Engineering Applications of Artificial Intelligence 81 (2019) 234–246

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Sequential hypothesis tests for streaming data via symbolic time-series analysis☆,☆☆

Nurali Virani [a], Devesh K. Jha [b], Asok Ray [c,*], Shashi Phoha [d]

[a] GE Global Research, Niskayuna, NY 12309, USA
[b] Mitsubishi Electric Research Labs, Cambridge, MA 02139, USA
[c] Department of Mechanical Engineering and Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA
[d] Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Keywords:
Sequential hypothesis testing
Symbolic dynamics
Markov modeling
Combustion instability

ABSTRACT

This paper addresses sequential hypothesis testing for Markov models of time-series data by using the concepts of symbolic dynamics. These models are inferred by discretizing the measurement space of a dynamical system, where the system dynamics are approximated as a finite-memory Markov chain on the discrete state space. The study is motivated by time-critical detection problems in physical processes, where a temporal model is trained to make fast and reliable decisions with streaming data. Sequential update rules have been constructed for log-posterior ratio statistic of Markov models in the setting of binary hypothesis testing and the stochastic evolution of this statistic is analyzed. The proposed technique allows selection of a lower bound on the performance of the detector and guarantees that the test will terminate in finite time. The underlying algorithms are first illustrated through an example by numerical simulation, and are subsequently validated on time-series data of pressure oscillations from a laboratory-scale swirl-stabilized combustor apparatus to detect the onset of thermo-acoustic instability. The performance of the proposed sequential hypothesis tests for Markov models has been compared with that of a maximum-likelihood classifier with fixed sample size (i.e., sequence length). It is shown that the proposed method yields reliable detection of combustion instabilities with fewer observations in comparison to a fixed-sample-size test.

1. Introduction

Recently there has been an unprecedented increase in the volume and speed of temporal data being generated by physical systems, due to the improvements in low-cost sensing and high-speed computation & communication. Typical machine learning tools, used for monitoring of physical processes, extract features from a fixed number of consecutive observations and pose a supervised learning problem for detection or classification using those features (Bishop, 2006). Similarly, fixed-sample-size (FSS) tests from the estimation theory literature (Poor, 1994) compute likelihood or posterior probability of an observed sequence of fixed length and select the class which maximizes this statistic. However, in order to perform well on all feasible sequences, the chosen sample length in these approaches is made longer than needed for most of the easily separable cases. For example, in time-critical online monitoring systems such as detection of combustion instabilities in aircraft gas-turbine engines, an early detection can enhance mitigation of structural damage in engines by avoiding the thermo-acoustic resonance and may even prevent accidents due to engine shutdown. A dynamic data-driven approach (Darema, 2004) for sequential detection and classification is needed, which can adapt the observed length of time series without any significant computational burden. Sequential hypothesis tests offer one such efficient and adaptive framework, which would allow online detection of anomalies, faults, and mode transitions in dynamical systems, where high-speed streaming data are generated. This paper presents a sequential hypothesis testing procedure for a class of Markov models of temporal data inferred by using symbolic time-series analysis (STSA) (Daw et al., 2003; Beim Graben, 2001).

A sequential detector is a statistical decision function that uses a random number of samples depending on the observation sequence to

☆ The work reported in this paper has been supported in part by U.S. Air Force Office of Scientific Research (AFOSR) under Grant Nos. FA9550-15-1-0400 and FA9550-18-1-0135 in the area of dynamic data-driven application systems (DDDAS). Any opinions, findings and conclusions in this paper are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

☆☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.engappai.2019.02.015.

* Corresponding author.
E-mail addresses: nurali.virani@ge.com (N. Virani), devesh.dkj@gmail.com (D.K. Jha), axr2@psu.edu (A. Ray), sxp26@arl.psu.edu (S. Phoha).

https://doi.org/10.1016/j.engappai.2019.02.015
Received 23 May 2018; Received in revised form 31 October 2018; Accepted 23 February 2019
Available online xxxx

detect the underlying hypothesis. A sequential detector, on the average, would need a much smaller sequence length than FSS tests (Poor, 1994). Sequential probability ratio test (SPRT) (Wald and Wolfowitz, 1948; Shiryaev, 1978) has been traditionally used for binary hypothesis testing and is known to be an optimal detector when the observation sequences are independent and identically distributed (IID) (Burkholder and Wijsman, 1963). However, sequential data from physical systems are typically not independent as causal relations, because of the underlying physics, may lead to statistical dependencies.

Hidden Markov models (HMMs) (Rabiner, 1989) undergo temporal evolution, which are learned from data using iterative techniques to capture some of the statistical dependencies (Bishop, 2006). In Grossi and Lops (2008), Chen and Willett (2000), Fuh (2003), the concept of SPRT was extended for data generated from a HMM. In contrast, symbolic time-series analysis-based Markov modeling makes use of concepts from symbolic dynamics to represent an observation sequence as a discrete Markov model. This technique consists of two critical steps, namely, discretization (also called partitioning) of the measurement space of a dynamical system (Kennel and Buhl, 2003; Virani et al., 2016) and memory-size estimation (Srivastav, 2014; Jha et al., 2015). The time-series data are then approximated as a D-Markov model, which is a Markov chain, with finite memory (or order which is denoted as D), over the discrete state-space of the system (Ray, 2004; Mukherjee and Ray, 2014); however, the D-Markov model is different from a hidden Markov model (HMM) in the sense that the state-space of a D-Markov model is always observable. The structure of a D-Markov model can be inferred by searching for the optimal discretization and the corresponding order of the Markov chain. Thus, the parameters associated with the model structure are: the size of partitioning set (also called alphabet in symbolic dynamics literature), location of partitioning segments, and the memory of the associated Markov chain. Once the memory of the discrete symbol sequence is estimated, the state-space of the Markov model is represented by the corresponding collection of memory words (or collections of symbols of length equal to the estimated memory for the discrete sequence). This paper presents a technique to sequentially estimate the log-posterior ratio (LPR) for a binary hypothesis test, where each hypothesis is represented by a D-Markov model.

Contributions: This paper extends the results of classical SPRT for IID observations to sequential hypothesis tests for observations from D-Markov models of time-series data. D-Markov modeling for time-series data via STSA is an existing technique (Ray, 2004) with applications in target detection and classification in surveillance (Virani et al., 2013; Mukherjee et al., 2011), prognostics and health monitoring of physical systems (e.g., nuclear plants (Jin et al., 2011) and electronic products (Kumar and Pecht, 2007)), and others. The novel contribution of this work is to enable the use of D-Markov models for streaming data analysis in detection and classification problems by formulating their real-time sequential estimation and hypothesis testing problem in a Bayesian framework. In this paper, the probability density of the D-Markov model is represented as a product of categorical distributions; and the parameters of this distribution themselves have a Dirichlet distribution. Then, using the fact that the Dirichlet distribution is the conjugate prior of the categorical distribution, a sequential update rule is developed for posterior probability ratios of D-Markov models to test a time-series model against an alternate time-series model for binary classification problems. Expected increment of the log-posterior ratios are explicitly provided under each hypothesis; and it is shown that the sequential tests for the Markov models terminate in finite time with probability one. The efficacy of the proposed method for online monitoring with streaming data is first illustrated with numerical simulation and then validated on experimental data from a lean-premixed swirl-stabilized combustor apparatus (Kim et al., 2010). The performance of the sequential tests is also compared with that of a (maximum likelihood classifier) FSS test to demonstrate that the proposed method is capable of making more accurate decisions using fewer observations.

Organization: The paper is organized in seven sections including the present one. Section 2 provides mathematical background for the rest of paper. Section 3 presents the problem formulation and explains the difference with respect to the SPRT for IID sequences. Section 4 explains the proposed approach to address the sequential hypothesis testing problem and shows the theoretical justification for the proposed technique. Section 5 presents the results and inferences from a simulation example to explain the underlying algorithms. Section 6 shows the validation results based on experimental data of combustion instability along with a description of the test apparatus. Section 7 concludes the paper along with recommendations for future research.

2. Mathematical preliminaries

This section introduces mathematical concepts that are used throughout the paper. Symbolic time-series analysis (STSA) (Beim Graben, 2001; Daw et al., 2003) partitions the measurement space of a dynamical system, where the partitioning is a mapping from a continuous space of measurements to a discrete space of symbols. In machine learning literature, discretization is generally studied as a feature extraction technique. In mathematics literature, the partitioning or discretization is characterized by the extent to which a dynamical system can be represented by a symbolic one. The data are symbolized based on the choice of a partitioning technique and then, the dynamics of the discrete process are studied. Upon symbolization of statistically stationary (or quasi-stationary) time-series data, the resulting symbol sequences are converted to probabilistic finite state automata (PFSA) for information compression. While the details are reported in Ray (2004), Mukherjee and Ray (2014), the essential information is presented below for completeness of the paper.

Definition 2.1 (*DFSA*)**.** A deterministic finite state automaton (DFSA) is a 3-tuple $\mathcal{G} = (\Sigma, Q, \delta)$ where:

- Σ is a non-empty finite set, called the alphabet, with cardinality $|\Sigma|$;
- Q is a non-empty finite set, called the set of states, with cardinality $|Q|$;
- $\delta : Q \times \Sigma \to Q$ is the state transition map.

It is noted that $\Sigma^\star$ is a countable collection of all finite-length strings with symbols from the alphabet Σ and the (zero-length) empty string ϵ.

Definition 2.2 (*PFSA*)**.** A probabilistic finite state automaton (PFSA) is constructed upon a DFSA $\mathcal{G} = (\Sigma, Q, \delta)$ as a pair $\mathcal{M} = (\mathcal{G}, M)$, that is, the PFSA $\mathcal{M}$ is a 4-tuple $\mathcal{M} = (\Sigma, Q, \delta, M)$, where:

- Σ, Q, and δ are the same as in Definition 2.1;
- $M : Q \times \Sigma \to [0, 1]$ is the morph function that satisfies the condition $\sum_{s\in\Sigma} m(q, s) = 1 \quad \forall q \in Q$. In matrix form M is called the morph matrix and its entries m_{ij} denotes the probability of emitting a symbol $s_j \in \Sigma$ from the state $q_i \in Q$.

A PFSA is viewed as a generative model capable of probabilistic generation of a symbol string through state transitions; it has found many applications ranging from pattern recognition and machine learning to computational linguistics (Vidal et al., 2005). For symbolic analysis of time-series data, a class of PFSAs called the D-Markov machine has been proposed in Ray (2004) as a sub-optimal but computationally efficient approach for encoding the dynamics of symbol sequences as a finite state automaton. The main assumption, which is the reason for sub-optimality, is that the symbolic process can be approximated as a D^{th}-order Markov chain. The assumption of finite memory (or order) is reasonable for stable and controlled engineering systems that tend to eventually forget their initial conditions. The states of this PFSA are words of length D (or less) over an alphabet Σ; and state transitions are described by a sliding block code of memory D and anticipation length of one or higher (Lind and Marcus, 1995).

Definition 2.3 (*D-Markov Machine (Ray, 2004; Mukherjee and Ray, 2014)*). A D-Markov machine is a statistically stationary stochastic process $S = \cdots s_{-1}s_0s_1\cdots$, where the probability of occurrence of a new symbol depends only on the last D symbols, that is,

$$P(s_n \mid \cdots s_{n-D}\cdots s_{n-1}) = P(s_n \mid s_{n-D}\cdots s_{n-1}) \tag{1}$$

where D is called the depth of the Markov machine.

