An Information Theoretic Approach to the Mode Estimation of Randomly Switching FIR Systems

Nuno C. Martins * and Munther A. Dahleh
Laboratory for Information and Decision Systems
Department of EECS, Massachusetts Institute of Technology, Cambridge MA 02139, USA
* Phone:+1 617 2532832, Email: nmartins@mit.edu

***Abstract*— In this paper we tackle the problem of estimating the mode of switching systems. From the theoretical point of view, our contribution is twofold: creating a framework that has a clear parallel with a communication paradigm and deriving an analysis of performance. In particular, our work is restricted to the class of systems that randomly switch among a finite alphabet of discrete-time finite impulse response linear operators, therein designated as modes. In our approach, the switching system is viewed as an encoder of the mode, which is interpreted as the message, while an external excitation process establishes a random code. Accordingly, the estimator, which observes the code and uses noisy measurements of the output, is constructed as a decoder whose properties can be studied by means of a modification of Shannon's theory. Using a distance function, we define an uncertainty ball where the estimates are guaranteed to lie with probability arbitrarily close to 1. The radius of the uncertainty ball is directly related to the rate of the switching process. It is shown that lower rates lead to smaller uncertainty. Such distance also reflects the informativity of the external excitation (code) and as such can be used as a guide on its choice.**

I. INTRODUCTION

The implications of modal estimation of hybrid systems span applications [8] [15] in Adaptive Control and fault detection. Dating back to more than 25 years [16], the investigation of such problems has generated a vast portfolio of algorithms and methods [13]. In this paper, we focus on the modal estimation of systems that randomly switch among a set of finite impulse response, or moving average, filters. A discrete stochastic process drives the switching, while the system is excited by a white Gaussian process.

By using finite impulse response filters, we expect to make our analysis and methods applicable to other classes of uniformly (over the switching sequences) stable [20] or stochastically stable switching systems [21]. As for the mode estimation, we recognize that there is a parallel with a communication setup that uses randomly generated codes [9]. Under that framework, the input defines a constrained code and the system is perceived as an encoder of the mode (message). Consequently, we adopt a decoder structure for the mode estimator. That allows the use of Shanon's theory in the analytical formulation of performance measures. In [10], it is mentioned the difficulty of computing measures of performance for the available suboptimal algorithms. In practice, such quality evaluation may have to resort to Monte Carlo simulations. We address that problem by defining a measure of distance $\mathcal{D}$ and computing the probability that a sequence of mode estimates $\bar{\hat{\mathbf{q}}}_{1,n}$ (generated by the decoder) is in a ball, of radius β, around the real one $\bar{\mathbf{q}}_{1,n}$. Explicit bounds are derived that depend on n, the size of the memory of the decoder, $r^q \in [0,1]$, the rate of the switching process, and the covariance matrices of the noise and the input. An interesting feature of this framework is that, for fixed $\beta > r^q + \frac{1}{2\ln m}$, the probability of $\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \bar{\hat{\mathbf{q}}}_{1,n}) < \beta$ converges to 1 as n tends to infinity. In that sense, as a theoretical result, our work relates to [19], where it is proven that the uncertainty [18] in identifying time-varying systems is directly related to the *speed* of variation. In our case, such quantity is given by the rate $r^q + \frac{1}{2\ln m}$. Information theoretically inspired distances, have been widely used in system identification and parameter estimation [3], [4], [5]. Further examples are the channel identification with finite observations, for a finite model set, studied in [7], and the identification of finite alphabet information sources[6]. The optimal solution to the mode estimation problem can be cast as a Bayesian Hypothesis testing. The search space of such approach grows, in general [17], with m^n, where m is the number of modes and n is the number of observations. That stimulated a quest for methods of merging and/or pruning hypothesis as a way to reduce computational load [10], [14]. The computation complexity of the estimator proposed in this paper grows with m^{nr^q}, where n (number of observations) is selected as a function of the performance specifications. Such dependence of the computational complexity on r^q is a result of restricting, the hypothesis testing, to a typical set [12]. If the switching rate r^q is low or n can be taken small, then such estimator is suitable for computer implementation. If that is not the case, we claim that, since the estimation scheme presented in this paper can be made arbitrarily close to the optimal, its performance analysis can be useful as a benchmark measure of other suboptimal methods. Also, the distance $\mathcal{D}$ can be used as a quality measure on the design of probing signals and can be viewed as a first step to the proper study of the effect of observed inputs on the mode observability of linear hybrid systems.

The paper is organized as follows: Section I-A introduces the notation used throughout the text. The problem is stated and its information theoretic equivalence established in section II, while section II-A provides a guide through the main results. The method is described in section III, while the performance analysis is carried out in section IV. Numerical examples are provided in [1].

