Statistica Sinica **14**(2004), 127-154

ON BAHADUR EFFICIENCY OF THE MAXIMUM LIKELIHOOD ESTIMATOR IN HIDDEN MARKOV MODELS

Cheng-Der Fuh

Academia Sinica, Taipei

Abstract: In this paper, we study large deviations of maximum likelihood and related estimators for hidden Markov models. A hidden Markov model consists of parameterized Markov chains in a Markovian random environment, with the underlying environmental Markov chain viewed as missing data. A difficulty with parameter estimation in this model is the non-additivity of the log-likelihood function. Based on a device used to represent the likelihood function as the L_1-norm of products of Markov random matrices, we investigate the tail probabilities for consistent estimators in hidden Markov models. The main result is that, under some regularity conditions, the maximum likelihood estimator is an asymptotically locally optimal estimator in Bahadur's sense. The results are applied to several types of hidden Markov models commonly used in speech recognition, molecular biology and economics.

Key words and phrases: Consistency, efficiency, hidden Markov models, large deviations, maximum likelihood, missing data, products of random matrices.

1. Introduction

A hidden Markov model (HMM) is, loosely speaking, a sequence $\{\xi_n\}_{n=1}^{\infty}$ of random variables obtained in the following way. First, a realization of a finite state Markov chain $\mathbf{X} = \{X_n\}$ is created. This chain is sometimes called the regime and is not observed. Then, conditioned on $\{X_n\}$, the ξ-variables are generated. Usually, the dependency of ξ_n on $\mathbf{X}$ is more or less local, as when $\xi_n = h(X_n, X_{n+1}, \eta_n)$ for some function h and random sequence $\{\eta_n\}$, independent of $\mathbf{X}$. ξ_n itself is generally not Markov and may, in fact, have a complicated dependency structure. A formal definition will be given at the end of this section.

The combination of rich probability structure and useful statistical analysis makes hidden Markov models a common tool for modeling dependent random variables, with applications in areas such as speech recognition (cf. Rabiner and Juang (1993)), signal processing (cf. Elliott, Aggoun and Moore (1995)), ion channels (cf. Ball and Rice (1992)), molecular biology (cf. Krogh, Brown, Mian, Sjolander and Haussler (1994)) and economics (cf. Hamilton (1994)). A good summary of these examples can be found in Künsch (2001). The main focus of these efforts has been state space estimation, algorithms for fitting these models, and the implementation of likelihood based methods. Sometimes the physical nature of the problem suggests the use of a hidden Markov model, while in other cases hidden Markov models simply provide a good fit to the observed data.

An important early work on the inferential problem for hidden Markov models is the 1966 paper of Baum and Petrie. They established the consistency and asymptotic normality of the maximum likelihood estimator (MLE) for a hidden Markov chain in the case where the observation is a deterministic function of the state space. Based on the results of Furstenberg and Kesten (1960) and Kingman's sub-additive ergodic theorem (1976), Leroux (1992) established the consistency of the MLE for general hidden Markov chains under mild conditions. By adding a few essential ideas to the penetrating analysis of Baum and Petrie (1966), Bickel and Ritov (1996) showed that the log likelihood for hidden Markov models obeys the local asymptotic normality condition of LeCam. Bickel, Ritov and Rydén (1998) later proved the asymptotic normality of the MLE under some regularity conditions.

A difficulty with analyzing hidden Markov models is that the likelihood function can only be expressed in additive form (cf. equation (1.8)). Fuh (1998) introduced a device used to represent the likelihood function as the L_1-norm of products of Markov random matrices in order to prove the existence of a consistent sequence of roots of the likelihood equations that is asymptotically efficient. Fuh (2003) also proved the asymptotic optimality of SPRT and CUSUM in hidden Markov models. This new representation enables us to apply limiting theorems in that area, to verify the asymptotic properties of the MLE in hidden Markov models.

In this paper, we study the properties of efficient parameter estimation for a general hidden Markov model. In contrast to the existence and construction of estimates that are optimal according to the asymptotic variance criterion (cf. Bickel and Ritov (1996); Fuh (1998); Bickel, Ritov and Rydén (1998)), the optimal criterion will be based on the inaccuracy rate. Thus, let $\xi_1, \xi_2, \ldots$ be a sequence of random variables, with the distribution determined by a parameter θ taking values in a parameter space Θ. Let h be a function on Θ into a metric space Γ with metric d, and assume that it is required to estimate h. For each n, let $T_n = T_n(\xi_1, \ldots, \xi_n)$ be an estimate, and for $\varepsilon > 0$ let

$$\alpha_n := \alpha_n(\varepsilon, \theta) = P_\theta\{d(T_n, h(\theta)) > \varepsilon\}. \tag{1.1}$$

Assume that T_n is consistent for h, i.e., $\alpha_n \to 0$ as $n \to \infty$ for each $\varepsilon > 0$ and $\theta \in \Theta$. In many important situations, $\alpha_n \to 0$ exponentially fast. A consistent estimator T_n is said to be locally optimal if

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{\varepsilon^2 n}\log\alpha_n(\varepsilon,\theta)=-\frac{I_h(\theta)}{2}, \tag{1.2}$$

where $I_h(\theta)$ is the Fisher information for estimating h. Note that the definition of local optimality in (1.2) is in the strict sense, the general definition of locally optimality can be found in Bahadur, Zabell and Gupta (1980) and Shen (2001).

When the ξ_n are independent and identically distributed random variables, Bahadur (1960) obtained global and local bounds for the best possible rate by applying the Neyman-Pearson Lemma. It was also shown by Bahadur (1960) that, under some regularity conditions, the local bound is attained by the MLE for small ε. Further investigations along this line were conducted by Bahadur (1967, 1983), Fu (1973, 1975, 1982), Bahadur, Zabell and Gupta (1980), Rukhin (1983), Kester (1985) and others. Shen (2001) generalized Bahadur's efficiency to general parameter spaces and discussed many recent developments.

Beside the well known results that the MLE for estimation based on independent and identically distributed (i.i.d.) observations $\{\xi_n, n\geq 0\}$ is efficient in the sense of (1.2), Bahadur (1983) generalized it to a finite state Markov chain. It has remained an open problem whether the MLE has the same optimality properties when $\{\xi_n, n\geq 0\}$ is a general state Markov chain or a hidden Markov model. The contribution of this paper is to provide a general framework for HMM, and to give sufficient conditions for asymptotic optimality of the MLE in the sense of (1.2) (Theorems 1 and 2). Thus, we answer a long-standing question and we illustrate the usefulness of the results.

The rest of this paper is organized as follows. We first give a formal definition of generalized hidden Markov models and provide a representation of the likelihood function. In Section 2, we give a brief summary of products of Markov random matrices and prove a large deviation theorem. In Section 3, we define Kullback-Leibler and Fisher information, and then provide sufficient conditions such that the MLE is locally optimal. Several examples of hidden Markov models commonly used in speech recognition, molecular biology and economics are illustrated in Section 4. Proofs are given in Section 5.

A hidden Markov model is defined as a parameterized Markov chain in a Markovian random environment (cf. Cogburn (1980)) with the underlying environmental Markov chain viewed as missing data. This setting generalizes the hidden Markov models considered by Leroux (1992), Bickel and Ritov (1996), Fuh (1998) and Bickel, Ritov and Rydén (1998), in order to cover several interesting examples of Gaussian regression and Gaussian autoregression studied by Rabiner and Juang (1993), Hamilton (1994) and Merhav (1991). We consider $\mathbf{X}=\{X_n, n\geq 0\}$ as a Markov chain on a finite state space $D=\{1,\ldots,d\}$, with transition matrix $P(\theta) = [p_{xy}(\theta)]_{x,y=1,\ldots,d}$ and stationary distribution $\pi(\theta) = (\pi_x(\theta))_{x=1,\ldots,d}$, where $\theta \in \Theta \subseteq R^q$ denotes an unknown parameter. Suppose that a random sequence $\{\xi_n\}_{n=0}^{\infty}$, taking values in R, is adjoined to the chain such that $\{(X_n, \xi_n), n \geq 0\}$ is a Markov chain on $D \times R$ and, conditioned on the full $\mathbf{X}$ sequence, ξ_n is a Markov chain with

$$P_\theta\{\xi_{n+1} \in B | X_0, X_1, \ldots; \xi_1, \ldots, \xi_n\} = P_\theta(X_{n+1} : \xi_n, B) \ a.s. \tag{1.3}$$

for each n and $B \in \mathcal{B}(R)$, the Borel σ-algebra of R. We further assume the existence of a transition probability density for the Markov chain $\{(X_n, \xi_n), n \geq 0\}$ with respect to a σ-finite measure μ on R such that

$$P_\theta\{X_1 \in A, \xi_1 \in B | X_0 = x, \xi_0 = s_0\} = \sum_{y \in A} \int_{s \in B} p_{xy}(\theta) f(s; \varphi_y(\theta) | s_0) d\mu(s), \tag{1.4}$$

where $f(\xi_k; \varphi_{X_k}(\theta)|\xi_{k-1})$ is the conditional density of ξ_k, given ξ_{k-1} and X_k, with respect to μ, $\theta \in \Theta$ is the unknown parameter, and $\varphi_y(\cdot)$ is a function defined on the parameter space Θ for each $y = 1, \ldots, d$. Here and in the sequel, we assume that the Markov chain $\{(X_n, \xi_n), n \geq 0\}$ is stationary with probability density $\pi_x(\theta) f(\cdot; \varphi_x(\theta))$ with respect to μ. Note that in (1.3), we assume that the distribution of the Markov chain ξ_n depends on ξ_{n-1} and X_n. It can also depend on $\xi_{n-p}, \ldots, \xi_{n-1}$ and $X_{n-p}, \ldots, X_{n-1}, X_n$ without causing any difficulty. The usual parameterization for $\theta \in \Theta$ is $\theta = (p_{11}, \ldots, p_{dd}, \theta_1, \ldots, \theta_d)$ with $p_{xy}(\theta) = p_{xy}$ and $\varphi_y(\theta) = \theta_y$. Here we consider $\theta = (\theta_1, \ldots, \theta_q) \in \Theta \subseteq R^q$ as the unknown parameter, the true parameter value is denoted by θ^0. For convenience of notation, we use π_x for $\pi_x(\theta)$ and p_{xy} for $p_{xy}(\theta)$, respectively.

Definition 1. A process $\{\xi_n, n \geq 0\}$ is called a hidden Markov model if there is a Markov chain $\{X_n, n \geq 0\}$ such that the process $\{(X_n, \xi_n), n \geq 0\}$ satisfies (1.3) and (1.4).

For given observations $\xi_0, \ldots, \xi_n$ from a hidden Markov chain $\{\xi_n, n \geq 0\}$, the likelihood function is

$$g_n(\xi_0, \ldots, \xi_n; \theta) = \sum_{x_0=1}^{d} \cdots \sum_{x_n=1}^{d} \pi_{x_0} f(\xi_0; \varphi_{x_0}(\theta)) \prod_{j=1}^{n} p_{x_{j-1}x_j} f(\xi_j; \varphi_{x_j}(\theta)|\xi_{j-1}). \tag{1.5}$$

For a given column vector $x = (x_1, \ldots, x_d)^t \in R^d$, the L_1-norm of x is $\|x\| = \sum_{i=1}^{d} |x_i|$. The likelihood function (1.5) can be represented as

$$g_n(\xi_0, \ldots, \xi_n; \theta) = \|M_n \cdots M_1 M_0 \pi\|, \tag{1.6}$$

where, for $k = 1, \ldots, n$,

$$M_0 = M_0(\theta) = \begin{bmatrix} f(\xi_0; \varphi_1(\theta)) & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & f(\xi_0; \varphi_d(\theta)) \end{bmatrix}, \tag{1.7}$$

$$M_k = M_k(\theta) = \begin{bmatrix} p_{11} f(\xi_k; \varphi_1(\theta)|\xi_{k-1}) & \cdots & p_{d1} f(\xi_k; \varphi_1(\theta)|\xi_{k-1}) \\ \vdots & \ddots & \vdots \\ p_{1d} f(\xi_k; \varphi_d(\theta)|\xi_{k-1}) & \cdots & p_{dd} f(\xi_k; \varphi_d(\theta)|\xi_{k-1}) \end{bmatrix}, \tag{1.8}$$

$$\pi = \pi(\theta) = \Big(\pi_1, \ldots, \pi_d\Big)^t. \tag{1.9}$$

2. Large Deviations for Products of Random Matrices

In the limit theory for products of Markov random matrices, a large deviation theorem can be found in Bougerol (1988) when the underlying Markov chain satisfies an *uniform ergodicity* condition. It is clear from the cocycle representation in V.1. of Bougerol and Lacroix (1985) that the limit theorems for products of random matrices are based on those for Markov chains. In fact, the proof of Bougerol's results is based on the perturbation theory for operators developed by Nagaev (1957) for Markov chains. Since Nagaev's representation theory and Bougerol's results work only for one dimensional deterministic functionals on uniformly ergodic Markov chains, we need more. Therefore, we give a brief summary of the extension of Nagaev's representation theory for Markov random walks satisfying the w-uniformly ergodicity condition (defined below), and provide propositions related to the large deviations theorem that can be used to prove Bahadur efficiency of the MLE in hidden Markov models.

