Optimal detection and error exponents for hidden semi-Markov models

Dragana Bajović, Kanghang He, Lina Stanković, Dejan Vukobratović, and Vladimir Stanković

[1]

Abstract. We study detection of random signals corrupted by noise that over time switch their values (states) between a finite set of possible values, where the switchings occur at unknown points in time. We model such signals as hidden semi-Markov signals (HSMS), which generalize classical Markov chains by introducing explicit (possibly non-geometric) distribution for the time spent in each state. Assuming two possible signal states and Gaussian noise, we derive optimal likelihood ratio test and show that it has a computationally tractable form of a matrix product, with the number of matrices involved in the product being the number of process observations. The product matrices are independent and identically distributed, constructed by a simple measurement modulation of the sparse semi-Markov model transition matrix that we define in the paper. Using this result, we show that the Neyman-Pearson error exponent is equal to the top Lyapunov exponent for the corresponding random matrices. Using theory of large deviations, we derive a lower bound on the error exponent. Finally, we show that this bound is tight by means of numerical simulations.

Keywords. Multi-state processes, hidden semi Markov models, explicit random duration, hypothesis testing, error exponent, large deviations principle, threshold effect, Lyapunov exponent.

I. Introduction

The problem of detecting a signal hidden in noise is investigated. The signal to be detected is characterised as having a constant magnitude in any one state and can transition to multiple states over time. Each occurrence of a particular state has a random duration, modelled as a discrete random variable which takes values from the finite set of integers, according to a certain probability mass function (pmf) associated with that state. Signal models of this kind are known in the literature as hidden semi-Markov models (HSMM) [1][2], which differ from the standard hidden Markov models in that in each state, the process can emit more than one observation. The underlying unobservable process in this case is called semi-Markov [3], and is defined as a sequence of pairs of two random variables – one from the Markov chain evolving sequence of states, and the other being the time spent in each visited state, the statistics of which is described by a pmf (or pdf, in the continuous time case). A related, more general class of random multi-state signals are Markov switching models [4] (or Markov jump processes) generally used to model time series and other kinds of signals, where the signal parameters of a certain model (e.g., moving average [4]) switch over time in a Markov fashion. When the durations of each parameter regime are modelled explicitly by a pmf (or pdf), the corresponding model is called explicit-duration Markov switching models – of which HSMMs are a special case.

Our main motivation for studying the described model comes from non intrusive appliance load monitoring (NILM) problem, i.e., detecting one or more particular appliance states, each of unknown duration, within an aggregate power signal, as obtained from smart meters. With the large-scale roll-out of smart meters worldwide, there has been increased interest in NILM, i.e., disaggregating total household energy consumption measured by the smart meter down to appliance level using purely software tools [5]. NILM can enrich energy feedback, it can support smart home automation [6], appliance retrofit decisions, and demand response measures [7].

NILM is an NP-hard problem [5], and an exact solution can only be found via exhaustive search: in practice, it would take over 1700 years to disaggregate 30 appliances using exhaustive search from a months data with top current GPUs [8, p. 124].

Since NILM boils down to identifying unknown sources that go through a sequence of (hidden) states (ON to OFF for single state loads), Hidden Markov Models (HMM), have become popular for this time-series data, with a number of extensions proposed over the past few years, including factorial HMM (FHMM), conditional FHMM, etc. [9], [10], [11], [12]. However some appliances violate the Markovian assumption [8, p.142], as the durations that appliances are on and off are not geometrically distributed, as occurs with HMMs. Further, the duration of appliance runs are not captured, which is the key difference w.r.t speech applications where durations of sounds are approximately equal.

NILM can also be seen as a pure signal waveform, or pattern recognition problem, with solutions drawn from a rich field of audio signal processing and speech recognition, including Dynamic Time Warping [7], rule-based and dictionary-based approaches. With a vast amount of different formulations, many signal processing and machine learning techniques have been proposed in the literature, including k-means, SVM [13], neural network [14], kNN, Generalized Viterbi [5], naïve Bayes, Genetic Algorithms, Graph Signal

[1]D. Bajovic and D. Vukobratovic are with Department of Power, Electronic and Communications Engineering, Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia (e-mail: {dbajovic, dejanv@uns.ac.rs}) K. He, L. Stankovic and V. Stankovic are with Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, G1 1XW, UK (e-mail: {kanghang.he, lina.stankovic, vladimir.stankovic}@strath.ac.uk)

Processing (GSP) [15], [16], Decision Trees [7], particle filtering, evolutionary algorithms [17], etc., but without measurable and convincing evidence of reliability, acceptable accuracy, and scalability.

Despite significant research efforts in developing efficient NILM algorithms (see [7], [15], [16], [9], [10] and references therein), NILM is still a challenge, especially at low sampling rates, in the order of seconds and minutes. One obstacle is the lack of standardised performance measures and appropriate theoretical bounds of detectability of appliance usage, which can help estimating performance of various algorithms. A particularly challenging problem is the detection of multi-state appliances, i.e., appliances whose power consumption switches over one appliance runtime through several different values. Examples of such appliances are a dishwasher or a washing machine, where the load or the chosen program or setting determines duration that the appliance spends in each state. The difficulty there arises from the fact that the program and the load, unknown from the perspective of NILM, are non-deterministic, i.e., vary each time the same appliance is run resulting in difficulty in detecting in which state the appliance is. In this work we propose to use HSMM as a model for multi-state appliances, where we have the full freedom to describe the state durations statistics, and thus obtain a better fit for multi-state appliance signals than with HMMs, which allow only for geometrically decaying pmfs on the state durations. The aggregate load minus the load of the appliance to be detected, consisting of other appliances being switched on and off randomly over time, is well modelled as Gaussian additive noise, as shown in [11].

HSMM is also representative of signals occurring in a range of other applications. In econometrics, examples of explicit duration signals include marital or employment status, or in general the time an individual spends in a certain state [18]. Further examples from econometrics are time to currency alignment or time to transactions in stock market [19]. In biometrics, HSMM is used to model forest tree growth and identify individual growth components [20]. In communication systems theory, pulse-duration modulated (PDM) signals for transmitting information encoded into the pulse duration have two possible signal states: the positive value state is a pulse whose duration is proportional to the information symbol to be encoded, and the zero-value state in between any two pulses. The probability distribution of the state duration is then controlled by the probability distribution on the set of information symbols to be transmitted. Further binary state examples are random telegraph signals, where the signal switches between two values in a random manner[2], and the activity pattern of a certain mobile user in a cellular communication system. We refer the reader to references [2], [4], [1] for detailed accounts on various other applications of HSMMs.

In this paper we focus on detection of binary signals of random state durations, hidden in noise, modelled as (binary) HSMMs. While the problem of detecting multi-state signals hidden in noise has been presented in [21], [23] and [24], the latter model the signal as hidden Markov chains unlike our proposed approach which adopts HSMM, with an explicit duration model for each of the states. Specifically, in [21] random telegraph signals are modelled as binary Markov chains and the corresponding optimal detection test is derived in the form of a product of certain measurement defined matrices. Detection of a random walk on a graph is considered in [23], where bounds on the error exponent for the Neyman-Pearson detection test are derived. The method of types is used in [24] to generalize the results from [23] to non-homogeneous setting where different nodes have different signal-to-noise ratios (SNR) with respect to the walk. Furthermore, proof is given in [24] that the derived bound on the error exponent has a convex optimization form.

Assuming Neyman-Pearson setting, we are interested in detection performance characterization, through computing the corresponding error exponent – the decay rate of the probability of a miss, under a constraint on the probability of false alarm, for given HSMM model parameters. It is well-known that when observations are independent and identically distributed (i.i.d.) both in the presence and absence of the signal (e.g., when the signal value is constant and known and the noise realizations are i.i.d.), the Neyman-Pearson error exponent is given by the Kullback-Leibler divergence between the corresponding two hypotheses, see Stein's lemma in [25], [26], and also [27]. This property, in a sense, extends to non-i.i.d. signal models of certain classes (such as, for example, ergodic models), in which case the error exponent is given by the asymptotic Kullback-Leibler rate [28], [29] (see the expression in (7) in Section II further ahead). Computing this limit is a difficult problem in general, but, for certain cases, solutions are known.

We briefly review the literature on error exponents for signals with Markovian structure. In [30] error exponent is computed for testing between two different Markov sources (without additive noise in the observations); for applications and extensions of this result in Markov source-coding see [31], [32], [33], [34]. Error exponents for HMMs are considered in [23] and [24], as detailed above. Error exponent is also shown to be computable for the problem of discriminating between two autoregressive processes (AR) of different parameters [35], [36]. For Gauss-Markov models, represented as AR process of order 1 with Gaussian noise, [37] finds a closed form for the error exponent via spectral domain characterization of the observed process. To the best of our knowledge, there are no results on the error exponent for HSMMs.

Contributions. In this paper, we first show that the optimal detection test, seemingly combinatorial in nature, admits a simple, linear recursion form of a product of matrices of dimension equal to the sum of the duration spreads for the two states. Using the preceding result, we show that the Neyman-Pearson error exponent for this problem is given by the top Lyapunov exponent [38] for the matrices that define the recursion. Each matrix involved in the product is of dimension equal to the sum of durations spreads of the two states, and it can be decomposed as a product of a

[2]We remark that there are other stochastic models in the literature for the random telegraph signal, e.g., the Poisson model, or the hidden Markov chain model [21], [22].

diagonal random matrix controlled by the process observations and a sparse constant matrix which governs transitions in the sequence of states of different durations. Thus, we reveal that a similar structural effect as with the error exponent for hidden Markov processes occurs here as well [21], [24]. This result is of immediate interest for inference in HSMM, as it allows extension and application to HSMM of certain algorithms designed for HMM that specifically rely on matrix product representation of the likelihood, see [20], [39]. Further, using the introduced transition matrix for the semi-Markov model, we find explicitly an upper bound on the error exponent, equal to the expected SNR of the process. This bound has an intuitive physical interpretation: it is the error exponent for the detection test which has information on the exact locations of all state transitions in the observed sequence of measurements. Finally, using the theory of large deviations [26], we derive a lower bound on the error exponent and demonstrate by numerical simulations that the derived bound is very close to the true error exponent.

Paper outline. Section II states the problem setup and Section III gives the preliminaries. Section IV gives main results on the form of the optimal likelihood ratio test. Section V provides the lower bound on the error exponent, while Section VI proves this result. Finally, numerical results are given in Section VII and Section VIII concludes the paper.

Notation. For an arbitrary integer n, $\mathbb{S}^{n-1}$ denotes the probability simplex in $\mathbb{R}^n$; e_1 denotes the first canonical vector (the n dimensional vector with 1 only in the first position, and having zeros in all other positions), and $\mathbb{1}$ the vector of all ones, where we remark that the dimension should be clear from the context; A_0 denotes the lower shift matrix (the $0/1$ matrix with ones only on the first subdiagonal); $\|\cdot\|$ denotes the spectral norm. We denote Gaussian distribution of mean value μ and standard deviation σ by $\mathcal{N}(\mu,\sigma^2)$; by $p[1,n]$ an arbitrary distribution over the first n integers; by $\mathcal{U}[1,n]$ the uniform distribution over the first n integers; $\log$ denotes the natural logarithm.

II. Problem setup

We consider the problem of detecting a signal corrupted by noise that randomly switches from one state m to another, where $m = 1, 2, ..., M$ and in each state the signal has a certain magnitude μ_m. The duration that the signal spends in a given state m is modelled as a discrete random variable on a given support set $[1, \Delta_m]$, and with a certain pmf defined by vector $p_m \in \mathbb{S}^{\Delta_m - 1}$. In this work, we consider the case when $M = 2$ and we assume that for each state m we know the corresponding value of the observed signal μ_m. Without loss of generality, we will assume that $\mu_2 > \mu_1 \geq 0$. For each sampling time $t = 1, 2, ...$, let $S^t = \{S_1, ..., S_t\}$ denote the sequence of states until time t of the signal that we wish to detect, where for each $k = 1, ..., t$, $S_k \in \{1, 2\}$; similarly, we denote $S^\infty = \{S_1, S_2, ...\}$. Let also $\mathcal{S}^t$ denote the set of all feasible sequences of states s^t of length t. We assume that, with probability one, the first state is $S_1 \equiv 1$, and, for the purpose of analysis, we set $S_0 \equiv 2$. Let X_k denote the signal measurement for sample time k, $k = 1, ..., t$, and, for each t, collect all measurements up to time t in vector $X^t = (X_1, ..., X_t)$. We assume that each measurement is corrupted by a zero mean additive Gaussian noise $\mathcal{N}(0, \sigma^2)$, where standard deviation $\sigma > 0$.

The sequence of switching times. For the sequence of states $S_1, S_2, ...$, we define the sequence of times $\{T_1, T_2, ...\}$, when the signal in the sequence switches from one state to another, i.e.,

$$T_{i+1} = \max\{k \geq T_i + 1 : S_k = S_{T_i+1}\}, \text{ for } i = 0, 1, 2, ... \tag{1}$$

where we set $T_0 \equiv 0$. We call a phase each time window $[T_i + 1, T_{i+1}]$, $i = 0, 1, 2, \ldots$, and note that during any phase, the sequence S^∞ stays in the same state. Since $S_1 \equiv 1$, all odd-numbered intervals $[T_0 + 1, T_1]$, $[T_2 + 1, T_3]$,..., where the ordering is with respect to the order of appearance, are state 1 phases, and all even-numbered intervals $[T_1 + 1, T_2]$, $[T_3 + 1, T_4]$,... are state 2 phases.

Random duration model. For $n = 1, 2, ...$, we denote by $D_{1,n}$ the difference process

$$D_{1,n} = T_{2n-1} - T_{2n-2}, \tag{2}$$

or, in words, for each n, $D_{1,n}$ is the duration of the n-th state-1 phase in the sequence S^∞. We assume that durations of state-1 phases are independent and identically distributed (i.i.d.), with support set of all integers in the finite interval $[1, \Delta_1]$, and with pmf given by vector $p_1 = (p_{11}, p_{12}, ..., p_{1\Delta_1}) \in \mathbb{S}^{\Delta_1 - 1}$, which we denote by $D_{1,n} \sim p_1\,(1, \Delta_1)$. Similarly, we define

$$D_{2,n} = T_{2n} - T_{2n-1} \tag{3}$$

to be the duration of the n-th state-2 phase in the sequence $S_1, S_2, ...$, for $n = 1, 2, ...$; we assume that the $D_{2,n}$'s are i.i.d., with support set of all integers in the interval $[1, \Delta_2]$, and pmf given by vector $p_2 = (p_{21}, p_{22}, ..., p_{2\Delta_2}) \in \mathbb{S}^{\Delta_2 - 1}$, i.e., $D_{2,n} \sim p_2\,(1, \Delta_2)$. We also assume that durations of state-1 and state-2 phases are mutually independent.