A. Notation

The following notation is adopted: Large caps letters are used to indicate vectors and matrices. Small caps letters are reserved for real scalars and discrete variables. In addition, p is reserved to represent probability distributions. Discrete-time sequences are indexed by time using integer subscripts, such as x_k, $k \geq 1$. Finite segments of discrete-time sequences are indicated with a *bar* on the top, e.g. $\bar{x}$. Whenever it is relevant, the range of their time-indexing must be indicated as in the following example:

$$\bar{q}_{k,n} = (q_k, \dots, q_n) \tag{1}$$

Superscripts are reserved for distinguishing different variables and functions according to their meaning. Random variables are represented using boldface letters and follow the conventions above. As an illustration, $\mathbf{x}$ is a valid representation for a scalar random variable, while a sample (or realization) is written as x. Also, a finite segment of a discrete-time sequence of random variables, would be $\bar{\mathbf{q}}_{1,k}$. A realization of such process would be indicated as $\bar{q}_{1,k}$. The probability of an *event* is indicated by $\mathcal{P}(event)$. We use the entropy function, of a random variable $\mathbf{Z}$, given by:

$$\mathcal{H}[\mathbf{Z}] = \mathcal{E}[-\log_m p^Z(\mathbf{Z})] \tag{2}$$

where $m \in \mathbb{N}$, p^Z is the p.d.f. of $\mathbf{Z}$ and $\mathcal{E}[\cdot]$ is the expected value, taken over $\mathbf{Z}$. Similarly, the conditional entropy is given by $\mathcal{H}[\mathbf{Z}^1|\mathbf{Z}^2] = \mathcal{H}[\mathbf{Z}^1, \mathbf{Z}^2] - \mathcal{H}[\mathbf{Z}^2]$ where, in this case, the expectation is taken with respect to $\mathbf{Z}^1$ and $\mathbf{Z}^2$. The covariance matrix of a random variable $\mathbf{Z}$, with $Z \in \mathbb{R}^n$, is given by:

$$\Sigma_Z = \mathcal{E}\left[(\mathbf{Z} - \mathcal{E}[\mathbf{Z}])(\mathbf{Z} - \mathcal{E}[\mathbf{Z}])^T\right] \tag{3}$$

II. Problem Statement

Consider the mode alphabet $\mathbb{A} = \{1, \dots, m\}$ and the random process $\mathbf{F}_k$, with $F_k \in \mathbb{R}^{n^F}$, described by:

$$\mathbf{F}_k = \mathbf{Y}_k + \mathbf{W}_k, \ k \geq \alpha \tag{4}$$

$$\mathbf{Y}_k = \sum_{i=0}^{\alpha} G_i(\bar{\mathbf{q}}_{k-\alpha,k})\mathbf{V}_{k-i} \tag{5}$$

where $k, \alpha \in \mathbb{N}$ and $G_i : \mathbb{A}^{\alpha+1} \to \mathbb{R}^{n^F \times n^V}$ are matrices that specify the switching system. The stochastic processes $\mathbf{V}_k$ (test signal), $\mathbf{W}_k$ and $\mathbf{q}_k$ are mutually independent and satisfy:

- $\mathbf{V}_k$ and $\mathbf{W}_k$ are Gaussian zero mean i.i.d. processes, with $V_k \in \mathbb{R}^{n^V}$ and $W_k \in \mathbb{R}^{n^F}$. In order to avoid degeneracy problems, we assume $\Sigma_W > 0$. In contrast to $\mathbf{W}_k$, the test signal $\mathbf{V}_k$ is assumed to be observed by the estimator. Examples where this is a realistic assumption are: when $\mathbf{V}_k$ is generated by the estimator; when it is an exogenous process that can be observed by the estimator or a combination of both.
- $\mathbf{q}_k$ is a discrete stationary stochastic process with alphabet $\mathbb{A} = \{1, 2, \dots, m\}$. Although our results are valid for Markovian $\mathbf{q}_k$ (see subsection IV-A), we reduce the technical complexity of this paper by considering the i.i.d. case. While keeping the same structure, such simplification makes proofs more clear. For a given $n \in \mathbb{N}$, we write the p.d.f. of $\bar{\mathbf{q}}_{1,n}$ as $p^q(\bar{q}_{1,n})$, regardless of n. The rate [12] of $\mathbf{q}_k$, designated by r^q, is computed as:

$$r^q = \mathcal{E}[-log_m p^q(\mathbf{q})] \tag{6}$$

For a given n, we wish to use a test signal $\mathbf{V}_k$ and a decision system that, by means of the measurement of $\bar{\mathbf{F}}_{1,n}$, produces an estimate $\hat{\bar{\mathbf{q}}}_{1,n}$ of $\bar{\mathbf{q}}_{1,n}$. The estimation method must have an associated measure of distance $\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \hat{\bar{\mathbf{q}}}_{1,n})$ that allows the specification of a ball around $\bar{\mathbf{q}}_{1,n}$ where the estimates $\hat{\bar{\mathbf{q}}}_{1,n}$ will lie with a given probability. In particular, for a given $\beta > 0$, we are interested in computing $\mathcal{P}(\mathcal{D}(\bar{q}_{1,n}, \hat{\bar{q}}_{1,n}) > \beta)$. Such probability is expected to depend on β and n. By using an information theoretic formulation, this characterization must also reflect the informativity of $\mathbf{V}_k$ and r^q, the rate of the switching process.

We would like to stress that, without loss of generality, we develop our analysis using the origin of time as $k = 1$. In real applications, this setup can be used in a **sliding window** scheme (see figure 1) and n *should be viewed as the memory of the estimator*. This way, once we determine all the parameters of the estimation process, for every $k > n$ we use measurements $\bar{\mathbf{F}}_{k-n+1,k}$ to produce $\hat{\bar{\mathbf{q}}}_{k-n+1,k}$.

