Input design in worst-case system identification with quantized measurements

Marco Casini ⋆, Andrea Garulli, Antonio Vicino,

Dipartimento di Ingegneria dell'Informazione
Centro per lo Studio dei Sistemi Complessi
Università di Siena
Via Roma 56, 53100 Siena, Italy

Abstract

This paper addresses the problem of set membership system identification with quantized measurements. Following the work developed for binary measurements, the problem of optimal input design with multiple sensor thresholds is tackled. For a FIR model of order n, the problem is decomposed into n static gain problems. The one-step optimal input problem is solved both for equispaced and generic sensor threshold distribution. Moreover, the N-step optimal input problem for the case of equispaced thresholds is addressed, and a solution is provided under a suitable assumption on the sensor range and resolution. The obtained results allow us to construct an upper bound on the time complexity of the FIR identification problem for the case of equispaced thresholds. Numerical application examples are reported to show the effectiveness of the proposed algorithms.

Key words: Quantized measurements, FIR models, input design, set membership identification

1 Introduction

System identification with quantized measurements is a research area which is receiving increasing attention. This topic draws motivation from several engineering fields. Communication systems show contexts like ATM networks where traffic information, e.g., bit rate, queue length, is measured through sensors characterized by appropriate thresholds. Typical sensors used in monitoring and control systems of industrial production plants are binary or quantized devices, popular examples being chemical process sensors in gas and oil industry, or sensors monitoring liquid or pressure levels. Quantized measurements are usually found in a number of automotive applications, including sensors for exhaust gas oxygen, shift-by-wire and ignition systems, photoelectric sensors for position detection, Hall-effect sensors for speed and acceleration measurement, ABS (anti-lock braking system), and others. More generally, all networked control systems feature a quantized data flow in input and output channels, and hence they call for system identification techniques taking into account the errors introduced by the sensor quantization.

In [20], Wang, Zhang and Yin introduced a new framework for addressing system identification problems with binary sensors, both in a stochastic and in a deterministic setting. Later on, they extended their approach to the case of quantized measurements and gave several contributions concerning space and time complexity [18], consistency and asymptotic efficiency of identification algorithms [16,17]. Least squares and instrumental variable methods for quantized data have been analyzed in [12]. Expectation maximization techniques have been employed in [7]. Optimal quantization schemes have been proposed in [1,15]. In [8], the use of dithering noise has been suggested to improve the performance of identification algorithms tailored to quantized data. Non asymptotic confidence sets for the estimated parameters have been studied in [21]. An extensive collection of results on identification with binary or quantized sensors can be found in [19].

Most contributions cited so far concern the stochastic setting. However, since the quantization introduces an error which is inherently unknown-but-bounded, it is quite natural to formulate the identification problem in a worst-case deterministic setting. In this context, the

⋆ Corresponding author. Phone: +39-0577-1912440; Fax: +39-0577-233602.

Email addresses: casini@ing.unisi.it (Marco Casini), garulli@ing.unisi.it (Andrea Garulli), vicino@ing.unisi.it (Antonio Vicino).

optimal input design problem for binary sensors has been studied in [5]. Preliminary results in the case of quantized sensors have been presented in [3,4]. Recent attempts to combine the stochastic and the deterministic approach can be found in [14, 22].

In this paper, the problem of system identification in presence of quantized measurements is addressed in a worst-case setting, based on the set-membership paradigm of uncertainty representation [6, 9, 10]. The first contribution concerns the optimal input design problem. The exact solution is provided for the case of a static gain, over a time horizon of length 1, for both generic and equally spaced quantization thresholds. The case of a time horizon of length N is also treated: although the complete solution remains an open problem, the optimal input sequence of length N is provided in the case that the sensor range and resolution satisfy a mild technical assumption. These results can be employed to construct efficient algorithms for identification of FIR models of arbitrary order, by choosing the class of input signals proposed in [2]. A further contribution of the paper concerns time complexity, i.e., the minimum number of measurements to be collected to reduce the estimation uncertainty below a given threshold [11, 13]. An upper bound to the time complexity, in the case of equally spaced thresholds, is provided for both static gains and for generic FIR models.

