Rate Distortion in the Modal Estimation of Switching FIR Linear Systems

Nuno C. Martins and Munther A. Dahleh

Abstract— This paper is about the mode estimation of randomly switching linear systems. We follow up on previous results, where we posed the problem in a way which is well fitted for a worst-case analysis. The aforementioned investigation relies on a divergence measure and several other quantities, such as the entropy rate of the mode switching as well as the power ratio between the system's excitation and the measurement noise. In this article we ascribe a natural interpretation to such divergence measure, while establishing a connection to rate-distortion theory. We also explore why and in what form estimation fidelity is affected by the entropy rate of the mode switching. Additionally, this analysis suggests that such entropy rate must be regarded as a fundamental quantity in the determination of the estimation fidelity.

I. INTRODUCTION

Motivated by a vast portfolio of applications, the challenge of estimating and tracking the mode, of switching linear systems, has received consistent attention over the last years. Such problem is important for several reasons, which include failure detection and adaptive control. Additionally, in adversarial environments, the mode sequence may represent the maneuvers executed by an external opponent. This information may be a vital intelligence for determining what will be the agent's next step or it's intent. Several publications have explored the most diverse applications, while stressing the importance of the general problem[3].

Depending on the formulation, modal estimation may be realized as part of a hybrid state estimation paradigm[8], comprising discrete (modes) and continuous states. Although the investigation of such problems has generated a variety of algorithms and methods, [9] suggests the need to devote more attention to the modal estimation alone. By adopting the framework used in the identification of FIR-LPV systems [13], [4], the estimation of the continuous state can be avoided. Along these lines, in [2], as well as in this paper, we focus on the modal estimation problem by considering systems that switch among a set of modes which comprise finite impulse response, or moving average, filters. A discrete stochastic process $\mathbf{q}_k$ drives the switching, while the system is excited by a white Gaussian process (probing signal).

In [2] we recognize that the modal estimation problem is equivalent to a communication setup in the presence of randomly generated codes [6]. Under that framework, the input probing signal defines a constrained code and the system is perceived as an encoder of the mode sequence $\mathbf{q}_k$. Motivated by that, we adopt a decoder structure for the mode estimator. In the same publication, we also present an alternative low complexity estimation algorithm that converges in probability, as k tends to infinity, to the decoder proposed.

Although using a different formulation, we refer to [7] where it is exposed the difficulty of computing measures of quality for the available modal estimation algorithms. In practice, such quality evaluation may have to resort to Monte Carlo simulations. The reason behind such difficulties is the fact that some of these estimators rely on the propagation of probability densities. Moreover, the propagation itself is very intricate and its computational implementation requires additional approximations and simplifications. In [2], we address that problem by means of a measure of divergence $\mathcal{D}_d$ which allows a probabilistic worst-case analysis. Still, [2] uses $\mathcal{D}_d$ as a distance to characterize a set that comprises all the mode estimation uncertainty. In particular, it is shown that the estimates of the mode sequences are, with probability arbitrarily close to 1, in a ball around the true sequence. The radius of the ball is an affine function of the entropy rate of the switching process. In that sense, as a theoretical result, the work in [2] relates to [11], where it is proven that the uncertainty [10] in identifying time-varying systems is directly related to the *speed (or rate)* of variation. In this paper we devote attention to the meaning of $\mathcal{D}_d$ as a measure of estimation fidelity. We also expound why the entropy rate of $\mathbf{q}_k$ can be regarded as a fundamental quantity.

The distance $\mathcal{D}_d$ can also 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. We also stress that our analysis and methods are applicable to other classes of stable switching systems [12]. In these cases, the moving average coefficients should be the truncation of the impulse response of such stable linear systems. Worst-case LPV-FIR system identification was also broached by [14] in the case where the coefficients have a specific functional form.

The paper is organized as follows: Section I-A introduces the notation used throughout the text. The technical framework is formulated, the main results of [2] are reviewed and the contributions of this paper are stated in section II, while section III presents rate-distortion theory in a way which is relevant in the present context. The detailed study, as well the implications in estimation fidelity, of the divergence measure introduced in [2] is performed in IV.

Nuno C. Martins and Munther A. Dahleh are at LIDS, Dept. of EECS, Massachusetts Institute of Technology nmartins@mit.edu, dahleh@mit.edu

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. Finite segments of discrete-time sequences are indicated by the range of their time-indexing as in the following example ($k \geq n$):

$$q_{n,k} = (q_n, \ldots, q_k) \tag{1}$$

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 $\mathbf{q}_{1,k}$. A realization of such process would be indicated as $q_{1,k}$. The probability of an *event* is indicated by $\mathcal{P}(event)$. We use the entropy function, of a random variable $\mathbf{Z}$, given by:

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

where $p^{\mathbf{Z}}$ 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 matrix $\mathbf{Z}$, with $Z \in \mathbb{R}^{n^1 \times n^2}$, is given by:

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

If Σ_Z is a scalar we write, instead, σ_z^2. The i-th eigenvalue of a matrix $B \in \mathbb{R}^{n \times n}$ is represented as $\lambda_i(B)$.

