod_{i=k+2}^{0} \exp(-2\mu_0(Y_i)),$$

$$\max_a \left|P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)\right| \le \prod_{i=-n+1}^{k-1} \exp(-2\mu_0(Y_i)),$$

$$\max_{a,b} \left|P_\vartheta(X_k = a, X_{k+1} = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a,\ X_{k+1} = b \mid \mathbf{Y}_{-m}^1)\right| \le \prod_{i=-n+1}^{k-1} \exp(-2\mu_0(Y_i)).$$

The first two conclusions hold true P_ϑ-a.s. also if $-n$ is replaced by $-\infty$, and the last two conclusions hold true P_ϑ-a.s. also if $-m$ is replaced by $-\infty$.

In the last two parts we may also replace $\mathbf{Y}_{-n}^1$ and $\mathbf{Y}_{-m}^1$ by $\mathbf{Y}_{-n}^0$ and $\mathbf{Y}_{-m}^0$, respectively, and also, as above, extend these statements to infinite m.

PROOF. First assume that n and m are finite. The first part of the lemma can be proved using Lemma 4 and arguing as in (h), (i) and (j) in the proof of Lemma 2.3 in Bickel and Ritov (1996).

For the second part, note that

$$\begin{aligned}
&\left|P_\vartheta(X_k = a,\ X_{k+1} = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a,\ X_{k+1} = b \mid \mathbf{Y}_{-n}^0)\right| \\
&\quad = \left|P_\vartheta(X_k = a \mid X_{k+1} = b, \mathbf{Y}_{-n}^1) P_\vartheta(X_{k+1} = b \mid \mathbf{Y}_{-n}^1)\right. \\
&\qquad \left. - P_\vartheta(X_k = a \mid X_{k+1} = b, \mathbf{Y}_{-n}^0) P_\vartheta(X_{k+1} = b \mid \mathbf{Y}_{-n}^0)\right| \\
&\quad = \left|P_\vartheta(X_k = a \mid X_{k+1} = b, \mathbf{Y}_{-n}^k) P_\vartheta(X_{k+1} = b \mid \mathbf{Y}_{-n}^1)\right. \\
&\qquad \left. - P_\vartheta(X_k = a \mid X_{k+1} = b, \mathbf{Y}_{-n}^k) P_\vartheta(X_{k+1} = b \mid \mathbf{Y}_{-n}^0)\right| \\
&\quad \le \left|P_\vartheta(X_{k+1} = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_{k+1} = b \mid \mathbf{Y}_{-n}^0)\right|
\end{aligned}$$

and use the first part (for $k = 0$ this argument is not valid, but the result is then trivially true).

Since

$$\begin{aligned}
&\left|P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1)\right| \\
&\quad = \left| \sum_{b=1}^{K} P_\vartheta(X_k = a \mid X_{-n} = b, \mathbf{Y}_{-n+1}^1) P_\vartheta(X_{-n} = b \mid \mathbf{Y}_{-n}^1)\right. \\
&\qquad \left. - \sum_{c=1}^{K} P_\vartheta(X_k = a \mid X_{-n} = c, \mathbf{Y}_{-n+1}^1) P_\vartheta(X_{-n} = c \mid \mathbf{Y}_{-m}^1)\right|
\end{aligned}$$

$$\le \max_{b,c} \left| P_\vartheta(X_k = a \mid X_{-n} = b, \mathbf{Y}_{-n+1}^1) - P_\vartheta(X_k = a \mid X_{-n} = c, \mathbf{Y}_{-n+1}^1) \right|$$
$$\le \prod_{i=-n+1}^{k-1} \exp(-2\mu_0(Y_i)),$$

the third part holds; the last inequality follows from Lemma 3. When $\mathbf{Y}_{-n}^1$ and $\mathbf{Y}_{-m}^1$ are replaced by $\mathbf{Y}_{-n}^0$ and $\mathbf{Y}_{-m}^0$, respectively, the bound follows in a completely similar fashion.

The last part is proved using part three and an argument like the one used to prove part two. Finally, if n or m is infinite, use the fact that $P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1) \to P_\vartheta(X_k = a \mid \mathbf{Y}_{-\infty}^1)$ P_ϑ-a.s. and so on. □

We are now ready to prove the following result.