The paper is organized as follows. In Section 2, the input design problem is formulated in the worst-case deterministic setting. The 1-step optimal design problem with generic thresholds is considered in Section 3. Section 4 addresses the special case of equally spaced thresholds which is interesting both from a practical point of view (in many sensors the discretization is uniform) and because it brings a remarkable simplification of the problem solution. Section 5 concerns the N-step optimal design problem. Time complexity of the identification procedure is studied in Section 6. Numerical examples showing the performance of the proposed input design approach are reported in Section 7, while some concluding remarks are given in Section 8. In order to streamline the presentation, proofs of theorems and lemmas are reported in Appendix B.

2 Problem formulation

Let $\mathbb{R}^N$ denote the N-dimensional Euclidean space. A sequence of real numbers $\{x(t), t = 1, \ldots, N\}$ is identified by a vector $x \in \mathbb{R}^N$ and $||x||_p$ is 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

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

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 $\{\theta_i, i = 1, \ldots \ n\}$ represent the truncated system impulse response. The noise $d(t)$ 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. 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 $\{\theta_i, i = n + 1, \ldots\}$) 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 $C_1, \ldots, C_P$, 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 $C_0 \triangleq -\infty$, $C_{P+1} \triangleq +\infty$. If the quantization is uniform in the range $[C_1, C_P]$, i.e. the sensor thresholds satisfy $C_i = C_{i-1} + Q$, $i = 2, \ldots, P$, for some $Q > 0$, we will refer to this setting as *equispaced thresholds*.

Let $\theta^T = [\theta_1, \ldots, \theta_n] \in \mathbb{R}^n$ denote the FIR parameter vector and $\phi^T(t) = [u(t), \ldots, u(t - n + 1)]$ the regressor vector. Then, (1) can be expressed in regression form as $y(t) = \phi^T(t)\,\theta + d(t)$. Let Θ_0 represent the prior information available on the FIR parameter vector. Let us denote by $u, s \in \mathbb{R}^N$ the input signal $\{u(t), \ t = 1, \ldots, N\}$ and the sequence of discrete measurements $\{s(t), \ t = 1, \ldots, N\}$, respectively. For a given input-output realization $\{u, s\}$ of length N, the problem feasible parameter set is defined as:

$$\mathcal{F}_N = \bigcap_{t=1}^{N} \mathcal{S}_t$$

where

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

The worst-case local identification error is defined as

$$e_p(N,u,s) = \mathrm{rad}(\mathcal{F}_N) \triangleq \inf_{c\in\mathbb{R}^n} \sup_{\theta\in\mathcal{F}_N} \|\theta - c\|_p \ . \quad (3)$$

For a fixed input sequence u, the global worst-case error with respect to the disturbance realization is defined as

$$e_p(N,u) = \sup_s e_p(N,u,s) \ .$$

The aim of the optimal input design problem is to compute an input sequence providing the minimum worst-case identification error [13], i.e.,

$$e_p(N) = \inf_{u:\|u\|_\infty \le U} e_p(N,u) \ . \quad (4)$$

For a given level of accuracy ε, we define the *time complexity* of Θ_0 as the minimum time length of the experiment such that the optimal worst-case error reaches the accuracy ε, i.e. $N(\varepsilon) = \min_{e_p(N)\le\varepsilon} N$. The aim of the input design problem (4) is to choose the optimal input sequence u over a time horizon of length N. We will refer to this problem as *N-step optimal input design*. This is, in general, an intractable problem because it requires the solution of a nested sequence of N min-max optimizations. In fact, at each time t, one aims at minimizing with respect to the current input value $u(t)$, which is chosen as a function of the available information up to time $t-1$, i.e.

$$u(t) = \eta(\mathcal{F}_{t-1}; t). \quad (5)$$

However, this must be done with respect to the *worst-case* $\mathcal{F}_{t-1}$, i.e., by maximizing over all possible output signals $s(t-1)$, and so on backwards, for $t = N, N-1, \ldots, 1$.

