Input Design in Worst-case System Identification using Binary Sensors

Marco Casini, Andrea Garulli, Antonio Vicino

***Abstract*—This paper addresses system identification using binary-valued sensors in a worst-case set-membership setting. The main contribution is the solution of the optimal input design problem for identification of scalar gains, which is instrumental to the construction of suboptimal input signals for identification of FIR models of arbitrary order. Two different cost functions are considered for input design: the maximum parametric identification error and the relative uncertainty reduction with respect to the minimum achievable error. It is shown that in the latter case, the solution enjoys the property of being independent of the length of the identification experiment and as such it can be implemented as an optimal recursive procedure over a time interval of arbitrary length.**

I. Introduction

Recent years have witnessed a considerable amount of work on identification of systems with binary valued measurements. Binary sensors are devices characterized by a threshold according to which the output is digitized. The basic reason for the widespread diffusion of binary sensors is mainly related to their relative low cost and to the fact that, in spite of the simplicity of the control laws adopted, monitoring and control of industrial plants often calls for measurement of a large number of variables. Binary sensors are also present in several communication systems, like ATM networks, and automotive applications.

In their pioneering work, Wang, Zhang and Yin [1] introduced a general framework to deal with system identification with binary data. The main difficulty in this problem is to tackle a discontinuous nonlinearity present in the sensor, which reduces drastically the information conveyed by measurements. While in [1] several important results have been obtained for FIR models either in a stochastic setting or in a worst-case setting, in [2] further results on output error models have been pursued in the stochastic framework. Consistency properties of weighted least-squares algorithms for parameter estimation based on binary data have been studied in [3]. The stochastic approach introduced in [1] has been applied to cope with several problems involving binary valued observations, including identification of Wiener or Hammerstein systems, and of systems with time-varying parameters (see [4] for an extensive overview). On the other hand, when the identification problem is cast in a worst-case deterministic setting, key issues such as optimal input design are still open problems. Preliminary results in this respect have been presented in [5]. An attempt to introduce a joint framework taking into account the features of both the stochastic and the deterministic setting has been recently made in [6].

The aim of the present paper is to tackle the input design problem in a worst-case setting, based on the set-membership paradigm of uncertainty representation (see e.g. [7], [8]). Specifically, for the case of identification of a scalar gain, an input is devised which minimizes the worst-case identification error. This result can be fruitfully used to construct efficient suboptimal procedures for identification of FIR models of arbitrary order.

Since the optimal input sequence for identification of a scalar gain derived by minimizing the worst-case error depends on the length of the sequence itself, the resulting excitation experiment and related estimation algorithm turn out to be a batch procedure, where the input sequence length is a design parameter, which cannot be tuned during the execution of the experiment. Actually, in several contexts, one wants to keep the option of increasing the input length, without running a new identification experiment from the beginning. In these cases, it is fairly natural to look for an optimal recursive algorithm, where the parameter estimate updating can be performed for an arbitrary number of steps, preserving optimality of the solution achieved. In this perspective, a further contribution of the paper consists in introducing a new worst-case cost function, such that the resulting optimal input selection strategy is recursive. Roughly speaking, this cost function is a normalized error representing the worst-case relative reduction of the gain uncertainty estimate with respect to the asymptotic error, i.e., the minimum achievable uncertainty as the input length tends to infinity. Beyond the recursive nature of the optimal input, numerical experiments show that this algorithm is quite effective for FIR model identification, providing accuracy results which outperform the input design algorithm based on the actual identification error, in several test cases.

The paper is organized as follows. Section II introduces notation and problem formulation. Section III addresses the problem of optimal input design with respect to the actual worst-case identification error, while Section IV provides the alternative optimal input design strategy based on the worst-case normalized identification error. A numerical example highlighting the comparison between the two input design approaches is reported in Section V, while concluding remarks are reported in Section VI.

II. Problem formulation

Let $\mathbb{R}^N$ denote the N-dimensional Euclidean space. A sequence of real numbers $\{x(t), t = 1, \ldots, N\}$ 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 denote by $\log(x)$ the base 2 logarithm of x, and by $\lceil x \rceil$ the minimum integer greater or equal to x.

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

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

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

Observations at the system output are taken by a binary sensor with a known threshold C, such that

$$s(t) = \begin{cases} 1 & \text{if } y(t) \leq C \\ 0 & \text{if } y(t) > C. \end{cases} \tag{2}$$

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

Let $\mathcal{F}_0 = B_p(c, \varepsilon_0)$ represent the prior information available on the parameter vector. Let us now denote by $u, s \in \mathbb{R}^N$ the input signal $\{u(t), \ t = 1, \ldots, N\}$ and the sequence of binary measurements $\{s(t), \ t = 1, \ldots, N\}$, respectively. For a given input-output realization $\{u, s\}$ of length N, the feasible parameter set is defined as:

$$\begin{aligned} \mathcal{F}_N = \{\theta \in \mathcal{F}_0 : \ & \phi^T(t)\, \theta \leq C + \delta \ \text{ if } s(t) = 1; \\ & \phi^T(t)\, \theta > C - \delta \ \text{ if } s(t) = 0; \ t = 1, \ldots, N\}. \end{aligned} \tag{3}$$

Authors are with the Dipartimento di Ingegneria dell'Informazione and the Centro per lo Studio dei Sistemi Complessi, Università di Siena, 53100 Siena, Italy (email: casini@ing.unisi.it; garulli@ing.unisi.it; vicino@ing.unisi.it).

The worst-case local identification error is defined as

$$e_p(N,u,s) = \inf_{c\in\mathbb{R}^n} \sup_{\theta\in\mathcal{F}_N} \|\theta - c\|_p \ . \tag{4}$$

For a fixed input sequence u, let us define the global worst-case error with respect to the disturbance realization as $e_p(N,u) = \sup_{s\in\mathcal{S}_0} e_p(N,u,s)$, where $\mathcal{S}_0 = \{s\colon \mathcal{F}_N \neq \emptyset\}$.