LEMMA 6. *There exist constants $\beta_0 \in [0, 1)$ and C_0 such that*

$$\|D \log p_{\vartheta_0}(Y_1 | Y_0, \dots, Y_{-n}) - \eta_1\|_2 \le C_0 \beta_0^n.$$

PROOF. Comparing (5) and (6), we see that it is sufficient to prove that there are $\beta_0 \in [0, 1)$ and C_0 such that

$$\left\| E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^0] \right\|_2 \le C_0 \beta_0^n, \tag{8}$$

$$\left\| E_0[\gamma_0(Y_1 | X_1) \mid \mathbf{Y}_{-n}^1] - E_0[\gamma_0(Y_1 | X_1) \mid \mathbf{Y}_{-\infty}^1] \right\|_2 \le C_0 \beta_0^n, \tag{9}$$

$$\left\| \sum_{k=-\lfloor n/2 \rfloor}^{0} \left\{ E_0[\gamma_0(Y_k | X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k | X_k) \mid \mathbf{Y}_{-\infty}^j] \right\} \right\|_2 \le C_0 \beta_0^n, \tag{10}$$

$$\left\| \sum_{k=-\lfloor n/2 \rfloor}^{0} \left\{ E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^j] - E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-\infty}^j] \right\} \right\|_2 \le C_0 \beta_0^n, \tag{11}$$

$$\left\| \sum_{k=-n}^{-\lfloor n/2 \rfloor - 1} \left\{ E_0[\gamma_0(Y_k | X_k) \mid \mathbf{Y}_{-n}^1] - E_0[\gamma_0(Y_k | X_k) \mid \mathbf{Y}_{-n}^0] \right\} \right\|_2 \le C_0 \beta_0^n, \tag{12}$$

$$\left\| \sum_{k=-n}^{-\lfloor n/2 \rfloor - 1} \left\{ E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^1] - E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-n}^0] \right\} \right\|_2 \le C_0 \beta_0^n, \tag{13}$$

$$\left\| \sum_{k=-\infty}^{-\lfloor n/2 \rfloor - 1} \left\{ E_0[\gamma_0(Y_k | X_k) \mid \mathbf{Y}_{-\infty}^1] - E_0[\gamma_0(Y_k | X_k) \mid \mathbf{Y}_{-\infty}^0] \right\} \right\|_2 \le C_0 \beta_0^n, \tag{14}$$

$$\left\| \sum_{k=-\infty}^{-\lfloor n/2 \rfloor - 1} \left\{ E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-\infty}^1] - E_0[\lambda_0(X_k, X_{k+1}) \mid \mathbf{Y}_{-\infty}^0] \right\} \right\|_2 \le C_0 \beta_0^n \tag{15}$$

for $j = 0, 1$, where $\lfloor \cdot \rfloor$ denotes the integer part.

We start with (8). By the first part of Lemma 5 we have

$$\begin{aligned}
&\bigl|E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^0]\bigr| \\
&\quad = \left| \sum_{a=1}^{K} \tau_0(a)\,[P_0(X_{-n} = a \mid \mathbf{Y}_{-n}^1) - P_0(X_{-n} = a \mid \mathbf{Y}_{-n}^0)] \right| \\
&\quad \le \max_a \tau_0(a)\, C \prod_{i=-n+1}^{0} \exp(-2\mu_0(Y_i)).
\end{aligned}$$

Thus, by the definition of an HMM,

$$\begin{aligned}
&\bigl|\bigl|E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^1] - E_0[\tau_0(X_{-n}) \mid \mathbf{Y}_{-n}^0]\bigr|\bigr|_2^2 \\
&\quad \le CE_0\left[\prod_{i=-n+1}^{0} \exp(-4\mu_0(Y_i)) \right] \\
&\quad = CE_0\left[E_0\left[\prod_{i=-n+1}^{0} \exp(-4\mu_0(Y_i)) \,\middle|\, \mathbf{X}_{-n+1}^0 \right] \right] \\
&\quad = CE_0\left[\prod_{i=-n+1}^{0} E_0\bigl[\exp(-4\mu_0(Y_i)) \,\big|\, X_i\bigr] \right] \\
&\quad \le CE_0\left[\prod_{i=-n+1}^{0} \max_a E_0\bigl[\exp(-4\mu_0(Y_i)) \,\big|\, X_i = a\bigr] \right] \\
&\quad = C\beta^n
\end{aligned}$$

for some $\beta \in [0, 1)$ and (8) follows. A similar argument shows (9).

We now turn to (10). By the third part of Lemma 5, with $m = \infty$,

$$\begin{aligned}
&\bigl|E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^j]\bigr| \\
&\quad = \left| \sum_{a=1}^{K} \gamma_0(Y_k|a)\,[P_0(X_k = a \mid \mathbf{Y}_{-n}^j) - P_0(X_k = a \mid \mathbf{Y}_{-\infty}^j)] \right| \\
&\quad \le \max_a |\gamma_0(Y_k|a)|\, C \prod_{i=-n+1}^{k-1} \exp(-2\mu_0(Y_i))
\end{aligned}$$