A relaxed version of problem (4) is the so-called 1-*step optimal input design*, in which at each time t, one chooses $u(t)$ in order to minimize the worst-case parametric error at time t, i.e.

$$u(t) = \arg \inf_{u(t)=\eta(\mathcal{F}_{t-1};t)} \sup_{s(t)} \mathrm{rad}\left(\mathcal{F}_{t-1} \bigcap \mathcal{S}_t\right). \quad (6)$$

In the literature on system identification with binary or quantized measurements, it is common practice to consider classes of input signals which excite each FIR parameter individually (see e.g., [2,20]). In [2], the shortest input sequence exciting any FIR parameter individually has been provided. More precisely, it has been proven that to excite individually k times the n coefficients of a FIR, one needs an input sequence of length N equal to:

$$N = k(n+1)\frac{n}{2}. \quad (7)$$

An algorithm for constructing such an input sequence is reported in Appendix A. In this paper, we will assume to use this type 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 1-step and N-step optimal input design problems will be addressed for FIR systems of order 1 (static gains), for both generic and equispaced thresholds.

3 Input design with generic thresholds

Let us consider a FIR model of order $n = 1$, i.e.,

$$y(t) = au(t) + d(t) \quad (8)$$

with prior information $a \in [\underline{a}_0, \overline{a}_0]$ and $|d(t)| \le \delta$. We assume that the sign of a is known a priori (here a is assumed positive, w.l.o.g., and hence $\underline{a}_0 > 0$). In fact, the signs of the parameters θ_i of a generic FIR model (1) can be detected by performing a suitable preliminary identification experiment (see e.g. [20]). Moreover, let $C_P > 0$. The following assumption is enforced throughout the paper:

$$\underline{a}_0 > C_P/U \quad (9)$$

Assumption (9) guarantees that, even with the smallest admissible value of a, the input signal is able to excite all the thresholds (otherwise, some thresholds are useless and the problem can be reformulated with a smaller number of thresholds). Indeed, if $\underline{a}_0 U < C_P$, then for $a = \underline{a}_0$ and $d(t) = 0\ \forall t$, the sensor output can never be $s(t) = P$, no matter which input $u(t)$ is chosen. Given the above discussion, in order to simplify the treatment we assume $C_1 > 0$, and hence one can assume without loss of generality $0 < u(t) \le U$.

We start by addressing the 1-step optimal input design problem in the noise-free case. Let us denote by $\mathcal{F}_t = [\underline{a}_t, \overline{a}_t]$ the feasible parameter set at time t. The aim of the 1-step optimal design problem (6) is to find the input signal $u(t)$ such that

$$u^*(t) = \arg \inf_{0<u\le U} D_t(u) \quad (10)$$

where

$$D_t(u) = \sup_{\substack{s:\ s=\sigma(a\,u)\\ a\in\mathcal{F}_{t-1}}} (\overline{a}_t - \underline{a}_t) \quad (11)$$

and $D_t^* = D_t(u^*)$ is the optimal diameter of the feasible set.

Remark 1 *In [20] it has been shown that the 1-step optimal input at each time t in presence of binary measurements is*

$$u^*(t) = \frac{2C}{\underline{a}_{t-1} + \overline{a}_{t-1}} \quad (12)$$

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., $\overline{a}_t - \underline{a}_t = \frac{1}{2}(\overline{a}_{t-1} - \underline{a}_{t-1})$. The above input sequence is clearly also N-step optimal in the noise-free case, while this is not true in the noisy setting, as it has been shown in [5].

