The Annals of Statistics
2003, Vol. 31, No. 3, 942–977

SPRT AND CUSUM IN HIDDEN MARKOV MODELS

BY CHENG-DER FUH[1]

Academia Sinica

In this paper, we study the problems of sequential probability ratio tests for parameterized hidden Markov models. We investigate in some detail the performance of the tests and derive corrected Brownian approximations for error probabilities and expected sample sizes. Asymptotic optimality of the sequential probability ratio test for testing simple hypotheses based on hidden Markov chain data is established. Next, we consider the cumulative sum (CUSUM) procedure for change point detection in this model. Based on the renewal property of the stopping rule, CUSUM can be regarded as a repeated one-sided sequential probability ratio test. Asymptotic optimality of the CUSUM procedure is proved in the sense of Lorden (1971). Motivated by the sequential analysis in hidden Markov models, Wald's likelihood ratio identity and Wald's equation for products of Markov random matrices are also given. We apply these results to several types of hidden Markov models: i.i.d. hidden Markov models, switch Gaussian regression and switch Gaussian autoregression, which are commonly used in digital communications, speech recognition, bioinformatics and economics.

1. Introduction. A hidden Markov model is a doubly stochastic process with an underlying stochastic process that is not directly observable (it is hidden) but can be observed only through another set of stochastic processes that produces the sequence of observations. The hidden Markov model has become important in a number of application areas, such as speech recognition [Rabiner and Juang (1993)], molecular biology [Krogh, Brown, Mian, Sjolander and Haussler (1994)], ion channel [Ball and Rice (1992)], economics [Hamilton (1989, 1994)] and digital communications over unknown channels [Elliott, Aggoun and Moore (1995)]. The main focuses of these efforts have been state space estimation, algorithms for fitting these models and the implementation of likelihood based methods. The statistical inference for hidden Markov models was first studied by Baum and Petrie (1966), and more recently by Leroux (1992), Bickel and Ritov (1996), Fuh (1998) and Bickel, Ritov and Rydén (1998).

The issue of hypothesis testing for hidden Markov models, whose statistics are not explicitly given, is of considerable importance in speech recognition applications [Merhav (1991) and Rabiner and Juang (1993)], in digital communications over unknown channels [Ziv (1985) and Csiszár and Narayan (1988)],

Received November 1999; revised February 2002.

[1]Supported in part by NSC 90-2118-M-001-033.

AMS 2000 subject classifications. Primary 60B15; secondary 60F05, 60K15.

Key words and phrases. Brownian approximation, change point detection, CUSUM, first passage time, products of random matrices, renewal theory, Wald's identity, Wald's equation.

in bioinformatics [Churchill (1989) and Liu, Neuwald and Lawrence (1999)], and in economics [Hamilton (1996)]. Sequence alignments and sequential algorithms for the on-line estimation are among the most commonly used methods, and some of these can be formulated as the binary hypothesis testing problem in the Neyman–Pearson setting. Merhav (1991) proposed an asymptotically optimal decision rule, and presented several types of hidden Markov models commonly used in speech recognition and communication applications. Another important subject related to this area is quick detection, with a low alarm rate, of parameter changes in state space models (hidden Markov models) on the basis of sequential observations from systems. This has numerous applications in statistical quality control, edge detection in images and the diagnosis of faults in the elements of computer communication networks. A comprehensive summary of this area was given by Basseville and Nikiforov (1993). See also Lai (1995) for a recent survey.

Motivated by the analysis of the cumulative sum (CUSUM) procedure for change point detection in hidden Markov models, we first study the fundamental issue of the performance of sequential probability ratio tests for parameterized hidden Markov models. The hidden Markov model considered here is in the general sense so as to cover the examples of switch Gaussian regression and switch Gaussian autoregression. It is well known [Wald and Wolfowitz (1948)] that Wald's sequential probability ratio test (SPRT) for testing simple hypotheses based on independent and identically distributed (i.i.d.) observations $\{\xi_n, n \geq 0\}$ is uniformly most efficient; that is, it simultaneously minimizes the expected sample sizes under both the null and the alternative hypotheses among all tests with the same or smaller error probabilities and with finite expected sample sizes under the two hypotheses. Their argument breaks down when the ξ_n are not i.i.d., and it has remained an open problem whether the SPRT has any optimality properties when $\{\xi_n, n \geq 0\}$ is a Markov chain or a hidden Markov model. Theorems 1 and 2 provide basic tools for an asymptotic solution of this long-standing problem. Before that, we analyze some basic properties of SPRT with simple hypotheses for hidden Markov chain data, and also for approximating error probabilities with composite statistical hypotheses, that is, hypotheses consisting of classes of hidden Markov chain distributions.

Next, we will consider the problem of change point detection via the CUSUM procedure for hidden Markov models, in the case of the distribution before and after change is given. As noted by Basseville and Nikiforov (1993) in their monograph, there is a great deal of literature on detection algorithms in complex systems but relatively little on the statistical properties and optimality theory of detection procedures beyond very simple models. The primary goal of this paper is to investigate the theoretical aspects of this procedure in hidden Markov models. Based on renewal property of the stopping rule, CUSUM can be regarded as a repeated one-sided sequential probability ratio test. Asymptotic optimality of the CUSUM procedure is proved in the sense of Lorden (1971). Motivated by

the sequential analysis, Wald's likelihood ratio identity and Wald's equation for products of Markov random matrices are also given.

A difficulty in analyzing likelihood based inferences for hidden Markov models is partly due to the nonadditive form [see (1.3)] of the log-likelihood function. A key idea to get rid of this difficulty is the representation of the likelihood function as the L_1-norm of products of Markov random matrices. This device has been proposed by Fuh (1998) to study efficient likelihood estimation for hidden Markov models. Here, we modify that method to represent the likelihood ratio as the ratio of the L_1-norm of products of Markov random matrices, and then to have the additive form of the log-likelihood ratio.

This paper is organized as follows. In the remainder of this section we give a formal definition of hidden Markov models and provide a representation of the likelihood ratio. In Section 2 we give the necessary definitions in products of Markov random matrices and state Wald's likelihood ratio identity and Wald's equation; their proofs will be deferred to the Appendix. In Section 3 we investigate in some detail the performance of the sequential probability ratio tests and derive corrected Brownian approximations for the error probabilities and expected sample sizes. In Section 4 we establish the asymptotic optimality of the sequential probability ratio test with simple hypotheses, based on hidden Markov chain data. In Section 5 we study the properties of the CUSUM procedure and provide an asymptotic lower bound under the average run length (ARL) constraint. This result is shown to imply the asymptotic optimality of the CUSUM scheme under the ARL constraint. In Section 6 we apply these results to several types of hidden Markov models: i.i.d. hidden Markov chains, switch Gaussian regression and switch Gaussian autoregression, which are commonly used in digital communications, speech recognition, bioinformatics and economics.

A hidden Markov model is defined as a parameterized Markov chain in a Markovian random environment [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 switch Gaussian regression and switch Gaussian autoregression studied by Merhav (1991), Rabiner and Juang (1993) and Hamilton (1994). That is, for each $\theta \in \Theta \subset R^q$, the unknown parameter, we consider $\mathbf{X} = \{X_n, n \geq 0\}$ as an ergodic (positive recurrent, irreducible and aperiodic) Markov chain on a finite state space $D = \{1, 2, \ldots, d\}$, with transition probability matrix $P(\theta) = [p_{xy}(\theta)]_{x,y=1,\ldots,d}$ and stationary distribution $\pi(\theta) = (\pi_x(\theta))_{x=1,\ldots,d}$. Suppose that an additive component $\Xi_n = \sum_{k=0}^{n} \xi_k$, 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 conditioning on the full $\mathbf{X}$ sequence, ξ_n is a Markov chain with probability

$$P^{(\theta)}\{\xi_{n+1} \in B | X_0, X_1, \ldots; \xi_0, \xi_1, \ldots, \xi_n\} = P^{(\theta)}(X_{n+1} : \xi_n, B) \qquad \text{a.s.} \tag{1.1}$$

for each n and $B \in \mathcal{B}(R)$, the Borel σ-algebra of R. Furthermore, we 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

$$\begin{aligned} &P^{(\theta)}\{X_1 \in A, \xi_1 \in B | X_0 = x, \xi_0 = s_0\} \\ &\quad = \sum_{y \in A} \int_B p_{xy}(\theta) f\big(s; \varphi_y(\theta) | s_0\big)\, d\mu(s), \end{aligned} \tag{1.2}$$

where $f(\xi_k; \varphi_{X_k}(\theta)|\xi_{k-1})$ is the transition probability density of ξ_k given ξ_{k-1}, 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 the Markov chain $\{(X_n, \xi_n), n \geq 0\}$ has stationary probability Γ with probability density $\pi_x(\theta) f(\cdot; \varphi_x(\theta))$ with respect to μ. Note that in (1.2), we assume that the distribution of the Markov chain ξ_n depends on ξ_{n-1} and X_n. It can be generalized to depend on $\xi_{n-p}, \ldots, \xi_{n-1}$ and $X_{n-p}, \ldots, X_{n-1}, X_n$ without 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 $\varphi_y(\theta) = \theta_y$. In this paper, we assume that only one parameter is of interest and treat the other parameters as nuisance parameters. That is, for simplicity, we consider $\theta \in \Theta \subseteq R$ as a one-dimensional unknown parameter. For convenience of notation, we will use π_x for $\pi_x(\theta)$ and p_{xy} for $p_{xy}(\theta)$, respectively, in the sequel. We give a formal definition of a hidden Markov model as follows:

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 (1.1) and (1.2).

Note that if ξ_n are conditionally independent given the full sequences $\mathbf{X}$, then the Markov chain $\{(X_n, \Xi_n), n \geq 0\}$ is called a *Markov random walk*, and $\{\xi_n, n \geq 0\}$ is the hidden Markov model studied by Leroux (1992), Bickel and Ritov (1996), Fuh (1998) and Bickel, Ritov and Rydén (1998).

Now, let $\xi_0, \xi_1, \ldots, \xi_n$ be the observations from the hidden Markov model $\{\xi_n, n \geq 0\}$ with an unknown parameter θ. Let

$$\begin{aligned} S_n &:= \frac{p_n(\xi_0, \xi_1, \ldots, \xi_n; \theta_1)}{p_n(\xi_0, \xi_1, \ldots, \xi_n; \theta_0)} \\ &:= \frac{\sum_{x_0=1}^d \cdots \sum_{x_n=1}^d \pi_{x_0}(\theta_1) f(\xi_0; \varphi_{x_0}(\theta_1))}{\sum_{x_0=1}^d \cdots \sum_{x_n=1}^d \pi_{x_0}(\theta_0) f(\xi_0; \varphi_{x_0}(\theta_0))} \\ &\quad \times \frac{\prod_{k=1}^n p_{x_{k-1}x_k}(\theta_1) f(\xi_k; \varphi_{x_k}(\theta_1)|\xi_{k-1})}{\prod_{k=1}^n p_{x_{k-1}x_k}(\theta_0) f(\xi_k; \varphi_{x_k}(\theta_0)|\xi_{k-1})} \end{aligned} \tag{1.3}$$

for fixed $\theta_0, \theta_1 \in \Theta$.

Let $\theta_0 \in \Theta^0$ (the interior of Θ) and consider the problem of testing hypothesis $\theta \leq \theta_0$. Given $\theta_1 > \theta_0$, we can construct a sequential probability ratio test of

$\theta = \theta_0$ versus $\theta = \theta_1$ and use it to test the composite hypothesis $\theta \le \theta_0$. Then, the sequential probability ratio test of $\theta = \theta_0$ versus $\theta = \theta_1$ stops sampling at stage

$$T := \inf\{n : \log S_n \le a \text{ or } \log S_n \ge b\} \tag{1.4}$$

for $a \le 0 < b$ and accepts the null hypothesis that $\theta = \theta_0$ (or the alternative hypothesis that $\theta = \theta_1$) according to $\log S_T \le a$ (or $\log S_T \ge b$). When it is regarded as a test of $\theta \le \theta_0$, the SPRT rejects $\theta \le \theta_0$ if and only if $\log S_T \ge b$. The problem of interest here is to approximate the type I error probability $\alpha = P^{(\theta_0)}\{\log S_T \ge b\}$, the type II error probability $\beta = P^{(\theta_1)}\{\log S_T \le a\}$ and the expected sample sizes $E^{(\theta_0)}T$ $(E^{(\theta_1)}T)$ of the test, where $P^{(\theta)}(E^{(\theta)})$ refers to the probability (expectation) with initial distribution as the stationary distribution $\pi_x(\theta) f(\cdot; \varphi_x(\theta))$.

Given a column vector $u = (u_1, \dots, u_d)^t \in R^d$, where t denotes the transpose of the underlying vector in R^d, define the L_1-norm of u as $\|u\| = \sum_{i=1}^d |u_i|$. The likelihood ratio (1.3) then can be represented as

$$S_n = \frac{p_n(\xi_1, \dots, \xi_n; \theta_1)}{p_n(\xi_1, \dots, \xi_n; \theta_0)} = \frac{\|M_n(\theta_1) \cdots M_1(\theta_1) M_0(\theta_1) \pi(\theta_1)\|}{\|M_n(\theta_0) \cdots M_1(\theta_0) M_0(\theta_0) \pi(\theta_0)\|}, \tag{1.5}$$

where

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

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

for $k = 1, \dots, n$, and

$$\pi(\theta) = \big(\pi_1(\theta), \dots, \pi_d(\theta)\big)^t. \tag{1.6}$$

Note that each component $p_{xy} f(\xi_k; \varphi_y(\theta)|\xi_{k-1})$ in M_k represents $X_{k-1} = x$ 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 $k = 1, \dots, n$, and therefore the M_k are random matrices. Since $\{(X_n, \xi_n), n \ge 0\}$ is a Markov chain by definition (1.1) and (1.2), this implies that $\{M_k, k = 1, \dots, n\}$ is a sequence of Markov random matrices (see Section 2 for a formal definition). Hence, S_n is the ratio of the L_1-norm of the products of Markov random matrices via representation (1.5). Note that π is fixed in (1.5).

In this paper, we assume that the parameter θ is identifiable in the sense that if, for some $\theta, \theta' \in \Theta$, $P_\theta^{(n)} = P_{\theta'}^{(n)}$ for all n, then $\theta = \theta'$. A necessary and sufficient condition was given by Itô, Amari and Kobayashi (1992) for deterministic functionals of Markov chains.

2. Wald's equation for products of Markov random matrices. Since $\{(X_n, \xi_n), n \geq 0\}$ considered in (1.1) is a Markov chain on a general state space $D \times R$, for simplicity of the notation, we let $\{X_n, n \geq 0\}$ be a Markov chain on a general state space D with σ-algebra $\mathcal{D}$, which is irreducible with respect to a maximal irreducibility measure on $(D, \mathcal{D})$ and is aperiodic. Let $P(\cdot, \cdot)$ denote the transition probability kernel and assume that X_n has stationary measure $\pi(\cdot)$. Define $Gl(d, R)$ as the set of invertible $d \times d$ matrices with real entries and let $\tilde{M}$, M be functions from $D \times D$ to $Gl(d, R)$ with

$$M_0 = \tilde{M}(X_0, X_0), \quad M_1 = M(X_0, X_1), \quad \ldots, \quad M_{n+1} = M(X_n, X_{n+1}).$$

Denote $\mathbb{S}_n = M_n \cdots M_0$. Then the system $\{(X_n, \mathbb{S}_n), n \geq 0\}$ is called a product of Markov random matrices on $D \times Gl(d, R)$ [cf. it is called a multiplicative Markov random walk in Bougerol (1988)]. Denote $\mathbb{P}_x$ as the probability of $\{(X_n, \mathbb{S}_n), n \geq 0\}$ with $X_0 = x$ and $M_0 = I_d$, the identity matrix, and $\mathbb{E}_x$ as the expectation under $\mathbb{P}_x$.