A. Main results and Implementation Issues

Among the results presented in this paper, the following are central to answering the questions posed above: In definition 4.1 a measure of distance $\mathcal{D}$ is introduced. It is through this function that the informativity of $\mathbf{V}_k$ can be gauged. If $\mathbf{V}_k$, or only part of it, are generated by the estimator then that generated portion of Σ_V may be tuned to achieve a desired distance function. Similar methods for the distance-based design of probing functions were derived in [2]. In lemma 5.8 we show that, as n increases, $\mathcal{P}(\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \hat{\bar{\mathbf{q}}}_{1,n}) > \beta)$ decreases. Theorem 4.1 proves that if $\beta > r^q + \frac{1}{2 \ln m}$ then $\mathcal{P}(\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \hat{\bar{\mathbf{q}}}_{1,n}) > \beta)$ can be made arbitrarily small by increasing n. The explicit bounds given in lemma 5.8 answer the question: "given $\beta > 0$ and $\delta > 0$, what is n that guarantee that $\mathcal{P}(\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \hat{\bar{\mathbf{q}}}_{1,n}) > \beta) < \delta$ ". In [1] we address these issues in the more general setting where observations are generated through:

$$\mathbf{F}_k = \mathbf{Y}_k + \tilde{\mathbf{Y}}_k + \mathbf{W}_k(\mathbf{q}_k), \ k \geq \alpha \tag{7}$$

where, for any given $q \in \mathbb{A}$, $\mathbf{W}_k(q)$ is a zero-mean white Gaussian process, with covariance matrix depending on q, independent of $\mathbf{V}_k$ and $\mathbf{q}_k$. Moreover, in [1], we assume that $\tilde{\mathbf{Y}}_k$ is the output observation of another autonomous moving average switching process whose switching is governed by $\mathbf{q}_k$.

B. Posing the Problem Statement as a Coding Paradigm

As the generality of the estimation paradigms [3], the estimation of the mode of (4)-(5) can be interpreted as a problem of communication through a noisy channel (see Figure 1). The message to be transmitted is q_k, the test signal $\mathbf{V}_k$ specifies the code while the measurement noise $\mathbf{W}_k$ completes the setup of such channel. The decoder is bound to use $\bar{\mathbf{V}}_{1,n}$ and the noisy measurements $\bar{\mathbf{F}}_{1,n}$ to make a decision as to which is the best estimate $\bar{\hat{q}}_{1,n}$.

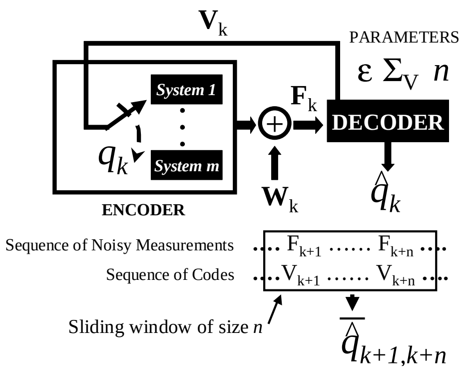

Fig. 1. Depiction of the communications setup interpretation for mode estimation.

III. Channel interpretation and decoding

A specific feature of this communications setup is that the *encoder* does not know the message to be transmitted. This is a problem if one wants to adopt the approach of coding long words as a way to reduce the probability of error (channel coding theorem). In order to circumvent this difficulty, we follow the procedure in the proof of Shannon's channel coding theorem, i.e., the use of random coding [9]. Consequently, we consider the process $\mathbf{V}_k$ as establishing a constrained random code specified by Σ_V.

In the decoding process we will use the following estimate:

$$\hat{\mathbf{Y}}(\bar{\hat{q}}_{k^1-\alpha,k^2}) = \begin{bmatrix} \sum_{i=0}^{\alpha} G_i(\bar{\hat{q}}_{k^2-\alpha,k^2})\mathbf{V}_{k^2-i} \\ \vdots \\ \sum_{i=0}^{\alpha} G_i(\bar{\hat{q}}_{k^1-\alpha,k^1})\mathbf{V}_{k^1-i} \end{bmatrix} \tag{8}$$

The decoding process has a hypothesis testing structure. The likelihood of a given candidate sequence $\bar{\hat{q}}_{1,n}$ is gauged by means of the estimation error $\bar{\mathbf{F}}_{1+\alpha,n} - \hat{\mathbf{Y}}(\bar{\hat{q}}_{1,n})$. The following definition, describing a *conditional* test random variable, will facilitate the definition and the performance analysis of the decoder.