Let us define

$$v(t) \triangleq \frac{1}{u(t)} \tag{13}$$

and let the sensor output corresponding to the input $u(t)$ be $s(t) = i$. This means that $C_i < y(t) \leq C_{i+1}$, i.e., since $u(t) > 0$,

$$C_i\, v(t) < a \leq C_{i+1}\, v(t). \tag{14}$$

Thus, the posterior feasible set will be [1]

$$\begin{aligned}\mathcal{F}_t &= \mathcal{F}_{t-1} \cap [C_i\, v(t),\, C_{i+1}\, v(t)] \\ &= [\underline{a}_{t-1},\, \overline{a}_{t-1}] \cap [C_i\, v(t),\, C_{i+1}\, v(t)] \triangleq [\underline{a}_t\,,\, \overline{a}_t].\end{aligned} \tag{15}$$

From (13)-(15), we can rewrite problem (10)-(11) as $u^*(t) = 1/v^*(t)$ where

$$\begin{aligned}v^*(t) = \arg \Big\{ \inf_{v \geq 1/U} \max_{i=0,\ldots,P} (\min\{\overline{a}_{t-1}, C_{i+1}\, v\} \\ -\max\{\underline{a}_{t-1}, C_i\, v\}) \Big\}.\end{aligned} \tag{16}$$

Since we focus on the 1-step 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 $\mathcal{F} = [\underline{a}, \overline{a}]$. Let us define

$$\underline{v}_i \triangleq \frac{\underline{a}}{C_i}\ ,\ \overline{v}_i \triangleq \frac{\overline{a}}{C_i}\ ,\ i = 1, \ldots, P. \tag{17}$$

Thus, let $V^* \triangleq [\underline{v}_P,\, \overline{v}_1]$ and $H_i \colon a = C_i\, v\ ,\ i = 1, \ldots, P$. In Fig. 1, the values $\underline{v}_i$, $\overline{v}_i$ and the functions H_i are depicted for an example with $P = 3$.
We can now state the following lemma.

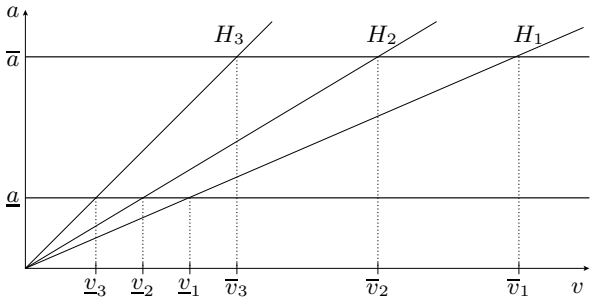

Fig. 1. Example of $\underline{v}_i$, $\overline{v}_i$ and functions H_i for $P = 3$.

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

Lemma 1 *Let v^* be an optimal solution of* (16)*. Then, $v^* \in V^*$.*

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

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

From (9), one has $U \geq \frac{C_P}{\underline{a}_0}$. This guarantess that the optimal input is always feasible, i.e. $u^*(t) \leq U$.
Let us sort the values $\underline{v}_i$, $\overline{v}_i$, $i = 1, \ldots, P$, in increasing order and rename them as $\hat{v}_1 \leq \hat{v}_2 \leq \ldots \leq \hat{v}_{2P}$. By construction one has $\hat{v}_1 = \underline{v}_P$ and $\hat{v}_{2P} = \overline{v}_1$. Let us define the intervals

$$W_1 = [\hat{v}_1, \hat{v}_2],\ \ldots\ ,\ W_{2P-1} = [\hat{v}_{2P-1}, \hat{v}_{2P}]. \tag{18}$$

By construction $\bigcup_{j=1}^{2P-1} W_j = V^*$. For $j = 1, \ldots, 2P-1$, let us define

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

and $v^{(j)}$ be the argument where the infimum in (19) is achieved.
Let us now analyze problem (19), i.e., the original problem (16) 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

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

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

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 $C_k v \geq \overline{a}$.*

According to (19), 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 $\mathcal{F}$ can be obtained by choosing 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 provides the explicit solution of problem (23).