Let $u \in R^d$ be a d-dimensional vector, $\bar{u} := u/\|u\|$ the normalization of u ($\|u\| \neq 0$), and denote $P(R^d)$ as the projection space of R^d which contains all elements $\bar{u}$. For given $\bar{u} \in P(R^d)$ and $M \in Gl(d, R)$, denote $M \cdot \bar{u} = \overline{Mu}$ and let

$$W_0 = (X_0, \mathbb{S}_0 \cdot \bar{u}), \; W_1 = (X_1, \mathbb{S}_1 \cdot \bar{u}), \ldots, W_n = (X_n, \mathbb{S}_n \cdot \bar{u}). \tag{2.1}$$

Then, $W_0, W_1, \ldots, W_n$ is a Markov chain on the state space $D \times P(R^d)$, with the transition kernel $\mathbf{P}((x, \bar{u}), A \times B) := \mathbb{E}_x(I_{A \times B}(X_1, M_1 \cdot \bar{u}))$ 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 $P(R^d)$. For simplicity, we let $\mathbf{P}_{(x,\bar{u})} := \mathbf{P}(\cdot, \cdot)$ and denote $\mathbf{E}_{(x,\bar{u})}$ as the expectation under $\mathbf{P}_{(x,\bar{u})}$. Under Condition A given below, an argument similar to Lemma 3.5 of Bougerol (1988) results in the Markov chain $\{W_n, n \geq 0\}$ having an invariant probability measure m on $D \times P(R^d)$. Now, for $x, y \in D$, $\bar{u} \in P(R^d)$ and $M = M(x, y) \in Gl(d, R)$, let $\sigma : (D \times P(R^d)) \times (D \times P(R^d)) \to R$ be $\sigma((x, \bar{u}), (y, M \cdot \bar{u})) = \log \frac{\|Mu\|}{\|u\|}$. Then, for $\bar{u} \in P(R^d)$ and $\|u\| = 1$,

$$\begin{aligned} \log \|\mathbb{S}_n u\| &= \log \frac{\|\mathbb{S}_n u\|}{\|\mathbb{S}_{n-1} u\|} + \cdots + \log \frac{\|\mathbb{S}_1 u\|}{\|\mathbb{S}_0 u\|} + \log \frac{\|\mathbb{S}_0 u\|}{\|u\|} \\ &= \sigma(W_{n-1}, W_n) + \cdots + \sigma(W_0, W_1) + \sigma(W_0, W_0) \end{aligned} \tag{2.2}$$

is an additive functional of the Markov chain $\{W_n, n \geq 0\}$, where $\sigma(W_0, W_0) = \log \frac{\|\mathbb{S}_0 u\|}{\|u\|}$.

Previous work on limiting theory for products of Markov random matrices was done by Bougerol (1988) via perturbation theory. It is clear from representation (2.2) that limit theorems for products of random matrices are based on limit theorems for Markov chains. In fact, Bougerol's results are based on the perturbation theory for operators developed by Nagaev (1957) for Markov chains. Since Nagaev's representation theory and hence, Bougerol's results, require *uniform ergodicity* for the underlying Markov chain, it excludes many important

time series models and stochastic systems. For example, the autoregressive model $X_n = \alpha X_{n-1} + \varepsilon_n$ with $0 < |\alpha| < 1$ and i.i.d. standard normal ε_n does not satisfy such a uniformity condition. In this section we extend Nagaev's representation theory in two directions. First we replace the one-dimensional partial sum $\sum_{t=1}^{n} g(X_t)$ in his paper by the additive component, induced by products of Markov random matrices. The second extension is to relax the uniformity in his ergodicity condition, by imposing a w-uniform ergodicity condition (defined in A1 below).

We first define the necessary terminology and give a brief summary of the spectral decomposition theory for products of Markov random matrices. Definition 2 is taken from Bougerol (1988).

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 $\|M_n\| = \sup\{\|M_n u\|; u \in R^d, \|u\| = 1\}$. A product of Markov random matrices $\{(X_n, \mathbb{S}_n), n \geq 0\}$ on $D \times Gl(d, R)$ is said to be contracting if $\pi\{x \in D; \Omega_x \text{ is contracting}\} = 1$, where Ω_x is the smallest closed semigroup in $Gl(d, R)$ which contains the support of $\mathbb{P}_x((X_1, M_1) \in D \times \cdot)$.

(ii) A product of Markov random matrices $\{(X_n, S_n), n \geq 0\}$ 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

$$\mathbb{S}_n V(X_0) = V(X_n), \qquad \mathbb{P}_\pi\text{-a.s. for all } n = 1, 2, \ldots.$$

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

A1. $\{X_n, n \geq 0\}$ is w-uniformly ergodic, that is, there exists a measurable function $w : D \to [1, \infty)$, with $\int w(y)\, d\pi(y) < \infty$, such that, for any Borel measurable function h on D satisfying $\sup_x |h(x)|/w(x) < \infty$, we have

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

$$\sup_x \left\{ \frac{E_x(w(X_1))}{w(x)} \right\} < \infty.$$

A2. There exist $a, C > 0$, such that $\mathbf{E}_x(\exp\{a\chi(M_1)\}) \leq C$ for all $x \in D$.

A3. The system $\{(X_n, \mathbb{S}_n), n \geq 0\}$ is strongly irreducible and contracting.

REMARK. Under irreducibility and aperiodicity, Condition A1 implies that there exist $r > 0$ and $0 < \rho < 1$ such that for all h and $n \geq 1$,

$$\sup_{x \in D} \frac{|E(h(X_n)|X_0 = x) - \int h(y)\, d\pi(y)|}{w(x)} \leq r\rho^n \sup_{x \in D} \frac{h(x)}{w(x)};$$

see pages 382–383 of Meyn and Tweedie (1993). When w is 1, this reduces to the classical uniform ergodic condition. Note that the assumption of weight function w allows us to develop the perturbation theory for Markovian operators, and to have the twisting transformation (2.5) of the transition probability. The w-uniform ergodicity assumption also allows us to study the example of general hidden Markov models, including switch autoregression with Markov regime [Hamilton (1989, 1994)]. The exponential moment assumption A2 is suitable for the log likelihood function studied in this paper. A positivity hypothesis on the matrices in the support of the Markov chain leads to contraction and irreducibility properties A3. Note that the elements in random matrices (1.5) are probability densities and, hence, are nonnegative $\mathbb{P}_x$-almost surely.

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

Let ν be an initial distribution of W_0, and let $x \in D$, $\bar{u} \in P(R^d)$, $\theta \in \mathbf{C}$, $M_1 \in Gl(d, R)$ and φ be a Hölder continuous function in $H(\alpha)$. We define linear operators $\mathbf{T}(\theta)$, $\mathbf{T}$, $\nu(\theta)$ and $\mathbf{T}_0$ on the space $H(\alpha)$ as follows:

$$
\begin{aligned}
\mathbf{T}(\theta)\varphi(x, \bar{u}) &= \mathbf{E}_{(x,\bar{u})}\{e^{\theta \log \|M_1 u\|}\varphi(X_1, M_1 \cdot \bar{u})\}, \\
\mathbf{T}\varphi(x, \bar{u}) &= \mathbf{E}_{(x,\bar{u})}\{\varphi(X_1, M_1 \cdot \bar{u})\}, \\
\nu(\theta)\varphi(x, \bar{u}) &= \mathbf{E}_\nu\{e^{\theta \log \|M_0 u\|}\varphi(X_0, M_0 \cdot \bar{u})\}, \\
\mathbf{T}_0\varphi(x, \bar{u}) &= \mathbf{E}_m\{\varphi(X_0, M_0 \cdot \bar{u})\}.
\end{aligned} \tag{2.3}
$$

Recall that m is the invariant probability measure for the Markov chain $\{W_n, n \geq 0\}$ defined at (1.1) and (1.2). Condition A2 ensures that $\mathbf{T}(\theta)$, $\mathbf{T}$, $\nu(\theta)$ and $\mathbf{T}_0$ are bounded linear operators on the Banach space $H(\alpha)$ with the Hölder continuous norm $\|\cdot\|_\alpha$. By making use of A1 and an argument similar to Theorem 3.7 of Bougerol (1988), there exist constants $\gamma_* > 0$ and $0 < \rho_* < 1$, such that

$$
\|\mathbf{T}^n - \mathbf{T}_0\|_\alpha := \sup_{\varphi(x,\bar{u}) \in H(a),\ \|\varphi(x,\bar{u})\|_\alpha = 1} \|\mathbf{T}^n\varphi(x, \bar{u}) - \mathbf{T}_0\varphi(x, \bar{u})\|_\alpha < \gamma_* \rho_*^n.
$$

For a bounded linear operator $\mathbf{L} : H(\alpha) \to H(\alpha)$, the resolvent set is defined as $\{z \in C : (\mathbf{L} - zI)^{-1} \text{ exists}\}$ and $(\mathbf{L} - zI)^{-1}$ is called the resolvent (when the inverse exists). From the result of the geometric bound of $\|\mathbf{T}^n - \mathbf{T}_0\|_\alpha$, it follows that for $z \neq 1$ and $|z| > \rho_*$,

$$
R(z) := \mathbf{T}_0/(z - 1) + \sum_{n=0}^{\infty} (\mathbf{T}^n - \mathbf{T}_0)/z^{n+1}
$$

is well defined. Since $R(z)(\mathbf{T}-zI) = -I = (\mathbf{T}-zI)R(z)$, the resolvent of $\mathbf{T}$ is $-R(z)$. Moreover, by A2 and an argument similar to the proof of Lemma 2.2 of Jensen (1987), there exist $K > 0$ and $\eta > 0$ such that for $|\theta| \le \eta$, $|z-1| > (1-\rho_*)/6$ and $|z| > \rho_* + (1-\rho_*)/6$,

$$\|\mathbf{T}(\theta) - \mathbf{T}\|_\alpha \le K|\theta|,$$

$$R_\theta(z) := \sum_{n=0}^{\infty} R(z)\big\{\big(\mathbf{T}(\theta)-\mathbf{T}\big)R(z)\big\}^n \qquad \text{is well defined.}$$

Since $R_\theta(z)(\mathbf{T}(\theta)-\mathbf{T}) = R_\theta(z)\{(\mathbf{T}(\theta)-\mathbf{T}) + (\mathbf{T}-zI)\} = -I = (\mathbf{T}(\theta)-zI)R_\theta(z)$, the resolvent of $\mathbf{T}(\theta)$ is $-R_\theta(z)$. Moreover, there exists sufficiently small $\eta > 0$ such that for $|\alpha| \le \eta$, the spectrum of $\mathbf{T}(\theta)$ lies inside the two circles

$$C_1 = \{z : |z-1| = (1-\rho_*)/3\}, \qquad C_2 = \{z : |z| = \rho_* + (1-\rho_*)/3\}.$$

Hence, by the Riesz's spectral decomposition theorem [cf. page 421 of Riesz and Sz.-Nagy (1955)], $H(\alpha) = H_1(\alpha) \oplus H_2(\alpha)$, the direct sum of $H_1(\alpha)$ and $H_2(\alpha)$, and

$$\mathbf{Q}(\theta) := \frac{1}{2\pi i}\int_{C_1} R_\theta(z)\,dz, \qquad I - \mathbf{Q}(\theta) := \frac{1}{2\pi i}\int_{C_2} R_\theta(z)\,dz$$

are parallel projections of $H(\alpha)$ onto the subspaces $H_1(\alpha)$ and $H_2(\alpha)$, respectively. Moreover, by an argument similar to the proof of Lemma 1.1 of Nagaev (1957), there exists $0 < \delta \le \eta$ such that $H_1(\alpha)$ is one-dimensional for $|\alpha| \le \delta$ and

$$\sup_{|\theta|\le\delta} \|\mathbf{T}(\theta) - \mathbf{T}\|_\alpha < 1.$$

For $|\theta| \le \delta$, let $\lambda(\theta)$ be the eigenvalue of $\mathbf{T}(\theta)$ with corresponding eigenspace $H_1(\alpha)$. Letting ν denote the initial distribution of W_0, and defining the operator $\nu(\theta)$ by (2.3), we then have for $\varphi \in H(\alpha)$,

$$\begin{aligned}
&\mathbf{E}_\nu\big\{e^{\theta \log \|\mathbb{S}_n u\|}\varphi(X_n, \mathbb{S}_n\cdot\bar{u})\big\} \\
&\qquad = \nu(\theta)\mathbf{T}^n(\theta)\varphi = \nu(\theta)\mathbf{T}^n(\theta)\big\{\mathbf{Q}(\theta) + \big(I-\mathbf{Q}(\theta)\big)\big\}\varphi \\
&\qquad = \lambda^n(\theta)\nu(\theta)\mathbf{Q}(\theta)\varphi + \nu(\theta)\mathbf{T}^n(\theta)\big(I-\mathbf{Q}(\theta)\big)\varphi.
\end{aligned} \tag{2.4}$$

An argument similar to (1.22) of Nagaev (1957) shows that there exist $K^* > 0$ and $0 < \delta^* < \delta$ such that for $|\theta| \le \delta^*$,

$$\big\|\nu(\theta)\mathbf{T}^n(\theta)\big(I-\mathbf{Q}(\theta)\big)\varphi\big\|_\alpha \le K^*\|\varphi\|_\alpha|\theta|\{(1+2\rho_*)/3\}^n.$$

We next consider the summand $\lambda^n(\theta)\nu(\theta)\mathbf{Q}(\theta)\varphi$ which appeared in (2.4). Since A2 holds, then for any integer $r \ge 3$, analogous to Lemma 1.2 of Nagaev (1957), $\lambda(\theta)$ has the Taylor expansion

$$\lambda(\theta) = 1 + \sum_{k=1}^{r} i^k\lambda_k\theta^k/k! + \Delta(\theta)$$

in some neighborhood of the origin, where $\Delta(\theta) = O(|\theta|^r)$ as $|\theta| \to 0$.

Let $h_1 \in H(\alpha)$ be the identity function $h_1 := \mathbf{1}$, and consider the case that the initial distribution ν is degenerate at x, so that $\nu(\theta)\mathbf{Q}(\theta)h_1$ has continuous partial derivatives of order $r-2$ in some neighborhood of the origin. Let $r(\cdot;\theta) := (\mathbf{Q}(\theta)h_1)(\cdot)$. From (2.4) and $\mathbf{Q}(\theta)$ is a parallel projection of $H(\alpha)$ onto $H_1(\alpha)$, it follows that $r(\cdot;\theta)$ is an right eigenfunction of $\mathbf{T}(\theta)$ associated with the eigenvalue $\lambda(\theta)$; for example, $r(\cdot;\theta)$ generates the one-dimensional eigenspace $H_1(\alpha)$. This result is due to Nagaev (1957) in the special case of uniform ergodic Markov chains and $\xi_n = g(X_n)$, for which Jensen (1987) has given the full details of the argument that can clearly be extended to general ξ_n (not necessarily of the form $g(X_n)$). The following proposition generalizes Proposition 3.8 of Bougerol (1988).

PROPOSITION 1. *Let $\{(X_n, \mathbb{S}_n), n \geq 0\}$ be a sequence of products of Markov random matrices satisfying Condition* A. *Then, there exists $\delta > 0$ such that for $|\theta| < \delta$, $\nu(\theta)\mathbf{T}(\theta) = \nu(\theta)\lambda(\theta)\mathbf{Q}(\theta) + \nu(\theta)\mathbf{T}(\theta)(I - \mathbf{Q}(\theta))$, and*:

(i) *$\lambda(\theta)$ is the unique eigenvalue of maximal modulus of $\mathbf{T}(\theta)$;*

(ii) *$\mathbf{Q}(\theta)$ is a rank-one projection such that $\mathbf{Q}(\theta)(I - \mathbf{Q}(\theta)) = (I - \mathbf{Q}(\theta)) \times \mathbf{Q}(\theta) = 0$;*

(iii) *the mappings $\lambda(\theta), \mathbf{Q}(\theta)$ and $I - \mathbf{Q}(\theta)$ are analytic for $|\theta| < \delta$;*

(iv) *there exists $0 < \rho_* < 1$ such that $|\lambda(\theta)| > \rho_*$ and for each $p \in N$, there exists $c > 0$ such that for each $n \in N$,*

$$\left\| \frac{d^p}{d\theta^p} \big(I - \mathbf{Q}(\theta)\big)^n \right\|_\alpha \leq c\rho_*^n;$$

(v) *defining $\gamma := \lim_{n\to\infty}(1/n)\mathbf{E}_x \log \|\mathbb{S}_n\|$ as the upper Lyapunov exponent, it follows that $\gamma = \frac{\partial\lambda(\theta)}{\partial\theta}|_{\theta=0} = \int \mathbf{E}_{(x,\bar{u})}(\log \|M_1 u\|/\|u\|)\, dm(x,\bar{u})$ and with probability 1, $\lim_{n\to\infty}(1/n)\log\|\mathbb{S}_n u\| = \gamma$.*

Our main results in this section are stated in Lemma 1, Theorems 1 and 2. Their proofs will be deferred to the Appendix.