II. TECHNICAL FRAMEWORK

Consider the mode alphabet $\mathbb{A} = \{1, \ldots, m\}$, with $m \geq 2$, and the random process $\mathbf{F}_k$, with $F_k \in \mathbb{R}^{n^F}$, described by:

$$\mathbf{F}_k = \mathbf{Y}_k + \mathbf{W}_k, \ k \geq 1 \tag{4}$$

$$\mathbf{Y}_k = \sum_{i=0}^{\alpha} G_i(\mathbf{q}_k)\mathbf{V}_{k-i}, \ k \geq 1 \tag{5}$$

where $k, \alpha \in \mathbb{N}$ and $G_i : \mathbb{A} \to \mathbb{R}^{n^F \times n^V}$ are matrices that specify the switching system. The stochastic processes $\mathbf{V}_k$ (probing 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 $\Sigma_W > 0$. Additionally, aiming at simplifying the presentation, we adopt $\Sigma_W = \sigma_w^2 I$. In contrast to $\mathbf{W}_k$, the probing 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 and ergodic Markovian stochastic process with alphabet $\mathbb{A}$. For a given $k \in \mathbb{N}$, we write the probability mass function of $\mathbf{q}_{1,k}$ as $p^{\mathbf{q}}(q_{1,k})$ and the conditional probability is expressed as $p^{\mathbf{q}}(q_k|q_{k-1}) = \frac{p^{\mathbf{q}}(q_k, q_{k-1})}{p^{\mathbf{q}}(q_{k-1})}$. The entropy rate of q_k is given by:

$$r_m^q = \mathcal{E}[-\log_m p^{\mathbf{q}}(\mathbf{q}_k|\mathbf{q}_{k-1})] \tag{6}$$

where we use a non-standard scaling. In [18], the entropy rate is defined, in *nats*, as $\mathcal{E}[-\ln p^{\mathbf{q}}(\mathbf{q}_k|\mathbf{q}_{k-1})]$.

A. Review of previous results

For a given k, we use the probing signal $\mathbf{V}_{1-\alpha,k}$ and a decision system that, by means of the measurement of $\mathbf{F}_{1,k}$, produces an estimate $\hat{\mathbf{q}}_{1,k}$ of $\mathbf{q}_{1,k}$. Given a realization $q_{1,k}$, the following is a statistic for gauging the likelihood of a candidate sequence $\hat{q}_{1,k}$:

Definition 2.1: Given $k \in \mathbb{N}$, $k \geq 1$ and $q_k, \hat{q}_k \in \mathbb{A}$, define the random variable $\mathbf{T}_k(q_k, \hat{q}_k)$ as:

$$\mathbf{T}_k(q_k, \hat{q}_k) = \sum_{i=0}^{\alpha} (G_i(q_k) - G_i(\hat{q}_k))\mathbf{V}_{k-i} + \mathbf{W}_k \tag{7}$$

For any given indices k^1, k^2 and $q_{k^1, k^2}, \hat{q}_{k^1, k^2} \in \mathbb{A}^{k^2 - k^1 + 1}$, we also adopt the following abuse of notation:

$$\mathbf{T}_{k^1, k^2}(q_{k^1, k^2}, \hat{q}_{k^1, k^2}) = (\mathbf{T}_{k^1}(q_{k^1}, \hat{q}_{k^1}), \ldots, \mathbf{T}_{k^2}(q_{k^2}, \hat{q}_{k^2})) \tag{8}$$

In [2], we have studied an estimator as well as the following associated measure [1] of distance $\mathcal{D}_d(q_{1,k}, \hat{q}_{1,k})$, which is related to the concept of divergence as in [5].

Definition 2.2: (**Measure of Distance**) The distance $\mathcal{D}_d : \bigcup_{k=1}^{\infty} \mathbb{A}^k \times \mathbb{A}^k \to \mathbb{R}$ is given by:

$$\mathcal{D}_d(q_{1,k}, \hat{q}_{1,k}) = \frac{\mathcal{H}(\mathbf{T}_{1,k}(q_{1,k}, \hat{q}_{1,k}))}{k} - \mathcal{H}(\mathbf{W}_k) \tag{9}$$

Such distance is well fitted for a worst case performance analysis, as shown in the following theorem:

Theorem 2.1: (**Main theorem of [2]**) For any given $\delta > 0$, there exists $k \in \mathbb{N}$ and an estimator which produces $\hat{\mathbf{q}}_{1,k}$, satisfying:

$$\mathcal{P}\left(\mathcal{D}_d(\mathbf{q}_{1,k}, \hat{\mathbf{q}}_{1,k}) > \ln(m)r^q + \frac{n^F}{2} + n^F\delta\right) < \delta \tag{10}$$

$$\mathcal{P}(\mathcal{D}_d(\mathbf{q}_{1,k}, \hat{\mathbf{q}}_{1,k}) > n^F\delta) < \delta, \text{ if } r_m^q = 0 \tag{11}$$

Theorem 2.1 shows that there exists an estimator for which the estimates are contained in a ball, centered around the true sequence, with probability arbitrarily close to 1. The *radius* of the ball can be made as close to $\ln(m)r^q + \frac{n^F}{2}$ as desired. We also emphasize that in [2] it is assumed that the estimator knows the probability transition matrix of $\mathbf{q}_k$. Such assumption can be dropped, by means of

---

[1] In [2], the distance is defined with an extra factor $\frac{1}{n^F}$. We have such scaling in mind when using the results of [2]
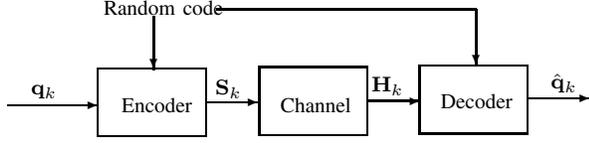
Fig. 1. Framework for a communication system

using an estimator constructed as a hypothesis testing over a polynomially growing set of sequences. An identical argument is used in [1] to prove the universal coding theorem.

The quantity $\ln(m)r^q$ is the entropy rate of $\mathbf{q}_k$, in nats, and, according to theorem 2.1, it modulates an upper-bound for the worst-case uncertainty ball; smaller $\ln(m)r^q$ leading to the assurance of less uncertainty. One might think of this property as the stochastic and countable alphabet analog to the characterization given by [11]. In common, they have the fact that the *rate* of time-variation impacts the magnitude of the uncertainty, in a robust identification framework.

B. Contributions of the paper

In this paper we devote attention to the meaning of $\mathcal{D}_d$ as a measure of estimation fidelity. We also expound why the entropy rate of $\mathbf{q}_k$ can be regarded as a fundamental quantity. We appeal to *rate distortion* theory [18] as a way to bring together these aspects and the information theoretical viewpoint, presented in [2]. Moreover, we investigate if the worst-case radius of theorem 2.1, given by $\ln(m)r^q + \frac{n^F}{2}$, is conservative.

III. RATE DISTORTION THEORY IN THE PRESENT CONTEXT

Consider the setup depicted in Figure 1. That is the basic communications scheme, where the channel is a memoryless map which is completely described by $p^{\mathbf{H}|\mathbf{S}}$, the probability of $\mathbf{H}_k$ conditioned on $\mathbf{S}_k$.

From standard texts in Information Theory, such as [16], we know that any map $\mathcal{D}: \bigcup_{k=1}^{\infty} \mathbb{A}^k \times \mathbb{A}^k \to \mathbb{R}_+$ qualifies as a distortion measure. Like the cost functional in an optimization procedure, a meaningful distortion measure should be used to quantify the desired properties of a communication scheme. In rate distortion theory, it is customary to look at $d^* = \mathcal{E}[\mathcal{D}(\mathbf{q}_{1,k}, \hat{\mathbf{q}}_{1,k})]$. The smaller $d^\star$ the better the quality of the communication scheme.

An example of a distortion measure is the following, based on the Hamming distance:

$$\mathcal{D}_h(q_{1,k}, \hat{q}_{1,k}) = \frac{1}{k}\sum_{i=1}^{k}\mathcal{I}(q_i \neq \hat{q}_i) \qquad (12)$$

where $\mathcal{I}$ is the indicator function. Accordingly, $\mathcal{D}_h$ is the relative frequency of errors over the horizon $\{1, \ldots, k\}$.

A. Capacity of the channel

We denote the capacity of the channel [18], in the m-ary basis, by C_m. Such quantity represents the maximum transfer rate at which information can be reliably transmitted through the channel. The analysis we develop in this article can be generalized, but we adopt the following additive white Gaussian channel:

$$\mathbf{H}_k = \mathbf{S}_k + \mathbf{W}_k \qquad (13)$$

Given $\Sigma_Y > 0$, if the eigenvalues of Σ_S are restricted by $\lambda_i(\Sigma_S) \leq \lambda_i(\Sigma_Y)$ then C_m satisfies [18]:

$$C_m = \sum_{i=1}^{n^F}\frac{1}{2}\log_m\left(1 + \frac{\lambda_i(\Sigma_Y)}{\sigma_w^2}\right) \qquad (14)$$

