Learning Discrete Time Markov Chains under Concept Drift

Manuel Roveri, *Senior Member, IEEE*

***Abstract*—Learning under concept drift is a novel and promising research area aiming at designing learning algorithms able to deal with nonstationary data-generating processes. In this research field, most of the literature focuses on learning nonstationary probabilistic frameworks, while some extensions about learning graphs and signals under concept drift exist. For the first time in the literature, this paper addresses the problem of learning Discrete-Time Markov Chains (DTMCs) under concept drift. More specifically, following a hybrid active/passive approach, this work introduces both a family of change-detection mechanisms (differing in the required assumptions and performance) for detecting changes in DTMCs and an adaptive learning algorithm able to deal with DTMCs under concept drift. The effectiveness of both the proposed change detection mechanisms and the adaptive learning algorithm has been extensively tested on synthetically-generated experiments and real datasets.**

***Index Terms*—Concept drift, learning in nonstationary environments, discrete time Markov chains, change detection mechanisms, adaptation.**

I. Introduction

In the recent years the research interest about learning under concept drift is significantly increased leading to a wide range of machine learning solutions able to deal with nonstationary learning problems [1]–[4]. Such solutions allow to weaken the stationary hypothesis on the process generating the data, which is generally implicitly or explicitly assumed in traditional machine-learning techniques [5]. In this way, machine learning solutions meant to operate in nonstationary environments are able to learn from data-generating processes that evolve over time due to variations in the environment in which a system is operating (e.g., seasonality or periodicity, ageing effects), changes in the interaction between the environment and the system (e.g., cyber-attacks or changes in the users' habits) or faults/malfunctioning affecting the system [6].

The literature about learning under concept drift is very wide and several families of solutions exist. Such solutions differ in the considered approach (e.g., active vs. passive), encompassed learning mechanism (e.g., single vs. ensemble solutions), and required assumptions (e.g., abrupt changes vs. drift) [3], [4]. Despite the heterogeneity of these solutions, most of the research focused on probabilistic frameworks (e.g., regression or classification) under concept drift. In this scenario data are modelled as random variables and concept drift refers to changes in the posterior probability or the marginal distribution [2], [4]. Extensions to such a probabilistic framework have been proposed in the field of learning nonstationary signals [7] or graph representations under concept drift [8].

For the first time in the literature, this paper focuses on the learning of Discrete Time Markov Chains (DTMCs) under concept drift. DTMCs are stochastic models describing data-generating processes, characterized by a discrete set of states and discrete time, following the Markov property [9], [10]. DTMCs have been extensively studied for decades [9], [11] and represent the theoretical basis of a wide range of real-world applications and tools (e.g., web search engines, natural language recognition and hidden Markov models). The change of state in a DTMC is called *transition* and the probability of moving from one state to another is called *transition probability*. Typically, the transition probabilities of DTMCs are assumed to be known or estimated from data [11], [12]. Such transition probabilities are time-independent as in *homogeneous* DTMCs or time-dependent as in *non-homogeneous* DTMCs (where the transition probabilities evolve over time according to a fixed law). Concept drift could affect both time-independent and time-dependent transition probabilities leading to a variation in the transition probabilities in case of homogeneous DTMCs or to a change in the time-dependency characterizing the transition probabilities in case of non-homogeneous ones. In order to react and adapt to such concept drift, the transition probabilities of DTMCs must be adapted over time following a learning-under-concept-drift approach [3], [4].

In this paper we focus on homogeneous DTMCs[1] and we introduce a family of change-detection mechanisms and an adaptive algorithm, called "ADaptive Algorithm for Markov chains" (*ADAM*), for learning DTMCs under concept drift.

The proposed change detection mechanisms aim at sequentially analyzing observations coming from the data-generating process looking for changes in the associated DTMCs [4]. Inspired by the well-known and theoretically-grounded CUmulative SUM (CUSUM) test [13], three versions of the change-detection mechanism for DTMCs are here introduced. These three versions differ in the a-priori knowledge they require to operate and performance. The first version, called "parameteric", relies on the knowledge of the transition probabilities of the DTMC before and after the change. Asymptotic properties for this parametric change-detection mechanism are derived, i.e., the Average Run Length to a false positive detection (ARL_0) and to a correct detection (ARL_1). The second one, called "non-parametric", does not require any a-priori knowledge about the DTMC before or after the

The author is with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy, e-mail: manuel.roveri@polimi.it

[1]The proposed solutions could be extended to the case on non-homogeneous DTMCs by modelling concept drift as a change in the law that drives the evolution of transition probabilities over time.

change. For this version, an approximated asymptotic ARL_0 is derived. In addition, following the approach proposed in [14], a hierarchical version of the non-parametric change-detection mechanism is introduced to improve the trade-off between false-positive detections and detection delay.

The proposed algorithm for the learning of homogeneous DTMCs under concept drift follows a hybrid active-passive approach [15]: the DTMC is adapted at each observation gathered from the data-generating process (as in passive approaches), while a change-detection mechanism is used to trigger the retraining of the DTMC when needed (as in active approaches). The core of *ADAM* is the joint use of a change-detection mechanism monitoring the stationarity of the data-generating process and an adaptive window over the recently acquired observations to estimate the transition probabilities of the DTMC over time. In stationary conditions, thanks to a change-detection index provided by the change-detection mechanism, such an adaptive window is enlarged to improve the transition-probability estimation of the DTMC over time. In nonstationary conditions, the change-detection index drives the reduction of the adaptive window to react to a possible concept drift. When a change is detected, the DTMC is retrained as in active approaches [16] on the new state of the data-generating process on an adaptive window of recently-acquired observations. The length of this adaptive window is automatically defined by means of a novel procedure to estimate the time instant a concept drift affected a DTMC.

The novel contributions of the paper can be summarized as:

- a hybrid active-passive adaptive algorithm called *ADAM* for the learning of DTMCs under concept drift;
- three different change-detection mechanisms for detecting changes in DTMCs differing in the required a-priori knowledge about the data-generating process they require to operate and the provided trade-off between false positive detections and detection delay;
- a change-detection index aiming at measuring the stationarity of the data-generating process and triggering the adaptation of DTMCs under concept drift;
- a procedure to estimate the time instant a concept drift affected a DTMC.

Both the proposed family of change-detection mechanisms and the *ADAM* framework are made available to the scientific community as a Matlab toolbox[2].

The effectiveness of what proposed has been tested on a wide synthetic experimental campaign and two real datasets, i.e., a dataset from the Australia New South West (ANSW) electricity market for electricity demand prediction and a dataset from the National Oceanic and Atmospheric Administration (NOAA) about annual hurricane rates for understanding global climate processes.

This work is organized as follows. Section II describes the related literature. Section III formulates the problem of learning DTMCs under concept drift. The parametric, non-parametric and hierarchical change-detection mechanisms are described in Section IV, while the proposed adaptive algorithm *ADAM* for learning DTMCs under concept drift is detailed in Section V. Experimental results are given in Section VI and conclusions are finally drawn in Section VII.

[2]The toolbox can be downloaded from IEEE Code Ocean from the following link https://codeocean.com/2018/12/06/adaptive-algorithm-for-markov-chains/code

II. Related literature

Estimating the unknown transition matrix of a DTMC from the observations generated by a stochastic process has been extensively studied in the literature [9] [11] [12] [17]. These solutions are based on the maximum-likelihood principle and generally rely on counting the times the stochastic process moves from one state to another. To achieve this goal, one or more sequences of observations can be used [11] and consistency properties and bounds have been derived for both homogeneous and non-homogenous DTMCs [9] [11] [10] [18] [19]. As stated in Section I, concept drift could affect both types of DTMCs by breaking the stationary assumption in the homogeneous case and the time-invariance of the stochastic process in the non-homogenous one. Interestingly, the problem of learning DTMCs in presence of concept drift has been rarely addressed in the literature and only few application-specific solutions exist. For example, [20] introduces a discrete-time Markov model aiming at investigating treatment-intervention and death in patients affected by diabetic retinopathy. Here, concept drift is explicitly introduced by combining two Markov chains in the considered stochastic process to model the progression of the disease over time.

A relatively larger literature about detecting changes in DTMCs exists. More specifically, the problem of detecting changes in Markov chains has been initially defined in [21] under a Bayesian formulation of geometric priors about the concept-drift time instant. That work introduces an optimal detection scheme for DTMCs based on the Shyrayev-Robert formulation [22] under the assumption of a-priori knowing the distribution of the concept-drift time instant as well as the parameters of the transition matrix before and after the change. Differently, [23] introduces a non-Bayesian framework for change detection in DTMC. This framework does not require any a-priori knowledge about the change-time distribution but assumes specific dependency structures of the transition matrices of DTMCs (i.e., symmetric variations of the transition probabilities). Even in this case, the DTMCs before and after the changes are assumed to be known. An interesting approach is proposed in [24] for detecting changes in hidden Markov models (HMMs). The change-detection mechanism is reformulated as a sequential probability ratio-test [25], whose log-likelihood ratio mechanism has been approximated to take into account the fact that processed data are not independent and identically distributed. HMMs are assumed to be known before and after the change to compute the approximated log-likelihood ratio. Nonstationary Markov models have been also introduced in the literature. For example, [26] proposes a nonstationary extension of HMMs to deal with time-varying transition-probability parameters. Similarly, [27] introduces HMMs able to model time-varying state durations. These models represent extensions of traditional HMMs but, unfortunately, they do not provide mechanisms for detecting changes in the associated data-generating process.

The problem of detecting changes in Markov chains has been also addressed in application-specific scenarios. For example, [28] reformulates the problem of change detection into a change-point analysis aiming at detecting the presence of a change-point into a fixed-length sequence of data. This approach, which is not truly sequential, has been applied to detect shifts in hurricane rates. Similarly, [29] proposes a video-segmentation mechanism based on Markov chains and change-point analysis. The proposed mechanism relies on a Bayesian framework, while the segmentation assumes the a-priori knowledge of the distribution of the number of scenes within the video. Differently, [30] introduces a sequential mechanism based on DTMCs for intrusion detection in computer and network systems. This mechanism relies on the estimation of a DTMC modelling the nominal behavior of the computer/network system by means of an "intrusion-free" training sequence. An intrusion is detected when the trained DTMC is no more able to explain the recently-acquired data (through the analysis of the likelihood). Similarly, [31] introduces a sequential mechanism based on DTMCs for detecting "unusual" human behaviors in intelligent houses. The "usual" behavior is modelled through the learning of a DTMC on a training sequence, while the change-detection phase relies on the analysis of the likelihood computed on acquired data.

