Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Optimal input design for identification of systems with quantized measurements

Marco Casini, Andrea Garulli, Antonio Vicino

Abstract— **This paper addresses system identification of FIR models with quantized measurements in a worst-case setting. It is assumed that measurements are collected through a multi-threshold sensor and that the system output is corrupted by unknown but bounded noise. The main contribution of the paper consists in the solution of the optimal input design problem for identification of a scalar gain. This result allows one to design a suboptimal input for a FIR model of arbitrary order. Moreover, for a selected configuration of the sensor thresholds, an upper bound on the time complexity of the identification problem is derived.**

I. Introduction

System identification with quantized measurements is gaining increasing attention both for the number of contexts where analog-to-digital conversion is needed and for the attractive theoretical developments attained in recent years. Typical contexts involving quantized measurements are sensor networks and networked control systems. Since data transmission band limitations are usual constraints in large/integrated systems interconnected through communication channels, the need for studying the quality of identification algorithms as related to the measurement quantization has become a typical issue in networked control systems. Moreover, most sensors used in monitoring and control systems for industrial production plants or chemical processes are quantized devices, in the sense that they are characterized by threshold values, according to which the output is digitized. Quantized measurements occur when several single-threshold sensors are used to measure the same quantity or the sensor is a multi-threshold device. Binary or quantized measurements are often used in automotive applications, where optimization of the ignition system and combustion is an important task. The basic reason for the widespread diffusion of binary/quantized sensors is mainly related to their relative low cost and to the fact that, no matter how simple are the control laws adopted, monitoring and control of industrial plants often calls for the measurement of a large number of variables.

The seminal paper [1] introduced and motivated very clearly the basic identification problem, providing a framework to deal with the binary information case. Of course, while the literature on regular measurements provides exact or approximated solutions to almost all of the basic issues, like estimation quality evaluation, model complexity and unmodeled dynamics, optimal input design, time complexity and estimate convergence (see, e.g. [2]), the scenario is quite different when dealing with quantized measurements. The main difficulty in this case is to deal with a discontinuous nonlinearity present in the sensor, which reduces drastically the information conveyed by measurements. In [1] several important results on time complexity and input design have been obtained for FIR models both in a stochastic and a worst-case setting. The quantized measurement case has been addressed in [3], [4] in a stochastic setting. In the second paper a complete characterization of optimal estimators for system gains is provided together with a suboptimal input design strategy for FIR models in the multi-threshold sensor case. In [5] a worst-case setting is taken based on the set membership paradigm of uncertainty representation [6]–[8], and time complexity for system gain estimates is computed exactly for the binary case. Also, a suboptimal input design strategy for FIR models is devised in the same paper.

The aim of the present paper is to extend the analysis in [5] to the multi-threshold sensor case. More specifically, system gains estimation is dealt with and an algorithm for one step optimal input selection is provided for such problem. This result can be used to devise a suboptimal input for a generic FIR model in the same way as in [5]. As expected, the time complexity issue reveals a formidable task for the quantized measurement case. For a specific threshold selection strategy, an upper bound on the time complexity is derived, showing the improvement achievable with respect to the binary case.

The paper is organized as follows. Section II introduces notation and problem formulation. In Section III the problem of optimal input design for the noise-free case is tackled. These results are extended to the noisy measurements case in Section IV. In Section V an upper bound on the time complexity is reported. Numerical examples are presented in Section VI, while concluding remarks and future perspectives are reported in Section VII.

II. Problem formulation

Let $\mathbb{R}^N$ denote the N-dimensional Euclidean space. A sequence of real numbers $\{x(t), t = 1, \dots, N\}$ will be identified with a vector $x \in \mathbb{R}^N$ and $||x||_p$ will be the standard ℓ_p norm. Let $B_p(c, r) = \{x \in \mathbb{R}^N : ||x - c||_p \leq r\}$ be the ball of radius $r \geq 0$ and center $c \in \mathbb{R}^N$ in the ℓ_p norm.

Let us consider an n-th order FIR SISO linear time-invariant model

$$y(t) = \sum_{i=1}^{n} \theta_i \, u(t - i + 1) + d(t) \qquad (1)$$

The authors are with the Dipartimento di Ingegneria dell'Informazione, Università di Siena, 53100 Siena, Italy.
Email: {casini,garulli,vicino}@ing.unisi.it

where $u(t)$ is the input signal, bounded in the max norm $||u||_\infty \leq U$, and $d(t)$ denotes the output disturbance. The model parameters $\{\theta_i, i = 1, \dots\ n\}$ represent the truncated system impulse response. The disturbance $d(t)$ is assumed to be bounded by a known quantity, i.e., $|d(t)| \leq \delta,\ t = 1, 2, \dots$. The true system generating the data is assumed to be exponentially stable. Notice that due to exponential stability of the system and boundedness of the input $u(t)$, unmodeled dynamics (i.e., the system impulse response tail $\{\theta_i, i = n+1, \dots\}$) can be easily accounted for by suitably tuning the noise bound δ.

Observations at the system output are taken by a multi-valued sensor with P known thresholds $C_1, \dots, C_P$, such that

$$s(t) = \sigma(y(t)) \triangleq \begin{cases} 0 & \text{if } C_0 < y(t) \leq C_1 \\ 1 & \text{if } C_1 < y(t) \leq C_2 \\ & \vdots \\ P & \text{if } C_P < y(t) \leq C_{P+1} \end{cases} \tag{2}$$

where $C_0 \triangleq -\infty$, $C_{P+1} \triangleq +\infty$.