A D-Markov machine is thus a D^{th}-order Markov approximation of the discrete symbolic process. It is also noted that the presented formalism for statistical learning is different from the standard hidden Markov models in the following sense:

- The underlying algebraic structure of a PFSA (and hence that of a D-Markov machine) is restricted to be deterministic, whereas a hidden Markov machine is not subjected to this restriction (Vidal et al., 2005).
- The state space of these models is inferred by partitioning the observed measurement space of the dynamical system, while in HMM, the state space may never be observed and thus it is not straight-forward to infer the underlying structure of the model. It is noted that, in the considered class of models, the states as well as transitions between the states are observed, in contrast to HMMs where only emissions are observed.
- Under the assumption that the observed sequence is a finite-order Markov chain, the estimation of parameters is simplified. The sufficient statistics could be obtained by estimating the order of the Markov chain and the symbol emission probabilities are conditioned on the memory words of the discrete sequence.

Although the framework of a D-Markov machine is restrictive due to the constraint of (finitely many) observable states, it can be conveniently used for online in-situ monitoring applications due to the simplicity of the inference algorithms. For finite-order and finite-state Markov chains, the conditional symbol emission probabilities and the initial state summarize all of the relevant information supplied by an appropriate sample (i.e., time series) (Laurence, 1993). Under stationarity assumptions, the initial state becomes unnecessary and thus, the sufficient statistic is provided by the emission probabilities given by the morph matrix.

In information theory, Kullback–Leibler divergence (or K–L divergence) between two probability mass functions (Cover and Thomas, 1991) is defined as:

$$\mathrm{d}_{\mathrm{KL}}\big(p(q) \parallel \bar{p}(q)\big) = \sum_{q\in Q} p(q)\log\left(\frac{p(q)}{\bar{p}(q)}\right).$$

In this paper, a measure of difference between two morph matrices, which are conditional distributions, is obtained in terms of the conditional relative entropy (Cover and Thomas, 1991) as:

$$\mathrm{d}\big(p(s|q) \parallel \bar{p}(s|q)\big) = \sum_{q\in Q} p(q)\sum_{s\in\Sigma} p(s|q)\log\left(\frac{p(s|q)}{\bar{p}(s|q)}\right). \tag{2}$$

3. Problem formulation

Sequential detectors for independent and identically distributed (IID) observations to infer which of the known probability densities is generating the data have been well-studied in literature (Poor, 1994). This section first addresses the sequential detection problem for discrete-valued IID observations and then extends it to the case in which the observation sequence is generated by a D-Markov model.

3.1. Sequential detection: IID Case

The discrete-valued IID observations $\{S_t;\ t = 1, 2, \dots\}$ are tested according to the following binary hypotheses:

$$\mathcal{H}_0 : S_t \sim P_0,\ t = 1, 2, \dots$$

versus

$$\mathcal{H}_1 : S_t \sim P_1,\ t = 1, 2, \dots,$$

where P_0 and P_1 are two known probability distributions (or measures) on the measurable space $(\Sigma, 2^{\Sigma})$. Here, Σ is a finite set of symbols, each symbol $s \in \Sigma$ is a discrete observation, and 2^{Σ} denotes the power set of Σ, which consists of all possible subsets of Σ including the empty set and the set Σ itself.

A sequential decision rule is formulated as a pair of (time-dependent) sequences $(\bar{\phi}, \bar{\mu})$. In this setting, $\bar{\phi} \triangleq \{\phi_j : j = 0, 1, 2, \dots\}$ is called a stopping rule with $\phi_j : \Sigma^j \to \{0, 1\}$ and $\bar{\mu} \triangleq \{\mu_j : j = 0, 1, 2, \dots\}$ is called a terminal decision rule with $\mu_j : \Sigma^j \to \mathcal{X}$; for binary hypothesis (i.e., either $\mathcal{H}_0$ or $\mathcal{H}_1$) testing, $\mathcal{X}$ is represented as $\{0, 1\}$. For an observed symbol string of finite length $S_t = \{s_i \in \Sigma : i = 1, 2, \dots, t\}$ at time t, either the stopping rule is executed if $\phi_t(S_t) = 1$) to make a decision, or sampling is continued if $\phi_t(S_t) = 0$) to observe the next symbol s_{t+1}. If the decision is to stop sampling (i.e., $\phi_t(S_t) = 1$), then the terminal decision rule selects the hypothesis for S_t, which yields the decision $\mu_t(S_t) \in \mathcal{X}$.

3.2. Sequential detection: D-Markov case

In the hypothesis testing problem under consideration, the observations may not necessarily be IID as they are often generated from a physical process. Therefore, a D-Markov model is assigned corresponding to each hypothesis, instead of a probability distribution. As stated earlier in Section 2, these D-Markov models are represented as $\mathcal{M}_k = (Q, \Sigma, \delta, M_k)$ for each hypothesis $\mathcal{H}_k$ with $k \in \mathcal{X}$, where Q represents the set of states, Σ is the alphabet set, $\delta : Q \times \Sigma \to Q$ is the transition function, and M_k is the class-specific morph matrix, which consists of entries $[M_k]_{ij} = m_{ij}^k = p_k(s_j|q_i)$. Each row of the matrix M_k represents the probability of emitting a symbol $s_j \in \Sigma$ from a particular state $q_i \in Q$ in the model $k \in \mathcal{X}$. Thus, each row of the morph matrix is a categorical distribution (Papoulis and Pillai, 2002).

It is noted that all of these models have a similar algebraic structure and they differ only in the symbol emission probabilities. It is assumed that the training instances for nominal and anomalous conditions have been obtained as sufficiently long symbol sequences, which allows the convergence of the morph matrix parameters in the training phase (Meyn and Tweedie, 2012). However, in the operation phase, it is necessary to make accurate decisions with (possibly) fewer streaming observations, because early detection (or prediction using precursors) of anomalies/faults and instabilities are crucial for taking corrective actions in the physical process.

Let a Markov model over the state-symbol pair, which generates the symbol sequence, be given as:

$$\begin{aligned} p(q_{t+1}, s_{t+1} \mid q_t, s_t) &= p(s_{t+1} \mid q_t, s_t, q_{t+1})\, p(q_{t+1} \mid q_t, s_t) \\ &= p(s_{t+1} \mid q_{t+1})\, p(q_{t+1} \mid q_t, s_t) \end{aligned} \tag{3}$$

where $p(q_{t+1} \mid q_t, s_t)$ is obtained from the transition function δ as:

$$p(q_{t+1} \mid q_t, s_t) = \begin{cases} 1 & \text{if } \delta(q_t, s_t) = q_{t+1} \\ 0 & \text{otherwise} \end{cases}$$

In Eq. (3), $p(s_{t+1} \mid q_{t+1})$ is obtained from the morph matrix M. Since, the state transition is deterministic (see Definition 2.1), the emission of a symbol sequence from hypothesis $\mathcal{H}_k$ is governed only by the morph matrix $M_k[i, j] = [p_k(s_{t+1} = j \mid q_{t+1} = i)]$. Thus, the hypothesis test is given as follows:

The discrete-valued observation sequence $S_t = \{s_i \in \Sigma : i = 1, 2, \dots, t\}$ is generated according to

$$\mathcal{H}_0 : S_t \sim M_0,\ t = 1, 2, \dots$$

versus

$$\mathcal{H}_1 : S_t \sim M_1,\ t = 1, 2, \dots,$$

where M_k is the morph matrix of the PFSA model $\mathcal{M}_k$ for $k \in \{0, 1\}$. It is noted that, in a D-Markov model, each state is represented by a finite

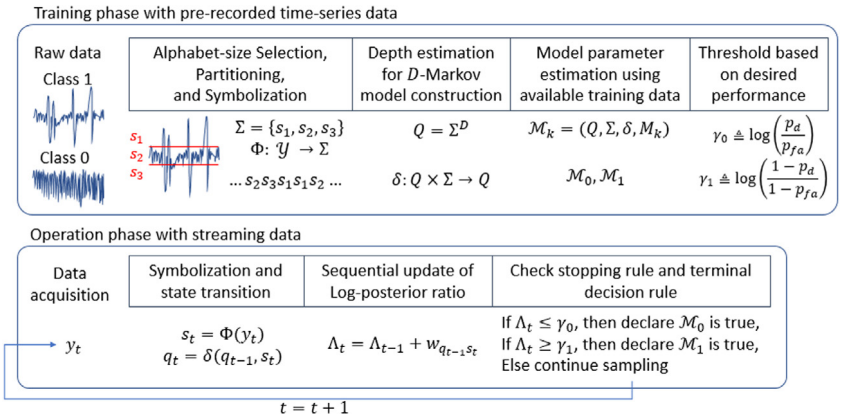


Fig. 1. Proposed scheme for training and during operation.

history of at most D symbols, thus after observing the first at most D symbols, the initial state will be known. Hence, the sequential detector is initiated after at most D symbols have been observed. The sequential detector in this case is also defined as the pair of sequences of stopping rule and terminal decision rule as discussed in Section 3.1.

3.3. Problem statement

In this paper, a sequential detector is constructed to identify the D-Markov model that generates symbol sequences, as shown in the hypothesis test in Section 3.2. Given these models, the problem is to execute the following tasks:

(1) To construct a sequential detector to guarantee a desired probability, p_d, of successful detection as well as the probability, p_{fa}, of false alarms;
(2) To prove that the sequential detector terminates with a decision in finite time with probability one and to obtain an estimate of the stopping time $\tau(S_t) \triangleq \min\{t : \phi_t(S_t) = 1\}$.
(3) To elucidate the developed sequential detector with numerical simulation.
(4) To validate the proposed sequential testing procedure for fault detection on an experimental apparatus.

4. Technical approach

This section first summarizes a time-series modeling method by using symbolic analysis along with the process for hyper-parameter estimation and a Bayesian approach for parameter estimation of these time-series models. The posterior distribution of the PFSA model $\mathcal{M}_k$, given the observed symbol sequence S_t, is then computed. Subsequently, the sequential decision rule is developed for a binary hypothesis testing problem. A summary of the proposed approach during training and during operation with streaming data is also provided. Some analysis is shown to study the evolution of log-posterior ratio (LPR) statistic and to obtain estimates of stopping time and performance bounds.

4.1. Time-series modeling and hyper-parameter estimation

Symbolic time-series analysis (STSA) has been adopted as a tool for time-series modeling, which consists of two critical steps: (i) *discretization*, where the continuous attributes of the sequential data are projected onto a symbolic space, which is followed by (ii) *identification of concise probabilistic patterns* to compress the information embedded in the discretized symbol sequences. Specifically, a finite-memory Markov model is built upon the memory estimate of the symbol sequence, which is represented by a state transition matrix. This leads to identification of the causal dynamical structure, which is intrinsic to the symbolic process under consideration; such a dynamical structure is suitable for feature extraction and pattern classification.

The hyper-parameters for model inference under the current class of Markov models include: (i) alphabet size and location of partition boundaries for discretization (i.e., symbolization) of the continuous real-valued signals, and (ii) memory of the corresponding discrete data. Symbolization is carried out via partitioning of the phase-space of the system within which the system dynamics evolves. In general, there are two main lines of thought behind the symbolization process, one inspired by dynamical systems theory and the other inspired by machine learning objectives. A brief formal introduction of discretization is presented next.