Interestingly, a relatively wide literature about learning classification and regression models under concept exists [3] [4]. Examples of these families of solutions are the Adaptive windowing algorithms [32] [33], the Just-In-Time Adaptive Classifiers [34] [35], and the Ensemble-based Algorithms [36] [37] [38]. In this research field the literature about change detection mechanisms for detecting changes in random variables is large and well established [4] [39]. Relevant and well known examples of change-detection mechanisms are Drift Detection Method (DDM) [32], Early Drift Detection Method (EDDM) [40] and Exponentially Weighted Moving Average (EWMA) [41], while other interesting change-detection mechanisms can be found in [42], [43], and [44], just to name a few. We emphasize that these solutions are meant to operate in a probabilistic framework, hence they cannot be directly applied to the scenario of DTMCs under concept drift.

Summarizing, for the first time in the literature, this paper introduces an adaptive algorithm for the learning of DTMCs under concept drift as well as three different mechanisms (differing in the a-priori knowledge they require to operate and the trade-off between detection delays and false positive detections) for detecting changes in DTMCs. In addition, this paper introduces a procedure to estimate the time instant a concept drift affected a DTMC, a precious information to support the adaptation of DTMCs over time.

III. Problem Formulation

Let $\mathcal{P}$ be a data-generating process generating a sequence of observations $\mathcal{T} = \{s_1, s_2, \ldots, s_t, \ldots, s_T\}$ over discrete time instants $t = 1, 2, \ldots, T$. The time-horizon T could be finite, i.e., $T < +\infty$, or infinite, i.e., $T = +\infty$.

Each observation s_t belongs to a finite state space, i.e., $s_t \in \Omega = \{\omega_1, \ldots, \omega_N\}$ being N the finite number of states. We assume that Ω does not change over time.

We also assume that $\mathcal{P}$ can be modelled as a DTMC $\Theta = \{\pi, P\}$, where π is the initial distribution of the states and P the transition matrix. We model the concept drift in $\mathcal{P}$ as an abrupt change in the transition matrix P:

$$P = \begin{cases} P_0 & t < t^* \\ P_1 & t \geq t^* \end{cases}, \tag{1}$$

where P_0 refers to the transition matrix of $\mathcal{P}$ before the change ($t < t^*$),

$$P_0 = \begin{bmatrix} p^0_{1,1} & p^0_{1,2} & \cdots & p^0_{1,N} \\ p^0_{2,1} & p^0_{2,2} & \cdots & p^0_{2,N} \\ \vdots & \vdots & & \vdots \\ p^0_{N,1} & p^0_{N,2} & \cdots & p^0_{N,N} \end{bmatrix},$$

P_1 refers to the transition matrix of $\mathcal{P}$ after the change ($t \geq t^*$),

$$P_1 = \begin{bmatrix} p^1_{1,1} & p^1_{1,2} & \cdots & p^1_{1,N} \\ p^1_{2,1} & p^1_{2,2} & \cdots & p^1_{2,N} \\ \vdots & \vdots & & \vdots \\ p^1_{N,1} & p^1_{N,2} & \cdots & p^1_{N,N} \end{bmatrix},$$

being $p^0_{i,j}$ and $p^1_{i,j}$ the probability to move from state ω_i to ω_j before and after the change, respectively and $t^* \leq T$ refers to the time instant the concept drift occurs (the case where $t^* = T$ refers to a stationary DTMC in the considered time horizon). We emphasize that t^* in Eq. (1) is a-priori unknown.

Since concept drift refers to a change in the transition matrix P, the data-generating process $\mathcal{P}$ before and after the concept drift is defined as $\Theta_0 = \{\pi, P_0\}$ and $\Theta_1 = \{\pi, P_1\}$, respectively.

We assume that the first L observations $TS = \{s_1, s_2, \ldots, s_L\}$ of $\mathcal{T}$ have been generated in stationary conditions, i.e., $L < t^*$. This is reasonable since concept drift generally occur with a large time constant, hence not affecting $\mathcal{P}$ in the early stages of operation[3].

The aim of the proposed change-detection mechanisms and *ADAM* is to detect changes and learn DTMCs under concept drift defined as in Eq. (1).

We emphasize that the solutions described in this paper could be easily extended to the case of *drift changes*, where P_1 is time-dependent whose probability $p^1_{i,j}$s slowly vary over time for $t \geq t^*$. In fact, the three proposed change-detection mechanisms are already ready to detect this type of changes, while, to be effective in case of drift changes, the proposed *ADAM* should be endowed with a non-homogeneous learning mechanism since the DTMC is time-dependent after t^*.

IV. The proposed Change-Detection Mechanisms: parametric, non-parametric and hierarchical

The goal of the proposed parameteric, non-parameteric and hierarchical change-detection mechanisms is to promptly and effectively detect concept drift, as defined in Eq. (1), affecting DTMCs. These three change-detection mechanisms differ in

[3]For example, this reflects the scenario where historical data are available to researches and practitioners to initially estimate the transition matrix of the DTMC.

the amount of a-priori knowledge they require to operate and the trade-off between detection promptness and false positive detection. More specifically, the parameteric change-detection mechanism assumes the knowledge of P_0 and P_1 to operate, while the non-parametric and the hierarchical ones do not. In particular, the hierarchical change-detection mechanism extends the non-parameteric change-detection mechanism by introducing a validation layer to reduce false positive detections.

Besides being stand-alone tools for the analysis of DTMCs, the proposed change detection mechanisms play a crucial role in triggering the adaptation phase in the proposed *ADAM*, as detailed in Section V.

ALGORITHM 1: The parametric change-detection mechanism *P-CDM* for detecting changes in DTMCs.

Input: $\mathcal{T}$, $\Theta_0 = \{\pi_0, P_0\}$, $\Theta_1 = \{\pi_1, P_1\}$ **and** K ;
Compute $\widetilde{\pi_0}$ and $\widetilde{\pi_1}$;
$m_0 = 0$;
while *(w_i is available)* **do**
  Compute $\widetilde{P}_{\Theta_1}(w_i)$ and $\widetilde{P}_{\Theta_0}(w_i)$ as in Eq. (7) and (8);
  $\widetilde{l_i} = log\left(\widetilde{P}_{\Theta_1}(w_i)/\widetilde{P}_{\Theta_0}(w_i)\right)$;
  $m_i = max\left(0, m_{i-1} + sign\left(\widetilde{l_i}\right)\right)$;
  if *($m_i \geq K$)* **then**
    Change detection in the i-th subsequence w_i;
  end
end

A. The parametric change-detection mechanism: algorithm, ARL_0 and ARL_1

This section details the proposed parametric mechanism, called *P-CDM*, for detecting changes in DTMCs. The proposed mechanism operates by sequentially analysing observation in $\mathcal{T}$ inspecting for changes defined in Eq. (1). More specifically, the mechanism operates on non-overlapping subsequences of length W of $\mathcal{T}$, defined as

$$w_i = \{s_{W(i-1)+1}, \ldots, s_{Wi}\} \tag{2}$$

where w_i is the i-th subsequence of $\mathcal{T}$. Inspired by the CUSUM approach [13], the core of the proposed parametric change-detection mechanism is the computation of the log-likelihood ratio

$$l_i = log\left(\frac{\mathbb{P}_{\Theta_1}(w_i)}{\mathbb{P}_{\Theta_0}(w_i)}\right) \tag{3}$$

where $\mathbb{P}_{\Theta_1}(w_i)$ and $\mathbb{P}_{\Theta_0}(w_i)$ represent the probability that w_i is generated by Θ_1 and Θ_0, respectively. Following the parametric approach, here Θ_1 and Θ_0 are assumed to be known. $\mathbb{P}_{\Theta_1}(w_i)$ and $\mathbb{P}_{\Theta_0}(w_i)$ are defined as follows [9]:

$$\mathbb{P}_{\Theta_1}(w_i) = \pi_1^{W(i-1)+1}(s_{W(i-1)+1}) \prod_{j=W(i-1)+1}^{Wi-1} p^1_{s_j,s_{j+1}} \tag{4}$$

and

$$\mathbb{P}_{\Theta_0}(w_i) = \pi_0^{W(i-1)+1}(s_{W(i-1)+1}) \prod_{j=W(i-1)+1}^{Wi-1} p^0_{s_j,s_{j+1}} \tag{5}$$

where $\pi_1^t(s_t)$ and $\pi_0^t(s_t)$ represent the probability of being in state s_t at time t by Θ_1 and Θ_0, respectively, while $p^1_{\omega_i,\omega_j}$ and $p^0_{\omega_i,\omega_j}$ are the transition probability from state ω_i to ω_j of Θ_1 and Θ_0, respectively.

Since we are interested in changes in the transition matrix P, we approximate $\pi_1^t(\bullet)$ and $\pi_0^t(\bullet)$ with the asymptotic distributions of the states $\widetilde{\pi_1}(\bullet)$ and $\widetilde{\pi_0}(\bullet)$ that can be easily computed from P_0 and P_1 [9]. In this way we are removing the dependency from the initial state distribution by assuming that enough time passed to achieve the stationary state of the DMTC [9]. This assumption is in line with the fact that changes rarely occur in the early stages of $\mathcal{P}$ (as commented in Section III). Hence, we can rewrite Eq. (3) as

$$\widetilde{l_i} = log\left(\frac{\widetilde{\mathbb{P}}_{\Theta_1}(w_i)}{\widetilde{\mathbb{P}}_{\Theta_0}(w_i)}\right) \tag{6}$$

being

$$\widetilde{\mathbb{P}}_{\Theta_1}(w_i) = \widetilde{\pi_1}(s_{W(i-1)+1}) \prod_{j=W(i-1)+1}^{Wi-1} p^1_{s_j,s_{j+1}} \tag{7}$$

and

$$\widetilde{\mathbb{P}}_{\Theta_0}(w_i) = \widetilde{\pi_0}(s_{W(i-1)+1}) \prod_{j=W(i-1)+1}^{Wi-1} p^0_{s_j,s_{j+1}}. \tag{8}$$

In order to support the sequential analysis of $\mathcal{T}$, we define the following figure of merit

$$m_i = max\left(0, m_{i-1} + sign\left(\widetilde{l_i}\right)\right) \tag{9}$$

being $sign(\bullet)$ the sign function and $m_0 = 0$. A change is detected in the i-th subsequence w_i of $\mathcal{T}$ when

$$m_i \geq K \tag{10}$$

being $K \in \mathbb{N}^+$ a user-defined parameter. The algorithm of the proposed *P-CDM* for detecting changes in DTMCs is detailed in Algorithm 1.