Let $\theta^T = [\theta_1, \theta_2, \dots, \theta_n] \in \mathbb{R}^n$ denote the FIR parameter vector and $\phi^T(t) = [u(t), \dots, u(t-n+1)]$ the regressor vector. Then, (1) can be expressed as

$$y(t) = \phi^T(t)\,\theta + d(t). \tag{3}$$

Let $\Theta_0 = B_p(c_0, \varepsilon_0)$ represent the prior information available on the FIR parameter vector. Let us denote by $u, s \in \mathbb{R}^N$ the input signal $\{u(t),\ t = 1, \dots, N\}$ and the sequence of discrete measurements $\{s(t),\ t = 1, \dots, N\}$, respectively. For a given input-output realization $\{u, s\}$ of length N, the problem feasible parameter set is defined as:

$$\begin{aligned} \mathcal{F}_N = \{\theta \in \Theta_0 :\ & \phi^T(t)\,\theta \leq C_1 + \delta \ \text{ if } s(t) = 0; \\ & C_1 - \delta < \phi^T(t)\,\theta \leq C_2 + \delta \ \text{ if } s(t) = 1; \\ & \vdots \\ & C_P - \delta < \phi^T(t)\,\theta \ \text{ if } s(t) = P;\ t = 1, \dots, N\}. \end{aligned} \tag{4}$$

The worst-case local identification error is defined as

$$e_p(N, u, s) = \inf_{c \in \mathbb{R}^n} \sup_{\theta \in \mathcal{F}_N} \|\theta - c\|_p\ . \tag{5}$$

For a fixed input sequence u, let us define the global worst-case error with respect to the disturbance realization as

$$e_p(N, u) = \sup_s e_p(N, u, s)\ . \tag{6}$$

The aim of the optimal input design problem is to compute an input sequence providing the minimum worst-case identification error [9], i.e.,

$$e_p(N) = \inf_{u: \|u\|_\infty \leq U} e_p(N, u)\ . \tag{7}$$

For a given level of accuracy $\varepsilon < \varepsilon_0$, we define the time complexity of $B_p(c_0, \varepsilon_0)$ as the minimum time length of the experiment such that the optimal worst-case error reaches the accuracy ε, i.e.

$$N(\varepsilon) = \min_{e_p(N) \leq \varepsilon} N. \tag{8}$$

In the literature on identification with quantized measurements, it is customary to restrict the class of input signals by requiring that the input sequence excite any FIR parameter independently (see e.g., [1], [5]). In [5], the shortest input sequence exciting any FIR parameter independently has been provided. More precisely, it has been proven that to excite independently k times the n coefficients of a FIR, one needs an input sequence of length N equal to:

$$N = k(n+1)\frac{n}{2}. \tag{9}$$

In this paper, we will assume to use such kind of input sequence. This allows one to focus on the optimal excitation of a single FIR parameter, in order to build effective suboptimal procedures for FIR models of arbitrary order. Hence, in the next sections, the problem of optimal input design for identification of FIR systems of order 1 (gains) for both the noise-free and noisy case will be addressed.

III. Identification of gains: noise-free case

In this section, we consider a FIR of order $n = 1$ in the noise-free case, i.e.,

$$y(t) = a u(t).$$

Let us assume that the sign of a is known. If it is unknown, it can be easily detected by performing an initial testing condition as reported in [1]. Hereafter, we assume $a > 0$, and so the prior information is $a \in [\underline{a}_0, \overline{a}_0]$, with $\underline{a}_0 > 0$. Moreover, let the input signal be bounded, i.e., $0 < u(t) \leq U$.

Let us denote by $\mathcal{F}_t = [\underline{a}_t, \overline{a}_t]$ the feasible parameter set at time t. The aim of the input design problem (7) is to choose the input signal $u(t)$ as a function of the available information up to time $t-1$, i.e.,

$$u(t) = \eta(\mathcal{F}_{t-1}; t)$$

in order to minimize the size of the feasible set at time t, that is

$$u^*(t) = \arg\ D_t^* \tag{10}$$

where D_t^* is the optimal *diameter* of the feasible set:

$$D_t^* = \inf_{u:\ 0<u\leq U} \sup_{\substack{s:\ s=\sigma(y) \\ y \in u \cdot \mathcal{F}_{t-1}}} (\overline{a}_t - \underline{a}_t). \tag{11}$$

Remark 1: In [1] it has been shown that the optimal input at each time t in presence of binary measurements is

$$u^*(t) = \frac{2C}{\underline{a}_{t-1} + \overline{a}_{t-1}}$$

where C denotes the binary threshold value. By applying such an input, the feasible set size is reduced by a factor $1/2$ at each time t, i.e., $\overline{a}_t - \underline{a}_t = \frac{1}{2}(\overline{a}_{t-1} - \underline{a}_{t-1})$.

Let us define

$$v(t) \triangleq \frac{1}{u(t)}\ . \tag{12}$$

Let us suppose to apply an input $u(t)$, and let the sensor response be $s(t) = i,\ i = 0, \dots, P$. This means that $C_i < y(t) \leq C_{i+1}$, i.e., (since $u(t) > 0$)

$$C_i\, v(t) < a \leq C_{i+1}\, v(t). \tag{13}$$

Thus, the posterior feasible set will be[1]

$$\begin{aligned} \mathcal{F}_t &= \mathcal{F}_{t-1} \cap [C_i\, v(t),\, C_{i+1}\, v(t)] \\ &= [\underline{a}_{t-1},\, \overline{a}_{t-1}] \cap [C_i\, v(t),\, C_{i+1}\, v(t)] \triangleq [\underline{a}_t\,,\, \overline{a}_t]. \end{aligned} \tag{14}$$

From (12)-(14), we can rewrite problem (10)-(11) as

$$u^*(t) = \frac{1}{v^*(t)}$$

where

$$\begin{aligned} v^*(t) = \arg \Big\{ \inf_{v \geq 1/U} \max_{i=0,\dots,P} \\ \big(\min\{\overline{a}_{t-1}, C_{i+1}\, v\} - \max\{\underline{a}_{t-1}, C_i\, v\}\big) \Big\}. \end{aligned} \tag{15}$$

Let us define

$$\underline{v}_i(t-1) \triangleq \frac{\underline{a}_{t-1}}{C_i}\ ,\ \overline{v}_i(t-1) \triangleq \frac{\overline{a}_{t-1}}{C_i}\ ,\ i = 1, \dots, P. \tag{16}$$

Since we focus on the optimal input design at a generic time t, for ease of notation the dependance on time will be omitted when it is clear from the context. So, the feasible set at time $t-1$ will be denoted by $\mathcal{F} = [\underline{a}, \overline{a}]$. Thus, let $V^* \triangleq [\underline{v}_P,\, \overline{v}_1]$ and

$$H_i \colon a = C_i\, v \quad , \quad i = 1, \dots, P.$$

In Fig. 1, the values $\underline{v}_i$, $\overline{v}_i$ and the lines H_i are depicted for an example with $P = 3$.