Since $\{(X_n, \xi_n), n \geq 0\}$ considered in (1.3) and (1.4) is a Markov chain on a general state space $D \times R$, by abusing the notation a little, we let $\{X_n, n \geq 0\}$ be a Markov chain on a general state space D with σ-algebra $\mathcal{D}$, irreducible with respect to a maximal irreducibility measure on $(D, \mathcal{D})$ and aperiodic. The transition probability kernel will be denoted by $P(\cdot, \cdot)$. Let $w : D \to [1, \infty)$ be a measurable function, and let B be the Banach space of measurable functions $h : D \to \mathcal{C}$ (:= set of complex numbers) with $|h|_w := \sup_x |h(x)|/w(x) < \infty$. We assume the following conditions on the Markov chain: $\{X_n, n \geq 0\}$ has an invariant probability measure π such that $\int w(y) d\pi(y) < \infty$, and for every $h \in B$ we have

$$\lim_{n\to\infty} \sup_{x \in D} \left\{ \frac{|E(h(X_n)|X_0 = x) - \int h(y) d\pi(y)|}{w(x)} : x \in D, |h| \leq w \right\} = 0, \tag{2.1}$$

$$\sup_{x\in D}\left\{E[w(X_1)|X_0=x]/w(x)\right\}<\infty, \tag{2.2}$$

$$\sup_{x\in D}\left\{E[|h(X_1)|^2 w(X_1)|X_0=x]/w(x)\right\}<\infty. \tag{2.3}$$

Condition (2.1) says that the chain is w-uniformly ergodic, and this implies that there exist $\gamma>0$ and $0<\rho<1$ such that for all $h\in B$ and $n\geq 1$,

$$\sup_{x\in D}|E[h(X_n)|X_0=x]-\int h(y)d\pi(y)|/w(x)\leq \gamma\rho^n|h|_w. \tag{2.4}$$

(pp.382-383 and Proposition 16.1.3 of Meyn and Tweedie (1993).) When w is 1, this reduces to the classical uniformly ergodic condition.

Let $Gl(d,R)$ be the set of invertible $d\times d$ matrices with real entries, and let M be a function from $D\times D$ to $Gl(d,R)$. For $A\in Gl(d,R)$, define

$$M_0=A,\ M_1=M(X_0,X_1),\ldots,\ M_{n+1}=M(X_n,X_{n+1}),\ \text{and}\, T_n=M_n\cdots M_1M_0. \tag{2.5}$$

The system $\{(X_n,T_n),n\geq 0\}$ is said to consist of products of Markov random matrices on $D\times Gl(d,R)$ (cf. Bougerol (1988)). Let $\mathbf{P}_x$ denote the probability of $\{(X_n,T_n),n\geq 0\}$ with $X_0=x$ and $M_0=I$, the identity matrix, and let $\mathbf{E}_x$ denote the expectation under $\mathbf{P}_x$. We say that two non-zero vectors $u,v\in R^d$ have the same direction if for some $\lambda\in R$, $u=\lambda v$. This defines an equivalence relation Γ on $R^d-\{0\}$. The set of directions in R^d is the projection space $P(R^d)$ defined as the quotient space $R^d-\{0\}/\Gamma$. For $u\in R^d-\{0\}$, $\bar{u}$ denotes its direction, i.e., the class in $P(R^d)$.

For given $\{(X_n,T_n),n\geq 0\}$ as in (2.5), $M\in Gl(d,R)$ and $\bar{u}\in P(R^d)$, let $M\cdot\bar{u}=\overline{Mu}$ and define $W_0=(X_0,\bar{u}),W_1=(X_1,M_1\cdot\bar{u}),\ldots,W_n=(X_n,T_n\cdot\bar{u})$. Then, $W_0,\ldots,W_n$ is a Markov chain on the state space $D\times P(R^d)$, with transition kernel $\mathcal{P}((x,\bar{u}),A\times B):=\mathbf{E}_x(I_{A\times B}(X_1,M_1\cdot\bar{u}))$ for all $x\in D,\ \bar{u}\in P(R^d),\ A\in\mathcal{D}$, and $B\in\mathcal{B}(P(R^d))$, the Borel σ-algebra of $P(R^d)$. Under Condition K given below, it follows from a simple modification of Lemma 3.5 of Bougerol (1988) that the Markov chain $\{W_n,n\geq 0\}$ has an invariant probability measure m on $D\times P(R^d)$. Let $\mathcal{P}_{x,\bar{u}}$ denote the probability of $\{W_n,n\geq 0\}$ with $W_0=(x,\bar{u})$, and let $\mathcal{E}_{x,\bar{u}}$ denote the expectation under $\mathcal{P}_{x,\bar{u}}$.

Definition 2. (i) A subset Ω of $Gl(d,R)$ is said to be contracting if there exists a sequence $\{M_n,n\geq 0\}$ in Ω for which $||M_n||^{-1}M_n$ converges to a rank 1 matrix, where $\|M_n\|=\sup\{||Mu||;u\in R^d,\|u\|=1\}$. A product of Markov random matrices $\{(X_n,T_n),n\geq 0\}$ on $D\times Gl(d,R)$ is said to be contracting if $\pi\{x\in D;\Omega_x$ is contracting$\}=1$, where Ω_x is the smallest closed semigroup in $Gl(d,R)$ which contains the support of $\mathbf{P}_x((X_1,M_1)\in D\times\cdot)$, and π is the invariant measure of $\{X_n,n\geq 0\}$.

(ii) A product of Markov random matrices $\{(X_n, T_n), n \geq 0\}$ on $D \times Gl(d, R)$ is strongly irreducible if for all p with $1 \leq p < d$, there does not exist a family of p-dimensional linear subspaces of R^d, $V_1(x), \ldots, V_k(x)$ such that $V(x) = V_1(x) \cup \cdots \cup V_k(x)$ and $T_n V(X_0) = V(X_n)$, $\mathbf{P}_x$ a.s. for all $n = 1, 2, \ldots$.

Let $\chi(M) = \sup(\log \|M\|, \log \|M^{-1}\|)$. The following Condition K will be assumed throughout this section.

K1. The underlying Markov chain $\{X_n, n \geq 0\}$ satisfies conditions (2.1)−(2.3).
K2. There exist $a, B > 0$, such that $\mathbf{E}_x(exp\{a\chi(M_1)\}) \leq B$ for all $x \in D$.
K3. The system $\{(X_n, T_n), n \geq 0\}$ is strongly irreducible and contracting.

Definition 3. Given $a > 0$, for any continuous functions $\varphi : D \times P(R^d) \to C$, define $|\varphi|_w := \sup\{|\varphi(x, \bar{u})|/w(x) : x \in D, \bar{u} \in P(R^d)\}$, and $m_a(\varphi) := \sup\{|\varphi(x, \bar{u}) - \varphi(x, \bar{v})|/\delta(\bar{u}, \bar{v})^a; x \in D, \bar{u}, \bar{v} \in P(R^d)\}$, where $\delta(\bar{u}, \bar{v}) := |\sin\{\text{angle}(\bar{u}, \bar{v})\}|$. Define $H(a)$ as the set of Hölder continuous functions φ on $D \times P(R^d)$ for which $\|\varphi\|_a = |\varphi|_w + m_a(\varphi)$ is finite.

For a Hölder continuous function $\varphi \in H(a)$, let $x \in D$, $\bar{u} \in P(R^d)$, $\alpha \in C$, and $M_1 \in Gl(d, R)$, and define linear operators P_α, P on the space $H(a)$ as

$$P_\alpha \varphi(x, \bar{u}) = \mathbf{E}_x\{e^{\alpha \log \|M_1 u\|} \varphi(X_1, M_1 \cdot \bar{u})\}, \quad P\varphi(x, \bar{u}) = \mathbf{E}_x\{\varphi(X_1, M_1 \cdot \bar{u})\}. \tag{2.6}$$

By an argument similar to the spectral decomposition theorem for operators given by Bougerol (1988), and that given by Fuh (1999) for w-uniformly ergodic Markov chains, we have that P_α and P are bounded linear operators on the Banach space $H(a)$ with norm $\|\cdot\|_a$. Moreover, there exists a sufficiently small $\eta > 0$ such that for $|\alpha| \leq \eta$, and with ρ defined as in (2.4), the spectrum of P_α lies inside the two circles

$$C_1 = \{z : |z - 1| = (1 - \rho)/3\}, \quad C_2 = \{z : |z| = \rho + (1 - \rho)/3\}.$$

Hence, by the spectral decomposition theorem in Bougerol (1988), $H(a) = H_1(a) \oplus H_2(a)$, and there exists $0 < \delta \leq \eta$ such that for $|\alpha| \leq \delta$, $H_1(a)$ is one-dimensional and

$$P_\alpha \pi_\alpha h = \lambda(\alpha) \pi_\alpha h \ \text{ for } h \in H(a), \tag{2.7}$$

where $\lambda(\alpha)$ is the eigenvalue of P_α with a corresponding eigenspace $H_1(a)$ and π_α is the parallel projection of $H(a)$ onto the subspace $H_1(a)$ in the direction of $H_2(a)$. Let $h_1 \in H(a)$ be the constant function $h_1 \equiv 1$, and let $r((x, \bar{u}); \theta) = (\pi_\alpha h_1)(x, \bar{u})$. From (2.7), it follows that $r(\cdot; \alpha)$ is an eigenfunction of P_α associated with the eigenvalue $\lambda(\alpha)$; i.e., $r(\cdot; \alpha)$ generates the one-dimensional eigenspace $H_1(a)$. The following proposition generalizes Proposition 3.8 in Bougerol (1988). Since the proof is similar, it will not be repeated here.

Proposition 1. *Let $\{(X_n, T_n), n \geq 0\}$ be a product of Markov random matrices defined in* (2.5) *and assume it satisfies condition K. Then there exists $\delta > 0$ such that for $|\alpha| < \delta$, $P_\alpha = \lambda(\alpha)N_\alpha + Q_\alpha$, and*

(i) *$\lambda(\alpha)$ is the unique eigenvalue of the maximal modulus of P_α;*

(ii) *N_α is a rank-one projection such that $N_\alpha Q_\alpha = Q_\alpha N_\alpha = 0$;*

(iii) *the mappings $\lambda(\alpha), N_\alpha$ and Q_α are analytic, and $\log \lambda(\alpha)$ is strictly convex for $|\alpha| < \delta$;*

(iv) *$|\lambda(\alpha)| > (2+\rho)/3$ and for each $p \in N$, there exists $c > 0$ such that for each $n \in N$,*

$$\|\frac{d^p}{d\alpha^p} Q_\alpha^n\|_a \leq c(\frac{1+2\rho}{3})^n;$$

(v) *let $\gamma = \lim_{n\to\infty}(1/n)\mathbf{E}_x \log ||T_n||$ be the upper Lyapunov exponent. Then*

$$\gamma = \frac{\partial \lambda(\alpha)}{\partial \alpha}|_{\alpha=0} = \int \mathcal{E}_{x,\bar{u}}(\log \|M_1 u\|/||u||)dm(x,\bar{u}).$$

The following large deviation result will be used to prove the local optimality of the MLE in hidden Markov models. The proof is similar to that in Theorem 4.3 of Bougerol (1988) and will not be repeated here. In the following propositions, recall that q is the dimension of the Euclidean space in which the parameter space Θ resides, and consider $\alpha \in R$.

Proposition 2. *For each $j = 1, \ldots, q$, let $\{(X_n, T_n^{(j)}), n \geq 0\}$ be a product of Markov random matrices on $D \times Gl(d,R)$ satisfying condition K. Then there exist $A, B > 0$ such that, for any initial value x, unit vector u and $0 < \varepsilon < B$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}_{x,\bar{u}}\{\log \|T_n^{(j)} u\| - n\gamma_j > n\varepsilon\} = \varphi_j(\varepsilon), \tag{2.8}$$

where $\varphi_j(\varepsilon) = -\sup_{0<\alpha<A}\left(\alpha\varepsilon - \log \lambda_j(\alpha) + \alpha\gamma_j\right) < 0$, $\lambda_j(\alpha)$ is the eigenvalue defined in (2.7) *and γ_j is the upper Lyapunov exponent of $\{(X_n, T_n^{(j)}), n \geq 0\}$.*

Proposition 3. *For each $j = 1, \ldots, q$, let $\{(X_n, T_n^{(j)}), n \geq 0\}$ be a product of Markov random matrices on $D \times Gl(d,R)$ satisfying condition K. Assume the transition probability $P(\cdot,\cdot)$ for the Markov chain X_n has a density with respect to the Haar measure of $Gl(d,R)$. Denote $W_n^{(j)}$ as the induced Markov chain on $D \times Gl(d,R)$. Let $f : D \times Gl(d,R) \to R$ be an additive functional such that $\mathcal{E}_m \exp(af(W_1^{(j)})) < \infty$ for some $a > 0$, and for each $j = 1, \ldots, q$, where m is the stationary distribution on $D \times P(R^d)$. Denote $\mu_j = \mathcal{E}_m f(W_1^{(j)})$. Then there exist $A, B > 0$ such that, for any initial value x, unit vector u and $0 < \varepsilon < B$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathcal{P}_{x,\bar{u}}\{\sum_{k=0}^{n} f(W_k^{(j)}) - n\mu_j > n\varepsilon\} = \eta_j(\varepsilon), \tag{2.9}$$

where $\eta_j(\varepsilon) = -\sup_{0<\alpha<A}(\alpha\varepsilon - \log\lambda_j(\alpha) + \alpha\mu_j) < 0$.