Hypothesis testing problem. Using the preceding definitions, we model the signal detection problem as the following binary hypothesis testing problem:

$$\mathcal{H}_0 : X_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{4}$$

$$\mathcal{H}_1 : X_k | S^t \overset{\text{indep.}}{\sim} \begin{cases} \mathcal{N}(\mu_1, \sigma^2), & \text{if } S_k = 1 \\ \mathcal{N}(\mu_2, \sigma^2), & \text{if } S_k = 2 \end{cases}, \text{ for } k = 1, ..., t,$$

where we assume $S_1 \equiv 1$. We remark that the model above easily generalizes to the case when the signals X_k are under both hypotheses shifted for some $\mu_0 \in \mathbb{R}$, i.e., when, under $\mathcal{H} = \mathcal{H}_0$, $X_k \sim \mathcal{N}(\mu_0, \sigma^2)$ and, under $\mathcal{H} = \mathcal{H}_1$, $X_k \sim \mathcal{N}(\mu_{S_k} + \mu_0, \sigma^2)$; see the example of appliance detection problem later in this section. The latter hypothesis testing problem reduces to the one in (4) by means of the change of variables $Y_k = X_k - \mu_0$.

Illustration: Multiphase appliance detection. Suppose that we wish to detect an event that a certain appliance in a household is switched on. We consider classes of appliances

whose signature signals exhibit a multistate (multiphase) type of behavior, such as switching from high to low signal values, where the durations of phases of the same signal level can be different across a single appliance run-time and also in different run-times of the same appliance. Examples of appliances whose signatures fall into this class are, e.g., a dishwasher and a washer-dryer. This problem can be modelled by the hypothesis testing problem (4) where μ_1 corresponds to the appliance consumption when in low state and μ_2 corresponds to the appliance consumption when in high state. In this scenario, there is an underlying baseline load which can also be modelled as a Gaussian random variable of expected value μ_0 and standard deviation σ^2. Since the same baseline load is present both under $\mathcal{H}_0$ and $\mathcal{H}_1$, to cast the described appliance detection problem in the format given in (4), we simply subtract the value μ_0 from the observed consumption signal X_k.

Comparison with random telegraph signals. The signal model that we consider is structurally similar to the random telegraph signal, modelled as a hidden binary Markov chain. The random telegraph signal switches between two opposite signal values, $\mu_1 = +\mu$ and $\mu_2 = -\mu$, where the transitions are governed by a certain transition matrix, which we denote by $P_{\mathrm{RT}} = [q\ (1-q); (1-q)\ q]$ (assuming, for simplicity, symmetry in the two states). Given that the random telegraph signal has just entered, say, state 1, we look at the probability that the signal stays in this state for d time instants, where d is arbitrary. It is easy to show that this probability equals $(1-q)q^{d-1}$, for arbitrary $d \geq 1$. That is, with the random telegraph signal, the distribution on the durations of states is geometric – thus, it decays with d exponentially. On the other hand, with the binary semi-Markov model that we consider, there is a complete freedom in setting the distribution on the time that the signal spends in either of the states, provided that the maximal state duration is bounded by some finite Δ. When Δ is large, and these pmfs are quasi (truncated) geometric, $p_1 = p_2 = 1/(1-q^{\Delta})\left(1-q, q(1-q), \ldots, q(1-q)^{\Delta-1}\right)$, the semi-Markov model can be approximated by the random telegraph signal, which has a simpler parametric representation. However, when the two pmfs are, for example, uniform, or even when the longer state durations in the studied signal are much more likely than the shorter ones (consider $p_1 = p_2 = (\epsilon, \epsilon, \ldots, 1-(\Delta-1)\epsilon)$), then the semi-Markov model is a much better alternative to the random telegraph signal. With multi-state appliances, once entered, any state is likely to last for a certain time (usually much longer than the unit, sampling period time), and hence the motivation to use semi-Markov models over Markov chains. See also Section VII and Figure 8 for a numerical illustration of the comparison of the two models.

Likelihood ratio test and Neyman-Pearson error exponent. We denote the probability laws corresponding to $\mathcal{H}_0$ and $\mathcal{H}_1$ by $\mathbb{P}_0$ and $\mathbb{P}_1$, respectively. Similarly, the expectations with respect to $\mathbb{P}_0$ and $\mathbb{P}_1$ are denoted by $\mathbb{E}_0$ and $\mathbb{E}_1$, respectively. The probability density functions of X^t under $\mathcal{H}_1$ and $\mathcal{H}_0$ are denoted by $f_{1,t}(\cdot)$ and $f_{0,t}(\cdot)$, respectively. It will also be of interest to introduce the conditional probability density function of X^t given $S^t = s^t$ (i.e., the likelihood functions), which we denote by $f_{1,t|S^t}(\cdot|s^t)$, for any s^t. Finally, the likelihood ratio at time t denoted by L_t, and at a given realization of X^t is computed by $L_t(X^t) = \frac{f_{1,t}(X^t)}{f_{0,t}(X^t)}$.

It is well known that the optimal detection test (both in Neyman-Pearson and Bayes sense) for problem (4) is the likelihood ratio test. Conditioning on the state realizations until time t, $S^t = s^t$, and denoting shortly $P(s^t) = \mathbb{P}_1(S^t = s^t)$, we have

$$\begin{aligned} L_t(X^t) &= \sum_{s^t \in \mathcal{S}^t} P(s^t) \frac{f_{1,t|S^t}(X^t|s^t)}{f_{0,t}(X^t)} \\ &= \sum_{s^t \in \mathcal{S}^t} P(s^t) \frac{\prod_{k=1}^{t} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mu_{s_k} - X_k)^2}{2\sigma^2}}}{\prod_{k=1}^{t} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{X_k^2}{2\sigma^2}}}, \end{aligned} \tag{5}$$

where, we recall, $\mathcal{S}^t$ is the set of all feasible sequences – for which $\mathbb{P}_1(\mathcal{S}^t = s^t) > 0$. In this paper our goal is to find a computationally tractable form for the optimal, likelihood ratio test and also to characterize its asymptotic performance, when the number of samples X_k grows large. In particular, with respect to performance characterization, we wish to compute the error exponent for the probability of a miss, under a given bound α on the probability of false alarm:

$$\lim_{t \to +\infty} -\frac{1}{t} \log P^{\alpha}_{\mathrm{miss},t} =: \zeta, \tag{6}$$

where $P^{\alpha}_{\mathrm{miss},t}$ is the minimal probability of a miss among all decision tests that have probability of false alarm bounded by α. By results from detection theory, e.g., [28], [29], the ζ in (6) is given by the asymptotic Kullback-Leibler rate in (7), provided that this limit exists

$$\zeta = \lim_{t \to +\infty} -\frac{1}{t} \log L_t(X^t). \tag{7}$$

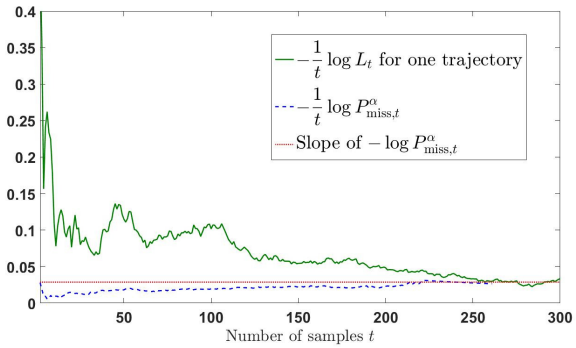

Fig. 1: Simulation setup: $\Delta = 3$, $p_1, p_2 \sim \mathcal{U}([1, \Delta])$, $\mu_1 = 2$, $\mu_2 = 5$, $\sigma = 10$, $\alpha = 0.01$. Green full line plots the evolution of $-\frac{1}{t}\log L_t$; blue dotted line plots the evolution of $-\frac{1}{t}\log P^{\alpha}_{\mathrm{miss},t}$, and red dashed line plots the estimated slope of the probability of a miss values (in the logarithmic scale) calculated for values until $t = 300$ observations.

We prove the existence of the limit in (7) in Lemma 9 in Section V further ahead. An illustration of the identity (6) is given in Figure 1, which clearly shows that both sequences

$-\frac{1}{t}\log P^{\alpha}_{\text{miss},t}$ and $-\frac{1}{t}\log L_t(X^t)$ are convergent and moreover that they converge to the same value – the asymptotic Kullback-Leibler rate for the two hypotheses defined in (4). For further details on this simulation see Section VII.

III. Preliminaries

In this section we now introduce a number of quantities related with the sequences $s^t \in \mathcal{S}^t$, $t = 1, 2, ...$, and give certain results pertaining to these quantities that will be useful for our analysis.

Statistics for the durations of phases.

For each t, for each s^t, we introduce N_1 and N_2 to count the number of state-1 and state-2 phases, respectively, in the sequence s^t:

$$N_1(s^t) = |\{1 \le k \le t : \, s_{k-1} = 2, \, s_k = 1\}| \tag{8}$$

$$N_2(s^t) = |\{1 \le k \le t : \, s_{k-1} = 1, \, s_k = 2\}|, \tag{9}$$

where, since the first phase is state-1 phase, we set $s_0 \equiv 2$. Note that functions N_1 and N_2 are, strictly speaking, dependent on time t (this dependence is observed in their domain sets $\mathcal{S}^t$ which clearly change with time t). However, for reasons of easier readibility, we suppress this dependence in the notation, as we also do for all the subsequently defined quantities. We remark that, for any sequence s^t, if the last state $s_t = 2$, then $N_1(s^t) = N_2(s^t)$, and if $s_t = 1$, then $N_1(s^t) = N_2(s^t) + 1$. Finally, $N(s^t)$ is the total number of phases in s^t, $N \equiv N_1 + N_2$.

We further define the sets $\mathcal{T}_{mn}(s^t)$ that contain time indices for the n-th state-m phase, $n = 1, ..., N_m(s^t)$, $m = 1, 2$; to compactly express the likelihood ratio (see expression (27) further ahead), it will also be of interest to group the $\mathcal{T}_{m,n}$s to $\mathcal{T}_m(s^t) := \cup_{n=1}^{N_m(s^t)} \mathcal{T}_{mn}(s^t)$, with its cardinality denoted by $\tau_m(s^t)$, for $m = 1, 2$. We now go over each state phase $\mathcal{T}_{m,n}$, $m = 1, 2$, and increase the counter corresponding to this phase duration, $d = |\mathcal{T}_{m,n}|$,

$$N_{1d}(s^t) = \sum_{n=1}^{N_1(s^t)} 1_{\{|\mathcal{T}_{1n}|=d\}}(s^t), \text{ for } d = 1, ..., \Delta_1, \tag{10}$$

$$N_{2d}(s^t) = \sum_{n=1}^{N_2(s^t)} 1_{\{|\mathcal{T}_{2n}|=d\}}(s^t), \text{ for } d = 1, ..., \Delta_2; \tag{11}$$

i.e., in words, vectors $(N_{m1}, ..., N_{m\Delta_m})$, $m = 1, 2$, represent histograms of phase 1 and phase 2 durations. It is easy to see that $N_m = \sum_{d=1}^{\Delta_m} N_{md}$, for $m = 1, 2$. Also, for each time t and each sequence s^t, the total number of state 1 and state 2 occurrences must sum up to t, and therefore $\sum_{d=1}^{\Delta_1} d\, N_{1d}(s^t) + \sum_{d=1}^{\Delta_2} d\, N_{2d}(s^t) = t$.

Figure 2 shows an example of simulation signals under Hypothesis $\mathcal{H}_1$ with $\Delta_1 = \Delta_2 = 10, \mu_1 = 3, \mu_2 = 5$ and $\sigma = 0.05$ using random duration model for various switching times T, difference process durations $D_{m,n}$ and numbers of different state-phases with fixed duration $N_{m,d}$. We can see from the figure that $D_{1,1} = T_1 - T_0 = 8$ as shown in eq. (2) and there is only one state-phase 1 last for 8 samples, hence $N_{1,8} = 1$. Again, from eq. (3) we can see from the figure again that $D_{2,1} = T_2 - T_1 = 8$ and $D_{2,3} = T_6 - T_5 = 8$. Thus $N_{2,8} = 2$ for there are two state-phase 2 that last for 8 samples.

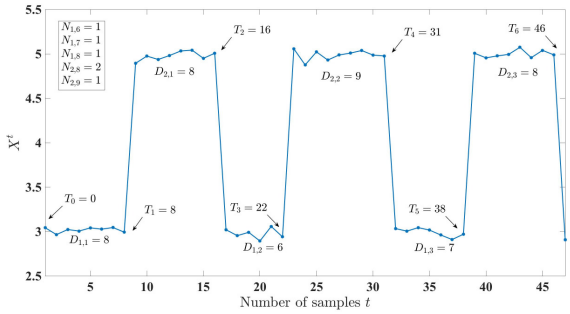

Fig. 2: Example of simulation signals with $\Delta_1 = \Delta_2 == 10, \mu_1 = 3, \mu_2 = 5$ and $\sigma = 0.05$ and various T, $D_{k,i}$, and $N_{k,d}$.

To simplify the notation, let $o(s^t)$ return the duration of the last phase in the sequence s^t, and note also that s_t returns the type of the last phase in s^t. The next lemma computes the probability of a given sequence s^t, $P(s^t) = \mathbb{P}_1(S^t = s^t)$.

Lemma 1. *For any sequence s^t, there holds*

$$P(s^t) = \frac{p^+_{s_t\, o(s^t)}}{p_{s_t\, o(s^t)}} \prod_{d=1}^{\Delta_1} p_{1d}^{N_{1d}(s^t)} \prod_{d=1}^{\Delta_2} p_{2d}^{N_{2d}(s^t)}, \tag{12}$$

where by p^+_{ml} we shortly denote $p^+_{ml} = p_{ml} + p_{ml+1} + ... + p_{m\Delta_m}$, for $l = 1, 2, ..., \Delta_m$ and $m = 1, 2$.