P_0-a.s. Thus,

$$\begin{aligned}
&\bigl|\bigl|E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^j]\bigr|\bigr|_2^2 \\
&\quad \le E_0\left[C \max_a |\gamma_0(Y_k|a)|^2 \prod_{i=-n+1}^{k-1} \exp(-4\mu_0(Y_i)) \right] \\
&\quad \le CE_0\left[E_0\left[\max_a |\gamma_0(Y_k|a)|^2 \prod_{i=-n+1}^{k-1} \exp(-4\mu_0(Y_i)) \,\middle|\, \mathbf{X}_{-n+1}^k \right] \right]
\end{aligned}$$

$$= CE_0\Big[E_0\Big[\max_a |\gamma_0(Y_k|a)|^2 \mid X_k\Big] \prod_{i=-n+1}^{k-1} E_0\big[\exp(-4\mu_0(Y_i)) \mid X_i\big]\Big]$$
$$\le C \max_b E_0\Big[\max_a |\gamma_0(Y_k|a)|^2 \mid X_k = b\Big]\beta^{k-1+n},$$

so that

$$\Big|\Big| \sum_{k=-\lfloor n/2\rfloor}^{0} \{E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-n}^j] - E_0[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^j]\}\Big|\Big|_2$$
$$\le C \sum_{k=-\lfloor n/2\rfloor}^{0} \beta^{(k-1+n)/2} \le C\beta^{(-\lfloor n/2\rfloor-1+n)/2},$$

and (10) follows. Also (11)–(15) follow in an entirely similar fashion, using other parts of Lemma 5. Note that (14) and (15) show that $\eta_1 \in \mathbb{L}_2(P_0)$. □

PROOF OF LEMMA 1. Let $\xi_k = D \log p_{\vartheta_0}(Y_k|Y_{k-1}, \ldots, Y_1)$, so that $\dot{L}_n(\vartheta_0) = \sum_{k=1}^n \xi_k$, and let

$$\eta_k = \sum_{i=-\infty}^{k-1} \Big\{E_0\big[\gamma_0(Y_i|X_i) + \lambda_0(X_i, X_{i+1}) \mid \mathbf{Y}_{-\infty}^k\big] - E_0\big[\gamma_0(Y_i|X_i) + \lambda_0(X_i, X_{i+1}) \mid \mathbf{Y}_{-\infty}^{k-1}\big]\Big\} + E_0\big[\gamma_0(Y_k|X_k) \mid \mathbf{Y}_{-\infty}^k\big].$$

Using (A3)(iii), it readily follows that

$$E_0\big[\gamma_0(Y_1|X_1) \mid \mathbf{Y}_{-\infty}^0\big] = E_0\big[E_0[\gamma_0(Y_1|X_1) \mid \mathbf{Y}_{-\infty}^0, X_1] \mid \mathbf{Y}_{-\infty}^0\big] = E_0\big[E_0[\gamma_0(Y_1|X_1) \mid X_1] \mid \mathbf{Y}_{-\infty}^0\big] = 0,$$

so that $\{\eta_k\}$ is a stationary and ergodic (because $\{Y_k\}$ is ergodic) martingale increment sequence with respect to $\{\sigma(Y_{-\infty}^k)\}$ in $\mathbb{L}_2(P_0)$. Its covariance matrix is $\mathscr{J}_0$. By the central limit theorem for martingales [see, e.g., Durrett (1991), page 375], we obtain

$$n^{-1/2} \sum_{k=1}^{n} \eta_k \to \mathscr{N}(0, \mathscr{J}_0). \tag{16}$$

Finally, Lemma 6 shows that

$$\Big|\Big|n^{-1/2} \sum_{k=1}^{n} \xi_k - n^{-1/2} \sum_{k=1}^{n} \eta_k\Big|\Big|_2 \le n^{-1/2} \sum_{k=1}^{n} ||\xi_k - \eta_k||_2 = n^{-1/2} \sum_{k=1}^{n} ||D \log p_{\vartheta_0}(Y_1|Y_0, \ldots, Y_{-k+2}) - \eta_1||_2,$$

where the last equality follows by stationarity. By Lemma 6, the expression on the right-hand side tends to zero as $n \to \infty$, whence the result follows from (16). □

5. A law of large numbers for the observed information. In this section we prove Lemma 2 via a uniform law of large numbers for the Hessian of the loglikelihood. Our approach is similar to the one used in Section 4, but the derivation is more delicate. First, again by a general identity for models with missing data [see Louis (1982), page 227], valid in our case because the X's take values in a finite set,