Definition 3.1: Given $k^1, k^2 \in \mathbb{N}$, $k^2 \geq k^1 > \alpha$ and $\bar{q}_{k^1-\alpha,k^2}, \bar{\hat{q}}_{k^1-\alpha,k^2} \in \mathbb{A}^{k^2-k^1+1+\alpha}$, define the random variable $\mathbf{T}(\bar{q}_{k^1-\alpha,k^2}, \bar{\hat{q}}_{k^1-\alpha,k^2})$ as:

$$\mathbf{T}(\bar{q}_{k^1-\alpha,k^2}, \bar{\hat{q}}_{k^1-\alpha,k^2}) = \begin{bmatrix} \sum_{i=0}^{\alpha} \left(G_i(\bar{q}_{k^2-\alpha,k^2}) - G_i(\bar{\hat{q}}_{k^2-\alpha,k^2})\right)\mathbf{V}_{k^2-i} + \mathbf{W}_{k^2} \\ \vdots \\ \sum_{i=0}^{\alpha} \left(G_i(\bar{q}_{k^1-\alpha,k^1}) - G_i(\bar{\hat{q}}_{k^1-\alpha,k^1})\right)\mathbf{V}_{k^1-i} + \mathbf{W}_{k^1} \end{bmatrix} \tag{9}$$

Notice that, in the previous definition, for any given indices k^1, k^2 and $\bar{q}_{k^1,k^2}, \bar{\hat{q}}_{k^1,k^2} \in \mathbb{A}^{k^2-k^1+1}$, we have:

$$T(\bar{q}_{k^1,k^2}, \bar{\hat{q}}_{k^1,k^2}) \in \mathbb{R}^{n^F(k^1-k^2+1-\alpha)} \tag{10}$$

If $\bar{\hat{q}}_{k^1,k^2} = \bar{q}_{k^1,k^2}$, then we write $\mathbf{T}(\bar{q}_{k^1,k^2}, \bar{\hat{q}}_{k^1,k^2})$ in abbreviate form as $\mathbf{T}(\bar{\hat{q}}_{k^1,k^2})$.

A. Description of the estimator (decoder)

By following the approach that leads to a standard decoder [12], in this section we construct a mode estimator.

In the subsequent analysis we use the following result:

Remark 3.1: If $\mathbf{X}$ is a Gaussian, zero mean random variable with covariance matrix $\Sigma_X \in \mathbb{R}^{n^X \times n^X}$, then:

$$\mathcal{H}(\mathbf{X}) = \frac{1}{2}\log_m\left((2\pi e)^{n^X}|\Sigma_X|\right) \tag{11}$$

The following is the definition of the estimation error rate r^{Av}. Such quantity is important in the construction of the estimator.

Definition 3.2: Define the estimation error rate r^{Av} as:

$$r^{Av} = \frac{\mathcal{H}(\mathbf{W}_k)}{n^F} = \frac{\log_m\left((2\pi e)^{n^F}|\Sigma_W|\right)}{2n^F} \tag{12}$$

The following definitions complete the list of mathematical objects needed to describe the decoder.

Definition 3.3: Consider $\bar{\hat{q}}_{1,n} \in \mathbb{A}^n$ and realizations $\bar{F}_{1,n}$ and $\hat{Y}(\bar{\hat{q}}_{1,n})$. Given parameters $n \in \mathbb{N}$ and $\epsilon \in (0,1)$, define the selection indicator $s^{\epsilon,n} : \mathbb{A}^n \to \{True, False\}$ as:

$$s^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True, \text{ if}$$
$$\left|\frac{\log_m p^{T(\bar{\hat{q}}_{1,n})}(\bar{F}_{1+\alpha,n} - \hat{Y}(\bar{\hat{q}}_{1,n}))}{n^F(n-\alpha)} + r^{Av}\right| < \epsilon \tag{13}$$

and $s^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = False$ otherwise.

The decoding process is carried out by recognizing which sequences, in the typical set defined bellow, can be invalidated.

Definition 3.4: Given the parameters $n \in \mathbb{N}$ and $\epsilon \in (0,1)$, the set of typical sequences $\mathbb{T}^{\epsilon,n}$ is defined as:

$$\mathbb{T}^{\epsilon,n} = \{\bar{\hat{q}}_{1,n} \in \mathbb{A}^n : \left|\frac{\log_m p^q(\bar{\hat{q}}_{1,n})}{n} + r^q\right| < \epsilon\} \tag{14}$$

It is the cardinality of $\mathbb{T}^{\epsilon,n}$ that determines the computational complexity of the estimator. According to [12] such quantity is bounded by $m^{n(r^q+\epsilon)}$. The structure of the decoder is the following:

Definition 3.5: (**Decoder**) Given the realizations $\bar{F}_{1,n}$ and $\bar{V}_{1,n}$, parameters $n \in \mathbb{N}$ and $\epsilon \in (0,1)$, define the decoder as a search in $\mathbb{T}^{\epsilon,n}$ that generates $\bar{\hat{q}}_{1,n}$ satisfying:

$$\bar{\hat{q}}_{1,n} \in \mathbb{T}^{\epsilon,n} \text{ and } s^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True \tag{15}$$

If, for a given realization, there is no such $\bar{\hat{q}}_{1,n}$, then the decoder generates an arbitrary $\bar{\hat{q}}_{1,n} \in \mathbb{T}^{\epsilon,n}$. The decoding process defines a random variable, which we designate by $\bar{\hat{\mathbf{q}}}_{1,n}$.

IV. Performance Analysis: Characterization of the Uncertainty Sets by means of a Measure of Distance

The quality evaluation, of the coding/decoding process, is carried out by computing the probability that $\bar{\hat{\mathbf{q}}}_{1,n}$ is in a ball around $\bar{\mathbf{q}}_{1,n}$. Such *uncertainty* set is specified by means of the distance $\mathcal{D}$ defined bellow. This function is related to the concept of divergence as in [2].