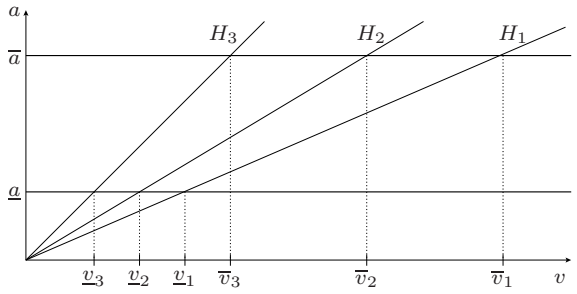

Fig. 1. Example of $\underline{v}_i$, $\overline{v}_i$ and H_i for $P = 3$.

We can now state the following lemma.

Lemma 1: There exists an optimal solution of (15) such that $v^* \in V^*$.

Proof: By contradiction, let us assume that $\nexists v^* \in V^*$. For example, let $v^* < \underline{v}_P$. One has

$$y = a\, u^* = \frac{a}{v^*} > \frac{a}{\underline{a}}\, C_P > C_P.$$

So, the sensor output is $s = P$ independently of the real position of a, and hence no reduction of the feasible set is obtained. Thus, either v^* is not optimal or any element of V^* is also optimal. Both cases lead to a contradiction. A similar reasoning can be repeated for the case $v^* > \overline{v}_1$. ■

[1]With a slight abuse of notation we will always denote feasible sets by closed intervals.

By using Lemma 1, the optimal input is such that:

$$u^*(t) \leq \sup_{v^* \in [\underline{v}_P(0),\, \overline{v}_1(0)]} \frac{1}{v^*} = \frac{1}{\underline{v}_P(0)} = \frac{C_P}{\underline{a}_0}.$$

Hence, a sufficient condition for the optimal input to satisfy $u^*(t) \leq U$, is $U \geq \frac{C_P}{\underline{a}_0}$. From now on, we will enforce this hypothesis.

Let us sort the values $\underline{v}_i,\ \overline{v}_i,\ i = 1, \dots, P$, in increasing order and rename them as $\hat{v}_1 \leq \hat{v}_2 \leq \dots \leq \hat{v}_{2P}$. By construction one has $\hat{v}_1 = \underline{v}_P$ and $\hat{v}_{2P} = \overline{v}_1$. Let us define the intervals

$$W_1 = [\hat{v}_1, \hat{v}_2],\ W_2 = [\hat{v}_2, \hat{v}_3],\ \dots\ ,\ W_{2P-1} = [\hat{v}_{2P-1}, \hat{v}_{2P}]. \tag{17}$$

By construction

$$\bigcup_{j=1}^{2P-1} W_j = V^*. \tag{18}$$

For $j = 1, \dots, 2P-1$, let us define

$$D^{(j)} = \inf_{v \in W_j} \max_{i=0,\dots,P} (\min\{\overline{a},\, C_{i+1}\, v\} - \max\{\underline{a},\, C_i\, v\}) \tag{19}$$

and $v^{(j)}$ be the argument where the infimum in (19) is achieved.

Let us now analyze problem (19), i.e., the original problem (15) whose admissible solution set is restricted to an interval W_j. To simplify notation, let us drop the index j and denote the left and right bounds of the interval by v_L and v_R, respectively, i.e., $W_j = [v_L,\, v_R]$.

Let us define

$$m = \arg \min_{i=1,\dots,P} \{i \colon \underline{v}_i \leq v_L\} \tag{20}$$

$$M = \arg \max_{i=1,\dots,P} \{i \colon \overline{v}_i \geq v_R\}. \tag{21}$$

The following lemma holds.

Lemma 2: Let $v \in [v_L, v_R]$. For each $k < m$ one has $C_k v \leq \underline{a}$, while for each $k > M$ one has $C_k v \geq \overline{a}$.

Proof: Let us assume $k < m$. By (20) one has $v_L < \underline{v}_k$. Since by construction $\underline{v}_k$ can only be an extremal point of the interval $[v_L,\, v_R]$, one has $\underline{v}_k \geq v_R$. So, it follows that

$$C_k\, v \leq C_k\, v_R \leq C_k\, \underline{v}_k = C_k \frac{\underline{a}}{C_k} = \underline{a}.$$

Let us now consider $k > M$; hence, one has $v_R > \overline{v}_k$. Following the same reasoning, one has $\overline{v}_k \leq v_L$, giving

$$C_k\, v \geq C_k\, v_L \geq C_k\, \underline{v}_k = C_k \frac{\overline{a}}{C_k} = \overline{a}.$$ ■

Lemma 2 states that for $v \in [v_L,\, v_R]$, the only thresholds that are able to reduce the size of the feasible set are C_i such that $m \leq i \leq M$. The other thresholds do not provide any additional information and so can be neglected when addressing problem (19). For instance, referring to Fig. 1, let us suppose $v_L = \overline{v}_3$ and $v_R = \overline{v}_2$. Then, one has $m = 1$ and $M = 2$. In fact, only the functions H_1 and H_2 take on values in the interval $[\underline{a},\, \overline{a}]$, for $v \in [v_L,\, v_R]$. Notice that, by Lemma 2, if $m > M$ it follows that no reduction of $\mathcal{F}$ would be obtained for u such that $v \in [v_L,\, v_R]$.

Lemma 2 allows us to compute the feasible set $\mathcal{F}$ at time t as a function of the measurements $s(t)$, when $v \in [v_L,\, v_R]$:

$$\mathcal{F} = \begin{cases} [\underline{a}\,,\, C_m\, v] & \text{if } s = m-1 \\ [C_s\, v\,,\, C_{s+1}\, v] & \text{if } m \le s \le M-1 \\ [C_M\, v\,,\, \overline{a}] & \text{if } s = M. \end{cases}$$

Let us define the functions $F_i(v)$, $i = m-1, \ldots, M$, as follows

$$\begin{aligned} F_{m-1}(v) &= C_m\, v - \underline{a} \\ F_i(v) &= (C_{i+1} - C_i)\, v\ ,\ i = m, \ldots, M-1 \\ F_M(v) &= \overline{a} - C_M\, v\ . \end{aligned} \tag{22}$$

Notice that functions F_i represent the size of $\mathcal{F}$ depending on s and on the input applied at time t such that $v \in [v_L,\, v_R]$. Then, (19) can be rewritten as

$$D^{(j)} = \inf_{v \in [v_L,\, v_R]} \max_{i=m-1,\ldots,M} F_i(v). \tag{23}$$

The next lemma provide the explicit solution of problem (23).