Proof. We first show that if $\{X_n, n \geq 0\}$ is a Markov chain as in the second paragraph of this section, then given X_n takes values on the whole real line R, the induced Markov chain $\{T_n \cdot \bar{u}\}$ satisfies the Doeblin's condition.

By means of the Iwasawa decomposition of $Gl(d,R)$ (cf. Lemma 6.1.1 of Bougerol and Lacroix (1985)), we have that any matrix M in $Gl(d,R)$ can be written as $M = s(M)k(M)$, where $k(M)$ is orthogonal and $S(M)$ is lower triangular with positive diagonal entries. Let S be the set of $s(M)$, and let K be the set of $k(M)$ for all $M \in Gl(d,R)$.

The existence of the transition probability density of the Markov chain $\{X_n, n \geq 0\}$ with respect to the Lebesgue measure implies that M_k has a density $p(u)$ with respect to the Haar measure m_G on $Gl(d,R)$, for each $k = 1,\dots,n$. Let m_S be the measure on S, and let m_K be the measure on K. Let μ' be the stationary measure of (X_k, M_k) on $R \times Gl(d,R)$. For any $\varepsilon > 0$, there is a measure $\tilde{\mu}$ on $R \times Gl(d,R)$, $d\tilde{\mu}(R \times M)/dm_G = \tilde{p}(M)$, such that $\tilde{p}(M) \leq c$, $\mathrm{var}(\mu', \tilde{\mu}) < \varepsilon/2$ and the support of $\tilde{\mu}(R \times \cdot)$ is contained in some compact set Γ of $Gl(d,R)$. Without loss of generality, we can assume that $K\Gamma K = \Gamma$.

It is well known (cf. p.407 in Helgason (1962)) that under suitable norming of m_G and m_S, $m_G(dM) = m_G(d(sk)) = m_S(ds)m_K(dk)$. Then, we have

$$\begin{aligned}\mathcal{P}\{(x,\bar{u}), R \times B\} &= \mu'\{(R,M) : M \cdot \bar{u} \in B\} \\ &= \int_B p(M \cdot \bar{u}) dm_G \leq \int_B \tilde{p}(M \cdot \bar{u}) dm_G + \varepsilon/2 \\ &= \int_B \int_{S\cap C} \tilde{p}(sk \cdot \bar{u}) dm_S dm_K + \varepsilon/2 \leq c m_S(S \cap C) m_K(B) + \varepsilon/2.\end{aligned}$$

Since Γ is compact, $m_S(S \cap C) < \infty$. This implies that the desired Doeblin's condition holds if X_n takes values on the whole real line R. Also by A1, $\{X_n, n \geq 0\}$ is a w-uniformly ergodic Markov chain. Combining these two properties, the Markov chain $\{W_n, n \geq 0\}$ is v-uniformly ergodic, with $v : D \times Gl(d,R) \to [1,\infty)$ and $v(x,\bar{u}) = w(x)$.

Denote $M_1^{(j)} = M^{(j)}(X_0, X_1)$ with $M^{(j)} \in Gl(d,R)$ for each $j = 1,\dots,q$. For all $x \in D$, $\bar{u} \in P(R^d)$ with $\|u\| = 1$, and $\alpha \in C^q$ (by a slight abuse of notation). For a bounded measurable function g, we define linear operators Q_α, Q on the Banach space B with norm $|\cdot|_v$ as

$$\begin{aligned}Q_\alpha g(x,\bar{u}) &= \mathcal{E}_x\{e^{(\alpha_1,\dots,\alpha_q)(g(X_1, M_1^{(1)}\cdot\bar{u}),\dots,g(X_1,M_1^{(q)}\cdot\bar{u}))^t}\}, \\ Qg(x,\bar{u}) &= \mathcal{E}_x\{g(X_1, M_1^{(1)} \cdot \bar{u}),\dots,g(X_1, M_1^{(q)} \cdot \bar{u})\}.\end{aligned} \tag{2.10}$$

Recall that $\mathcal{E}_x$ is the expectation defined on $D \times P(R^d)$ with initial distribution (x,π). Since $\{W_n, n \geq 0\}$ is v-uniformly ergodic for some $v : D \times Gl(d,R) \to$

$[1,\infty)$, by an argument similar to (2.6) through (2.7), we can let $\lambda_1(\alpha)$ be the eigenvalue of Q_α with corresponding one-dimensional eigenspace, such that Lemma 1 still holds for Q_α and $\lambda_1(\alpha)$. The rest of the proof follows by an argument similar to Theorem 4.3 of Bougerol (1988), and is omitted.

3. Main Results

Let $\{\xi_n, n \geq 0\}$ be the hidden Markov model defined in Section 1. Given a Markov chain $Z_n := (X_n, \xi_n)$ as defined in (1.3) and (1.4), let $D' := D \times R$ and M_k be the random matrix from $D' \times D'$ to $Gl(d,R)$, as defined in (1.7) and (1.8). Let $T_n = M_n \cdots M_1 M_0$; then, $\{(Z_n, T_n), n \geq 0\}$ is a product of Markov random matrices on $D' \times Gl(d,R)$ as defined in Section 2. Since M_0 is fixed, we simply let $\mathbf{P}_z := \mathbf{P}_z^\theta$ denote the probability of $\{(Z_n, T_n), n \geq 0\}$ with $Z_0 = z$, and let $\mathbf{E}_z := \mathbf{E}_z^\theta$ denote the expectation under $\mathbf{P}_z$.

For $\bar{u} \in P(R^d)$, $M \in Gl(d,R)$, and $\pi = \pi(\theta) = (\pi_1, \ldots, \pi_d)^t \in S^{d-1}$, the unit sphere with respect to the L_1-norm $\|\cdot\|$ in R^d, we have

$$\log \|T_n \pi\| = \log \frac{\|T_n \pi\|}{\|T_{n-1}\pi\|} + \cdots + \log \frac{\|T_0 \pi\|}{\|\pi\|}. \tag{3.1}$$

Define

$$W_0 = (Z_0, \overline{T_0\pi}),\ W_1 = (Z_1, \overline{T_1\pi}), \ldots, W_n = (Z_n, \overline{T_n\pi}). \tag{3.2}$$

Then, $W_0, \ldots, W_n$ is a Markov chain on the state space $E := D' \times P(R^d)$ with the transition kernel

$$\mathcal{P}((z,\bar{u}), A \times B) := \mathcal{P}^\theta((z,\bar{u}), A \times B) := \mathbf{E}_z(I_{A\times B}(Z_1, \overline{M_1 u})) \tag{3.3}$$

for all $z \in D'$, $\bar{u} \in P(R^d)$, $A \in \mathcal{D} \times \mathcal{B}(R)$, and $B \in \mathcal{B}(P(R^d))$, the Borel σ-algebra of $P(R^d)$. Note that the initial distribution of W_0 depends on Z_0 only, and Z_0 has the stationary distribution $\pi_x f(\xi; \varphi_x(\theta))$ as its initial distribution. We note that $\mathcal{P}_z := \mathcal{P}(\cdot,\cdot)$ in (3.3) depends only on z and let $\mathcal{E}_z := \mathcal{E}_{(z,\bar{u})}$ denote the expectation under $\mathcal{P}_z$. By (1.4), the Markov chain $\{(X_n, \xi_n), n \geq 0\}$ has transition density $p_{xy}(\theta) f(s; \varphi_y(\theta)|s_0)$ with respect to μ. Therefore, the induced transition probability $\mathcal{P}(\cdot,\cdot)$ has a probability density $p(\cdot,\cdot)$ with respect to μ. Under condition K in Section 2, it follows from Lemma 3.5 of Bougerol (1988) that the Markov chain W_n has an invariant probability measure $m := m_\theta$ on E. Note that m is a product measure on $D' \times P(R^d)$, and the first component has probability density $\pi_x f(\xi; \varphi_x(\theta))$ with respect to μ. Now, for $M \in Gl(d,R)$, let $\sigma : E \times E \to R$ be $\sigma((z_0, \bar{u}), (z_1, \overline{Mu})) = \log(\|Mu\|/\|u\|)$; then for $\pi = \pi(\theta)$ defined in (1.9),

$$\log \|T_n \pi\| = \sigma(W_{n-1}, W_n) + \cdots + \sigma(W_0, W_1) + \sigma(W_0, W_0) \tag{3.4}$$

is an additive functional of the Markov chain $\{W_n, n \geq 0\}$, where $\sigma(W_0, W_0) = \log(\|T_0\pi\|/\|\pi\|)$.

It is known that Kullback-Leibler divergence $K(\theta', \theta)$ is a major quantity for the development of Bahadur efficiency. When $\{\xi_n\}$ are independent and identically distributed random variables, it is easy to see that $K(\theta', \theta) = \int p_{\theta'}(\xi) \log(p_{\theta'}(\xi)/p_\theta(\xi)) d\xi$, the usual Kullback-Leibler information number. For the case of a hidden Markov chain $\{\xi_n, n \geq 0\}$, for given observations ξ_n from $\{\xi_n, n \geq 0\}$, Rabiner and Juang (1993) defined a generalized Kullback-Leibler divergence $K(\theta^0, \theta)$ as $H(\theta^0, \theta^0) - H(\theta^0, \theta)$ with $H(\theta^0, \theta) = \lim_{n\to\infty} n^{-1} E_{\theta^0}(\log g_n(\xi_0, \ldots, \xi_n; \theta))$. See also pp.134-136 in Leroux (1992). By the ergodic theorem for products of Markov random matrices, it is easy to see that $H(\theta^0, \theta)$ is just the upper Lyapunov exponent γ for $\{(Z_n, T_n), n \geq 0\}$ under the parameter θ^0. Therefore, the Kullback-Leibler information number can be defined as

$$K(\theta^0, \theta) := \int \mathcal{E}_z\Big(\log \frac{\|M_1(\theta^0) M_0(\theta^0) \pi(\theta^0)\|}{\|M_1(\theta) M_0(\theta) \pi(\theta)\|}\Big) dm_{\theta^0}(w), \tag{3.5}$$

where $\mathcal{E}_z$ is the expectation for the induced Markov chain W_n starting at z. Recall that $W_0 = w = (z, \overline{M_0\pi})$ from definition (3.2), and $M_0(\theta), M_1(\theta)$ and $\pi(\theta)$ are defined in (1.7)$-$(1.9).

Suppose that Θ is an open set in R^q. Based on the definition of Bahadur (1967) and Shen (2001), a parameter space $\bar{\Theta}$ is a suitable compactification of Θ if Θ is dense in $\bar{\Theta}$. A continuity condition (cf. condition A3 on Shen (2001)) for the likelihood function on the boundary of the parameter space is also assumed to hold. The following conditions will be used throughout the rest of this paper.

C1. The true parameter θ^0 is an interior point of Θ, and the transition probability matrix $[p_{xy}(\theta)]$ is ergodic (irreducible, aperiodic and positive recurrent) for all $\theta \in \Theta$. Moreover, the Markov chain $\{(X_n, \xi_n), n \geq 0\}$ has an invariant measure and satisfies conditions (2.1)$-$(2.3), and $M_0(\theta), M_1(\theta)$ defined in (1.7) and (1.8) are invertible $\mathbf{P}^\theta$ almost surely, for all $\theta \in \Theta$.

C2. θ is identifiable, i.e., for each $\theta, \theta' \in \Theta$, $\theta \neq \theta'$, and all $n = 1, 2, \ldots$, $g_n(\xi_0, \ldots, \xi_n; \theta)$ and $g_n(\xi_0, \ldots, \xi_n; \theta')$ are not equal μ-$a.s.$

C3. For each $\theta \in \Theta$, there exists $u = u(\theta) > 0$ such that for all $x \in D$,

$$\sup_{(x,\xi_0)\in D\times R} E_\theta\Big(\Big[\sup_{\eta\in N_h(\theta)} \frac{g_1(\xi_0, \xi_1; \eta)}{g_1(\xi_0, \xi_1; \theta)} \Big]^u |Z_0 = (x, \xi_0)\Big) < \infty,$$

where $N_h(\theta)$ is a h-neighborhood of θ and $E_\theta(\cdot|Z_0 = (x, \xi_0))$ denotes the expectation defined under P_θ in (1.3) with initial state $(x, \xi_0) \in D \times R$.