Let the time-series data be denoted as the sequence $\{y_t\}_{t\in\mathbb{N}}$ where $y_t \in \mathcal{Y} \subseteq \mathbb{R}$ and let $\boldsymbol{\Phi}$ represent the partitioning function such that $\boldsymbol{\Phi} : y_t \mapsto s_t$ where $s_t \in \Sigma$ for all $t \in \mathbb{N}$ and the alphabet size $|\Sigma| \in \mathbb{N}$ is known and fixed (see Fig. 1). Then the partitioning function, which is determined by the set $\mathcal{R}_{\boldsymbol{\Phi}} = \{R_1, \ldots, R_{|\Sigma|}\}$, has mutually exclusive and exhaustive segments $\overline{R_i}$, $i = 1, \ldots, |\Sigma|$ (i.e., $\mathcal{Y} = \overline{R_1} \cup \cdots \cup \overline{R_{|\Sigma|}}$ and $\overline{R_i} \cap \overline{R_j} = \emptyset \ \forall i \neq j$). The dynamics of the discrete system is governed by the function $\boldsymbol{\Phi}$ and thus, to find a useful discrete stochastic system, it is important to find a good partitioning function. Some approaches inspired by the dynamical systems theory could be found in Buhl and Kennel (2005), Adler (1998), Kennel and Buhl (2003). Several other partitioning methods have been proposed in literature such as maximum entropy partition (MEP) (Rajagopalan and Ray, 2006), symbolic aggregate approximation (SAX) (Lin et al., 2007), maximally-bijective partition (Sarkar et al., 2013), and sparse density estimation-based partition (Virani et al., 2016). In general, most of the machine learning applications use a cross-validation-based approach to select the partitioning size and boundaries. This paper has used (unsupervised) maximum entropy partitioning (MEP) (Rajagopalan and Ray, 2006), where approximately equal number of points are assigned to each of the partitioning segments; thus, it leads to a unique solution for a one-dimensional time-series. For its simplicity, MEP is the most widely used partitioning technique. For a detailed discussion on partitioning techniques, interested readers are referred to Jha (2016).

Working in the symbolic domain, the task is to identify concise probabilistic models that relate the past, present and the future states of the system under consideration. For Markov modeling of the symbol sequence, this is achieved by first estimating the depth (or size of memory) for the discrete symbolic process and then, estimating the approximate stochastic model from the observed sequence. Various approaches have been reported in literature for order estimation of Markov chains (Jha, 2016; Jha et al., 2018). For machine learning applications, the estimation process follows the *wrapper* approach (Bishop, 2006), where a search algorithm with a certain stopping criterion calls the main modeling module to build several temporal models with varying depths; and the search is stopped when the stopping criterion (e.g., information gain or entropy rate) show marginal

improvement for the added complexity. This paper does not address optimization of the depth of the underlying models — it is assumed to be known. Interested readers are referred to Jha et al. (2015, 2018) for a detailed discussion on estimation of depth for D-Markov models. For completeness of the paper, a recent and computation-efficient algorithm based on spectral properties of the Markov model has been provided in Appendix.

The work presented in this paper is independent of the choice of hyper-parameters of the model; in other words, the presented results hold true for all choices of hyper-parameters. Once these hyper-parameters are estimated, the time-series data are represented by a stochastic matrix of the inferred Markov chain. The stochastic matrix is estimated in a sequential fashion by a Bayesian approach with conjugate priors as explained in the next section.

4.2. Bayesian approach for parameter estimation

The D-Markov model for each hypothesis is represented by a morph matrix, as discussed in Section 2. Each row of the morph matrix of a D-Markov model is a discrete probability mass function denoting the probability of emission of a symbol from a given state. Given the states, the rows of a morph matrix represent a set of independent categorical distributions (Papoulis and Pillai, 2002) over the alphabet set. It is known that the Dirichlet distribution is the conjugate prior of the categorical distribution (Bishop, 2006), i.e., if the prior density over parameters of the categorical distribution is represented as a Dirichlet distribution, then the posterior density after a Bayesian update with a new observation follows a Dirichlet distribution too. Thus, for a given state, the density over parameters of a particular row of the morph matrix is given by Dirichlet distribution (Wen et al., 2013).

For any $t \in \mathbb{N}$, let S_t denote the symbol sequence $(s_1, s_2, \dots, s_t)$ of length t. The posterior distribution of the model parameters of $\mathcal{M}_k$, given an observed symbol sequence S_t, is computed as:

$$P(\mathcal{M}_k \mid S_t) = \prod_{i=1}^{|Q|} \left(\frac{\prod_{j=1}^{|\Sigma|} (p_{ij}^k)^{\alpha_{ij}^t - 1}}{\mathrm{B}(\bar{\alpha}_i^t)} \right) \tag{4}$$

where the hyper-parameter α_{ij}^t is initialized at α_{ij}^0; the value of $(\alpha_{ij}^t - \alpha_{ij}^0)$ is the count of occurrences of the symbol s_j at state q_i in a symbol sequence S_t; the vector $\bar{\alpha}_i^t$ represents $[\alpha_{i1}^t, \alpha_{i2}^t, \dots, \alpha_{i|\Sigma|}^t]$ for all $i \in \{1, 2, \dots, |Q|\}$; the scalar $\mathrm{B}(\bar{\alpha}_i^t) = \frac{\prod_{j=1}^{|\Sigma|} \Gamma(\alpha_{ij}^t)}{\Gamma(\sum_{j=1}^{|\Sigma|} \alpha_{ij}^t)}$; and $\Gamma(\cdot)$ is the standard Gamma function. Using the posterior distribution and given S_t, the expected value of the morph map for the PFSA $\mathcal{M}$ is:

$$m^t(q_i, s_j) = m_{ij}^t \triangleq \frac{\alpha_{ij}^t}{\sum_{\ell=1}^{|\Sigma|} \alpha_{i\ell}^t} = \frac{\alpha_{ij}^0 + n_{ij}^t}{\sum_{\ell=1}^{|\Sigma|} \alpha_{i\ell}^0 + \sum_{\ell=1}^{|\Sigma|} n_{i\ell}^t} \tag{5}$$

where n_{ij}^t is the count of occurrence of symbol s_j at state q_i in a sequence S_t; and α_{ij}^0 is the initial value of the hyper-parameter α_{ij}^t of the Dirichlet distribution. During parameter estimation in training phase, the following observations are made:

- The uniform prior is obtained by usage of Eq. (5) to yield $m_{ij}^0 = \frac{1}{|\Sigma|}$ by setting $n_{ij}^0 = 0$ and $\alpha_{ij}^0 = 1$ for all i and j. In that case, the morph matrix is given by:

$$m_{ij}^t \triangleq \frac{1 + n_{ij}^t}{|\Sigma| + \sum_{\ell=1}^{|\Sigma|} n_{i\ell}^t} \tag{6}$$

  In the absence of any prior information, Eq. (6) is used as the estimate of the model parameter from the training data.
- If no symbol is generated from a specific state $q \in Q$ for the entire symbol sequence, then there should be no preference to any particular symbol, for which a uniform prior distribution is assumed for symbol emission from that state q.
- This procedure guarantees that the PFSA constructed from a finite-length symbol string must have an (element-wise) strictly positive morph map M. This property will be important in construction of the sequential update rule.

This parameter estimation process is repeated to obtain a model for each of the hypotheses. One may also create reduced-order Markov models with fewer states for each hypothesis which can lead to faster training and potentially quicker termination of sequential test by using the tools prescribed in Jha et al. (2018). In the next part, these models are used to obtain a test statistic and their sequential update rules for use in the testing phase.

4.3. Initialization and sequential update during testing

In binary hypothesis testing, such as Bayes or Neyman–Pearson hypothesis testing, the log-likelihood ratio is computed from observed data and compared with a predetermined threshold to choose a likely hypothesis (Poor, 1994). The proposed method uses Eq. (4) to compute the log-posterior ratio for as follows:

$$\begin{aligned}
\Lambda_t &= \log\left(\frac{P(\mathcal{M}_1 \mid S_t)}{P(\mathcal{M}_0 \mid S_t)} \right) = \log\left(\frac{\prod_{i=1}^{|Q|} \prod_{j=1}^{|\Sigma|} (m_{ij}^1)^{\alpha_{ij}^t - 1}}{\prod_{i=1}^{|Q|} \prod_{j=1}^{|\Sigma|} (m_{ij}^0)^{\alpha_{ij}^t - 1}} \right) \\
&= \log\left(\prod_{i=1}^{|Q|} \prod_{j=1}^{|\Sigma|} \left(\frac{m_{ij}^1}{m_{ij}^0} \right)^{\alpha_{ij}^t - 1} \right) \\
&= \sum_{i=1}^{|Q|} \sum_{j=1}^{|\Sigma|} (\alpha_{ij}^t - 1) \log\left(\frac{m_{ij}^1}{m_{ij}^0} \right), \\
\Lambda_t &= \sum_{i=1}^{|Q|} \sum_{j=1}^{|\Sigma|} w_{ij} (\alpha_{ij}^t - 1),
\end{aligned} \tag{7}$$

where the constant w_{ij} denotes $\log(m_{ij}^1 / m_{ij}^0)$. The intuition behind this statistic is as follows:

- The observed symbol sequence is used to obtain a posterior distribution over the model parameter space.
- Using that posterior distribution and the known model parameters from the training phase for each hypothesis, the probability of each model can be computed.
- The ratio of these posterior probabilities will be high (resp. low) if $\mathcal{M}_1$ is more (resp. less) likely to be the model generating the symbol sequence.

In the absence of any initial count of emissions of symbols from states, the hyper-parameters α_{ij}^t in Eq. (7) are initialized to be 1, i.e., $\alpha_{ij}^0 = 1 \ \forall i, j$. This is consistent with the uniform prior assumption to obtain an estimate of the Markov model parameters as discussed in Section 4.2. However, if prior observations of the number, N_{ij}^0, of emissions of symbol s_j from state q_i are available before the start of the hypothesis test, then the initial value is assigned as $\alpha_{ij}^0 = N_{ij}^0 + 1$. Specifically, in nested binary classification problems for anomaly/fault detection and then anomaly/fault classification, the Bayesian formulation allows to use observation counts from the first hypothesis test in the Dirichlet priors during initialization of the second test for faster classification. The fact that the Dirichlet class allows usage of non-informative priors as well as previous symbol counts to impose priors in a straightforward way is another reason for adopting the Bayesian approach.

Alternatively, one may derive an expression for log-likelihood ratio using the D-Markov models and choose to use log-likelihood ratio as the test statistic too. The log-likelihood ratio statistic is derived next.

Let S_t for $t \in \mathbb{N}$ denote the symbol sequence $(s_1, s_2, \dots, s_t)$ of length t in the testing phase. The likelihood of an observed symbol sequence S_t for a given model $\mathcal{M}_k$ is computed as:

$$P(S_t \mid \mathcal{M}_k) = m_{q_0 s_{D+1}}^k m_{q_1 s_{D+2}}^k \cdots m_{q_{t-D} s_t}^k = \prod_{i=1}^{|Q|} \prod_{j=1}^{|\Sigma|} (m_{ij}^k)^{n_{ij}^t} \tag{8}$$

where the state q_i represents the D-length word $(s_{i+1}s_{i+2}\dots s_{i+D})$, the value of n_{ij}^t is the count of occurrences of symbol s_j at state q_i in a sequence S_t. Similar to the derivation of Eq. (7) from Eq. (4), a log-likelihood ratio is derived from Eq. (8) as:

$$\widetilde{\Lambda}_t = \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij} n_{ij}^t. \tag{9}$$

A simple relation to compute the log-posterior ratio was obtained in Eq. (7). Its relationship to log-likelihood ratio from Eq. (9) will be shown ahead. It follows from Eq. (7), the relationship $\alpha_{ij}^t = \alpha_{ij}^0 + n_{ij}^t$, and Eq. (9) that:

$$\begin{aligned}\Lambda_t &= \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij}(\alpha_{ij}^0 + n_{ij}^t - 1)\\ &= \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij} n_{ij}^t + \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij}(\alpha_{ij}^0 - 1)\\ &= \widetilde{\Lambda}_t + \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij}(\alpha_{ij}^0 - 1).\end{aligned} \tag{10}$$

The above relationship highlights that, without any informative prior, where $\alpha_{ij}^0 = 1$, the statistic log-likelihood ratio and log-posterior ratio are identical and initialize at zero. However, any prior symbol occurrence counts can be used to bias the test statistic initially by assigning a non-zero value. The sequential update rule for log-posterior ratio is formulated next.