Lemma 3: Let $Q = \max\limits_{i=m,\ldots,M-1} (C_{i+1} - C_i)$. Then, the solution of (23) is given by

$$D^{(j)} = \begin{cases} \max\{C_m\, v_L - \underline{a},\, Q\, v_L\} & \text{if } v_c \le v_L \\ \max\left\{\frac{\overline{a}\, C_m - \underline{a}\, C_M}{C_m + C_M},\, \frac{Q\,\overline{a}}{Q + C_M}\right\} & \text{if } v_L < v_c < v_R \\ \overline{a} - C_M\, v_R & \text{if } v_c \ge v_R \end{cases}$$

where

$$v_c = \min\left\{\frac{\underline{a} + \overline{a}}{C_m + C_M},\, \frac{\overline{a}}{Q + C_M}\right\}.$$

Moreover, the argument at which the solution of (23) is attained is

$$v^{(j)} = \begin{cases} v_L & \text{if } v_c \le v_L \\ v_c & \text{if } v_L < v_c < v_R \\ v_R & \text{if } v_c \ge v_R. \end{cases}$$

Proof: Let

$$q = \arg \max_{i=m,\ldots,M-1} (C_{i+1} - C_i). \tag{24}$$

Since for any $v \in [v_L,\, v_R]$ one has $F_i(v) \le F_q(v) = Q\, v$, for any $i = m, \ldots, M-1$, it is possible to rewrite (23) as

$$\begin{aligned} D^{(j)} &= \inf_{v \in [v_L,\, v_R]} \max\{F_{m-1}(v),\, F_q(v),\, F_M(v)\} \\ &= \inf_{v \in [v_L,\, v_R]} \max\{C_m\, v - \underline{a},\, Q\, v,\, \overline{a} - C_M\, v\}\,. \end{aligned} \tag{25}$$

Being $F_i(v)$ linear in v, the $v^{(j)}$ at which the infimum in (25) is achieved lies either at the extremes v_L, v_R, or at one of the intersections between $F_{m-1}(v)$, $F_q(v)$, $F_M(v)$.
First, let us suppose that the optimum is achieved at some $\widetilde{v} \in (v_L,\, v_R)$. We want to show that $\widetilde{v}$ cannot be such that $F_{m-1}(\widetilde{v}) = F_q(\widetilde{v}) > F_M(\widetilde{v})$. Indeed, being $F_{m-1}(v)$ and $F_q(v)$ increasing functions of v, there exists $\varepsilon > 0$ such that $\widetilde{v} - \varepsilon \in (v_L,\, v_R)$, $F_M(\widetilde{v} - \varepsilon) < F_{m-1}(\widetilde{v} - \varepsilon) < F_{m-1}(\widetilde{v})$ and $F_M(\widetilde{v} - \varepsilon) < F_q(\widetilde{v} - \varepsilon) < F_q(\widetilde{v})$. This leads to a contradiction, because

$$\begin{aligned} &\max\{F_{m-1}(\widetilde{v} - \varepsilon),\, F_q(\widetilde{v} - \varepsilon),\, F_M(\widetilde{v} - \varepsilon)\} < \\ &\qquad < \max\{F_{m-1}(\widetilde{v}),\, F_q(\widetilde{v}),\, F_M(\widetilde{v})\}. \end{aligned}$$

Now, let us define v_{m-1} and v_q satisfying respectively $F_{m-1}(v_{m-1}) = F_M(v_{m-1})$, $F_q(v_q) = F_M(v_q)$. It is immediate to check that

$$v_{m-1} = \frac{\underline{a} + \overline{a}}{C_m + C_M} \quad , \quad v_q = \frac{\overline{a}}{Q + C_M}.$$

According to the above reasoning the only candidate solutions $v^{(j)}$ are v_L, v_R, v_{m-1}, v_q. By noticing that $v_c = \min\{v_{m-1},\, v_q\}$, one has that only the following three cases can occur.
i) $v_c \le v_L$. Being $F_{m-1}(v)$ and $F_q(v)$ increasing functions of v and $F_M(v)$ a decreasing function of v, this means that $\max\{F_{m-1}(v),\, F_q(v)\} \ge F_M(v)$, $\forall v \in [v_L,\, v_R]$. Hence, the minimum is attained at v_L and takes on the value $\max\{F_{m-1}(v_L),\, F_q(v_L)\} = \max\{C_m\, v_L - \underline{a},\, Q\, v_L\}$.
ii) $v_L < v_c < v_R$. In this case the minimum is attained at v_c and the corresponding feasible set size turns out to be $\max\{F_M(v_{m-1}),\, F_M(v_q)\} = \max\left\{\frac{\overline{a}\, C_m - \underline{a}\, C_M}{C_m + C_M},\, \frac{Q\,\overline{a}}{Q + C_M}\right\}$.
iii) $v_c \ge v_R$. This means $F_M(v) \ge \max\{F_{m-1}(v),\, F_q(v)\}$, $\forall v \in [v_L,\, v_R]$ and then the minimum is attained at v_R and takes on the value $F_M(v_R) = \overline{a} - C_M\, v_R$. ■

For any fixed j, Lemma 3 gives the solution of problem (19). The following theorem providing the optimal solution of the original problem (10) is a direct consequence of Lemma 3.