C4. For all $x \in D$, $\xi_0, \xi_1 \in R$, $\theta \in \Theta$, and for $i, j, k = 1, \ldots, q$, the partial derivatives $\partial l / \partial \theta_i$, $\partial^2 l / \partial \theta_i \partial \theta_j$, $\partial^3 l / \partial \theta_i \partial \theta_j \partial \theta_k$ exist and are continuous in θ, where $l(\xi_0, \xi_1; \theta) = \log g(\xi_0, \xi_1; \theta)$, and

$$\mathcal{E}_{m_\theta}(\frac{\partial \log g_1(\xi_0, \xi_1; \theta)}{\partial \theta_i}) = \int \mathcal{E}_z \Big(\frac{\partial \log \|M_1(\theta) M_0(\theta) \pi(\theta)\|}{\partial \theta_i} \Big) dm_\theta(w) = 0,$$

where $\mathcal{E}_{m_\theta}$ is defined as the expectation under $\mathcal{P}_{m_\theta}$ in (3.3), with the initial measure as the stationary probability measure m_θ. For all $x, y \in D$, $\theta \to p_{xy}(\theta)$ and $\theta \to \pi_x(\theta)$ have two continuous derivatives for $\theta \in \Theta$.

C5. For all $\theta \in \Theta$, the Fisher information matrix

$$\begin{aligned} \mathbf{I}(\theta) = (I_{ij}(\theta)) &= \Big(\mathcal{E}_{m_\theta} \Big[\Big(\frac{\partial \log g_1(\xi_0, \xi_1; \theta)}{\partial \theta_i} \Big) \Big(\frac{\partial \log g_1(\xi_0, \xi_1; \theta)}{\partial \theta_j} \Big) \Big] \Big) \qquad (3.6) \\ &= \Big(\int \mathcal{E}_z \Big[\Big(\frac{\partial \log \|M_1(\theta) M_0(\theta) \pi(\theta)\|}{\partial \theta_i} \Big) \Big(\frac{\partial \log \|M_1(\theta) M_0(\theta) \pi(\theta)\|}{\partial \theta_j} \Big) \Big] dm_\theta(w) \Big) \end{aligned}$$

is positive definite in a neighborhood of θ^0.

Let $m_i(\theta, \xi, u) = \sup_\eta \{ | \frac{\partial l}{\partial \eta_i} | : d(\eta, \theta) < u \}$ and $m_{ij}(\theta, \xi, u) = \sup_\eta \{ | \frac{\partial^2 l}{\partial \eta_i \partial \eta_j} | : d(\eta, \theta) < u \}$.

C6. There exist $u = u(\theta^0) > 0$ and $t = t(\theta^0) > 0$ such that, for all $z = (x, \xi_0) \in D \times R$ and $\xi_1 \in R$, $\sup_{z \in D \times R} E_\theta \exp(t m_i(\theta^0, \xi, u) | Z_0 = z) < \infty$, $\sup_{z \in D \times R} E_\theta \exp(t m_{ij}(\theta^0, \xi, u) | Z_0 = z) < \infty$, for $i, j = 1, \ldots, q$. Furthermore, $N_h(\theta^0)$ is non-empty and there exists a measurable function $w(\cdot, \cdot | \theta^0)$ such that

$$\sup_{z \in D \times R} E_\theta (\sup_{\theta \in N_h(\theta^0)} | \frac{\partial^3 l}{\partial \theta_i \partial \theta_j \partial \theta_k} | | Z_0 = z) < w(\xi_0, \xi_1; \theta^0) \qquad (3.7)$$

for all $i, j, k = 1, \ldots, q$. The moment generating function of w exists when θ^0 obtains.

Discussion of Assumptions:

1. We consider hidden Markov chains with a finite state space; therefore, the requirement that $[p_{xy}(\theta)]$ be ergodic is equivalent to requiring that there exists $r > 0$, for all $x, y \in D$, $\theta \in \Theta$, such that $p^r_{xy}(\theta) \geq \gamma(\theta) > 0$, where $p^r_{xy}(\theta)$ denotes the rth step transition. For simplicity, we assume $r = 1$ throughout this paper. That $\{(X_n, \xi_n), n \geq 0\}$ satisfies (2.1)$-$(2.3), the w-uniform ergodicity condition, is quite general and covers several examples of Gaussian autoregression presented in Section 4. The reader is referred to Meyn and Tweedie (1993) for more details about w-uniformly ergodic Markov chains. The invertible condition is a technicality. C2 is the identifiability condition

for hidden Markov chains, which can be relaxed to the condition given by Leroux (1992). For the case where ξ_n is a deterministic function of X_n, the reader is referred to Ito, Amari and Kobayashi (1992) for necessary and sufficient conditions. C3 and C6 are Bahadur-type moment conditions for large deviations of the MLE. C4 amounts to standard smoothing conditions. To be more specific, for fixed $\theta' \in \Theta$, and for $i, j = 1, \ldots, q$, the partial derivatives $\partial K(\theta, \theta')/\partial \theta_i$, $\partial^2 K(\theta, \theta')/\partial \theta_i \partial \theta_j$ exist and are continuous in θ. The reader is referred to Shen (2001) for the case of general parameter spaces.

2. In order to define the Fisher information (3.6), we need to verify that the score function $\partial \log g_1(\xi_0, \xi_1; \theta)/\partial \theta$ is in $L_2(\mathcal{P}_{m_\theta})$ for $\theta \in N_h(\theta^0) :=$ a h-neighborhood of θ^0, since the definition is based on the law of large numbers for products of Markov random matrices developed from Proposition 2. That is, we need to verify that, for $\theta \in N_h(\theta^0)$,

$$\mathcal{E}_{m_\theta}\left(\frac{\partial \log g_1(\xi_0, \xi_1; \theta)}{\partial \theta}\right)^2 < \infty. \tag{3.8}$$

A sufficient condition for (3.8) to hold is that

$$\sup_{\bar{u} \in S^{d-1}} \sup_{(x, \xi_0) \in D'} \mathcal{E}_\theta\left(\left[\frac{\partial \log \sum_{x,y=1}^d \nu_x f(\xi_0; \varphi_x(\theta)) p_{xy} f(\xi_1; \varphi_y(\theta)|\xi_0)}{\partial \theta}\right]^2 \right.$$
$$\left. |W_0 = ((x, \xi_0), \bar{u})\right) < \infty,$$

for $\theta \in N_h(\theta^0)$. By definitions at (1.8) and (1.10), this reduces to

$$\sup_{(x,\xi_0) \in D'} E_\theta\left(\left[\frac{\partial \log \sum_{x,y=1}^d \nu_x f(\xi_0; \varphi_x(\theta)) p_{xy} f(\xi_1; \varphi_y(\theta)|\xi_0)}{\partial \theta}\right]^2 |Z_0 = (x, \xi_0)\right) < \infty,$$

for $\theta \in N_h(\theta^0)$. From condition C3, simple calculation shows that

$$\sup_{(x,\xi_0) \in D'} E_{\theta^0}\left(\left[\sup_{|\theta - \theta^0| < h} \max_{y, y' \in D} \frac{f(\xi_1; \varphi_y(\theta)|\xi_0)}{f(\xi_1; \varphi_{y'}(\theta)|\xi_0)}\right]^2 |Z_0 = (x, \xi_0)\right) < \infty \tag{3.9}$$

is a sufficient condition.

Remark 1. Let $p(\xi_1|\xi_0, \xi_{-1}, \ldots; \theta)$ be the conditional probability of ξ_1 given $\xi_0, \xi_{-1}, \ldots$, and $P_\theta(X_0 = \cdot|\xi_0, \xi_{-1}, \ldots; \theta)$ be the filtered probabilities given $\xi_0, \xi_{-1}, \ldots$. The Kullback-Leibler distance is $K(\theta^0, \theta) = H(\theta^0, \theta^0) - H(\theta^0, \theta)$ with $H(\theta^0, \theta)$ the relative entropy, intuitively defined as

$$\begin{aligned} H(\theta^0, \theta) &= E_{\theta^0}(\log p(\xi_1|\xi_0, \xi_{-1}, \ldots; \theta)) \\ &= E_{\theta^0}\left(\log \sum_{x=1}^d \sum_{y=1}^d f(\xi_1|\xi_0; \varphi_y(\theta)) p_{xy} P_\theta(X_0 = x|\xi_0, \xi_{-1}, \ldots; \theta)\right) \\ &= E_{\theta^0}(\log \|M_1(\theta) P_\theta(X_0 = x|\xi_0, \xi_{-1}, \ldots; \theta)\|). \end{aligned}$$

This shows that the definition of $H(\theta^0, \theta)$ is the expectation under the stationary distribution $P_\theta(X_0 = \cdot | \xi_0, \xi_{-1}, \ldots; \theta)$, that is, $\overline{T_n(\theta)\pi}$, when run under a process having parameter θ^0. A similar interpretation can be made for (3.5), where the Kullback-Leibler information number is defined under m_{θ^0}, the stationary distribution of the Markov chain $W_n = ((X_n, \xi_n), \overline{T_n \pi})$ defined in (3.2) and (3.3). Note that $\overline{T_n(\theta)\pi}$ itself does not form a Markov chain, but $\{((X_n, \xi_n), \overline{T_n \pi}), n \geq 0\}$ is a Markov chain. It is worth mentioning that only ξ_n appears in $\overline{T_n \pi}$, in which it reflects the nature of hidden Markov models. We also note that the Markov chain $\{W_n, n \geq 0\}$ with stationary distribution m_{θ^0} as its initial distribution involves the entire past $\{\xi_0, \xi_{-1}, \ldots\}$.

Remark 2. A similar interpretation can be made for the Fisher information $\mathbf{I}(\theta)$ in (3.6). Note that $\mathbf{I}(\theta)$ is defined as the expected value under the stationary distribution m_{θ^0} of the Markov chain $\{W_n, n \geq 0\}$. This is the key idea for the representation at (3.6), since the log likelihood function is an additive functional of the Markov chain $\{W_n, n \geq 0\}$, as shown in (1.6)$-$(1.9) and (3.4). See Lemma 2 in Section 5 for the approximation of $K(\theta', \theta)$ by $\mathbf{I}(\theta)$ as $\theta' \to \theta$.

For the estimation of $h(\theta)$, a real valued function of the unknown parameter θ, let $\Delta(\varepsilon, \theta) \equiv \{\theta' : \theta' \in \Theta,\ |h(\theta') - h(\theta)| > \varepsilon\}$. If $\Delta(\varepsilon, \theta)$ is not empty, define $b(\varepsilon, \theta) \equiv \inf\{K(\theta', \theta) : \theta' \in \Delta(\varepsilon, \theta)\}$, where $K(\theta', \theta)$ is defined in (3.5). If $\Delta(\varepsilon, \theta)$ is empty, define $b(\varepsilon, \theta) = \infty$.

For each $i, j = 1, \ldots, q$, let $[I^{ij}(\theta)] = [I_{ij}(\theta)]^{-1}$ which exists by condition C5. Let h be a real-valued smooth function from Θ to R, and let $I_h(\theta)$ be the Fisher information matrix for estimating $h(\theta)$. It is easy to see that

$$I_h(\theta) = \Big(\sum_{i,j=1}^{q} \frac{\partial h(\theta)}{\partial \theta_i} I^{ij}(\theta) \frac{\partial h(\theta)}{\partial \theta_j} \Big)^{-1}.$$

Theorem 1. *For a hidden Markov chain as at* (1.3) *and* (1.4), *let $\theta \in \Theta$ be the unknown parameter, let h be a real-valued smooth function from Θ to R, and let T_n be a consistent estimate of $h(\theta)$. Then if $b(\varepsilon, \theta) = (1/2) I_h(\theta) \varepsilon^2 + o(\varepsilon^2)$ as $\varepsilon \to 0$, we have*

$$\lim_{\varepsilon \to 0} \liminf_{n \to \infty} \frac{1}{\varepsilon^2 n} \log P_\theta(|T_n - h(\theta)| \geq \epsilon) \geq -\frac{I_h(\theta)}{2}. \tag{3.10}$$

Our main result shows that the maximum likelihood estimator of $h(\theta)$ can actually attain the lower bound in (3.10).