LEMMA 1. *Let $\{(X_n, \mathbb{S}_n), n \geq 0\}$ be a sequence of products of Markov random matrices satisfying Condition* A. *Let $\lambda(\theta)$ be defined as* (2.4), *and let $\Lambda(\theta)$ be (the principal branch of) $\log\lambda(\theta)$ so that $\lambda(\theta) = e^{\Lambda(\theta)}$. For any $\bar{u} \in P(R^d)$, let $G_n^{(\theta)} = r(W_n;\theta)\exp\{\theta \log\|\mathbb{S}_n u\| - n\Lambda(\theta)\}$ and let $\mathcal{F}_n$ be the σ-algebra generated by $\{(X_k, \mathbb{S}_k), k \leq n\}$. Then, for $\delta > 0$ small enough, $|\theta| \leq \delta$, $\{G_n^{(\theta)}, \mathcal{F}_n, n \geq 0\}$ is a martingale for any initial distribution ν on W_0.*

Now, for a given $\delta > 0$ which is small enough and $|\theta| \leq \delta$, we define the "twisting" transformation for the transition probability of $\{W_n, n \geq 0\}$ as

$$\mathbf{P}^{(\theta)}\big((x,\bar{u}), d(y,\bar{v})\big) = \frac{r((y,\bar{v});\theta)}{r((x,\bar{u});\theta)} e^{-\Lambda(\theta)+\theta\sigma((x,\bar{u}),(y,\bar{v}))}\mathbf{P}\big((x,\bar{u}), d(y,\bar{v})\big). \tag{2.5}$$

Let $\{W_n^{(\theta)}, n \geq 0\}$ be the Markov chain with transition probability kernel $\mathbf{P}^{(\theta)}$. Call this the θ-conjugate Markov chain. If the function $\Lambda(\theta)$ is normalized so that $\Lambda(0) = \Lambda'(0) = 0$, then $\mathbf{P}^{(0)} = \mathbf{P}$ is the transition probability kernel of the Markov chain $\{W_n, n \geq 0\}$, with invariant measure m. Let $\mathbf{P}_\nu^{(\theta)}$ denote the probability measure under which $\{W_n^{(\theta)}, n \geq 0\}$ has initial distribution ν, and let $\mathbf{E}_\nu^{(\theta)}$ denote the expectation under $\mathbf{P}_\nu^{(\theta)}$ here and in the sequel.

Taking $\mathcal{F}_\infty = \sigma(X_0, (X_1, \mathbb{S}_1), (X_2, \mathbb{S}_2), \ldots)$, we say that an integer-valued random variable $N \geq 0$ is a stopping time if, for each $n = 0, 1, 2, \ldots$, the conditional probability $P(N = n|\mathcal{F}_\infty)$ is $\mathcal{F}_n$-measurable. That is, there exists a Borel measurable function α_n such that $P(N = n|\mathcal{F}_\infty) = \alpha_n(X_0, (X_1, \mathbb{S}_1), (X_2, \mathbb{S}_2), \ldots, (X_n, \mathbb{S}_n))$. Denote

$$\mathcal{F}_N = \big\{A \in \mathcal{F}_\infty : A \cap \{N = n\} \in \mathcal{F}_n \text{ for all } n \geq 0\big\}.$$

The following theorems are Wald's likelihood ratio identity and Wald's equation for products of Markov random matrices.

THEOREM 1. *Let $\{(X_n, \mathbb{S}_n), n \geq 0\}$ be a sequence of products of Markov random matrices satisfying Condition* A. *Let N be any stopping time. Then, for any $x \in D$, $\bar{u} \in P(R^d)$, $B \in \mathcal{F}_N$ and $|\theta| \leq \delta$ for $\delta > 0$ small enough,*

$$\begin{aligned} &\mathbf{P}_{(x,\bar{u})}^{(\theta)}\big\{B \cap \{N < \infty\}\big\} \\ &\quad = \int_{B \cap \{N < \infty\}} \frac{r(W_N; \theta)}{r((x, \bar{u}); \theta)} \exp\big(\theta \log \|\mathbb{S}_N u\| - N\Lambda(\theta)\big)\, d\mathbf{P}_{(x,\bar{u})}. \end{aligned} \tag{2.6}$$

THEOREM 2. *Let $\{(X_n, \mathbb{S}_n), n \geq 0\}$ be a sequence of products of Markov random matrices satisfying Condition* A. *Let ν be an initial distribution of W_0, and let N be a stopping time such that $\mathbf{E}_\nu N < \infty$. Suppose $\sup_{(x,\bar{u})} \mathbf{E}_{(x,\bar{u})} |\log \|\mathbb{S}_1 u\|| < \infty$, and let γ be the upper Lyapunov exponent defined in Proposition* 1(v). *Then, for any $\bar{u} \in P(R^d)$,*

$$\mathbf{E}_\nu \log \|\mathbb{S}_N u\| = \gamma \mathbf{E}_\nu N - \mathbf{E}_\nu\{r'(W_N; 0) - r'(W_0; 0)\}, \tag{2.7}$$

where $r'(\cdot; \theta)$ denotes the derivative of $r(\cdot; \theta)$ with respect to θ. Furthermore, $r'((x, \bar{u}); 0)$ is the solution g of the following Poisson equation:

$$(I - \mathbf{T})g = \mathbf{E}_{(x,\bar{u})} \log \|\mathbb{S}_1 u\| - \gamma, \tag{2.8}$$

where I is the identity operator and $\mathbf{T}$ is the operator defined in (2.3).

REMARK. Wald's equation for uniformly ergodic Markov chains was employed by Fuh and Lai (1998). The uniform ergodicity assumption in that paper guarantees that $r'(\cdot; 0)$ defined in (2.7) is uniformly bounded. Theorem 2 generalizes that to products of Markov random matrices. Since the induced Markov chain $\{W_n, n \geq 0\}$ is not uniformly ergodic in general, the results obtained by Fuh and

Lai (1998) can not be applied. As an alternative, in Theorem 2, we characterize $r'(\cdot;0)$ as the solution of the Poisson equation (2.8) and then apply Foster's drift criterion for Markov chains to ensure the boundedness of $r'(\cdot;0)$. A weaker assumption, based on the existence of solutions for the Poisson equation, to ensure Wald's equation in Theorem 2 is valid can be found in Theorem 4 of Fuh and Zhang (2000).

3. Performance analysis of SPRT. To analyze the properties of the sequential probability ratio tests, the following Condition C will be assumed throughout this paper.

C1. For each $\theta \in \Theta$, the Markov chain $\mathbf{X} = \{X_n, n \geq 0\}$ is ergodic (positive recurrent, irreducible and aperiodic) on a finite state space $D = \{1, \ldots, d\}$. Moreover, the Markov chain $\{(X_n, \xi_n), n \geq 0\}$ satisfies Condition A1 and has stationary probability Γ with probability density $\pi_x(\theta) f(\cdot; \varphi_x(\theta))$ with respect to μ.

C2. For any $\theta \in \Theta$, the random matrices $M_0(\theta)$ and $M_1(\theta)$ defined in (1.5) are invertible $\mathbf{P}^{(\theta)}$ almost surely and

$$\sup_{(x,\xi_0)\in D\times R} E_x^{(\theta)} \left| \sum_{x,y=1}^{d} \pi_x(\theta) f\big(\xi_0; \varphi_x(\theta)\big) p_{xy}(\theta) \xi_1 f\big(\xi_1; \varphi_y(\theta)|\xi_0\big) \right| < \infty.$$

REMARK. The ergodicity Condition C1 for Markov chains is quite general and covers several interesting examples, such as i.i.d. hidden Markov models, switch Gaussian regression and switch Gaussian autoregression, considered in Section 6. Condition C2 is a moment condition for the likelihood function and a technical condition for the parameter space. For instance, the classical mixture models are excluded, but the log-likelihood function in mixture models is already in the additive form.

By using representation (1.5), the analysis of the likelihood ratio for hidden Markov models reduces to that of products of Markov random matrices. In order to apply the results developed in Section 2, we first need to check that Condition A holds.

Recall that the state space $D = \{1, \ldots, d\}$ is finite, and $\{(X_n, \xi_n), n \geq 0\}$ defined in (1.1) and (1.2) is a Markov chain on $D \times R$. Therefore, each component $p_{xy}(\theta) f(\xi_k; \varphi_y(\theta)|\xi_{k-1})$ in the matrix M_k represents that $X_{k-1} = x$ and $X_k = y$, and ξ_k is a Markov chain with transition probability density $f(\xi_k; \varphi_y(\theta)|\xi_{k-1})$. Hence, M_k is a Markov random matrix for $k = 1, \ldots, n$, and $\{((X_n, \xi_n), M_n \cdots M_0), n \geq 0\}$ is a product of Markov random matrices. Note that Condition C1 implies that the w-uniform ergodicity Condition A1 holds for all $\theta \in \Theta$, and C2 implies the moment Condition A2 holds.

In order to check Condition A3 holds, we may assume without loss of generality that $p_{xy}(\theta) \geq \gamma(\theta) > 0$ for all $x, y \in D$ due to the ergodicity

condition in C1. Conditioned on the full sequence $\mathbf{X}$, $f(\xi_k; \varphi_y(\theta)|\xi_{k-1})$ is a transition probability density and, hence, is positive $P^{(\theta)}$ almost surely for any $\theta \in \Theta$. As a result, all entries in M_1 are positive $P^{(\theta)}$ almost surely; therefore, $\{((X_n, \xi_n), M_n \cdots M_0), n \geq 0\}$ on $(D \times R) \times Gl(d, R)$ is strongly irreducible for all $\theta \in \Theta$.

It is known that a product of Markov random matrices $\{((X_n, \xi_n), M_n \cdots M_0), n \geq 0\}$ on $(D \times R) \times Gl(d, R)$ is contracting if for $\pi(\theta) f(\cdot, \varphi_1(\theta))$-almost all (x, s_0), there exists a matrix M in the smallest closed semigroup in $Gl(d, R)$ which contains the support of $P^{(\theta)}((D \times R) \times \cdot | X_0 = x, \xi_0 = s_0)$, such that M has a unique largest absolute eigenvalue. [This is an easy generalization of Corollary IV. 2.2 of Bougerol and Lacroix (1985).] Since the dimension of the matrix is finite and $f(\xi_1; \varphi_y(\theta)|\xi_0)$ is a transition probability density for all $y \in D$, it follows that for all $\theta \in \Theta$, there exists $\xi_1 \in R$ such that $f(\xi_1; \varphi_y(\theta)|\xi_0) > 0$ for all $y \in D$. Therefore, without loss of generality, we may let $f(\xi_1; \varphi_y(\theta)|\xi_0) = 1$ and consider the matrix $P_\theta = [p_{xy}(\theta)]$. By Condition C1, there exists $n_0 > 0$ such that $[p_{xy}(\theta)]^{n_0}$ is positive for all $\theta \in \Theta$, where $[\,\cdot\,]^{n_0}$ denotes n_0 multiplications of the matrix; hence, by the Perron–Frobenius theorem for positive matrices, $[p_{xy}(\theta)]^{n_0}$ has a unique largest eigenvalue. This implies that $\{((X_n, \xi_n), M_n \cdots M_0), n \geq 0\}$ on $(D \times R) \times Gl(d, R)$ is contracting for all $\theta \in \Theta$.

Now, let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from the hidden Markov model $\{\xi_n, n \geq 0\}$. Here and in the sequel, we always assume the initial distribution of (X_0, ξ_0) is the stationary distribution $\pi(\theta) f(\xi_0; \varphi_{X_0}(\theta))$. Let $\mathbf{P}^{(\theta)} := \mathbf{P}_\pi^{(\theta)}$ be the probability measure of $\{W_n, n \geq 0\}$, induced by the hidden Markov model, and let $\mathbf{E}^{(\theta)} := \mathbf{E}_\pi^{(\theta)}$ be the expectation under the probability $\mathbf{P}^{(\theta)}$. The sequential probability ratio test for testing a simple hypothesis versus a simple hypothesis is simply based on observation of the first exit of $\log S_n$ from an interval (a, b), for $-\infty < a \leq 0 < b < \infty$, where S_n is defined in (1.3). The stopping time, T, which is the time of first exit, is the sample size required for a decision. If $\log S_n$ exits (a, b) above, that is, if $\log S_n \geq b$, then the decision is in favor of the statistical hypothesis H_1. Likewise, if $\log S_n$ exits below, that is, $\log S_n \leq a$, then the decision is in favor of H_0. It turns out that the drift of $\log S_n$ is determined by $\Lambda'(0)$. This follows from a large deviations result for products of Markov random matrices; $\log S_n / n$ converges in probability to $\Lambda'(0)$ (at an exponential rate). Hence, $\mathbf{E}^{(\theta_0)} \log S_n \sim n\Lambda'(0)$, where $\Lambda'(\theta)$ denotes the derivative of $\Lambda(\theta)$. Any reasonable sequential test will tend to have $\log S_n$ exit (a, b) below under hypothesis H_0, and above under hypothesis H_1. Thus, we shall characterize the statistical hypotheses as follows:

$$H_0\colon\ \Lambda'(0) < 0 \qquad \text{(negative drift)}, \tag{3.1}$$

$$H_1\colon\ \Lambda'(0) > 0 \qquad \text{(positive drift)}. \tag{3.2}$$

LEMMA 2. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov model $\{\xi_n, n \geq 0\}$ satisfying Condition* C. *Assume that $\Lambda'(0) \neq 0$; then,*

$\mathbf{P}^{(\theta_0)}(T < \infty) = 1$, *where* T *is defined in* (1.4). *In general, if there exist* $\delta > 0$ *small enough and* $|\theta| < \delta$ *such that* $\Lambda'(\theta) \neq 0$, *then* $\mathbf{P}^{(\theta)}(T < \infty) = 1$.

PROOF. Consider only the case of $\Lambda'(0) < 0$. For fixed $\Lambda'(0) < \delta < 0$, by a proof similar to the large deviations result of Theorem 4.3 given by Bougerol (1988), we have that $\mathbf{P}^{(\theta_0)}(\log S_n \geq n\delta)$ vanishes exponentially fast. Based on this exponential rate of convergence, it follows from the Borel–Cantelli lemma that the event $\{\log S_n \geq n\delta\}$ occurs only finitely often $\mathbf{P}^{(\theta_0)}$ almost surely, and this implies that $\{\log S_n < n\delta\}$ occurs infinitely often $\mathbf{P}^{(\theta_0)}$ almost surely; thus, $\log S_n$ must eventually cross the threshold a. This proves $\mathbf{P}^{(\theta_0)}(T < \infty) = 1$. The general statement holds under (2.5) and the same argument. □

In particular, for the case of $\theta \in R$, it is well known that if $|\theta| < \delta$, $\Lambda(\theta)$ is a lower semicontinuous, strictly convex function with essential domain $\Theta = \{\theta : |\theta| < \delta,\ \Lambda(\theta) < \infty\}$ being an interval containing the point $\theta = 0$. For an H_0 distribution, we define

$$\theta^* = \sup_{|\theta|<\delta} \{\theta : \Lambda(\theta) \leq 0\}, \tag{3.3}$$

and for an H_1 distribution, we define

$$\theta_* = \inf_{|\theta|<\delta} \{\theta : \Lambda(\theta) \leq 0\}. \tag{3.4}$$

The lower semicontinuity of $\Lambda(\theta)$ implies that $\Lambda(\theta^*) \leq 0$ ($\Lambda(\theta_*) \leq 0$). Subject to the drift conditions (3.1) and (3.2), we can have $\theta^* = 0$ ($\theta_* = 0$) only if 0 is a boundary point of Θ. In particular, if Θ has a nonempty interior containing $\theta = 0$, then $\theta^* \neq 0$ ($\theta_* \neq 0$). From (3.1) and (3.2), we have $\theta^* > 0$ for H_0 and $\theta_* < 0$ for H_1. We may have $\theta^* = \pm\delta$ ($\theta_* = \pm\delta$).