Lemma 3 *Let* $Q = \max_{i \in \{m, m+1, \ldots, M-1\}} (C_{i+1} - C_i)$, *and*

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

Then, the solution of (23) *is given by*

$$D^{(j)} = \begin{cases} \max\{C_m\, v_L - \underline{a},\, Q\, v_L\} & \text{if } v_c \le v_L \\ \max\left\{ \frac{\overline{a}\, C_m - \underline{a}\, C_M}{C_m + C_M},\, \frac{Q\, \overline{a}}{Q + C_M} \right\} & \text{if } v_L < v_c < v_R \\ \overline{a} - C_M\, v_R & \text{if } v_c \ge v_R. \end{cases} \tag{24}$$

Moreover, the optimal solution of (23) *is attained at*

$$v^{(j)} = \begin{cases} v_L & \text{if } v_c \le v_L \\ v_c & \text{if } v_L < v_c < v_R \\ v_R & \text{if } v_c \ge v_R. \end{cases} \tag{25}$$

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 the above reasoning.

Theorem 1 *Let* $D^{(j)}$ *and* $v^{(j)}$ *be given by* (24)-(25), *respectively. At a given time* t, *the 1-step optimal solution* $u^*(t)$ *in* (10) *is given by* $u^*(t) = \dfrac{1}{v^*}$, *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* $u^* = \hat{u}$.

The result of Theorem 2 will be instrumental to define a specific set of thresholds C_i for which the time complexity can be computed exactly (see Section 6).

Now, let us consider the noisy case, i.e. model (8). The optimal input procedure can be formulated in a similar manner w.r.t. the noise-free case, by suitably amending the previous treatment. Due to the presence of noise, (14) becomes

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

Thus, the posterior feasible set is

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

Taking into account the presence of noise, let us rewrite (17) as

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

and define the functions

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

The 1-step optimal input design problem can be reformulated as in (16)

$$v^*(t) = \arg\Big\{ \inf_{v\geq 1/U} \max_{i:\, i=0,\dots,P} (\min\{\overline{a}_{t-1}, (C_{i+1}+\delta)v\} - \max\{\underline{a}_{t-1}, (C_i-\delta)v\})\Big\}. \quad (27)$$

Following the same reasoning as in the noise-free case, one can introduce the restricted optimization problems

$$D^{(j)} = \inf_{v\in W_j} \max_{i=0,\dots,P} (\min\{\overline{a}, (C_{i+1}+\delta)v\} - \max\{\underline{a}, (C_i-\delta)v\}) \quad (28)$$

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

Lemma 4 *Let m and M be defined as in* (20)-(21), $Q = \max_{i\in\{m,m+1,\dots,M-1\}} (C_{i+1} - C_i + 2\,\delta)$ *and*

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

Then, the solution of problem (28) *is given by*

$$D^{(j)} = \begin{cases} \max\{(C_m+\delta)\,v_L - \underline{a},\, Q\,v_L\} & \text{if } v_c \leq v_L \\ \max\left\{ \frac{\overline{a}\,C_m - \underline{a}\,C_M + \delta(\underline{a}+\overline{a})}{C_m + C_M},\ \frac{Q\,\overline{a}}{Q + C_M - \delta} \right\} & \text{if } v_L < v_c < v_R \\ \overline{a} - (C_M - \delta)\,v_R & \text{if } v_c \geq v_R. \end{cases} \quad (29)$$

Moreover, the optimal solution is attained at

$$v^{(j)} = \begin{cases} v_L & \text{if } v_c \leq v_L \\ v_c & \text{if } v_L < v_c < v_R \\ v_R & \text{if } v_c \geq v_R. \end{cases} \quad (30)$$

From (27)-(28) and Lemma 4, it is clear that Theorem 1 applies as well to the noisy case with $D^{(j)}$ and $v^{(j)}$ given by (29) and (30), respectively. This result provides the 1-step optimal input u^* at each time t, in the noisy case.