Since we are interested in optimal input design, the aim is to compute the minimum worst-case identification error [9]

$$e_p(N) = \inf_{u:\|u\|_\infty\leq U} e_p(N,u) \ . \tag{5}$$

Let us define

$$\gamma \triangleq \min_{i=1,\dots,n} \{|\theta_i| :\ \theta\in\mathcal{F}_0\}. \tag{6}$$

The following assumptions are enforced throughout the paper.

Assumption 1: i)

1) $\gamma > C/U$.
2) The signs of the FIR parameters θ_i are known.

The first assumption is necessary to guarantee that the initial uncertainty on each parameter can be reduced by using the information from the binary sensor (see [1, Prop. 1]). The signs of the parameters θ_i can be easily detected by performing a suitable preliminary identification experiment involving at most $2n$ samples.

In [5], [10], a strategy has been proposed for tackling the input design problem (5). The idea is to select the input samples which individually excite the parameter θ_i in order to achieve the maximum uncertainty reduction for that parameter. Optimal input design for each scalar parameter θ_i will result in an effective suboptimal solution for FIR models of arbitrary order. For this reason, the exact solution of the optimal input design problem for scalar gains is provided in the next section.

III. Input design for identification of gains

In this section, we address the optimal input design problem for the special case of FIR systems of order $n = 1$. The following setting will be considered

$$y(t) = a\,u(t) + d(t) \ , \quad s(t) = \begin{cases} 1 & \text{if } y(t)\leq C \\ 0 & \text{if } y(t) > C, \end{cases}$$

with $|u(t)|\leq U$, $|d(t)|\leq\delta$ and $a\in\mathcal{F}_0 = [\underline{a}_0, \overline{a}_0]$.

Let us denote by $\mathcal{F}_t = [\underline{a}_t, \overline{a}_t]$ the feasible parameter set at time t. The aim is to choose the input signal $u(t)$ at time t as a function of the available information up to time $t-1$, i.e. $u(t) = \eta(\mathcal{F}_{t-1}; t)$, $t = 1,2,\dots,N$. In order to stress that the set-theoretic information is exploited in the input choice, the optimal input design problem (5) is rewritten as

$$u^* = \arg \inf_{\substack{u:\ \|u\|_\infty\leq U \\ u(t)=\eta(\mathcal{F}_{t-1};t)}} e_p(N,u) \ . \tag{7}$$

According to Assumption 1, one has $\underline{a}_0 > C/U$. Moreover, let $C > \delta > 0$ and define $\beta \triangleq \delta/C$. The following proposition clarifies under which conditions, reduction of the feasible parameter set is possible [1, Thm. 14].

Proposition 1: The following statements hold:

a) At a given time t, uncertainty reduction is possible if and only if $\frac{\overline{a}_{t-1}}{\underline{a}_{t-1}} > \frac{1+\beta}{1-\beta}$.
b) If $\frac{\overline{a}_0}{\underline{a}_0} > \frac{1+\beta}{1-\beta}$, then there exists an input sequence such that $\frac{\overline{a}_\infty}{\underline{a}_\infty} \triangleq \lim_{N\to\infty} \frac{\overline{a}_N}{\underline{a}_N} = \frac{1+\beta}{1-\beta}$.
c) If $\frac{\overline{a}_{t-1}}{\underline{a}_{t-1}} > \frac{1+\beta}{1-\beta}$, then there does not exist any input $u(t)$ such that $\frac{\overline{a}_t}{\underline{a}_t} \leq \frac{1+\beta}{1-\beta}$.

By Proposition 1, it follows that the condition

$$\frac{\overline{a}_0}{\underline{a}_0} > \frac{1+\beta}{1-\beta} \tag{8}$$

is necessary in order to reduce the radius of the initial feasible parameter set, and it is also sufficient to guarantee that uncertainty reduction is possible at any time $t > 0$. Hereafter, the assumption (8) will be enforced. The following result provides the solution to problem (7).

Theorem 1: Let (8) hold and N be the length of the input signal to be applied. Then, the optimal input solving problem (7) is given by $u^*(t) = C/\widetilde{a}_{t|N},\ t = 1,2,\dots,N$ where

$$\widetilde{a}_{t|N} = \frac{\overline{a}_{t-1}\,(1+\beta)^{(2^{N-t}-1)} + \underline{a}_{t-1}\,(1-\beta)^{(2^{N-t}-1)}}{(1+\beta)^{(2^{N-t})} + (1-\beta)^{(2^{N-t})}}, \tag{9}$$

with estimate updating laws:

$$\begin{cases} \underline{a}_t = \underline{a}_{t-1}\ ,\ \overline{a}_t = \widetilde{a}_{t|N}\,(1+\beta) & \text{if } s(t) = 1 \\ \underline{a}_t = \widetilde{a}_{t|N}\,(1-\beta)\ ,\ \overline{a}_t = \overline{a}_{t-1} & \text{if } s(t) = 0. \end{cases}$$

Moreover, the radius of the feasible parameter set $\mathcal{F}_N$ is

$$\varepsilon_N = \beta\,\frac{\overline{a}_0\,(1+\beta)^{(2^N-1)} - \underline{a}_0\,(1-\beta)^{(2^N-1)}}{(1+\beta)^{(2^N)} - (1-\beta)^{(2^N)}}. \tag{10}$$

Proof: Let us define

$$P_N \triangleq \prod_{j=0}^{N-1} \left[(1+\beta)^{(2^j)} + (1-\beta)^{(2^j)}\right].$$