Theorem 2. *For a hidden Markov chain as at* (1.3) *and* (1.4) *satisfying* C1$-$C6, *let $\theta \in \Theta$ be the unknown parameter, let h be a real-valued smooth function from Θ to R, and let $\hat{\theta}_n$ be the MLE of θ. Then*

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{\varepsilon^2 n} \log P_\theta(|h(\hat{\theta}_n) - h(\theta)| \geq \varepsilon) = -\frac{I_h(\theta)}{2}. \tag{3.11}$$

Remark 3. If h is a real-valued measurable function from Θ to R and if $\hat{\theta}_n$ is the MLE of θ, then (3.11) can be replaced by

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{nb(\varepsilon,\theta)}\log P_\theta(|h(\hat{\theta}_n)-h(\theta)|\geq\varepsilon)=-1. \tag{3.12}$$

Hence $h(\hat{\theta}_n)$ is asymptotically locally optimal in the sense of Bahadur, Zabell and Gupta (1980) and Bahadur (1983).

4. Examples

Several types of hidden Markov models fit within the framework of (1.3) and (1.4). In this section, we apply Theorem 2 to these models, to include Markov chains, i.i.d. hidden Markov models, and Gaussian autoregression.

Example 1. Markov Chains

When ξ_n is equal to x_n in (1.3), one has a Markov chain. For the finite state space case, Bahadur (1983) investigated the local optimality for the maximum likelihood estimator $\hat{\theta}_n$. Consider the following characterization of the canonical measure $K(\hat{\theta}_n,\theta)$, where $K(\theta',\theta)$ is defined as $\sum_{x\in D}\pi_x(\theta')\sum_{y\in D}p_{xy}(\theta')\log(p_{xy}(\theta')/p_{xy}(\theta))$. Bahadur showed that if $\limsup_{\varepsilon\to 0}\ \limsup_{n\to\infty}\ (1/\varepsilon n)\log P_\theta(K(\hat{\theta}_n,\theta)\geq\varepsilon)\leq -1$ then, for each measurable function h from Θ to R, $h(\hat{\theta}_n)$ is locally optimal. Our conditions C1$-$C6 differ slightly from assumptions (i)$-$(iii) on p.279 of Bahadur (1983) and we obtain the same result in terms of $\mathbf{I}(\theta)$ for small ε, where $\mathbf{I}(\theta)=\sum_{x,y\in D}(\pi_x(\theta)/p_{xy}(\theta))(\partial p_{xy}(\theta)/\partial\theta)^2$ for $\theta\in\Theta\subseteq R$. Note that conditions C1$-$C6 require that the Markov chain be ergodic under the true parameter θ^0, and that the transition probabilities $p_{xy}(\theta)$ satisfy some smoothness and moment conditions. The sufficient condition (3.9) reduces to the standard Markov chain condition.

Example 2. i.i.d. Hidden Markov Chains

When the ξ_n at (1.3) are conditionally independent given $\mathbf{X}$, one has the so-called mixture model with a Markov regime. Let X_n be a two-state ergodic Markov chain and, conditional on X_n, let the ξ_n have normal densities with means and variances $\mu_1=2,\ \mu_2=0,\ \sigma_1^2=\sigma_2^2=1$. Suppose C2 holds. The requirement $p_{xy}(\theta)>0$ for all $x,y=1,2$, and for all $\theta\in\Theta$ is a sufficient condition of C1. By simple calculation, C3 reduces to $Ee^{2\xi}<\infty$. By using similar arguments, C4$-$C6 also hold in this case.

When the p_{xy} are known for $x,y=1,2$, $\mu_2=0$, $\sigma_1^2=\sigma_2^2=1$, and μ_1 is the only unknown parameter, (3.9) reduces to

$$\sup_{x\in D}E_{\theta^0}\left(\left[\sup_{|\theta-\theta^0|<h}\max_{y,y'\in D}\frac{f(\xi_1;\varphi_y(\theta))}{f(\xi_1;\varphi_{y'}(\theta))}\right]^2|X_0=x\right)<\infty, \tag{4.1}$$

which holds for the normal mixture. Then the Fisher information $\mathbf{I}(\mu_1)$ is equal to $\mathcal{E}_{m_{\mu_1}}[\sum_{x=1}^{2} \pi_x p_{x1}(\xi_1 - \mu_1)\varphi(\xi_1 - \mu_1)/g_1(\xi_1; \mu_1)]^2$, where $\varphi(\cdot)$ is the standard normal density and $g_1(\xi_1; \mu_1) = \sum_{x,y=1}^{2} \pi_x p_{xy} \varphi(\xi_1 - \mu_x)$. In general, for a finite ergodic Markov chain X_n, let $f(\cdot; \varphi_x(\theta))$ be a Lebesgue density on R^1 for each x, and assume that f is continuous and positive with $\lim_{\xi \to \pm\infty} f(\xi; \varphi_x(\theta)) = 0$, $\int_{-\infty}^{\infty} f^{\alpha}(\xi; \varphi_x(\theta)) d\xi < \infty$ for some $\alpha < 1$, and that $f(\cdot; \varphi_x(\theta))$ has at least three continuous derivatives and is bounded. Suppose the identifiability condition C2 holds. By an argument similar to the one above, conditions C1, C3−C6 hold, so Theorem 2 implies that, for each smooth function h from Θ to R, the MLE $h(\hat{\theta}_n)$ of $h(\theta)$ is locally optimal for each $\theta \in \Theta$.

Example 3. Gaussian Autoregression

We start with a simple scalar-valued fourth-order autoregression around one of two constants μ_1 or μ_2:

$$\begin{aligned} \xi_n - \mu_{x_n} = {} & \varphi_1(\xi_{n-1} - \mu_{x_{n-1}}) + \varphi_2(\xi_{n-2} - \mu_{x_{n-2}}) + \varphi_3(\xi_{n-3} - \mu_{x_{n-3}}) \\ & + \varphi_4(\xi_{n-4} - \mu_{x_{n-4}}) + \varepsilon_n, \end{aligned} \tag{4.2}$$

where $\varepsilon_n \sim N(0, \sigma^2)$ and $\theta = (\varphi_1, \ldots, \varphi_4, \mu_1, \mu_2, \sigma^2)$ is the unknown parameter. This model was studied by Hamilton (1989) in order to analyze the behavior of U.S. real GNP. The likelihood function for given $X_n = x_n$, $n \geq 0$, is

$$f(\xi_n | x_n; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- [(\xi_n - \mu_{x_n}) - \sum_{k=1}^{4} \varphi_k (\xi_{n-k} - \mu_{x_{n-k}})]^2 / 2\sigma^2 \Big). \tag{4.3}$$

Assume that all the roots of $1 - \sum_{k=1}^{4} \varphi_k z^k = 0$ are outside the unit circle, and that there exists a constant $c > 0$ such that $\sigma^2 > c$. Suppose the identifiability condition C2 holds. Suppose $p_{xy}(\theta) > 0$ for all $x, y \in D$, for all $\theta \in \Theta$, and that the condition C1 holds when $w(x) = x^2$. It is easy to see that conditions C3−C6 are satisfied in this model. Condition (3.9) reduces to

$$\sup_{(x,\xi_0) \in D'} E_{\theta^0} \Big(\Big[\sup_{|\theta - \theta^0| < h} \max_{y, y' \in D} \frac{f(\xi_1; \varphi_y(\theta) | \xi_0)}{f(\xi_1; \varphi_{y'}(\theta) | \xi_0)} \Big]^2 | Z_0 = (x, \xi_0) \Big) < \infty. \tag{4.4}$$

Since the maximum over x and y is applied to a finite set D, and f defined in (4.3) is a normal density, it is easy to check that (4.4) is satisfied in model (4.2). Although the random variables ξ_n depend on ξ_{n-1} and X_n only in Theorem 2, the result can be extended to depend on $\xi_{n-p}, \ldots, \xi_{n-1}$ and $X_{n-p}, \ldots, X_{n-1}, X_n$ without any difficulty. Therefore, the MLE in model (4.2) is locally optimal.

When $\xi_n = X_n$ in (4.2) and $\mu_1 = \mu_2 = \mu$, one has the classical autoregressive model with unknown parameters $\theta = (\varphi_1, \ldots, \varphi_4, \sigma^2)$. The Fisher information matrix is

$$\mathbf{I}(\theta) = \begin{pmatrix} \sigma^{-2}\Gamma & 0 \\ 0 & 2(\sigma^4)^{-1} \end{pmatrix},$$

where $\Gamma = (\gamma_{i-j})_{4\times 4}$ for $1 \le i, j \le 4$, with $\gamma_k = EX_nX_{n+k}$.

Engel and Hamilton (1990) considered switching autoregression model in which both the mean vector and the variance-covariance matrix were functions of the state: $\xi_n|x_n \sim N(\mu_{x_n}, \Omega_{x_n})$, for $x_n = 1, 2$, where $\theta = (\mu_1, \mu_2, \Omega_1, \Omega_2)$ are the unknown parameters. In this case, the likelihood function for given $X_n = x_n$, $n \ge 0$, is

$$f(\xi_n|x_n;\theta) = \frac{1}{2\pi|\Omega_{x_n}|^{1/2}} \exp\left(\frac{-(\xi_n - \mu_{x_n})'\Omega_{x_n}^{-1}(\xi_n - \mu_{x_n})}{2}\right),$$

where $|\Omega_x|$ denotes the determinant of Ω_x. Suppose C2 holds. Assume there exists a constant c such that $0 < c < |\Omega_x|$ for each $x = 1, 2$, and suppose that μ_1, μ_2 are in R; simple calculation shows that conditions C1, C3−C6 are satisfied.

In general, let $\xi_1, \ldots, \xi_n$ be a sample from the model

$$\xi_n = \sum_{k=1}^{p-1} a_{x_n}^k \xi_{n-k} + \sigma_{x_n}\varepsilon_n, \tag{4.5}$$

where ε_n is a normal random variable with zero mean and unit variance, and $a_x = (a_x^1, \ldots, a_x^{p-1}, \sigma_x)$ are the unknown parameters. In this case, the likelihood function is

$$f(\xi_n|a_{x_n}) = (2\pi\sigma_{x_n})^{-1/2} \cdot \exp\left\{-\frac{1}{2\sigma_{x_n}^2}\left(\xi_n - \sum_{k=1}^{p-1} a_{x_n}^k \xi_{n-k}\right)^2\right\}.$$

Assume that all the roots of $1 - \sum_{k=1}^{p} a_x^k z^k = 0$ are outside the unit circle, and that there exists a constant c with $0 < c < \sigma_x^2$ for $x = 1, \ldots, d$. Suppose the identifiability condition C2 holds. By a simple calculation, conditions C1 and C3−C6 hold so Theorem 2 implies that, for each smooth function h from Θ to R, the MLE $h(\hat{\theta}_n)$ of $h(\theta)$ is locally optimal for each $\theta \in \Theta$.

5. Proof of the Main Results

By (1.6), the analysis of likelihood estimation for hidden Markov models is reduced to that of products of Markov random matrices. In order to apply the large deviations result of Propositions 2 and 3 obtained in Section 2, it is necessary to check whether conditions C1−C6 imply conditions K1−K3. The proof of the following proposition is included here for completeness.

Proposition 4. *Consider a hidden Markov chain as at* (1.3) *and* (1.4) *and that satisfies* C1−C6*, and let $\theta \in \Theta$ be the unknown parameter. The induced product of Markov random matrices $\{(Z_n, T_n), n \ge 0\}$ satisfies conditions* K1−K3.

Proof. First, we note that C1 implies K1. Because each component $p_{xy}f(\xi_k; \varphi_y(\theta)|\xi_{k-1})$ in M_k has $X_{k-1} = x$ and $X_k = y$, and ξ_k is a Markov chain with transition probability density $f(\xi_k;\varphi_y(\theta)|\xi_{k-1})$, for each $k=1,\ldots,n$, M_k is a Markov random matrix with the underlying Markov chain $\{(X_n,\xi_n),n\geq 0\}$ having transition probability (1.3). It is easy to see that the moment conditions C3 and C6 imply condition K2.

Next we need to verify K3. Because the Markov chain $\{X_n,n\geq 0\}$ is ergodic (C1), there exists an $r>0$ such that $p^r_{xy}(\theta)>0$ for all $x,y\in D$ and for all $\theta\in\Theta$, where $p^r_{xy}(\theta)$ denotes the rth step transition probability. We have taken $r=1$ to have $p_{xy}(\theta)>0$ for all $x,y\in D$ and for all $\theta\in\Theta$. Conditioned on $\mathbf{X}$, $f(\xi_k;\varphi_y(\theta)|\xi_{k-1})$ is a transition probability density and hence is nonnegative for any $\theta\in\Theta$. Therefore, the strongly irreducible condition is satisfied.