The choice of K is critical to balance the trade-off between false positives and detection delay. For this reason, the rest of this subsection is devoted to analyse the performance of the proposed mechanism in terms of the average time to a false positive detection ARL_0 and to a correct change detection ARL_1 w.r.t. K.

More specifically, given W, the set $\mathcal{U} = \{u_1, \ldots, u_{|\mathcal{U}|}\}$ of all the possible state sequences of length W is finite. The cardinality $|\mathcal{U}|$ of $\mathcal{U}$ is the total number of permutations with repetitions of N states over a sequence of length W, i.e., $|\mathcal{U}| = N^W$.

To compute the ARL_0, which is the average time to a false positive detection, we assume that $t^* = +\infty$. Hence, the whole $\mathcal{T}$ is generated by Θ_0. The probability $q_j^{\Theta_0|\Theta_0}$ that the j-th state sequence $u_j \in \mathcal{U}$, with $j = 1, \ldots, |\mathcal{U}|$, has been generated by Θ_0 given the fact that it has been generated by Θ_0 is defined as

$$q_j^{\Theta_0|\Theta_0} = \widetilde{\mathbb{P}}_{\Theta_0}(u_j).$$

Similarly, we can define

$$q_j^{\Theta_1|\Theta_0} = \widetilde{\mathbb{P}}_{\Theta_1}(u_j).$$

as the probability that u_j has been generated by Θ_1 given the fact that it has been generated by Θ_0. Obviously $\bigcup_{j=1}^{|\mathcal{U}|} u_i = \mathcal{U}$ and $\sum_{j=1}^{|\mathcal{U}|} q_j^{\Theta_0|\Theta_0} = 1$.

Then, $\mathcal{U}$ is partitioned into two subsets $\{\mathcal{U}_0, \mathcal{U}_1\}$ as follows

$$\begin{cases} \mathcal{U}_1 = \{u_j \text{ s.t. } q_j^{\Theta_1|\Theta_0} > q_j^{\Theta_0|\Theta_0}, \quad i = 1, \dots, |\mathcal{U}|\} \\ \mathcal{U}_0 = \mathcal{U} \diagdown \mathcal{U}_1 \end{cases}$$

where $\mathcal{U}_1$ contains all the state sequences of $\mathcal{U}$ that are more likely to be generated by Θ_1 than Θ_0 and $\mathcal{U}_0$ is the complement set of $\mathcal{U}_1$ w.r.t. $\mathcal{U}$. We can now define

$$Q_0^{\Theta_1} = \sum_{v=1}^{|\mathcal{U}_1|} q_v^{\Theta_1|\Theta_0}$$

being u_v the v-th element of $\mathcal{U}_1$. $Q_0^{\Theta_1}$ represents the probability of generating a state sequence of length W by Θ_0 that is more likely to be generated by Θ_1 than Θ_0. Similarly, we can define $Q_0^{\Theta_0}$ as the probability of generating a state sequence of length W by Θ_0 that is more likely to be generated by Θ_0 than Θ_1 (or where the probability is equal). Obviously $Q_0^{\Theta_0} + Q_0^{\Theta_1} = 1$.

We can now compute the ARL_0 as follows:

Theorem 1: Let $\bar{t}$ be the detection time of the proposed change-detection mechanism, $t^* = +\infty$ and $Q_0^{\Theta_1} > 0$,

$$ARL_0 = \mathbb{E}_{\mathcal{T}}[\bar{t}] = \underline{u}(I - P_Z^0)^{-1}\underline{1} \tag{11}$$

where I is the $(K+1) \times (K+1)$ identity matrix, P_Z^0 is the $(K+1) \times (K+1)$ matrix defined as follows

$$P_Z^0 = \begin{bmatrix} 1 - Q_0^{\Theta_1} & Q_0^{\Theta_1} & 0 & \dots & 0 \\ 1 - Q_0^{\Theta_1} & 0 & Q_0^{\Theta_1} & \dots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix},$$

$\underline{1}$ is the $(K+1)$-dimensional vector of ones, and $\underline{u}$ is the $(K+1)$-dimensional vector defined as $\underline{u} = [1, 0, \dots, 0]$.

Proof: The demonstration is based on the fact that the detection mechanism in Eq. (10) can be modelled as a discrete time birth-death Markov chain with $K+1$ states $\{0, 1, \dots, K\}$ defined by the following $(K+1) \times (K+1)$ transition matrix

$$P_{BD} = \begin{bmatrix} 1 - Q_0^{\Theta_1} & Q_0^{\Theta_1} & 0 & \dots & 0 \\ 1 - Q_0^{\Theta_1} & 0 & Q_0^{\Theta_1} & \dots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & \dots & 1 - Q_0^{\Theta_1} & 0 & Q_0^{\Theta_1} \\ 0 & \dots & 0 & 1 - Q_0^{\Theta_1} & Q_0^{\Theta_1} \end{bmatrix},$$

while the corresponding initial distribution vector is the $(K+1)$-dimensional vector defined as $[1, 0, \dots, 0]$.

A detection occurs when the state K is achieved starting from state 0. Given this formalization we can resort on the theory of DTMCs to compute $\underline{\mu}$ that is the vector of the mean first-time passages from the states $\{0, 1, \dots, K\}$ to the state K. $\underline{\mu}$ is computed by solving the following equation

$$(I - P_Z^0)\underline{\mu} = \underline{1}. \tag{12}$$

The first element of $\underline{\mu}$ represents the ARL_0. ■

We can similarly define ARL_1 as the average time to the first detection when $t^* = 0$, i.e., the whole $\mathcal{T}$ is generated by Θ_1. Even in this case we can compute

$$q_j^{\Theta_0|\Theta_1} = \widetilde{\mathbb{P}}_{\Theta_0}(u_j) \tag{13}$$

and

$$q_j^{\Theta_1|\Theta_1} = \widetilde{\mathbb{P}}_{\Theta_1}(u_j). \tag{14}$$

Then, we partition $\mathcal{U}$ as follows

$$\begin{cases} \mathcal{U}_1 = \{u_j \text{ s.t. } q_j^{\Theta_1|\Theta_1} > q_j^{\Theta_0|\Theta_1}, \quad i = 1, \dots, |\mathcal{U}|\} \\ \mathcal{U}_0 = \mathcal{U} \diagdown \mathcal{U}_1 \end{cases} \tag{15}$$

and we can compute

$$Q_1^{\Theta_1} = \sum_{v=1}^{|\mathcal{U}_1|} q_v^{\Theta_1|\Theta_1} \tag{16}$$

that is the probability of generating a state sequence of length W by Θ_1 that is more likely to be generated by Θ_1 than Θ_0

We can now compute the ARL_1 as follows:

Lemma 2: Let $\bar{t}$ be the detection time of the proposed change-detection mechanism, $t^* = 0$ and $Q_1^{\Theta_1} > 0$,

$$ARL_1 = \mathbb{E}_{\mathcal{T}}[\bar{t}] = \underline{u}(I - P_Z^1)^{-1}\underline{1} \tag{17}$$

where I is the $(K+1) \times (K+1)$ identity matrix, P_Z^1 is the $(K+1) \times (K+1)$ matrix defined as follows

$$P_Z^1 = \begin{bmatrix} 1 - Q_1^{\Theta_1} & Q_1^{\Theta_1} & 0 & \dots & 0 \\ 1 - Q_1^{\Theta_1} & 0 & Q_1^{\Theta_1} & \dots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \tag{18}$$

and $\underline{1}$ is the $(K+1)$-dimensional vector of ones, while $\underline{u}$ is the $(K+1)$-dimensional vector defined as $\underline{u} = [1, 0, \dots, 0]$.

Proof: The demonstration relies on the same procedure used for *Theorem 1* and is omitted for brevity. ■

B. The non-parametric change-detection mechanism

The parametric change-detection mechanism described above assumes the a-priori knowledge of Θ_0 and Θ_1. Unfortunately, in real-world conditions, this assumption rarely holds. To overcome this limitation, non-parametric solutions should be considered [16]. In this section, we present the non-parametric extension of the parametric change-detection mechanism described in Section IV-A.

The core of the proposed non-parametric change detection mechanism, called *NP-CDM*, is that, being not a-priori known, Θ_0 and Θ_1 are estimated from data.

More specifically, under the assumption that the first L samples of $\mathcal{T}$ have been generated in stationary conditions, $TS = \{s_1, \dots, s_L\}$ is used to compute an estimate $\widetilde{\Theta}_0$ of Θ_0. Several techniques are available for this purpose and we opted

ALGORITHM 2: The non-parametric change-detection mechanism *NP-CDM* for detecting changes in DTMCs.