Definition 4.1: (**Measure of Distance**)The distance $\mathcal{D} : \mathbb{A}^n \times \mathbb{A}^n \to \mathbb{R}$ is given by:

$$\mathcal{D}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) = \frac{\mathcal{H}(\mathbf{T}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}))}{n^F(n-\alpha)} - r^{Av} \tag{16}$$

The following theorem is the main result of this paper.

Theorem 4.1: (**Main result**) Let $\bar{\hat{\mathbf{q}}}_{1,n}$ be determined according to the decoding process described in the definition 3.5. For any given $\delta > 0$, there exists $n \in \mathbb{N}$ and $\epsilon \in (0,1)$ such that:

$$\mathcal{P}(\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \bar{\hat{\mathbf{q}}}_{1,n})) > \frac{r^q}{n^F} + \frac{1}{2\ln m} + \delta) < \delta, \text{ if } r^q \geq 0 \tag{17}$$

$$\mathcal{P}(\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \bar{\hat{\mathbf{q}}}_{1,n})) > \delta) < \delta, \text{ if } r^q = 0 \tag{18}$$

Proof:
The result for $r^q \geq 0$ follows directly from lemma 5.8 in section V, while (18) is proven in [1].

□

According to [1], if, for every $\bar{q}_{1,n}, \bar{\hat{q}}_{1,n} \in \mathbb{A}^n$ and $k > 2\alpha$, the following approximation holds:

$$\begin{aligned}\mathcal{H}\left[\mathbf{T}(\bar{q}_{k-\alpha,k}, \bar{\hat{q}}_{k-\alpha,k}) | \mathbf{T}(\bar{q}_{1,k-1}, \bar{\hat{q}}_{1,k-1})\right] \simeq \\ \mathcal{H}\left[\mathbf{T}(\bar{q}_{k-\alpha,k}, \bar{\hat{q}}_{k-\alpha,k}) | \mathbf{T}(\bar{q}_{k-2\alpha,k-1}, \bar{\hat{q}}_{k-2\alpha,k-1})\right]\end{aligned} \tag{19}$$

then, for large n, we can approximate the distance $\mathcal{D}$ as:

$$\mathcal{D}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) \simeq \sum_{i=1+2\alpha}^{n} \frac{\mathcal{R}(\bar{q}_{i-2\alpha,i}, \bar{\hat{q}}_{i-2\alpha,i}) - r^{Av}}{n-2\alpha} \tag{20}$$

where $\mathcal{R} : \mathbb{A}^{2\alpha+1} \times \mathbb{A}^{2\alpha+1} \to \mathbb{R}$, is given by [1]:

$$\mathcal{R}(\bar{q}^1, \bar{q}^2) = \frac{1}{2n^F} \log_m \left((2\pi e)^{n^F} \frac{\left|\Sigma_{\mathbf{T}(\bar{q}_{1,2\alpha+1}, \bar{\hat{q}}_{1,2\alpha+1})}\right|}{\left|\Sigma_{\mathbf{T}(\bar{q}_{1,2\alpha}, \bar{\hat{q}}_{1,2\alpha})}\right|} \right) \tag{21}$$

with $\bar{q}_{1,2\alpha+1} = \bar{q}^1$ and $\bar{\hat{q}}_{1,2\alpha+1} = \bar{q}^2$. The determination of $\Sigma_{\mathbf{T}(\cdot,\cdot)}$ is a standard procedure, which is further simplified by the fact that $\mathbf{V}_k$ and $\mathbf{W}_k$ are independent. Also, the resulting matrices are affine on the elements of Σ_V. It is through this decomposition that the informativity of the input can be accessed. For any given subsequences $\bar{q}^1, \bar{q}^2 \in \mathbb{A}^{2\alpha+1}$, the *higher* $\mathcal{R}(\bar{q}^1, \bar{q}^2)$ is the more efficiently the decoder can distinguishing between $\bar{q}^1$ and $\bar{q}^2$.

A. Extension to Markovian Switching

The main result of this paper (Theorem 4.1) holds if $\mathbf{q}_k$ is a Markovian process with rate $r^q > 0$. The fact that theorem 4.1 remains valid can be obtained by continuity arguments [1]. In that situation, $\frac{p^q(q_k, q_{k-1})}{p^q(q_{k-1})}$ defines the transition probabilities of the Markov process. The rate of $\mathbf{q}_k$, designated by r^q, is computed as $r^q = \mathcal{E}[-log_m \frac{p^q(\mathbf{q}_k, \mathbf{q}_{k-1})}{p^q(\mathbf{q}_{k-1})}]$.

V. Auxiliary results leading to the Proof of Theorem 4.1

The main result of this section is lemma 5.8, where explicit expressions are given to bound the probability (17).

In order guarantee notational simplicity, we start with the following definitions:

Definition 5.1: For any given $\bar{q}_{1,n}, \bar{\hat{q}}_{1,n} \in \mathbb{A}^n$ we define the following random variable:

$$\mathbf{z}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) = -\frac{\log_m p^{T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})}(\mathbf{T}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}))}{n^F(n-\alpha)} \tag{22}$$

Lemma 5.1: Given $n > \alpha$ and $\bar{q}_{1,n}, \bar{\hat{q}}_{1,n} \in \mathbb{A}^n$

$$\mathcal{E}[\mathbf{z}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})] = \mathcal{D}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) + r^{Av} \tag{23}$$

Proof:

The proof follows the from:

$$\mathcal{H}\left(\mathbf{T}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})\right) = \mathcal{E}\left[-\log_m p^{T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})}\left(\mathbf{T}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})\right)\right] \tag{24}$$

and the formula for $\mathcal{D}$ given in definition 4.1 .