Proposition 4.1 (*Sequential Update*). *Given that the log-posterior ratio at time t is Λ_t and the state at time $t+1$ is q_{t+1}, if s_{t+1} is the emitted symbol, then the updated log-posterior ratio Λ_{t+1} is given by :*

$$\Lambda_{t+1} = \Lambda_t + w_{q_{t+1}s_{t+1}}, \tag{11}$$

where $w_{q_{t+1}s_{t+1}} = \{w_{ij} : q_{t+1} = i \text{ and } s_{t+1} = j\}$.

Proof. Recalling the hyper-parameter $\alpha_{ij}^{t+1} - \alpha_{ij}^0$ to be the count of occurrence of symbol s_j at the state q_i in the observed symbol sequence S_{t+1}, it follows that:

$$\alpha_{ij}^{t+1} = \alpha_{ij}^t + \mathbb{1}_{\{i\}}(q_{t+1})\mathbb{1}_{\{j\}}(s_{t+1}), \tag{12}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function with set A, that is, $\mathbb{1}_A(x) = 1$, if x belongs to A; otherwise, $\mathbb{1}_A(x) = 0$. Then, by using the log-posterior ratio given in Eq. (7) for time $t+1$, and Eq. (12), it follows that:

$$\begin{aligned}\Lambda_{t+1} &= \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij}(\alpha_{ij}^t + \mathbb{1}_{\{i\}}(q_{t+1})\mathbb{1}_{\{j\}}(s_{t+1}) - 1)\\ &= \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij}(\alpha_{ij}^t - 1) + \sum_{i=1}^{|Q|}\sum_{j=1}^{|\Sigma|} w_{ij}\mathbb{1}_{\{i\}}(q_{t+1})\mathbb{1}_{\{j\}}(s_{t+1})\\ &= \Lambda_t + w_{q_{t+1}s_{t+1}}.\end{aligned}$$

□

The simplicity of sequential update of the log-posterior ratio for D-Markov models enables online update of the statistic $\Lambda_t(S_t)$. The time complexity of this sequential update is $O(\log(|\Sigma|))$ as explained below.

Using binary search, the new measurement is assigned a symbol s_{t+1} by efficiently identifying its location in a sorted sequence of real numbers. This sorted sequence represents a partition of the measurement range. The time complexity of symbolization via binary search is known to be $O(\log(|\Sigma|))$. Given the symbol s_{t+1} and previous state q_{t+1}, the relation $w_{q_{t+1}s_{t+1}} = w_{ij}\mathbb{1}_{\{i\}}(q_{t+1})\mathbb{1}_{\{j\}}(s_{t+1})$ implies that $w_{q_{t+1}s_{t+1}}$ can be obtained directly from a look-up table and added to previous statistic Λ_t in constant time. Thus, time complexity of sequential update is given by $O(\log(|\Sigma|))$.

Since alphabet size $|\Sigma|$ is typically small, the time complexity of update shows that the approach is feasible for real-time monitoring of physical systems. It is noted that, under the current framework of Dirichlet distribution for the class of D-Markov models, the expression for likelihood ratio computation as well as its update is significantly simplified when compared to the same for hidden Markov models as proposed in Fuh (2003). This simplification is attributed to the perfect information of the states being observed in the current class of models. In the next part, the computed log-posterior ratio is used to obtain a sequential decision rule.

4.4. Sequential hypothesis test for D-Markov models

As mentioned in Section 3.2, a sequential decision rule is a pair of sequences $(\bar{\phi}, \bar{\mu})$, where $\bar{\phi} \triangleq \{\phi_j : j = 0, 1, 2, \dots\}$ is called a stopping rule and $\bar{\mu} \triangleq \{\mu_j : j = 0, 1, 2, \dots\}$ is called a terminal decision rule. By choosing two thresholds γ_0 and γ_1 with $\gamma_0 < 0 < \gamma_1$, the sequential hypothesis test SHT(γ_0, γ_1) is constructed from the sequence of stopping rule ϕ_j and terminal decision rule μ_j as follows:

$$\phi_j(S_j) = \begin{cases} 0; & \text{if } \gamma_0 < \Lambda_j(S_j) < \gamma_1,\\ 1; & \text{otherwise.}\end{cases} \tag{13a}$$

$$\mu_j(S_j) = \begin{cases} 0; & \text{if } \Lambda_j(S_j) \le \gamma_0,\\ 1; & \text{if } \Lambda_j(S_j) \ge \gamma_1.\end{cases} \tag{13b}$$

Let p_d be the probability of detection and p_{fa} be the probability of false alarm of the detector. In order to choose the thresholds, the following relations from Wald's SPRT (Poor, 1994) are used:

$$\gamma_1 \le \log\left(\frac{p_d}{p_{fa}}\right) \text{ and } \gamma_0 \ge \log\left(\frac{1-p_d}{1-p_{fa}}\right). \tag{14}$$

These relations hold even without the assumption of independence (Grossi and Lops, 2008) and the underlying inequalities not only relate the performance with the chosen thresholds but also help to choose the threshold to guarantee a particular desired performance.

4.5. Summary of the proposed approach

The proposed approach is summarized in this subsection and a schematic overview is provided in Fig. 1. In the training phase, the pre-recorded time-series data are used to learn the corresponding D-Markov models. In this learning process, the model hyper-parameters (i.e., alphabet size, boundary locations of partition segments, and depth) are estimated using the techniques describes in Section 4.1. The parameters of the time-series model, i.e. the entries of the morph matrix, are then obtained using a Bayesian approach given in Section 4.2. The maximum *a posteriori* probability (MAP) estimates of the model parameters for each hypothesis are then used to compute a weight matrix $W = [w_{ij}]$, where $w_{ij} \triangleq \log(m_{ij}^1)/\log(m_{ij}^0)$, which is used in the sequential update step during operation. Finally, the user-defined performance is used to obtain thresholds for the sequential test.

During operation with streaming data at initialization of the test, the log-posterior ratio is initialized using suitable hyper-parameters α_{ij}^0 for all $i \in Q$ and $j \in \Sigma$. Based on the first D symbols, the state q_{D+1} is initialized to be $(s_1 s_2 \dots s_D)$. After initialization, each observation from the data acquisition system is symbolized, used to obtain next state, and also update the log-posterior ratio statistic using Eq. (11). The statistic is then compared with the predefined thresholds in the stopping rule and terminal decision rule to either declare the hypothesis or continue sampling. These computations enable online monitoring and hypothesis testing with streaming data.

Fig. 1 summarizes the process during training and during operation. In the next part of this section, the sequential detector is analyzed in detail and certain convergence and test completion results have been derived.

4.6. Analysis of the sequential detector

The sequential detector SHT(γ_0, γ_1) has been shown in Eq. (13). This part explores the stochastic evolution and asymptotic behavior of the log-posterior ratio (LPR) statistic for D-Markov models as well as prove that the sequential detector will terminate in finite time with probability one.

The sequential update rule for LPR in Eq. (11) suggests that its stochastic evolution has some resemblance with a random walk on the real line (Papoulis and Pillai, 2002). This stochastic evolution is characterized in the following Proposition.

Proposition 4.2 (*Random Walk on a DFSA*)**.** *Given that DFSA $\mathcal{G}$ is the common algebraic structure in both models $\mathcal{M}_0$ and $\mathcal{M}_1$, and S_t is a (finite-length) symbol sequence over the alphabet Σ, then the stochastic evolution of the log-posterior ratio (LPR) $\Lambda_t = \log\left(\frac{P(\mathcal{M}_1|S_t)}{P(\mathcal{M}_0|S_t)}\right)$, is given by a weighted random walk on the DFSA $\mathcal{G}$.*

Proof. Using Eq. (11), it follows that:

$$\begin{aligned}\Lambda_{t+2} &= \Lambda_{t+1} + w_{q_{t+2}s_{t+2}},\\ &= \Lambda_t + w_{q_{t+1}s_{t+1}} + w_{q_{t+2}s_{t+2}}.\end{aligned}$$

Given q_{t+1} and s_{t+1}, the next state $q_{t+2} = \delta(q_{t+1}, s_{t+1})$, thus, q_{t+2} is not random and is governed by the algebraic structure of the DFSA. The symbol emission probabilities are obtained from the morph matrix of true model, given the current state. The step size $w_{q_t s_t}$ at time step t can take $|Q||\Sigma|$ different values depending on the state and emitted symbol pair. Thus, the evolution of LPR is a weighted random walk over a DFSA. □

Given the stochastic evolution of the LPR statistic from Proposition 4.2, the expected value of the LPR statistic is presented in the following theorem.

Theorem 4.3 (*Expected Value of Log-Posterior Ratio (LPR)*)**.** *Given that $p_k(q_1)$ is a row vector of initial state probabilities, T_k is the state transition probability matrix and M_k is the morph matrix for model $\mathcal{M}_k$, and $W = [w_{ij}] = [\log(m^1_{ij}/m^0_{ij})]$, then the expectation (over all possible symbol sequences) of the log-posterior ratio (LPR) statistic after t updates is given as:*

$$\mathrm{E}[\Lambda_t \mid \mathcal{M}_k] = \sum_{l=1}^{t} p_k(q_1) T_k^{l-1} (M_k \odot W) \mathbf{1}_{|\Sigma|}, \tag{15}$$

where $\odot$ is element-wise multiplication operator and $\mathbf{1}_n$ is column vector of n ones.

Proof. Using $\Lambda_{t+1} = \Lambda_t + w_{q_{t+1}s_{t+1}}$ from Eq. (11), it follows that:

$$\begin{aligned}&\mathrm{E}[\Lambda_{t+1} - \Lambda_t \mid \mathcal{M}_k] = \mathrm{E}[w_{s_{t+1}q_{t+1}} \mid \mathcal{M}_k]\\ &= \mathrm{E}\left[\sum_{j=1}^{|Q|}\sum_{i=1}^{|\Sigma|} w_{ij} \mathbb{1}_{\{i\}}(q_{t+1}) \mathbb{1}_{\{j\}}(s_{t+1}) \mid \mathcal{M}_k\right]\\ &= \sum_{j=1}^{|Q|}\sum_{i=1}^{|\Sigma|} w_{ij} p_k(q_{t+1}=j) p_k(s_{t+1}=i|q_{t+1}=j)\\ &= \sum_{j=1}^{|Q|} p_k(q_{t+1}=j) \sum_{i=1}^{|\Sigma|} w_{ij} p_k(s_{t+1}=i|q_{t+1}=j)\\ &= p_k(q_{t+1})(M_k \odot W)\mathbf{1}_{|\Sigma|}.\end{aligned} \tag{16}$$

Using $p_k(q_{t+1}) = p_k(q_1)T_k^t$, it follows that:

$$\mathrm{E}[\Lambda_{t+1} - \Lambda_t \mid \mathcal{M}_k] = p_k(q_1) T_k^t (M_k \odot W)\mathbf{1}_{|\Sigma|}.$$

Thus, the expression for the expected increment in LPR is derived after t updates. Finally, adding up the expected increments, the desired result is obtained for the expected LPR as:

$$\mathrm{E}[\Lambda_t \mid \mathcal{M}_k] = \sum_{l=1}^{t} p_k(q_1) T_k^{l-1} (M_k \odot W)\mathbf{1}_{|\Sigma|}. \quad \square$$

The information-theoretic interpretation of the result in Theorem 4.3 is explained by the following remark.