First, let us prove that

$$P_N = \frac{(1+\beta)^{(2^N)} - (1-\beta)^{(2^N)}}{2\,\beta}. \tag{11}$$

For $N = 1$, (11) holds trivially. Moreover, if (11) holds for a generic N, then it holds also for $N+1$, because

$$\begin{aligned} P_{N+1} &= P_N \cdot \left[(1+\beta)^{(2^N)} + (1-\beta)^{(2^N)}\right] \\ &= \frac{\left[(1+\beta)^{(2^N)} - (1-\beta)^{(2^N)}\right]\cdot\left[(1+\beta)^{(2^N)} + (1-\beta)^{(2^N)}\right]}{2\,\beta} \\ &= \frac{(1+\beta)^{(2^{N+1})} - (1-\beta)^{(2^{N+1})}}{2\,\beta}. \end{aligned}$$

Let us now demonstrate (9) and (10) by induction. First, let us show that $\widetilde{a}_{1|1} = \frac{\overline{a}_0+\underline{a}_0}{2}$. In fact, by applying $u(1) = C/\widetilde{a}_{1|1}$, one obtains two possible feasible sets at time 1, namely $\mathcal{F}_1^{(0)} = [\widetilde{a}_{1|1}(1-\beta),\ \overline{a}_0]$ or $\mathcal{F}_1^{(1)} = [\underline{a}_0,\ \widetilde{a}_{1|1}(1+\beta)]$, depending on the value of $s(1)$ being respectively 0 or 1. Notice that (8) guarantees the existence of an $\widetilde{a}_{1|1}$ such that both $\mathcal{F}_1^{(0)}$ and $\mathcal{F}_1^{(1)}$ are strictly contained in $\mathcal{F}_0$. The maximum worst-case uncertainty reduction is thus obtained by choosing $\widetilde{a}_{1|1}$ so that the size of both intervals is the same, thus leading to $\widetilde{a}_{1|1} = \frac{\overline{a}_0+\underline{a}_0}{2}$. By substituting this value of $\widetilde{a}_{1|1}$ in $\mathcal{F}_1^{(0)}$ or $\mathcal{F}_1^{(1)}$, one gets $\varepsilon_1 = \frac{\overline{a}_0(1+\beta)-\underline{a}_0(1-\beta)}{4}$ which corresponds to (10). It is worth remarking that the resulting input $u(1) = \frac{2\,C}{\underline{a}_0+\overline{a}_0}$ satisfies $|u(1)| < U$ thanks to Assumption 1.

Now, let us assume that (9) and (10) hold for a given N. We want to show that they hold also for an input of length $N+1$. By (11), (10) can be rewritten as

$$\varepsilon_N = \frac{\overline{a}_0\,(1+\beta)^{(2^N-1)} - \underline{a}_0\,(1-\beta)^{(2^N-1)}}{2P_N}. \tag{12}$$

Observe that the last N samples of the optimal input of length $N+1$ can be chosen by applying to the feasible set at time 1, i.e. $[\underline{a}_1, \overline{a}_1]$, the same input selection strategy adopted for computing the optimal input of length N for the initial feasible set $[\underline{a}_0, \overline{a}_0]$. In other words,

the second optimal input sample for an input of length $N+1$ depends on $[\underline{a}_1, \overline{a}_1]$ in the same way as the first optimal sample of an input of length N depends on $[\underline{a}_0, \overline{a}_0]$, and so on. Indeed, if in (9) the time indices of the feasible set bounds $\underline{a}$ and $\overline{a}$ are suitably rearranged, one has $\widetilde{a}_{t+1|N+1} = \widetilde{a}_{t|N}$ for all $t = 1, \ldots, N$. Moreover, since (10) holds for an input of length N, the radius of the final feasible set $\mathcal{F}_{N+1}$ for the optimal input sequence of length $N+1$ can be expressed according to (12) as

$$\varepsilon_{N+1} = \frac{\overline{a}_1 (1+\beta)^{(2^N-1)} - \underline{a}_1 (1-\beta)^{(2^N-1)}}{2P_N}. \tag{13}$$

We have now to select the first input sample $u^*(1) = C/\widetilde{a}_{1|N+1}$. The choice of $\widetilde{a}_{1|N+1}$ leads to two possibile feasible sets at time 1: $\mathcal{F}_1^{(0)} = [\widetilde{a}_{1|N+1}(1-\beta),\ \overline{a}_0]$ if $s(1) = 0$, or $\mathcal{F}_1^{(1)} = [\underline{a}_0,\ \widetilde{a}_{1|N+1}(1+\beta)]$ if $s(1) = 1$. From (13), the two corresponding final radii turn out to be

$$\varepsilon_{N+1}^{(0)} = \frac{\overline{a}_0(1+\beta)^{(2^N-1)} - \widetilde{a}_{1|N+1}(1-\beta)^{(2^N)}}{2P_N}$$

$$\varepsilon_{N+1}^{(1)} = \frac{\widetilde{a}_{1|N+1}(1+\beta)^{(2^N)} - \underline{a}_0(1-\beta)^{(2^N-1)}}{2P_N}.$$

The optimal choice of $\widetilde{a}_{1|N+1}$ is such that $\varepsilon_{N+1}^{(0)} = \varepsilon_{N+1}^{(1)}$, i.e.,

$$\begin{aligned}\overline{a}_0(1+\beta)^{(2^N-1)} - \widetilde{a}_{1|N+1}(1-\beta)^{(2^N)} \\ = \widetilde{a}_{1|N+1}(1+\beta)^{(2^N)} - \underline{a}_0(1-\beta)^{(2^N-1)}\end{aligned}$$

which leads to

$$\widetilde{a}_{1|N+1} = \frac{\overline{a}_0(1+\beta)^{(2^N-1)} + \underline{a}_0(1-\beta)^{(2^N-1)}}{(1+\beta)^{(2^N)} + (1-\beta)^{(2^N)}}$$

that is in accordance with (9).
Finally, to compute ε_{N+1} one has to substitute $\widetilde{a}_{1|N+1}$ in $\varepsilon_{N+1}^{(0)}$ (or in $\varepsilon_{N+1}^{(1)}$), thus obtaining