The proof of Lemma 1 is given in the extended version of the paper [40]. Further, to simplify the analysis, in what follows we will assume that $\Delta_1 = \Delta_2 =: \Delta$.

Let C_t denote the cardinality of the set of all feasible sequences of states $\mathcal{S}^t$. When p_1 and p_2 are strictly greater than zero, it can be shown that C_t equals the number of ways in which integer t can be partitioned with parts bounded by Δ. This number is known as the Δ-generalized Fibonacci number, and is computed via the following recursion:

$$C_t = C_{t-1} + \ldots + C_{t-\Delta}, \tag{13}$$

with the initial condition $C_1 = 1$. The recursion in (13) is linear and hence can be represented in the form $\widetilde{C}_t = A\widetilde{C}_{t-1}$, where $\widetilde{C}_t = [C_t\, C_{t-1}\, \ldots\, C_{t-\Delta+1}]$ and A is a square, $\Delta \times \Delta$ matrix; it can be shown that A is equal to $A = e_1 \mathbb{1}^\top + A_0$, where, we recall, A_0 is the lower shift matrix of dimension Δ. The growth rate of C_t is given by the largest zero of the characteristic polynomial of A, as the next result, which we borrow from [41] asserts.

Lemma 2. *[Asymptotics for Δ-generalized Fibonacci number [41]] For any ϵ, there exists $t_0 = t_0(\epsilon)$ such that for every $t \ge t_0$,*

$$e^{t(\log\psi - \epsilon)} \le C_t \le e^{t(\log\psi + \epsilon)}, \tag{14}$$

where ψ is the unique positive zero of the following polynomial $\psi^\Delta - \psi^{\Delta-1} - \ldots - 1 = 0$.

A. Sequence types

Duration fractions. For $d = 1, 2, ..., \Delta$, let $V_{m,d}$ denote the number of times along a given sequence of states that state-m phase had length d, normalized by time t, i.e.,

$$V_{m,d}(s^t) = \frac{N_{m,d}(s^t)}{t},\ m = 1, 2. \tag{15}$$

For each sequence s^t, we define its type as the $2 \times \Delta$ matrix $V(s^t) := \left((V_1(s^t))^\top ; (V_2(s^t))^\top \right)$, where $V_m(s^t) = (V_{m,1}(s^t), ..., V_{m,\Delta}(s^t))$, for $m = 1, 2$. Recalling N_1 and N_2 (8), which, respectively, count the number of state-1 and state-2 phases along s^t, we see that $N_m = t\mathbb{1}^\top V_m$, $m = 1, 2$.

It will also be of interest to define the fractions of times Θ_1 and Θ_2 that a given sequence of states was in states 1 and 2, respectively,

$$\Theta_m(s_t) = \frac{\tau_m(s^t)}{t},\ m = 1, 2. \tag{16}$$

It is easy to verify that $\Theta_m = \sum_{d=1}^{\Delta} d\, V_{m,d}$, for $m = 1, 2$.

Let $\mathcal{V}_t$ denote the set of all $2 \times \Delta$-tuples of feasible occurrence of type V at time t

$$\mathcal{V}_t = \left\{ \nu = (\nu_1, \nu_2) : \nu = V(s^t),\ \text{for some } s^t \right\}. \tag{17}$$

Note that, as they are defined as normalized versions of quantities $N_{md}(s^t)$, $V_{md}(s^t)$'s also inherit the properties of N_{md}'s:

$$\sum_{d=1}^{\Delta} dV_{1d}(s^t) + dV_{2d}(s^t) = 1;$$
$$0 \le \mathbb{1}^\top V_1(s^t) - \mathbb{1}^\top V_2(s^t) \le 1/t.$$

As $t \to +\infty$, for every $s^t \in \mathcal{S}^t$, the difference between $\mathbb{1}^\top V_1(s^t)$ and $\mathbb{1}^\top V_2(s^t)$ decreases. Motivated by this, we introduce the set

$$\mathcal{V} = \left\{ \nu \in \mathbb{R}_+^{2\times\Delta} :\ \mathbb{1}^\top \nu_1 = \mathbb{1}^\top \nu_2,\ q^\top \nu_1 + q^\top \nu_2 = 1 \right\}, \tag{18}$$

where $q = [1\, 2\, \ldots\, \Delta]^\top$.

For each t, $\nu \in \mathcal{V}_t$, define the set $\mathcal{S}_\nu^t$ that collects all sequences $s^t \in \mathcal{S}^t$ whose type is ν:

$$\mathcal{S}_\nu^t = \left\{ s^t \in \mathcal{S}^t : V(s^t) = \nu \right\} \tag{19}$$

(note that if $\nu \notin \mathcal{V}_t$, then set $\mathcal{S}_\nu^t$ would be empty). Set $\mathcal{S}_\nu^t$ therefore consists of all sequences with the following properties: 1) the first phase is state-1 phase; 2) the total number of state-1 phases is $\mathbb{1}^\top \nu_1\, t$, where the total number of such phases of duration exactly d is given by $\nu_{1,d}\, t$; and 3) the total number of state-2 phases is $\mathbb{1}^\top \nu_2\, t$, where the total number of such phases of duration exactly d is given by $\nu_{2,d}\, t$.

Let $C_{t,\nu}$ denote the cardinality of $\mathcal{S}_\nu^t$. This number is equal to the number of ways in which one can order $\mathbb{1}^\top \nu_1 t$ state-1 phases (of different durations), where each new ordering has to give rise to a different pattern of state occurrences, times the corresponding number for state-2 phases. Since for any d, any permutation of $\nu_{m,d} t$ phases, each of which is of length d, gives the same sequence pattern, $C_{t,\nu}$ is given by the number of permutations with repetitions for state-1 phases times the number of permutations with repetitions for state-2 phases:

$$C_{t,\nu} = \frac{\left(\mathbb{1}^\top \nu_1 t\right)!}{(\nu_{1,1}t)! \cdot \ldots \cdot (\nu_{1,\Delta_1}t)!} \frac{\left(\mathbb{1}^\top \nu_2 t\right)!}{(\nu_{2,1}t)! \cdot \ldots \cdot (\nu_{2,\Delta_2}t)!}. \tag{20}$$

From (20) the following result regarding the growth rate of $C_{t,\nu}$ easily follows (e.g., by Stirling's approximation bounds).

Lemma 3. *For any $\epsilon > 0$ there exists $t_1 = t_1(\epsilon)$ such that for all $t \ge t_1$*

$$e^{t(H(\nu_1)+H(\nu_2)-\epsilon)} \le C_{t,\nu} \le e^{t(H(\nu_1)+H(\nu_2)+\epsilon)}, \tag{21}$$

where $H : \mathbb{R}_+^\Delta \mapsto \mathbb{R}$ is defined as

$$H(\lambda) = -\sum_{d=1}^{\Delta} \frac{\lambda_d}{\mathbb{1}^\top \lambda} \log \frac{\lambda_d}{\mathbb{1}^\top \lambda}, \tag{22}$$

where λ_d denotes the d-th element of an arbitrary vector $\lambda \in \mathbb{R}_+^\Delta$.

We end this section by giving some well-known results from the theory of large deviations that we will use in our analysis of detection problem (4).

B. Varadhan's lemma and large deviations principle

We first state the definition of the large deviations principle (LDP) for an arbitrary sequence of random measures (see eq. (51) further ahead in Section IV for the sequence of random measures that will be analyzed in the paper). We remark that this definition differs from the standard LDP (i.e., the LDP for a deterministic sequence of measures). In particular, we require that, for every large deviation set, there exists a probability one set (with respect to the probability space that generates the random sequence of measures) such that, on this set, the corresponding lower and upper large deviations bounds hold with a certain rate function. Or, alternatively put, for every large deviation set, the two LDP bounds hold with probability one (and, of course, with the same rate function).

Large deviations principle.

Definition 4 (Large deviations principle [26] with probability 1)**.** *Let $\mu_t^\omega : \mathcal{B}\left(\mathbb{R}^D\right)$ be a sequence of Borel random measures defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, μ_t^ω, $t = 1, 2, ...$ satisfies the large deviations principle with probability one, with rate function I if the following two conditions hold:*

1) *for every closed set F there exists a set $\Omega_F^\star \subseteq \Omega$ with $\mathbb{P}\left(\Omega_F^\star\right) = 1$, such that for each $\omega \in \Omega_F^\star$,*

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mu_t^\omega(F) \le -\inf_{x\in F} I(x); \tag{23}$$

2) *for every open set E there exists a set $\Omega_E^\star \subseteq \Omega$ with $\mathbb{P}\left(\Omega_E^\star\right) = 1$, such that for each $\omega \in \Omega_E^\star$,*

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mu_t^\omega(E) \ge -\inf_{x\in E} I(x). \tag{24}$$

We give here the version of the Varadhan's lemma which involves sequence of random probability measures and large deviations principle (LDP) with probability one[3].

Lemma 5 (Varadhan's lemma [26])**.** *Suppose that the random sequence of measures μ_t^ω satisfies the LDP with probability one, with rate function I, as defined in Def. 4. Then, if for function F the tail condition below holds with probability one,*

$$\lim_{B\to+\infty} \limsup_{t\to+\infty} \frac{1}{t} \log \int_{x:F(x)\geq B} e^{tF(x)} d\mu_t^\omega(x) = -\infty, \quad (25)$$

then, with probability one,

$$\lim_{t\to+\infty} \frac{1}{t} \log \int_x e^{tF(x)} d\mu_t^\omega(x) = \sup_{x\in\mathbb{R}^D} F(x) - I(x). \quad (26)$$

IV. Linear recursion for the LLR and the Lyapunov exponent

From (5) and (12), it is easy to see that the likelihood ratio can be expressed through the defined quantities as:

$$\begin{aligned} L_t(X^t) &= \sum_{s^t\in\mathcal{S}^t} P(s^t) e^{\frac{1}{\sigma^2}\sum_{m=1}^2 \mu_m \sum_{k\in\mathcal{T}_m(s^t)} X_k - \tau_m(s^t)\frac{\mu_m^2}{2\sigma^2}} \\ &= \sum_{s^t\in\mathcal{S}^t} \frac{p^+_{s_t,o(s^t)}}{p_{s_t,o(s^t)}} e^{\sum_{m=1}^2\sum_{d=1}^{\Delta_m} N_{md}(s^t)\log p_{md}} \times \\ &\quad e^{\frac{1}{\sigma^2}\sum_{m=1}^2 \mu_m \sum_{k\in\mathcal{T}_m(s^t)} X_k - \tau_m(s^t)\frac{\mu_m^2}{2\sigma^2}}. \end{aligned} \quad (27)$$

The expression in (27) is combinatorial, and its straightforward implementation would require computing $C_t \approx e^{\psi t}$ summands. This is prohibitive when the observation interval t is large. In this paper, we unveil a simple, linear recursion form for the likelihood $L_t(X^t)$, for $t = 1, 2, ...$ We give this result in the next lemma. To shorten the notation, we introduce functions $f_m : \mathbb{R} \mapsto \mathbb{R}$, which we define by $f_m(x) := \frac{1}{\sigma^2}\mu_m x - \frac{1}{2\sigma^2}\mu_m^2$, for $x \in \mathbb{R}$ and $m = 1, 2$. Recall that e_1 denotes the first canonical vector in $\mathbb{R}^\Delta$ (the Δ dimensional vector with 1 only in the first position, and having zeros in all other positions), and $\mathbb{1}$ denotes the vector of all ones in $\mathbb{R}^\Delta$.

Lemma 6. *Let $\Lambda_k = \left(\Lambda_k^{1\top}, \Lambda_k^{2\top}\right)^\top$ evolve according to the following recursion*

$$\Lambda_{k+1} = A_{k+1}\Lambda_k, \quad (28)$$

with the initial condition $\Lambda_1 = \left(e^{f_1(X_k)}e_1^\top,\, e^{f_2(X_k)}e_1^\top\right)^\top$, and where, for $k \geq 2$, matrix A_k is given by

$$A_k = \left[\begin{array}{c|c} e^{f_1(X_k)}A_0 & e^{f_1(X_k)}e_1 p_2^\top \\ \hline e^{f_2(X_k)}e_1 p_1^\top & e^{f_2(X_k)}A_0 \end{array}\right], \quad (29)$$

and A_0 is, we recall, the lower shift matrix of dimension Δ. Then, for each $t \geq 1$, the likelihood ratio $L_t(X^t)$ is computed by

$$L_t(X^t) = \sum_{d=1}^{\Delta} p_{1d}^+ \Lambda_{t,d}^1 + p_{2d}^+ \Lambda_{t,d}^2, \quad (30)$$

where $\Lambda_{t,d}^m$ is the d-th element of Λ_t^m, for $d = 1, ..., \Delta$ and $m = 1, 2$.

Remark. We note that the matrix A_k can be further decomposed as

$$A_k = D_k P \quad (31)$$

$$D_k = \operatorname{diag}\left(\left(e^{f_1(X_k)}\mathbb{1}^\top,\, e^{f_2(X_k)}\mathbb{1}^\top\right)^\top\right),\ k = 1, 2, ...,$$

$$P = \left[\begin{array}{cccc:cccc} 0 & & \dots & 0 & p_{21} & p_{22} & \dots & p_{2\Delta} \\ 1 & 0 & \dots & 0 & & & & \\ 0 & 1 & \dots & 0 & & & 0 & \\ \vdots & \vdots & \ddots & \vdots & & & & \\ 0 & 0 & \dots & 1 & 0 & & & \\ \hdashline p_{21} & p_{22} & \dots & p_{2\Delta} & 0 & & \dots & 0 \\ & & & & 1 & 0 & \dots & 0 \\ & & 0 & & 0 & 1 & \dots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & 0 & 0 & \dots & 1 \end{array}\right]$$

i.e., D_k is a random diagonal matrix of size 2Δ, modulated by the k-th measurement X_k, and P is a sparse, constant matrix of the same dimension, which defines transitions from the current state pattern to the one in the next time step.