For $|\theta| < \delta$, define

$$r_a(\theta) = \frac{\mathbf{E}^{(\theta_0)}[\exp(\theta(\log S_T - a))r(W_T;\theta) | \log S_T \leq a]}{\mathbf{E}^{(\theta_0)}[r(W_0;\theta)]} \tag{3.5}$$

and

$$r_b(\theta) = \frac{\mathbf{E}^{(\theta_1)}[\exp(\theta(\log S_T - b))r(W_T;\theta) | \log S_T \geq b]}{\mathbf{E}^{(\theta_1)}0[r(W_0;\theta)]}. \tag{3.6}$$

The next result is a Chernoff bound for sequential probability ratio tests with hidden Markov chain data.

LEMMA 3. *Let* $\xi_0, \xi_1, \ldots, \xi_n$ *be a sequence of random variables from a hidden Markov model* $\{\xi_n, n \geq 0\}$ *satisfying Condition* C. *Let* α *and* β *be the type* I *and type* II *error probabilities, respectively. Under* H_0, *for any* $\theta \in [0, \theta^*]$ *such that* $\mathbf{E}^{(\theta_1)}[r(W_0;\theta)] < \infty$,

$$\alpha \leq r_b(\theta)^{-1} e^{-\theta b}, \tag{3.7}$$

and under H_1, *for any* $\theta \in [\theta_*, 0]$ *such that* $\mathbf{E}^{(\theta_0)}[r(W_0;\theta)] < \infty$,

$$\beta \le r_a(\theta)^{-1} e^{-\theta a}. \tag{3.8}$$

PROOF. Consider H_0 only. Since $\mathbf{P}^{(\theta)}(T < \infty) = 1$, from (2.6) in Theorem 1 we have

$$\begin{aligned} 1 &= \frac{\mathbf{E}^{(\theta_1)}[\exp(\theta \log S_T - T\Lambda(\theta))r(W_T;\theta)]}{\mathbf{E}^{(\theta_1)}[r(W_0;\theta)]} \\ &\ge \frac{\mathbf{E}^{(\theta_1)}[\exp(\theta \log S_T - T\Lambda(\theta))r(W_T;\theta) | \log S_T \ge b]}{\mathbf{E}^{(\theta_1)}[r(W_0;\theta)]}\alpha. \end{aligned}$$

Since $\Lambda(\theta) \le 0$, it follows that $-T\Lambda(\theta) \ge 0$. Hence, from the above inequalities we obtain $1 \ge r_b(\theta)e^{\theta b}\alpha$. □

In many cases, when $\log S_n\pi$ crosses one of the thresholds at time T, the difference between $\log S_T$ and the threshold is negligible. That is, either $\log S_T \approx a$ or $\log S_T \approx b$. These are called Wald's approximations [see Chapter 3 of Woodroofe (1982)]. Applying Wald's approximations to (3.5) and (3.6), we have

$$r_a(\theta) \approx \hat{r}_a(\theta) = \frac{\mathbf{E}^{(\theta_0)}[r(W_T;\theta) | \log S_T \le a]}{\mathbf{E}^{(\theta_0)}[r(W_0;\theta)]} \tag{3.9}$$

and

$$r_b(\theta) \approx \hat{r}_b(\theta) = \frac{\mathbf{E}^{(\theta_1)}[r(W_T;\theta) | \log S_T \ge b]}{\mathbf{E}^{(\theta_1)}[r(W_0;\theta)]}. \tag{3.10}$$

These approximations may be applied to (3.7) or (3.8) to obtain useful error probability approximations by ignoring the overshoot. However, the overshoot may play an important factor in some situations. In the case of i.i.d. observations, with an additional assumption that the drift tends to zero at a certain rate, Siegmund (1979) derived a corrected Brownian approximation for the error probabilities in an exponential family by means of conjugate transform and renewal theory. This method has been extended by Asmussen (1989) and Fuh (1997) for ruin probabilities in finite state Markov chains. In this section, we will apply the renewal theorem from Kesten (1973), and Fuh and Lai (2001) to approximate the overshoot $\mathbf{E}^{(\theta_0)}(\log S_T - b)$ and to obtain more accurate approximations for the error probabilities and expected sample sizes of the tests. Although Kesten's result was developed for products of i.i.d. random matrices under different assumptions, and Fuh and Lai's result was developed for w-uniformly ergodic Markov chains (note that W_n is uniformly ergodic under Hölder continuous norm), it can be extended to products of Markov random matrices satisfying Condition A without any difficulty.

Assume that $b > 0$; we define the stopping time $\tau_b = \inf\{n : \log S_n > b\}$ and define $\tau_+ = \tau_0$ as the first ladder time, and let $\tau_n = \inf\{n \geq \tau_{n-1} : \log S_n > b\}$ be the consecutive descending ladder time. Let

$$\mathbf{P}_+^{(\theta)} = \mathbf{P}_\pi^{(\theta)}\{\log S_{\tau_+} \leq s, \tau_+ < \infty, W_{\tau_+} \in dy\} \tag{3.11}$$

denote the transition probability associated with the products of Markov random matrices based on the ascending ladder point $(\tau_+, \log S_{\tau_+})$. Let $\{W_{\tau_n}, n \geq 0\}$ be the ladder Markov chain defined as in (3.11) associated with $\{W_n, n \geq 0\}$. Under Condition C and $\gamma > 0$, $\{W_n, n \geq 0\}$ is aperiodic. An argument similar to Lemma 2 of Guivarch and Raugi (1986) leads to the conclusion that it is irreducible. Also, Lemma 4 in the Appendix shows that $\{W_n, n \geq 0\}$ is w-uniformly ergodic. Therefore, Theorem 9.1.8 of Meyn and Tweedie (1993) implies that $\{W_n, n \geq 0\}$ is Harris recurrent. And by making use of Theorem 1 of Alsmeyer (2000), we have that $\{W_{\tau_n}, n \geq 0\}$ is also Harris recurrent. Hence, $\mathbf{P}_+^{(\theta)}$ has an invariant measure π_+.

It is known that there exists $\delta > 0$ small enough such that $0 \neq |\theta| < \delta$, and that $\Lambda(\theta)$ is a strictly convex and real analytic function for which $\Lambda'(\theta) = \mathbf{E}^{(\theta)} \log S_1$. Therefore,

$$\mathbf{E}^{(\theta)} \log S_1 <, = \text{ or } > 0 \quad \Longleftrightarrow \quad \theta <, = \text{ or } > 0.$$

For any $\delta > 0$ small enough and $0 \neq |\theta| < \delta$, there is at most one value θ', necessarily of opposite sign, for which $\Lambda(\theta) = \Lambda(\theta')$. Assume that such θ' exists; we may let $\theta_0 = \min(\theta, \theta')$ and $\theta_1 = \max(\theta, \theta')$ such that $\theta_0 < 0 < \theta_1$ and $\Lambda(\theta_0) = \Lambda(\theta_1)$, and let $\Delta = \theta_1 - \theta_0$. Since the normalization of the mean is zero and the variance is 1, we can also assume without loss of generality that $\Lambda''(0) = \sigma^2 = 1$, where $\Lambda''(\theta)$ denotes the second derivative of $\Lambda(\theta)$ with respect to θ. Let $\Theta = \{|\theta| < \delta : \mathbf{E}^{(\theta)}(e^{\theta \log S_1}) < \infty\}$ such that Θ is open and $\Theta \neq \{0\}$.

The following theorem gives corrected Brownian approximations for error probability and expected sample size of the SPRT. The proof will be given in the Appendix.

THEOREM 3. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov model $\{\xi_n, n \geq 0\}$ satisfying Condition* C. *Denote ξ^- as the negative part of ξ. Assume further that there exists $\varepsilon > 0$ such that $\inf_{x,\xi} \mathbf{P}^{(\theta_0)}\{\log \|S_1\| \leq -\varepsilon | W_0 = ((x, \xi), \overline{S_0 \pi})\} > 0$. Suppose $a \to -\infty$, $b \to \infty$ and $\theta_0 \uparrow 0$ in such a way that $|a|/b \to \eta \in (0, \infty)$ and $\theta_0 b \to -\delta \leq 0$; then,*

$$\begin{aligned} &\mathbf{P}^{(\theta_0)}\{\log S_T > b\} \\ &\quad = \frac{1 - \exp\{\Delta(a + \rho_- - c_-)\}}{\exp\{\Delta(b + \rho_+ - c_+)\} - \exp\{\Delta(a + \rho_- - c_-)\}} + o(\Delta), \end{aligned} \tag{3.12}$$

where

$$\rho_+ = \frac{\mathbf{E}^{(\theta_0)}(\log S_{\tau_+})^2}{2\mathbf{E}^{(\theta_0)} \log S_{\tau_+}},$$

$$\rho_- = \frac{\mathbf{E}^{(\theta_1)}(\log S_{\tau_-})^2}{2\mathbf{E}^{(\theta_1)} \log S_{\tau_-}},$$

$$c_+ = \mathbf{E}^{(\theta_0)} r'(W_0; 0) - \frac{\mathbf{E}^{(\theta_0)} \log S_{\tau_+} r'(W_{\tau_+}; 0)}{\mathbf{E}^{(\theta_0)} \log S_{\tau_+}},$$

$$c_- = \mathbf{E}^{(\theta_1)} r'(W_0; 0) - \frac{\mathbf{E}^{(\theta_1)} \log S_{\tau_-} r'(W_{\tau_-}; 0)}{\mathbf{E}^{(\theta_1)} \log S_{\tau_-}},$$

where τ_- is the first descending ladder time and π_- is the stationary distribution of $\mathbf{P}^{(\theta_1)}\{W_{\tau_-} \in dy,\ \tau_- < \infty | W_0 = ((x, \xi), \overline{S_0 \pi})\}$. Also,

$$\begin{aligned} \mathbf{E}^{(\theta_0)} T = \theta_0^{-1}[(a + \rho_- - c_-) + (b - a + \rho_+ - \rho_- - c_+ + c_-) p^*] \\ + o(\Delta^{-1}), \end{aligned} \tag{3.13}$$

where p^* denotes the right-hand side of (3.12).

REMARKS. 1. If $\Lambda(\cdot)$ is a function symmetric about zero, then $\Delta = 2\theta_0$. Also, by definition, T is the first boundary crossing time with two barrier linear boundary (a, b). Therefore, it is easy to see that T tends to ∞ as b approaches ∞ and a approaches $-\infty$.

2. Note that (3.12) and (3.13) are just Wald's approximations but with b and a replaced by $b + \rho_+ - c_+$ and $a + \rho_- - c_-$, where ρ_+ (ρ_-) is the correction of the overshoot for discrete time and c_+ (c_-) reflects the Markovian dependence.

3. Numerical calculation of ρ_+ (ρ_-) and c_+ (c_-) involves ladder variables and Markovian Wiener–Hopf factorization, which will be published in a separate paper.

4. Asymptotic optimality of SPRT. Let $\{\xi_n, n \geq 0\}$ be the hidden Markov model with transition kernel (1.1) and satisfying Condition C. Consider the problem of testing the simple hypothesis sequentially to see whether $\mathbf{Q} := \mathbf{Q}_\pi := P_\pi^{(\theta_0)}$ or $\mathbf{P} := \mathbf{P}_\pi := P_\pi^{(\theta_1)}$ is the transition kernel of $\{\xi_n, n \geq 0\}$. Let S_n be the likelihood ratio based on $\xi_0, \xi_1, \ldots, \xi_n$ as defined in (1.3). The SPRT stops sampling at stage $T := \inf\{n : \log S_n \leq a \text{ or } \log S_n \geq b\}$, for $a \leq 0 < b$, and accepts the null hypothesis that $\mathbf{Q}$ (or the alternative hypothesis that $\mathbf{P}$) is the actual density according as $\log S_T \leq a$ (or $\log S_T \geq b$). The type I and type II error probabilities of the test are

$$\alpha = \mathbf{Q}\{\log S_T \geq b\}, \qquad \beta = \mathbf{P}\{\log S_T \leq a\}. \tag{4.1}$$

Let $\mathcal{J}(\alpha, \beta)$ denote the class of tests (N, δ) whose type I and type II error probabilities are bounded above by α and β, respectively, where N denotes the stopping rule and δ the terminal decision rule of a test. For the case where $\xi_n = X_n$ are i.i.d. random variables, Wald and Wolfowitz (1948) showed that the SPRT is optimal in the sense that it minimizes both $\int N\, dP$ and $\int N\, dQ$ in all tests

$(N, \delta) \in \mathcal{J}(\alpha, \beta)$. When $\{\xi_n, n \geq 0\}$ is a hidden Markov model, we will show that the SPRT is asymptotically optimal, in the sense of asymptotically attains [with at most a $O(1)$ discrepancy] the infima of $\int N\, dP_\nu$ and $\int N\, dQ_\nu$ over all $(N, \delta) \in \mathcal{J}(\alpha, \beta)$, as $b \to \infty$ and $a \to -\infty$.

Now, let π denote the stationary distribution of $\{X_n, n \geq 0\}$ under $\mathbf{P}$, and let π' denote the stationary distribution of $\{X_n, n \geq 0\}$ under $\mathbf{Q}$. Define the Kullback–Leibler information numbers

$$K(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_{\mathbf{P}}\left[\log \frac{\|M_1 M_0 \pi\|}{\|M_1' M_0' \pi'\|}\right], \qquad K(\mathbf{Q}, \mathbf{P}) = \mathbf{E}_{\mathbf{Q}}\left[\log \frac{\|M_1' M_0' \pi'\|}{\|M_1 M_0 \pi\|}\right], \tag{4.2}$$

where $\mathbf{E}_{\mathbf{P}}$ ($\mathbf{E}_{\mathbf{Q}}$) refers to the expectation for the induced products of Markov random matrices under probability $\mathbf{P}$ ($\mathbf{Q}$).

Note that the above defined Kullback–Leibler information number (4.2) is based on representation (1.5) and the ergodic theorem for products of random matrices in Proposition 1(v). Without this, Juang and Rabiner (1985) used the large sample average Kullback–Leibler divergence per observation between $p_n(\xi_0, \ldots, \xi_n; \theta)$ and $q_n(\xi_0, \ldots, \xi_n; \theta')$ in a numerical study on the effects of starting values and the observation sequence length on maximum likelihood estimates for hidden Markov models. [See also (14) in Leroux (1992) and (2.3) in Liu and Narayan (1994).] With this abstract definition of $K(\mathbf{P}, \mathbf{Q})$ and $K(\mathbf{Q}, \mathbf{P})$ in (4.2), we can apply Theorem 2 of Wald's equation for products of Markov random matrices to have the following theorem.

THEOREM 4. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov chain $\{\xi_n, n \geq 0\}$ satisfying Condition* C. *Assume that $K(\mathbf{P}, \mathbf{Q}) > 0$, $K(\mathbf{Q}, \mathbf{P}) > 0$ and $\sup_x\{\int (\log S_1)^2\, d\mathbf{P} + \int (\log S_1)^2\, d\mathbf{Q}\} < \infty$. Then, as $b \to \infty$ and $a \to -\infty$,*

$$\int T\, d\mathbf{P} = \inf_{(N,\delta)\in\mathcal{J}(\alpha,\beta)} \int N\, d\mathbf{P} + O(1) = (1-\beta)b/K(\mathbf{P}, \mathbf{Q}) + O(1),$$

$$\int T\, d\mathbf{Q} = \inf_{(N,\delta)\in\mathcal{J}(\alpha,\beta)} \int N\, d\mathbf{Q} + O(1) = (1-\alpha)|a|/K(\mathbf{Q}, \mathbf{P}) + O(1).$$

REMARK. Under the additional assumption of $|a|/b \to \eta > 0$, by Lemma 3, we have that $\beta b \leq e^{a/2}$ for b sufficiently large. This implies that $\beta b = o(1)$ and, consequently, the statement in Theorem 4 becomes $\int T\, d\mathbf{P} = b/K(\mathbf{P}, \mathbf{Q}) + O(1)$ and $\int T\, d\mathbf{Q} = |a|/K(\mathbf{Q}, \mathbf{P}) + O(1)$.