Input: $\mathcal{T}$, L **and** K ;
Estimate $\widetilde{\Theta}_0$ and $\widetilde{\Theta}_1$ on $TS = \{s_1, \ldots, s_L\}$;
$i = 0$, $\widetilde{m}_i = 0$;
while *(s_t is available)* **do**
  if $(mod(t, W) == 0)$ **then**
    $i = i + 1$;
    $w_i = \{s_{t-W+1}, \ldots, s_t\}$;
    Compute $\widetilde{P}_{\widetilde{\Theta}_1}(w_i)$ and $\widetilde{P}_{\widetilde{\Theta}_0}(w_i)$ as described in Section IV-B;
    $l_i^{np} = log\left(\widetilde{P}_{\widetilde{\Theta}_1}(w_i)/\widetilde{P}_{\widetilde{\Theta}_0}(w_i)\right)$;
    $\widetilde{m}_i = max\left(0, \widetilde{m}_{i-1} + sign\left(l_i^{np}\right)\right)$;
    if *($\widetilde{m}_i \geq K$)* **then**
      Change detection in the i-th subsequence w_i;
    end
  end
  Estimate $\widetilde{\Theta}_1$ on $\{s_{t-L+1}, \ldots, s_t\}$;
end

for the statistical procedure based on *maximum likelihood* described in [9]. Similarly, to compute an estimate $\widetilde{\Theta}_1$ of Θ_1, the *NP-CDM* relies on a sliding window of length L over the latest acquired observations from $\mathcal{P}$. Hence, at time t, $\widetilde{\Theta}_1$ is estimated from the state sub-sequence $\{s_{t-L+1}, \ldots, s_t\}$.

Similarly to its parametric version, the *NP-CDM* operates by analysing non-overlapping sequences of length W defined in Eq. (2) and where Θ_0 and Θ_1 are approximated with $\widetilde{\Theta}_0$ and $\widetilde{\Theta}_1$. In particular, the proposed *NP-CDM* approximates the log-likelihood ratio defined in Eq. (6) with

$$l_i^{np} = log\left(\frac{\widetilde{\mathbb{P}}_{\widetilde{\Theta}_1}(w_i)}{\widetilde{\mathbb{P}}_{\widetilde{\Theta}_0}(w_i)}\right) \tag{19}$$

where $\widetilde{\mathbb{P}}_{\widetilde{\Theta}_1}(w_i)$ and $\widetilde{\mathbb{P}}_{\widetilde{\Theta}_0}(w_i)$ represent the probability that w_i is generated by $\widetilde{\Theta}_1$ and $\widetilde{\Theta}_0$, respectively. $\widetilde{\mathbb{P}}_{\widetilde{\Theta}_1}(w_i)$ and $\widetilde{\mathbb{P}}_{\widetilde{\Theta}_0}(w_i)$ are defined as in Eq. (7) and (8) by replacing Θ_1 and Θ_0 with $\widetilde{\Theta}_1$ and $\widetilde{\Theta}_0$, respectively.

The non-parametric sequential analysis of $\mathcal{T}$ is performed by analysing

$$\widetilde{m}_i = max\left(0, \widetilde{m}_{i-1} + sign\left(l_i^{np}\right)\right) \tag{20}$$

with $\widetilde{m}_0 = 0$. A change is detected in the i-th subsequence w_i when

$$\widetilde{m}_i \geq K. \tag{21}$$

The proposed non-parametric mechanism for detecting changes in DTMCs is detailed in Alg. 2.

The choice of K is more critical in this case since in stationary conditions, i.e., before the change, $\widetilde{\Theta}_0$ and $\widetilde{\Theta}_1$ represent two realizations of the same random variable modelling the unknown DTMC Θ_0. This could lead to a larger probability of FPs than the parametric case given the same K. This aspect is explored in the rest of the subsection.

We approximate the ARL_0^{NP} of the *NP-CDM* by assuming that, in stationary conditions, $Q_0^{\Theta_1} \approx 0.5$ meaning that, before the change, the subsequence w_i could be assigned with equal probability to $\widetilde{\Theta}_0$ or $\widetilde{\Theta}_1$. More specifically, let $t^* = 0$ and $Q_0^{\Theta_1} \approx 0.5$, the approximated ARL_0^{NP} $\bar{t}$ of the proposed *NP-CDM* is computed as

$$ARL_0^{NP} = \underline{u}(I - P_{NP}^0)^{-1}\underline{1} \tag{22}$$

where I is the $(K+1) \times (K+1)$ identity matrix, P_Z^{NP} is the $(K+1) \times (K+1)$ matrix defined as follows

$$P_{NP}^0 = \begin{bmatrix} 0.5 & 0.5 & 0 & \ldots & 0 \\ 0.5 & 0 & 0.5 & \ldots & 0 \\ \vdots & \vdots & & \vdots & \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \tag{23}$$

and $\underline{1}$ is the $(K+1)$-dimensional vector of ones, while $\underline{u}$ is the $(K+1)$-dimensional vector defined as $\underline{u} = [1, 0, \ldots, 0]$.

ALGORITHM 3: The hierarchical non-parametric change-detection mechanism *H-NPCDM* for detecting changes in DTMCs.

Input: $\mathcal{T}$, L, K, α **and** N;
Estimate $\widetilde{\Theta}_0$ and $\widetilde{\Theta}_1$ on $TS = \{s_1, \ldots, s_L\}$;
$i = 0$, $\widetilde{m}_i = 0$;
while *(s_t is available)* **do**
  if *(mod(t,W)==0)* **then**
    $i = i + 1$;
    $w_i = \{\omega_{t-W+1}, \ldots, \omega_t\}$;
    Compute $\widetilde{P}_{\widetilde{\Theta_1}}(w_i)$ and $\widetilde{P}_{\widetilde{\Theta_0}}(w_i)$ as described in Section IV-B;
    $l_i^{np} = log\left(\widetilde{P}_{\widetilde{\Theta}_1}(w_i)/\widetilde{P}_{\widetilde{\Theta}_0}(w_i)\right)$;
    $\widetilde{m}_i = max\left(0, \widetilde{m}_{i-1} + sign\left(l_i^{np}\right)\right)$;
    if *($\widetilde{m}_i \geq K$)* **then**
      $temp = 0$, $j = 1$;
      while *($j \leq N$)* **do**
        $temp = temp + \chi^2(\underline{p_j^0}, \underline{p_j^1}, \alpha/N)$;
        $j = j + 1$;
      end
      if *($temp > 0$)* **then**
        Change detection at time t;
      end
    end
  end
  Estimate $\widetilde{\Theta}_1$ on $\{s_{t-L+1}, \ldots, s_t\}$
end

C. The hierarchical non-parametric change-detection mechanism

The choice of K defines a trade-off between false positive detections and detection delay in both the parametric and non-parametric mechanism detailed above. Following the hierarchical approach for change-detection tests proposed in [14], we coupled the non-parametric change detection algorithm described in Section IV-B with an hypothesis test. This allowed us to define a two-layer hierarchical non-parametric change detection mechanism, called *H-NPCDM*, for detecting changes in DTMCs.

More specifically, the proposed *H-NPCDM* relies on the following two-layer architecture:

1) The first layer comprises the *NP-CDM* described above. When a change is detected, the first layer triggers the activation of the second layer of analysis;
2) The second layer relies on a multiple *two-sample* χ^2 *hypothesis test* on the estimated DTMCs to reduce false

positive detections. This second layer of analysis aims at confirming (or not) the change detected at the first layer by inspecting variations in the frequency distribution of the estimated DTMCs.

More specifically, the second layer operates as a multiple hypothesis test applied to the corresponding rows of the estimated transition matrices $\widetilde{P}_{\widetilde{\Theta}_0}$ and $\widetilde{P}_{\widetilde{\Theta}_1}$ of $\widetilde{\Theta}_0$ and $\widetilde{\Theta}_1$, respectively, as follows

$$\begin{cases} h_1 = \chi^2(\underline{p_1^0}, \underline{p_1^1}, \alpha/N) \\ h_2 = \chi^2(\underline{p_2^0}, \underline{p_2^1}, \alpha/N) \\ \dots \\ h_N = \chi^2(\underline{p_N^0}, \underline{p_N^1}, \alpha/N) \end{cases} \tag{24}$$

where $\chi^2(\underline{a}, \underline{b}, \gamma)$ is the two-sample χ^2 hypothesis test [45] applied to vectors $\underline{a}$ and $\underline{b}$ with confidence γ, and $\underline{p_j^0}$ and $\underline{p_j^1}$ are the j-th row of the transition matrix $\widetilde{P}_{\widetilde{\Theta}_0}$ and $\widetilde{P}_{\widetilde{\Theta}_1}$, respectively. The output of $\chi^2(\underline{a}, \underline{b}, \gamma)$ is 0 when the null hypothesis is accepted (i.e., no change in distribution between $\underline{a}$ and $\underline{b}$ is detected), and 1 when rejected. Please note that the Bonferroni correction α/N is considered in Eq. (24) to take under control false positive detections occurring in multiple hypothesis testing. The detected change is confirmed by the second layer when at least one of the hypothesis tests in Eq. (24) rejects the null hypothesis, i.e., when $\sum_{i=1}^{N} h_i > 0$.

The hierarchical non-parametric change-detection mechanism is detailed in Alg. 3.

V. The proposed Adaptive Algorithm for Learning Discrete-Time Markov chains under Concept Drift

The Adaptive Algorithm for Markov chains (*ADAM*) proposed in this paper aims at learning and tracking the evolution of the transition matrix of a DTMC under concept drift. The proposed *ADAM*, which is detailed in Alg. 4, is based on a hybrid active-passive approach [15] where the transition matrix is continuously adapted as new observations become available as in passive approaches, while the re-training of the transition matrix is triggered by the change-detection mechanism in response to concept drift as in active ones.

More specifically, *ADAM* initially estimates the transition matrix $\hat{P}$ on TS during the initial training phase. Then, during the operational life, $\hat{P}$ is updated at each new observation s_t provided by $\mathcal{P}$ by relying on an adaptive window over the recently acquired observations. The core of *ADAM* is the adaptive definition of the length L_{adapt} of such a window that is widened in stationary conditions to improve the estimation of $\hat{P}$ [9] and reduced in nonstationary ones to remove out-to-date knowledge from $\hat{P}$ and adapt it to the concept drift.

This widening/reduction mechanism, which is activated for every subsequence w_i of observations, is driven by the change-detection index $\widetilde{m}_i \in \{0, 1, \dots, K\}$ defined in Eq. (20). When $\widetilde{m}_i < K/2$, $\mathcal{P}$ can be safely associated to the stationary state and L_{adapt} can be increased. On the contrary, when $\widetilde{m}_i > K/2$, $\mathcal{P}$ could approach a concept drift and L_{adapt} is reduced to remove obsolete knowledge from $\hat{P}$.

ALGORITHM 4: The ADaptive Algorithm for Markov chains (ADAM) for learning DTMCs under concept drift.