Remark 4.4 (*Information-Theoretic Interpretation*)**.** Using Eq. (16) and the relation for conditional relative entropy Eq. (2), it follows that:

$$\begin{aligned}&\mathrm{E}[\Lambda_{t+1} - \Lambda_t \mid \mathcal{M}_1] = p_1(q_{t+1})(M_1 \odot W)\mathbf{1}_{|\Sigma|}\\ &= \sum_{j=1}^{|Q|}\sum_{i=1}^{|\Sigma|} p_1(q_{t+1}=j) m^1_{ij} \log\left(\frac{m^1_{ij}}{m^0_{ij}}\right)\\ &= \mathrm{E}_{p_1(s_{t+1},q_{t+1})}\left[\log\left(\frac{p_1(s_{t+1} \mid q_{t+1})}{p_0(s_{t+1} \mid q_{t+1})}\right)\right]\\ &= \mathrm{d}_{t+1}\big(p_1(s_{t+1} \mid q_{t+1}) \parallel p_0(s_{t+1} \mid q_{t+1})\big)\\ &= \mathrm{d}_{t+1}(M_1 \parallel M_0).\end{aligned}$$

Similarly, using non-negativity of conditional relative entropy, it follows that:

$$\mathrm{E}[\Lambda_{t+1} - \Lambda_t \mid \mathcal{M}_0] = -\mathrm{d}_{t+1}(M_0 \parallel M_1).$$

Thus, one may infer that expected increment in log-posterior ratio at time $t+1$ is the conditional relative entropy between the given models at time t+1, where time dependence is introduced because of state probability vector $p_1(q_{t+1})$. Thus, considering conditional relative entropy to be a distance function, the expected increment is the statistical distance between the two different models. Moreover, the derived relationship shows that the expected value of the statistic at time t is

$$\mathrm{E}[\Lambda_t \mid \mathcal{M}_k] = (-1)^{1-k} \sum_{l=1}^{t} \mathrm{d}_l(M_k \parallel M_{1-k})$$

Here the magnitude of the expected value of the statistic is directly proportional to the distance between the models. Thus, this result implies that models which are closer will need more incremental updates to cross a prescribed threshold. In other words, models which are closer will need larger number of observations to terminate the sequential hypothesis test for a given desired performance.

There are two consequences of the result in this remark:

- The expected increment and the rate of sequential update depends directly on the statistical distance between models of the hypotheses. Hence, the speed of sequential update cannot be tuned with any parameter. However, an alternative approach to create reduced-order Markov models (Jha et al., 2018) of time-series data with smaller state set can be employed. It can provide sufficiently accurate models that maybe further apart and yield high expected increment in test statistic leading to faster termination of the test.
- The work in Beim Graben (2001) showed that symbolic dynamics approach shows good robustness to measurement noise as additive noise leads to uncertainty in the assigned symbol for a measurement, which are near the partition boundaries only. The noise statistics observed in the training data sequence are already captured in the Markov models. Addition of noise in data for each hypothesis leads to reduction in conditional relative entropy between models giving shorter expected increments. Thus, noisy signals will lead to slower termination of test to guarantee similar performance.

The results from Theorem 4.3 are used in next two remarks to derive the asymptotic behavior of the test statistic and to obtain insights on special case where the D-Markov models are stationary.

Remark 4.5 (*Asymptotic Behavior of Log-Posterior Ratio*)**.** Using the result from Theorem 4.3 and computing the limit as number of observations tend to infinity, it follows that:

$$\lim_{t\to\infty} \frac{\mathrm{E}[\Lambda_t \mid \mathcal{M}_k]}{t} = \tilde{\pi}_k (M_k \odot W)\mathbf{1}_{|\Sigma|} \tag{17}$$

where $\tilde{\pi}_k = \lim_{t\to\infty}\left(\frac{1}{t}\sum_{l=1}^{t} p_k(q_1)T_k^{l-1}\right)$ is the Césaro limit. The constant value on the right hand side of this result implies that the growth of log-posterior ratio becomes linear asymptotically.

Remark 4.6 (*Stationary Markov Models*)**.** Let two Markov models $\mathcal{M}_0$ and $\mathcal{M}_1$ be stationary and let their stationary distributions be given as π_0 and π_1, respectively. Thus, the state probability vector is the same as the stationary distribution of the Markov model at all time epochs, i.e., $p_k(q_t) = \pi_k$ for all $t \in \mathbb{N}$ and $k = 0, 1$. Using π_k in Eq. (16), it follows that:

$$\mathrm{E}[\Lambda_{t+1} - \Lambda_t \mid \mathcal{M}_k] = \pi_k(M_k \odot W)\mathbf{1}_{|\Sigma|}, \tag{18}$$

which has a constant value. Hence, the expectation under model $\mathcal{M}_k$ is given by:

$$\mathrm{E}[\Lambda_t \mid \mathcal{M}_k] = t\pi_k(M_k \odot W)\mathbf{1}_{|\Sigma|}. \tag{19}$$

Since, it is known that the Césaro limit is equal to the stationary distribution for irreducible aperiodic Markov models (Meyn and Tweedie, 2012), this result matches the intuition from Remark 4.5 with $\tilde{\pi}_k = \pi_k$. Using the interpretation from Remark 4.4, it follows that:

$$\mathrm{E}[\Lambda_t \mid \mathcal{M}_k] = (-1)^{1-k} t\, \mathrm{d}(M_k \parallel M_{1-k}), \tag{20}$$

for $k = 0, 1$. This relation for the stationary case shows a direct relationship between the expected value of log-posterior ratio and the distance between the models.

4.7. Completion of the test and estimation of the sequence length

This subsection proves that the sequential hypothesis test for D-Markov models terminates in finite time with probability one and also provides an estimate of the average observation sequence length. Similar to the methodology in Grossi and Lops (2008), the notion of stopping time under each hypothesis for a symbol sequence S_t is given as follows:

$$\tau_0(S_t) = \inf\{t \in \mathbb{N} : \Lambda_t(S_t) \le \gamma_0\}, \tag{21a}$$

$$\tau_1(S_t) = \inf\{t \in \mathbb{N} : \Lambda_t(S_t) \ge \gamma_1\}. \tag{21b}$$

Then, the stopping time for the sequential test is given by $\tau(S_t) = \min\{\tau_0(S_t), \tau_1(S_t)\}$. The formal result for the guarantee of completion of the sequential test in finite time is given below as Theorem 4.7.

Theorem 4.7 (*Completion of the Test*)**.** *If the discrete finite-order Markov process generating the symbols is ergodic, then the sequential test with decision rule shown in Section 4.4 terminates in finite time almost surely under both hypothesis, i.e.,* $P(\{\tau(S_t) < +\infty\}|\mathcal{M}_k) = 1$*, for* $k = 0, 1$.

Proof. Ergodicity of the symbol sequence generated by a Markov model implies that the sequence of log-posterior ratio (LPR) is also ergodic for the model, because the LPR is a deterministic function of the stochastic sequence. Thus, the sequence of time averages of the LPR converges to the ensemble average (i.e., expected value) of the LPR almost surely, that is,

$$\lim_{t\to\infty} \frac{\mathrm{E}[\Lambda_t \mid \mathcal{M}_k]}{t} \overset{as}{=} \lim_{t\to\infty} \frac{1}{t}\Lambda_t(S_t) \tag{22}$$

under hypothesis $k \in \mathcal{X}$. Using Eq. (17) for $k = 1$, it follows that:

$$\lim_{t\to\infty} \frac{1}{t}\Lambda_t(S_t) \overset{as}{=} \tilde{\pi}_1(M_1 \odot W)\mathbf{1}_{|\Sigma|} \tag{23}$$

The constant value in this result implies that asymptotic growth rate of test statistic sequence is linear. Thus, eventually the statistic will be greater than any finite threshold, i.e., $P(\{\lim_{t\to\infty} \Lambda_t(S_t) > \gamma_1\}|\mathcal{M}_1) = 1$. Similarly, for $k = 0$, it follows that:

$$P(\{\lim_{t\to\infty} \Lambda_t(S_t) < \gamma_0\}|\mathcal{M}_0) = 1$$

Table 1
D-Markov models used in the numerical simulation.

Case	M_0	M_1	$\mathrm{d}(M_0 \parallel M_1)$	$\mathrm{d}(M_1 \parallel M_0)$
1	$\begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \\ 0.6 & 0.4 \\ 0.7 & 0.3 \end{bmatrix}$	$\begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \\ 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix}$	0.037	0.075
2	$\begin{bmatrix} 0.1 & 0.9 \\ 0.7 & 0.3 \\ 0.6 & 0.4 \\ 0.7 & 0.3 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \\ 0.4 & 0.6 \\ 0.3 & 0.7 \end{bmatrix}$	0.561	0.989

Using Eq. (21) for $k = 0, 1$, it follows that:

$$P(\{\tau_k(S_t) < +\infty\}|\mathcal{M}_k) = 1$$

The proof follows from $\tau(S_t) = \min\{\tau_0(S_t), \tau_1(S_t)\}$. □

It follows from Theorem 4.7 that the sequential tests for D-Markov machines terminate in finite time with probability one. Now an estimate of this stopping time is obtained with respect to the expected behavior of the LPR sequence. For each hypothesis, this sequence length estimate is defined by using $\mathrm{E}[\Lambda_t \mid \mathcal{M}_k]$ from Eq. (15) as:

$$\widetilde{\tau}_1 \triangleq \inf_{t\in\mathbb{N}}\left\{t : \sum_{l=1}^{t} p_1(q_1)T_1^{l-1}(M_1 \odot W)\mathbf{1}_{|\Sigma|} \ge \gamma_1\right\}, \tag{24a}$$

$$\widetilde{\tau}_0 \triangleq \inf_{t\in\mathbb{N}}\left\{t : \sum_{l=1}^{t} p_0(q_1)T_0^{l-1}(M_0 \odot W)\mathbf{1}_{|\Sigma|} \le \gamma_0\right\}. \tag{24b}$$

The utility of these estimates and other theoretical results from this section are elucidated by numerical simulation and validation with experimental data in the next two sections.

5. Numerical simulation

A theoretical framework for sequential hypothesis testing with D-Markov models has been developed in the previous sections. This section presents the results of numerical simulation to elucidate the underlying principles of the proposed sequential hypothesis testing procedure.

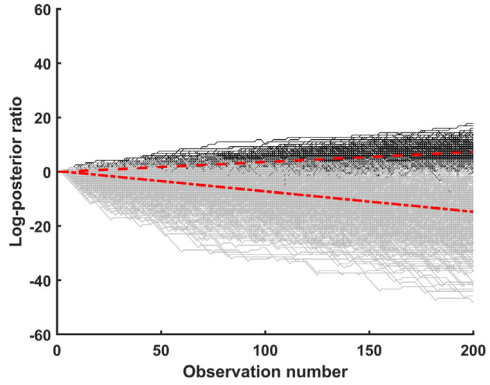
5.1. Description of the simulation scenarios

An anomaly detection scenario is simulated by two D-Markov models, having a common binary alphabet set (i.e., $\Sigma = \{0, 1\}$) and the same depth $D = 2$. These models have been trained with sufficiently long data. The first model, $\mathcal{M}_0$, represents the nominal behavior and the second model, $\mathcal{M}_1$, represents the anomalous behavior. The state set Q for each model is given by the words $\{00, 01, 10, 11\}$. The simulation has been performed for two different cases: *Case 1*, where models $\mathcal{M}_0$ and $\mathcal{M}_1$ are largely similar; and *Case 2*, where models $\mathcal{M}_0$ and $\mathcal{M}_1$ are reasonably different. The morph matrices for the models from both the cases are shown in Table 1. The distance between the models, in terms of conditional relative entropy, using the stationary distribution for the state probabilities, is also given in Table 1. It is verified that *Case 1* models are closer than the models in *Case 2*.