PROOF OF THEOREM 4. We shall only consider the assertion concerning $\int N\, d\mathbf{P}$ and $\int T\, d\mathbf{P}$ as the other assertion can be proved similarly. Define $\tau_b = \inf\{n \geq 1 : \log S_n > b\}$ for $b > 0$ and $\tau_a = \inf\{n \geq 1 : \log S_n < a\}$ for $a < 0$. Since

$T = \min(\tau_b, \tau_a)$, under the assumption of $\sup_x\{\int(\log S_1)^2\,d\mathbf{P} + \int(\log S_1)^2\,d\mathbf{Q}\} < \infty$, it is easy to see that $\mathbf{E}_{\mathbf{P}}(T) < \infty$. Hence, we can apply Theorem 2 to conclude that

$$\begin{aligned} &K(\mathbf{P},\mathbf{Q})\mathbf{E}_{\mathbf{P}}(T) + O(1) \\ &\qquad = \mathbf{E}_{\mathbf{P}} \log S_T \pi \le b\mathbf{P}\{\log S_T \ge b\} + \mathbf{E}_{\mathbf{P}}(\log S_{\tau_b} - b), \end{aligned} \tag{4.3}$$

noting that $\log S_{\tau_b} \ge b > 0$. An argument similar to Lemma 3 of Fuh (1997) results in $\mathbf{E}_{\mathbf{P}}(\log S_{\tau_b} - b) = O(1)$; therefore, (4.3) implies that

$$K(\mathbf{P},\mathbf{Q})\mathbf{E}_{\mathbf{P}}(T) \le b\mathbf{P}\{\log S_T \ge b\} + O(1) = b(1-\beta) + O(1). \tag{4.4}$$

It follows from Theorem 1 that

$$\begin{aligned} \alpha = \mathbf{Q}\{\log S_T \ge b\} &= \int_{\{\log S_T \pi \ge b\}} e^{-\log S_T}\,d\mathbf{P} \\ &= e^{-b}\mathbf{E}_{\mathbf{P}}\big\{e^{-(\log S_{\tau_b} - b)} I(\tau_b < \tau_a)\big\} = e^{-b+O(1)}, \end{aligned} \tag{4.5}$$

in view of the tightness of $\log S_{\tau_b} - b$ which follows from the Markov renewal theory for ladder variables referred to Kesten (1974) and Alsmeyer (1994). Hence, $\log\alpha = -b + O(1)$ and, similarly, it can be shown that $\log\beta = a + O(1)$, implying that $\max(\alpha,\beta) \to 0$.

The same argument involving Wald's likelihood ratio identity, like that in the proof of Theorem 2.39 of Siegmund (1985), can be used to show that for any $(N,\delta) \in \mathcal{J}(\alpha,\beta)$,

$$\mathbf{E}_{\mathbf{P}} \log S_N \ge (1-\beta)\log\big((1-\beta)/\alpha\big) + \beta\log\big(\beta/(1-\alpha)\big). \tag{4.6}$$

Applying Theorem 2 to the left-hand side of (4.6) yields

$$\inf_{(N,\delta)\in\mathcal{J}(\alpha,\beta),\ \mathbf{E}_{\mathbf{P}}N<\infty} K(\mathbf{P},\mathbf{Q})\mathbf{E}_{\mathbf{P}}N \ge (1-\beta)\log\alpha^{-1} + O(1). \tag{4.7}$$

Since $\log\alpha^{-1} = b + O(1)$ by (4.5), combining (4.4) with (4.7) proves the first assertion of the theorem. □

5. CUSUM procedures. The cumulative sum (CUSUM)-type procedure is one of the most popular change point detection algorithms used to detect a possible change from a given process to another given process. It was proposed by Page (1954) in setting that θ_n are θ_0 so that the sample statistics ξ_j are independent and identically distributed (i.i.d.) with common density f_{θ_0} when the process is in control, and that there is at most one change point ω after which θ_n are θ_1 so that ξ_j are again i.i.d. with common density f_{θ_1}. Lorden (1971) showed that subject to the "average run length" (ARL) constraint, the CUSUM procedure asymptotically minimizes the "worst case" detection delay defined in (5.1) below. Instead of studying the optimal detection problem via sequential testing theory, Moustakides (1986) formulated the worst case detection delay problem subject to an ARL constraint as an optimal solution to the optimal stopping problem. Ritov

(1990) later gave a simpler proof. For change point detection in complex dynamic systems beyond the i.i.d. setting, Bansal and Papantoni-Kazakos (1986) extended Lorden's asymptotic theory to the case where ξ_j are stationary ergodic sequences, under the condition that $\{\xi_j, j < \omega\}$ (before the change point) and $\{\xi_j, j \geq \omega\}$ (after the change point) are independent, and proved the asymptotic optimality of the CUSUM algorithm. Further extensions to general stochastic sequences ξ_n were obtained by Lai (1995, 1998). Moreover, using a change-of-measure argument, Lai (1998) also established the asymptotic optimality of the CUSUM rule under several alternative performance criteria. Finally, we mention that Yakir (1994) studied Bayesian optimal detection for a finite state Markov chain.

It is known that Lorden's method relates the CUSUM procedure to certain one-sided sequential probability ratio tests which are optimal for testing simple hypotheses. Based on the representation (1.5) of the likelihood ratio, the renewal property of the stopping rule and Wald's equation for products of Markov random matrices, we generalize Lorden's asymptotic theory to that of hidden Markov models. Our method relates the CUSUM procedure to certain one-sided sequential probability ratio tests in hidden Markov models, which have been shown in Section 4 to be asymptotically optimal for testing simple hypotheses.

Let $\xi_0, \xi_1, \ldots, \xi_{\omega-1}$ be the observations from the hidden Markov model $\{\xi_n, n \geq 0\}$ with unknown parameter θ_0, and let $\xi_\omega, \xi_{\omega+1}, \ldots$ be the observations from the hidden Markov model $\{\xi_n, n \geq 0\}$ with unknown parameter θ_1. We shall use $\mathbf{P}^{(\omega)}$ to denote such a probability measure (with change time ω) and use $\mathbf{P}_0$ to denote the case $\omega = \infty$ (no change point). Recall that $\mathbf{P}^{(\omega)}$ and $\mathbf{P}_0$ denote the probabilities for the induced products of Markov random matrices $\{((X_n, \xi_n), M_n \cdots M_0), n \geq 0\}$. Let $S_n\pi$ be defined as (1.3) for all $\theta_0, \theta_1 \in \Theta$. Define the CUSUM scheme

$$\tau := \inf\left\{n : \max_{1\leq k\leq n} (\log S_n - \log S_k) \geq c_\gamma\right\}, \tag{5.1}$$

where c_γ is chosen such that $\mathbf{E}_0\tau = \gamma$. Here and in the sequel, we define $\inf \varnothing = \infty$. When ω is finite, we are concerned with the conditional expected delay $\mathbf{E}^{(\omega)}[(\tau - \omega + 1)^+ | \xi_0, \xi_1, \ldots, \xi_{\omega-1}]$, whose supreme over $\{\omega, \xi_0, \xi_1, \ldots, \xi_{\omega-1}\}$ represents the worst case delay. More precisely, we want to show that the CUSUM scheme τ minimizes asymptotically as $\gamma \to \infty$:

$$\bar{\mathbf{E}}_1 N = \sup_{\omega\geq 1} \operatorname{esssup} \mathbf{E}^{(\omega)}\big[(N - \omega + 1)^+ | \xi_0, \xi_1, \ldots, \xi_{\omega-1}\big] \tag{5.2}$$

over all schemes N with $\mathbf{E}_0 N \geq \gamma$.

Now let $\xi_0, \xi_1, \ldots, \xi_{\omega-1}$ be the observations from the hidden Markov model $\{\xi_n, n \geq 0\}$ with probability $\mathbf{P}$ and $\xi_\omega, \xi_{\omega+1}, \ldots$ be the observations from the hidden Markov model $\{\xi_n, n \geq 0\}$ with probability $\mathbf{Q}$. Let K be the maximum likelihood estimate of the time ω. Following the idea of Lorden (1971) and its extension by Bansal and Papantoni-Kazakos (1986), the likelihood ratio

CUSUM scheme (5.1) corresponds to stopping when a one-sided SPRT with log-boundary c_γ based on $\xi_K, \xi_{K+1}, \ldots$ shows significant evidence against the null hypothesis H_0: $\mathbf{P}^{(\theta)} = \mathbf{P}$. Since the initial distribution of (X_0, ξ_0) is the stationary distribution $\pi_x f(\cdot; \varphi_x(\theta))$, it follows that $\xi_0, \xi_1, \ldots, \xi_n$ forms a stationary sequence, and that the CUSUM scheme (5.1) can be expressed as

$$\tau := \min_{k \geq 1}\{T_k + k - 1\}, \tag{5.3}$$

where T_k is the stopping time of the one-sided SPRT applied to $\xi_k, \xi_{k+1}, \ldots$.

By using the asymptotic representation of the expected sample sizes in Theorem 4 and Wald's equation for products of Markov random matrices in Theorem 2, we obtain the following theorem which establishes the asymptotic lower bound $(K(\mathbf{P}, \mathbf{Q})^{-1} + o(1)) \log \gamma$ of $\bar{\mathbf{E}}_1 N$ under the ARL constraint $\mathbf{E}_0 N \geq \gamma$.

THEOREM 5. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov model $\{\xi_n, n \geq 0\}$ satisfying Condition* C. *Then, as $\gamma \to \infty$,*

$$\inf\{\bar{\mathbf{E}}_1 N : \mathbf{E}_0 N \geq \gamma\} \geq \big(K(\mathbf{P}, \mathbf{Q})^{-1} + o(1)\big) \log \gamma. \tag{5.4}$$

PROOF. Let $K_1 = K(\mathbf{P}, \mathbf{Q})$. It suffices to show that for any given $0 < \varepsilon < 1$, there is a $C(\varepsilon) < \infty$ such that for all stopping times N,

$$K_1 \bar{\mathbf{E}}_1 N \geq (1 - \varepsilon) \log \mathbf{E}_0 N - C(\varepsilon). \tag{5.5}$$

We fix ε and define stopping times $T_0 = 0 < T_1 < T_2 < \cdots$ as follows: $T_{j+1} (j = 0, 1, \ldots)$ is the smallest n (or ∞ if there is no n) such that $n > T_j$ and

$$p_n(\xi_0, \xi_1, \ldots, \xi_n; \theta_1) \leq \varepsilon p_n(\xi_0, \xi_1, \ldots, \xi_n; \theta_0). \tag{5.6}$$

By an argument similar to the estimation of the error probabilities of the SPRT in Lemma 3 in Section 3, we have $\mathbf{P}_1(T_1 < \infty) \leq \varepsilon$, and the same argument is easily modified to yield that if $\mathbf{P}_1(D_{rk}) > 0$, then $\mathbf{P}_{k+1}(T_r < \infty | D_{rk}) \leq \varepsilon$, where $D_{rk} = \{T_{r-1} = k < N\}$, which depends only on $\xi_0, \xi_1, \ldots, \xi_k$, and where $\mathbf{P}_k$ refers to the probability based on the observations $\xi_k, \xi_{k+1}, \ldots$.

Consider all D_{rk} for which $\mathbf{P}_0(D_{rk}) > 0$ and, hence, $\mathbf{P}_{k+1}(D_{rk}) > 0$; since $\xi_0, \xi_1, \ldots, \xi_n$ is a stationary sequence, $\mathbf{P}_{k+1}$ gives the same distribution of $\xi_0, \xi_1, \ldots, \xi_k$ as does $\mathbf{P}_0$. On the subset D_{rk}, N and T_r determine the following sequential test based on $\xi_{k+1}, \xi_{k+2}, \ldots$: stop at $\min(N, T_r)$ and

$$\text{decide } \mathbf{P}_{k+1} \text{ is true if } N \leq T_r;$$

$$\text{decide } \mathbf{P}_0 \text{ is true if } N > T_r.$$

The number of observations taken is $\min(N, T_r) - k$, whose (conditional) expectation under $\mathbf{P}_{k+1}(\cdot | D_{rk})$ is at most $\bar{E}_1 N$. (D_{rk} belongs to the σ-algebra

of events determined by $\xi_0, \xi_1, \ldots, \xi_k$). Applying Theorem 4 and (4.5) with $\alpha = \mathbf{P}_0(N \le T_r | D_{rk})$ and $1-\beta = \mathbf{P}_{k+1}(N \le T_r | D_{rk})$, there exists a finite constant C:

$$\begin{aligned} K_1 \bar{\mathbf{E}}_1 N &= \mathbf{P}_{k+1}(N \le T_r | D_{rk}) |\log \mathbf{P}_0(N \le T_r | D_{rk})| + C \\ &\ge (1-\varepsilon) |\log \mathbf{P}_0(N \le T_r | D_{rk})| + C, \end{aligned} \tag{5.7}$$

where the latter inequality of (5.7) follows from $\mathbf{P}_{k+1}(N \le T_r | D_{rk}) \ge \mathbf{P}_{k+1}(T_r = \infty | D_{rk}) \ge 1-\varepsilon$.

Let R be the smallest $r \ge 1$ (or ∞ if there is no r) such that $T_r > N$. If $\mathbf{P}_0(R \ge r) > 0$, then $\mathbf{P}_0(R < r+1 | R \ge r)$ is well defined and equals $\mathbf{P}_0(N \le T_r | T_{r-1} < N)$, which is an average (over k) of the probabilities $\mathbf{P}_0(N \le T_r | T_{r-1} = k < N)$ satisfying (5.7). Therefore, $\mathbf{P}_0(R \ge r) > 0$ implies that

$$K_1 \bar{\mathbf{E}}_1 N \ge (1-\varepsilon) |\log \mathbf{P}_0(R < r+1 | R \ge r)| + C. \tag{5.8}$$

Since $\{(W_{T_j}, T_j), n \ge 0\}$ is a Markov chain, simple calculation shows that a lower bound of the form

$$\mathbf{P}_0(R < r+1 | R \ge r) \ge q \qquad \text{for} \quad r = 1, 2, \ldots, \text{ such that } \quad \mathbf{P}_0(R \ge r) > 0$$

implies that $\mathbf{P}_0(R \ge r+1) \le (1-q)^r$ for $r = 1, 2, \ldots$ and, hence, that $\mathbf{E}_0 R \le q^{-1}$. Thus, (5.8) yields

$$K_1 \bar{\mathbf{E}}_1 N \ge (1-\varepsilon) \log \mathbf{E}_0 R + C. \tag{5.9}$$

When $\mathbf{P}_0$ is true, we let $\rho_1^{(\varepsilon)} = \rho_+^{(\varepsilon)} = \inf\{n : \log S_n \pi \le \varepsilon\}$ be the first descending ladder time, and let $\rho_n^{(\varepsilon)} = \inf\{n \ge \rho_{n-1}^{(\varepsilon)} : \log S_n \pi \le \varepsilon\}$ be the consecutive descending ladder time. Let $\{W_{\rho_n^{(\varepsilon)}}, n \ge 0\}$ be the ladder Markov chain defined as in (3.11) associated with $\{W_n, n \ge 0\}$. Then, $\{T_j\}$ is a Markov random walk defined on $\{W_{\rho_n^{(\varepsilon)}}, n \ge 0\}$, distributed like T_1, as can be verified from (5.6). Note that $\{W_n, n \ge 0\}$ is aperiodic. An argument similar to Lemma 2 of Guivarch and Raugi (1986) leads to the conclusion that it is irreducible. Also, Lemma 4 in the Appendix shows that $\{W_n, n \ge 0\}$ is w-uniformly ergodic. Therefore, Theorem 9.1.8 of Meyn and Tweedie (1993) implies that $\{W_n, n \ge 0\}$ is Harris recurrent. By making use of Theorem 1 in Alsmeyer (2000), we have that $\{W_{\rho_n^{(\varepsilon)}}, n \ge 0\}$ is also Harris recurrent.