Proof intuition. The intuition behind this recursive form is the following. We break the sum in (27) into sequences s^t whose last phases are of the same type. For sequences that end with state $m = 1$, $\Lambda_{t,d}^1$ represents the contribution to the overall likelihood ratio $L_t(X^t)$ of all such sequences whose last phase is of length d, and similarly for $\Lambda_{t,d}^2$. Once the vectors $\Lambda_{t,d}^1$ and $\Lambda_{t,d}^2$ are defined, their update is simple. Consider the value $\Lambda_{t+1,d}^1$, where $d > 1$; this value corresponds to the likelihood ratio contribution of all sequences s^{t+1} that end with state-1 phase of duration d. Since $d > 1$, the only possible way to get a sequence of that form is to have a sequence at time t that ends with the same state, where the duration of the last phase is $d-1$. This translates to the update $\Lambda_{t+1,d}^1 = e^{f_1(X_{t+1})}\Lambda_{t,d-1}^1$, where the choice of f_1 in the exponent is due to the fact that the last state is $s_{t+1} = 1$; see also the first line in (29). On the other hand, if $d = 1$, then the state at time t must have been $m = 2$. The duration of this previous phase could have been arbitrary from $d = 1$ to $d = \Delta$. Hence $\Lambda_{t+1,1}^1$ is computed as the sum $\Lambda_{t+1,1}^1 = \sum_{d=1}^{\Delta} p_{2d} e^{f_1(X_{t+1})}\Lambda_{t,d}^2$, where the probabilities p_{2d} are used to mark that the previous phase is completed, see the second line in (29). The analysis for $\Lambda_{t+1,d}^2$ is similar. The formal proof of Lemma 6 is given in the extended version of the paper [40].

[3]We note one technical subtlety in Def. 4. It would be analytically "cleaner" to require the existence of a probability one set, say $\Omega^\star \subseteq \Omega$, on which the LDP bounds hold for an arbitrary large deviation set. This is, however, too restrictive for our purposes, and thus we relax this condition to the existence of such a set for each given large deviation set, but requiring, of course, that we have the same rate function for each of the obtained large deviation probabilities. As it turns, this condition is sufficient to yield Varadhan's lemma with probability 1; see [40] for details.

A. Transition matrix P and error exponent upper bound

The matrix P defined in (31) has a nice physical interpretation. Namely, define the probabilities that the transition from one state to the other occurs exactly at time t, $q_{t,1} = \mathbb{P}(S_t = 1, S_{t-1} = 2)$ and $q_{t,2} = \mathbb{P}(S_t = 2, S_{t-1} = 1)$, for $t \geq 1$. Conditioning on the duration of the state that just ended, it is easy to see that these two probabilities, in the next time step, are computed by

$$\begin{aligned} q_{t+1,1} &= \sum_{d=1}^{\Delta} \mathbb{P}\left(S_{t+1} = 1, S_t = \ldots = S_{t-d+1} = 2 \right| \\ &\qquad \left. S_{t-d+1} = 2, S_{t-d} = 1\right) \mathbb{P}\left(S_{t-d+1} = 2, S_{t-d} = 1\right) \\ &= \sum_{d=1}^{\Delta} p_{2d} q_{t-d+1,2}, \end{aligned} \tag{32}$$

and similarly for $q_{t+1,2}$. Since we assume that the first state is always state 1, and taking for convenience that the Δ states preceding S_1 are state 2, i.e., $S_0 = S_1 = ... = S_{-\Delta+1} = 2$, we have initialization $q_{1,1} = 1$ and $q_{1,1-d} = 0$, for $d = 1, ..., \Delta - 1$, and $q_{2,1-d} = 0$, for $d = 0, ..., = \Delta - 1$. Forming the 2Δ vector $\mathbf{q}_t = [\mathbf{q}_{t,1}^\top \mathbf{q}_{t,2}^\top]^\top$, where $\mathbf{q}_{t,m} = [q_{t,m}, q_{t-1,m}, ..., q_{t-\Delta+1,m}]^\top$, for $m = 1, 2$, we have the following transition relation

$$\mathbf{q}_{t+1} = P \mathbf{q}_t, \tag{33}$$

for $t = 0, 1, 2, ...$, where $\mathbf{q}_0 = [e_1^\top, 0_\Delta^\top]^\top$. It is easy to verify that the transition matrix P satisfies the following properties.

Proposition 7. 1) *P is stochastic and irreducible;*
2) *the left Perron eigenvector of P is the vector $p^+ = \left[{p_1^+}^\top, {p_2^+}^\top\right]^\top$, where the d-th entry of p_m^+ equals p_{md}^+, for $m = 1, 2$, $d = 1, 2, ..., \Delta$.*

The fact that P is stochastic follows directly from the structure of P, by using the fact that vectors p_1 and p_2 have entries that sum up to one, and irreducibility follows by the assumption that $p_1, p_2 > 0$ (entry-wise). Property 2 can be verified directly (note that $p_{11}^+ = p_{21}^+ = 1$).

Upper bound on the error exponent. We use the transition formula (33), together with the properties of P, to derive an upper bound on the error exponent (6), which we give in the following lemma and prove in the Appendix.

Lemma 8. *There holds*

$$\zeta \leq \frac{q^\top p_1}{q^\top p_1 + q^\top p_2} \frac{\mu_1^2}{2\sigma^2} + \frac{q^\top p_2}{q^\top p_1 + q^\top p_2} \frac{\mu_2^2}{2\sigma^2}. \tag{34}$$

Expected SNR interpretation. Interpretation of the upper bound (34) is highly intuitive. The factor $q^\top p_1/(q^\top p_1 + q^\top p_2)$ represents the fraction of times that the process spends in state 1, and similarly for $q^\top p_2/(q^\top p_1 + q^\top p_2)$. Thus, the right hand side of (34) is in that sense the average SNR of the observed signal sequence. If we consider any typical sequence of states, and if we assumed the perfect knowledge of this sequence, then the error exponent would be given by the right hand side of (34) (we remark that any typical sequence of states will have approximately the same SNR, as given in (34)). Since in our scenario we have a more complex problem where we only have the observations (and not the underlying states), it is natural to expect that the corresponding error exponent is upper bounded by the error exponent for the case when both the observations and the states are available – equal to the right hand side of (34).

B. Error exponent ζ as Lyapunov exponent

From Lemma 6 we see that L_t can be represented as a linear function of the matrix product $\Pi_t := A_t \cdot \ldots \cdot A_1$,

$$L_t = {p^+}^\top \Pi_t \Lambda_0, \tag{35}$$

where A_k are matrices of the form (29). Each A_k is modulated by the measurement X_k obtained at time k. Since X_k's, $k = 1, 2, ...$, are i.i.d., it follows that the matrices A_k are i.i.d. as well. Applying a well-known result from the theory of random matrices, see Theorem 2 in [42], to sequence A_k it follows that the sequence of the negative values of the normalized log-likelihood ratios $-\frac{1}{t}\log L_t$, $t = 1, 2, ...$, converges to the Lyapunov exponent of the matrix product Π_t. This result is given in Lemma 9 and proven in Appendix.

Lemma 9. *With probability one,*

$$\lim_{t\to+\infty} \frac{1}{t} \log \|\Pi_t\| = \lim_{t\to+\infty} \frac{1}{t} \mathbb{E}_0 \left[\log \|\Pi_t\|\right], \tag{36}$$

and thus, with probability one,

$$\zeta = \lim_{t\to+\infty} -\frac{1}{t} \log \|\Pi_t\| = \lim_{t\to+\infty} -\frac{1}{t} \mathbb{E}_0 \left[\log L_t\right]. \tag{37}$$

Lemma 9 asserts that the error exponent for hypothesis testing problem (4) equals the top Lyapunov exponent for the sequence of products Π_t. Computation of the Lyapunov exponent (e.g., for i.i.d. matrices) is a well-known problem in random matrix theory and theory of random dynamical systems, proven to be very difficult to solve, see, e.g., [38]. We instead search for tractable lower bounds that tightly approximate ζ. We base our method for approximating ζ on the right hand-side identity in (37).

V. Main result

Our first step for computing the limit in (37) is a natural one. Since $\mu_1 \geq 0$ is the guaranteed signal level (recall that $\mu_2 > \mu_1 \geq 0$), we assume that the signal was at all times at state 1, and remove the corresponding components of the signal to noise ratio (SNR) $\frac{\mu_1^2}{2\sigma^2}$ and the signal sum $\sum_{k=1}^t X_k$ from the likelihood ratio. This manipulation then gives us a lower bound on the error exponent. By doing so, we arrive at an equivalent problem to problem (4) just with $\mu_1 = 0$. Mathematically, we have

$$\begin{aligned} L_t(X^t) = &\sum_{s^t \in \mathcal{S}^t} P(s^t) e^{\frac{1}{\sigma^2}\mu_1 \left(\sum_{k=1}^t X_k - \sum_{k\in\mathcal{T}_2(s^t)} X_k\right) - (t - \tau_2(s^t))\frac{\mu_1^2}{2\sigma^2}} \times \\ &\times e^{\frac{1}{\sigma^2}\mu_2 \sum_{k\in\mathcal{T}_2(s^t)} X_k - \tau_2(s^t)\frac{\mu_2^2}{2\sigma^2}} = e^{\frac{1}{\sigma^2}\mu_1 \sum_{k=1}^t X_k - t\frac{\mu_1^2}{2\sigma^2}} \times \\ &\times \sum_{s^t \in \mathcal{S}^t} P(s^t) e^{\frac{1}{\sigma^2} \sum_{k\in\mathcal{T}_2(s^t)} (\mu_2 - \mu_1) X_k - \tau_2(s^t)\frac{\mu_2^2 - \mu_1^2}{2\sigma^2}}. \end{aligned} \tag{38}$$

Taking the logarithm, dividing by t, and computing the expectation with respect to hypothesis $\mathcal{H}_0$, we get

$$\frac{1}{t}\mathbb{E}_0\left[\log L_t(X^t)\right] = -\frac{\mu_1^2}{2\sigma^2} + \frac{1}{t}\mathbb{E}_0\left[\log \sum_{s^t\in\mathcal{S}^t} P(s^t)\times \right.$$
$$\left. \times\ e^{\frac{1}{\sigma^2}\sum_{k\in\mathcal{T}_2(s^t)}(\mu_2-\mu_1)X_k - \tau_2(s^t)\frac{\mu_2^2-\mu_1^2}{2\sigma^2}}\right], \quad (39)$$

where we used that $\mathbb{E}_0[X_k]=0$, for all k, see (4). Taking the limit as $t\to+\infty$, we obtain

$$\zeta = \frac{\mu_1^2}{2\sigma^2} + \eta, \quad (40)$$

where η is given by the following limit

$$\eta = \lim_{t\to+\infty} -\frac{1}{t}\mathbb{E}_0\left[\log \sum_{s^t\in\mathcal{S}^t} P(s^t)\times \right.$$
$$\left. \times e^{\frac{1}{\sigma^2}\sum_{k\in\mathcal{T}_2(s^t)}(\mu_2-\mu_1)X_k - \tau_2(s^t)\frac{\mu_2^2-\mu_1^2}{2\sigma^2}}\right], \quad (41)$$

the existence of which is guaranteed by (37), in Lemma 9. From now on, we focus on computing η.

For $\lambda\in\mathbb{R}^\Delta$, and $p\in\mathbb{S}^{\Delta-1}$, introduce the relative entropy function $D(\lambda||p) := \sum_{d=1}^{\Delta}\frac{\lambda_d}{\mathbb{1}^\top\lambda}\log\frac{\lambda_d/(\mathbb{1}^\top\lambda)}{p_d}$.

Theorem 10. *There holds $\underline{\eta}+\frac{\mu_1^2}{2\sigma^2}\le\zeta$, where $\underline{\eta}$ is the optimal value of the following optimization problem*

$$\begin{array}{ll} \text{minimize} & G(\nu,\xi) \\ \text{subject to} & H(\nu_1)+H(\nu_2)\ge\frac{\xi^2}{2\theta_2\sigma^2} \\ & \theta_2 = q^\top\nu_2 \\ & \nu\in\mathcal{V} \\ & \xi\in\mathbb{R}. \end{array}, \quad (42)$$

where $G(\nu) = D(\nu_1||p_1) + D(\nu_2||p_2) + \frac{\theta_2}{2\sigma^2}\left(\frac{\xi}{\theta_2}-(\mu_2-\mu_1)\right)^2 + \theta_2\frac{\mu_1(\mu_2-\mu_1)}{\sigma^2}$, for $\nu\in\mathbb{R}_+^{2\Delta}$, $\xi\in\mathbb{R}$.

Guaranteed error exponent. Since each of the terms in the objective function of (42) is non-negative, its optimal value is lower bounded by 0. Using relation (40), we obtain that the value of the error exponent is lower bounded by the value of SNR in state-1, $\frac{\mu_1^2}{2\sigma^2}$, i.e.,

$$\zeta \ge \frac{\mu_1^2}{2\sigma^2}. \quad (43)$$

The preceding bound holds for any choice of parameters Δ, p_1, p_2, μ_1 and μ_2. This result is very intuitive, as it mathematically formalizes the reasoning that, no matter which configuration of states occurs, signal level μ_1 is always guaranteed, and hence the corresponding value of error exponent $\frac{\mu_1^2}{2\sigma^2}$ is ensured. In that sense, any appearance of state 2 (i.e., signal level $\mu_2>\mu_1$) can only increase the error exponent.