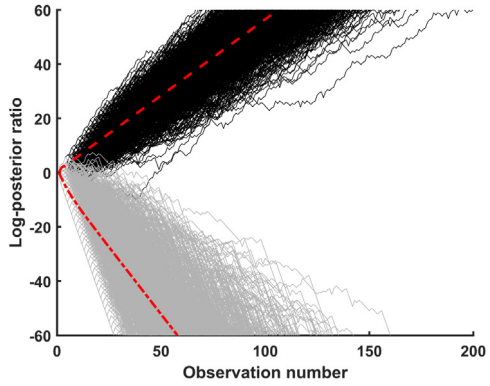
Each model simulates 2500 symbol sequences of length 1000 and uses the sequential detector algorithms from Section 4.4 to make decisions. The results from this analysis are discussed next.

5.2. Simulation results

Simulated symbol sequences have been used to sequentially estimate the log-posterior ratio (LPR) statistic. In order to visualize the behavior of the LPR statistics under different hypotheses, 1000 LPR trajectories are shown under each hypothesis for both cases in Fig. 2.

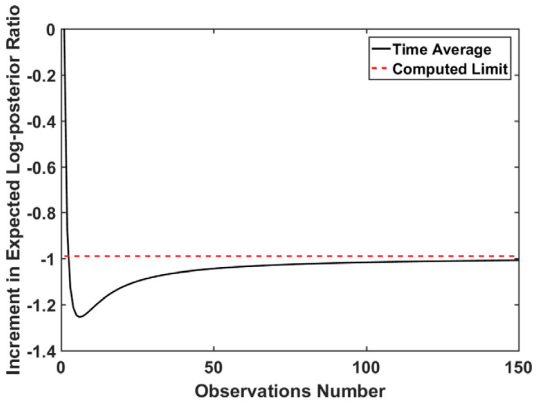


(a) *Case 1*: Models $\mathcal{M}_0$ and $\mathcal{M}_1$ being largely similar

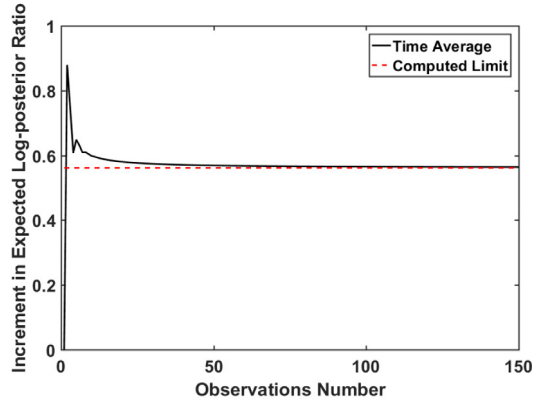


(b) *Case 2*: Models $\mathcal{M}_0$ and $\mathcal{M}_1$ being different

Fig. 2. Log-posterior ratio trajectories, when $\mathcal{H}_0$ is true (shown in gray) and when $\mathcal{H}_1$ is true (shown in black), for (a) *Case 1* and (b) *Case 2*. The expected value of LPR in each case is shown for $\mathcal{H}_0$ (with red –· line) and $\mathcal{H}_1$ (with red –– line). The log-posterior ratio grows much faster when models are different (*Case 2*) as compared to when models are largely similar (*Case 1*).



(a) Under $\mathcal{H}_0$



(b) Under $\mathcal{H}_1$

Fig. 3. Verification of convergence of the expected log-posterior ratio to the computed limit (from Remark 4.5) under each hypothesis.

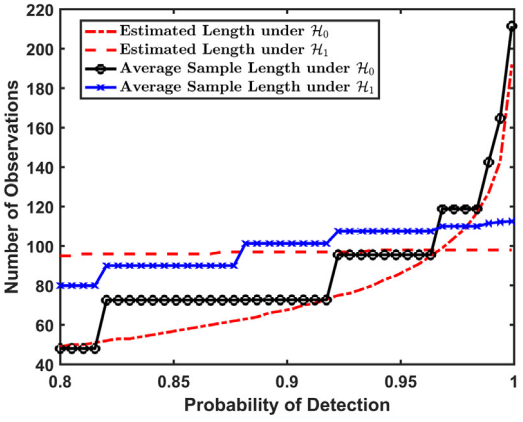


Fig. 4. Average sample length and estimated length for different values of desired probability of detection (p_d) and constant probability of false alarm (p_{fa}) = 0.001.

The expected value of the LPR is computed using the derived equation Eq. (15) and it is also shown in Fig. 2. It can be qualitatively verified that the LPR trajectories under the two hypothesis begin to differ as more observations are used in both cases. Moreover, if the models are further apart (as in *Case 2*), then the distinction between the LPR trajectories is discernible using fewer observations. The models in *Case 1* have identical emission probabilities for 3 out of 4 states (see Table 1), thus the discriminatory information is only available when the model is in state 4. Hence, in Fig. 2(a), it is seen that the LPR value remains constant for several consecutive observations.

In Remark 4.5, the asymptotic value of the average increment of expected value LPR has been derived under each hypothesis. Fig. 3 qualitatively verifies the convergence of the time average of increment in expected LPR to the limit computed in Remark 4.5 under both hypotheses for *Case 2*.

An estimate of the sequence length is given in Eq. (24), which is shown to be a reasonable estimate of the average sample length; this is seen in Fig. 4 that compares the estimated length with the average sample length by repeating the simulation for 40 different choices of probability of detection between 0.8 and 0.999, while keeping probability of false alarm (p_{fa}) fixed at 0.001. Fig. 4 also verifies the intuition that more observations are needed to achieve better performance.

Fig. 5 shows that the histograms of stopping length under $\mathcal{H}_0$ and that the histograms of stopping length under $\mathcal{H}_1$ for Case 1 of the simulation models. From Remark 4.4, it follows that magnitude of expected increment in LPR, under a given hypothesis, say $\mathcal{H}_k$, is equal to the distance of the true model from the alternate model, i.e., $\mathrm{d}(M_k \parallel M_{1-k})$. Thus, if $\mathrm{d}(M_1 \parallel M_0) > \mathrm{d}(M_0 \parallel M_1)$ (see Table 1 for *Case 1*), then on average it is expected the sequential test to terminate with fewer observations under hypothesis $\mathcal{H}_1$. This can be verified in Fig. 5, where the distribution for $\mathcal{H}_1$ is to the left of $\mathcal{H}_0$. It is noted that the empirical distribution is not symmetric, thus, the characterization of the stopping time for the test requires further investigation by estimating the higher moments of the respective distributions (Grossi and Lops, 2008).

The efficacy of a sequential detector is characterized by the average sample length and the detection performance. The desired performance in terms of probability of detection and false alarm is used to design the sequential detector, as shown in Section 4.4. In Fig. 6, it

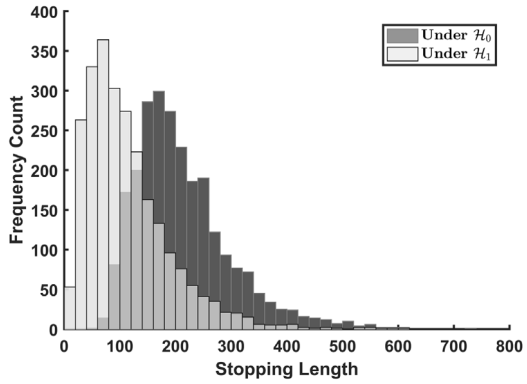


Fig. 5. Histograms of sample length under each hypothesis with $p_d = 0.999$ and constant $p_{fa} = 0.001$.

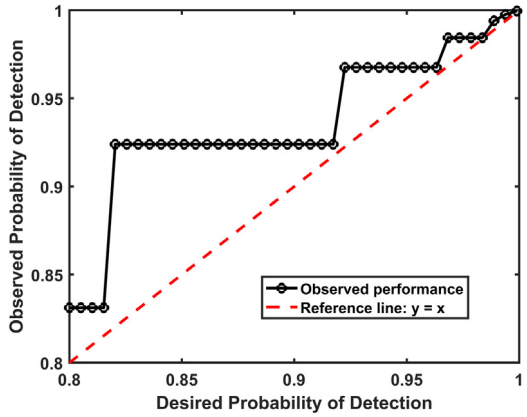


Fig. 6. Observed probability of detection for different values of desired probability of detection (p_d) and constant probability of false alarm (p_{fa}) = 0.001 using simulated data.

is seen that the probability of detection of the sequential detector is better than the desired p_d and the probability of false alarm was also lesser than 0.001. Thus, the claim that sequential detector with D-Markov models achieves the desired performance has been verified with simulation.

The performance of the proposed sequential approach is compared with the maximum likelihood classifier, which is a fixed-sample-size (FSS) test. The classification rule for a symbol sequence S_N of length N is given in terms of the likelihood as follows: $x^*(S_N) = \text{argmax}_{x\in\mathcal{X}}\, p(S_N|\mathcal{M}_x)$. The comparison of performance of the classifier with the developed approach for different probability of detection (p_d) is shown in Table 2. It is inferred that, for the same p_d, on average the FSS test needs several more observations as compared to the proposed sequential detector. It is noted that the desired p_{fa} is chosen to be 0.001 for the SHT and the observed p_{fa} is 0.0004, which is lower than all cases of ML classifier. Thus, better performance is achieved with lesser number of observations (on average) with the proposed sequential detector.

In summary, the inferences from the results of numerical simulation are given as follows:

- The LPR trajectories differ under different hypothesis as more observations are used and the distance between the models governs the expected LPR increment.
- The average increment of expected LPR from the simulated models converges to the limit derived in Eq. (17).
- The estimate of the sequence length provided in Eq. (24) is a reasonable estimate of the average sample length.

Table 2
Comparison of the performance (p_{fa}) and average sample length (ASL) of the proposed sequential hypothesis test (SHT) and maximum likelihood (ML) classifier (with fixed sequence length (N)) for the same p_d.

p_d	SHT		ML	
	p_{fa}	ASL	p_{fa}	N
0.960	0.0004	101.5	0.0360	140
0.980	0.0004	114.4	0.0260	170
0.995	0.0004	138.4	0.0044	280
0.999	0.0004	162.0	0.0016	380

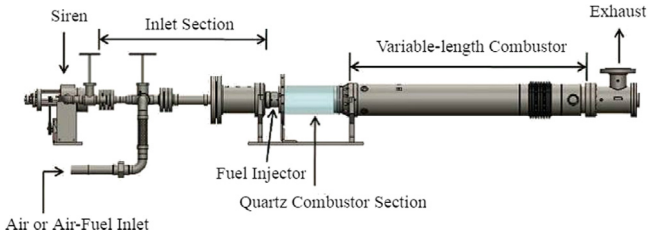


Fig. 7. Schematic diagram of the combustor apparatus.

- The models that are further apart need fewer observations for termination of the sequential detector.
- The thresholds in the sequential detector can be chosen based on any desired performance and that performance is guaranteed by the detector at the expense of sampling more observations.
- The sequential detector gives better performance with fewer observations (on average) as compared to fixed-sample-size tests in this study.