$$
\begin{aligned}
&D^2 \log p_\vartheta(Y_1|Y_0,\dots,Y_{-n}) \\
&\quad = D^2 \log p_\vartheta(Y_{-n},\dots,Y_1) - D^2 \log p_\vartheta(Y_{-n},\dots,Y_0) \\
&\quad = E_\vartheta\big[D^2 \log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_1) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad + E_\vartheta\big[(D \log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_1))^2 \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad - \big\{E_\vartheta\big[D \log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_1) \mid \mathbf{Y}_{-n}^1\big]\big\}^2 \\
&\qquad - E_\vartheta\big[D^2 \log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_0) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad - E_\vartheta\big[(D \log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_0))^2 \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad + \big\{E_\vartheta\big[D \log p_\vartheta(X_{-n},\dots,X_1,Y_{-n},\dots,Y_0) \mid \mathbf{Y}_{-n}^0\big]\big\}^2 \\
&\quad = \sum_{k=-n}^{0} \big\{E_\vartheta\big[\dot\gamma_\vartheta(Y_k|X_k) + \dot\lambda_\vartheta(X_k,X_{k+1}) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - E_\vartheta\big[\dot\gamma_\vartheta(Y_k|X_k) + \dot\lambda_\vartheta(X_k,X_{k+1}) \mid \mathbf{Y}_{-n}^0\big]\big\} \\
&\qquad + E_\vartheta\big[\dot\gamma_\vartheta(Y_1|X_1) \mid \mathbf{Y}_{-n}^1\big] + E_\vartheta\big[\dot\tau_\vartheta(X_{-n}) \mid \mathbf{Y}_{-n}^1\big] - E_\vartheta\big[\dot\tau_\vartheta(X_{-n}) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad + \sum_{k=-n}^{0}\sum_{l=-n}^{0} \Big\{E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad\qquad + E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^0\big]E_\vartheta\big[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad\qquad + E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1})\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1}) \mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1})\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad\qquad + E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1}) \mid \mathbf{Y}_{-n}^0\big]E_\vartheta\big[\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad\qquad + 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^1\big] \\
&\qquad\qquad - 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^0\big] \\
&\qquad\qquad + 2E_\vartheta\big[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-n}^0\big]E_\vartheta\big[\lambda_\vartheta(X_l,X_{l+1}) \mid \mathbf{Y}_{-n}^0\big]\Big\} \\
&\qquad + E_\vartheta\big[\gamma_\vartheta^2(Y_1|X_1) \mid \mathbf{Y}_{-n}^1\big] - \big\{E_\vartheta\big[\gamma_\vartheta(Y_1|X_1) \mid \mathbf{Y}_{-n}^1\big]\big\}^2 \qquad (17)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{k=-n}^{0}\Big\{2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad -2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad +2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\lambda_\vartheta(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad -2E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^1\big]\Big\}\\
&+E_\vartheta\big[\tau_\vartheta^2(X_{-n})\mid \mathbf{Y}_{-n}^1\big]-\big\{E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^1\big]\big\}^2\\
&-E_\vartheta\big[\tau_\vartheta^2(X_{-n})\mid \mathbf{Y}_{-n}^0\big]+\big\{E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^0\big]\big\}^2\\
&+\sum_{k=-n}^{0}\Big\{2E_\vartheta\big[\tau_\vartheta(X_{-n})\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^0\big]\\
&\qquad +2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^0\big]E_\vartheta\big[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^0\big]\\
&\qquad +2E_\vartheta\big[\tau_\vartheta(X_{-n})\lambda_\vartheta(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^1\big]\\
&\qquad -2E_\vartheta\big[\tau_\vartheta(X_{-n})\lambda_\vartheta(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^0\big]\\
&\qquad +2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^0\big]E_\vartheta\big[\lambda_\vartheta(X_k,X_{k+1})\mid \mathbf{Y}_{-n}^0\big]\Big\}\\
&+2E_\vartheta\big[\tau_\vartheta(X_{-n})\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^1\big]\\
&-2E_\vartheta\big[\tau_\vartheta(X_{-n})\mid \mathbf{Y}_{-n}^1\big]E_\vartheta\big[\gamma_\vartheta(Y_1|X_1)\mid \mathbf{Y}_{-n}^1\big].
\end{aligned}
$$

Again, we need some additional lemmas before we look closer at this expression.

LEMMA 7. *Let $-m\le -n\le k$, $l\le 0$. Then for any ϑ such that $|\vartheta-\vartheta_0|<\delta$,*

$$
\begin{aligned}
&\max_{a,b}\big|P_\vartheta(X_k=a,\ X_l=b\mid \mathbf{Y}_{-n}^1)-P_\vartheta(X_k=a,\ X_l=b\mid \mathbf{Y}_{-n}^0)\big|\\
&\qquad\le\prod_{i=k\vee l+1}^{0}\exp(-2\mu_0(Y_i)),\\
&\max_{a,b}\big|P_\vartheta(X_k=a,\ X_l=b\mid \mathbf{Y}_{-n}^1)-P_\vartheta(X_k=a,\ X_l=b\mid \mathbf{Y}_{-m}^1)\big|\\
&\qquad\le\prod_{i=-n+1}^{k\wedge l-1}\exp(-2\mu_0(Y_i)).
\end{aligned}
$$

The second conclusion holds true also if $\mathbf{Y}_{-n}^1$ and $\mathbf{Y}_{-m}^1$ are replaced by $\mathbf{Y}_{-n}^0$ and $\mathbf{Y}_{-m}^0$, respectively.

The proof is entirely similar to the proofs of parts two and four of Lemma 5.

LEMMA 8. *Let $-n \le k, l \le 0$. Then for any ϑ such that $|\vartheta - \vartheta_0| < \delta$,*

$$\max_{a,b}\left|P_\vartheta(X_k = a,\ X_l = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1)\right| \le \prod_{i=k\wedge l+1}^{k\vee l-1} \exp(-2\mu_0(Y_i)).$$

The conclusion holds true also if $\mathbf{Y}_{-n}^1$ is replaced by $\mathbf{Y}_{-n}^0$.