A. Special case $\mu_1 = 0$ and detectability condition

When the signal level in state 1 equals zero, then, since the statistics of X_k for $S_k=1$ is the same as its statistics under $\mathcal{H}_0$, effectively we can have information on the state of nature $\mathcal{H}_1$ only when state $S_k=2$ occurs. Denoting $\mu=\mu_2$, optimization problem (42) then simplifies to:

$$\begin{array}{ll} \text{minimize} & D(\nu_1||p_1)+D(\nu_2||p_2)+\frac{\theta_2}{2\sigma^2}\left(\frac{\xi}{\theta_2}-\mu\right)^2 \\ \text{subject to} & H(\nu_1)+H(\nu_2)\ge\frac{\xi^2}{2\theta_2\sigma^2} \\ & \theta_2 = q^\top\nu_2 \\ & \nu\in\mathcal{V} \\ & \xi\in\mathbb{R}. \end{array}. \quad (44)$$

From (44) we obtain the following condition for detectability of process S_k:

$$H(p_1)+H(p_2) \ge \frac{q^\top p_2}{q^\top p_1 + q^\top p_2}\frac{\mu^2}{2\sigma^2}, \quad (45)$$

i.e., if the inequality above holds, then the optimal value of optimization problem (44) is zero. To see why this holds, note that the point $(\nu_1,\nu_2,\xi)\in\mathbb{R}^{2\Delta+1}$, where $\nu_m = p_m/(q^\top p_1 + q^\top p_2)$, $m=1,2$, and $\xi = q^\top p_2/((q^\top p_1+q^\top p_2))\mu$ under which the cost function of (44) vanishes, under condition (45) belongs to the constraint set of (44). Thus, under condition (45), the lower bound on the error exponent $\underline{\eta}$ is zero, indicating that the process S_k is not detectable. To further illustrate this condition, note that the left hand-side corresponds to the entropy of the process S_k, and the right hand-side corresponds to the expected, i.e. – long-run SNR of the measured signal ($q^\top p_2/\left(q^\top p_1+q^\top p_2\right)$ is the expected fraction of times that the process was in state 2, and $\frac{\mu^2}{2\sigma^2}$ is the SNR for this state). Condition (45) therefore asserts that, if the entropy of the process S_k is too high compared to the expected, or long-run, SNR, then it is not possible to detect its presence. Intuitively, if the dynamics of the phase durations is too stochastic, then it is not possible to estimate the locations of state 2 occurrences, in order to perform the likelihood ratio test. However, on the other hand, if the SNR is very high (e.g., the level μ is high compared to the process noise σ^2) then, whenever state 2 occurs, the signal will make a sharp increase and can therefore be easily detected. The condition in this sense quantitatively characterizes the threshold between the two physical quantities which makes detection possible.

Reformulation of (44)**.** In this subsection we show that optimization problem (44) admits a simplified form, obtained by suppressing the dependence on ξ through inner minimization over this variable. To simplify the notation, introduce $H(\nu) = H(\nu_1)+H(\nu_2)$ and $R(\nu) = q^\top\nu_2\frac{\mu^2}{2\sigma^2}$; note that the function R has the physical meaning of the expected SNR of the S_t process that we wish to detect, for a given sequence type ν.

Lemma 11. *Suppose that $H(p_1)+H(p_2) < q^\top p_2/\left(q^\top p_1+q^\top p_2\right)\frac{\mu^2}{2\sigma^2}$. Then, optimization problem* (44) *is equivalent to the following optimization problem:*

$$\begin{array}{ll} \text{minimize} & D(\nu_1||p_1)+D(\nu_2||p_2)+\left(\sqrt{H(\nu)}-\sqrt{R(\nu)}\right)^2 \\ \text{subject to} & H(\nu)\le R(\nu) \\ & \nu\in\mathcal{V} \end{array}. \quad (46)$$

The proof is given in the Appendix.

VI. Proof of Theorem 10

Sum of conditionals as an expectation. For each $s^t \in \mathcal{S}_t$, introduce

$$\mathcal{X}_{s^t} = \frac{1}{t}\sum_{k\in\mathcal{T}_2} X_k, \tag{47}$$

and note that, for each s^t and under $\mathcal{H} = \mathcal{H}_0$, $\mathcal{X}_{s^t}$ is Gaussian random variable of mean zero and variance equal to $\sigma^2\tau_2(s^t)/t^2 = \sigma^2\theta_2(s^t)/t$. The idea is to view the sum in (41) as an expectation of a certain function $g_{\mathcal{X}} : \mathcal{S}_t \mapsto \mathbb{R}$ defined over the set $\mathcal{S}_t$ of all possible sequences s^t, parameterized by random family (i.e., vector) $\mathcal{X} = \{\mathcal{X}_{s^t} : s^t \in \mathcal{X}^t\}$. More precisely, consider the probability space with the set of outcomes $\mathcal{S}_t$ and where an element s^t of $\mathcal{S}_t$ is drawn uniformly at random – and hence with probability $1/C_t$, where, we recall $C_t = |\mathcal{S}^t|$; denote the corresponding expectation by $\mathbb{E}_U$. We see that the sum under the logarithm in (41) equals

$$\begin{aligned}&\sum_{s^t\in\mathcal{S}^t} P(s^t)e^{t\frac{(\mu_2-\mu_1)}{\sigma^2}\mathcal{X}_{s^t}-\tau_2(s^t)\frac{\mu_2^2-\mu_1^2}{2\sigma^2}}\\ &= C_t\sum_{s^t\in\mathcal{S}^t}\frac{1}{C_t}g_{\mathcal{X}}(s^t) = C_t\,\mathbb{E}_U\left[g_{\mathcal{X}}(s^t)\right],\end{aligned} \tag{48}$$

where it is easy to see that $g_{\mathcal{X}}(s^t) = P(s^t)e^{t\frac{(\mu_2-\mu_1)}{\sigma^2}\mathcal{X}_{s^t}-\tau_2(s^t)\frac{\mu_2^2-\mu_1^2}{2\sigma^2}}$, for $s^t \in \mathcal{S}_t$.

Using further the type V defined in Subsection III-A, we can express $g_{\mathcal{X}}(s^t)$ as

$$g_{\mathcal{X}}(s^t) = e^{t\frac{(\mu_2-\mu_1)}{\sigma^2}\mathcal{X}_{s^t}-t\Theta_2(s^t)\frac{\mu_2^2-\mu_1^2}{2\sigma^2}+t\sum_{m=1}^{2}\sum_{d=1}^{\Delta}V_{md}(s^t)\log p_{md}}, \tag{49}$$

where we assumes that $o(s^t) = \Delta$, in which case the first factor on the right hand side of (49) equals 1, but we remark that the claims that follow can be derived even without this assumption, by a slightly more technical proof path – we refer the reader to the extended version of the paper [40].

Induced measure. We see that function $g_{\mathcal{X}}$ essentially depends on s^t only through type V of the sequence and the values of vector $\mathcal{X}$. More precisely, define $F : \mathbb{R}^{2\Delta}\times\mathbb{R}\mapsto\mathbb{R}$ as

$$F(\nu,\xi) = \frac{\mu_2-\mu_1}{\sigma^2}\xi - \theta_2\frac{\mu_2^2-\mu_1^2}{2\sigma^2} + \sum_{m=1}^{2}\sum_{d=1}^{\Delta}\nu_{md}\log p_{md}. \tag{50}$$

Then, for any s^t, $g_{\mathcal{X}}(s^t) = e^{F(V(s^t),\mathcal{X}_{s^t})}$. For each vector $\mathcal{X}$, let then $Q_t^{\mathcal{X}} : \mathcal{B}\big(\mathbb{R}^{2\Delta+1}\big)\mapsto\mathbb{R}$ denote the probability measure induced by $(V(s^t),\mathcal{X}(s^t))$, for the assumed uniform measure on $\mathcal{S}_t$:

$$Q_t^{\mathcal{X}}(B) := \frac{\sum_{s^t\in\mathcal{S}^t}1_{\{(V,\mathcal{X})\in B\}}(s^t)}{C_t}, \tag{51}$$

for arbitrary $B \in \mathcal{B}\Big(\mathbb{R}^{N^2+N}\Big)$. It is easy to verify that $Q_t^{\mathcal{X}}$ is indeed a probability measure. Also, we note that, for any fixed t and $\mathcal{X}$, $Q_t^{\mathcal{X}}$ is discrete, supported on the discrete set $\{(V(s^t),\mathcal{X}_{s^t}) : s^t \in \mathcal{S}_t\}$; note that the latter set is a subset of $\mathcal{V}^t\times\cup_{s^t\in\mathcal{S}_t}X_{s^t}$ – the Cartesian product of the set of all feasible types at time t with the set of all elements of vector $\mathcal{X}$.

Let $\mathbb{E}_Q$ denote the expectation with respect to measure $Q_t^{\mathcal{X}}$. Then, we have $\mathbb{E}_U\left[g_{\mathcal{X}}(S^t)\right] = \mathbb{E}_Q\left[e^{tF(V,\mathcal{X})}\right]$. Going back to (48), and using the result of Lemma 2, we obtain for η given in (41):

$$\eta = -\log\psi + \lim_{t\to+\infty} -\frac{1}{t}\mathbb{E}_0\left[\log\mathbb{E}_Q\left[e^{tF(V,\mathcal{X})}\right]\right], \tag{52}$$

where, we recall $\mathbb{E}_0$ is the expectation with respect to probability $\mathbb{P}_0$ that corresponds to $\mathcal{H}_0$ state of nature, under which measurements X_k – and hence vector $\mathcal{X}$ are generated.

If the measures $Q_t^{\mathcal{X}}$ were sufficiently nice such that they satisfied the LDP and the moderate growth condition (25), then one could apply Varadhan's lemma to compute the exponential growth of the expectation in the right hand side of (52). However, the measures $Q_t^{\mathcal{X}}$ are very difficult to analyze due to the correlations in different elements of $\mathcal{X}$ which couple the indicator functions in (51). Hence, we resort to an upper bound of η which we derive by replacing vector $\mathcal{X}$ by vector $\mathcal{Z}$ with the same statistical properties, but with an added feature that its elements are mutually independent. More precisely, for each t we introduce a family of independent Gaussian variables $\mathcal{Z} = \{\mathcal{Z}_{s^t} : s^t \in \mathcal{S}^t\}$. Further, for each s^t the corresponding element of the family $\mathcal{Z}_{s^t}$ is Gaussian with the same mean and variance as $\mathcal{X}_{s^t}$: expected value equal to 0, and variance equal to $\mathrm{Var}\left[\mathcal{Z}_{s^t}\right] = \sigma^2\theta_2(s^t)/t$. Denote by $\mathbb{P}$ and $\mathbb{E}$, respectively, the probability function and the expectation corresponding to the family $\{\{\mathcal{Z}_{s^t} : s^t \in \mathcal{S}^t\} : t = 1,2,\ldots\}$. Then, the following result holds; the proof is based on Slepian's lemma [43], and it can be found in an extended version of this paper [40].

Lemma 12. *For each t, there holds,*

$$\mathbb{E}\left[\log\mathbb{E}_Q\left[e^{tF(V,\mathcal{Z})}\right]\right] \geq \mathbb{E}_0\left[\log\mathbb{E}_Q\left[e^{tF(V,\mathcal{X})}\right]\right], \tag{53}$$

where the inner left hand side expectation is with respect to the measures $Q_t^{\mathcal{X}}$ and the inner right hand-side expectation is with respect to the measures $Q_t^{\mathcal{Z}}$.

The next result asserts that $Q_t^{\mathcal{Z}}$ satisfies the LDP with probability one and computes the corresponding rate function.

Theorem 13. *For every measurable set G, the sequence of measures $Q_t^{\mathcal{Z}}$, $t = 1,2,...$, with probability one satisfies the LDP upper bound* (23) *and the LDP lower bound* (24)*, with the same rate function $I : \mathbb{R}^{2\Delta+1}\mapsto\mathbb{R}$, equal for all sets G, which for $\nu\in\mathcal{V}$ for which $H(\nu_1)+H(\nu_2) \geq J_\nu(\xi)$ is given by*

$$I(\nu,\xi) = \log\psi - H(\nu_1) - H(\nu_2) + J_\nu(\xi), \tag{54}$$

and equals $+\infty$ otherwise, and where, for any $\nu\in\mathcal{V}$, function $J_\nu : \mathbb{R}\mapsto\mathbb{R}$ is defined as $J_\nu(\xi) := \frac{1}{q^\top\nu_2}\frac{\xi^2}{2\sigma^2}$.

Having the large deviations principle for the sequence $Q_t^{\mathcal{Z}}$, we can invoke Varadhan's lemma to compute the limit of the scaled values in (52). Applying Lemma 5 (the details of the moderate growth condition (25) for $Q_t^{\mathcal{Z}}$ are given in the extended version of this paper [40]), we obtain that, with probability one,

$$\lim_{t\to+\infty}\frac{1}{t}\log\mathbb{E}_Q\left[e^{tF(V,\mathcal{Z})}\right] = \sup_{(\nu,\xi)}F(\nu,\xi) - I(\nu,\xi). \tag{55}$$

It can be shown that the sequence under the preceding limit is uniformly integrable, the proof of which can be found in the extended version of this paper [40]. Thus, the limit of the sequence values and the limit of their expected values coincide, i.e.,

$$\lim_{t\to+\infty}\frac{1}{t}\mathbb{E}\left[\log\mathbb{E}_Q\left[e^{tF(V,\mathcal{Z})}\right]\right]=\lim_{t\to+\infty}\frac{1}{t}\log\mathbb{E}_Q\left[e^{tF(V,\mathcal{Z})}\right]. \tag{56}$$

Combining with (52), (53), and (55), we finally obtain

$$\eta\geq-\log\psi-\sup_{(\nu,\xi)\in\mathbb{R}^{2\Delta+1}}F(\nu,\xi)-I(\nu,\xi). \tag{57}$$

It remains to show that the value of the above supremum equals the value of the optimization problem (42). Using the definition of I, we have that $I(\nu,\xi)=+\infty$ for any (ν,ξ) such that $H(\nu)<J_\theta(\xi)$ or such that $\nu\notin\mathcal{V}$. Since the supremum is surely not achieved at these points, set $\mathbb{R}^{2\Delta+1}$ in (57) can be replaced by $\{(\nu,\xi)\in\mathcal{V}\times\mathbb{R}:H(\nu)<J_\theta(\xi)\}$. Using the definitions of F and I, we have

$$\begin{aligned}F(\nu,\xi)-I(\nu,\xi)=&\sum_{m=1}^{2}\sum_{d=1}^{\Delta}\nu_{md}\log p_{md}-\nu_{md}\log\nu_{md}\\&+\frac{\mu_2-\mu_1}{\sigma^2}\xi-\theta_2\frac{\mu_2^2-\mu_1^2}{2\sigma^2}-\frac{1}{\theta_2}\frac{\xi^2}{2\sigma^2}-\log\psi.\end{aligned} \tag{58}$$

Cancelling out the term $\log\psi$ in the preceding equation with the one in (57), and recognizing that $\sum_{d=1}^{\Delta}\nu_{md}\log p_{md}-\nu_{md}\log\nu_{md}=-D(\nu_m||p_m)$, we see that problem (42) is equivalent to the one in (57). This completes the proof of Theorem 10.

VII. Numerical results

In this section we report our numerical results to demonstrate tightness of the developed performance bounds. We also illustrate our methodology on the problem of detecting one single run of a dish-washer, where we use real-world data to estimate the state values for a dish-washer.