It is known that for all $\theta\in\Theta$, the product of random matrices $\{(Z_n,T_n),n\geq 0\}$ on $D'\times Gl(d,R)$ is contracting if there exists a matrix M in the smallest closed semigroup in $Gl(d,R)$ which contains the support of $\mathbf{P}(D'\times\cdot)$, and such that M has a unique largest absolute eigenvalue. (This is an easy generalization of Corollary IV. 2.2 of Bougerol and Lacroix (1988).) Since the dimension of the matrix is finite and $f(\xi_1;\varphi_y(\theta)|\xi_0)$ is a conditional transition probability density, we have that for all $\theta\in\Theta$, there exists $\xi_1\in R$, with $f(\xi_1;\varphi_y(\theta)|\xi_0)>0$ for all $y\in D$. Let $Q_\theta=[p_{xy}(\theta)f(\xi_1;\varphi_y(\theta)|\xi_0)]$. Based on the assumption that the transition probability matrix $[p_{xy}(\theta)]$ of $\{X_n,n\geq 0\}$ is positive for all $\theta\in\Theta$, and according to the Perron-Frobenius theorem for a positive matrix, Q_θ has a unique largest eigenvalue. This implies that $\{(Z_n,T_n),n\geq 0\}$ on $D'\times Gl(d,R)$ is contracting for all $\theta\in\Theta$.

Proof of Theorem 1. This follows from Theorem 4.1 of Bahadur, Gupta and Zabell (1983) and the expansion of $b(\varepsilon,\theta)$ in terms of $I_h(\theta)$.

In the following, we consider hidden Markov models (1.3) and (1.4) that satisfy conditions C1−C6. Note that M (π) depends on θ, and we write M (π) or $M(\theta)$ $(\pi(\theta))$, respectively, for convenience. By using the definition of Kullback-Leibler information number at (3.5), a simple application of Jensen's inequality and condition C2 yields the following.

Lemma 1. *For any $\theta',\theta\in\Theta$, $K(\theta',\theta)\geq 0$ with equality if and only if $\theta'=\theta$.*

In the following proofs, for simplicity, we first consider a one-dimensional parameter $\theta\in\Theta\subseteq R$ and $h(x)=x$. The parallel development when θ is in a q-dimensional parameter space Θ with real valued function h on Θ is discussed at the end of this section.

Lemma 2. *Let $\theta,\theta'\in\Theta\subset R^q$ and let h be a smooth real-valued function such that $|h(\theta)-h(\theta')|<\varepsilon$. Then as $\varepsilon\to 0$, $b(\varepsilon,\theta)=(1/2)I_h(\theta)\varepsilon^2+o(\varepsilon^2)$, where $I_h(\theta)$ is defined in Theorem* 1.

Proof. Note that in the one dimensional case with $h(x) = x$, we have $I_h(\theta) = \mathbf{I}(\theta)$ and $b(\varepsilon, \theta) = K(\theta', \theta)$ with $\theta' = \theta + \varepsilon$.

Recall that $W_0, \ldots, W_n$ is a Markov chain on the state space $E := D' \times P(R^d)$ with the transition kernel $\mathcal{P}((z, \bar{u}), A \times B) := \mathcal{P}^\theta((z, \bar{u}), A \times B) := \mathbf{E}_z(I_{A\times B}(Z_1, \overline{M_1 u}))$ for all $z \in D',\ \bar{u} \in P(R^d),\ A \in \mathcal{D} \times \mathcal{B}(R)$ and $B \in \mathcal{B}(P(R^d))$. By (1.4), the Markov chain $\{Z_n = (X_n, \xi_n), n \geq 0\}$ has transition probability with density $p_{xy}(\theta) f(s; \varphi_y(\theta)|s_0)$ with respect to μ. Therefore, the induced transition probability $\mathcal{P}^\theta(\cdot, \cdot)$ of W_n has a probability density $p_\theta(\cdot, \cdot)$ with respect to μ and the invariant measure m_θ has a probability density with respect of μ. With an abuse of notation, we still denote the last by m_θ. (Note that we write $p_{xy}(\theta)$ $(\pi_x(\theta))$ as the transition probability (stationary distribution) of the Markov chain $\{X_n, n \geq 0\}$, and $p_\theta(\cdot, \cdot)$ $(m_\theta(\cdot))$ as the transition probability (stationary distribution) of the Markov chain $\{W_n, n \geq 0\}$.)

The Kullback-Leibler information number at (3.5) is defined in the framework of products of Markov random matrices; therefore, by letting $w_0 = (z_0, \overline{M_0 \pi})$, $w_1 = (z_1, \overline{T_1 \pi})$, we have

$$
\begin{aligned}
&K(\theta', \theta) \\
&= \int_{(z_0, \overline{M_0\pi})} \mathcal{E}_{z_0}\bigg(\log \frac{\|M_1(\theta') M_0(\theta') \pi(\theta')\|}{\|M_1(\theta) M_0(\theta) \pi(\theta)\|}\bigg) m_{\theta'}(z_0, \overline{M_0\pi}) d\mu(z_0, \overline{M_0\pi}) \\
&= \int_{w_0} \int_{w_1} \log\bigg(\frac{\|M_1(\theta') M_0(\theta') \pi(\theta')\|}{\|M_1(\theta) M_0(\theta) \pi(\theta)\|}\bigg) p_{\theta'}(w_0, w_1) m_{\theta'}(w_0) d\mu(w_1) d\mu(w_0), \quad (5.1)
\end{aligned}
$$

where $\mathcal{E}_z$ denotes the expectation under $\mathcal{P}$.

Let $F(\theta) := \|M_1(\theta) M_0(\theta) \pi(\theta)\|$, $F'(\theta) := \partial F(\theta)/\partial\theta$ and $F''(\theta) := \partial^2 F(\theta)/\partial\theta^2$, these exist by C4. Also by condition C4, the derivative $m'_\theta(\cdot)$ of $m_\theta(\cdot)$ with respect to θ, and the derivative $p'_\theta(\cdot, \cdot)$ of $p_\theta(\cdot, \cdot)$ with respect to θ exist. Using condition C4, we can write a derivative exists of the likelihood function, and second derivatives exist of $m_\theta(\cdot)$ and $p_\theta(\cdot, \cdot)$, Taylor expansion of (5.1) as

$$
\begin{aligned}
&K(\theta', \theta) \\
&= \int_{w_0} \int_{w_1} \log\Big[1 + \frac{F'(\theta)}{F(\theta)}\varepsilon + \frac{1}{2}\frac{F''(\theta)}{F(\theta)}\varepsilon^2 + o(\varepsilon^2)\Big] \\
&\quad \times \Big[p_\theta(w_0, w_1) + p'_\theta(w_0, w_1)\varepsilon + o(\varepsilon)\Big]\Big[m_\theta(w_0) + m'_\theta(w_0)\varepsilon + o(\varepsilon)\Big] d\mu(w_1) d\mu(w_0) \\
&= \int_{w_0} \int_{w_1} \Big[\log\Big(1 + \frac{F'(\theta)}{F(\theta)}\varepsilon + \frac{1}{2}\frac{F''(\theta)}{F(\theta)}\varepsilon^2\Big) + o(\varepsilon^2)\Big] \\
&\quad \times \Big[p_\theta(w_0, w_1) + p'_\theta(w_0, w_1)\varepsilon + o(\varepsilon)\Big]\Big[m_\theta(w_0) + m'_\theta(w_0)\varepsilon + o(\varepsilon)\Big] d\mu(w_1) d\mu(w_0) \\
&= \int_{w_0} \int_{w_1} \Big[\frac{F'(\theta)}{F(\theta)}\varepsilon + \frac{1}{2}\frac{F''(\theta)}{F(\theta)}\varepsilon^2 - \frac{1}{2}\Big(\frac{F'(\theta)}{F(\theta)}\Big)^2\varepsilon^2 + o(\varepsilon^2)\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\times\Big[p_\theta(w_0,w_1)+p'_\theta(w_0,w_1)\varepsilon+o(\varepsilon)\Big]\Big[m_\theta(w_0)+m'_\theta(w_0)\varepsilon+o(\varepsilon)\Big]d\mu(w_1)d\mu(w_0)\\
&=\int_{w_0}\int_{w_1}\Big[\varepsilon\frac{F'(\theta)}{F(\theta)}p_\theta(w_0,w_1)+\frac{\varepsilon^2}{2}\frac{F''(\theta)}{F(\theta)}p_\theta(w_0,w_1)-\frac{\varepsilon^2}{2}\Big(\frac{F'(\theta)}{F(\theta)}\Big)^2 p_\theta(w_0,dw_1)\\
&\quad+\varepsilon^2\frac{F'(\theta)}{F(\theta)}p'_\theta(w_0,w_1)\Big]\Big[m_\theta(w_0)+m'_\theta(w_0)\varepsilon\Big]d\mu(w_1)d\mu(w_0)+o(\varepsilon^2)\\
&=\ \varepsilon\int_{w_0}\int_{w_1}\frac{F'(\theta)}{F(\theta)}p_\theta(w_0,w_1)m_\theta(w_0)d\mu(w_1)d\mu(w_0)\\
&\quad+\frac{\varepsilon^2}{2}\int_{w_0}\int_{w_1}\frac{F''(\theta)}{F(\theta)}p_\theta(w_0,w_1)m_\theta(w_0)d\mu(w_1)d\mu(w_0)\\
&\quad-\frac{\varepsilon^2}{2}\int_{w_0}\int_{w_1}\Big(\frac{F'(\theta)}{F(\theta)}\Big)^2 p_\theta(w_0,w_1)m_\theta(w_0)d\mu(w_1)d\mu(w_0)\\
&\quad+\varepsilon^2\int_{w_0}\int_{w_1}\frac{F'(\theta)}{F(\theta)}\frac{m'_\theta(w_0)}{m_\theta(w_0)}p_\theta(w_0,w_1)m_\theta(w_0)d\mu(w_1)d\mu(w_0)\\
&\quad+\varepsilon^2\int_{w_0}\int_{w_1}\frac{F'(\theta)}{F(\theta)}p'_\theta(w_0,w_1)m_\theta(w_0)d\mu(w_1)d\mu(w_0)+o(\varepsilon^2)\\
&=J_1(\theta)\varepsilon+\frac{1}{2}J_2(\theta)\varepsilon^2-\frac{1}{2}\mathbf{I}(\theta)\varepsilon^2+J_3(\theta)\varepsilon^2+J_4(\theta)\varepsilon^2+o(\varepsilon^2).
\end{aligned}
$$

Note that the third equation above comes from a Taylor expansion of the log function.

To calculate each term in the above equation. By C4 and C6, we can interchange integration and differentiation to obtain

$$
\begin{aligned}
J_1(\theta)&=\mathcal{E}_m\Big[\frac{\partial\log\left(\|M_1(\theta)M_0(\theta)\pi(\theta)\|\right)}{\partial\theta}\Big]\\
&=\mathcal{E}_m\Big[\mathcal{E}_m\Big(\frac{\partial\log\left(\|M_1(\theta)M_0(\theta)\pi(\theta)\|\right)}{\partial\theta}|(\mathbf{X},\pi(\theta))\Big)\Big]=0.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
J_2(\theta)&=\mathcal{E}_m\Big[\frac{F''(\theta)}{F(\theta)}\Big]=\mathcal{E}_m\Big[\mathcal{E}_m\Big(\frac{F''(\theta)}{F(\theta)}|(\mathbf{X},\pi(\theta))\Big)\Big]=0,\\
J_3(\theta)&=\mathcal{E}_m\Big[\frac{\partial\log\left(\|M_1(\theta)M_0(\theta)\pi(\theta)\|\right)}{\partial\theta}\frac{\partial\log dm_\theta(w_0)}{\partial\theta}\Big]\\
&=\mathcal{E}_m\Big[\mathcal{E}_m\Big(\frac{\partial\log\left(\|M_1(\theta)M_0(\theta)\pi(\theta)\|\right)}{\partial\theta}|(\mathbf{X},\pi(\theta))\Big)\frac{\partial\log dm_\theta(w_0)}{\partial\theta}\Big]=0.
\end{aligned}
$$

A simple calculation shows that $J_4(\theta)=\mathbf{I}(\theta)\varepsilon^2$. Hence, $K(\theta',\theta)=(1/2)\mathbf{I}(\theta)\varepsilon^2+o(\varepsilon^2)$.

The following lemma proves that the MLE is consistent in the large deviation sense under conditions C1−C6. It also implies that the MLE $\hat{\theta}_n$ exists and is strongly consistent.

Lemma 3. *Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ^0. Then for any $\varepsilon > 0$, there exists ρ, $0 < \rho < 1$, such that for sufficiently large n,*

$$P_\theta(|\hat{\theta}_n - \theta^0| \geq \varepsilon) < \rho^n. \tag{5.2}$$

Proof. Let $\theta \in \Theta := (a, b)$ and $\theta \neq \theta^0$, set $z(\xi_0, \xi_1|\theta, \theta^0) := \log[g_1(\xi_0, \xi_1; \theta)/g_1(\xi_0, \xi_1; \theta^0)]$. By an abusing notation, let P_α be given by (2.6) with $\log \|M_1 u\|$ replaced by $z(\xi_0, \xi_1|\theta, \theta^0)$, and let $\lambda(\alpha)$ be correspondingly defined at (2.7). Note that $\lambda(\alpha)$ depends on θ and θ^0 here. $\{W_n, n \geq 0\}$ is v-uniformly ergodic as shown in the proof of Proposition 3. By C3, C6 and Theorem 4.1 of Ney and Nummelin (1987), $\lambda(\alpha)$ is well defined and Proposition 1 still holds for $0 < \alpha < 1$. By (2.6) and (2.7), we have $\log \lambda(0) = \log \lambda(1) = 0$. Now, since $\log \lambda(\alpha)$ is strictly convex in $[0, 1]$ by Proposition 1(iii),

$$\log \lambda(\alpha) < 0 \quad \text{for} \quad 0 < \alpha < 1. \tag{5.3}$$

Under the same boundary condition for the parameter space as that in (5.17) of Bahadur (1960) (or A3 in Shen (2001)), a similar argument as that of Lemma 5.2 in Bahadur (1960) shows that (5.3) holds for $\theta = a$ or b. So (5.3) holds for each $\theta \neq \theta^0$ and $\theta \in \bar{\Theta} := [a, b]$.