4 Input design with equispaced thresholds

In this section, we assume that the sensor thresholds are equispaced, i.e., $C_i = C_{i-1} + Q$, $i = 2, \dots, P$, $Q > 0$. In this setting, the solution of the 1-step optimal design problem turns out to be much simpler than that found in Section 3 for the case of generic thresholds. In particular, while the solution presented in Theorem 1 requires the application of an algorithmic procedure to compute the optimal input value, in the case of equispaced thresholds an analytic solution depending only on the quantities $\underline{a}$, $\overline{a}$, C_1, C_P and Q is provided. We start by considering the noise-free case.

Let us address problem (10)-(11) in the setting of equispaced thresholds. Let $\underline{v}_i$ and $\overline{v}_i$ be defined as in (17). Let $u^* = 1/v^*$ be the optimal input and define

$$m \triangleq \arg \min_{i=1,\dots,P} \{i\colon \underline{v}_i \leq v^*\} \quad (31)$$

$$M \triangleq \arg \max_{i=1,\dots,P} \{i\colon \overline{v}_i \geq v^*\}. \quad (32)$$

Notice that definitions (17), (31) and (32) imply that at the considered time t, only the thresholds C_i, $i = m, \dots, M$, can contribute to the reduction of the feasible set, since they satisfy $\underline{a} \leq C_i\, v^* \leq \overline{a}$. Clearly, $M \geq m$ holds (otherwise, there would be no reduction of the feasible set). It is worth remarking that (31)-(32) differ from (20)-(21) because they do not depend on a specific interval in v, but on the actual optimal solution v^*.

Lemma 5 *At a given time t, let $u^* = 1/v^*$ be the solution of problem* (10)-(11)*. It follows that*

- *If $m = M$ then*

$$v^* = \frac{\underline{a}+\overline{a}}{2\,C_M} \quad (33)$$

and the resulting optimal diameter is

$$D^* = \overline{a} - C_M\, v^* = C_m\, v^* - \underline{a} = \frac{\overline{a}-\underline{a}}{2}.$$

- *If $m < M$ then*

$$v^* = \begin{cases} \dfrac{\overline{a}}{C_M + Q} & \text{if } \dfrac{\underline{a}}{\overline{a}} \geq \dfrac{C_m - Q}{C_M + Q} \\ \dfrac{\underline{a}+\overline{a}}{C_m + C_M} & \text{if } \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_m - Q}{C_M + Q} \end{cases} \quad (34)$$

and

$$D^* = \overline{a} - C_M\, v^* = \max\{Q\, v^*,\, C_m\, v^* - \underline{a}\} = \begin{cases} \dfrac{Q\,\overline{a}}{C_M + Q} & \text{if } \dfrac{\underline{a}}{\overline{a}} \geq \dfrac{C_m - Q}{C_M + Q} \\ \dfrac{\overline{a}\,C_m - \underline{a}\,C_M}{C_m + C_M} & \text{if } \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_m - Q}{C_M + Q}. \end{cases} \quad (35)$$

Notice that Lemma 5 does not provide a constructive procedure for computing v^*, as the indexes m and M depend on v^* itself. Nevertheless, it will be exploited in the following to derive analytic expressions for v^* and D^*. Observe also that when only one threshold is active at the optimum, i.e. $m = M$, the feasible set is reduced by $1/2$ as in presence of binary sensors. The next lemma bounds the range where the optimal solution lies.

Lemma 6 *An optimal solution* $u^* = 1/v^*$ *of problem* (10)-(11) *satisfies* $v^* \in V^* \triangleq \left[\frac{\underline{v}_P+\overline{v}_P}{2}, \overline{v}_P\right] = \left[\frac{\underline{a}+\overline{a}}{2\,C_P}, \frac{\overline{a}}{C_P}\right]$.

The following theorem provides the solution of the 1-step optimal input design problem for the case of equispaced thresholds.