In the first set of simulations, we consider the setup in which $\mu_1>0$ and we compare the error exponents obtained via simulations to the guaranteed lower bound (43). We simulate a two-state signal, X^t, as an i.i.d. Gaussian random variable with standard deviation σ and mean $\mu_1=2$ and $\mu_2=5$ in states 1 and 2, respectively. We take the maximal duration to be $\Delta=3$. The observation interval is $t\in[1,T]$, where $T=200$. In the absence of the signal, the data is distributed according to the Gaussian distribution with mean $\mu_0=0$ and the same standard deviation σ.

To estimate the receiver operating characteristics (ROC) curves, we use $J=100000$ Monte Carlo simulation runs for each hypothesis. For each hypothesis and each simulation run, we compute the values $L_t(X^t)$, for $t=1,2,...,T$, using the linear recursion from Lemma 6. Then, for each t, to obtain the corresponding ROC curve, we first find the minimal and maximum value $\underline{L}_{t,m}$ and $\overline{L}_{t,m}$, respectively, across J runs for each hypothesis m, and change the detection threshold γ with a small step size from $\underline{L}_{t,1}-\beta$ to $\overline{L}_{t,0}+\beta$, where β is a carefully chosen bound. For each t and γ the probability of false alarm P_{fa} or false positive, i.e., wrongly determining that the signal is present, is calculated as

$$P_{\text{fa},t}^{\gamma}=\frac{\sum_{j=1}^{J}\mathbb{1}(L_t(X_{(j)}^t)\geq\gamma)}{J}$$

where $\mathbb{1}$ is an indicator function that returns 1 if the corresponding condition is true and 0 otherwise, and $X_{(j)}^t$ is the j-th realisation of the sequence X^t under $\mathcal{H}_0$. The probability of a miss P_{miss} or false negative, that is, declaring that the signal is not present, though it is, is calculated as:

$$P_{\text{miss},t}^{\gamma}=\frac{\sum_{j=1}^{J}\mathbb{1}(L_t(X_{(j)}^t)<\gamma)}{J}.$$

We set the bound $\alpha=0.01$ and find $P_{\text{miss},t}^{\alpha}=P_{\text{miss},t}^{\gamma^\star}$ where $\gamma^\star$ resulted in the highest probability of a miss that satisfied $P_{\text{fa},t}^{\gamma^\star}\leq\alpha$.

Error exponents for uniform and concentrated distributions. In the first set of experiments, we investigate the dependence of the slope both on the noise variance σ^2 and also on the pmfs p_1 and p_2, for fixed signal levels μ_1 and μ_2. With respect to p_1 and p_2, we start with the uniform distribution, in which case the signal is the most difficult to detect, as each of the state durations is equally likely (the sequence of states has the highest entropy), and thus it is very difficult to detect the locations of state transitions. Then we gradually shift towards the distribution which has the probability of 0.9 on the duration $d=2$ of both states; it is intuitive that with the latter distribution the signal should be easier to detect than with the uniform, as we know that, in any state, the transition occurs, with high probability, after two sampling periods. More precisely, we consider five different cases with respect to the two pmfs: 1) $p_1=p_2=[1/3,\,1/3,\,1/3]$ (uniform distribution); 2) $p_1=p_2=[0.25,\,0.5,\,0.25]$; 3) $p_1=p_2=[0.15,\,0.7,\,0.15]$; 4) $p_1=p_2=[0.1,\,0.8,\,0.1]$; and 5) $p_1=p_2=[0.05,\,0.9,\,0.05]$.

For each of the five cases above, and each different value of σ, we compute the values of $P_{\text{miss},t}^{\alpha}$, for $t=1,...,T$, and apply linear regression on the sequence of values $-\log P_{\text{miss},t}^{\alpha}$ for all observation times t for which the probability of a miss was non-zero. For each of the five cases, this gives an estimate for the error exponent (i.e., the slope) for the probability of a miss under a fixed value of σ, which we denote by $S_\sigma^{(k)}$, $k=1,...,5$.

Figure 3 plots the probability of a miss curve (in the logarithmic scale) vs. the number of samples t for five different values of σ, namely $\sigma=10,15,20,25,30$, for the case when the distributions p_1 and p_2 are uniform, $p_1=p_2=[1/3,\,1/3,\,1/3]$. We observe that for large observation intervals t the curves are close to linear, as predicted by the theory, see Lemma 9. Further, as σ increases the magnitude of the slope decreases becoming very close to 0 for large values of σ.

Figure 4 compares the five error exponents curves $S_\sigma^{(k)}$, $k=1,...,5$, obtained numerically. As expected, as σ increases, each of the curves tends to zero, and they also become closer. Comparing the five curves, we see that, for any fixed noise variance, the lowest curve is always the one with the

uniform pmfs. As the pmf gradually becomes more and more concentrated, the error exponents monotonically increase, until the highest error exponent curve, corresponding to the most concentrated pmf of the five, with state duration $d = 2$ occurring most of the time. This result is expected, as it is easiest to detect the process with the lowest entropy, and hence the corresponding error exponent should be the highest.

Figure 4 also plots the theoretical upper and lower bounds in (43) and (34), respectively; we note that, since $p_1 = p_2$ in each of the five simulation setups, the same upper bound– equal to $1/2\mu_1^2/(2\sigma^2) + 1/2\mu_2^2/(2\sigma^2)$– applies, see eq. (34). The lower bound, equal to $\mu_1^2/(2\sigma^2)$, is plotted in blue dotted line, while the upper bound is plotted in red dashed line. It can be seen from the figure that each of the five numerical error exponent curves is at all points sandwiched between the lower bound (43) curve $\mu_1^2/(2\sigma^2)$ and the upper bound curve $1/2\mu_1^2/(2\sigma^2)+1/2\mu_2^2/(2\sigma^2)$. Further, the closest curve to the lower bound is the error exponent for the uniform distribution, $p_1 = p_2 = [1/3,\ 1/3,\ 1/3]$, which is intuitively expected, as also explained in the above paragraph. The curve closest to the upper bound is the most skewed, i.e., sharpest distribution, $p_1 = p_2 = [0.05,\ 0.9,\ 0.05]$.

In order to get further closer to the theoretical error exponent limit, we shift the probability mass from state duration $d = 2$ to $d = 3$, and simulate the case $p_1 = p_2 = [0.05,\ 0.05,\ 0.9]$. The reasoning is the following: the longer the process stays in the same state, it should be easier to detect it. For completeness, we also simulate the case $p_1 = p_2 = [0.9,\ 0.05,\ 0.05]$, i.e., when the process often switches from one state to the other. The results, shown in Figure 5, are well aligned with the intuition. The lowest of the three curves is the curve corresponding to the fastest switching process, with most of the mass on the shortest possible duration, $d = 1$. The highest curve (and the one closest to the theoretical upper bound) is the curve corresponding to the most inert process, when most of the mass is on the longest state duration, $d = \Delta = 3$, while the curve with the mass concentrated on $d = 2$ is in the middle of the two.

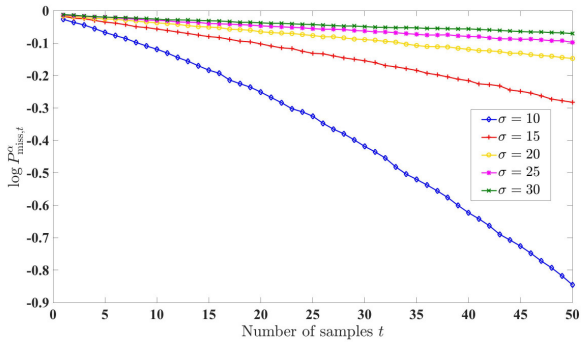

Fig. 3: Simulation setup: $\Delta = 3$, $p_1 = p_2 = [1/3,\ 1/3,\ 1/3]$, $\mu_1 = 2$, $\mu_2 = 5$, $\alpha = 0.01$. Evolution of probability of a miss, in the logarithmic scale, for $\sigma = 10, 15, 20, 25, 30$.

In the second set of experiments, we consider the setup where the signal level in state 1 is zero, $\mu_1 = 0$, and $\mu_2 = \mu = 1$; similarly as in the previous setup, we consider uniform

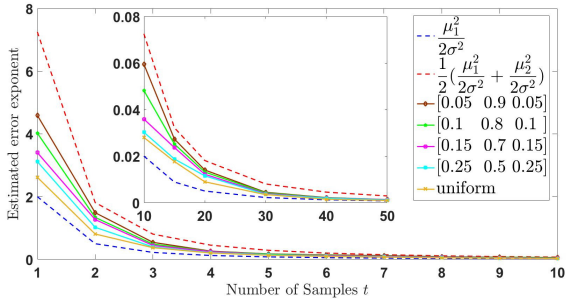

Fig. 4: Simulation setup: $\Delta = 3$, $\mu_1 = 2$, $\mu_2 = 5$, $\alpha = 0.01$. σ varies from 5 to 50. The five middle full lines plot the numerical error exponents estimated from slope of $\log P^{\alpha}_{\text{miss},t}$ vs. σ, for 1) $p_1 = p_2 = [1/3,\ 1/3,\ 1/3]$ (yellow); 2) $p_1 = p_2 = [0.25,\ 0.5,\ 0.25]$ (turquoise); 3) $p_1 = p_2 = [0.15,\ 0.7,\ 0.15]$ (pink); and 4) $p_1 = p_2 = [0.1,\ 0.8,\ 0.1]$ (light green); and 5) $p_1 = p_2 = [0.05,\ 0.9,\ 0.05]$ (brown). Blue dotted line plots the theoretical lower bound $\mu_1^2/(2\sigma^2)$ in (43) and red dashed line plots the upper bound $1/2\mu_1^2/(2\sigma^2) + 1/2\mu_2^2/(2\sigma^2)$ in (34)

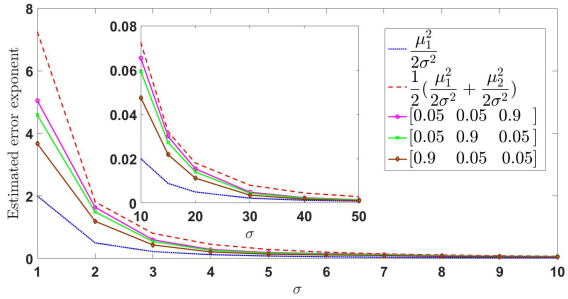

Fig. 5: Simulation setup: $\Delta = 3$, $\mu_1 = 2$, $\mu_2 = 5$, $\alpha = 0.01$. σ varies from 5 to 50. The three middle full lines plot the numerical error exponents estimated from slope of $\log P^{\alpha}_{\text{miss},t}$ vs. σ, for 1) $p_1 = p_2 = [0.9,\ 0.05,\ 0.05]$ (brown); 2) $p_1 = p_2 = [0.05,\ 0.9,\ 0.05]$ (light green); and 3) $p_1 = p_2 = [0.05,\ 0.05,\ 0.9]$ (pink). Blue dotted line plots the theoretical bound $\mu_1^2/(2\sigma^2)$ in (43) and red dashed line plots the upper bound $1/2\mu_1^2/(2\sigma^2) + 1/2\mu_2^2/(2\sigma^2)$ in (34).

distributions $p_1, p_2 \sim \mathcal{U}([1, \Delta])$, with $\Delta = 2$. We compare the numerical error exponent with the one obtained as a solution to optimization problem (46). To solve (46), we apply random search over 10^6 different vectors from set $\mathcal{V}$, and pick the point which gives the smallest value of the objective (and satisfies the constraint in (46)).

Figure 6 plots probability of a miss vs. number of samples t for 5 different values of σ, in the interval from 0.2 to 0.6. Again, we can observe that linearity emerges with the increase of σ. Figure 7, top, compares error exponent estimated from the slope in Figure 6 with the theoretical bound calculated from solving (46). We can see from the plot that the two lines are very close to each other. In fact, we have that the numerical values are slightly below the lower bound values. This seemingly contradictory effect is a consequence of the following. As the probability of a miss curves have a concave shape in this simulation setup (which can be observed from

Figure 6) their slopes continuously increase with the increase of the observation interval. As a consequence, the linear fitting performed on the whole observation interval is underestimating the slope, as it is trying to fit also the region of values where concavity is more prominent. To further investigate this effect, we performed linear fitting of probability of a miss curves only for a region of higher values of t, where emergence of linearity is already evident. In particular, for each different value of σ, we apply linear fitting for $[4/5\, t_{\max}, t_{\max}]$, where $t_{\max}$ is the maximal t for which the probability of a miss is non-zero, and we plot the results in Figure 7, bottom. It can be seen from the figure that the numerical curve got closer to the theoretical curve, indicating that the bound in (46) is very tight or even exact. Finally, it can be seen from Figure 7 (top and bottom) that the value of σ for which the error exponent is equal to zero matches the threshold predicted by the theory, $\sigma^\star = \mu/(2\sqrt{2\log\Delta}) = 0.4247$, obtained from detectability condition (45).

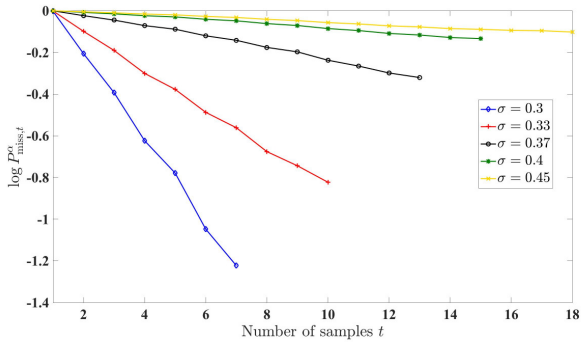

Fig. 6: Simulation setup: $\Delta = 2$, $p_1, p_2 \sim \mathcal{U}([1, \Delta])$, $\mu_1 = 0$, $\mu_2 = 1$, $\alpha = 0.01$. Plots of probability of a miss in the logarithmic scale for $\sigma = 0.3, 0.33, 0.37, 0.4, 0.45$

Comparison with the HMM detector. To illustrate the difference between the HSMM and the HMM, we compare the performance of the optimal HSMM detector derived here with HMM-based detector, derived in [22], see Proposition 1. Namely, we run both detectors on the same data generated by an HSMM model, with certain pmfs. In particular, we set $\Delta = 5$ and consider two sets of simulations: 1) truncated geometric pmfs $p_{1,\text{g}} = p_{2,\text{g}} = 1/(1-q^{\Delta})((1-q),\, q(1-q),\, q^2(1-q),\, q^3(1-q),\, q^4(1-q)) \in \mathbb{R}^5$, where $q = 0.8$; and 2) concentrated pmfs $p_{1,\text{c}} = p_{2,\text{c}} = (0.025,\, 0.025,\, 0.025,\, 0.025,\, 0.9) \in \mathbb{R}^5$, see paragraph on the comparison with random telegraph signal in Section II. In the first case we set the HMM transition matrix as $P_{\text{HMM}} = [q,\, (1-q);\, (1-q),\, q]$, which ensures that the resulting distribution of the state durations will be close to $p_{1,\text{g}} = p_{2,\text{g}}$. Since the data is in this case well fitted by the HMM, we expect that the non-optimal but tuned HMM-based detector will behave close to the optimal HSMM-based detector. In the second case, the pmfs cannot be fitted by a geometric distribution, but we keep the same transition matrix P_{HMM}, as it describes well the property that the process stays the same time in both states. Since in this case the data is far from an HMM, we expect that the optimal HSMM-based detector will outperform the HMM-based detector.