Input: $\mathcal{T}$, L, K **and** γ ;
Estimate $\hat{P}$ on $TS = \{s_1, \dots, s_L\}$
$\widetilde{m}_i = 0$;
$L_{adapt}^t = L$;
$i = L/W$;
while *(s_t is available)* **do**
    if *(mod(t,W)==0)* **then**
        $i = i + 1$;
        $w_i = \{\omega_{t-W+1}, \dots, \omega_t\}$;
        Compute $\widetilde{m}_i$ as described in Section IV;
        $\Delta_L = -\lfloor \eta W \left(\sigma(\widetilde{m}_i - K/2) - 0.5\right) \rfloor$;
        $L_{adapt}^i = L_{adapt}^{i-1} + \Delta_L$;
        if *($\widetilde{m}_i == K$)* **then**
            $i^0 = \max_{i=1,\dots,\bar{i}}\{i \mid \widetilde{m}_i == K/2\}$;
            $t^0 = W(i^0 - 1) + 1$;
            $L_{adapt}^i = t - t^0 + 1$;
        end
    end
    Estimate $\hat{P}$ on $\{s_{t-L_{adapt}^i+1}, \dots, s_t\}$;
end

This widening/reduction mechanism is formalized through the following adaptive definition of L_{adapt}, i.e.,

$$L_{adapt}^i = L_{adapt}^{i-1} + \Delta_L \tag{25}$$

where L_{adapt}^i is the value of L_{adapt} at the i-th subwindow w_i and

$$\Delta_L = -\lfloor \eta W \left(\sigma(\widetilde{m}_i - K/2) - 0.5\right) \rfloor \tag{26}$$

being η a user-defined learning-rate parameter, $\lfloor \bullet \rfloor$ the floor function and $\sigma(\bullet)$ the log-sigmoidal function. Bounds on Δ_L can be easily defined since $\widetilde{m}_i \in \{0, 1, \dots, K\}$, hence

$$\lfloor -\gamma W \sigma(K/2) \rfloor \le \Delta_L \le \lfloor \gamma W \sigma(K/2) \rfloor \tag{27}$$

that can be approximated with

$$\lfloor -\gamma W \rfloor \le \Delta_L \le \lfloor \gamma W \rfloor \tag{28}$$

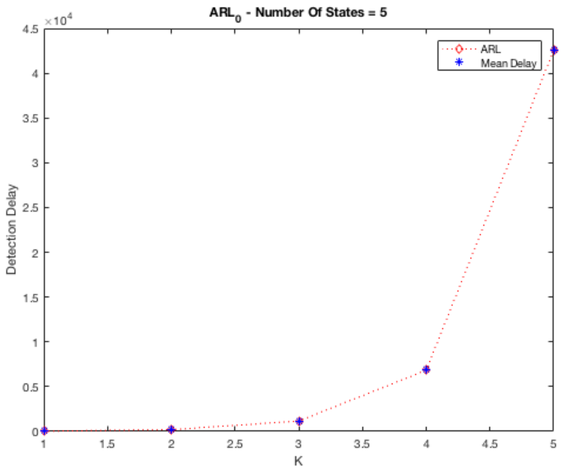
when $K \gg 2$. Hence, the widening/reduction of L_{adapt}^i is adaptive and strictly depends on how far $\widetilde{m}_i$ is from $K/2$. Given Eq. (26) and (28), Δ_L is equal to $+\lfloor \gamma W \rfloor$, 0 and $-\lfloor \gamma W \rfloor$ when $\widetilde{m}_i$ is equal to 0, $K/2$ and K, respectively. In addition, a maximum L_{MAX} and minimum L_{MIN} value (suitably defined by the user) can be set to bound L_{adapt}^i during the operational life.

When $\widetilde{m}_i == K$, the considered change-detection mechanism (i.e., the non-parameteric *NP-CDM* or the hierarchical *H-NPCDM* confirmed by the second layer of analysis) detects a change in $\mathcal{P}$. Let w_i be the subwindow where a change is detected (corresponding to time instant t equal to Wi), *ADAM* triggers the re-training of $\hat{P}$ by relying on an estimate t^0 of the time instant t^* the drift occurred. Such an estimate t^0 is computed as

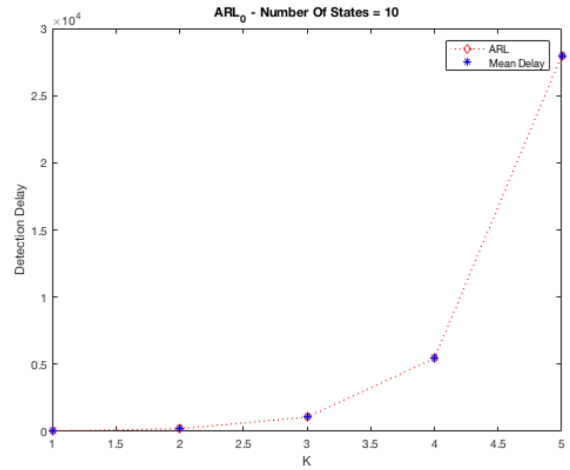
$$t^0 = W(i^0 - 1) + 1 \tag{29}$$

where

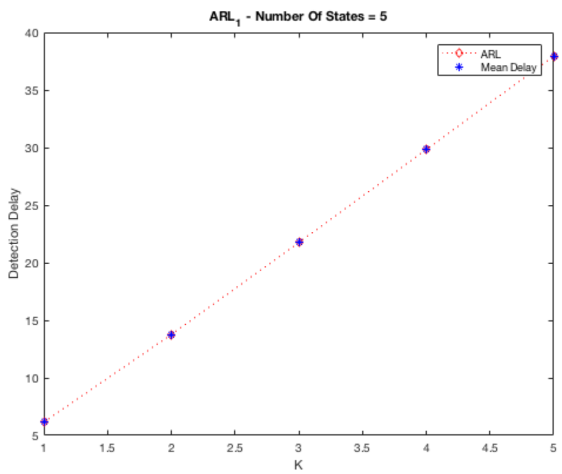
$$i^0 = \max_{j=1,\dots,i}\{j \mid \widetilde{m}_j == K/2\} \tag{30}$$



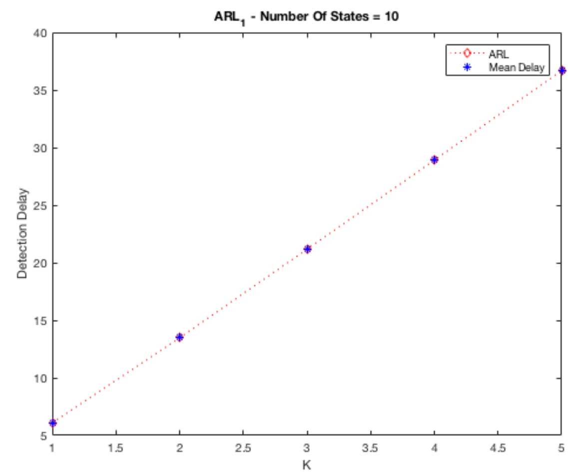
(a) ARL_0 and $N = 5$



(b) ARL_0 and $N = 10$



(c) ARL_1 and $N = 5$



(d) ARL_1 and $N = 10$

Fig. 1. The comparison between theoretical and estimated ARL_0 and ARL_1 of the proposed parametric change-detection mechanism *P-CDM* for $N = 5$ and $N = 10$ w.r.t. K

being i^0 the largest subwindow index such that $\widetilde{m}_i$ is equal to $K/2$. t^0 represents an estimate of the time instant t^* the concept drift occurred and all the observations acquired from t^0 to $t = Wi$ can be safely associated to the new state of the DTMC, following the formalization in Eq. (1). These observations are used to re-estimate $\hat{P}$, hence neglecting all the observations acquired before t^0.

In this way, the DTMC is adapted to the new state of $\mathcal{P}$ and, after that, it is ready to operate, being able to detect and adapt to further concept drift affecting $\mathcal{P}$.

The estimation of $\hat{P}$, during both the training and the operational phase, is carried out through the maximum likelihood procedure (as described in the previous section).

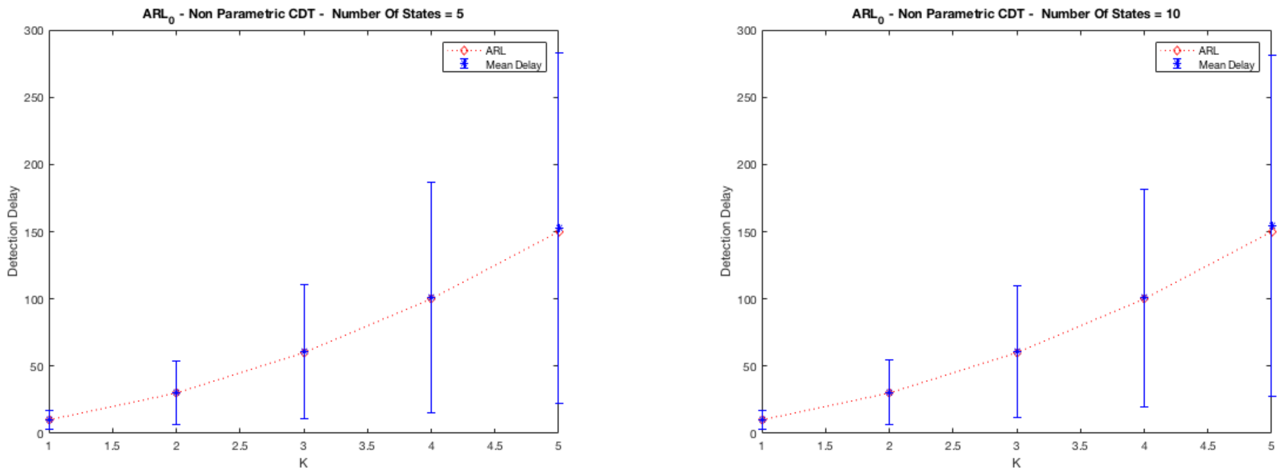
VI. Experimental Results

The experimental campaign described in this section aims at evaluating both the ability of the proposed change-detection mechanisms in correctly detecting changes in DTMCs (see Section VI-A) and the capability of the proposed *ADAM* in learning DTMCs under concept drift (see Section VI-B).