□

Lemma 5.2: For any given positive definite $\Sigma_X \in \mathbb{R}^{n^X}$, designate by $\mathbf{X}$ the zero-mean Gaussian random variable with covariance matrix Σ_X. Then, the following holds:

$$\sup_{\Sigma_X > 0} \left(\mathcal{E}[\left(\frac{\log_m p^X(\mathbf{X})}{n^X}\right)^2] - \left(\mathcal{E}[\frac{\log_m p^X(\mathbf{X})}{n^X}]\right)^2 \right) < \frac{a}{n^X} \tag{25}$$

where $p^X(X) = \frac{e^{-\frac{1}{2}X^T \Sigma_X^{-1} X}}{\left((2\pi)^{n^X} |\Sigma_X|\right)^{1/2}}$ and $a = \frac{1}{2(\ln m)^2}$.

[1] If $\alpha = 0$, then take $\left|\Sigma_{\mathbf{T}(\bar{q}_{1,0}, \bar{\hat{q}}_{1,0})}\right| = 1$.

Proof: The result is obtained by evaluating the expected values in (25) and applying the change of variables $\tilde{X} = \Sigma_X^{-1/2} X$.

$\square$

Lemma 5.3: Given $\epsilon > 0$, $n > \alpha$ and sequences $\bar{q}_{1,n}, \bar{\hat{q}}_{1,n} \in \mathbb{A}^n$, the following holds:

$$\mathcal{P}\left(\left|\mathbf{z}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) - \mathcal{E}\left[\mathbf{z}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})\right]\right| > \epsilon\right) < \frac{\frac{1}{2(\ln m)^2}}{\epsilon^2 n^F (n-\alpha)} \tag{26}$$

Proof:

Since $\mathbf{W}_k$ is i.i.d. and $\Sigma_W > 0$, we conclude that $\Sigma_{\mathbf{T}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})} > 0$, so that lemma 5.2 and the fact that $T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) \in \mathbb{R}^{n^F(n-\alpha)}$ lead to:

$$Var(\mathbf{z}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})) \leq \frac{\frac{1}{2(\ln m)^2}}{n^F(n-\alpha)} \tag{27}$$

The proof is concluded by a direct application of the Chebyshev-Bienaymé inequality [11].

$\square$

Lemma 5.4: Given $\epsilon > 0$ and $n > \alpha$ the following holds:

$$\mathcal{P}\left(\left|\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) - r^{Av})\right| > \epsilon\right) \quad < \quad \frac{\frac{1}{2(\ln m)^2}}{\epsilon^2 n^F (n-\alpha)} \tag{28}$$

Proof: We start by noticing that:

$$\begin{aligned}\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) - r^{Av} &= \\ = \mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) - \mathcal{E}\left[\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n})\right] &+ \mathcal{D}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n})\end{aligned} \tag{29}$$

and as such, since $\mathcal{D}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) = 0$:

$$\begin{aligned}&\mathcal{P}\left(\left|\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) - r^{Av}\right| > \epsilon\right) = \\ &\quad = \mathcal{P}\left(\left|\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) - \mathcal{E}\left[\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n})\right]\right| > \epsilon\right)\end{aligned} \tag{30}$$

The final result follows by direct substitution of the expressions of lemma 5.3.

$\square$

Definition 5.2: (**Event** Γ - The decoder's search space is empty) Given $\epsilon \in (0,1)$ and $n \in \mathbb{N}$, define Γ as the event that:

$$\{\bar{\hat{q}}_{1,n} \in \mathbb{T}^{\epsilon,n} : s^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True\} = \emptyset \tag{31}$$

Lemma 5.5: For any given $\epsilon \in (0,1)$, if $n > \alpha$, then the following holds:

$$P(\Gamma) < \frac{\varrho^2}{n\epsilon^2} + \frac{\frac{1}{2(\ln m)^2}}{\epsilon^2 n^F (n-\alpha)} \tag{32}$$

where

$$\varrho = \max_{q \in \mathbb{A}, p^q(q) \neq 0} \left|log_m p^q(q) - r^q\right| \tag{33}$$

Proof:

Notice that, from definition 3.5, if a given realization $\bar{q}_{1,n}$ satisfies $\bar{q}_{1,n} \in \mathbb{T}^{\epsilon,n}$ and $s^{\epsilon,n}(\bar{q}_{1,n}) = True$, then Γ is false because $\bar{q}_{1,n}$ is itself a valid choice for $\bar{\hat{q}}_{1,n}$. That leads to the following inequality:

$$\mathcal{P}(\Gamma) \leq \mathcal{P}(|\mathbf{z}(\bar{\mathbf{q}}_{1,n}, \bar{\mathbf{q}}_{1,n}) - r^{Av}| > \epsilon) + \mathcal{P}(\bar{\mathbf{q}}_{1,n} \notin \mathbb{T}^{\epsilon,n}) \tag{34}$$