Next, we consider the Poisson equation

$$(I - \mathbf{P})\Delta = T_1 - \mathbf{E}_0 T_1. \tag{5.10}$$

From (5.9), we have $\mathbf{E}_0 R < \infty$. By Theorem 14.0.1 and Theorem 17.4.2 of Meyn and Tweedie (1993), (5.10) has a bounded solution. Therefore, Wald's equation still holds for $\{W_{\rho_n^{(\varepsilon)}}, n \ge 0\}$ by Theorem 4 in Fuh and Zhang (2000); that is, we have $\mathbf{E}_0 T_R = \mathbf{E}_0 R \cdot E_0 T_1 + C_1$, for some constant C_1. Hence,

$$\log \mathbf{E}_0 N \le \log \mathbf{E}_0 T_R = \log \mathbf{E}_0 R + B(\varepsilon), \tag{5.11}$$

where $B(\varepsilon) = \log \mathbf{E}_0 T_1 + C'$, which is finite for all ε and does not depend on N. Relation (5.5) follows from (5.9) and (5.11), and the proof is complete. $\square$

Based on (5.3) and the upper bound for the one-sided SPRT $\mathbf{P}_0(T_1 < \infty) \le \gamma^{-1}$, the following theorem yields that $\mathbf{E}_0\tau \ge \gamma$.

THEOREM 6. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov model $\{\xi_n, n \ge 0\}$ satisfying Condition* C. *Let N be an extended stopping variable with respect to $\xi_0, \xi_1, \xi_2, \ldots$ such that $\mathbf{P}_0(N < \infty) \le \alpha$. For $k = 1, 2, \ldots$, let N_k denote the stopping time obtained by applying N to $\xi_k, \xi_{k+1}, \ldots$ and define $N^* = \min_{k\ge1}\{N_k + k - 1\}$. Then, N^* is a stopping time, and*

$$\mathbf{E}_0 N^* \ge 1/\alpha. \tag{5.12}$$

PROOF. For $k = 1, 2, \ldots$, define $\eta_k = 1$ if $N_k < \infty$ and $\eta_k = 0$ if $N_k = \infty$. By the ergodic property of $\xi_1, \xi_2, \ldots$ in Proposition 1(v), we have

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} \eta_k = \mathbf{E}_0\eta_1 = \mathbf{P}_0(N_1 < \infty) \le \alpha \qquad \text{a.s. } (\mathbf{P}_0). \tag{5.13}$$

Note that the last inequality in (5.13) is due to the stationarity of $\xi_0, \xi_1, \xi_2, \ldots$. Assume that $\mathbf{E}_0 N^* < \infty$ [otherwise, we have (5.12)]. Let $N_0^* = 0$ and define $N_1^* < N_2^* < \cdots$ recursively as follows. If $N_{m-1}^* = n$, then for each $r = 1, 2, \ldots$, apply N to $\xi_{n+r}, \xi_{n+r+1}, \ldots$ and let N_m^* be the first time stopping occurs for some r. Then $N_1^* = N^*$ and $N_1^*, N_2^* - N_1^*, N_3^* - N_2^*, \ldots$ forms a Markov chain. Clearly, for $m = 0, 1, \ldots$, there exists r such that $\eta_{N_{m^*}+r} = 1$, and this causes the stop at N_{m+1}^*. Therefore, $\eta_{N_{m^*}+1} + \cdots + \eta_{N_{m+1}^*} \ge 1$. Hence, $\eta_1 + \cdots + \eta_{N_{m^*}} \ge m$ for $m = 0, 1, \ldots$, so that

$$\frac{\eta_1 + \cdots + \eta_{N_{m^*}}}{N_{m^*}} \ge \frac{m}{N_{m^*}}. \tag{5.14}$$

The last paragraph in the proof of Theorem 5 shows that $\{W_n, n \ge 0\}$ is Harris recurrent. The strong law of large numbers for Harris chains [cf. Theorem 17.0.1 of Meyn and Tweedie (1993)] implies that as $m \to \infty$, the right-hand side of (5.14) approaches $(\mathbf{E}_0 N^*)^{-1}$ and the left-hand side tends to a limit $\le \alpha$ by (5.13), thus proving (5.12). □

The proof of the upper bound of $\bar{\mathbf{E}}_1 N^*$ in Theorem 2 of Lorden (1971) depends heavily on the independence structure and is difficult to generalize to dependent data. The extension of Lorden's method and the results obtained by Bansal and Papantoni-Kazakos (1986) for the case of stationary ergodic sequences ξ_j involve a strong assumption that requires independence between $\{\xi_j, j < \omega\}$ and $\{\xi_j, j \ge \omega\}$. Instead of using the independence assumption employed in previous studies, we apply the uniform strong law of large numbers to the induced Markov chains $\{W_n, n \ge 0\}$ to get the following theorem.

THEOREM 7. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov model $\{\xi_n, n \geq 0\}$ satisfying Condition* C. *Let N^* be defined as in Theorem* 6. *Then, as $\gamma \to \infty$,*

$$\bar{\mathbf{E}}_1 N^* \leq (K(\mathbf{P}, \mathbf{Q})^{-1} + o(1)) \log \gamma. \tag{5.15}$$

PROOF. Let $K_1 = K(\mathbf{P}, \mathbf{Q})$. To prove (5.15), it suffices to show that for any $0 < \delta < 1$, as $\gamma \to \infty$,

$$\begin{aligned} &\sup_{\omega \geq 1} \operatorname{esssup} \mathbf{E}^{(\omega)} \{(N - \omega + 1)^+ | \mathcal{F}_{\omega-1}\} \\ &\qquad \leq (1 + o(1))(1 - \delta)^{-1} K_1^{-1} \log \gamma, \end{aligned} \tag{5.16}$$

where $\mathcal{F}_{\omega-1}$ is the σ-algebra generated by $\{\omega, \xi_1, \ldots, \xi_\omega\}$. For $c \sim \log \gamma$, let n_c be the largest integer $\leq (1 - \delta)^{-1} K_1^{-1} c$. The last paragraph in the proof of Theorem 5 shows that $\{W_n, n \geq 0\}$ is a Harris recurrent Markov chain. By the strong law of large numbers for the additive component of W_n [cf. Theorem 17.0.1 of Meyn and Tweedie (1993)], we have with probability 1, $\mathbf{P}_1\{\log S_n \pi \leq (K_1 - \delta)n | W_0 = ((x, \xi), \overline{S_0 \pi})\} \to 0$ as $n \to \infty$. Along with this Condition C2 implies that $\sup_{x,\xi} \mathbf{P}_1\{\log S_n \leq (K_1 - \delta)n | W_0 = ((x, \xi), \overline{S_0 \pi})\} \to 0$ as $n \to \infty$. Hence,

$$\begin{aligned} &\lim_{n \to \infty} \sup_{t \geq \omega \geq 1} \operatorname{esssup} \mathbf{P}^{(\omega)} \left\{ n^{-1} \sum_{j=t}^{t+n} \sigma(W_{j-1}, W_j) < (K_1 - \delta) | \xi_0, \xi_1, \ldots, \xi_{t-1} \right\} \\ &\qquad = 0, \end{aligned} \tag{5.17}$$

and this implies that

$$\sup_{t \geq \omega \geq 1} \operatorname{esssup} \mathbf{P}^{(\omega)} \left\{ \sum_{t \leq j \leq t + n_c - 1} \sigma(W_{j-1}, W_j) < c | \xi_0, \xi_1, \ldots, \xi_{t-1} \right\} \leq \delta \tag{5.18}$$

for all large c. Hence, it follows that, for all sufficiently large c, for any $\omega \geq 1$ and $k \geq 1$,

$$\begin{aligned} &\operatorname{esssup} \mathbf{P}^{(\omega)} \{N - \omega + 1 > k n_c | \mathcal{F}_{\omega-1}\} \\ &\leq \operatorname{esssup} \mathbf{P}^{(\omega)} \left\{ \sum_{j=\omega+(l-1)n_c}^{\omega + l n_c - 1} \sigma(W_{j-1}, W_j) < c \text{ for all } 1 \leq l \leq k | \mathcal{F}_{\omega-1} \right\} \\ &\leq \delta^k. \end{aligned} \tag{5.19}$$

Applying (5.19) and conditioning on $\xi_0, \xi_1, \ldots, \xi_{\omega+(l-1)n_c-1}$ for $l = k$, $k-1, \ldots, 1$ in succession, we have, for all sufficiently large c,

$$\sup_{\omega \geq 1} \operatorname{esssup} \mathbf{E}^{(\omega)} \{n_c^{-1} (N - \omega + 1)^+ | \mathcal{F}_{\omega-1}\} \leq \sum_{k=0}^{\infty} \delta^k = (1 - \delta)^{-1}. \tag{5.20}$$

Since $n_c \sim (1 - \delta)^{-1} K_1^{-1} \log \gamma$, (5.20) implies (5.15). □

Theorem 7 establishes that the CUSUM procedure attains the asymptotic lower bound for detection delay in (5.4). Along with Theorems 5 and 6, we have proved the asymptotic optimality of the CUSUM procedure under the ARL constraint.

6. Examples and applications. Several interesting examples which are widely used in speech recognition, digital communications, bioinformatics and economics can be formalized as the hidden Markov model (1.1) and (1.2). In this section, we will demonstrate the application of our results to models of i.i.d. hidden Markov models, switch Gaussian regression and switch Gaussian autoregression. Note that finite state space models can also be formulated in this framework.

EXAMPLE 1 (i.i.d. hidden Markov models). When ξ_n defined in (1.1) are conditionally independent given the full sequences $\mathbf{X}$, the hidden Markov model was considered by Leroux (1992), Bickel and Ritov (1996), Fuh (1998) and Bickel, Ritov and Rydén (1998). It is also called the mixture model with Markovian regimes.

We first consider a simple case that $\{X_n, n \geq 0\}$ is a two-state ergodic Markov chain and conditional on X_n, ξ_n have normal densities with means and variances $\mu_1 = 2$, $\mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$. The requirement of $p_{xy}(\theta) > 0$ for all $x, y = 1, 2$ and for all $\theta \in \Theta$ is a sufficient condition of C1. Simple calculation leads to the conclusion that the moment condition in C2 reduces to $Ee^{2\xi_1} < \infty$. The inverse matrix condition of the random matrix in C2 indicates that it is not an independent mixture. When $p_{xy}(\theta)$ are known, positive for $x, y = 1, 2$ and $p_{11}(\theta) \neq p_{21}(\theta)$, the second mean μ_2 is 0, $\sigma_1^2 = \sigma_2^2 = 1$ and μ_1 is the only unknown parameter, Conditions C1 and C2 are satisfied.

In general, we consider a finite state ergodic Markov chain $\{X_n, n \geq 0\}$ with transition probability matrix $P(\theta)$. The parameter θ is chosen such that the determinant of $P(\theta)$ is not zero. For each fixed $x \in D$, let $f(\cdot; \varphi_x(\theta))$ be a Lebesgue density on R, and assume further that f is continuous and positive with $\lim_{\xi \to \pm\infty} f(\xi; \varphi_x(\theta)) = 0$ and $\int_{-\infty}^{\infty} f^{\alpha}(\xi; \varphi_x(\theta))\, d\xi < \infty$ for some $\alpha < 1$. Suppose the identifiability holds; a simple argument suffices to show that Conditions C1 and C2 hold.

When ξ_n is equal to x_n in (1.1), this reduces to the classical example of Markov chains. Sadowsky (1989) investigated Wald's likelihood ratio identity and Wald's equation for *uniformly recurrent* Markov chains, in the sense that there exist $c_2 > c_1 > 0$, $n \geq 1$ and a probability measure μ^* on $\mathcal{D}$ such that

$$c_1 \mu^*(A) \leq P\{X_n \in A | X_0 = x\} \leq c_2 \mu^*(A) \tag{6.1}$$

for all measurable subsets $A \in \mathcal{D}$ and all $x \in D$. He also assumed the exponential moment condition and the boundedness of $r'(\cdot; 0)$ which appears in (2.7). Fuh and Lai (1998) generalized the result of Wald's equation to *uniformly ergodic* ($w = 1$) Markov chains, and dropped the exponential moment condition assumption as well

as the assumption of boundedness of $r'(\cdot;0)$. By using an argument similar to that in Theorems 2 and 4, we generalize the results of Wald's equation and the asymptotic optimality of the SPRT to an irreducible w-uniformly ergodic Markov chain. Note that the Kullback–Leibler information number (4.2) in this case is $K(\theta',\theta)=\int_{x\in D} d\pi_x(\theta')\int_{y\in D} p_{xy}(\theta')\log p_{xy}(\theta')/p_{xy}(\theta)$.

Consider the following example which involves change in the mean value θ of a stable autoregressive sequence:

$$x_n=\sum_{k=1}^{p} a_k x_{n-k}+v_k+\left(1-\sum_{k=1}^{p} a_k\right)\theta, \tag{6.2}$$

where $a_1,\dots,a_p$ are autoregressive coefficients and v_k is a Gaussian sequence with zero mean and variance σ^2. By Theorem 16.5.1 of Meyn and Tweedie (1993), x_n defined in (6.2) is a w-uniformly ergodic Markov chain with $w(x)=x^2$. And Example 1 of Fuh and Lai (2001) shows that the ladder Markov chain is still w-uniformly ergodic. The existence of exponential moments for the normal distribution (with mean zero and finite variance σ^2) implies that the moment Condition C2 holds. This example can be generalized to the case of random coefficient autoregression as on page 404 of Meyn and Tweedie (1993).

EXAMPLE 2 (Gaussian regression). Let X_n be an ergodic Markov chain with finite state space $D=\{1,2,\dots,d\}$. Given that $X_n=x$, let

$$\xi_n=\sum_{k=1}^{p-1} a_x^k r_k+\sigma_x v_n, \tag{6.3}$$

where $r_k\in R$, for $k=1,2,\dots,p-1$, are deterministic values, v_n is a normal random variable with zero mean and unit variance and $a_x=(a_x^1,\dots,a_x^{p-1},\sigma_x)$ are the unknown parameters. In this case, the likelihood ratio for ξ_n given $X_n=x$ is

$$\frac{f(\xi_n|a_x,\sigma_x)}{f(\xi_n|a'_x,\sigma'_x)}=\left(\frac{\sigma'_x}{\sigma_x}\right)\exp\left\{-\frac{1}{2\sigma_x^2}\left(\xi_n-\sum_{k=1}^{p-1}a_x^k r_k\right)^2+\frac{1}{2\sigma'^2_x}\left(\xi_n-\sum_{k=1}^{p-1}a'^k_x r_k\right)^2\right\}.$$

Suppose that the transition probabilities $p_{xy}(\theta)$ are known and the determinant of the matrix $[p_{xy}(\theta)]$ is not zero for $\theta\in\Theta$. Assume further that $-\infty<a_k<\infty$, and that there exists a constant c such that $0<c<\sigma_k$, for $k=1,\dots,d$. Let θ^0 be the true parameter in some closure of Θ which does not contain $\sigma_k=0$ for $k=1,\dots,d$. Since X_n is a finite state ergodic Markov chain and ξ_n are conditionally independent normal random variables given the full sequence $\mathbf{X}$, Conditions C1 and C2 hold by straightforward calculations.

One application of this model, when combined with the hidden Markov models, is in detection of signals from a finite-state additive Gaussian channel, where each state is characterized by a different noise level [Merhav (1991)]. A useful special case is a channel with two states, one state with small σ_x^2 and another state with large σ_x^2. The transition probabilities, associated with the hidden Markov model in this case, are closely related to the amount of time the channel spends in each state. Theorems 5–7 lead to the conclusion that the CUSUM procedure is asymptotically optimal, provided there exists a constant c such that $0 < c < \sigma_k$ for $k = 1, 2, \ldots, d$.

This model is also useful for capturing occasional but recurrent regime shifts in empirical macroeconomics and dynamic econometrics. Goldfeld and Quandt (1973) studied a model for a housing market in disequilibrium, in which the demand and supply functions were specified as switching regressions as in (6.3). The reader is referred to that paper for details.