$$\begin{aligned}\varepsilon_{N+1} &= \frac{1}{2P_N}\Bigg\{\overline{a}_0(1+\beta)^{(2^N-1)} \\ &\quad - \frac{\overline{a}_0(1+\beta)^{(2^N-1)} + \underline{a}_0(1-\beta)^{(2^N-1)}}{(1+\beta)^{(2^N)} + (1-\beta)^{(2^N)}}(1-\beta)^{(2^N)}\Bigg\} \\ &= \frac{\overline{a}_0(1+\beta)^{(2^{N+1}-1)} - \underline{a}_0(1-\beta)^{(2^{N+1}-1)}}{2P_N\left[(1+\beta)^{(2^N)} + (1-\beta)^{(2^N)}\right]} \\ &= \frac{\overline{a}_0(1+\beta)^{(2^{N+1}-1)} - \underline{a}_0(1-\beta)^{(2^{N+1}-1)}}{2P_{N+1}} \\ &= \beta\,\frac{\overline{a}_0(1+\beta)^{(2^{N+1}-1)} - \underline{a}_0(1-\beta)^{(2^{N+1}-1)}}{(1+\beta)^{(2^{N+1})} - (1-\beta)^{(2^{N+1})}} \end{aligned} \tag{14}$$

which concludes the proof. ■

Remark 1: In [1], it is proposed as optimal input the signal

$$u(t) = \frac{2C}{\overline{a}_{t-1} + \underline{a}_{t-1}}. \tag{15}$$

Such signal actually minimizes the size of the worst-case $\mathcal{F}_t$ according to the available information up to time $t-1$, but it does not provide the optimal input sequence of length N minimizing the size of $\mathcal{F}_N$ (and hence $e_p(N,u)$) in the worst-case sense (7). In other words, if one applies the input signal (15), it is always possible to find a FIR parameter $a \in \mathcal{F}_0$ and a disturbance sequence d such that the radius of the resulting $\mathcal{F}_N$ is larger than ε_N in (10).

IV. Input design for relative uncertainty reduction

So far, the considered input design problem has addressed the minimization of the actual parameter uncertainty, i.e. the size of the feasible set. The exact solution for identification of gains, provided by Theorem 1, has shown that the optimal input sequence depends on the length N of the sequence itself, which is clearly an undesirable feature. In fact, one may want to adapt on-line the length of the identification experiment without loosing optimality. Moreover, it would be more sensible to measure the performance of the identification procedure with respect to the minimum error achievable in the considered setting, i.e. to optimize the relative uncertainty reduction instead of the absolute one, with the aim of designing a *recursive* optimal input design strategy. This would also give a rationale for devising a stopping criterion for the identification experiment.

Motivated by the above reasons, in this section a new worst-case error cost function is introduced and the input signal minimizing such a functional is derived for the identification of a scalar gain. The same setting and assumptions as in Section III are adopted.

Let $a \in \mathcal{F}$ be the true parameter and let $\mathcal{F} = [\underline{a}, \overline{a}]$ at a given time instant. Let us denote by $D_{min}(\underline{a}, \overline{a}, a)$ the minimum size of the worst-case feasible set, obtainable by applying an infinite sequence of inputs. In other words, D_{min} represents the irreducible size of the worst-case feasible set. The following result holds.

Lemma 1: Let $\mathcal{F} = [\underline{a}, \overline{a}]$, and let $a \in \mathcal{F}$ be the true parameter. One has:

$$D_{min}(\underline{a}, \overline{a}, a) = \begin{cases} \frac{2\,\beta}{1-\beta}\,a & \text{, if } \underline{a} \le a \le \frac{1-\beta}{1+\beta}\,\overline{a} \\ \frac{2\,\beta}{1+\beta}\,\overline{a} & \text{, if } \frac{1-\beta}{1+\beta}\,\overline{a} < a \le \overline{a} \end{cases}.$$

Proof: Let us define as $[\underline{a}_\infty, \overline{a}_\infty]$ the feasible set obtained after applying an appropriate infinite sequence of inputs. By the definition of D_{min} and Proposition 1, one can write

$$\begin{aligned} & D_{min}(\underline{a}, \overline{a}, a) = \sup_{\underline{a}_\infty, \overline{a}_\infty} (\overline{a}_\infty - \underline{a}_\infty) \\ & s.t.: \\ & \quad \frac{\overline{a}_\infty}{\underline{a}_\infty} \ge \frac{1+\beta}{1-\beta}\ ; \quad \underline{a}_\infty \le a \le \overline{a}_\infty \\ & \quad \underline{a}_\infty \ge \underline{a}\ ; \quad \overline{a}_\infty \le \overline{a}. \end{aligned} \tag{16}$$

By exploiting the first constraint and rearranging the others, problem (16) can be rewritten as

$$\begin{aligned} & D_{min}(\underline{a}, \overline{a}, a) = \sup_{\underline{a}_\infty} \frac{2\,\beta}{1-\beta}\,\underline{a}_\infty \\ & s.t.:\ \underline{a}_\infty \le a\ ;\ \underline{a}_\infty \le \tfrac{1-\beta}{1+\beta}\,\overline{a} \end{aligned} \tag{17}$$

whose solution is achieved at $\underline{a}_\infty = \min\left\{a,\ \frac{1-\beta}{1+\beta}\,\overline{a}\right\}$. The result follows by substitution. ■

Let us introduce the following cost function:

$$J(\underline{a}, \overline{a}, a) = \frac{\overline{a} - \underline{a}}{D_{min}(\underline{a}, \overline{a}, a)}.$$