Theorem 3 *Let* $C_1, \ldots, C_P$ *be equispaced thresholds. At a given time* t*, the solution of problem* (10)-(11) *is* $u^*(t) = 1/v^*$ *where*

$$v^* = \begin{cases} \dfrac{\underline{a}+\overline{a}}{2\,C_P} & \text{if } \dfrac{\underline{a}}{\overline{a}} \geq \dfrac{C_P-Q}{C_P+Q} \\ \dfrac{\overline{a}}{C_P+Q} & \text{if } \dfrac{C_1-Q}{C_P+Q} \leq \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_P-Q}{C_P+Q} \\ \dfrac{\underline{a}+\overline{a}}{C_1+C_P} & \text{if } \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_1-Q}{C_P+Q}\,. \end{cases}$$

The corresponding size of the feasible set turns out to be

$$D^* = \begin{cases} \dfrac{\overline{a}-\underline{a}}{2} & \text{if } \dfrac{\underline{a}}{\overline{a}} \geq \dfrac{C_P-Q}{C_P+Q} \\ \dfrac{\overline{a}\,Q}{C_P+Q} & \text{if } \dfrac{C_1-Q}{C_P+Q} \leq \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_P-Q}{C_P+Q} \\ \dfrac{C_1\,\overline{a}-C_P\,\underline{a}}{C_1+C_P} & \text{if } \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_1-Q}{C_P+Q}\,. \end{cases}$$

Remark 3 *For a binary sensor with threshold* $C_P = C_1 = C$*, it is immediate to see that the* 1*-step optimal input provided by Theorem 3 boils down to* $u^* = \frac{2C}{\underline{a}+\overline{a}}$*, which corresponds to* (12) *in Remark 1.*

Let us now consider the case of noisy data. Then, problem (6) can be cast as (10) with

$$D_t(u) = \sup_{\substack{s:\ s=\sigma(au+d) \\ a\in\mathcal{F}_{t-1};\ |d|\leq\delta}} (\overline{a}_t - \underline{a}_t). \tag{36}$$

The solution of the 1-step optimal input design problem in presence of noise is given by the next theorem.

Theorem 4 *Let* $C_1, \ldots, C_P$ *be equispaced thresholds. At a given time* t*, the solution of problem* (10),(36) *is* $u^* = 1/v^*$ *where*

$$v^* = \begin{cases} \dfrac{\underline{a}+\overline{a}}{2\,C_P} & \text{if } \dfrac{C_P-Q-\delta}{C_P+Q+\delta} \leq \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_P-\delta}{C_P+\delta} \\ \dfrac{\overline{a}}{C_P+Q+\delta} & \text{if } \dfrac{C_1-Q-\delta}{C_P+Q+\delta} \leq \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_P-Q-\delta}{C_P+Q+\delta} \\ \dfrac{\underline{a}+\overline{a}}{C_1+C_P} & \text{if } \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_1-Q-\delta}{C_P+Q+\delta}\,. \end{cases}$$

The corresponding size of the feasible set turns out to be

$$D^* = \begin{cases} \dfrac{\overline{a}-\underline{a}}{2} + \dfrac{\delta\,(\underline{a}+\overline{a})}{2\,C_P} & \text{if } \dfrac{C_P-Q-\delta}{C_P+Q+\delta} \leq \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_P-\delta}{C_P+\delta} \\ \dfrac{(Q+2\delta)\,\overline{a}}{C_P+Q+\delta} & \text{if } \dfrac{C_1-Q-\delta}{C_P+Q+\delta} \leq \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_P-Q-\delta}{C_P+Q+\delta} \\ \dfrac{\overline{a}\,C_1-\underline{a}\,C_P+\delta(\underline{a}+\overline{a})}{C_1+C_P} & \text{if } \dfrac{\underline{a}}{\overline{a}} \leq \dfrac{C_1-Q-\delta}{C_P+Q+\delta}\,. \end{cases}$$

5 N-step optimal input design

Let us now consider the *N-step optimal input design* problem for equispaced thresholds, in the noise-free case. According to (4) and (5), the aim is to choose the input signal sequence $u^*(t)$, $t = 1, \ldots, N$ such that

$$u^* = \arg \inf_{\substack{u:\ \|u\|_\infty \leq U \\ u(t)=\eta(\mathcal{F}_{t-1};t)}} e_p(N,u) \tag{37}$$

Let us introduce a key concept that will be useful to solve problem (37).