EXAMPLE 3 (Gaussian autoregression). We start with a simple scalar valued fourth-order autoregression around one of two constants μ_1 or μ_2:

$$\begin{aligned}\xi_n - \mu_{x_n} = {} & \varphi_1(\xi_{n-1} - \mu_{x_{n-1}}) + \varphi_2(\xi_{n-2} - \mu_{x_{n-2}}) \\ & + \varphi_3(\xi_{n-3} - \mu_{x_{n-3}}) + \varphi_4(\xi_{n-4} - \mu_{x_{n-4}}) + \varepsilon_n,\end{aligned} \tag{6.4}$$

where $\varepsilon_n \sim N(0, \sigma^2)$, x_n is a two-state ergodic Markov chain and $\theta = (\varphi_1, \varphi_2, \varphi_3, \varphi_4, \mu_1, \mu_2, \sigma^2)$ are the unknown parameters. This model has been studied by Hamilton (1989) in an attempt to analyze the behavior of U.S. real GNP. We may assume that only one parameter is of interest and treat the other parameters as nuisance parameters. In this case, the likelihood ratio for ξ_n given $X_n = x_n$, $n \geq 0$, is

$$\begin{aligned}\frac{f(\xi_n|x_n;\theta)}{f(\xi_n|x_n;\theta')} = \left(\frac{\sigma'}{\sigma}\right) \exp\Bigg\{ & -\left[(\xi_n - \mu_{x_n}) - \frac{1}{2\sigma^2}\sum_{k=1}^{4}\varphi_k(\xi_{n-k} - \mu_{x_{n-k}})\right]^2 \\ & + \left[(\xi_n - \mu'_{x_n}) - \frac{1}{2\sigma'^2}\sum_{k=1}^{4}\varphi'_k(\xi_{n-k} - \mu'_{x_{n-k}})\right]^2\Bigg\}.\end{aligned} \tag{6.5}$$

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 holds. Assume that for all $\theta \in \Theta$, $p_{xy}(\theta) > 0$ for all $x, y \in \{1, 2\}$, and $p_{11}(\theta) \neq p_{21}(\theta)$. Then, $\{(X_n, \xi_n), n \geq 0\}$ is a w-uniformly ergodic Markov chain with $w(x) = x^2$. And Example 1 of Fuh and Lai (2001) shows that the ladder Markov chain is still w-uniformly ergodic. This implies that Condition C1 holds. The assumption of $\varepsilon_n \sim N(0, \sigma^2)$ also implies that Condition C2 is satisfied in model (6.4). Although the random variables ξ_n depend on ξ_{n-1} and X_n only in Theorems 3 to 7, the results can be extended to dependence on $\xi_{n-4}, \ldots, \xi_{n-1}$ and $X_{n-4}, \ldots, X_{n-1}, X_n$ without any difficulty. Therefore, the SPRT for testing simple

hypothesis vs. simple hypothesis in model (6.4) is asymptotically optimal, and the CUSUM algorithm for change point detection is also asymptotically optimal.

Engel and Hamilton (1990) considered another switching autoregression model in which both mean vectors and variance-covariance matrices are functions of states:

$$\xi_n | x_n \sim N(\mu_{x_n}, \Omega_{x_n}) \qquad \text{for } x_n = 1, 2, \tag{6.6}$$

where $\theta = (\mu_1, \mu_2, \Omega_1, \Omega_2)$ are unknown parameters. In this case the likelihood ratio for ξ_n given $X_n = x_n$, $n \geq 0$ is

$$\frac{f(\xi_n | x_n; \theta)}{f(\xi_n | x_n; \theta')} = \left(\frac{|\Omega'_{x_n}|^{1/2}}{|\Omega_{x_n}|^{1/2}} \right) \exp \left\{ \frac{-(\xi_n - \mu_{x_n})^t \Omega_{x_n}^{-1} (\xi_n - \mu_{x_n})}{2} + \frac{(\xi_n - \mu'_{x_n})^t \Omega'^{-1}_{x_n} (\xi_n - \mu'_{x_n})}{2} \right\}, \tag{6.7}$$

where $|\Omega_x|$ denotes the determinant of Ω_x. Assume that there exists a constant c such that $0 < c < |\Omega_x|$ for each $x = 1, 2$. Suppose the identifiability condition holds and that μ_1, μ_2 are in R; the ergodicity assumption for the Markov chain $\{X_n, n \geq 0\}$, and the normal assumption for ξ_n given $X_n = x_n$ gives that Conditions C1 and C2 hold.

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

$$\xi_n = \sum_{k=1}^{p-1} a_{x_n}^k \xi_{n-k} + \sigma_{x_n} v_n, \tag{6.8}$$

where v_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 unknown parameters. In this case, the likelihood ratio for ξ_n given $X_n = x_n$, $n \geq 0$, is

$$\frac{f(\xi_n | a_{x_n})}{f(\xi_n | a'_{x_n})} = \left(\frac{\sigma'_{x_n}}{\sigma_{x_n}} \right) \exp \left\{ -\frac{1}{2\sigma_{x_n}^2} \left(\xi_n - \sum_{k=1}^{p-1} a_{x_n}^k \xi_{n-k} \right)^2 + \frac{1}{2\sigma'^2_{x_n}} \left(\xi_n - \sum_{k=1}^{p-1} a'^k_{x_n} \xi_{n-k} \right)^2 \right\}.$$

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 $x = 1, \ldots, d$. The same argument as that in (6.4) and (6.5) shows that Conditions C1 and C2 hold for model (6.8).

The formulation (6.8) also includes a generalization of Engle's ARCH model [Engle (1982)] to allow for occasional discrete shifts in the ARCH parameters. As another example of (6.8), we can generalize a vector autoregression so as to allow the constant terms, the autoregressive coefficients and the innovation variance-covariance matrix to be functions of the state x_n. This model also has many applications in speech recognition [Rabiner and Juang (1993)] where the only essential difference in this example is the dependence of each state.

APPENDIX

We will use the same notation as that in Section 2 unless otherwise mentioned.

Recall that $\{(X_n, \xi_n), n \geq 0\}$ defined in (1.1) and (1.2) is a Markov chain on a state space $D \times R$, where $D = \{1, \ldots, d\}$ is a finite set. In the proof of Lemma 4, for simplicity, we assume the state space of the Markov chain $\{X_n, n \geq 0\}$ is R and consider the case where the associated transition probability has transition probability density with respect to the Lebesgue measure on R. Without this constraint, the result is still correct, and the proof based on Markovian iterated random functions will be published in a separate paper.

LEMMA 4. *Let $\{X_n, n \geq 0\}$ be the Markov chain defined in Section 2 and satisfying Condition* A. *Then, the induced Markov chain W_n on $D \times P(R^d)$ defined in* (2.1) *is v-uniformly ergodic for some $v : D \times P(R^d) \to [1, \infty)$.*

PROOF. We first want to show that the Markov chain $\{(X_n, S_k \cdot \bar{u}), n \geq 0\}$, satisfies Doeblin's condition if X_n takes values on the whole real line R. 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 $k = 1, \ldots, n$. 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. page 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

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

Since Γ is compact, $m_S(S \cap C) < \infty$. This implies that the desired Doeblin condition holds if X_n takes values on the whole real line R. Note that $\{X_n, n \geq 0\}$

is w-uniformly ergodic by Condition A1. Define $v: D \times P(R^d) \to [1, \infty)$ by $v(x, \bar{u}) = w(x)$. Then, W_n is v-uniformly ergodic and we complete the proof. $\square$

PROOF OF LEMMA 1. Let h_1 be the identity function in $H(\alpha)$. By the Markovian property of W_n, we have

$$\begin{aligned}
\mathbf{E}_v\big(G_{n+1}^{(\theta)} | \mathcal{F}_n\big) &= e^{\theta \log \|\mathbb{S}_n u\| - (n+1)\Lambda(\theta)} \mathbf{E}_{W_n}\big\{r(W_1; \theta) e^{\theta \sigma(W_1, W_0)}\big\} \\
&= e^{\theta \log \|\mathbb{S}_n u\| - (n+1)\Lambda(\theta)} \big(\mathbf{T}(\theta)\mathbf{Q}(\theta) h_1\big)(W_n) \\
&\qquad\qquad \text{[by (2.3) and } \mathbf{Q}(\theta)h_1 = r(\cdot; \theta)\text{]} \\
&= e^{\theta \log \|\mathbb{S}_n u\| - n\Lambda(\theta)} (\lambda(\theta))^{-1} \{\lambda(\theta)\mathbf{Q}(\theta) h_1(W_n)\} \qquad \text{[by (2.4)]} \\
&= e^{\theta \log \|\mathbb{S}_n u\| - n\Lambda(\theta)} r(W_n; \theta) = G_n^{(\theta)}. \qquad \square
\end{aligned}$$

PROOF OF THEOREM 1. Let α_n be the Borel measurable function defined in the paragraph before Theorem 1. By using the twisting formula (2.5),

$$\begin{aligned}
&\mathbf{P}_{(x,\bar{u})}^{(\theta)}\big\{B \cap \{N = n\}\big\} \\
&\quad = \mathbf{E}_{(x,\bar{u})}^{(\theta)}[\alpha_n(M_0, M_1, \ldots, M_n)] \\
&\quad = \mathbf{E}_{(x,\bar{u})}\bigg[\exp\big(\theta \log \|\mathbb{S}_n u\| - n\Lambda(\theta)\big) \frac{r(W_n; \theta)}{r((x, \bar{u}); \theta)} \alpha_n(M_0, M_1, \ldots, M_n)\bigg].
\end{aligned}$$

Let $I_{B \cap \{N=n\}}$ denote the indicator random variable of the event $B \cap \{N = n\}$. For any $\mathcal{F}_\infty$-measurable random variable $Z \geq 0$, we have

$$\mathbf{E}[Z I_{B \cap \{N=n\}}] = \mathbf{E}\big[Z \mathbf{E}[I_{B \cap \{N=n\}} | \mathcal{F}_\infty]\big] = \mathbf{E}[Z \alpha_n(M_0, M_1, \ldots, M_n)].$$

Thus, from the above display, we have

$$\mathbf{P}_{(x,\bar{u})}^{(\theta)}\big\{B \cap \{N = n\}\big\} = \mathbf{E}_{(x,\bar{u})}\bigg[\exp\big(\theta \log \|\mathbb{S}_n u\| - n\Lambda(\theta)\big) \frac{r(W_n; \theta)}{r((x, \bar{u}); \theta)} I_{B \cap \{N=n\}}\bigg]$$

and, hence,

$$\begin{aligned}
&\mathbf{P}_{(x,\bar{u})}^{(\theta)}\big\{B \cap \{N < n\}\big\} \\
&\quad = \sum_{n=0}^{\infty} \mathbf{P}_{(x,\bar{u})}^{(\theta)}\big\{B \cap \{N = n\}\big\} \\
&\quad = \sum_{n=0}^{\infty} \mathbf{E}_{(x,\bar{u})}\bigg[\exp\big(\theta \log \|\mathbb{S}_n u\| - n\Lambda(\theta)\big) \frac{r(W_n; \theta)}{r((x, \bar{u}); \theta)} I_{B \cap \{N=n\}}\bigg] \\
&\quad = \mathbf{E}_{(x,\bar{u})}\bigg[\exp\big(\theta \log \|\mathbb{S}_N u\| - N\Lambda(\theta)\big) \frac{r(W_N; \theta)}{r((x, \bar{u}); \theta)} I_{B \cap \{N<\infty\}}\bigg]. \qquad \square
\end{aligned}$$

PROOF OF THEOREM 2. We will prove (2.8) first. Since $r(\cdot;\theta)$ is an eigenfunction of $\lambda(\theta)$ with respect to the operator $\mathbf{T}(\theta)$, we have

$$\mathbf{T}(\theta)r((x,\bar{u});\theta) = \lambda(\theta)r((x,\bar{u});\theta),$$

which implies that

$$\mathbf{E}_{(x,\bar{u})}\{e^{\theta\log\|\mathbb{S}_1u\|}r(W_1;\theta)\} = \lambda(\theta)r((x,\bar{u});\theta).$$

By one-term Taylor expansion of $\lambda(\theta)$ and $r((x,\bar{u});\theta)$ with respect to θ around 0, we have $\lambda(\theta) \cong 1 + \gamma\theta + o(\theta)$ and $r((x,\bar{u});\theta) \cong 1 + r'((x,\bar{u});0) + o(\theta)$. Therefore,

$$\begin{aligned}
&\mathbf{E}_{(x,\bar{u})}\{(1+\log\|\mathbb{S}_1u\|\theta+o(\theta))(1+r'(W_1;0)\theta+o(\theta))\}\\
&\quad= (1+\gamma\theta+o(\theta))(1+r'((x,\bar{u});0)\theta+o(\theta))\\
&\quad\Longrightarrow\quad \mathbf{E}_{(x,\bar{u})}\log\|\mathbb{S}_1u\| + \mathbf{E}_{(x,\bar{u})}r'(W_1;0) = \gamma + r'((x,\bar{u});0)\\
&\quad\Longrightarrow\quad (I-\mathbf{T})r'((x,\bar{u});0) = \mathbf{E}_{(x,\bar{u})}\log\|S_1u\| - \gamma.
\end{aligned}$$

Next, we want to show that

$$\sup_{(x,\bar{u})} r'((x,\bar{u});0) < \infty. \tag{A.1}$$

By Lemma 4, the Markov chain $\{(X_n, \mathbb{S}_n\cdot\bar{u}), n\ge 0\}$ on $D\times P(R^d)$ is w-uniformly ergodic, and an argument similar to Lemma 2 of Guivarch and Raugi (1986) leads to the conclusion that it is irreducible. Hence, the drift criterion of Theorem 17.4.2 of Meyn and Tweedie (1993) implies that there exists $K>0$ such that

$$r'((x,\bar{u});0) \le K(\mathbf{E}_{(x,\bar{u})}\log\|\mathbb{S}_1u\| + 1).$$

Since $\mathbf{E}_{(x,\bar{u})}\log\|\mathbb{S}_1u\| < B$ for all $x\in D$ and $\bar{u}\in P(R^d)$ by Condition A2, we have (A.1).

For the proof of (2.7), let $T(n) = \min(N,n)$. By Lemma 1 and Doob's optional stopping theorem, for all sufficiently small $|\theta|$,

$$\mathbf{E}_{(x,\bar{u})}\{e^{\theta\log\|\mathbb{S}_{T(n)}u\| - \Lambda(\theta)T(n)}r(W_{T(n)};\theta)\} = r((x,\bar{u});\theta).$$

Taking derivatives with respect to θ on both sides yields

$$\begin{aligned}
&\mathbf{E}_{(x,\bar{u})}\{(\log\|\mathbb{S}_{T(n)}u\| - \Lambda'(\theta)T(n))e^{\theta\log\|\mathbb{S}_{T(n)}u\| - \Lambda(\theta)T(n)}r(W_{T(n)};\theta)\\
&\qquad + e^{\theta\log\|\mathbb{S}_{T(n)}u\| - \Lambda(\theta)T(n)}r'(W_{T(n)};\theta)\} = r'((x,\bar{u});\theta).
\end{aligned} \tag{A.2}$$

From (A.1), we can interchange the expectation and differentiation by the dominated convergence theorem since $T(n)\le n$ and $\sup_{|\theta|\le\delta, x\in D,\bar{u}\in P(R^d)}\{|\Lambda(\theta)| + |\Lambda'(\theta)| + |r((x,\bar{u});\theta)| + |r'((x,\bar{u});\theta)| + \mathbf{E}_{(x,\bar{u})}\log\|\mathbb{S}_1u\|\} < \infty$ for sufficiently

small $\delta > 0$. Setting $\theta = 0$ in (A.2) and noting that $\gamma = \Lambda'(0)$, we obtain by integrating with respect to ν that

$$\mathbf{E}_\nu\{(\log\|\mathbb{S}_{T(n)}u\| - \Lambda'(0)T(n)) + r'(W_{T(n)};0)\} = \mathbf{E}_\nu r'(W_0;0).$$

Note that $\mathbf{E}_\nu T(n) \to \mathbf{E}_\nu N$, and that $\mathbf{E}_\nu r'(W_{T(n)};0) \to \mathbf{E}_\nu r'(W_N;0)$ as $n \to \infty$ by the dominated convergence theorem. By the monotone convergence theorem, we have $\mathbf{E}_\nu(\sum_{j=1}^{T(n)} \sigma(W_{j-1}, W_j)^+) \to \mathbf{E}_\nu(\sum_{j=1}^{N} \sigma(W_{j-1}, W_j)^+)$ and $\mathbf{E}_\nu(\sum_{j=1}^{T(n)} \sigma(W_{j-1}, W_j)^-) \to \mathbf{E}_\nu(\sum_{j=1}^{N} \sigma(W_{j-1}, W_j)^-)$ as $n \to \infty$. Hence, applying the preceding argument separately to $\sum_{j=1}^{T(n)} \sigma(W_{j-1}, W_j)^+$ and $\sum_{j=1}^{T(n)} \sigma(W_{j-1}, W_j)^-$ gives the desired conclusion. □

The next lemma provides a renewal theorem for conditional Markov random walks. Let $\tau_- = \inf\{n \geq 0; \log S_n < 0\}$, and $N = N_b = \inf\{n \geq 1 : \log S_n \notin (0, b)\}$ for $0 < b < \infty$. Under the assumption of $\mu \leq 0$, by using an argument similar to (3.11), we can define

$$\mathbf{P}_- = \mathbf{P}_\pi\{\log S_{\tau_-} \leq s, \tau_- < \infty, W_{\tau_-} \in dy\},$$

and denote π_- as its stationary distribution.