Theorem 1: At a given time, the optimal solution u^* of (10) is given by $u^* = \dfrac{1}{v^*}$, where

$$v^* = v^{(j^*)}$$

and

$$j^* = \arg \min_{j=1,\ldots,2P-1} D^{(j)}.$$

The corresponding size of the feasible set turns out to be

$$D^* = D^{(j^*)}.$$

Remark 2: Note that in the case $P = 1$, i.e., binary measurements, $V^* = W_1 = [\underline{v}_1, \overline{v}_1]$, and so the optimal solution of (10) coincides with that provided by Lemma 3. In this case, one has $m = M = 1$, $F_0(v) = C_1\, v - \underline{a}$ and $F_1(v) = \overline{a} - C_1\, v$. The candidate minimizer is $v_c = \frac{\underline{a} + \overline{a}}{2}\,\frac{1}{C_1}$ which by construction belongs to $[v_L,\, v_R] = [\frac{\underline{a}}{C_1}, \frac{\overline{a}}{C_1}]$. So, the optimal input is $u^* = \frac{2\, C_1}{\underline{a} + \overline{a}}$, in accordance with what reported in Remark 1.

Notice that the maximum reduction rate of the feasible set achievable in one step is $\frac{1}{P+1}$, i.e. $diam(\mathcal{F}_t) \ge \frac{1}{P+1}\, diam(\mathcal{F}_{t-1})$. In fact, for any fixed v, the P functions H_i, $i = 1, \ldots, P$, can divide the interval $[\underline{a}, \overline{a}]$ at most in $P+1$ subintervals. Since in a worst-case setting, the output is such to choose the larger subinterval, one has $diam(\mathcal{F}_t) \ge \frac{1}{P+1}\, diam(\mathcal{F}_{t-1})$.

Based on the above observation, we can now state the following result.

Theorem 2: There exists an input $u^* \le U$ such that $diam(\mathcal{F}_t) = \frac{1}{P+1}\, diam(\mathcal{F}_{t-1})$ if and only if there exists $\hat{u} \le U$ such that

$$\hat{u} = C_i \left[\frac{P+1-i}{P+1}\,\underline{a} + \frac{i}{P+1}\,\overline{a}\right]^{-1}\ ,\ i = 1, \ldots, P. \tag{26}$$

Moreover, if (26) holds, then $u^* = \hat{u}$.

Proof: Let us assume that (26) holds and let $\hat{v} = 1/\hat{u}$. Since

$$\underline{v}_1 = \frac{\underline{a}\,\hat{v}}{\frac{P}{P+1}\underline{a} + \frac{1}{P+1}\overline{a}} \leq \hat{v}$$

and

$$\overline{v}_P = \frac{\overline{a}\,\hat{v}}{\frac{1}{P+1}\underline{a} + \frac{P}{P+1}\overline{a}} \geq \hat{v}$$

one has $\underline{v}_1 \leq \overline{v}_P$. By (16), there cannot be other breakpoints $\hat{v}_i$ in the interval $[\underline{v}_1,\, \overline{v}_P]$. Hence, according to (20) and (21), one has $m = 1$ and $M = P$. By (22) and (26) one has

$$\begin{aligned} F_0(\hat{v}) &= C_1\,\hat{v} - \underline{a} = \frac{\overline{a} - \underline{a}}{P+1} \\ F_i(\hat{v}) &= (C_{i+1} - C_i)\,\hat{v} = \frac{\overline{a} - \underline{a}}{P+1}\ ,\ i = 1, \ldots, P-1 \quad (27) \\ F_P(\hat{v}) &= \overline{a} - C_P\,\hat{v} = \frac{\overline{a} - \underline{a}}{P+1}. \end{aligned}$$

Since the maximum possible reduction of the feasible set is by a factor of $P+1$, one has $D^* = \frac{\overline{a}-\underline{a}}{P+1}$ and $v^* = \hat{v}$. Conversely, assume that the maximum reduction is achieved at some $\hat{u} = \frac{1}{\hat{v}}$. Then, the relationships (27) must hold and (26) easily follows. ■

IV. Identification of gains: noisy case

In this section, let us consider the noisy case, i.e.,

$$y(t) = au(t) + d(t)$$

Let $s(t)$ denote the sensor output and let $0 < u(t) \leq U$, $\delta > 0$ and $a \in [\underline{a}_0, \overline{a}_0]$. Assume that $U \geq (C_P - \delta)/\underline{a}_0 > 0$ and

$$C_1 > \delta. \quad (28)$$

It will be shown that the optimal input procedure for the noisy case can be formulated in a similar manner w.r.t. the noise-free case. So, all the quantities defined in Section III will be redefined accordingly.

Due to the presence of noise, (13) becomes

$$(C_i - \delta)\,v(t) < a \leq (C_{i+1} + \delta)\,v(t). \quad (29)$$

Thus, the posterior feasible set is

$$\begin{aligned} \mathcal{F}_t &= \mathcal{F}_{t-1} \cap [(C_i - \delta)\,v(t)\,,\,(C_{i+1} + \delta)\,v(t)] \\ &= [\underline{a}_{t-1},\, \overline{a}_{t-1}] \cap [(C_i - \delta)\,v(t)\,,\,(C_{i+1} + \delta)\,v(t)] \\ &\triangleq [\underline{a}_t,\, \overline{a}_t]. \quad (30) \end{aligned}$$

Let us define

$$\underline{v}_i \triangleq \frac{\underline{a}}{C_i - \delta} \quad , \quad \overline{v}_i \triangleq \frac{\overline{a}}{C_i + \delta} \quad , \quad i = 1, \ldots, P \quad (31)$$

and

$$H_i^+ \colon a = (C_i + \delta)\,v\ ,\ H_i^- \colon a = (C_i - \delta)\,v\ ,\ i = 1, \ldots, P. \quad (32)$$

The optimal input design problem can be reformulated as in (15)

$$v^*(t) = \arg\Big\{ \inf_{v \geq 1/U} \max_{i:\, i=0,\ldots,P} (\min\{\overline{a}_{t-1},\, (C_i + \delta)\,v\} - \max\{\underline{a}_{t-1},\, (C_i - \delta)\,v\})\Big\}. \quad (33)$$

Following the same reasoning as in Section III, one can introduce the restricted optimization problems

$$D^{(j)} = \inf_{v \in W_j} \max_{i=0,\ldots,P} (\min\{\overline{a}, (C_i+\delta)v\} - \max\{\underline{a}, (C_i-\delta)v\}) \quad (34)$$

We can now state the following lemma, which is the counterpart of Lemma 3 in the presence of noisy measurements.