PROOF. Assume that $k \ge l$. Then

$$\begin{aligned}
&\left|P_\vartheta(X_k = a,\ X_l = b \mid \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1)\right| \\
&\quad = \Big|P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1)P_\vartheta(X_l = b \mid \mathbf{Y}_{-n}^1)\Big| \\
&\quad \le \left|P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-n}^1)\right| \\
&\quad = \Big|\sum_{c=1}^{K}[P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid X_l = c, \mathbf{Y}_{-n}^1)]P_\vartheta(X_l = c \mid \mathbf{Y}_{-n}^1)\Big| \\
&\quad \le \max_{a,b,c}\left|P_\vartheta(X_k = a \mid X_l = b, \mathbf{Y}_{-n}^1) - P_\vartheta(X_k = a \mid X_l = c, \mathbf{Y}_{-n}^1)\right| \\
&\quad \le \prod_{i=l+1}^{k-1} \exp(-2\mu_0(Y_i)),
\end{aligned}$$

where the last inequality follows from Lemma 3. The proof with $\mathbf{Y}_{-n}^0$ is analogous. □

Let G denote the neighborhood $\{\vartheta\colon |\vartheta - \vartheta_0| < \delta\}$ of ϑ_0.

LEMMA 9. *As $m, n \to \infty$,*

$$\left|\left|\sup_{\vartheta\in G}\left|D^2 \log p_\vartheta(Y_1|\mathbf{Y}_{-m}^1) - D^2 \log p_\vartheta(Y_1|\mathbf{Y}_{-n}^1)\right|\right|\right|_1 \to 0.$$

PROOF. Considering (17), we see that we must prove, for example,

$$\begin{aligned}
\left|\left|\sup_{\vartheta\in G}\right|\sum_{k=-m}^{0}\sum_{l=-m}^{0}\Big\{&E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^1] \\
&- E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^1]E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^1] \\
&- E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^0] \\
&+ E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^0]E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^0]\Big\} \qquad (18)
\end{aligned}$$

$$
\begin{aligned}
-\sum_{k=-n}^{0}\sum_{l=-n}^{0}\Big\{ & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^1]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^0] \\
& + E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^0]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^0]\Big\}\Big|\Big|\Big|_1 \to 0
\end{aligned}
$$

as $m, n \to \infty$. Other statements, similar to (18) and which together with (18) prove the lemma, can be shown using slight variations of the technique used below. In order to prove (18), it is sufficient to show that (assuming $m \geq n$) for $j = 0, 1$,

$$
\begin{aligned}
\sum_{k=-m}^{-\lfloor n/2\rfloor}\sum_{l=k}^{\lfloor k/2\rfloor}\Big|\Big| \sup_{\vartheta\in G}\big| & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^0]\big|\Big|\Big|_1 \to 0,
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
\sum_{k=-m}^{-\lfloor n/2\rfloor}\sum_{l=k}^{\lfloor k/2\rfloor}\Big|\Big| \sup_{\vartheta\in G}\big| & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-m}^1]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-m}^0]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^0]\big|\Big|\Big|_1 \to 0,
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
\sum_{k=-n}^{-\lfloor n/2\rfloor}\sum_{l=k}^{\lfloor k/2\rfloor}\Big|\Big| \sup_{\vartheta\in G}\big| & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^0]\big|\Big|\Big|_1 \to 0,
\end{aligned} \tag{21}
$$

$$
\begin{aligned}
\sum_{k=-n}^{-\lfloor n/2\rfloor}\sum_{l=k}^{\lfloor k/2\rfloor}\Big|\Big| \sup_{\vartheta\in G}\big| & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^1]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^1] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^0]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^0]\big|\Big|\Big|_1 \to 0,
\end{aligned} \tag{22}
$$

$$
\begin{aligned}
\sum_{k=-\lfloor n/2\rfloor}^{0}\sum_{l=-\lfloor n/2\rfloor}^{0}\Big|\Big| \sup_{\vartheta\in G}\big| & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^j] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^j]\big|\Big|\Big|_1 \to 0,
\end{aligned} \tag{23}
$$

$$
\begin{aligned}
\sum_{k=-\lfloor n/2\rfloor}^{0}\sum_{l=-\lfloor n/2\rfloor}^{0}\Big|\Big| \sup_{\vartheta\in G}\big| & E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-m}^j]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^j] \\
& - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^j]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^j]\big|\Big|\Big|_1 \to 0,
\end{aligned} \tag{24}
$$

$$\sum_{k=-m}^{-\lfloor n/2\rfloor}\sum_{l=-\lfloor k/2\rfloor}^{0}\Big|\Big|\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{j}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-m}^{j}]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{j}]\big|\Big|\Big|_1 \to 0, \tag{25}$$

$$\sum_{k=-n}^{-\lfloor n/2\rfloor}\sum_{l=-\lfloor k/2\rfloor}^{0}\Big|\Big|\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^{j}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\mid \mathbf{Y}_{-n}^{j}]E_\vartheta[\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-n}^{j}]\big|\Big|\Big|_1 \to 0, \tag{26}$$

as $m, n \to \infty$; compare Figure 1. The idea of splitting up the sum (18) goes back to Baum and Petrie (1966).