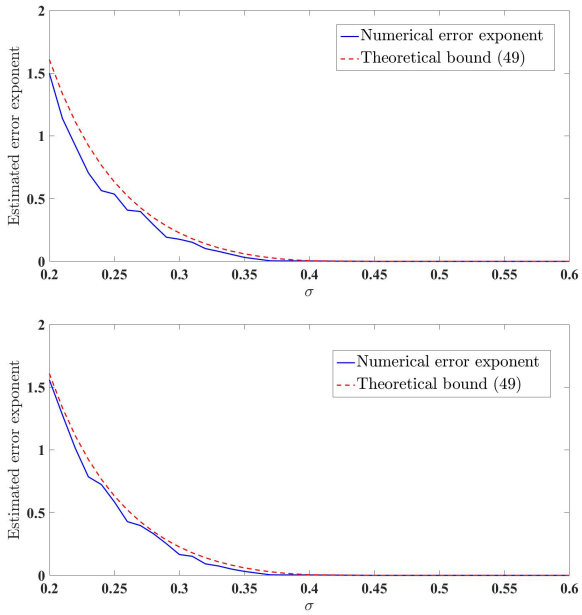

Fig. 7: Simulation setup: $\Delta = 2$, $p_1, p_2 \sim \mathcal{U}([1, \Delta])$, $\mu_1 = 0$, $\mu_2 = 1$, $\alpha = 0.01$. σ varies from 0.2 to 0.6. Blue full line plots the numerical error exponent estimated from slope of $\log P^{\alpha}_{\text{miss},t}$ vs. σ by linear fitting. **Top**: linear fitting performed on the whole interval $[1, t_{\max}]$; **bottom**: linear fitting performed on $[4/5\, t_{\max}, t_{\max}]$. Red dashed line plots the theoretical bound calculated by solving (46)).

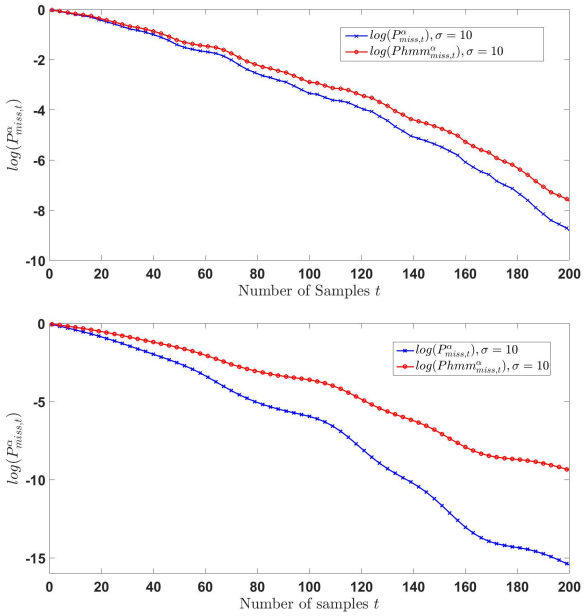

Fig. 8: Simulation setup: (upper) $\Delta = 5$, $\sigma = 10$, $\mu_1 = 2$, $\mu_2 = 5$, $\alpha = 0.01$. **Top**: $p_{1,\text{g}} = p_{2,\text{g}} = 1/(1-q^{\Delta})((1-q),\, q(1-q),\, q^2(1-q),\, q^3(1-q),\, q^4(1-q))$, where $q = 0.8$; **bottom**: $p_{1,\text{c}} = p_{2,\text{c}} = (0.025,\, 0.025,\, 0.025,\, 0.025,\, 0.9)$. The curves plot the probability of a miss $\log P^{\alpha}_{\text{miss},t}$ (in the log scale) vs. number of samples t. Blue full lines with "x" markers plot the probability of a miss for the HSMM-based detector, and red full lines with "o" markers plot the probability of a miss for the HMM-based detector.

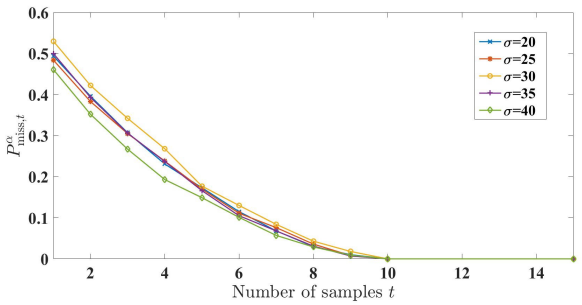

Fig. 9: Simulation setup: $\Delta = 10$, $p_1, p_2 \sim \mathcal{U}([1, \Delta])$, $\mu_1 = 66$, $\mu_2 = 2200$, $\sigma = 90$, $\alpha = 0.01$. Plots of probability of a miss for 5 different σ values.

The results shown in Figure 8 (top and bottom) corroborate the preceding intuition. Indeed, in the first case, the two error probability curves are very close to each other, but the HSMM detector still has an edge over the HMM-based one, which is of course expected, as HSMM-based detector is the optimal one. On the other hand, in the second case, when the state duration distribution is concentrated on the highest duration $\Delta = 5$, the advantage of applying the optimal HSMM-based detector is evident: for the value of error probability of $e^{-5} = 0.0067$, the HSMM takes about 80 samples, while the HMM takes more than 120 samples, hence requiring 50% more resources in terms of measurements. Note that the HMM model is not very adequate for the concentrated pmf $p_{1,\text{c}} = p_{2,\text{c}}$, while the HSMM model developed here is capable of accommodating this type of distributions.

NILM simulation. In the final set of simulations, we demonstrate applicability of the results to estimate the number of samples needed to detect an appliance run from the smart meter data. To do that, we use measurements of a dishwasher from the REFIT dataset [6]. REFIT dataset contains 2 years of appliance measurements from 20 houses. The monitored dishwasher is a two-state appliance, with mean power values of $\mu_1 = 2200W$, $\mu_2 = 66W$ and standard deviation of $\sigma_1 = 36.6W$ and $\sigma_2 = 18.2W$, in states 1 and 2, respectively. The mean value of background noise which is also base-load in that house is $\mu_0 = 90$ and with standard deviation $\sigma_0 = 16.6W$. We downsampled dishwasher data with $\Delta = 10$ to simulate the influence of noise, including base-load and unknown appliances on detecting the appliance. The simulation results are shown in Figure 9 as plots of $P^{\alpha}_{\text{miss},t}$ vs. t for several values of σ between the measured σ_1 and σ_2.

As expected, the probability of a miss decreases with the increase of number of samples t. Furthermore, the number of samples needed for successful detection is about 10.

VIII. Conclusion

We studied the problem of detecting a multi-state signal hidden in noise, where the durations of state occurrences vary over time in a nondeterministic manner. We modelled such a process via a random duration model that, for each state, assigns a (possibly distinct) probability mass function to the duration of each occurrence of that state. Assuming Gaussian noise and a process with two possible states, we derived optimal likelihood ratio test and showed that it has a form of a linear recursion of dimension equal to the sum of the duration spreads of the two states. Using this result, we showed that the Neyman-Pearson error exponent is equal to the top Lyapunov exponent for the linear recursion, the exact computation of which is a well-known hard problem. Using the theory of large deviations, we provided a lower bound on the error exponent. We demonstrated the tightness of the bound with numerical results. Finally, we illustrated the developed methodology in the context of NILM, applying it on the problem of detecting multi-state appliances from the aggregate load signal.

Acknowledgement

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 734331.

Appendix

Proof of Lemma 9. To prove the claim, we apply Theorem 2 from [42]. Note that since matrices A_k are i.i.d., they are stationary and ergodic, and hence they are also metrically transitive, see, e.g., [44]. Therefore the assumptions of the theorem are fulfilled. We now show that the condition of the theorem holds, i.e., we show that

$$\mathbb{E}_0 \left[\log^+ \|A_k\|\right] < +\infty, \tag{59}$$

where $\log^+ = \max\{\log, 0\}$. It is easy to verify that $\|A_k\| \leq e^{\max_{m=1,2}|f_m(X_k)|} C_{M_0}$, where $C_{M_0} = \|M_0\|$. Thus, we have

$$\begin{aligned}\log^+ \|A_k\| &\leq \log^+ C_{M_0} e^{\max_{m=1,2}|f_m(X_k)|} \\ &\leq \log^+ C_{M_0} + \max_{m=1,2} |f_m(X_k)| \\ &\leq \log^+ C_{M_0} + |f_1(X_k)| + |f_2(X_k)|\,. \end{aligned} \tag{60}$$

Since X_k is Gaussian, and f_1 and f_2 are linear functions, we have that $f_1(X_k)$ and $f_2(X_k)$ are Gaussian. Therefore, the expectation of the right hand side of the preceding equation is finite (which can be seen by bounding $\mathbb{E}_0\left[|f_1(X_k)|\right] \leq \sqrt{\mathbb{E}_0\left[f_1^2(X_k)\right]} \leq +\infty$, and similarly for $m = 1$). Hence, the condition (59) follows. By Theorem 2 from [42] we therefore have that

$$\lim_{t\to+\infty} \frac{1}{t} \log \|\Pi_t\| = \lim_{t\to+\infty} \frac{1}{t} \mathbb{E}\left[\log \|\Pi_t\|\right], \tag{61}$$

which proves (36). To prove (37), we note that $L_t = {p^+}^\top \Pi_t \mathbb{1}_{2\Delta}$, where $p^+ > 0$. Thus, there exist constants c and C such that $c\|\Pi_t\| \leq L_t \leq C\|\Pi_t\|$ [45]. The claim now follows from the preceding sandwich relation between L_t and $\|\Pi_t\|$.

Proof of Lemma 8. Fix $t \geq 1$ and consider L_t as expressed in (27). Applying Jensen's inequality, and taking the logarithm, we get:

$$\log L_t \geq \sum_{s^t \in \mathcal{S}^t} P(s^t) \left(\frac{1}{\sigma^2} \sum_{m=1}^{2} \mu_m \sum_{k \in \mathcal{T}_m(s^t)} X_k - \tau_m(s^t) \frac{\mu_m^2}{2\sigma^2} \right). \tag{62}$$

Recall now Lemma 9, eq. (61). Taking the expectation w.r.t. $\mathcal{H}_0$ in (62), and expressing $\tau_m(s^t) = \sum_{k=1}^t 1_{\{S_k=m\}}$, we obtain

$$\mathbb{E}_0\left[\log L_t\right] \geq -\sum_{s^t\in\mathcal{S}^t} P(s^t)\left(\tau_1(s^t)\frac{\mu_1^2}{2\sigma^2}+\tau_2(s^t)\frac{\mu_2^2}{2\sigma^2}\right) = -\frac{\sum_{s^t\in\mathcal{S}^t}P(s^t)\mu_{s^t}^2}{2\sigma^2} = -\frac{\mathbb{E}_1\left[\mu_{S^t}^2\right]}{2\sigma^2}, \quad (63)$$

where $\mu_{s^t}^2 = \sum_{k=1}^t 1_{\{s_k=1\}}\mu_1^2 + \sum_{k=1}^t 1_{\{s_k=2\}}\mu_2^2$, and similarly for a random sequence S^t. Dividing both sides of (63) by t, inverting the sign, and taking the limit we get

$$\zeta \leq \liminf_{t\to+\infty}\frac{1}{2\sigma^2}\frac{\mathbb{E}_1\left[\mu_{S^t}^2\right]}{t}, \quad (64)$$

i.e., the error exponent is upper bounded by the expected, per sample SNR (we will show shortly that the above limit in fact exists). The right hand side in the above equation can be alternatively expressed as:

$$\mathbb{E}_1\left[\mu_{S^t}^2\right] = \sum_{m=1}^2 \mathbb{E}_1\left[\sum_{k=1}^t 1_{\{S_k=m\}}\right]\mu_m^2 = \sum_{m=1}^2\sum_{k=1}^t \mathbb{P}_1\left(S_k=m\right)\mu_m^2. \quad (65)$$

For an arbitrary time k, we now express the probability that $S_k = m$ via the vector $\mathbf{q}$, defined in Subsection IV-A:

$$\mathbb{P}_1\left(S_k=1\right) \dot{=} \sum_{d=1}^{\Delta}\mathbb{P}_1\left(S_k=\ldots=S_{k-d+1}=1, S_{k-d}=2\right)$$
$$= \sum_{d=1}^{\Delta} p_{1d}^+ \mathbb{P}_1\left(S_{k-d+1}=1, S_{k-d}=2\right) = \left(p_1^+\right)^\top \mathbf{q}_{k,1}. \quad (66)$$

Summing up over $k = 1,...,t$ and using the transition formula (33), we get:

$$\frac{1}{t}\sum_{k=1}^t \mathbb{P}_1\left(S_k=m\right) = \left[\left(p_1^+\right)^\top, 0_\Delta^\top\right]\frac{\sum_{k=1}^t P^k}{t}\mathbf{q}_0. \quad (67)$$

Using now proposition 7, it is easy to show that $P^k \to \frac{1}{q^\top p_1+q^\top p_2}[\mathbf{1}_\Delta^\top, 1_\Delta^\top]^\top \left(p^+\right)^\top$, see Theorem 8.5.1 in [45]. As the Cesaró averages must converge to the same matrix, we have

$$\lim_{t\to}\frac{1}{t}\sum_{k=1}^t \mathbb{P}_1\left(S_k=m\right) = \frac{q^\top p_1}{q^\top p_1+q^\top p_2}, \quad (68)$$

where the identity follows by the fact that only the first element of $\mathbf{q}_0$ is equal to one(the remaining ones being zero), and also the fact that $p_{11}^+ = 1$. Similar identity can be derived for the limit of $\frac{1}{t}\sum_{k=1}^t \mathbb{P}_1\left(S_k=m\right)$. Replacing the right hand-side of (68) for $m=1$ and $m=2$ in (64) we get the claim of the Lemma.