LEMMA 5. *Let $\xi_0, \xi_1, \ldots, \xi_n$ be a sequence of random variables from a hidden Markov model $\{\xi_n, n \geq 0\}$ satisfying Condition* C. *Denote ξ^- as the negative part of ξ. Assume further there exists $\varepsilon > 0$ such that* $\inf_{x,\xi} \mathbf{P}\{\log S_1 \leq -\varepsilon | W_0 = ((x,\xi), \overline{S_0\pi})\} > 0$. *Furthermore, we assume $\mu \leq 0$, $\mathbf{E}_m|\log S_1| > 0$. If $\mathbf{E}_{\pi_-} \log S_{\tau_-}^2 - < \infty$, then on $[\log S_n > b]$ as $b \to \infty$,*

$$\mathbf{E}_{(x,\bar{u})}[\log S_{\tau_-} \mid \log S_N] \to \frac{\mathbf{E}_{\pi_-} \log S_{\tau_-}^2}{2\mathbf{E}_{\pi_-} \log S_{\tau_-}}.$$

PROOF. Since $\mathbf{P}_{(x,\bar{u})}(N < \infty) = 1$, let $Z_n = \log S_{N+n}$ for $n \geq 1$. Put $V_n = \sum_{j=1}^n Z_j$ and let

$$\beta = \beta^{(1)} = \beta_1 = \inf\{n : V_n < 0\}, \qquad \beta_n = \beta^{(1)} + \cdots + \beta^{(n)},$$

where $\beta^{(2)}, \beta^{(3)}, \ldots$ are copies of β_1. Since $\{(X_n, \xi_n), n \geq 0\}$ is a stationary sequence, (β, V_β) and (τ_-, S_{τ_-}) have the same distribution.

Define $N_v = \inf\{n \geq 1; V_{\beta_n} < -\log S_N\}$ on $[\log S_N > b]$. Then $0 > \log S_N + V_{\beta_\gamma} = \log S_N + \sum_{j=1}^{\beta_\gamma} \log S_{N+j}$, and $\log S_k \geq 0$ for $k \leq N + \beta_\gamma$. Hence $\log S_{\tau_-} = \log S_N + V_{\beta_\gamma}$, and by Theorem 2, $\mathbf{E}_{(x,\bar{u})} V_{\beta_\gamma} = \mathbf{E}_m N_v \mathbf{E}_{(x,\bar{u})} V_\beta + k((x,\bar{u}))$, where $k((x,\bar{u})) = \mathbf{E}_{(x,\bar{u})}(r'(W_N;0) - r'(W_{N+\beta_\gamma};0))$, which approaches 0 as $b \to \infty$. On $\{\log S_N > b\}$,

$$\begin{aligned}\mathbf{E}_{(x,\bar{u})}(\log S_{\tau_-} \mid \log S_N \pi) &= \mathbf{E}_{(x,\bar{u})}(\log S_N + V_{\beta_r} \mid \log S_N) \\ &= \log S_N + \mathbf{E}_{(x,\bar{u})} N_v \mathbf{E}_{(x,\bar{u})} V_\beta + k((x,\bar{u})).\end{aligned}$$

By using Theorem 2.4 of Fuh and Lai (2001) along with Theorem 2 of Stone (1965) for the renewal function, we have as $b \to \infty$,

$$\mathbf{E}_{(x,\bar{u})} N_v - \frac{\log S_N}{\mathbf{E}_{\pi_-}|V_\beta|} \to \frac{\mathbf{E}_{\pi_-} V_\beta^2}{2\mathbf{E}_{\pi_-}^2 V_\beta}.$$

Hence, as $b \to \infty$,

$$\begin{aligned}
&\mathbf{E}_{(x,\bar{u})}(\log S_{\tau_-} \mid \log S_N) \\
&\quad = \log S_N + \mathbf{E}_{(x,\bar{u})} N_v \mathbf{E}_{(x,\bar{u})} V_\beta + k((x,\bar{u})) \\
&\quad = -\mathbf{E}_{(x,\bar{u})}|V_\beta| \left(\mathbf{E}_{(x,\bar{u})} N_v - \frac{\log S_N}{\mathbf{E}_{(x,\bar{u})}|V_\beta|} \right) + k((x,\bar{u})) \\
&\quad \to -\mathbf{E}_{\pi_-}|V_\beta| \left(\frac{\mathbf{E}_{\pi_-} V_\beta^2}{2\mathbf{E}_{\pi_-}^2 V_\beta} \right) = \frac{\mathbf{E}_{\pi_-} \log S_{\tau_-}^2}{2\mathbf{E}_{\pi_-} \log S_{\tau_-}}.
\end{aligned}$$

□

PROOF OF THEOREM 3. An essential part here is the verification of the uniform integrability condition for the zero drift Markov random walk, induced by the products of Markov random matrices. This can be done by a proof similar to Theorem 5 of Fuh and Lai (1998). Since

$$\mathbf{P}_{(x,\bar{u})}^{(\theta_0)}\{\log S_T \geq b\} = \mathbf{P}_{(x,\bar{u})}^{(\theta_0)}\{\tau < \infty\} - \int_{\{\log S_T \leq a\}} \mathbf{P}_{(x,\bar{u})}^{(\theta_0)}\{\tau < \infty \mid \log S_T\} \, d\mathbf{P}_{(x,\bar{u})}^{(\theta_0)},$$

we can approximate the first term via Theorem 2 of Fuh (1997). For the second term, since $\theta_1 > 0$ and $\Lambda'(\theta_1) = \mathbf{E}_{\pi(\theta_1)} \log S_1 > 0$, we have $\mathbf{P}_{(x,\bar{u})}^{(\theta_0)}\{\tau < \infty\} = 1$, and

$$\begin{aligned}
&\int_{\{\log S_T \leq a\}} \mathbf{P}_{(x,\bar{u})}^{(\theta_0)}\{\tau < \infty \mid \log S_T\} \, d\mathbf{P}_{(x,\bar{u})}^{(\theta_0)} \\
&\quad = \int_{\{\log S_T \leq a\}} \frac{r(W_\tau;\theta_0)}{r(W_0;\theta_0)} \frac{r(W_0;\theta_1)}{r(W_\tau;\theta_1)} \mathbf{E}_{(x,\bar{u})}^{(\theta_1)}\left[e^{-(\theta_1-\theta_0)\log S_\tau} \mid \log S_T \right] d\mathbf{P}_{(x,\bar{u})}^{(\theta_1)}.
\end{aligned}$$

Since $r((x,\bar{u});\theta_0)/r((x,\bar{u});\theta_1) \to 1$ as $\Delta \to 0$, therefore there exists a positive constant K such that $0 < r((x,\bar{u});\theta_0)/r((x,\bar{u});\theta_1) \leq K$ for $|\theta_1 - \theta_0|$ small enough and for all $(x,\bar{u}) \in D \times P(R^d)$. Hence it is easy to check that $g((x,\bar{u}),t) = \exp(-(\theta_1 - \theta_0)t)\frac{r((x,\bar{u});\theta_0)}{r((x,\bar{u});\theta_1)}$ is directly Riemann integrable. Therefore by the Markov renewal theorem in Theorem 2.4 of Fuh and Lai (2001), Theorem 2 of Fuh (1997) and a similar argument to that in Siegmund (1979), we have

$$\begin{aligned}
&\int_{\{\log S_T \leq a\}} \mathbf{P}_{(x,\bar{u})}^{(\theta_0)}\{\tau < \infty \mid \log S_T\} \, d\mathbf{P}_{(x,\bar{u})}^{(\theta_0)} \\
&\quad = \exp[-\Delta(b + \rho_+ - c_1)]\mathbf{P}_{(x,\bar{u})}^{(\theta_1)}\{\log S_T \leq a\} + o(\Delta).
\end{aligned}$$

On the other hand,

$$\mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\log S_T \le a\}$$
$$= \mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\tau^* < \infty\} - \int_{\{\log S_T \pi \ge b\}} \mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\tau^* < \infty \mid \log S_T\}\, d\mathbf{P}^{(\theta_1)}_{(x,\bar{u})},$$

where $\tau^* = \inf\{n : \log S_n < a\}$. By Taylor expansion for $\mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\tau^* < \infty \mid \log S_T\}$ on the set $\{\log S_T \ge b\}$, we have

$$\mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\tau^* < \infty \mid \log S_T\}$$
$$= e^{\Delta a}\mathbf{E}^{(\theta_0)}_{(x,\bar{u})}\big[1 + \Delta(\log S_{\tau^*} - a)$$
$$- \Delta\big(r'(W_0; 0) - r'(W_{\tau^*}; 0)\big) + o(\Delta) \mid \log S_T\big].$$

By Lemma 5, we have

$$\int_{\{\log S_T \ge b\}} \mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\tau^* < \infty \mid \log S_T\}\, d\mathbf{P}^{(\theta_1)}_{(x,\bar{u})}$$
$$= \exp[\Delta(a + \rho_- - c_-)]\mathbf{P}^{(\theta_0)}_{(x,\bar{u})}\{\log S_T \ge b\} + o(\Delta).$$

Hence, there are two relations for $\mathbf{P}^{(\theta_0)}_{(x,\bar{u})}\{\log S_T \ge b\}$ and $\mathbf{P}^{(\theta_1)}_{(x,\bar{u})}\{\log S_T \le a\}$. Solve the equations and take expectation under stationary distribution to prove the theorem. □

Acknowledgments. The author is grateful to the two referees and the Associate Editor for constructive comments and suggestions. He also thanks Professors Larry Shepp and John Marden, the Editor, for careful proofreading.

REFERENCES

ALSMEYER, G. (1994). On the Markov renewal theorem. *Stochastic Process. Appl.* **50** 37–56.

ALSMEYER, G. (2000). The ladder variables of a Markov random walk. *Probab. Math. Statist.* **20** 151–168.

ASMUSSEN, S. (1989). Risk theory in a Markov environment. *Scand. Actuar. J.* **1989**(2) 69–100.

BALL, F. and RICE, J. A. (1992). Stochastic models for ion channels: Introduction and bibliography. *Math. Biosci.* **112** 189–206.

BANSAL, R. K. and PAPANTONI-KAZAKOS, P. (1986). An algorithm for detecting a change in a stochastic process. *IEEE Trans. Inform. Theory* **32** 227–235.

BASSEVILLE, M. and NIKIFOROV, I. V. (1993). *Detection of Abrupt Changes*: *Theory and Application*. Prentice-Hall, Englewood Cliffs, NJ.

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 systèmes linéaires à coefficients markoviens. *Probab. Theory Related Fields* **78** 193–221.

BOUGEROL, P. and LACROIX, J. (1985). *Products of Random Matrices with Applications to Schrödinger Operators.* Birkhäuser, Boston.

CHURCHILL, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51** 79–94.

COGBURN, R. (1980). Markov chains in random environments: The case of Markovian environments. *Ann. Probab.* **8** 908–916.

CSISZÁR, I. and NARAYAN, P. (1988). Arbitrarily varying channels with constrained inputs and states. *IEEE Trans. Inform. Theory* **34** 27–34.

ELLIOTT, R., AGGOUN, L. and MOORE, J. (1995). *Hidden Markov Models*: *Estimation and Control.* Springer, New York.

ENGEL, C. and HAMILTON, J. D. (1990). Long swings in the dollar: Are they in the data and do markets know it? *American Economic Review* **80** 689–713.

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** 987–1008.

FUH, C. D. (1997). Corrected diffusion approximations for ruin probabilities in a Markov random walk. *Adv. in Appl. Probab.* **29** 695–712.

FUH, C. D. (1998). Efficient likelihood estimation of hidden Markov models. Technical report, Institute of Statistical Science, Taipei, Taiwan, ROC.

FUH, C. D. and LAI, T. L. (1998). Wald's equations, first passage times and moments of ladder variables in Markov random walks. *J. Appl. Probab.* **35** 566–580.

FUH, C. D. and LAI, T. L. (2001). Asymptotic expansions in multidimensional Markov renewal theory and first passage times for Markov random walks. *Adv. in Appl. Prob.* **33** 652–673.

FUH, C. D. and ZHANG, C.-H. (2000). Poisson equation, maximal inequalities and r-quick convergence for Markov random walks. *Stochastic Process. Appl.* **87** 53–67.

GOLDFELD, S. M. and QUANDT, R. E. (1973). A Markov model for switching regressions. *J. Econometrics* **1** 3–16.

GUIVARCH, Y. and RAUGI, A. (1986). Products of random matrices: Convergence theorems. In *Random Matrices and Their Applications* (J. E. Cohen, H. Kesten and C. M. Newman, eds.) 31–54. Amer. Math. Soc., Providence, RI.

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 Univ. Press.

HAMILTON, J. D. (1996). Specification testing in Markov-switching time series models. *J. Econometrics* **70** 127–157.

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

ITÔ, 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.

JENSEN, J. L. (1987). A note on asymptotic expansions for Markov chains using operator theory. *Adv. in Appl. Math.* **8** 377–392.

JUANG, B.-H. and RABINER, L. R. (1985). A probabilistic distance measure for hidden Markov models. *AT&T Tech. J.* **64** 391–408.

KESTEN, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta Math.* **131** 207–248.

KESTEN, H. (1974). Renewal theory for functionals of a Markov chain with general state space. *Ann. Probab.* **2** 355–386.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. and HAUSSLER, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Molecular Biology* **235** 1501–1531.

LAI, T. L. (1995). Sequential change point detection in quality control and dynamical systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 613–658.

LAI, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inform. Theory* **44** 2917–2929.

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

LIU, C. C. and NARAYAN, P. (1994). Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures. *IEEE Trans. Inform. Theory* **40** 1167–1180.

LIU, J. S., NEUWALD, A. F. and LAWRENCE, C. E. (1999). Markovian structures in biological sequence alignments. *J. Amer. Statist. Assoc.* **94** 1–15.

LORDEN, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **41** 1897–1908.

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, New York.

MOUSTAKIDES, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.* **14** 1379–1387.

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

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–114.

RABINER, L. R. and JUANG, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ.

RIESZ, F. and SZ.-NAGY, B. (1955). *Functional Analysis.* Ungar, New York.

RITOV, Y. (1990). Decision theoretic optimality of the CUSUM procedure. *Ann. Statist.* **18** 1464–1469.

SADOWSKY, J. S. (1989). A dependent data extension of Wald's identity and its application to sequential test performance computation. *IEEE Trans. Inform. Theory* **35** 834–842.

SIEGMUND, D. (1979). Corrected diffusion approximations in certain random walk problems. *Adv. in Appl. Probab.* **11** 701–719.

SIEGMUND, D. (1985). *Sequential Analysis. Tests and Confidence Intervals*. Springer, New York.

STONE, C. (1965). On characteristic functions and renewal theory. *Trans. Amer. Math. Soc.* **120** 327–342.

WALD, A. and WOLFOWITZ, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19** 326–339.

WOODROOFE, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis.* SIAM, Philadelphia.

YAKIR, B. (1994). Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space. In *Change-Point Problems* (E. Carlstein, H. G. Müller and D. Siegmund, eds.) 346–358. IMS, Hayward, CA.

ZIV, J. (1985). Universal decoding for finite-state channels. *IEEE Trans. Inform. Theory* **31** 453–460.

INSTITUTE OF STATISTICAL SCIENCE
ACADEMIA SINICA
TAIPEI, 11529
TAIWAN, R. O. C.
E-MAIL: stcheng@stat.sinica.edu.tw