Starting with (19), by the first part of Lemma 7 we have that

$$\begin{aligned}
&\sup_{\vartheta\in G}\big|E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{1}] - E_\vartheta[\gamma_\vartheta(Y_k|X_k)\gamma_\vartheta(Y_l|X_l)\mid \mathbf{Y}_{-m}^{0}]\big| \\
&\quad\le \sup_{\vartheta\in G}\sum_{a,\,b=1}^{K}|\gamma_\vartheta(Y_k|a)|\,|\gamma_\vartheta(Y_l|b)|\,\big|P_\vartheta(X_k=a,\ X_l=b\mid \mathbf{Y}_{-m}^{1}) \\
&\qquad\qquad - P_\vartheta(X_k=a, X_l=b\mid \mathbf{Y}_{-m}^{0})\big| \\
&\quad\le C\Big(\sup_{\vartheta\in G}\max_a|\gamma_\vartheta(Y_k|a)|\Big)\Big(\sup_{\vartheta\in G}\max_a|\gamma_\vartheta(Y_l|b)|\Big)\prod_{i=k\vee l+1}^{0}\exp(-2\mu_0(Y_i)).
\end{aligned}$$

By conditioning on the X's, we obtain that the $\mathbb{L}_1(P_0)$-norm of the above expression is bounded by $C\beta^{|k|\wedge|l|}$ for some $\beta\in[0,1)$, whence the left-hand

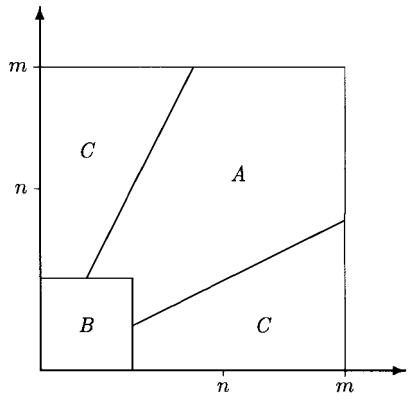

FIG. 1. *Illustration of how the sum in* (18) *is split into subregions. In region A,* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]$ *is compared to* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{0}]$ *etc. In region B,* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]$ *is compared to* $E_\vartheta[\cdot\mid \mathbf{Y}_{-n}^{1}]$ *etc. In region C,* $E_\vartheta[\cdot\times\cdot\mid \mathbf{Y}_{-m}^{1}]$ *is compared to* $E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]\times E_\vartheta[\cdot\mid \mathbf{Y}_{-m}^{1}]$ *and so on.*

side of (19) is bounded by

$$C \sum_{k=\lfloor n/2 \rfloor}^{m} \sum_{l=\lfloor k/2 \rfloor}^{m} \beta^{l} \le C \sum_{k=\lfloor n/2 \rfloor}^{m} \beta^{\lfloor k/2 \rfloor} \le C \beta^{\lfloor n/4 \rfloor}.$$

Here, the right-hand side tends to zero as $m, n \to \infty$, and (19) follows; (21) follows similarly.

For (20), the first part of Lemma 5 shows that for any $\vartheta \in G$,

$$\begin{aligned}
&\max_{a,b} \big| P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) \\
&\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \big| \\
&\le \max_{a,b} \big| P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \\
&\qquad\qquad + P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \\
&\qquad\qquad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \big| \\
&\le \max_b |P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) - P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0)| \\
&\qquad + \max_a |P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0)| \\
&\le 2 \prod_{i=k \vee l+1}^{0} \exp(-2\mu_0(Y_i)),
\end{aligned} \tag{27}$$

so that

$$\begin{aligned}
&\sup_{\vartheta \in G} \big| E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^1] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^1] \\
&\qquad - E_\vartheta[\gamma_\vartheta(Y_k|X_k) \mid \mathbf{Y}_{-m}^0] E_\vartheta[\gamma_\vartheta(Y_l|X_l) \mid \mathbf{Y}_{-m}^0] \big| \\
&\le \sup_{\vartheta \in G} \sum_{a,b=1}^{K} |\gamma_\vartheta(Y_k|a)|\, |\gamma_\vartheta(Y_l|b)| \\
&\qquad\qquad \times \big| P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^1) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^1) \\
&\qquad\qquad\quad - P_\vartheta(X_k = a \mid \mathbf{Y}_{-m}^0) P_\vartheta(X_l = b \mid \mathbf{Y}_{-m}^0) \big| \\
&\le C \Big(\sup_{\vartheta \in G} \max_a |\gamma_\vartheta(Y_k|a)| \Big) \Big(\sup_{\vartheta \in G} \max_a |\gamma_\vartheta(Y_l|a)| \Big) \prod_{i=k \vee l+1}^{0} \exp(-2\mu_0(Y_i)).
\end{aligned}$$

Now (20) follows as above, and (22) follows similarly.