The first term in the RHS of (34) can be bounded by means of lemma 5.4. The second term can be analyzed through the expansion:

$$\log_m p^q(\bar{q}_{1,n}) = \sum_{i=1}^{n} log_m p^q(q_i) \tag{35}$$

Standard results on the variance of bounded random variables, lead to:

$$Var\left(\log_m p^q(\bar{\mathbf{q}}_{1,n})\right) \leq n\varrho^2 \tag{36}$$

Using the Chebyshev-Bienaymé inequality [11], we get:

$$\mathcal{P}(\bar{\mathbf{q}}_{1,n} \notin \mathbb{T}^{\epsilon,n}) \leq \frac{n\varrho^2}{n^2\epsilon^2} \tag{37}$$

$\square$

Lemma 5.6: For any given $\bar{q}_{1,n}, \bar{\hat{q}}_{1,n} \in \mathbb{A}^n$ and $n > \alpha$ the following holds:

$$\begin{aligned}&\max_{X \in \mathbb{R}^{n^F(n-\alpha)}} p^{T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})}(X) = \\ &\qquad = m^{n^F(n-\alpha)(-\mathcal{D}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) - r^{Av} + \frac{1}{2\ln m})}\end{aligned} \tag{38}$$

Proof:

From the definition of the Gaussian distribution:

$$\max_{X \in \mathbb{R}^{n^F(n-\alpha)}} p^{T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})}(X) \;=\; \frac{\left|\Sigma_{T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})}\right|^{-1/2}}{(2\pi)^{\frac{n^F(n-\alpha)}{2}}} \tag{39}$$

or equivalently, using (11), it can also be written as:

$$\begin{aligned}&\max_{X \in \mathbb{R}^{n^F(n-\alpha)}} p^{T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})}(X) = \\ &\qquad = m^{-\mathcal{H}\left(T(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})\right) + \frac{n^F(n-\alpha)}{2\ln m}}\end{aligned} \tag{40}$$

The final result is achieved once we recall that $\mathcal{D}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n}) = \frac{\mathcal{H}\left(\mathbf{T}(\bar{q}_{1,n}, \bar{\hat{q}}_{1,n})\right)}{n^F(n-\alpha)} - r^{Av}$.

$\square$

Lemma 5.7: Let $n > \alpha$, $\epsilon \in (0,1)$, $\bar{q}_{1,n} \in \mathcal{A}^n$ and $\bar{\hat{q}}_{1,n} \in \mathbb{T}^{\epsilon,n}$. The following is an upper-bound for the conditional probability [2] that $\bar{\hat{q}}_{1,n}$ satisfies the decoding condition $s^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True$.

[2] Assuming that $\bar{q}_{1,n}$ is a realization of the switching process, this probability can be described in words as the probability that $\bar{\hat{q}}_{1,n}$ is not rejected by the decoder's search process.

$$P(\mathbf{s}^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True|\bar{q}_{1,n}) < \\ < m^{n^F(n-\alpha)(-\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n})+\frac{1}{2\ln m}+\epsilon)} \quad (41)$$

Proof: The structure of the decoder leads to :

$$P(\mathbf{s}^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True|\bar{q}_{1,n}) = \int_{\mathbb{S}} p^{T(\bar{q}_{1,n},\bar{\hat{q}}_{1,n})}(X)dX \quad (42)$$

where $\mathbb{S}$, the set of realizations leading to (13), is given by:

$$\mathbb{S} = \{X \in \mathbb{R}^{n^F(n-\alpha)} : \\ \left| \frac{\log_m p^{T(\bar{q}_{1,n},\bar{\hat{q}}_{1,n})}(X)}{n^F(n-\alpha)} + r^{Av} \right| < \epsilon\} \quad (43)$$

We can use lemma 5.6 to infer that:

$$P(\mathbf{s}^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True|\bar{q}_{1,n}) \leq \\ \int_{\mathbb{S}} m^{n^F(n-\alpha)\left(-\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n})-r^{Av}+\frac{1}{2\ln m}\right)} dX \leq \\ \leq Vol(\mathbb{S}) m^{n^F(n-\alpha)\left(-\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n})-r^{Av}+\frac{1}{2\ln m}\right)} \quad (44)$$

The proof is completed once we notice that the volume of $\mathbb{S}$ is upper-bounded [12] by:

$$Vol(\mathbb{S}) < m^{n^F(n-\alpha)(r^{Av}+\epsilon)} \quad (45)$$

□

Lemma 5.8: (*Main lemma*) Given $\beta \in \mathbb{R}_+$ and $\epsilon \in (0,1)$, the following holds:

$$\mathcal{P}(\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n}) > \beta) < \mathcal{P}(\Gamma) + \\ + m^{n^F(n-\alpha)\left(\frac{r^q}{n^F}-\beta+\frac{1}{2\ln m}+\frac{1+n^F}{n^F}\epsilon\right)+\alpha(r^q+\epsilon)} \quad (46)$$

where $\mathcal{P}(\Gamma)$ is given in lemma 5.5.