B. Rate Distortion trade-off

The fact that C_m establishes a fundamental limit, for reliable transmission, is explicit in the following relation [16], [17], which is true for any encoder, decoder and code [2]:

$$\frac{\Gamma(d_h^*(k,\Sigma_V))}{\ln(m)} + d_h^*(k,\Sigma_V)\frac{\ln(m-1)}{\ln(m)} \geq r_m^q - C_m \quad (15)$$

where $\Gamma : [0,1] \to \mathbb{R}_+$ is given by:

$$\Gamma(d_h^*(k,\Sigma_V)) = -d_h^*(k,\Sigma_V)\ln(d_h^*(k,\Sigma_V))$$
$$- (1 - d_h^*(k,\Sigma_V))\ln(1 - d_h^*(k,\Sigma_V)) \quad (16)$$

and $d_h^*(k,\Sigma_V)$ represents the following expected distortion:

$$d_h^*(k,\Sigma_V) = \mathcal{E}[\mathcal{D}_h(\mathbf{q}_{1,k}, \hat{\mathbf{q}}_{1,k})] \qquad (17)$$

Indeed, if $C_m < r_m^q$ then $d_h^*(k,\Sigma_V) > 0$, which indicates the existence of distortion.

C. Analogy to the mode estimation problem

The diagram in Figure 2 portrays the mode estimation problem described in section II. Yet, by inspection, one can notice the structural similarity with Figure 1. The following is a list of the equivalences we establish:

- The process $\mathbf{q}_k$ represents the sequence of modes, but it can, instead, be viewed as a string of symbols placed for transmission through the communication scheme of Figure 1. In such analogy, a random code is set by $\mathbf{V}_k$, while the system equation (5) lays down an encoding rule.
- Measurements are produced according to the equation $\mathbf{F}_k = \mathbf{Y}_k + \mathbf{W}_k$. Hence, the noise $\mathbf{W}_k$ acts by creating a white Gaussian additive channel, as described by (13).
- Finally, the estimator can be viewed as a decoder.

[2]Notice that (15) is proven in [16], [17] for i.i.d. sources. Some extra work is needed to prove it for a Markovian source
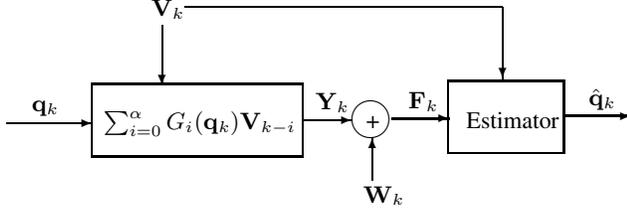
Fig. 2.   Analogy to the mode estimation problem

D. Preliminary Considerations

By means of the communication system analogy, we can use rate distortion theory to derive some of the fundamental limitations in mode estimation. From the outset, we perceive that (15) is a valid lower bound for the expected Hamming distortion, as a function of C_m and r_m^q. Doubtlessly, since (15) holds for any encoding strategy, it must hold for the encoder specified by (5). Such fact also exposes the influence of the Signal to Noise Ratios $\frac{\lambda_i(\Sigma_Y)}{\sigma_w^2}$ on $d_h^*(k, \Sigma_V)$. If C_m is small, when compared to r_m^q, then distortion will occur. As a final result of our analogy, we obtain the following, which holds for any estimator and probing signal:

$$d_h^*(k, \Sigma_V) \underset{m >> 1}{\approx} \frac{\Gamma(d_h^*(k, \Sigma_V))}{\ln(m)} + d_h^*(k, \Sigma_V) \frac{\ln(m-1)}{\ln(m)}$$
$$\geq r_m^q - C_m \quad (18)$$

where we use the fact that $\Gamma(d_h^*(k, \Sigma_V)) \leq 1$.

Notice that if a given agent does not want his mode to be accurately estimated then (18) suggests that it should keep C_m as low as possible. Equation (18) can also be used to compute what is the maximum C_m which guarantees a certain minimal level of distortion. In practice, there are two ways in which C_m can be lowered: by jamming the opponent's measurements, i.e., increasing Σ_W and by not responding to the opponent's probing signals, that is lowering Σ_Y. Alternatively, one can also use distraction (dummy) maneuvers to increase m or r_m^q.