6. Validation on a laboratory apparatus

This section describes and analyzes the data used for detection of unstable behavior in lean pre-mixed combustion for validation on a laboratory apparatus, where thermo-acoustic instabilities are a consequence of nonlinear coupling between the acoustic and thermal oscillations during combustion and have undesirable effects on durability of structural components as well as performance of gas-turbine engines. The dynamics of the thermo-acoustic phenomena are very fast and the underlying physics is still not completely understood (O'Connor et al., 2015; Jha et al., 2016). In this context, the detection system should be capable of identifying the onset of thermo-acoustic instabilities as early as possible with negligible false alarm rates.

6.1. Experimental apparatus and data

This subsection very briefly discusses the data collected for detection of thermo-acoustic instabilities in a laboratory environment. Fig. 7 presents a schematic diagram of the test apparatus that is a swirl-stabilized, lean-premixed, laboratory-scale combustor (see Kim et al. (2010) for details). The apparatus consists of an inlet section, an injector, a combustion chamber, and an exhaust section. The combustor chamber consists of an optically-accessible quartz section followed by a variable length steel section. Tests have been conducted at a nominal combustor pressure of 1 atmosphere over a range of operating conditions, as listed in Table 3.

In each test, dynamic pressure and global OH and CH chemiluminescence intensity in the combustion chamber were measured to study the mechanisms of combustion instability. The measurements were made simultaneously at a sampling rate of 8192 Hz (per channel), and data were collected for 8 s, which include a total of 65,536 measurements (per channel). A total of 780 cases for pressure data are collected

Table 3
Operating conditions of the combustor apparatus.

Parameters	Value
Equivalence ratio	0.525, 0.55, 0.60, 0.65
Combustor length	0. 635 m to 1.475 m in 25 mm increments

(each sequence 65,536 long); however, the label for stable/unstable case is not available. In this work, the labels are created by clustering the Markov models of pressure data and using some domain expertise. The Markov models are clustered in an unsupervised fashion using symmetric KL-distance and hierarchical clustering for the states of the Markov models. The states of the Markov model are, to some extent, similar to each other in the stable case — however, as the lean premixed combustion locks onto an unstable limit cycle, the states of the corresponding Markov models become more distinct. This behavior of the Markov models is used to cluster the stable and unstable classes and provides almost perfect separation. More details are available in a previous publication (Jha et al., 2018). Finally, 125 stable and 125 unstable cases have been used for analysis presented in this paper.

6.2. Time-series modeling and sequential tests on experimental data

The time-series data set is first normalized by subtracting the mean and then dividing by the standard deviation of its elements; this step corresponds to bias removal and variance normalization. Data from engineering systems are typically over-sampled to ensure that the underlying dynamics can be captured. Due to coarse-graining in the symbolization process, an over-sampled time-series may mask the true nature of the system dynamics in the symbolic domain (e.g., occurrence of self loops and irrelevant spurious transitions in the Markov chain). A time-series is first down-sampled to find the next crucial observation. The first minimum of auto-correlation function $R(\tau)$ generated from the observed time-series is obtained to find the uncorrelated samples in time. The data sets are then down-sampled by this lag (it is noted that different sequence will have different lags). Fig. 8 shows the auto-correlation function for a typical time-series in the region of unstable combustion, where the data are down-sampled by the lag marked in red rectangles. To avoid discarding significant amount of data due to down-sampling, the down-sampled data using different initial conditions (or offsets) are concatenated. Using this approach, the number of data points lost if reduced significantly to $\text{mod}(n, \tau) \leq \tau - 1$ where, n is the original data length and τ is the lag value. For example, if the original time-series is given by $x_1, x_2, x_3, \ldots$, after this pre-processing it could rearranged as $x_1, x_{\tau+1}, \ldots, x_{n-\tau+1}, x_2, \ldots, x_{n-\tau+2}, \ldots$. An example is shown in Fig. 9 for $\tau = 4$. Further details of this pre-processing are reported in Srivastav (2014), Jha et al. (2015).

The measurement space of the continuous time-series is then partitioned using maximum entropy partitioning (MEP) (Rajagopalan and Ray, 2006), where the information rich regions of the measurement space are partitioned finer and those with sparse information are partitioned coarser. In essence, each cell in the partitioned set contains (approximately) equal number of data points under MEP. A ternary alphabet with $\Sigma = \{0, 1, 2\}$ has been used to symbolize the data. As discussed in Section 6.1, data sets are analyzed from two different modes of combustion, stable and unstable, with the aim of detecting the unstable modes.

To perform the sequential ratio tests using the combustion data, a single-depth Markov model is first trained for stable and unstable cases on the discretized sequence. A single symbol sequence of length $\sim 65,000$ is used to train Markov models for the two hypotheses — stable and unstable. This model is then used to detect the unstable modes from a test data set with 250 discrete sequences with equal proportions of stable and unstable cases. Figs. 10 and 11 display the pertinent results of sequential tests. Fig. 10 compares the theoretical estimate of stopping

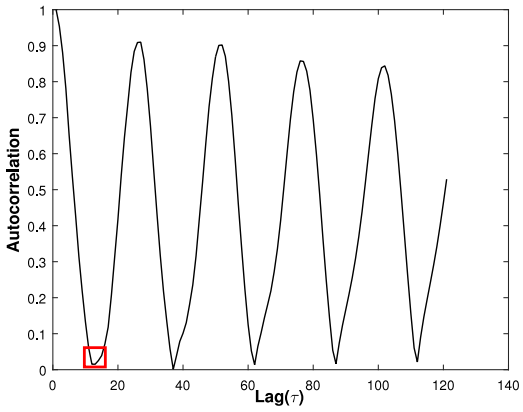


Fig. 8. Auto-correlation function of time-series data in the unstable phase of combustion. The time-series data is down-sampled by the lag marked in the red square. It is noted that the individual time-series have their own down-sampling lags.

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ x_5 & x_6 & x_7 & x_8 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n-3} & x_{n-2} & x_{n-1} & x_n \end{pmatrix} \longrightarrow x_1, x_5, \ldots, x_{n-3}, x_2, x_6, \ldots, x_{n-2}, \ldots$$

Fig. 9. A schematic demonstrating re-arrangement of data for $\tau = 4$. The data points x_i are first re-arranged row-wise and then the columns are concatenated. Only a maximum of $\text{mod}(n, \tau)$ data points are thrown away in the process.

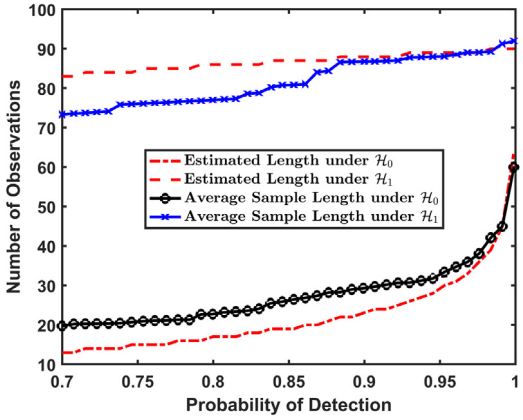


Fig. 10. Average sample length and estimated length for different values of p_d (max. 0.999) and constant $p_{fa} = 0.01$.

time for the sequential tests with the average sample length obtained over the test data set corresponding to different operating conditions specified by the desired probability of detection (p_d) and a constant false alarm rate, $p_{fa} = 0.01$. Fig. 11 shows the probability of detection achieved by the sequential tests for different desired detection rates. Observed probability of false alarm is 0.016 for all test cases, which is slightly higher than the desired value of $p_{fa} = 0.01$. Out of 125 stable and 125 unstable cases, there are 5 misdetections and 2 false alarms as the best performance. It is noted that the sequential tests for Markov models are able to achieve desired performance for most cases (except for cases when desired $p_d > 0.95$).

The results from sequential approach using experimental data are also compared with fixed-sample-size (FSS) maximum likelihood classifier. The desired p_d was chosen to be 0.95 for the SHT, the observed p_d is greater than 0.95 and it is also higher than all cases of ML classifier

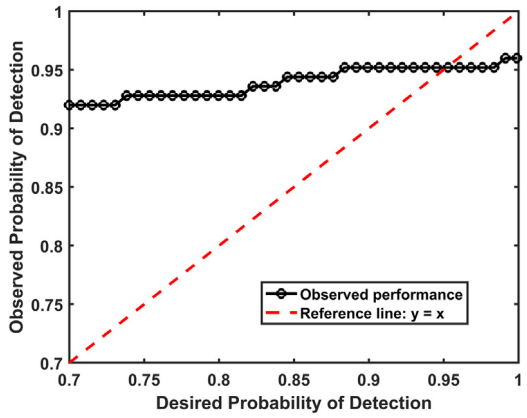


Fig. 11. Observed probability of detection for different values of desired probability of detection (p_d) and constant probability of false alarm (p_{fa}) = 0.01 using experiment data.

Table 4
Comparison of the performance (i.e., p_d for the for same p_{fa}) and average sample length (ASL) of the proposed sequential hypothesis test (SHT) with that of the (fixed length N) maximum likelihood (ML) classifier.

p_{fa}	SHT		ML	
	p_d	ASL	p_d	N
0.016	0.952	60.64	0.944	150
0.032	0.952	52.55	0.944	100
0.080	0.960	41.88	0.944	60
0.120	0.968	36.56	0.928	50

(See Table 4). Thus, the comparison shown in Table 4 verifies the claim with experiments that sequential tests need fewer observations than FSS tests on average to give the same (or even better) level of performance.

Remark 6.1. Even though the sequential detection procedure is tested using a dataset with equal number of stable and unstable samples (i.e., a balanced dataset), the Markov models are learned using just one such sample from the stable and unstable class (as mentioned earlier in the section that the models are trained using one long time-series). The theoretical results presented in the paper are independent of the number of samples available from the two classes (see Theorem 4.7). Hence, as long as one representative symbol sequence for each of the hypotheses under consideration is available, the proposed technique is expected to work well. Thus, the proposed sequential detection test is immune to data imbalance (which is common in a lot of industrial applications).

In contrast to the inferences from numerical simulation in Section 5.2, the sequential tests on experimental data could not always guarantee the desired performance. This observation can be attributed to the fact that the test set contains (experimental) data from a wide range of operating conditions as was discussed in Section 6.1. Thus, the test samples cannot be represented accurately by a single D-Markov model used for training.

Remark 6.2 (*Multi-Dimensional Time-Series:*)**.** The experimental example in this section considers only a single-dimensional time-series data for simplicity of presentation. However, the results presented in the paper are applicable to any symbol sequence with D-Markov property and thus can treat multi-dimensional data with any additional constraints. In essence, generating symbol sequences from multi-dimensional time-series is a non-trivial task and is a topic of ongoing research (Virani et al., 2016). Illustration of the proposed approach with multi-dimensional data is thus out-of-scope of the current paper and is recommended as a topic of future research.

7. Summary, conclusions, and future work

Symbolic time-series analysis-based Markov modeling provides learning and inference capabilities that can be used for on-line monitoring of critical physical systems. Sequential hypothesis tests are useful for fast and in-situ detection of anomalies and events in systems where measurement data are available in a streaming fashion. This paper formulates and validates a concept of sequential hypothesis testing for symbolic analysis-based Markov models of time-series data. First, a Bayesian approach for parameter estimation D-Markov models is presented using Dirichlet and categorical distributions. This approach has been used to estimate the log-posterior ratio (LPR) for hypothesis testing. The LPR is then used to develop a sequential hypothesis test for the D-Markov model. The stochastic evolution of the LPR statistic is studied and it is proved that the sequential test will terminate in finite time with probability one. The underlying concept is illustrated on two cases of PFSA models with a binary symbol alphabet and depth $D = 2$. Finally, the proposed sequential test has been used to detect occurrence of unstable combustion on experimental data of pressure time-series collected from a laboratory apparatus of swirl-stabilized combustion. The sequential tests are able to achieve 95% detection accuracy at 1.6% false alarm for detecting instabilities with short-length data over a wide range of operating conditions. The performance of the proposed sequential test for D-Markov models is also compared with fixed-sample-size test (maximum likelihood) to verify that sequential tests need fewer observations than FSS tests on average to give the same level of performance, which is important for real-time monitoring of time-critical processes.