Proof:

We separate the two events that potentially generate $\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n}) > \beta$, and write the following bound:

$$\mathcal{P}(\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n}) > \beta) \leq \mathcal{P}(\Gamma) + \\ + \sum_{\bar{q}_{1,n}\in\mathbb{A}^n} \sum_{\bar{\hat{q}}_{1,n}\in\mathbb{D}(\bar{q}_{1,n})} P(\mathbf{s}^{\epsilon,n}(\bar{\hat{q}}_{1,n}) = True|\bar{q}_{1,n}) p^q(\bar{q}_{1,n}) \quad (47)$$

where

$$\mathbb{D}(\bar{q}_{1,n}) = \{\bar{\hat{q}}_{1,n} \in \mathbb{T}^{\epsilon,n} : \mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n}) > \beta\} \quad (48)$$

Using lemma 5.7 and the inequality above, we get:

$$\mathcal{P}(\mathcal{D}(\bar{q}_{1,n},\bar{\hat{q}}_{1,n}) > \beta) \leq \\ (\sharp\mathbb{T}^{\epsilon,n})\, m^{n^F(n-\alpha)(-\beta+\epsilon+\frac{1}{2\ln m})} + \mathcal{P}(\Gamma) \quad (49)$$

The fact [12] that $\sharp\mathbb{T}^{\epsilon,n} \leq m^{n(r^q+\epsilon)}$ concludes the proof.

□

Acknowledgment

This work was sponsored by the University of California - Los Angeles, MURI project title: *"Cooperative Control of Distributed Autonomous Vehicles in Adversarial Envrionments"*, award: 0205-G-CB222. The first author was also partially supported by the Portuguese Foundation for Science and Technology, award: PRAXIS XXI, BD 19630/99.

References

[1] Martins, N. C. and Dahleh, M. A., *"An Information Theoretic Approach to the Mode Estimation of Randomly Switching Linear Systems"* MIT-LIDS Internal Publication, LIDS-P 2560, 2003

[2] Arimoto, S. and Kimura H., *"Optimum input test signals for system identification - an information-theoretical approach"* Int. J. Systems Sci., No 3, Vol I, pp. 279-290, 1971

[3] Weidemann, H.L., *"Entropy Analysis of Parameter Estimation"* Information and Control, 14, pp. 493-506, 1969

[4] Kailath, T., *"The Divergence and Bhattachryya Distance Measures in Signal Selection"* IEEE Transaction on Communication Technology, Vol. COM-15, No. 1, pp. 52-60, February 1967

[5] Upadhyaya, B. R., Sorenson, H. W., *"Bayesian discriminant approach to input signal selection in parameter estimation problems"* Information Sciences, 12(1), pp. 61-91, 1977

[6] Kieffer, J. *"Strongly consistent code-based identification and order estimation for constrained finite-state model classes"* IEEE Transaction on Information Theory, Vol. 39, No. 3, pp. 893-902, May 1993

[7] Al Jabri, A. Kh., *"Channel Identification: An Information Theoretic Viewpoint"* ISIT, pp. 329, August 1998

[8] Bar-Shalom, Y. and Li, X. R. *"Estimation and tracking: Principle, Techniques and Software"*. Boston, MA:Artech House, 1993

[9] Battail, G. *"On random-like codes."* Information Theory and Applications II, Lecture notes in Computer Science 1133, pp. 76-94, Springer 1995

[10] Blom, H. A. P. and Bar-Shalom, Y. *"The interacting multiple model algorithm for systems with Markovian switching coefficients."* IEEE Transactions on Automatic Control, Vol. 33, No. 8, pp. 780-783, 1988

[11] Billingsley, P. *"Probability and measure."* Wiley Inter-5Science, 3rd Ed., 1995

[12] Cover, T. and Thomas, J.A. *Elements of Information theory* Wiley series in Telecommunications, 1991

[13] Sworder, D.D.; Boyd J.E. *Estimation Problems in Hybrid Systems* Cambridge University Press, 1999

[14] Sworder, D. D.; Boyd J. E.; Elliott, R. J. *"Modal Estimation in Hybrid Systems"* Journal of Mathematical Analysis and Applications 245, 225-247 (2000)

[15] Mariton, M. *Jump Linear Systems in Automatic Control* Marcel Dekker, 1990

[16] Ackerson, G. A.; Fu, S. K. *"On State Estimation in Switching Environments"* IEEE Transactions on Automatic Control, vol 15, no 1, pp. 10-17, 1970

[17] Elliott, R.; Dufour F.; Sworder D. D. *"Exact Hybrid Filters in Discrete Time"* IEEE Transaction on Automatic Control, vol 41, no 12, pp. 1807-1810, 1996

[18] Venkatesh, S. R.; Dahleh, M. A. *"On System Identification of Complex Systems from Finite Data"* IEEE Transaction on Automatic Control, vol 46, no. 2, 235-257, 2001

[19] Zames, G.; Lin, L.; Wang L. Y. *"Fast Identification n-Widths and Uncertainty Principles for LTI and Slowly Varying Systems"* IEEE Transaction on Automatic Control, vol 39, no 9, 1827-1838, 1994

[20] Liberzon, D.; Morse A. S. *Basic Problems in Stability and Design of Switched Systems* IEEE Control Systems Magazine, pp. 1-19, Oct 1999

[21] Chizeck, H. J.; Willsky, A. S.; Castanon, D. *Discrete-Time Markovian-Jump quadratic optimal Control* International Journal of Control, vol 43, no 1, pp. 213-231, 1986