IV. COMPARATIVE STUDY OF THE MEASURE $\mathcal{D}_d$

In contrast to the Hamming distortion, that has immediate implications in the symbol error frequency, it is not clear, in general, how to interpret $\mathcal{D}_d$. In this section, we address that concern by comparing $\mathcal{D}_d$ with the following measure of distortion:

Definition 4.1: The following is a distortion measure, which is sensitive to a mismatch in the first coefficient of the switching FIR system:

$$\mathcal{D}_g(q_{1,k}, \hat{q}_{1,k}) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{J}(q_i, \hat{q}_i) \quad (19)$$

where $\mathcal{J} : \mathbb{A} \times \mathbb{A} \to \mathbb{R}_+$ is given by:

$$\mathcal{J}(q, \hat{q}) = \frac{1}{2} \sum_{j=1}^{n^F} \ln\left(1 + \frac{\lambda_j(\tilde{\Sigma}(q, \hat{q}))}{\sigma_w^2}\right) \quad (20)$$

$$\tilde{\Sigma}(q, \hat{q}) = (G_0(q) - G_0(\hat{q})) \Sigma_V (G_0(q) - G_0(\hat{q}))^T \quad (21)$$

We also use $d_g^*(k, \Sigma_V)$ to indicate the following expected distortion:

$$d_g^*(k, \Sigma_V) = \mathcal{E}[\mathcal{D}_g(\mathbf{q}_{1,k}; \hat{\mathbf{q}}_{1,k})] \quad (22)$$

The previously defined measure has a natural interpretation, which is suggested by the following properties of $\mathcal{J}$:

- $\mathcal{J}(q, \hat{q}) \geq 0$ and $q = \hat{q} \implies \mathcal{J}(q, \hat{q}) = 0$.
- If $n^F = 1$ and $n^V = 1$ then the FIR system is single-input single-output and

$$\mathcal{J}(q, \hat{q}) = \frac{1}{2} \ln\left(1 + (g_0(q) - g_0(\hat{q}))^2 \frac{\sigma_v^2}{\sigma_w^2}\right) \quad (23)$$

- Under ergodicity assumptions, the following also holds:

$$d_g^*(k, \Sigma_V) \underset{k >> 1}{\approx} \mathcal{E}\left(\mathcal{J}(\mathbf{q}_k, \hat{\mathbf{q}}_k)\right) \quad (24)$$

In reality, $\mathcal{D}_g$ has technical advantages that permit a direct comparison with $\mathcal{D}_d$, while providing an insight similar to $\mathcal{D}_h$.

Lemma 4.1: The following relation holds:

$$\mathcal{D}_d(q_{1,k}; \hat{q}_{1,k}) \geq \mathcal{D}_g(q_{1,k}; \hat{q}_{1,k}) \quad (25)$$

Proof: We start by noticing that the law of iterated entropy [18], leads to:

$$\mathcal{H}\left(\mathbf{T}_{1,k}(q_{1,k}; \hat{q}_{1,k})\right) = \mathcal{H}\left(\mathbf{T}_1(q_1; \hat{q}_1)\right) +$$
$$\sum_{i=2}^{k} \mathcal{H}\left(\mathbf{T}_i(q_i; \hat{q}_i) | \mathbf{T}_{1,i-1}(q_{1,i-1}; \hat{q}_{1,i-1})\right) \quad (26)$$

and that, from the definition of $\mathbf{T}_i$, we arrive at:

$$\mathcal{H}\left(\mathbf{T}_i(q_i; \hat{q}_i) | \mathbf{T}_{1,i-1}(q_{1,i-1}; \hat{q}_{1,i-1})\right) \geq$$
$$\mathcal{H}\left((G_0(q_i) - G_0(\hat{q}_i))\mathbf{V}_i + \mathbf{W}_i\right) \quad (27)$$

$$\mathcal{H}\left(\mathbf{T}_1(q_1; \hat{q}_1)\right) \geq \mathcal{H}\left((G_0(q_1) - G_0(\hat{q}_1))\mathbf{V}_1 + \mathbf{W}_1\right) \quad (28)$$

The standard formula for the differential entropy [18], of a normally distributed vector, leads to:

$$\mathcal{H}\left((G_0(q) - G_0(\hat{q}))\mathbf{V}_k + \mathbf{W}_k\right) - \mathcal{H}(\mathbf{W}_k) =$$
$$\frac{1}{2} \sum_{j=1}^{n^F} \ln\left(1 + \frac{\lambda_j(\tilde{\Sigma}(q, \hat{q}))}{\sigma_w^2}\right) \quad (29)$$

The proof follows after the substitution of (27)-(29) into (26) and by inspection, of the formulae for $\mathcal{D}_d$ and $\mathcal{D}_g$, corresponding to the definitions 2.2 and 4.1, respectively.
□

A. Worst case analysis using $\mathcal{D}_g$

At this point, we can use lemma 4.1 and theorem 2.1 to reach a worst case analysis for the estimator in [2], in terms of $\mathcal{D}_g$. The following theorem encompasses such result:

Theorem 4.2: [2] For any given $\delta > 0$, there exists $k \in \mathbb{N}$ and an estimator which produces $\hat{\mathbf{q}}_{1,k}$, satisfying:

$$\mathcal{P}\left(\mathcal{D}_g(\mathbf{q}_{1,k}, \hat{\mathbf{q}}_{1,k}) > \ln(m)r_m^q + \frac{n^F}{2} + n^F\delta\right) < \delta \quad (30)$$

$$\mathcal{P}\left(\mathcal{D}_g(\mathbf{q}_{1,k}, \hat{\mathbf{q}}_{1,k}) > n^F\delta\right) < \delta, \text{ if } r_m^q = 0 \quad (31)$$

We can elucidate the repercussions of theorem 4.2 by focusing on the non-zero entropy rate. From (30) we find that, given an arbitrary $\delta > 0$, there exists an estimator, and k sufficiently large, for which the following holds:

$$\mathcal{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{J}(\mathbf{q}_i, \hat{\mathbf{q}}_i) < \ln(m)r_m^q + \frac{n^F}{2} + \delta\right) \approx 1 \quad (32)$$

In order to simplify the sub-sequential exposé, we define the following quantity:

Definition 4.2: Let $\tilde{C}$ be given by:

$$\tilde{C} = \min_{q \neq \hat{q}} \frac{1}{2}\sum_{i=1}^{n^F} \ln\left(1 + \frac{\lambda_i(\tilde{\Sigma}(q, \hat{q}))}{\sigma_w^2}\right) \quad (33)$$

By inspection, we find that $\mathcal{J}(q, \hat{q})$ satisfies:

$$\mathcal{J}(q, \hat{q}) \geq \mathcal{I}(q \neq \hat{q})\tilde{C} \quad (34)$$

As a consequence, given $\delta > 0$, we can state that there exists an estimator and k sufficiently large so that the following holds:

$$\mathcal{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}(\mathbf{q}_i \neq \hat{\mathbf{q}}_i) < \frac{\ln(m)r_m^q + \frac{n^F}{2} + \frac{\delta}{\ln(m)}}{\tilde{C}}\right) \approx 1 \quad (35)$$

Equally meaningful interpretations can be given by means of the approximation (24). In addition, we can determine that information can be reliably transmitted from $\mathbf{q}_k$ to $\hat{\mathbf{q}}_k$. In particular, (35) implies:

$$P_e \leq \frac{\ln(m)r_m^q + \frac{n^F}{2} + \frac{\delta}{\ln(m)}}{\tilde{C}} \quad (36)$$

where $P_e = \mathcal{P}(\mathbf{q}_i \neq \hat{\mathbf{q}}_i)$. But, from Fano's inequality [1], we have that $H(\mathbf{q}|\hat{\mathbf{q}}) \leq \Gamma(P_e) + P_e \log(m-1)$.

B. Influence of the input to noise ratio on the estimation distortion

In the subsequent subject matter, we explore the role of the input to noise ratio (INR) in the estimation fidelity. We perform the previously characterized analysis for the single-input single-output case ($n^V = n^F = 1$). The conclusions extend to the multi-input multi-output case, which we did not consider because it is notationally cumbersome.

We initiate by following the same steps, mutates muntandis, of the analysis in (34)-(35) and using (23) to arrive at the following:

$$\mathcal{P}\left(\frac{\sum_{i=1}^{k}\mathcal{I}\left(|g_0(\mathbf{q}_i) - g_0(\hat{\mathbf{q}}_i)| \geq \eta\right)}{k} < \right.$$
$$\left. \frac{r_m^q + \frac{1}{\ln(m)2} + \frac{\delta}{\ln(m)}}{\frac{1}{2}\log_m\left(1 + \max\{\eta^2, \eta_{min}^2\}INR\right)}\right) \approx 1 \quad (37)$$

where $\delta > 0$ is given, k is sufficiently large, $\eta > 0$ is arbitrary, $\eta_{min} = \min_{q \neq \hat{q}}|g_0(q) - g_0(\hat{q})|$ and INR is given by:

$$INR = \frac{\sigma_v^2}{\sigma_w^2}$$

Also, notice that if we adopt $\eta = \eta_{min}$, then $\mathcal{I}\left(|g_0(\mathbf{q}_i) - g_0(\hat{\mathbf{q}}_i)| \geq \eta\right) = \mathcal{I}(q \neq \hat{q})$ and the previous approximation reduces to (35). Additionally, we can write C_m as:

$$C_m = \frac{1}{2}\log_m\left(1 + \mu^2 INR\right) \quad (38)$$

where $\mu^2 = \frac{\sigma_y^2}{\sigma_v^2}$

Remark 4.1: From (18) and the previous analysis, we can infer the facts listed below:

- If $C_m = \frac{1}{2}\log_m\left(1 + \mu^2 INR\right) < r_m^q$ then Hamming distortion will occur. In order to attain small distortion, it is necessary that $C_m = \frac{1}{2}\log_m\left(1 + \mu^2 INR\right) > r_m^q - \epsilon$, where ϵ is also small.
- In addition to the previous item, we know, from (37), that as the ratio $\frac{r_m^q + \frac{1}{\ln(m)2} + \frac{\delta}{\ln(m)}}{\frac{1}{2}\log_m\left(1 + \max\{\eta^2, \eta_{min}^2\}INR\right)}$ decreases, the smaller the distortion, as measured by $\sum_{i=1}^{k}\mathcal{I}\left(|g_0(\mathbf{q}_i) - g_0(\hat{\mathbf{q}}_i)| \geq \eta\right)$.

In rate distortion theory, the difference given by $C_m - r_m^q$ determines, not only, a lower bound for the Hamming distortion, but it also provides an encoding/decoding scheme for which the converse of (18) holds, i.e.:

$$\frac{\Gamma(d_h^*(k, \Sigma_V))}{\ln(m)} + d_h^*(k, \Sigma_V)\frac{\ln(m-1)}{\ln(m)} \leq r_m^q - C_m + \epsilon \quad (39)$$

where $\epsilon > 0$ is arbitrarily small. In contrast to that, the approximation in (37), provides a distortion upper-bound that depends on the ratio $\frac{r_m^q + \frac{1}{\ln(m)2} + \frac{\delta}{\ln(m)}}{\frac{1}{2}\log_m\left(1 + \max\{\eta^2, \eta_{min}^2\}INR\right)}$. Especially in the cases where $\eta \approx \mu$, such upper-bound is close to $\frac{r_m^q + \frac{1}{\ln(m)2} + \delta}{C_m}$ and that is much worst than what is dictated by (39). On the other hand, that should be expected, since the theory leading to (39) allows any code, encoder and decoder, while (37) is derived for the very particular case depicted in Figure 2. Indeed, when viewed as an encoding/decoding scheme, the mode estimation problem imposes the following constraints:

- Shannon's theory suggests that, by using long codewords, information can be transmitted, with arbitrarily low Hamming distortion, at rates which are arbitrarily

close to capacity. In such scheme, the information stream $\mathbf{q}_k$ is grouped into blocks. Each block is encoded as a whole and placed for transmission. The decoding is also performed block-wise, thereby introducing a delay in the communication system. In the mode estimation problem, we do not have such luxury. The stream $\mathbf{q}_k$ is encoded as it occurs and that rules out the use of Shannon's technique.

- The other limitation is that, when compared to other coding schemes, the FIR switching system encodes $\mathbf{q}_k$ in a constrained fashion. As an illustration, consider the 0-th order FIR with $g_0(1) = 1$ and $g_0(2) = -1$. In that case, the channel code for symbol 2, say $Y_k = g_0(2)v_k$, has always the same absolute value, but a different sign, when compared to the code for symbol 1 which is given by $Y_k = g_0(1)v_k$.

C. Lower bound for $d_d^*(k, \Sigma_V)$

We recall that the upper-bound radius in theorem 2.1 is, up to an arbitrarily small term, given by

$$\bar{d} = r_m^q \ln(m) + \frac{n^F}{2}$$

and that it does not depend on Σ_V nor the estimation horizon k. In the analysis below, we show that if $r_m^q > 0$ then, even if $\bar{d}$ might be conservative, it could never be zero. Since $d_d^*(k, \Sigma_V) \leq \bar{d}$, for all Σ_V and k, it suffices to look for a lower bound for $d_d^*(k, \Sigma_V)$. We denote such lower-bound by $\underline{d}^*(\Sigma_V)$.

We begin with the coming remark, which establishes an inequality between $\mathcal{D}_g$ and $\mathcal{D}_h$:

Remark 4.2: By means of (34), we can derive the following:

$$\mathcal{D}_g(q_{1,k}; \hat{q}_{1,k}) \geq \tilde{C}\mathcal{D}_h(q_{1,k}; \hat{q}_{1,k}) \qquad (40)$$

Consequently, $d_g^*(k, \Sigma_V)$ and $d_h^*(k, \Sigma_V)$ satisfy the inequality below:

$$d_d^*(k, \Sigma_V) \geq \tilde{C}d_h^*(k, \Sigma_V) \qquad (41)$$

The following theorem establishes that, if $r_m^q > 0$, and the FIR system has distinct zero order coefficients, then we can encounter Σ_V for which there exists $\underline{d}^*(\Sigma_V) > 0$ satisfying $d_d^*(k, \Sigma_V) \geq \underline{d}^*(\Sigma_V)$ for all k.