A. Evaluating the change-detection mechanisms

The ability in correctly detecting changes of the proposed change-detection mechanisms, i.e., *P-CDM*, *NP-CDM* and *H-NPCDM*, is tested through three different steps. At first, we experimentally evaluate ARL_0 and ARL_1 of *P-CDM* and ARL_0 of *NP-CDM*. Then, we experimentally show the ability of the proposed *H-NPCDM* in reducing false positive detections w.r.t. *NP-CDM*. Finally, we compare *H-NPCDM* with state-of-the-art change-detection mechanisms on both synthetic experiments and a real-world dataset.

1) Analysis of ARL_0 and ARL_1: We initially evaluated the expected ARL_0 and ARL_1 characterizing *P-CDM* as described in Section IV-A. To achieve this goal we generated $N_{couple} = 1000$ couples of DTMCs $\{\Theta_0, \Theta_1\}$ and, for each couple, $N_{seq} = 10000$ state sequences. In this set of experiments, $t^* = +\infty$ for ARL_0 and $t^* = 0$ for ARL_1. Following the parametric approach, $\{\Theta_0, \Theta_1\}$ are assumed to be known by *P-CDM*. The results are shown in Figure 1 for $N = 5$ and $N = 10$ w.r.t. K. These results corroborate the



(a) ARL_0 and $N = 5$

(b) ARL_0 and $N = 10$

Fig. 2. Approximated ARL_0 of the non-parametric change-detection mechanism *NP-CDM* with $N = 5$ and $N = 10$ w.r.t. K. The errorbar represents the standard deviation.

ability of Eq. (11) and (17) to correctly compute ARL_0 and ARL_1 for *P-CDM*. Moreover, as expected, ARL_0 and ARL_1 increase with K.

Similarly, we computed the approximated ARL_0 for the *NP-CDM*. Results are shown in Figure 2. Even in this case, the estimated ARL_0 well approximates the theoretical ARL_0 computed according to Eq. (22).

2) Comparing H-NPCDM with NP-CDM: We defined an experiment to measure the percentage of false positive (FP), false negative detections (FN) and mean detection delay (DD), measured as the mean delay between a correct detection $\bar{t} > t^*$ and t^*, for *NP-CDM* and its hierarchical extension *H-NPCDM*. These experiments have been organized by generating $N_{couple} = 1000$ couples of DTMCs $\{\Theta_0, \Theta_1\}$ and by considering $N_{seq} = 10000$ state sequences of length $T = 2000$ defined as follows

$$\mathcal{P} = \begin{cases} \Theta_0 & t < t^* = 1500 \\ \Theta_1 & t \geq t^* = 1500 \end{cases} \tag{31}$$

with $L = 1000$. W has been set to 5. Experimental results, averaged over the couples and the sequences, are shown in Table I and confirm the ability of the hierarchical approach to reduce FP detections not at the expenses of an increase in the DD.

K	NP-CDT FP (%)	NP-CDT FN(%)	NP-CDT DD	H-NPCDT FP(%)	H-NPCDT (%) FN	H-NPCDT DD
5	100.0	0.0	-	8.20	0.29	133.3
10	58.59	0.0	98.9	4.00	0.11	154.1
15	26.43	0.0	130.1	1.71	0.07	159.9
20	9.64	0.29	166.8	0.57	0.26	183.3

TABLE I
FALSE POSITIVE (%), FALSE NEGATIVE (%) AND DETECTION DELAYS OF THE PROPOSED NON-PARAMETRIC CDT (NP-CDT) AND ITS HIERARCHICAL EXTENSION (H-NPCDT).

3) Comparing H-NPCDT with state-of-the-art change-detection mechanisms: In order to show the effectiveness of the proposed *H-NPCDT*, we compared its performance, in terms of FP, FN and DD, with three state-of-the-art change-detection mechanisms: Drift Detection Method (DDM) [32], Early Drift Detection Method (EDDM) [40] and Exponentially Weighted Moving Average (EWMA) [41].

To achieve this goal, two sets of experiments have been considered. In both sets of experiments $W = 5$ and $N = 2$ to allow the comparison between *H-NPCDT* operating on DTMCs and the considered state-of-the-art change-detection mechanisms operating on sequences of random variables (i.e., the binary classification error over time).

The first one refers to the same experiment described in Eq. (31) where T has been set to 2000, 3000 and 5000. We considered two different configurations for the state-of-the-art change-detection mechanisms: in DDM, σ has been set to 3 and 2; in EDDM, β has been set to 0.95 and 0.9; in EWMA, L has been set as Table 1 -$ARL_0 = 1000$ of [41] and 5. As regards *H-NPCDT*, we considered K equal to 1 and 5.

The experimental results of this comparison are detailed in Table II. Several comments arise. It is worth noting that *H-NPCDT* provides the best performance in terms of FP and FN with respect to DDM, EDDM and EWMA. As expected, FN decreases for all the change-detection mechanisms when T increases (also leading to a corresponding increase in the DDs). Moreover, the configuration of *H-NPCDT* with $K = 5$ provides the lowest FPs at the expense of an increase of FNs and DDs. Similarly, the different configurations of DDM, EDDM and EWMA provide a trade-off among FP, FN and DD. We emphasize that, in all the experiments, the best configuration of DDM, EDDM and EWMA provides lower performance than *H-NPCDT*. Obviously, configurations providing the largest FPs are also characterized by the lowest DDs.

The proposed *H-NPCDT* is also lightweight in terms of computational load, making it suitable for streaming analysis. The last column of Table II shows the comparison of the execution times per iteration (in ms) of the four considered

			$t_{max} = 2000$			$t_{max} = 3000$			$t_{max} = 5000$			Exec. Time
			FP (%)	FN(%)	DD	FP(%)	(%) FN	DD	FP (%)	FN(%)	DD	(ms)
Synthetic	H-NPCDT	$K = 1$	0.69	12.93	211.5	0.78	2.22	252.2	0.96	1.68	275.6	0.31
		$K = 5$	0.39	14.28	218.1	0.45	2.71	260.6	0.52	1.54	282.6	
	DDM	$\sigma = 3$	17.45	56.83	156.7	16.26	53.90	320.9	17.83	53.34	636.6	0.29
		$\sigma = 2$	60.23	25.27	96.7	60.98	22.49	271.0	60.69	20.46	615.1	
	EDDM	$\beta = 0.95$	79.09	20.91	NaN	79.56	19.87	218.7	80.48	18.27	738.2	0.35
		$\beta = 0.9$	59.96	40.04	NaN	62.93	36.04	210.4	62.92	34.63	708.8	
	EWMA	L as in [41]	99.07	0.11	36.4	98.60	0.02	47.0	98.96	0.02	80.1	0.13
		$L = 5$	64.65	10.87	84.6	65.72	8.09	191.1	64.58	6.55	326.6	
			$\delta_{CD} = 0.5$			$\delta_{CD} = 0.25$			$\delta_{CD} = 0.1$			
			FP (%)	FN(%)	DD	FP(%)	(%) FN	DD	FP (%)	FN(%)	DD	
ELEC [32]	H-NPCDT	$K = 1$	0.00	0.00	1365.7	0.00	0.00	2458.3	0.00	0.00	3675.2	-
	DDM	$\sigma = 10$	0.00	0.00	1464.7	0.00	0.00	5193.0	0.00	100.00	NaN	-
	EDDM	$\beta = 0.2$	0.00	100.00	NaN	0.00	100.00	NaN	0.00	100.00	NaN	-
	EWMA	$L = 18$	0.00	100.00	NaN	0.00	100.00	NaN	0.00	100.00	NaN	-

TABLE II

FALSE POSITIVE (FP), FALSE NEGATIVE (FN) AND DETECTION DELAYS (DD): COMPARISON AMONG H-NPCDT, DDM, EDDM AND EWMA.

		$\bar{L}$				Δ_L (ADAM w.r.t.)		
N	Change	*Fixed*	*Active*	*Passive*	*ADAM*	*Fixed*	*Active*	*Passive*
2	S0	8.103630e-02	8.293574e-02	8.113073e-02	8.071688e-02	0.996100 + 0.01	0.973265 + 0.01	0.994917 + 0.01
2	S1	2.743110e-02	4.636851e-02	5.023415e-02	6.192471e-02	2.258169 + 0.06	1.335591 + 0.02	1.232837 + 0.02
2	S2	1.131956e-02	1.185468e-02	1.314581e-02	3.172813e-02	2.809538 + 0.27	2.676692 + 0.20	2.413166 + 0.15
5	S0	2.795334e-06	2.814755e-06	2.781474e-06	2.781308e-06	0.996342 + 0.06	0.988957 + 0.06	1.000145 + 0.02
5	S1	9.037156e-07	2.207568e-06	2.687480e-06	4.034711e-06	4.764445 + 1.65	1.849614 + 0.38	1.514499 + 0.27
5	S2	5.434207e-07	6.180067e-07	7.336953e-07	3.343776e-06	6.135610 + 9.45	4.930982 + 5.81	3.884824 + 3.35
10	S0	1.498579e-09	1.500390e-09	1.497509e-09	1.497632e-09	0.999571 + 0.02	0.998351 + 0.02	1.000093 + 0.01
10	S1	4.508413e-10	8.214064e-10	9.987959e-10	1.239264e-09	2.750571 + 0.08	1.508949 + 0.02	1.240864 + 0.02
10	S2	4.392315e-10	4.746833e-10	5.323545e-10	6.442083e-10	1.467239 + 0.04	1.357507 + 0.03	1.210345 + 0.02
2	Storm	1.439869e-01	3.054131e-01	3.798536e-01	4.170199e-01	2.896234	1.365429	1.097844
2	Hurricane	3.296223e-02	3.143671e-02	3.449289e-02	3.670496e-02	1.113546	1.167583	1.064131
2	Major Hurr.	8.953661e-02	1.518917e-01	2.617715e-01	3.018572e-01	3.371327	1.987318	1.153132

TABLE III

AVERAGE LIKELIHOOD $\bar{L}$ AND RATIO Δ_L FOR THE CONSIDERED SOLUTIONS IN SCENARIOS $S0$, $S1$ AND $S2$ WITH $N = 2, 5, 10$.

change-detection mechanisms. More specifically, to compute these values, we measured the execution time of 100 iterations without any detection by the change-detection mechanisms and we computed the median value to remove outliers. The considered hardware platform is a 2,5 GHz Intel Core i7 with 16 GB 2133 MHz LPDDR3. Interestingly, execution times of *H-NPCDT*, DDM and EDDM are similar, while EWMA is characterized by the lowest computational load.