Lemma 4: Let $Q = \max_{i=m,\ldots,M-1} (C_{i+1} - C_i + 2\,\delta)$. Then, the solution of (34) is given by

$$D^{(j)} = \begin{cases} \max\{(C_m + \delta)\,v_L - \underline{a}\,,\, Q\,v_L\} & \text{if } v_c \leq v_L \\ \max\left\{\frac{\overline{a}\,C_m - \underline{a}\,C_M + \delta(\underline{a}+\overline{a})}{C_m + C_M},\ \frac{Q\,\overline{a}}{Q + C_M - \delta}\right\} & \text{if } v_L < v_c < v_R \\ \overline{a} - (C_M - \delta)\,v_R & \text{if } v_c \geq v_R \end{cases}$$

where

$$v_c = \min\left\{\frac{\underline{a} + \overline{a}}{C_m + C_M},\ \frac{\overline{a}}{Q + C_M - \delta}\right\}.$$

Moreover, the argument at which the solution is attained is

$$v^{(j)} = \begin{cases} v_L & \text{if } v_c \leq v_L \\ v_c & \text{if } v_L < v_c < v_R \\ v_R & \text{if } v_c \geq v_R. \end{cases}$$

Proof: Analogous to the proof of Lemma 3. ■

By Lemma 4, Theorem 1 applies as well to the noisy case with $D^{(j)}$ given by (34), allowing the design of the optimal input u^* at each time t.

A necessary and sufficient condition for the optimal input to actually reduce the size of the feasible set is given next.

Proposition 1: At a given time t, a reduction of the feasible set is possible if and only if

$$\frac{\overline{a}_{t-1}}{\underline{a}_{t-1}} > \frac{C_P + \delta}{C_P - \delta}. \quad (35)$$

Proof: Sufficiency follows by an analogous result for the binary case, see Theorem 14 in [1].

Now, let us prove that if $\frac{\overline{a}_{t-1}}{\underline{a}_{t-1}} \leq \frac{C_P+\delta}{C_P-\delta}$ no reduction of the feasible set size is achievable. Notice that, since by (28), $C_P \geq C_i > \delta$, $i = 1, \ldots, P$, it follows that $\frac{\overline{a}_{t-1}}{\underline{a}_{t-1}} \leq \frac{C_P+\delta}{C_P-\delta} \leq \frac{C_i+\delta}{C_i-\delta}$ for all $i = 1, \ldots, P$, or equivalently, by (31)

$$\overline{v}_i \leq \underline{v}_i \quad \text{, for } i = 1, \ldots, P. \quad (36)$$

We know that the optimal v must belong to $V^* = [\underline{v}_P,\, \overline{v}_1]$, and so, by (36), it must belong also to $\overline{V} = [\underline{v}_P,\, \underline{v}_1]$. Let us divide the set $\overline{V}$ into subsets $W_i = [\underline{v}_i,\, \underline{v}_{i-1}]$, $i = 2, \ldots, P$. One has $\bigcup_{i=2}^{P} W_i = \overline{V}$. Let us assume that the optimal v belongs to the set W_{i^*}, i.e., $v^* \in W_{i^*}$, $i^* = 2, \ldots, P$, and that the sensor output is $s = i^* - 1$ (notice that this occurs e.g., if $a = \underline{a}$ and $d = \delta$). This means

$$(C_{i^*-1} - \delta)\,v^* \leq a < (C_{i^*} + \delta)\,v^*.$$

Moreover, one has

$$(C_{i^*} + \delta)\,v^* \geq (C_{i^*} + \delta)\,\underline{v}_{i^*} = \frac{(C_{i^*} + \delta)}{(C_{i^*} - \delta)}\,\underline{a}_{t-1} \geq \overline{a}_{t-1} \quad (37)$$

and

$$(C_{i^*-1} - \delta)v^* \leq (C_{i^*-1} - \delta)\underline{v}_{i^*-1} = \frac{(C_{i^*-1} - \delta)}{(C_{i^*-1} - \delta)}\,\underline{a}_{t-1} = \underline{a}_{t-1}. \quad (38)$$

By (37) and (38), one has

$$\mathcal{F}_t = [\underline{a}_{t-1}, \overline{a}_{t-1}] \cap [(C_{i^*-1} - \delta)v^*,\, (C_{i^*} + \delta)v^*] = [\underline{a}_{t-1}, \overline{a}_{t-1}]$$

and so no improvement can be obtained. Necessity follows from arbitrariness of i^*. ■

V. An upper bound on the time complexity

In this section, an upper bound on the time complexity is provided, for the case when the thresholds are chosen according to (26) and measurements are noise free.

Let $\mathcal{F}_0 = [\underline{a}_0, \overline{a}_0]$ and let D_t be the diameter of the feasible set at time t. Let us assume that the sensor has $P > 1$ thresholds satisfying

$$C_i = \left[\frac{P+1-i}{P+1}\,\underline{a}_0 + \frac{i}{P+1}\,\overline{a}_0\right]\hat{u} \quad , \; i = 1, \ldots, P. \tag{39}$$

for some $\hat{u}$. We will devise an input selection strategy to derive an upper bound on the time complexity. Let us start by considering an optimal input which purses the maximum reduction at time 1. Hence, let us choose $u^*(1)$ according to Theorem 2, i.e., $u^*(1) = \hat{u}$ with $\hat{u}$ satisfying (39). Thus, one has $D_1 = \frac{D_0}{P+1} = \frac{\overline{a}_0 - \underline{a}_0}{P+1}$ independently of the output of the sensor $s(1)$. Let us assume that $s(1) = P-1$. In the following, it will turn out that such an output represents the worst-case one. By (14), the feasible set at time 1 is

$$\mathcal{F}_1 = [C_{P-1}\, v^*(1),\, C_P\, v^*(1)].$$

Let us now switch to the case $t > 1$. By (16), for $i = 1, \ldots, P$, one has

$$\underline{v}_i(1) = \frac{\underline{a}_1}{C_i} = \frac{C_{P-1}\, v^*(1)}{C_i} \;,\; \overline{v}_i(1) = \frac{\overline{a}_1}{C_i} = \frac{C_P\, v^*(1)}{C_i}. \tag{40}$$

Let us introduce the following lemma.