At a given time instant t, $J(\underline{a}_t, \overline{a}_t, a)$ represents the ratio between the current feasible set size and the minimum size of the worst-case feasible set, achievable by applying an infinite input sequence starting at time $t+1$. Intuition suggests that minimizing J over a time horizon of length N corresponds to minimizing the "residual uncertainty" after the first N time instants, or equivalently, to maximizing the performance of the first N input samples. Unfortunately, the function J depends on the true value of a which is unknown. According to the worst-case approach taken throughout the paper, the objective will

be to minimize the worst-case value of J with respect to all feasible values of a. By using Lemma 1, it turns out that

$$\begin{aligned} J_{max}(\underline{a},\overline{a}) &\triangleq \sup_{a\in[\underline{a},\overline{a}]} J(\underline{a},\overline{a},a) = \frac{\overline{a}-\underline{a}}{\inf_{a\in[\underline{a},\overline{a}]} D_{min}(\underline{a},\overline{a},a)} \\ &= \frac{\overline{a}-\underline{a}}{D_{min}(\underline{a},\overline{a},\underline{a})} = \frac{\overline{a}-\underline{a}}{\frac{2\,\beta}{1-\beta}\,\underline{a}} \\ &= \frac{(1-\beta)(\overline{a}-\underline{a})}{2\,\beta\,\underline{a}} = \frac{1-\beta}{2\,\beta}\left(\frac{\overline{a}}{\underline{a}}-1\right). \end{aligned} \tag{18}$$

Since by hypothesis $\frac{\overline{a}}{\underline{a}} > \frac{1+\beta}{1-\beta}$, one has $J_{max}(\underline{a},\overline{a}) > 1$, as expected. Moreover, by Proposition 1, there exists an infinite input sequence such that $\lim_{t\to\infty}\frac{\overline{a}_t}{\underline{a}_t} = \frac{1+\beta}{1-\beta}$, and hence $\lim_{t\to\infty} J_{max}(\underline{a}_t,\overline{a}_t) = 1$.

Remark 2: To clarify the meaning of $J_{max}(\underline{a},\overline{a})$, notice that it allows one to quantify the worst-case relative reduction of the feasible set w.r.t. the minimum achievable uncertainty. For example, $J_{max} = 1.2$ means that the size of the feasible set is 20% greater than that of the minimum feasible set achievable by any input sequence of infinite length, in worst-case sense. Hence, J_{max} can be used to devise a rationale for stopping the identification experiment. For instance, one can decide to stop the experiment whenever $J_{max} < 1.01$, because it will not be possible to further reduce the feasible set by more than 1%.

In the following, the optimal input sequence minimizing J_{max} is derived, for the case of a scalar gain. In particular, the aim is to solve the input design problem

$$u^* = \arg \inf_{\substack{u:\ \|u\|_\infty \le U \\ u(t)=\eta(\mathcal{F}_{t-1};t)}} J_{max}(\underline{a}_N,\overline{a}_N) \tag{19}$$

Theorem 2: Let N be the length of the input signal to be applied. Then, the optimal input sequence solving problem (19) is given by $u^*(t) = C/\widetilde{a}_{t|N},\ t = 1,2,\ldots,N$ where

$$\widetilde{a}_{t|N} = \sqrt{\frac{\underline{a}_{t-1}\,\overline{a}_{t-1}}{1-\beta^2}} \tag{20}$$

with estimate updating laws:

$$\begin{cases} \underline{a}_t = \underline{a}_{t-1}\,,\ \overline{a}_t = \widetilde{a}_{t|N}\,(1+\beta) = \sqrt{\underline{a}_{t-1}\,\overline{a}_{t-1}\,\frac{1+\beta}{1-\beta}}\ , \text{ if } s(t)=1 \\ \underline{a}_t = \widetilde{a}_{t|N}\,(1-\beta) = \sqrt{\underline{a}_{t-1}\,\overline{a}_{t-1}\,\frac{1-\beta}{1+\beta}}\,,\ \overline{a}_t = \overline{a}_{t-1}\ , \text{ if } s(t)=0. \end{cases} \tag{21}$$

Moreover, the value of J_{max} after N input samples is

$$J_{max|N} = \frac{1-\beta}{2\,\beta}\left[\frac{1+\beta}{1-\beta}\ \sqrt[2^N]{\frac{\overline{a}_0}{\underline{a}_0}\frac{1-\beta}{1+\beta}} - 1\right]. \tag{22}$$

Proof: Let us prove (20) and (22) by induction, in a similar way as in the proof of Theorem 1. Let $N = 1$. By (18) one has $J_{max|0} = \frac{1-\beta}{2\,\beta}\left(\frac{\overline{a}_0}{\underline{a}_0}-1\right)$. By applying $u(1) = C/\widetilde{a}_{1|1}$ one obtains two possible feasible sets at time 1, namely $\mathcal{F}_1^{(0)} = [\widetilde{a}_{1|1}(1-\beta),\ \overline{a}_0]$ or $\mathcal{F}_1^{(1)} = [\underline{a}_0,\ \widetilde{a}_{1|1}(1+\beta)]$, depending on the value of $s(1)$ being respectively 0 or 1. The optimal value of $\widetilde{a}_{1|1}$ is such that the value of $J_{max|1}$ for both intervals is the same, i.e., $J^{(0)}_{max|1} = J^{(1)}_{max|1}$, thus obtaining

$$\frac{1-\beta}{2\,\beta}\left(\frac{\overline{a}_0}{\widetilde{a}_{1|1}(1-\beta)}-1\right) = \frac{1-\beta}{2\,\beta}\left(\frac{\widetilde{a}_{1|1}(1+\beta)}{\underline{a}_0}-1\right)$$

which leads to $\widetilde{a}^2_{1|1} = \frac{\overline{a}_0\,\underline{a}_0}{1-\beta^2}$. Since $0 \le \underline{a}_0 \le \widetilde{a}_{1|1} \le \overline{a}_0$ and $\beta < 1$, only the positive solution is feasible, and so $\widetilde{a}_{1|1} = \sqrt{\frac{\underline{a}_0\,\overline{a}_0}{1-\beta^2}}$. By substitution, one obtains $J_{max|1}$ as in (22).