By C3, there exists $\alpha \in (0, 1)$ such that $E\{\exp(\alpha \sup_{\theta \in N_h(\theta^0)} z(\xi_0, \xi_1|\theta, \theta^0))\} < \infty$. Then for each $\theta \in \bar{\Theta}$ with $\theta \neq \theta^0$, there exists an interval $\mathcal{I}(\theta)$ containing θ such that

$$z^*(\xi_0, \xi_1|\theta) = \sup\{z(\xi_0, \xi_1|\theta_1, \theta^0) : \theta_1 \in \mathcal{I}(\theta)\}. \tag{5.4}$$

For each fixed θ, define $\lambda^*_\theta(\alpha)$ as before, but with $z^*(\xi_0, \xi_1|\theta)$. Then by using (5.3), Proposition 1 and Lebesgue's Dominated Convergence Theorem (in terms of θ) for $P_\alpha = \lambda(\alpha)N_\alpha + Q_\alpha$, we have $\lambda^*_\theta(\alpha) < 1$ for $0 < \alpha < 1$, and such that $\mathcal{I}(\theta)$ is open in $\bar{\Theta}$.

Now given $\varepsilon > 0$, let $S_\varepsilon = \{\theta \in \bar{\Theta}; |\theta - \theta^0| < \varepsilon\}$ with S^c_ε as the compliment of S_ε. We need only consider the case of non-empty S^c_ε. Since S^c_ε is compact in $\bar{\Theta}$, and $\theta_0 \notin S^c_\varepsilon$, there exists a finite number of points in S^c_ε, say $\theta_1, \ldots, \theta_k$ such that $S^c_\varepsilon \subset \mathcal{I}(\theta_1) \cup \mathcal{I}(\theta_2) \cup \cdots \cup \mathcal{I}(\theta_k)$, where $\mathcal{I}(\theta)$ is defined for each $\theta \neq \theta^0$ as in the preceding paragraph.

For fixed n and $\xi_0, \ldots, \xi_n$, suppose that $|\hat{\theta}_n - \theta^0| \geq \varepsilon$. Suppose first that there exists $\theta_1 \in \Theta$ for which $\log g_n(\xi_0, \ldots, \xi_n; \theta_1) = \sup_{\theta \in \Theta} \log g_n(\xi_0, \ldots, \xi_n; \theta)$. Then

$$\begin{aligned} &\sup_{\theta \in S^c_\varepsilon} \log g_n(\xi_0, \ldots, \xi_n; \theta) \geq \log g_n(\xi_0, \ldots, \xi_n; \hat{\theta}_n) \\ &= \sup_{\theta \in \Theta} \log g_n(\xi_0, \ldots, \xi_n; \theta) \geq \log g_n(\xi_0, \ldots, \xi_n; \theta^0). \end{aligned}$$

Thus

$$\sup_{\theta\in S_\varepsilon^c}\log g_n(\xi_0,\ldots,\xi_n;\theta)\geq\log g_n(\xi_0,\ldots,\xi_n;\theta^0). \tag{5.5}$$

Suppose next that $\log g_n(\xi_0,\ldots,\xi_n;\hat{\theta}_n)\neq\sup_{\theta\in\Theta}\log g_n(\xi_0,\ldots,\xi_n;\theta)$. Then $\sup_{\theta\in\Theta}\log g_n(\xi_0,\ldots,\xi_n;\theta)=\max\{\log g_n(\xi_0,\ldots,\xi_n;a),\ \log g_n(\xi_0,\ldots,\xi_n;b)\}>\log g_n(\xi_0,\ldots,\xi_n;\theta^0)$; hence (5.5) still holds since a and b are included in S_ε^c. Thus $|\hat{\theta}_n-\theta^0|\geq\varepsilon$ implies (5.5).

Since $S_\varepsilon^c\subset\mathcal{I}(\theta_1)\cup\mathcal{I}(\theta_2)\cup\cdots\cup\mathcal{I}(\theta_k)$, (5.5) implies

$$\max_{1\leq j\leq k}\left\{\sup_{\theta\in\mathcal{I}(\theta_j)}\log g_n(\xi_0,\ldots,\xi_n;\theta)\right\}\geq\log g_n(\xi_0,\ldots,\xi_n;\theta^0). \tag{5.6}$$

Now, for any θ, by (3.4) and (5.4), we have $\sup_{\theta\in\mathcal{I}(\theta)}\log g_n(\xi_0,\ldots,\xi_n;\theta)-\log g_n(\xi_0,\ldots,\xi_n;\theta^0)\leq\sum_{i=1}^n z^*(\xi_{i-1},\xi_i|\theta)$. Consequently, (5.6) implies that

$$\max_{1\leq j\leq k}\left\{\sum_{i=1}^n z^*(\xi_{i-1},\xi_i|\theta_j)\right\}\geq 0. \tag{5.7}$$

Let $A_n^{(j)}$ denote the event that $\sum_{i=1}^n z^*(\xi_{i-1},\xi_i|\theta_j)\geq 0$. Then (5.7) is equivalent to $\cup_{j=1}^k A_n^{(j)}$. Therefore, by (5.4)−(5.7), we have

$$\mathcal{P}\{|\hat{\theta}_n-\theta^0|\geq\varepsilon\}\leq\sum_{j=1}^k\mathcal{P}(A_n^{(j)}). \tag{5.8}$$

It follows from the definition of $A_n^{(j)}$, (5.4) and Proposition 3, that $\mathcal{P}(A_n^{(j)})\leq[\lambda_{\theta_j}^*(\alpha)]^n$. If $\rho_0=\max\{\lambda_{\theta_1}^*(\alpha),\ldots,\lambda_{\theta_k}^*(\alpha)\}$, $\rho_0<1$ and the right hand side of (5.8) does not exceed $k\rho_0^n$. Choose a ρ such that $\rho_0<\rho<1$. Then $k\rho_0^n<\rho^n$ for all sufficient large n, and we have the proof.

The following lemma relates the behavior of the maximum likelihood estimator $\hat{\theta}_n$ to that of the score function.

Lemma 4. *Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ^0. Then for any given δ with $0<\delta<\mathbf{I}(\theta^0)$, there exists $\rho<1$ such that for any given $\varepsilon>0$,*

$$P_{\theta^0}(|\hat{\theta}_n-\theta^0|\geq\varepsilon)\leq P_{\theta^0}(|\varsigma_n|\geq\varepsilon[\mathbf{I}(\theta^0)-\delta])+\rho^n \tag{5.9}$$

for all sufficiently large n, where

$$\varsigma_n=\frac{1}{n}\sum_{i=1}^{n-1}\frac{\partial}{\partial\theta}\Big(\sigma(W_{i-1},W_i)+\sigma(W_i,W_{i+1})\Big)\Big|_{\theta=\theta^0}. \tag{5.10}$$

Proof. Let

$$\begin{aligned}\eta_n &= \mathbf{I}(\theta^0) + \frac{1}{n}\sum_{i=1}^{n-1} \frac{\partial^2}{\partial\theta^2}\Big(\sigma(W_{i-1}, W_i) + \sigma(W_i, W_{i+1})\Big)\Big|_{\theta=\theta^0}, \\ \zeta_n &= \frac{1}{n}\sum_{i=1}^{n} w(\xi_{i-1}, \xi_i; \theta^0),\end{aligned} \tag{5.11}$$

where $w(\cdot,\cdot;\theta^0)$ is defined in (3.7). For fixed $h > 0$, let $N_h := \{|\theta - \theta^0| < h\}$ such that (3.7) holds for all $\theta \in N_h$, and $\delta/h > E[w(\xi_{i-1}, \xi_i; \theta^0)]$. By C6 which holds for each $\theta \in N_h$, and (3.6), $\mathbf{I}(\theta^0) = -\mathcal{E}_{m^0_\theta}[(\partial^2/\partial\theta^2)(\sigma(W_0, W_1) + \sigma(W_1, W_2))|\theta = \theta^0]$, and it follows from (5.11) and Proposition 3 that there exist $0 < \rho_i < 1$ for $i = 1, 2, 3$ and all n with

$$\mathcal{P}(\eta_n \geq \frac{\delta}{2}) \leq \rho_1^n, \quad \mathcal{P}(\eta_n \leq -\frac{\delta}{2}) \leq \rho_2^n, \quad \mathcal{P}(\zeta_n \geq \frac{\delta}{h}) \leq \rho_3^n. \tag{5.12}$$

For given n and $\{\xi_0, \ldots, \xi_n\}$, assume that

$$|\hat{\theta}_n - \theta^0| < h, \quad \text{and} \quad \log g_n(\xi_0, \ldots, \xi_n; \hat{\theta}_n) = \sup_{\theta\in\Theta} \log g_n(\xi_0, \ldots, \xi_n; \theta). \tag{5.13}$$

Since $\hat{\theta}_n \in N_h$ and N_h is open, it follows from (5.13) by Taylor's expansion that there exists a θ^* with $|\theta^* - \theta^0| < |\hat{\theta}_n - \theta^0| < h$ such that

$$\begin{aligned}0 &= \frac{\partial}{\partial\theta}\log g_n(\xi_0, \ldots, \xi_n; \hat{\theta}_n) \\ &= \frac{\partial}{\partial\theta}\log g_n(\xi_0, \ldots, \xi_n; \theta^0) + (\hat{\theta}_n - \theta^0)\frac{\partial^2}{\partial\theta^2}\log g_n(\xi_0, \ldots, \xi_n; \theta^0) \\ &\quad + \frac{1}{2}(\hat{\theta}_n - \theta^0)^2 \frac{\partial^3}{\partial\theta^3}\log g_n(\xi_0, \ldots, \xi_n; \theta^*).\end{aligned}$$

By (5.10) and (5.11), we have

$$(\hat{\theta}_n - \theta^0)[\mathbf{I}(\theta^0) + r_n] = \varsigma_n, \quad \text{with} \quad |r_n| \leq |\eta_n| + \frac{1}{2}h\zeta_n. \tag{5.14}$$

Let $A_n = \{|\hat{\theta}_n - \theta^0| \geq h\}$, $B_n = \{\log g_n(\xi_0, \ldots, \xi_n; \hat{\theta}_n) \neq \sup_{\theta\in\Theta} \log g_n(\xi_0, \ldots, \xi_n; \theta)\}$ and $C_n = \{|\eta_n| + (1/2)h\zeta_n \geq \delta\}$. As shown in proof of Lemma 3, each of the events A_n and B_n implies (5.5) and the probability of (5.5) is $\leq \rho_0^n$ for all n large enough by (5.6), (5.7) and (5.8), where $\rho_0 < 1$. C_n implies at least one of the three events whose probabilities are considered in (5.3). Therefore, there exists a $\rho < 1$ and a measurable event E_n such that $A_n \cup B_n \cup C_n$ implies E_n, and such that $P(E_n) \leq \rho^n$ for all sufficiently large n.

For given $\varepsilon > 0$, $P_\theta(|\hat{\theta}_n - \theta^0| \geq \varepsilon) \leq P_\theta(|\hat{\theta}_n - \theta^0| \geq \varepsilon,\ (\xi_0, \ldots, \xi_n) \notin E_n) + P_\theta(E_n)$. Hence (5.9) follows from (5.14) and the preceding paragraph.