We also emphasize that, similarly to the other change detection mechanisms, the memory occupation of the *H-NPCDT* is very low, requiring only the storage of $2(N^2 + N)$ Float values and $W + 2N^2 + 4$ Integer values.

The second set of experiments refers to detection of changes in the ELEC2 benchmark that is typically used in the concept drift community [32]. This dataset contains 45312 records about the prediction of the electricity prices of the Australian New South West electricity market. The two classes are "UP" and "DOWN", representing the $N = 2$ states of the associated stochastic process. The experiment has been set-up by defining a dataset comprising the class labels of all the records. The first $L = 20000$ labels have been used for the training. The concept drift has been inserted at $t^* = 25000$ and modelled as a change of the label of δ_{CD}-percentage randomly-selected observations with "UP" label (that are transformed into "DOWN"). The parameters of DDM, EDDM and EWMA have been experimentally configured to avoid false-positive detections in case of no concept drift in the dataset. Three different values of δ_{CD} have been considered: 0.5, 0.25, and 0.1. Results, which are detailed in Table II, are particularly interesting and show that, even in this case, *H-NPCDT* provides the best trade-off between FP and DD, being able to detect the concept drift in all the three configurations of δ_{CD} without introducing false positive or negative detections. The DDM change-detection mechanism is able to detect all the concept drift in the configurations $\delta_{CD} = 0.5$ and $\delta_{CD} = 0.25$ but with larger DDs. The DDM is not able to detect any concept drift with $\delta_{CD} = 0.1$. Similarly, EDDM and EWMA are not able to detect any concept drift in any of the configurations of δ_{CD}.

B. Evaluating the Adaptive Algorithm for Markov Chains (ADAM)

In order to evaluate the ability of *ADAM* to learn DTMCs under concept drift, we defined the following set of synthetically-generated scenarios:

S0: *No concept drift*. The experiment lasts $T = 3000$ observations. The first $L = 1000$ samples represent the training sequence TS. Θ_0 is randomly generated and no concept drift occurs during the experiment;

S1: *Concept drift*. The experiment lasts $T = 3000$ observations. The first $L = 1000$ samples represent the training sequence TS. A concept drift occurs at time $t^* = 1500$. Θ_0 and Θ_1 are randomly generated;

S2: *Sequence of concept drift*. The experiment lasts $T = 3000$ observations. The first $L = 1000$ samples represent the training sequence TS. A sequence of three concept drift occurring at time $t^* = 1500$, $t^* = 2000$ and $t^* = 2500$ is here considered. Θ_0 and the three Θ_1s of the sequence of concept drift are randomly generated.

In addition, we also considered a public-available dataset from the National Oceanic and Atmospheric Administration [46] about storms, hurricanes and major hurricanes yearly registered in the Atlantic Basin from 1851 to 2016. Here, the problem has been reformulated as a two-state learning problem aiming at modelling the following three data sequences:

- in a year at least five storms were registered in the Atlantic basin (0: no/1: yes);
- in a year at least five hurricanes were registered in the Atlantic basin (0: no/1: yes);
- in a year at least one major hurricane was registered in the Atlantic basin (0: no/1: yes).

We compared *ADAM*, in the configuration encompassing the *H-NPCDT*, with the following learning solutions inspired by the literature of learning in presence of concept drift [4]:

- *Fixed:* $\hat{P}$ is estimated on TS and not updated during the experiment. This solution refers to a traditional not-adaptive learning approach;
- *Active:* $\hat{P}$ is estimated on TS. The hierarchical change-detection mechanism monitors the observations coming from $\mathcal{P}$. When a change is detected, $\hat{P}$ is estimated on the recently-acquired L observations;
- *Passive:* $\hat{P}$ is trained on TS and adapted over time by relying on a sliding window of length L on the recently-acquired observations .

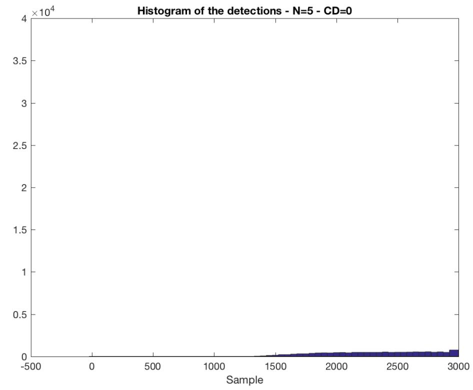
The considered figures of merit are the average likelihood $\bar{L}$ over the experiment, the ratio Δ_L between $\bar{L}$ provided by *ADAM* and that by the other solutions, and the length L_{adapt} of the adaptive window in *ADAM*. γ has been set to 2 and, even in this case, $W = 5$.

Experimental results are shown in Table III and Fig. 3. More specifically, Table III shows the average likelihood $\bar{L}$ and the ratio Δ_L for the considered experimental scenarios with $N = 2, 5, 10$. Three main comments arise. First, as expected, all the considered solutions provide similar likelihood $\bar{L}$s in scenario $S0$ for all the values of N (as emphasized by the values Δ_Ls that are close to 1). This is reasonable since, in stationary conditions, both adaptive and non-adaptive solutions are effective. Differently, in scenarios $S1$ and $S2$, adaptive solutions (i.e., *Active*, *Passive* and *ADAM*) outperform the *Fixed* solution. Second, *ADAM* clearly outperforms both the *Active* and the *Passive* solution in scenarios $S1$ and $S2$ and $N = 2, 5, 10$ (see values of Δ_L with related standard deviation). This corroborates the ability of the proposed solution to effectively adapt to concept drift affecting a DTMC. It is also worth noting that, as expected, the advantages provided by *ADAM* are even more evident in scenario $S2$, comprising a sequence of concept drift. Third, *ADAM* is also very effective with the real-world dataset about storms, hurricanes and major hurricanes from the NOAA. A particularly interesting result is that, as regards the major hurricane, the change has been detected by the *H-NPCDT* in the year 1950. This result is also confirmed by the climatological analysis described in [28], showing that a larger major-hurricane activity is present in North Atlantic in the decade 1940-1950. It is also worth noting that, in that climatological analysis, the activity of hurricanes in the considered period (1851-2016) revealed to be stationary. Even this fact is confirmed by the results of *ADAM* showing that no change-detection occurred after the training sequence during the analysis of the hurricane dataset.

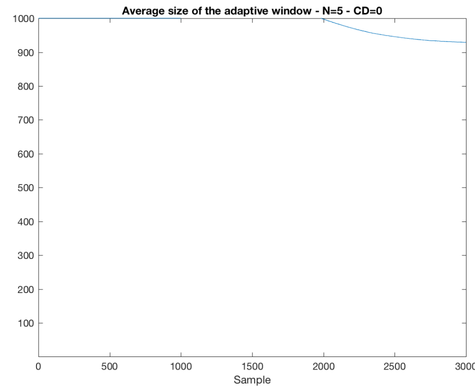
Results depicted in Fig. 3 show the ability of *ADAM* to adapt DTMCs in presence of concept drift in Scenario $S0$, $S1$ and $S2$ with $N = 5$. In particular, the histograms of the change-detection time instants $\bar{t}$, i.e., Fig. 3(a), (c) and (e), corroborate the expected behavior of detections. In fact, in Fig. 3(a) the number of detections is low and detections are distributed in the whole time-horizon of the experiment since no concept drift is here introduced (i.e., these detections are FPs); Fig. 3(c) shows a peak of detections between $t = 1500$ and 2000 and this is reasonable since in $S1$ the concept drift occurs at time $t^* = 1500$; Fig. 3(e) shows three peaks of detections between $t = 1500$ and $t = 3000$ and, again, this is reasonable since $S2$ encompasses three concept drift in that time horizon. Fig. 3(b), (d) and (f) show the average window size L_{adapt} of *ADAM* in the three considered scenarios. These results are particularly interesting since they show the ability of *ADAM* to adapt the window size to concept drift. In fact, as expected, L_{adapt} decreases after a concept drift and increases during the stationary periods (see Fig. 3 (d) and (f)). The reduction of L_{adapt} in $S0$ is due to the false positive detections.

VII. Conclusions

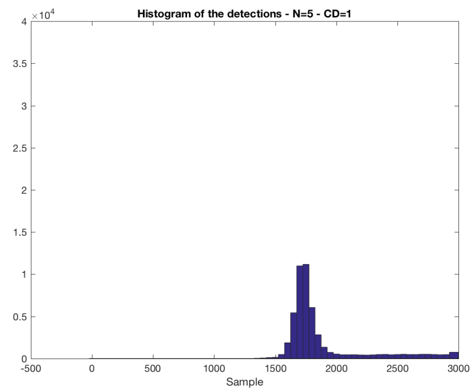
This paper introduces, for the first time in the literature, a family of change-detection mechanisms and a learning algorithm, called *ADAM*, to deal with DTMCs under concept drift. In particular, three different change-detection mechanisms have been proposed differing in required assumptions and performance. Theoretical properties have been derived for the parameteric change-detection mechanism. The proposed *ADAM* relies on a hybrid active-passive approach where the estimated transition matrix is adapted over time at each new observation (as in passive approaches), while the estimation of the transition matrix is triggered by the change-detection mechanism to react to concept drift to remove obsolete knowledge. The adaptive mechanism of *ADAM* relies on an adaptive window on the recently acquired observations whose length is widened or reduced according to a change-detection index extracted from the proposed change-detection mechanisms. Results on both synthetically-generated datasets and real-world datasets show the effectiveness of the proposed change detection mechanisms and *ADAM*.