Theorem 4.3: If $r_m^q > 0$ and $(q \neq \hat{q}) \implies (G_0(q) \neq G_0(\hat{q}))$ then there exists Σ_V and $\underline{d}^*(\Sigma_V) > 0$ such that $d_d^*(k, \Sigma_V) > \underline{d}^*(\Sigma_V)$ for all k.
Proof: Choose Σ_V such that $\tilde{C} > 0$ and, in addition, re-scale Σ_V such that $C_m \leq r_m^q$. The result follows by sequential substitution in (15) and (41). $\square$

D. Design of probing Signals by means of LMI's

Consider the following problem: given a constant $pwr > 0$ and a system in the form (5), design Σ_V under the power constraint $tr(\Sigma_V) \leq pwr$. Motivated by (35) and from the definition of $\tilde{C}$, we suggest the following optimization:

$$\begin{cases} \max \lambda \\ \text{subject to} \quad \min_{q \neq \hat{q}} \tilde{\Sigma}(q, \hat{q}) - \lambda I > 0, \ tr(\Sigma_V) < pwr \end{cases}$$

Notice that the constraints are linear in Σ_V, which makes the optimization above solvable by means of an LMI solver [15]. Let λ^{max} denote the optimal solution. By direct substitution on the definition of $\tilde{C}$, we find that the resulting probing signal leads to:

$$\tilde{C} \geq \frac{n^F}{2} \ln(1 + \frac{\lambda^{max}}{\sigma_w^2})$$

V. ACKNOWLEDGMENTS

REFERENCES

[1] Cover, T.M; Thomas, J. A.; *"Elements of Information Theory"*, Wiley-Iterscience Publication, 1991
[2] Martins, N. C. and Dahleh, M. A., *"An Information Theoretic Approach to the Mode Estimation of Switching FIR Linear Systems"*, IEEE Proc. of the CDC, pp. 4152-7, Maui, 2003
[3] *Mode Estimation of Probabilistic Hybrid Systems*, M. W. Hofbaur and B. C. Williams, HSCC 2002, LNCS 2289, pp. 253-266, Springer Verlag, 2002
[4] Campi, M. C., *"Exponentially Weighted Least Squares Identification of Time-Varying Systems with White Disturbances"* IEEE Tr. Sig. Pr., Vol. 42, pp. 2906-14, 1994
[5] Arimoto, S. and Kimura H., *"Optimum input test signals for system identification - an information-theoretical approach"* Int. J. Syst. Sci., No 3, Vol I, pp. 279-290, 1971
[6] Battail, G. *"On random-like codes."* Information Theory and Applications II, Lecture notes in Computer Science 1133, pp. 76-94, Springer 1995
[7] Blom, H. A. P. and Bar-Shalom, Y. *"The interacting multiple model algorithm for systems with Markovian switching coefficients."* IEEE Trans. A. C., Vol. 33, No. 8, pp. 780-783, 1988
[8] Sworder, D.D.; Boyd J.E. *Estimation Problems in Hybrid Systems* Cambridge University Press, 1999
[9] Sworder, D. D.; Boyd J. E.; Elliott, R. J. *"Modal Estimation in Hybrid Systems"* Journal of Math. Analis. and Ap. 245, 225-247 (2000)
[10] Venkatesh, S. R.; Dahleh, M. A. *"On System Identification of Complex Systems from Finite Data"* IEEE Trans. A. C., vol 46, no. 2, 235-257, 2001
[11] Zames, G.; Lin, L.; Wang L. Y. *"Fast Identification n-Widths and Uncertainty Principles for LTI and Slowly Varying Systems"* IEEE Trans. A. C., vol 39, no 9, 1827-1838, 1994
[12] Liberzon, D.; Morse A. S. *Basic Problems in Stability and Design of Switched Systems* IEEE Cont. Syst. Mag., pp. 1-19, Oct 1999
[13] Tsumura, K., Oishi Y., *Optimal Length of Data for Identification of Time Varying Systems* IEEE CDC'99, pp 3224-3229
[14] Belforte, G., Gay, P., *Optimal Worst Case Estimation for LPV-FIR Models with Bounded Errors* CDC2000, pp. 4573-77
[15] Boyd, S.;Ghaoui, L. El;Feron, E.; Balakrishnan, V., *"Linear Matrix Inequalities in System and Control Theory"*,SIAM Studies in Applied Mathematics, 1994
[16] Csiszar, I.; Korner J., *"Information Theory: Coding Theorems for Discrete Memoryless Systems"*, Academic Press, 1981
[17] Jerohin, V. D., *"ε entropy of discrete random objects (in Russian)"*, Teor. Veroyatnost. i Primenen. 3, 103-107, 1958
[18] Gallager, R., *"Information Theory and Reliable Communication"*, John Wiley and Sons, 1968