Now, let us assume that (20) and (22) hold for a given N. We want to show that they hold also for $N+1$. By following the same reasoning as in the proof of Theorem 1, we can state that

$$J_{max|N+1} = \frac{1-\beta}{2\,\beta}\left[\frac{1+\beta}{1-\beta}\ \sqrt[2^N]{\frac{\overline{a}_1}{\underline{a}_1}\frac{1-\beta}{1+\beta}} - 1\right]. \tag{23}$$

We have now to select the first input sample $u^*(1) = C/\widetilde{a}_{1|N+1}$. The choice of $\widetilde{a}_{1|N+1}$ leads to two possibile feasible sets at time 1: $\mathcal{F}_1^{(0)} = [\widetilde{a}_{1|N+1}(1-\beta),\ \overline{a}_0]$ if $s(1) = 0$, or $\mathcal{F}_1^{(1)} = [\underline{a}_0,\ \widetilde{a}_{1|N+1}(1+\beta)]$ if $s(1) = 1$. From (23), the two corresponding final values of J_{max} turn out to be

$$J^{(0)}_{max|N+1} = \frac{1-\beta}{2\,\beta}\left[\frac{1+\beta}{1-\beta}\ \sqrt[2^N]{\frac{\overline{a}_0}{\widetilde{a}_{1|N+1}}\ \frac{1}{1+\beta}} - 1\right]$$

$$J^{(1)}_{max|N+1} = \frac{1-\beta}{2\,\beta}\left[\frac{1+\beta}{1-\beta}\ \sqrt[2^N]{\frac{\widetilde{a}_{1|N+1}}{\underline{a}_0}(1-\beta)} - 1\right].$$

The optimal choice of $\widetilde{a}_{1|N+1}$ is such that $J^{(0)}_{max|N+1} = J^{(1)}_{max|N+1}$, i.e., $\frac{\overline{a}_0}{\widetilde{a}_{1|N+1}}\frac{1}{1+\beta} = \frac{\widetilde{a}_{1|N+1}}{\underline{a}_0}(1-\beta)$ which leads to $\widetilde{a}_{1|N+1} = \sqrt{\frac{\underline{a}_0\,\overline{a}_0}{1-\beta^2}}$. Finally, to compute $J_{max|N+1}$ one has to substitute $\widetilde{a}_{1|N+1}$ in $J^{(0)}_{max|N+1}$ (or in $J^{(1)}_{max|N+1}$), thus obtaining, after some algebra

$$J_{max|N+1} = \frac{1-\beta}{2\,\beta}\left[\frac{1+\beta}{1-\beta}\ \sqrt[2^{N+1}]{\frac{\overline{a}_0}{\underline{a}_0}\frac{1-\beta}{1+\beta}} - 1\right]$$

which concludes the proof. ∎

Remark 3: A key feature of Theorem 2 is that the optimal input sequence solving problem (19) does not depend on the length N of the identification experiment. This means that the length of the experiment can be adapted on-line, without loosing optimality with respect to the relative uncertainty reduction J_{max}.

In the following, we will denote by "Optimal Error Procedure" (OEP) and "Optimal Relative Error Procedure" (OREP) the input design procedures described in Theorems 1 and 2, respectively. With the aim of providing a quantitative comparison of the two approaches, the performance of the OREP will be evaluated in terms of the radius of the resulting feasible set. Let us now state a lemma which is instrumental to this purpose.

Lemma 2: Let N be the length of the input signal. Then, the maximum achievable value of $\underline{a}_N$ when applying the OREP is

$$\breve{\underline{a}}_N \triangleq \sup_s \underline{a}_N = \frac{1-\beta}{1+\beta}\,\overline{a}_0\ \sqrt[2^N]{\frac{\underline{a}_0}{\overline{a}_0}\frac{1+\beta}{1-\beta}}. \tag{24}$$

Proof: Since by Proposition 1 one has $\frac{\overline{a}_t}{\underline{a}_t} > \frac{1+\beta}{1-\beta}\ \forall t$, by (21) it follows that the maximum value for $\underline{a}_t$ is always obtained when $s(t) = 0$, being $\sqrt{\underline{a}_{t-1}\,\overline{a}_{t-1}\,\frac{1-\beta}{1+\beta}} > \underline{a}_{t-1}$. This condition holds, for instance, when $a = \overline{a}_0$. By defining $\mu \triangleq \overline{a}_0\frac{1-\beta}{1+\beta}$, and observing that if $s(t) = 0\ \forall t$, then $\overline{a}_t = \overline{a}_0\ \forall t$, one has $\breve{\underline{a}}_1 = \sqrt{\underline{a}_0\,\mu} = \mu\,\sqrt{\frac{\underline{a}_0}{\mu}}$; $\breve{\underline{a}}_2 = \sqrt{\underline{a}_1\,\mu} = \mu\,\sqrt[4]{\frac{\underline{a}_0}{\mu}}$; $\ldots$; $\breve{\underline{a}}_N = \sqrt{\underline{a}_{N-1}\,\mu} = \mu\ \sqrt[2^N]{\frac{\underline{a}_0}{\mu}}$. ∎

Theorem 3: Let N be the length of the OREP input signal. Then, the worst-case radius of the feasible parameter set $\mathcal{F}_N$ is

$$\epsilon_N = \frac{\overline{a}_0}{2}\left[1 - \frac{1-\beta}{1+\beta}\ \sqrt[2^N]{\frac{\underline{a}_0}{\overline{a}_0}\frac{1+\beta}{1-\beta}}\ \right] \tag{25}$$

Proof: By (18), one has $J_{max|N} = \frac{\overline{a}_N - \underline{a}_N}{\frac{2\beta}{1-\beta}\,\underline{a}_N}$ which can be rewritten as

$$\frac{\overline{a}_N - \underline{a}_N}{2} = J_{max|N}\cdot\frac{\beta}{1-\beta}\,\underline{a}_N. \tag{26}$$

Since the left-hand side of (26) denotes the radius of $\mathcal{F}_N$, and $J_{max|N}$ is given in (22), by Lemma 2 one has $\epsilon_N = \sup_{\underline{a}_N} J_{max|N}\cdot$