Although the performance of the proposed sequential hypothesis test is apparently promising, much theoretical and experimental research is necessary before its real-life applications. To this end, the authors propose the following topics for future research.

(1) *Modeling uncertainties, measurement noise, and/or process variability*: Although symbolization provides robustness to noise (Beim Graben, 2001), the effects of modeling uncertainty and process variability on the posterior density and the sequential test need to be explored in the future.
(2) *Optimality of sequential test with D-Markov Models*: The SPRT with IID observations is known to be optimal in the sense that it minimizes the expected sample size among all sequential tests. Future research may investigate whether this optimality holds for the proposed sequential test for D-Markov observations.
(3) *Multi-hypothesis testing* (Baum and Veeravalli, 1994), *truncated SPRT* (Tantaratana and Thomas, 1977), *and Markov mode-switching* (Zhang et al., 2010): Future research can extend the current approach for tests with multiple hypothesis and truncated tests with limited decision horizon for D-Markov models. Instead of classification, one can also do time-series segmentation with system mode estimation and tracking via Markov mode-switching, where D-Markov models represent dynamics of each discrete mode.

Acknowledgments

The authors thank Professor Domenic Santavicca and Dr. Jihang Li from Pennsylvania State University for kindly providing the experimental data that were used to validate the theoretical results.

Appendix. Depth estimation using spectral analysis

This appendix briefly describes a technique for depth estimation in D-Markov models using their spectral properties. Srivastav (2014) has interpreted the significance of depth D of a symbol sequence (see Definition 2.3) as the number n of time steps after which probability of current symbol is independent of any past symbol, i.e.,

$$\Pr(s_k|s_{k-n}) = \Pr(s_k)\ \forall n > D, \tag{25}$$

where the statistical dependence is evaluated on individual past symbols as $\Pr(s_k|s_{k-n})$ instead of assessing the dependence on words of length D as $\Pr(s_k|s_{k-1},\dots,s_{k-D})$. Since the equality in Eq. (25) may not strictly hold, the following approximation is made for estimation of depth D as follows:

$$|\Pr(s_k|s_{k-n}) - \Pr(s_k)| \leq \epsilon\ \forall n > D \tag{26}$$

where ϵ is a user-specified tolerance or convergence condition. It is noted that $\Pr(s_k)$ is the stationary distribution for the one-step transition matrix. After some simplification and using the distance of the state-transition probability matrix after steps from the stationary point (see Eq. (26)), depth D is obtained such that the following condition is satisfied (see Srivastav (2014), Jha et al. (2018) for more details):

$$|\text{trace}\,(\mathbf{\Pi}^n) - \text{trace}\,(\mathbf{\Pi}^\infty)| \leq \sum_{j=2}^{J} \left|\lambda_j\right|^n < \epsilon \quad \forall n > D, \tag{27}$$

where n is the number of iterations of the PFSA; J is number of (non-zero) eigenvalues of $\mathbf{\Pi}$; and ϵ is the user-specified threshold as an appropriate convergence condition, i.e., the depth D of the symbol sequence is estimated for a choice of ϵ by estimating the stochastic matrix for the one-step PFSA. A pseudo-code of the process is presented in Algorithm 1.

Algorithm 1: `DepthEstimation`

1 **Input:** The observed symbol sequence $S = \{s_t\}_{t\in\mathbb{N}}$ and parameter ϵ
2 **Output:** Depth estimate D
3 Estimate the Markov model $\mathcal{M} = (Q, \Sigma, \delta, M)$, where $D = 1$
4 Estimate the state transition matrix $\mathbf{\Pi} = [\pi_{ij}]$, where $\pi_{ij} = \sum_{s\in\Sigma} M(q, s)$
5 Estimate D using the following relationship:

$$|\text{trace}\,(\mathbf{\Pi}^n) - \text{trace}\,(\mathbf{\Pi}^\infty)| \leq \sum_{j=2}^{J} \left|\lambda_j\right|^n < \epsilon \quad \forall n > D,$$

where, λ_j are the eigenvalues of the state transition matrix $\mathbf{\Pi}$

References

Adler, R., 1998. Symbolic dynamics and Markov partitions. Bull. Amer. Math. Soc. 35 (1), 1–56.
Baum, C.W., Veeravalli, V.V., 1994. A sequential procedure for multihypothesis testing. IEEE Trans. Inf. Theory 40 (6).
Beim Graben, P., 2001. Estimating and improving the signal-to-noise ratio of time series by symbolic dynamics. Phys. Rev. E 64 (5), 051104.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.
Bishop, C.M., 2006. Pattern recognition. Mach. Learn.
Buhl, M., Kennel, M.B., 2005. Statistically relaxing to generating partitions for observed time-series data. Phys. Rev. E 71 (4), 046213.
Burkholder, D., Wijsman, R., 1963. Optimum properties and admissibility of sequential tests. Ann. Math. Stat. 34 (1), 1–17.
Chen, B., Willett, P., 2000. Detection of hidden Markov model transient signals. IEEE Trans. Aerosp. Electron. Syst. 36 (4), 1253–1268.
Cover, T.M., Thomas, J.A., 1991. Information theory and statistics. Elem. Inf. Theory 279–335.
Darema, F., 2004. Dynamic data driven applications systems: A new paradigm for application simulations and measurements. In: Computational Science-ICCS 2004. Springer, pp. 662–669.
Daw, C.S., Finney, C.E.A., Tracy, E.R., 2003. A review of symbolic analysis of experimental data. Rev. Sci. Instrum. 74 (2), 915–930.
Fuh, C.-D., 2003. SPRT and CUSUM in hidden Markov models. Ann. Statist. 942–977.
Grossi, E., Lops, M., 2008. Sequential detection of Markov targets with trajectory estimation. IEEE Trans. Inf. Theory 54 (9), 4144–4154.
Jha, D.K., 2016. Learning and Decision Optimization in Data-Driven Autonomous Systems. (Ph.D. Thesis), The Pennsylvania State University.
Jha, D.K., Srivastav, A., Mukherjee, K., Ray, A., 2015. Depth estimation in Markov models of time-series data via spectral analysis. In: American Control Conference (ACC), 2015. IEEE, pp. 5812–5817.
Jha, D.K., Srivastav, A., Ray, A., 2016. Temporal learning in video data using deep learning and Gaussian processes. In: Workshop on Machine Learning for Prognostics and Health Managament At 2016 KDD. San Francisco, CA.
Jha, D.K., Virani, N., Reimann, J., Srivastav, A., Ray, A., 2018. Symbolic analysis-based reduced order Markov modeling of time series data. Signal Process. 149, 68–81.
Jin, X., Guo, Y., Sarkar, S., Ray, A., Edwards, R.M., 2011. Anomaly detection in nuclear power plants via symbolic dynamic filtering. IEEE Trans. Nucl. Sci. 58 (1), 277–288.
Kennel, M.B., Buhl, M., 2003. Estimating good discrete partitions from observed data: Symbolic false nearest neighbors. Phys. Rev. Lett. 91 (8), 084102.
Kennel, M.B., Buhl, M., 2003. Estimating good discrete partitions from observed data: Symbolic false nearest neighbors. Phys. Rev. Lett. 91 (8), 084102.
Kim, K.T., Lee, J.G., Quay, B.D., Santavicca, D.A., 2010. Response of partially premixed flames to acoustic velocity and equivalence ratio perturbations. Combust. Flame 157 (9), 1731–1744.
Kumar, S., Pecht, M., 2007. Health monitoring of electronic products using symbolic time series analysis. In: AAAI Fall Symposium on Artificial Intelligence for Prognostics. pp. 73–80.
Laurence, B., 1993. A complete sufficient statistic for finite-state Markov processes with application to source coding. IEEE Trans. Inf. Theory 39 (3), 1047.
Lin, J., Keogh, E., Wei, L., Lonardi, S., 2007. Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Discov. 15 (2), 107–144.
Lind, D., Marcus, B., 1995. An Introduction to Symbolic Dynamics and Coding. Cambridge University Press.
Meyn, S.P., Tweedie, R.L., 2012. Markov Chains and Stochastic Stability. Springer Science & Business Media.
Mukherjee, K., Gupta, S., Ray, A., Phoha, S., 2011. Symbolic analysis of sonar data for underwater target detection. IEEE J. Ocean. Eng. 36 (2), 219–230.
Mukherjee, K., Ray, A., 2014. State splitting and merging in probabilistic finite state automata for signal representation and analysis. Signal Process. 104, 105–119.
O'Connor, J., Acharya, V., Lieuwen, T., 2015. Transverse combustion instabilities: Acoustic, fluid mechanic, and flame processes. Prog. Energy Combust. Sci. 49, 1–39.
Papoulis, A., Pillai, S.U., 2002. Probability, Random Variables, and Stochastic Processes. Tata McGraw-Hill Education.
Poor, H.V., 1994. An Introduction to Signal Detection and Estimation, second ed. Springer, New York, NY.
Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.
Rajagopalan, V., Ray, A., 2006. Symbolic time series analysis via wavelet-based partitioning. Signal Process. 86 (11), 3309–3320.
Ray, A., 2004. Symbolic dynamic analysis of complex systems for anomaly detection. Signal Process. 84 (7), 1115–1130.
Sarkar, S., Srivastav, A., Shashanka, M., 2013. Maximally bijective discretization for data-driven modeling of complex systems. In: American Control Conference. Washington, D.C., USA, pp. 2680–2685.
Shiryaev, A.N., 1978. Optimal Stopping Rules. Springer-Verlag, New York, NY.
Srivastav, A., 2014. Estimating the size of temporal memory for symbolic analysis of time-series data. In: American Control Conference (ACC), 2014. IEEE, pp. 1126–1131.
Tantaratana, S., Thomas, J.B., 1977. Truncated sequential probability ratio test. Inform. Sci. 13 (3), 283–300.
Vidal, E., Thollard, F., De La Higuera, C., Casacuberta, F., Carrasco, R.C., 2005. Probabilistic finite-state machines—Part I. IEEE Trans. Pattern Anal. Mach. Intell. 27 (7), 1013–1025.
Virani, N., Lee, J.-W., Phoha, S., Ray, A., 2016. Information-space partitioning and symbolization of multi-dimensional time-series data using density estimation. In: American Control Conference (ACC), 2016. IEEE, pp. 3328–3333.
Virani, N., Lee, J.-W., Phoha, S., Ray, A., 2016. Information-space partitioning and symbolization of multi-dimensional time-series data using density estimation. In: American Control Conference (ACC), 2016. IEEE, pp. 3328–3333.
Virani, N., Marcks, S., Sarkar, S., Mukherjee, K., Ray, A., Phoha, S., 2013. Dynamic data driven sensor array fusion for target detection and classification. Procedia Comput. Sci. 18, 2046–2055.
Wald, A., Wolfowitz, J., 1948. Optimum character of the sequential probability ratio test. Ann. Math. Stat. 326–339.
Wen, Y., Mukherjee, K., Ray, A., 2013. Adaptive pattern classification for symbolic dynamic systems. Signal Process. 93 (1), 252–260.
Zhang, L., Boukas, E.-K., Baron, L., Karimi, H.R., 2010. Fault detection for discrete-time *m*arkov jump linear systems with partially known transition probabilities. Internat. J. Control 83 (8), 1564–1572.