Lemma 5: Let $s(1) = P-1$. For $i = 2, \ldots, P$ one has $\overline{v}_i(1) < \underline{v}_{i-1}(1)$.

Proof: Since by construction the P thresholds (39) are equispaced, there exists q such that $C_i = C_{i-1} + q,\; i = 2, \ldots, P$.

By (40) one has

$$\begin{aligned}\frac{\underline{v}_{i-1}(1)}{\overline{v}_i(1)} &= \frac{C_{P-1}\, C_i}{C_{i-1}\, C_P} = \frac{C_{P-1}\, C_{i-1} + q\, C_{P-1}}{C_{P-1}\, C_{i-1} + q\, C_{i-1}} \\ &= 1 + \frac{q\,(C_{P-1} - C_{i-1})}{C_{P-1}\, C_{i-1} + q\, C_{i-1}} > 1. \end{aligned} \tag{41}$$

■

By Lemma 1, $v^*(2) \in [\underline{v}_P(1), \overline{v}_1(1)] \triangleq V^*(1)$, and since by construction $\underline{v}_i(1) < \overline{v}_i(1)$, for all $i = 1, \ldots, P$, by Lemma 5 one has

$$\underline{v}_P(1) < \overline{v}_P(1) < \underline{v}_{P-1}(1) < \ldots < \overline{v}_2(1) < \underline{v}_1(1) < \overline{v}_1(1).$$

Let us now define

$$\begin{aligned} W_1(1) &= [\underline{v}_P(1),\, \overline{v}_P(1)] \\ W_2(1) &= [\overline{v}_P(1),\, \underline{v}_{P-1}(1)] \\ &\vdots \\ W_{2P-2}(1) &= [\overline{v}_2(1),\, \underline{v}_1(1)] \\ W_{2P-1}(1) &= [\underline{v}_1(1),\, \overline{v}_1(1)]. \end{aligned} \tag{42}$$

It is straightforward to verify that $V^*(1) = \bigcup_{j=1}^{2P-1} W_j(1)$.

Let us now prove the next lemma for the case $t > 1$.

Lemma 6: Let $u^*(1) = \hat{u}$ satisfying (39), and assume $s(1) = P-1$. Then, the maximum reduction rate of the feasible set at any time $t > 1$ is $1/2$, i.e., $D_t = \frac{1}{2} D_{t-1}$.

Proof: Let us prove the result for $t = 2$. A similar reasoning can be repeated for any $t > 2$. By Theorem 1, it follows that the optimal input can be computed by solving problem (19) in each interval $W_j(1)$. So, let us evaluate $v^{(j)}$ for each subinterval $W_j(1)$.

First, let us consider the subintervals $W_j(1)$ where j is odd. By (42), it follows that all these subintervals are of the form $[\underline{v}_i(1),\, \overline{v}_i(1)],\; i = 1, \ldots, P$. By (20)-(21) one has $m = i$ and $M = i$, and so only one threshold (namely C_i) is active in $[\underline{v}_i(1),\, \overline{v}_i(1)]$. Hence, this case reduces to the binary case and one has $D^{(j)} = \frac{D_1}{2}$ (see Remark 1).

Let us now consider the subintervals $W_j(1)$ where j is even, i.e., of the form $[\overline{v}_i(1),\, \underline{v}_{i-1}(1)],\; i = 2, \ldots, P$. By (20)-(21) one has $m = i$ and $M = i-1$, and by Lemma 2 no reduction of the feasible set can be obtained.

Summarizing, by Theorem 1, the optimal input $u^*(2)$ provides a feasible set reduction rate of $1/2$. ■

Since, after time $t = 1$, the uncertainty reduction rate is $1/2$, the measurement $s(1) = P-1$ actually represents the worst-case sensor output realization at time $t = 1$.

In the following theorem an upper bound on the time complexity is provided.

Theorem 3: The time complexity for reducing the diameter of the feasible set from D_0 to $D_N > 0$ is

$$N(D_N) = 1 \quad , \text{ if } D_N \geq \frac{D_0}{P+1}$$

and

$$N(D_N) \leq \left\lceil 1 - \log_2(P+1) + \log_2\left(\frac{D_0}{D_N}\right)\right\rceil \text{, if } D_N < \frac{D_0}{P+1}.$$

Proof: If $D_N \geq \frac{D_0}{P+1}$, it follows immediately from Theorem 2. Let $D_N < \frac{D_0}{P+1}$. Following the above reasoning, if the thresholds are chosen according to (39), at $t = 1$ we obtain a reduction rate of $\frac{1}{P+1}$. By Lemma 6, for $t > 1$ the reduction rate is $\frac{1}{2}$. So,

$$D_1 = \frac{D_0}{P+1},\; D_2 = \frac{1}{P+1}\cdot\frac{D_0}{2}, \ldots,\; D_N = \frac{1}{P+1}\cdot\frac{D_0}{2^{N-1}}.$$

Hence, $2^{N-1} = \frac{1}{P+1} \cdot \frac{D_0}{D_N}$ and therefore

$$N - 1 = -\log_2(P+1) + \log_2\left(\frac{D_0}{D_N}\right) \tag{43}$$

which proves the theorem. ■

Remark 3: Since it is not guaranteed that the one step ahead optimal input $u^*(1)$ be the optimal input at a longer time horizon, (43) provides an upper bound on the time complexity. Moreover, notice that when $D_N < \frac{D_0}{P+1}$, the time complexity is reduced at least by $(\log_2(P+1) - 1)$ samples w.r.t. to the binary case.

VI. Numerical examples

Example 1: Let us consider a FIR of order 1, and let $\mathcal{F}_0 = [1, 21]$, $U = 10$, $\delta = 0$ (noise-free case). Let us assume the sensor has 4 thresholds $C_1 = 25$, $C_2 = 45$, $C_3 = 65$, $C_4 = 85$.

Notice that these thresholds satisfy (26) in Theorem 2 with $\hat{u} = 5$. In Figure 2, the size of the feasible set for different values of $a \in [1, 21]$ and different input lengths is reported for the optimal input u^* given by Theorem 1.