$\frac{\beta}{1-\beta}\,\underline{a}_N = J_{max|N} \cdot \frac{\beta}{1-\beta}\,\breve{\underline{a}}_N$. The result follows by substitution of $J_{max|N}$ and $\breve{\underline{a}}_N$ from (22) and (24), respectively. ■

Since ε_N reported in (10) denotes the radius of $\mathcal{F}_N$ when the OEP is applied, by construction one has $\varepsilon_N \leq \epsilon_N$. Figure 1 shows the comparison between the two radii ε_N and ϵ_N, for the setting $\beta = 0.1$, $\underline{a}_0 = 1$ and $\overline{a}_0 = 100$. Note that for $N \geq 12$ the two radii are very close. In fact, from (22) one has $J_{max|12} = 1.0059$. The comparison shows that, in a worst-case setting, if N is small, i.e., $J_{max|N}$ is much greater than 1, it is convenient to use the OEP, while if $J_{max|N} \simeq 1$, the two procedures are almost equivalent in terms of size of the final feasible set.

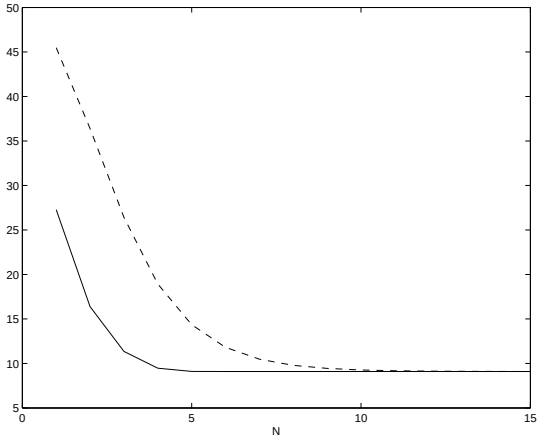

Fig. 1. Plot of ε_N (solid) and ϵ_N (dashed) for $N = 1, \ldots, 15$.

A different scenario arises when the true location of the parameter a is not the worst-case one. In this case, the OREP may lead to a much better performance than the OEP, in terms of the size of the actual resulting feasible set. The worst-case identification errors (with respect to the disturbance $d(t)$), for different values of a, are reported in Figures 2(a) and 2(b), in the cases $N = 10$ and $N = 4$, respectively. In both figures, $\beta = 0.1$ and $a \in [1, 100]$. It can be noticed that the error returned by the OEP does not depend on the actual value of a. Conversely, the radius of the feasible set provided by the OREP is a nondecreasing function of the true value of a. Clearly, if the maximum error also with respect to a is considered (as it is in the nature of the worst-case approach), the value returned by the OEP is smaller, by construction. However, Fig. 2(a) shows that for almost all values of a, OREP gives a smaller error than OEP. This is not the case when a shorter identification experiment is performed, as in Fig. 2(b) ($N = 4$).

Remark 4: From a practical point of view, the discussion above suggests that, if N is small (and then $J_{max|N}$ is significantly larger than 1) it is convenient to choose the OEP, which gives a guaranteed value of the worst-case radius, while if N is large (i.e., $J_{max|N} \simeq 1$), it is convenient to use the OREP, which has almost the same performance as the OEP for the worst values of the parameter a, but for other values of a may lead to a much smaller radius. If N is not given a priori, it is convenient to use the OREP and choose N such that $J_{max|N}$ satisfies the desired tolerance on the final radius reduction.

Remark 5: Notice that the OEP is highly sensitive to the a priori information, and in particular to $\overline{a}_0$. In fact, by (10) it follows that $\lim_{N\to\infty} \varepsilon_N = \frac{\beta}{1+\beta}\overline{a}_0$, independently of the value of true parameter a. So, if a priori bounds are chosen in a conservative way, such a conservatism will affect the final value of the radius, also if N is large. Instead, this fact is not true if one uses the OREP. In fact,

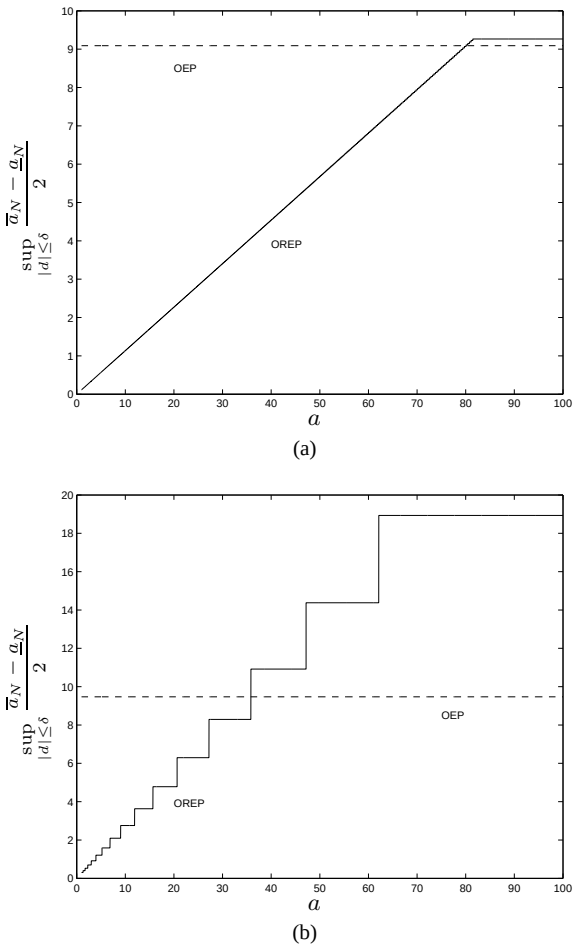

Fig. 2. Worst-case identification error for OREP (solid) and OEP (dashed) for $a \in [1, 100]$, $\beta = 0.1$, $N = 10$ (a) and $N = 4$ (b).

$D_{min}(\underline{a}_0, \overline{a}_0, a)$ is highly sensitive to the true parameter value and less sensitive to the initial feasible set.