Further, the second part of Lemma 7 shows that the left-hand side of (23) is bounded by

$$C \sum_{k=-\lfloor n/2\rfloor}^{0} \sum_{l=-\lfloor n/2\rfloor}^{0} \beta^{n+k\wedge l-1} = C \sum_{k=0}^{\lfloor n/2\rfloor} \sum_{l=0}^{\lfloor n/2\rfloor} \beta^{n-k\vee l-1}$$
$$\leq 2C \sum_{k=0}^{\lfloor n/2\rfloor} \sum_{l=k}^{\lfloor n/2\rfloor} \beta^{n-l-1}$$
$$\leq C \sum_{k=0}^{\lfloor n/2\rfloor} \beta^{\lfloor n/2\rfloor} \leq C(\lfloor n/2\rfloor + 1)\beta^{\lfloor n/2\rfloor}.$$

The right-hand side vanishes as $n \to \infty$, whence (23) follows; (24) follows using a bound similar to (27).

Finally, by Lemma 8 the left-hand side of (25) is bounded by

$$C \sum_{k=-m}^{-\lfloor n/2\rfloor} \sum_{l=\lfloor k/2\rfloor}^{0} \beta^{k\vee l-k\wedge l-1} = C \sum_{k=\lfloor n/2\rfloor}^{m} \sum_{l=0}^{\lfloor k/2\rfloor} \beta^{k\vee l-k\wedge l-1}$$
$$= C \sum_{k=\lfloor n/2\rfloor}^{m} \sum_{l=0}^{\lfloor k/2\rfloor} \beta^{k-l-1}$$
$$\leq C \sum_{k=\lfloor n/2\rfloor}^{m} \beta^{k-\lfloor k/2\rfloor-1} \leq C\beta^{\lfloor n/4\rfloor},$$

whence (25) follows; (26) follows similarly, and the proof is complete. □

Thus, $\{D^2 \log p_\vartheta(Y_1|Y_0, \ldots, Y_{-n})\}$ is a "uniform Cauchy sequence" in $\mathbb{L}_1(P_0)$, and the following result is then immediate.

LEMMA 10. *There is a continuous function $\zeta_1(\vartheta)$ from G to $\mathbb{L}_1(P_0)$ such that*

$$\Big|\Big| \sup_{\vartheta\in G} |D^2 \log p_\vartheta(Y_1|Y_0, \ldots, Y_{-n}) - \zeta_1(\vartheta)| \Big|\Big|_1 \to 0$$

as $n \to \infty$.

REMARK. Assuming the MLE to be consistent, that is, that (A6) holds, any subset of the sample space with P_ϑ-measure one for some $\vartheta \neq \vartheta_0$ has P_0-measure zero, whence Lemma 5 does not guarantee that any of the statements with infinite n or m holds P_0-a.s. for any ϑ other than ϑ_0. This is the reason for working with Cauchy sequences in this section, rather than with an explicit representation of $\zeta_1(\vartheta)$ similar to (6).

PROOF OF LEMMA 2. Define $\zeta_k(\vartheta)$ as the $\mathbb{L}_1(P_0)$-limit of

$$D^2 \log p_\vartheta(Y_k|\mathbf{Y}_{-n}^{k-1})$$

and let G' be an arbitrary neighborhood of ϑ_0 such that $G' \subseteq G$. We then have

$$\begin{aligned}
&\limsup_{n\to\infty} P_0\biggl(\Bigl|n^{-1}\ddot{L}_n(\vartheta_n^*) - n^{-1}\sum_{k=1}^n \zeta_k(\vartheta_0)\Bigr| > \varepsilon\biggr)\\
&\quad= \limsup_{n\to\infty} P_0\biggl(\Bigl|n^{-1}\sum_{k=1}^n \bigl\{D^2 \log p_{\vartheta_n^*}(Y_k|Y_{k-1},\ldots,Y_1) - \zeta_k(\vartheta_0)\bigr\}\Bigr| > \varepsilon\biggr)\\
&\quad\le \limsup_{n\to\infty} P_0\biggl(n^{-1}\sum_{k=1}^n \sup_{\vartheta\in G'}\bigl|D^2 \log p_{\vartheta}(Y_k|Y_{k-1},\ldots,Y_1) - \zeta_k(\vartheta_0)\bigr| > \varepsilon\biggr)\\
&\qquad+ \limsup_{n\to\infty} P_0(\vartheta_n^* \notin G')\\
&\quad\le \limsup_{n\to\infty} n^{-1}\varepsilon^{-1}\sum_{k=1}^n \Bigl|\Bigl| \sup_{\vartheta\in G'}\bigl|D^2 \log p_{\vartheta}(Y_1|Y_0,\ldots,Y_{-k+2}) - \zeta_1(\vartheta_0)\bigr| \Bigr|\Bigr|_1\\
&\quad\le \limsup_{n\to\infty} n^{-1}\varepsilon^{-1}\sum_{k=1}^n \Bigl|\Bigl| \sup_{\vartheta\in G'}\bigl|D^2 \log p_{\vartheta}(Y_1|Y_0,\ldots,Y_{-k+2}) - \zeta_1(\vartheta)\bigr| \Bigr|\Bigr|_1\\
&\qquad+ \limsup_{n\to\infty} n^{-1}\varepsilon^{-1}\sum_{k=1}^n \Bigl|\Bigl| \sup_{\vartheta\in G'}\bigl|\zeta_1(\vartheta) - \zeta_1(\vartheta_0)\bigr| \Bigr|\Bigr|_1\\
&\quad= \varepsilon^{-1}\Bigl|\Bigl| \sup_{\vartheta\in G'}\bigl|\zeta_1(\vartheta) - \zeta_1(\vartheta_0)\bigr| \Bigr|\Bigr|_1,
\end{aligned}$$