The diameter obtained by assuming only one threshold (binary case) is also reported in Figure 2. Notice that in this case the feasible set size is independent from the true parameter location. As expected, the information provided by the 4-thresholds sensor allows a faster reduction of uncertainty.

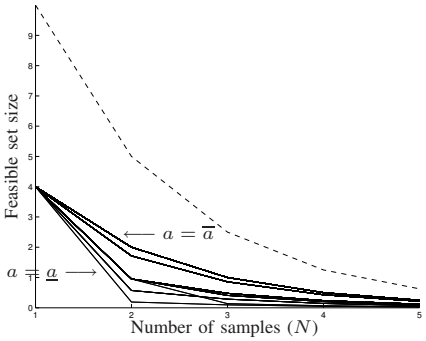

Fig. 2. Example 1: Feasible set size of a 4-thresholds sensor (solid) compared with a binary one (dashed) for different values of the true parameters a.

Example 2: Let us consider a FIR of order $n = 10$ and let us assume that the a priori information on the impulse response coefficients is $1 \leq \theta_i \leq M\rho^i,\ M = 100,\ \rho = 0.75,\ i = 1, \ldots, 10$. Moreover, let us assume $U = 50,\ \delta = 1$ and let the sensor have $P = 5$ equally-spaced thresholds, namely $C_j = 10\,j,\ j = 1, \ldots, 5$.

Let us suppose we want to independently excite each FIR coefficient 3 times. By applying the input strategy provided in [5], one needs $N = 3\,n(n+1)/2 = 165$ samples according to (9). Then, each FIR coefficient will be excited by the optimal input derived in Section IV.

Let us assume that the true parameter vector is $\theta^* = [58, 19, 31, 15, 6, 9, 10, 4, 3, 1]'$. Moreover, let us choose the noise signal $d(t)$ in order to maximize the size of the feasible set at time t. In Fig. 3 the feasible set bounds for each parameter after applying the designed input signal of length 165, are reported.

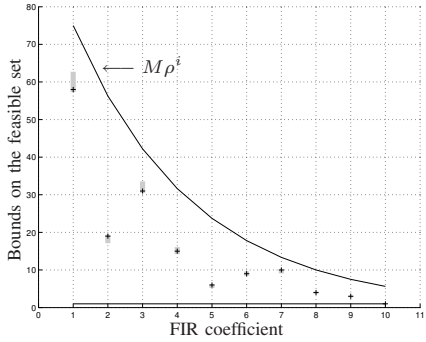

Fig. 3. Example 2: Feasible set bounds and true values of FIR parameters (crosses), for an input signal independently exciting each parameter 3 times. The system sensor has 5 thresholds.

TABLE I
NUMERICAL VALUES OF EXAMPLE 2 FOR EACH FIR COEFFICIENT

i	D_5	D_1	R_5	R_1	R_{min}
1	5.4111	11.3012	1.0944	1.2055	1.0408
2	2.2539	7.6403	1.1317	1.5163	1.0408
3	3.0437	6.1079	1.0998	1.2331	1.0408
4	1.8716	4.2528	1.1317	1.3431	1.0408
5	0.9241	3.0506	1.1769	1.7839	1.0408
6	1.1662	2.3474	1.1378	1.3242	1.0408
7	0.9407	1.8556	1.1033	1.2173	1.0408
8	0.5755	1.2756	1.1565	1.3935	1.0408
9	0.4316	0.9311	1.1565	1.3559	1.0408
10	0.1745	0.6498	1.1745	1.6498	1.0408

In Table I the final uncertainty of each parameter (diameter of $\mathcal{F}_N$) for the 5-thresholds sensor (D_5) is compared with that obtained with a binary sensor (D_1) with threshold $C = 50$. The last three columns denote the ratio $\frac{\overline{a}_N}{\underline{a}_N}$ for both the multi thresholds (R_5) and binary (R_1) sensors, and the minimum value achievable (R_{min}) as stated in (35). Notice that, since $C = C_P$, by (35) both cases have the same value of R_{min} (choosing a lower value of C would produce a higher value of R_{min} for the binary case). As expected, the table shows that a larger number of thresholds will lead to a faster reduction of uncertainty.

VII. CONCLUSIONS

In this paper, a solution to the one step ahead recursive optimal input design problem for identification of systems with quantized measurements has been proposed. The measurement noise is assumed unknown but bounded and a worst-case approach has been adopted. Improvements over the binary measurement case is illustrated through numerical examples and an analytical upper bound on the time complexity is devised. Further research should be directed to the optimal input design problem on a multi-step time horizon, which unlike the binary case, is still an unsolved problem. Likewise, the related issue of tight time complexity bounds both for the noise-free and the noisy measurements case needs a deeper investigation.

REFERENCES

[1] L. Y. Wang, J. F. Zhang, and G. G. Yin. System identification using binary sensors. *IEEE Transactions on Automatic Control*, 48(11):1892–1907, 2003.

[2] L. Ljung. *System Identification: Theory for the User, 2nd ed.* Prentice Hall, Upper Saddle River, NJ, 1999.

[3] L. Y. Wang, G. G. Yin, and J. F. Zhang. Joint identification of plant rational models and noise distribution functions using binary-valued observations. *Automatica*, 42(4):535–547, 2006.

[4] L. Y. Wang and G. G. Yin. Asymptotically efficient parameter estimation using quantized output observations. *Automatica*, 43(7):1178–1191, 2007.

[5] M. Casini, A. Garulli, and A. Vicino. Time complexity and input design in worst-case identification using binary sensors. In *Proc. of 46th IEEE Conf. on Decision and Control*, New Orleans, LA, December 2007.

[6] M. Milanese and A. Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. *Automatica*, 27(6):997–1009, 1991.

[7] M. Milanese and A. Vicino. Information-based complexity and nonparametric worst-case system identification. *Journal of Complexity*, 9:427–446, 1993.

[8] A. Garulli, A. Tesi, and A. Vicino, editors. *Robustness in Identification and Control*. Lecture Notes in Control and Information Sciences. Springer, London, 1999.

[9] D. N. C. Tse, M. A. Dahleh, and J. N. Tsitsiklis. Optimal asymptotic identification under bounded disturbances. *IEEE Transactions on Automatic Control*, 38(8):1176–1190, 1993.