V. Numerical example

Let us consider a system whose transfer function is

$$G(z) = \frac{38.4\,z^3 + 28.8\,z^2 + 35.2\,z - 20}{3.15\,z^3 - 0.08\,z^2 + 1.84\,z - 1}.$$

By observing that the impulse response of $G(z)$ can be considered negligible after $n = 30$ samples, the aim is to identify the first n samples of the impulse response. Let us assume that the bounds on the impulse response coefficients $-M\rho^i \leq \theta_i \leq M\rho^i$, $M = 20$, $\rho = 0.9$, $i = 1, \ldots, n$ are known a priori. Moreover, let us set $-500 \leq u(t) \leq 500$, $\delta = 3$ and $C = 20$, i.e., $\beta = 0.15$. Let us excite every FIR parameter $k = 4$ times, by applying the input design strategy proposed in [5]. This requires a total input length equal to $N = 1920$ (see [5], [10] for details). Both the OEP and the OREP are applied. In Fig. 3(a), a comparison between the resulting radii of the two approaches is reported. Both radii are computed for the worst-case realization of the disturbance $d(t)$ (i.e., the realization that produces the largest feasible set, for each input strategy). Since for some parameters $J_{max|k}$ is quite larger than 1, the corresponding OREP radii may result considerably bigger than the OEP radii (depending on the true value of the parameters).

Let us now repeat the experiment for $k = 10$. In this case, one needs $N = 4710$ samples. In Fig. 3(b), a comparison between the

resulting radii of the two approaches is reported. In this case, $J_{max|k}$ is close to 1, and the OREP strategy leads to better performance, as expected.

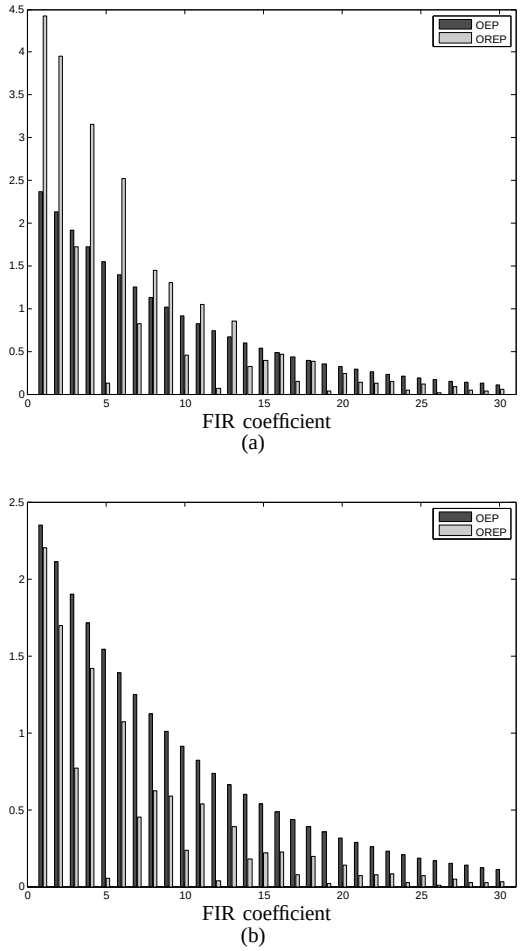

Fig. 3. Comparison between final radii obtained by using OEP (dark) and OREP (light), for $k = 4$ (a) and $k = 10$ (b).

VI. Conclusions

System identification based on binary-valued observations is an intriguing problem, due to the remarkable information reduction caused by the binary sensor. When a deterministic set-membership setting is considered, problems like optimal input design turn out to be even more challenging. The paper has provided new results for worst-case identification of systems equipped with binary sensors. The exact solution of the optimal input design problem for identification of gains allows one to devise an effective input design strategy for identification of general FIR systems. Moreover, an undesirable property of input design based on the worst-case identification error, namely the dependence of the optimal input sequence on the length of the identification experiment, has been overcome by introducing a new cost function. This led to the derivation of an optimal recursive design procedure, over a time interval of arbitrary length.

Several open issues deserve attention for future investigations. A key question is whether the input selection strategy based on independent excitation of each impulse response sample is optimal in general. Optimal input design for ARX models is another subject of ongoing research.

References

[1] L. Y. Wang, J. F. Zhang, and G. G. Yin, "System identification using binary sensors," *IEEE Transactions on Automatic Control*, vol. 48, no. 11, pp. 1892–1907, 2003.

[2] 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*, vol. 42, no. 4, pp. 535–547, 2006.

[3] J. Juillard, K. Jafaridinani, and E. Colinet, "Consistency of weighted least-square estimators for parameter estimation problems based on binary measurements," in *15th IFAC Symposium on System Identification*, Saint-Malo, France, July 2009, pp. 72–77.

[4] L. Y. Wang, G. G. Yin, J. F. Zhang, and Y. Zhao, *System Identification with Quantized Observations*. Springer, 2010.

[5] M. Casini, A. Garulli, and A. Vicino, "Time complexity and input design in worst-case identification using binary sensors," in *Proc. 46th IEEE Conference on Decision and Control*, New Orleans (USA), December 2007, pp. 5528–5533.

[6] Y. Zhao, L. Wang, J. Zhang, and G. Yin, "Jointly deterministic and stochastic identification of linear systems using binary-valued observations," in *15th IFAC Symposium on System Identification*, Saint-Malo, France, July 2009, pp. 60–65.

[7] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: an overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.

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

[9] D. N. C. Tse, M. A. Dahleh, and J. N. Tsitsiklis, "Optimal asymptotic identification under bounded disturbances," *IEEE Transactions on Automatic Control*, vol. 38, no. 8, pp. 1176–1190, 1993.

[10] M. Casini, A. Garulli, and A. Vicino, "Worst-case system identification using binary sensors: Input design and time complexity," Università di Siena, Tech. Rep. 2010-1, 2010, http://www.dii.unisi.it/priv/papers/papers_doc/42.pdf.