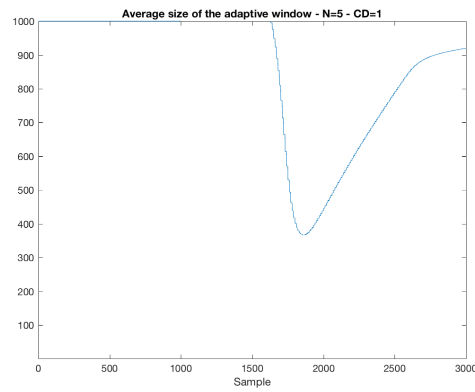
(a) Histogram of change detection $\bar{t}$ in $S0$



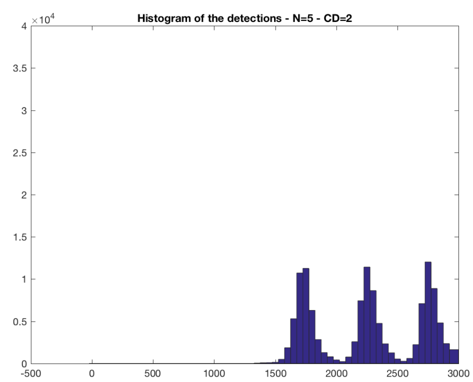
(b) Average window size L_{adapt} in $S0$



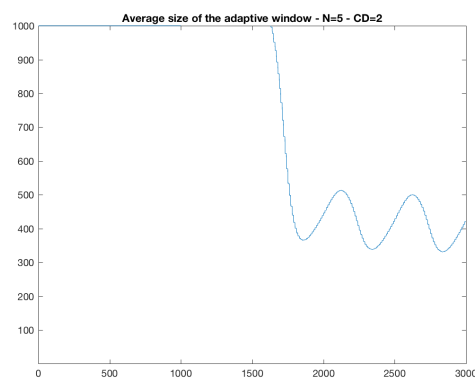
(c) Histogram of change detection $\bar{t}$ in $S1$



(d) Average window size L_{adapt} in $S1$



(e) Histogram of change detection $\bar{t}$ in $S2$



(f) Average window size L_{adapt} in $S2$

Fig. 3. Histogram of change-detection time instant $\bar{t}$s and average window size L_{adapt} for *ADAM* in Scenarios $S0$, $S1$ and $S2$ with $N = 5$.

Future works will encompass the integration of adaptive mechanisms to deal with gradual or intermittent concept drift, the extension of ADAM to non-homogeneous DTMCs and the introduction of the change detection and adaptation mechanisms in Hidden Markov Models.

Acknowledgement

This work was supported by the project Italian PRIN GAUChO Project 2015.

References

[1] R. Polikar and C. Alippi, "Guest editorial learning in nonstationary and evolving environments," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 9–11, Jan 2014.

[2] A. Tsymbal, "The problem of concept drift: definitions and related work," *C. S. Dept., Trinity College Dublin*, vol. 106, no. 2, 2004.

[3] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, p. 44, 2014.

[4] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in non-stationary environments: A survey," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, pp. 12–25, 2015.

[5] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[6] C. Alippi and M. Roveri, "The (not) far-away path to smart cyber-physical systems: An information-centric framework," *Computer*, vol. 50, no. 4, pp. 38–47, 2017.

[7] C. Alippi, W. Qi, and M. Roveri, "An improved hilbert-huang transform for non-linear and time-variant signals," in *Multidisciplinary Approaches to Neural Computing*. Springer, 2018, pp. 109–117.

[8] D. Zambon, C. Alippi, and L. Livi, "Concept drift and anomaly detection in graph streams," *arXiv preprint arXiv:1706.06941*, 2017.

[9] J. R. Norris, *Markov chains*. Cambridge university press, 1998, no. 2.

[10] M. Sharpe, *General theory of Markov processes*. Academic press, 1988, vol. 133.

[11] T. W. Anderson and L. A. Goodman, "Statistical inference about markov chains," *The Annals of Mathematical Statistics*, pp. 89–110, 1957.

[12] C. Zhang and D. Tao, "Generalization bounds of erm-based learning processes for continuous-time markov chains," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 23, no. 12, pp. 1872–1883, Dec 2012.

[13] M. Basseville, I. V. Nikiforov *et al.*, *Detection of abrupt changes: theory and application*. Prentice Hall Englewood Cliffs, 1993, vol. 104.

[14] C. Alippi, G. Boracchi, and M. Roveri, "Hierarchical change-detection tests," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 2, pp. 246–258, 2017.

[15] C. Alippi, W. Qi, and M. Roveri, "Learning in nonstationary environments: A hybrid approach," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2017, pp. 703–714.

[16] C. Alippi, *Intelligence for Embedded Systems: A Methodological Approach*. Springer, 2014.

[17] B. A. Craig and P. P. Sendi, "Estimation of the transition matrix of a discrete-time markov chain," *Health economics*, vol. 11, no. 1, pp. 33–42, 2002.

[18] J. Hajnal and M. Bartlett, "Weak ergodicity in non-homogeneous markov chains," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 54, no. 2. Cambridge University Press, 1958, pp. 233–246.

[19] ——, "The ergodic properties of non-homogeneous finite markov chains," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 52, no. 1. Cambridge University Press, 1956, pp. 67–77.

[20] B. A. Craig, D. G. Fryback, R. Klein, and B. E. Klein, "A bayesian approach to modelling the natural history of a chronic condition from observations with intervention," *Statistics in medicine*, vol. 18, no. 11, pp. 1355–1371, 1999.

[21] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2917–2929, 1998.

[22] A. N. Shiryaev, *Optimal stopping rules*. Springer Science & Business Media, 2007, vol. 8.

[23] G. V. Moustakides, "Quickest detection of abrupt changes for a class of random processes," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1965–1968, 1998.

[24] B. Chen and P. Willett, "Detection of hidden markov model transient signals," *IEEE Transactions on Aerospace and Electronic systems*, vol. 36, no. 4, pp. 1253–1268, 2000.

[25] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, pp. 326–339, 1948.

[26] B. Sin and J. H. Kim, "Nonstationary hidden markov model," *Signal Processing*, vol. 46, no. 1, pp. 31–46, 1995.

[27] P. M. Djuric and J.-H. Chun, "An mcmc sampling approach to estimation of nonstationary hidden markov models," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1113–1123, 2002.

[28] J. B. Elsner, X. Niu, and T. H. Jagger, "Detecting shifts in hurricane rates using a markov chain monte carlo approach," *Journal of climate*, vol. 17, no. 13, pp. 2652–2666, 2004.

[29] Y. Zhai and M. Shah, "Video scene segmentation using markov chain monte carlo," *IEEE Trans. on Multimedia*, vol. 8, no. 4, pp. 686–697, 2006.

[30] N. Ye *et al.*, "A markov chain model of temporal behavior for anomaly detection," in *Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, vol. 166. West Point, NY, 2000, p. 169.

[31] K. Hara, T. Omori, and R. Ueno, "Detection of unusual human behavior in intelligent house," in *Neural Networks for Signal Processing, 2002*. IEEE, 2002, pp. 697–706.

[32] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Brazilian symposium on artificial intelligence*. Springer, 2004, pp. 286–295.

[33] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 2007, pp. 443–448.

[34] C. Alippi and M. Roveri, "Just-in-time adaptive classifierspart ii: Designing the classifier," *IEEE Transactions on Neural Networks*, vol. 19, no. 12, pp. 2053–2064, 2008.

[35] C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 4, pp. 620–634, 2013.

[36] W. N. Street and Y. Kim, "A streaming ensemble algorithm (sea) for large-scale classification," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 377–382.

[37] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: A new ensemble method for tracking concept drift," in *Data Mining, 2003. ICDM*. IEEE, 2003, pp. 123–130.

[38] R. Elwell and R. Polikar, "Incremental learning of concept drift in nonstationary environments," *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1517–1531, 2011.

[39] C. Alippi, G. Boracchi, M. Roveri, G. Ditzler, and R. Polikar, "Adaptive classifiers for nonstationary environment," *Contemporary Issues in Systems Science and Engineering*, pp. 265–288, 2015.

[40] M. Baena-García, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavaldà, and R. Morales-Bueno, "Early drift detection method," 2006.

[41] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, "Exponentially weighted moving average charts for detecting concept drift," *Pattern recognition letters*, vol. 33, no. 2, pp. 191–198, 2012.

[42] P. Sobhani and H. Beigy, "New drift detection method for data streams," in *Adaptive and intelligent systems*. Springer, 2011, pp. 88–97.

[43] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004, pp. 180–191.

[44] G. Ditzler and R. Polikar, "Hellinger distance based drift detection for nonstationary environments," in *Computational Intelligence in Dynamic and Uncertain Environments (CIDUE), 2011 IEEE Symposium on*. IEEE, 2011, pp. 41–48.

[45] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[46] N. Oceanic and A. Administration. (2017) Atlantic basin. individual years with the numbers in each category. [Online]. Available: http://www.aoml.noaa.gov/hrd/tcfaq/E11.html

PLACE
PHOTO
HERE

Manuel Roveri Manuel Roveri received the Dr. Eng. degree in Computer Science Engineering from the Politecnico di Milano (Italy) in June 2003, the MS in Computer Science from the University of Illinois at Chicago (USA) in December 2003 and the Ph.D. degree in Computer Engineering from the Politecnico di Milano (Italy) in May 2007. He has been Visiting Researcher at Imperial College London (UK) in 2011. Currently, he is an Associate Professor at the Department of Electronics and Information of the Politecnico di Milano (Italy). Current research activity addresses include intelligent embedded and cyber-physical systems, learning in nonstationary-evolving environments and adaptive algorithms.

Manuel Roveri is a Senior Member of IEEE and an Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems. He served as Chair of the Neural Networks Technical Committee and Chair of the Task Force on Intelligent Cyber-Physical Systems of the IEEE Computational Intelligent Society. He also served as a Chair and Member in several IEEE subcommittees. He holds 1 patent and has published about 100 papers in international journals and conference proceedings.

Manuel Roveri received the following scientific awards: the Outstanding Transactions on Neural Networks and Learning Systems Paper Award from the IEEE Computational Intelligence Society in 2016; the Best Regular Paper Award at the INNS Conference on Big Data in 2016; the Outstanding Computational Intelligence Magazine Paper Award from the IEEE Computational Intelligence Society in 2018.