Proof of Lemma 11.

Proof. Fix $\nu\in\mathcal{V}$. To remove the dependence on ξ in (44), for any given fixed $\nu\in\mathcal{V}$, we need to solve

$$\begin{array}{ll}\text{minimize} & \theta_2\frac{\left(\frac{\xi}{\theta_2}-\mu\right)^2}{2\sigma^2} \\ \text{subject to} & H(\nu)\geq\frac{\xi^2}{2\theta_2\sigma^2} \\ & \xi\in\mathbb{R}\end{array}, \quad (69)$$

where, as before, we denote $\theta_2 = q^\top\nu_2$. Since $\mu>0$, and the constraint set is defined only through the square of ξ, the optimal solution of (69) is achieved for $\xi\geq 0$. Thus, (69) is equivalent to

$$\begin{array}{ll}\text{minimize} & \theta_2\frac{\left(\frac{\xi}{\theta_2}-\mu\right)^2}{2\sigma^2} \\ \text{subject to} & 0\leq\xi\leq\sigma\sqrt{2\theta_2 H(\nu)}\end{array}. \quad (70)$$

The solution of (70) is given by: 1) $\xi^\star = \theta_2\mu$, if $\theta_2\mu \leq \sigma\sqrt{2\theta_2 H(\nu)}$; and 2) $\xi^\star = \sigma\sqrt{2\theta_2 H(\nu)}$, otherwise. Hence, to solve (44) we can partition its constraint set $\mathcal{V} = \mathcal{V}_1\bigcup\mathcal{V}_2$ according to these two cases, where $\mathcal{V}_1 = \left\{\nu\in\mathcal{V}:\ H(\nu)\geq\theta_2\frac{\mu^2}{2\sigma^2}\right\}$ and $\mathcal{V}_2 = \left\{\nu\in\mathcal{V}:\ H(\nu)\leq\theta_2\frac{\mu^2}{2\sigma^2}\right\}$, solve the corresponding two optimization problems, and finally find the minimum among the two obtained optimal values.

Consider first the case $\nu\in\mathcal{V}_1$. Since in this case $\xi^\star=\theta_2\mu$, plugging in this value in (70), we have that the optimization problem (44) with $\mathcal{V}$ reduced to $\mathcal{V}_1$ simplifies to:

$$\begin{array}{ll}\text{minimize} & D(\nu_1||p_1)+D(\nu_2||p_2) \\ \text{subject to} & \nu\in\mathcal{V}_1.\end{array}. \quad (71)$$

If $H(p)\geq\frac{q^\top p_2}{q^\top p_1+q^\top p_2}\frac{\mu^2}{2\sigma^2}$, then the point $1/\left(q^\top p_1+q^\top p_2\right)p$ belongs to $\mathcal{V}$, where $p=(p_1,p_2)$ and hence the optimal solution to (71) equals $1/\left(q^\top p_1+q^\top p_2\right)p$ with the corresponding optimal value equal to 0. Suppose now that $H(p)<\frac{q^\top p_2}{q^\top p_1+q^\top p_2}\frac{\mu^2}{2\sigma^2}$. We show that in this case the solution to (71) must be at the boundary of the constraint set, in the set of points $\left\{\nu\in\mathcal{V}:\ H(\nu)=\theta_2\frac{\mu^2}{2\sigma^2}\right\}$.

We prove the above claim. Since the entropy function H, see eq. (22), is concave, the constraint set $\mathcal{V}_1$ is convex, and since KL divergence D is convex, we conclude that the problem in (71) is convex. Also, it can be shown that the Slater point exists [46]. Therefore, the solution to (71) is given by the corresponding Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{cases}(1+\lambda)\log\frac{\nu_{1d}}{\mathbb{1}^\top\nu_1}-\log p_{1d}=0,\ \text{for } d=1,...,\Delta \\ (1+\lambda)\log\frac{\nu_{2d}}{\mathbb{1}^\top\nu_2}-\log p_{2d}+\lambda d\frac{\mu^2}{2\sigma^2}=0,\ \text{for } d=1,...,\Delta \\ H(\nu)\geq q^\top\nu_2\frac{\mu^2}{2\sigma^2} \\ \lambda\geq 0 \\ \lambda\left(H(\nu)-q^\top\nu_2\frac{\mu^2}{2\sigma^2}\right)=0 \\ \nu\in\mathcal{V}\end{cases}. \quad (72)$$

From the fourth and fifth condition, we have that either $\lambda=0$, or that $\lambda>0$ and $H(\nu)=q^\top\nu_2\frac{\mu^2}{2\sigma^2}$. Suppose that $\lambda=0$. Then, from the first two KKT conditions we have that the solution ν must satisfy $\nu_{md}/\mathbb{1}^\top\nu_m = p_{md}$, for $m=1,2$, $d=1,...,\Delta$. However, this contradicts with the third condition (recall that we assumed that $H(p)<q^\top p_2\frac{\mu^2}{2}$). Therefore, the solution to (71) must belong to the set $\left\{\nu\in\mathcal{V}:\ H(\nu)=q^\top p_2/\left(q^\top p_1+q^\top p_2\right)\frac{\mu^2}{2\sigma^2}\right\}$. Since this set intersects with the set $\mathcal{V}_2$, we conclude that, when $H(p)<q^\top p_2/\left(q^\top p_1+q^\top p_2\right)\frac{\mu^2}{2\sigma^2}$, then the optimal solution to (44) is found by optimizing over the smaller set $\mathcal{V}_2\subseteq\mathcal{V}$, i.e., (44) is equivalent to

$$\begin{aligned} \text{minimize} \quad & D(\nu_1||p_1) + D(\nu_2||p_2) + \tfrac{\theta_2}{2\sigma^2}\left(\tfrac{\xi^\star}{\theta_2} - \mu\right)^2 \\ & \nu \in \mathcal{V}_2. \end{aligned}, \tag{73}$$

where $\xi^\star(\nu) = \sigma\sqrt{2\theta_2 H(\nu)}$. Simple algebraic manipulations reveal that the third term in the objective above is equal to $\left(\sqrt{H(\nu)} - \sqrt{R(\nu)}\right)^2$. Finally, set $\mathcal{V}_2$ is precisely the constraint set in (44), and hence the claim of the lemma follows. □

REFERENCES

[1] S.-Z. Yu, "Hidden semi-markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215 – 243, 2010, special Review Issue.

[2] V. S. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis*, ser. Lecture notes in statistics. Springer, 2008.

[3] S. M. Ross, *Stochastic Processes*. Wiley, 2nd Ed, February 1995.

[4] S. Chiappa, "Explicit-duration Markov switching models," *Foundations and Trends in Machine Learning*, vol. 7, no. 6, pp. 803–886, Dec. 2014.

[5] G. W. Hart, *Nonintrusive Appliance Load Data Acquisition Method: Progress Report*.

[6] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," *Scientific data*, vol. 4, p. 160122, 2017.

[7] J. Liao, G. Elafoudi, L. Stankovic, and V. Stankovic, "Non-intrusive appliance load monitoring using low-resolution smart meter data," in *2014 IEEE Int. Conf. on Smart Grid Comm.*, Nov 2014, pp. 535–540.

[8] J. Kelly, "Disaggregation on domestic smart meter energy data," Ph.D. dissertation, Imperial College London, London, 2017.

[9] O. Parson, S. Ghosh, M. J. Weal, and A. Rogers, "Non-intrusive load monitoring using prior models of general appliance types." in *26th Conf. Artificial Intelligence. AAAI. Toronto, ON, Canada*, 2012, pp. 356–362.

[10] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial HMMs with application to energy disaggregation," in *Int. Conf. Artificial Intelligence and Statistics. AISTATS. La Palma, Canary Islands*, 2012, pp. 1472–1482.

[11] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han, "Unsupervised disaggregation of low frequency power measurements," in *11th SIAM Int. Conf. on Data Mining*, 2011.

[12] M. Zhong, N. Goddard, and C. Sutton, "Signal aggregate constraints in additive factorial hmms, with application to energy disaggregation," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3590–3598.

[13] H. Altrabalsi, V. Stankovic, J. Liao, and L. Stankovic, "Low-complexity energy disaggregation using appliance load modelling," *AIMS Energy*, vol. 4, no. 1, pp. 1–21, 2016.

[14] H. Lange and M. Berges, "The neural energy decoder: Energy disaggregation by combining binary subcomponents," in *NILM Workshop, Vancouver, Canada*, May 2012, pp. 1472–1482.

[15] K. He, L. Stankovic, J. Liao, and V. Stankovic, "Non-intrusive load disaggregation using graph signal processing," *IEEE Trans. on Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[16] B. Zhao, L. Stankovic, and V. Stankovic, "On a training-less solution for non-intrusive appliance load monitoring using graph signal processing," *IEEE Access*, vol. 4, pp. 1784–1799, 2016.

[17] D. Egarter, M. Pöchacker, and W. Elmenreich, "Complexity of power draws for load disaggregation," Jan. 2015, arXiv:1501.02954.

[18] C. Uggen, "Work as a turning point in the life course of criminals: A duration model of age, employment, and recidivism," *American Sociological Review*, vol. 65, no. 4, pp. 529–546, Aug. 2000.

[19] J. R. Russell and R. F. Engle, "A discrete-state continuous-time model of financial transactions prices and times: The autoregressive conditional multinomial-autoregressive conditional duration model," *Journal of Business and Economic Statistics*, vol. 23, no. 2, pp. 166–180, April 2005.

[20] F. Chaubert-Pereira, Y. Guédon, C. Lavergne, and C. Trottier, "Markov and semi Markov switching linear mixed models used to identify forest tree growth components," *Biometrics, Wiley*, vol. 66, no. 3, pp. 753 – 762, 2010.

[21] M. Ting, A. O. Hero, D. Rugar, C.-Y. Yip, and J. A. Fessler, "Near-optimal signal detection for finite-state Markov signals with application to magnetic resonance force microscopy," *IEEE Trans. on Sig. Proc.*, vol. 54, no. 6, pp. 2049–2062, June 2006.

[22] M. Ting and A. O. Hero, "Detection of a random walk signal in the regime of low signal to noise ratio and long observation time," in *2006 IEEE Int. Conf. on Acoustics Speech and Signal Processing*, vol. 3, May 2006.

[23] A. Agaskar and Y. M. Lu, "Optimal detection of random walks on graphs: Performance analysis via statistical physics," April 2015, http://arxiv.org/abs/1504.06924.

[24] D. Bajović, J. M. F. Moura, and D. Vukobratovic, "Detecting random walks on graphs with heterogeneous sensors," July 2017, http://arxiv.org/abs/1707.06900.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: John Wiley and Sons, 2006.

[26] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, MA: Jones and Barlett, 1993.

[27] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[28] R. Bahadur, *Some Limit Theorems in Statistics*. Society for Industrial and Applied Mathematics, 1971.

[29] P.-N. Chen, "General formulas for the Neyman-Pearson type-II error exponent subject to fixed and exponential type-I error bounds," *IEEE Trans. on Inf. Theory*, vol. 42, no. 1, pp. 316–323, Jan 1996.

[30] L. H. Koopmans, "Asymptotic rate of discrimination for Markov processes," *The Annals of Mathematical Statistics*, vol. 31, no. 10, pp. 982–994, 1960.

[31] L. B. Boza, "Asymptotically optimal tests for finite Markov chains," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1992–2007, 1971.

[32] K. Vašek, "On the error exponent for ergodic Markov source," *Kybernetika*, vol. 16, no. 4, pp. 318–329, 1980.

[33] L. D. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. on Inf. Theory*, vol. 27, no. 4, pp. 431–438, 1981.

[34] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains," *IEEE Trans. on Inf. Theory*, vol. 31, no. 3, pp. 360–365, May 1985.

[35] H. Luschgy, A. L. Rukhkin, and I. Vajda, "Adaptive tests for stochastic processes in the ergodic case," *Stochastic Processes and their Applications*, vol. 25, pp. 47–59, 1990.

[36] H. Luschgy, "Asymptotic behavior of neyman-pearson tests for autoregressive processes," *Scandinavian Journal of Statistics*, vol. 21, no. 4, pp. 461–473, 1994.

[37] Y. Sung, L. Tong, and H. V. Poor, "Neyman-Pearson detection of Gauss-Markov signals in noise: closed-form error exponent and properties," *IEEE Trans. on Inf. Theory*, vol. 52, no. 4, pp. 1354–1365, April 2006.

[38] J. N. Tsitsiklis and V. D. Blondel, "The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate," *Mathematics of Control, Signals and Systems*, vol. 10, no. 1, pp. 31–40, March 1997.

[39] R. Altman, "Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting," *Journal of the American Statistical Association*, vol. 102, no. 15, pp. 201 – 210, 2007.

[40] D. Bajović, K. He, L. Stanković, D. Vukobratović, and V. Stanković, "Optimal detection and error exponents for hidden multi-state processes via random duration model approach," December 2017, arXiv preprint.

[41] I. Flores, "Direct calculation of k-generalized Fibonacci numbers," *Fibonacci Quarterly*, vol. 5, no. 3, pp. 259–266, 1967.

[42] H. Furstenberg and H. Kesten, "Products of random matrices," *Ann. Math. Statist.*, vol. 31, no. 2, pp. 457–469, 06 1960.

[43] O. Zeitouni, "Gaussian Fields," March 2016, Lecture notes. Courant institute, New York.

[44] C. Shalizi, "Stochastic processes," 2007, Lecture notes. Stochastic processes. Carnegie Mellon University.

[45] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, United Kingdom: Cambridge University Press, 1990.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, United Kingdom: Cambridge University Press, 2004.