Proof of Theorem 2. In the following, for simplicity, we assume the true parameter θ belongs to a one dimensional parameter space, and $\lambda''_\theta(0) = \sigma^2 = 1$. By Theorem 1 and Lemma 2, the lower bound $-b(\varepsilon, \theta)$ is approximated by $-\mathbf{I}(\theta)\varepsilon^2/2$ for small ε. Next, we want to prove that this is indeed an upper bound. We choose δ with $0 < \delta < \mathbf{I}(\theta)$ and write $a = \mathbf{I}(\theta) - \delta$. It follows from (5.10), Proposition 3 and C5 that, given $\rho < 1$, the first term on the right hand side of (5.9) tends to a limit $> \log \rho$ as $n \to \infty$, provided $\varepsilon > 0$ is sufficiently small. Lemma 4 then yields

$$\begin{aligned}&\limsup_{n\to\infty} n^{-1} \log P_\theta(|\hat{\theta}_n - \theta| \geq \epsilon)\\ &\leq \limsup_{n\to\infty} n^{-1} \log P_\theta\left(\frac{1}{n}\sum_{i=1}^{n-1} \frac{\partial}{\partial\theta}(\sigma(W_{i-1}, W_i) + \sigma(W_i, W_{i+1})) \geq \varepsilon a\right).\end{aligned}$$

By the large deviation result in Proposition 3 for products of Markov random matrices described in Section 2, there exists $A > 0$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log P_\theta(|\hat{\theta}_n - \theta| \geq \epsilon) \leq -\sup_{0<\alpha<A} (\alpha\varepsilon - \log \lambda_\theta(\alpha)), \tag{5.15}$$

where $\lambda_\theta(\alpha)$ is defined in (2.7). From C4, C6 and Proposition 1(iii), it is easy to see that $\lambda_\theta(\alpha)$ is analytic in $(\theta, \alpha) \in \Theta \times R$. Let $\theta' = \theta + \varepsilon$; by means of Taylor expansion of θ' around θ and α around zero, the right hand side of (5.15) is $-\sup_{0<\alpha<A}(-\log \lambda_{\theta'}(\alpha))$, up to $o(\varepsilon^2)$. We need to verify

(i) the approximation of the optimal point α_0 for small ε, employing the supremum on the right side of (5.15);

(ii) the approximation of $\lambda_{\theta'}(\alpha)$ at the optimal point α_0 for small ε, that is, we want to prove that, for θ' in an ε-neighborhood $N(\theta)$ of θ, $\lim_{\varepsilon\to 0}(1/\varepsilon^2) \log \lambda_{\theta'}(\alpha_0) = -(1/2)\mathbf{I}(\theta)$.

For the proof of (i), note that α_0 is defined as $\alpha_0\varepsilon - \log \lambda_\theta(\alpha_0) = \inf_\alpha(\alpha\varepsilon - \log \lambda_\theta(\alpha))$. That is, α_0 is the solution of $\lambda'_\theta(\alpha)/\lambda_\theta(\alpha) = \varepsilon$, where $\lambda'_\theta(\alpha)$ denotes the first derivative of $\lambda_\theta(\alpha)$ with respect to α. By C5, $\lambda'_\theta(0) = 0$, and the smoothness properties of $\lambda_\theta(\alpha)$ in Proposition 1(iii), we have as $\varepsilon \to 0$ that the optimal α_0 also $\to 0$. By Taylor expansion of α_0 around 0, $\lambda''_\theta(0)\alpha_0/\lambda_\theta(0) + o(\varepsilon) = \varepsilon$, which implies that $\alpha_0 = \varepsilon + o(\varepsilon)$.

(ii) As $\varepsilon \to 0$, by (i) and Taylor's expansion of $\lambda_{\theta'}(\alpha_0)$ for α_0 around 0, we have

$$\log \lambda_{\theta'}(\alpha_0) = k_1(\theta', \varepsilon)\alpha_0 + \frac{k_2(\theta', \varepsilon)}{2!}\alpha_0^2 + o(\alpha_0^2) = k_1(\theta', \varepsilon)\,\varepsilon + \frac{k_2(\theta', \varepsilon)}{2!}\varepsilon^2 + o(\varepsilon^2),$$

where $k_1(\theta', \varepsilon)$, and $k_2(\theta', \varepsilon)$ are the first two cumulants with

$$k_1(\theta', \varepsilon) = \mathcal{E}_m\Big(\frac{\partial \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta}|_{\theta=\theta'}\Big). \tag{5.16}$$

Since $\varepsilon \to 0$, by Taylor expansion of θ' around θ, by Lebesgue's Dominated Convergence Theorem via C6, and by integrating term by term, (5.16) becomes

$$\begin{aligned} &\mathcal{E}_m\Big(\frac{\partial \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta} + \frac{\partial^2 \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta^2}\varepsilon\Big) + o(\varepsilon^2) \\ &= 0 + \mathcal{E}_m\Big(\frac{\partial^2 \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta^2}\Big)\varepsilon + o(\varepsilon^2). \end{aligned}$$

In a similar way, we have

$$\begin{aligned} k_2(\theta',\varepsilon) &= Var_m\Big(\frac{\partial \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta}|_{\theta=\theta'}\Big) \\ &= \mathcal{E}_m\Big(\frac{\partial \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta}\Big)^2 + o(\varepsilon^2). \qquad (5.17) \end{aligned}$$

Through standard computation of additive functionals of the Markov chain W_n, we have $\mathcal{E}_m(\partial^2 \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)/\partial\theta^2) = -\mathbf{I}(\theta)$, and therefore

$$\begin{aligned} &\log\lambda_{\theta'}(\alpha_0) \\ &= \mathcal{E}_m\Big(\frac{\partial^2 \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta^2}\Big)\varepsilon^2 + \frac{1}{2}\mathcal{E}_m\Big(\frac{\partial \log(\|M_1(\theta)M_0(\theta)\pi(\theta)\|)}{\partial\theta}\Big)^2\varepsilon^2 \\ &\quad + o(\varepsilon^2) \\ &= -\frac{1}{2}\mathbf{I}(\theta)\varepsilon^2 + o(\varepsilon^2). \end{aligned}$$

Hence

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{\varepsilon^2 n}\log P_\theta\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \leq -\frac{\mathbf{I}(\theta)}{2}. \qquad (5.18)$$

To treat the general one-dimensional case, suppose $\hat{\theta}_n$ is the MLE of θ. Suppose that $h(\theta^0) \neq 0$. Choose $\lambda > 1$. It is easy to see that, for any sufficient small $\varepsilon > 0$, $|h(\theta) - h(\theta^0)| > \varepsilon$ implies $|\theta - \theta^0| > \delta$, where $\delta = \varepsilon/\lambda|h'(\theta^0)|$. Let $B_{n,\varepsilon} = \{(\xi_0, \ldots, \xi_n) : |h(\hat{\theta}_n) - h(\theta)| > \varepsilon\}$. Then $B_{n,\varepsilon}$ implies $\{|\hat{\theta}_n - \theta| \geq \varepsilon\}$, and hence $P_\theta(B_{n,\varepsilon}) \leq P_\theta(|\hat{\theta}_n - \theta| \geq \varepsilon)$. Since $\delta/\varepsilon = 1/\lambda|h'(\theta^0)|$, it follows from (5.18) that

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{\varepsilon^2 n}\log P_\theta\left(|h(\hat{\theta}_n) - h(\theta)| \geq \varepsilon\right) \leq -\frac{\mathbf{I}(\theta)}{2\lambda^2[h'(\theta)]^2}. \qquad (5.19)$$

Since $\lambda > 1$ is arbitrary, (5.19) implies

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{\varepsilon^2 n}\log P_\theta\left(|h(\hat{\theta}_n) - h(\theta)| \geq \varepsilon\right) \leq -\frac{\mathbf{I}(\theta)}{2[h'(\theta)]^2}. \qquad (5.20)$$

By using (5.20) and the lower bound (3.10) in Theorem 1, we have (3.11) for one dimensional case.

Now, suppose that Θ is an open set of the q dimensional Euclidean space of points, and let $\theta \in \Theta$ be the unknown parameter. To have (3.11), we need to show that for any smooth function h from Θ to R,

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{\varepsilon^2 n} \log P_\theta \left(|h(\hat{\theta}_n) - h(\theta)| \geq \varepsilon \right) \leq -\frac{I_h(\theta)}{2}. \tag{5.21}$$

The main difficulty of generalizing (5.20) to (5.21), as noted on pp.251-252 of Bahadur (1960), is in formulating a satisfactory boundary condition on the parameter space. The reader is referred to p.320 of Bahadur (1967), and p.484 of Shen (2001) for details.

Remark 4. Note that the trick used here is to analyze the right side of (5.15), which is a slightly different approach from those employed in Bahadur (1960) and Fu (1973). Bahadur's proof relies on the tail probability approximation of the normal distribution (see Lemma 2.4 for details); Fu's proof is based on the relationship between the MLE and the score function. In this paper, we have developed the relation via analytic properties of $\lambda_\theta(\alpha)$.

Remark 5. In light of (10) and (11) in Bahadur (1983), or (1.10) and (1.11) in Shen (2001), we have (5.21) if we can show that

$$\limsup_{\varepsilon \to 0} \limsup_{n \to \infty} \frac{1}{\varepsilon n} \log P_\theta (K(\hat{\theta}_n, \theta) \geq \varepsilon) \leq -1. \tag{5.22}$$

By using the decomposition of $K(\hat{\theta}_n, \theta)$ in terms of $\mathbf{I}(\hat{\theta}_n)$ in Lemma 2, and the large deviation result in Proposition 3, we have (5.22). Details are omitted.

Acknowledgements

The author is very grateful to an associate editor and the referee for their careful reading and constructive remarks on the manuscript. This research was partially supported by NSC 90-2118-M-001-011.

References

Bahadur, R. R. (1960). On the asymptotic efficiency of tests and estimates. *Sankhyā.* **22**, 229-252.

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Ann. Math. Statist.* **38**, 303-324.

Bahadur, R. R., Gupta, J. C. and Zabell, S. L. (1980). Large deviations tests and estimates. In *Asymptotic Theory of Statistical Tests and Estimation* (Edited by I. M. Chakravarti), 33-64. Academic Press, New York.

Bahadur, R. R. (1983). Large deviations of maximum likelihood estimates in the Markov chain case. In *Recent Advances in Statistics.* (Edited by M. H. Rizvi, J. S. Rustagi and D. Siegmund), 273-286. Academic, New York.

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

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

Bickel, P. and Ritov, Y. (1996). Inference in hidden Markov models I: local asymptotic normality in the stationary case. *Bernoulli* **2**, 199-228.

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

Bougerol, P. (1988). Théorèmes limite pour les sysèmes linéaires à coefficients markoviens. *Probab. Theory Related Fields* **78**, 193-221.

Bougerol, P. and Lacroix, J. (1985). *Product of Random Matrices with Application to Schrödinger Operators.* Birkhäuser, Boston.

Cogburn, R. (1980). Markov chains in random environments: the case of Markovian environments. *Ann. Probab.* **8**, 908-916.

Elliott, R., Aggoun, L. and Moore, J. (1995). *Hidden Markov Models: Estimation and Control.* Springer-Verlag, New York.

Engel, C. and Hamilton, J. D. (1990). Long swings in the dollar: are they in the data and do markets know it? *Amer. Econom. Rev.* **80**, 689-713.

Fu, J. C. (1973). On a theorem of Bahadur on the rate of convergence of point estimators. *Ann. Statist.* **1**, 741-749.

Fu, J. C. (1975). The rate of convergence of consistent estimators. *Ann. Statist.* **3**, 234-240.

Fu, J. C. (1982). Large sample point estimation: A large deviation theory approach. *Ann. Statist.* **10**, 762-771.

Fuh, C. D. (1998). Efficient likelihood estimation in hidden Markov models. Technical report, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, R.O.C.

Fuh, C. D. (1999). Paley-type inequalities related to the central limit theorem for Markov chains. *Sankhyā A* **61**, 81-100.

Fuh, C. D. (2003). SPRT and CUSUM in hidden Markov models. *Ann. Statist.* **31**, 942-977.

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

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357-384.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton University Press, Princeton, New Jersey.

Helgason, S. (1962). *Differential Geometry and Symmetric Spaces.* Academic Press, New York.

Ito, H., Amari, S. I. and Kobayashi, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Inform. Theory* **38**, 324-333.

Kester, A. D. (1985). *Some Large Deviations Results in Statistics.* CWI Trac 18. Mathematisch Centrum, Amsterdam.

Kingman, J. (1976). Subadditive processes. In *Ecole d'Eté de Probabilités de Saint-Flour* V-1976. Lectures Notes in Math. No. 539 (Edited by P. L. Hennequin), 167-223. Springer, Berlin.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). Hidden Markov models in computational biology: application to protein modeling. *J. Molecular Biology* **235**, 1501-1531.

Künsch, H. R. (2001). State space and hidden Markov models. In *Complex Stochastic Systems*, (Edited by Barndorff-Nielsen, Cox and Klüppelberg). Chapman and Hall/CRC.

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

Merhav, N. (1991). Universal classification for hidden Markov models. *IEEE Trans. Inform. Theory* **37**, 1586-1594.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer-Verlag, New York.

Nagaev, S. V. (1957). Some limit theorems for stationary Markov chains. *Theory Probab. Appl.* **2**, 378-406.

Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey.

Rukhin, A. L. (1983). Convergence rate of estimators of a finite parameter: how small can error probabilities be? *Ann. Statist.* **11**, 197-202.

Shen, X. L. (2001). On Bahadur efficiency and maximum likelihood estimation in general parameter spaces. *Statist. Sinica* **11**, 479-498.

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, R.O.C.

E-mail: stcheng@stat.sinica.edu.tw

(Received January 2002; accepted May 2003)