where the third step follows by Markov's inequality and stationarity, and the last one by Lemma 10. Let $G' \downarrow \{\vartheta_0\}$ and use continuity of $\zeta(\cdot)$ to conclude that

$$n^{-1}\ddot{L}_n(\vartheta^*) - n^{-1}\sum_{k=1}^n \zeta_k(\vartheta_0) \to 0 \text{ in } P_0\text{-probability} \tag{28}$$

as $n \to \infty$.

Now, because $\{Y_k\}$ is ergodic, so is $\{\zeta_k(\vartheta_0)\}$, whence $n^{-1}\sum_1^n \zeta_k(\vartheta_0) \to J$ P_0-a.s. for some matrix $J = E_0\zeta_1(\vartheta_0)$. The proof is thus complete if we can show that $J = -\mathscr{J}_0$.

Using (A3)(iii) it readily follows that

$$E_0\bigl[-D^2 \log g_{\vartheta_0}(Y_1|X_1)\bigr] = E_0\bigl[(D \log g_{\vartheta_0}(Y_1|X_1))^2\bigr],$$

which together with the representations (4) and (17) show that

$$E_0\bigl[D^2 \log p_{\vartheta_0}(Y_1|Y_0,\ldots,Y_{-n})\bigr] = -E_0\bigl[\bigl(D \log p_{\vartheta_0}(Y_1|Y_0,\ldots,Y_{-n})\bigr)^2\bigr]$$

for each n. Hence, by Lemma 6 and Lemma 10, $J = -\mathscr{J}_0$. □

Acknowledgments. Many thanks to Jens Ledet Jensen and Niels Væver Petersen, who did not only carefully read an earlier version of this paper and found four errors (in Assumption A2 and the proofs of Lemmas 1, 2 and 9), but who also provided solutions to these errors.

REFERENCES

ALBERT, P. S. (1991). A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics* **47** 1371–1381.

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

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

BICKEL, P. J. and RITOV, Y. (1996). Inference in hidden Markov models I: local asymptotic normality in the stationary case. *Bernoulli* **2** 199–228.

DURRETT, R. (1991). *Probability: Theory and Examples*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

FREDKIN, D. R. and RICE, J. A. (1992). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Royal Soc. London Ser. B* **249** 125–132.

GUTTORP, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman & Hall, London.

HEFFES, H. and LUCANTONI, D. (1986). A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Select. Areas Comm.* **4** 856–867.

JAMSHIDIAN, M. and JENNRICH, R. I. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *J. Royal Statist. Soc. Ser. B* **59** 569–587.

LE, N. D., LEROUX, B. G. and PUTERMAN, M. L. (1992). Reader reaction: exact likelihood evaluation in a Markov mixture model for time series of seizure counts. *Biometrics* **48** 317–323.

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

LEROUX, B. G. and PUTERMAN, M. L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48** 545–558.

LINDGREN, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Statist.* **5** 81–91.

LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. Royal Statist. Soc. Ser. B* **44** 226–233.

MACDONALD, I. L. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall, London.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a new fast tune (with discussion). *J. Royal Statist. Soc. Ser. B* **59** 511–567.

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

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1989). *Numerical Recipes*. Cambridge Univ. Press.

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

RITOV, Y. (1996). Uniform convergence of quasi-convex functions with applications to missing data and hidden Markov models. Preprint.

RYDÉN, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Stochastic Models* **10** 795–829.

SHIRYAYEV, A. N. (1984). *Probability*. Springer, New York.

P. J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
EVANS HALL
BERKELEY, CALIFORNIA 94720

Y. RITOV
DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM 91905
ISRAEL

T. RYDÉN
DEPARTMENT OF MATHEMATICAL STATISTICS
LUND UNIVERSITY
BOX 118
S-221 00 LUND
SWEDEN
E-MAIL: tobias@maths.lth.se