Definition 1 *At a given time* t*, we say that the selection of the* N*-step optimal input* $u^*(t)$ *is a* binary design problem *if, for each* v *satisfying* $\underline{a} \leq C_i\,v \leq \overline{a}$ *for some* $i \in \{1, \ldots, P\}$*, it holds* $C_{i+1}\,v \geq \overline{a}$ *and* $C_{i-1}\,v \leq \underline{a}$*.*

The name binary design comes from the fact that when the corresponding condition holds, each input will generate at most two different sensor outputs (depending on the value of a) and hence the maximum reduction rate of the feasible set at time t will be $1/2$, as it is always the case in system identification with binary sensors (see Remark 1). The next result provides a necessary and sufficient condition for the input design problem to be a binary design.

Lemma 7 *At a given time* t*, the choice of* $u^*(t)$ *is a binary design problem if and only if* $\overline{v}_P \leq \underline{v}_{P-1}$*.*

According to Lemma 7, if the problem is a binary design one has

$$\frac{\underline{a}}{\overline{a}} \geq \frac{C_P-Q}{C_P} > \frac{C_P-Q}{C_P+Q}$$

and Theorem 3 states that the 1-step optimal input reduces the feasible set by 1/2. Moreover, since $\underline{a}_t$ is a non decreasing function and $\overline{a}_t$ is a non increasing function, it follows that $\frac{\underline{a}_t}{\overline{a}_t}$ is a non decreasing function. Therefore, if at time $\overline{t}$ one has a binary design problem, this will hold also for any time $t \geq \overline{t}$. Since it is always possible to select an input that reduces the feasible set by 1/2 in one step, it can be concluded that the binary design condition corresponds to the worst possible situation in the N-step input design problem.

The theorems given next allow one to derive the solution of problem (37), for different a priori information on the uncertain parameter.

Theorem 5 *Let $C_1, \ldots, C_P$ be equispaced thresholds. Let N be the length of the input signal to be applied and let $\frac{\underline{a}_0}{\overline{a}_0} \geq \frac{C_P - Q}{C_P + Q}$. Then, the optimal input solving problem* (37) *is given by $u^*(t) = \frac{1}{v^*(t)}$ where:*

$$v^*(t) = \frac{\underline{a}_{t-1} + \overline{a}_{t-1}}{2\, C_P} \quad , \qquad t = 1, 2, \ldots, N. \tag{38}$$

Moreover, the resulting diameter of the feasible set $\mathcal{F}_N$ at time N turns out to be

$$D_N^* = (\overline{a}_0 - \underline{a}_0) \left(\frac{1}{2}\right)^N .$$

Theorem 6 *Let $C_1, \ldots, C_P$ be equispaced thresholds. Let N be the length of the input signal to be applied and let $\frac{C_1 - Q}{C_P + Q} \leq \frac{\underline{a}_0}{\overline{a}_0} \leq \frac{C_P - Q}{C_P + Q}$. Then, the optimal input solving problem* (37) *is given by $u^*(t) = \frac{1}{v^*(t)}$ where:*

$$v^*(1) = \frac{\overline{a}_0}{C_P + Q} \tag{39}$$

and $v^*(t)$, $t = 2, \ldots, N$, are chosen according to (38)*. Moreover, the diameter of the resulting feasible set $\mathcal{F}_N$ is*

$$D_N^* = \frac{Q\, \overline{a}_0}{C_P + Q} \left(\frac{1}{2}\right)^{N-1} . \tag{40}$$

It is worth observing that the optimal input sequences provided by Theorems 5 and 6 coincide with those given by the 1-step optimal strategy in Theorem 3. In particular, the optimal input provided by Theorem 5 is similar to the optimal input for the case of noise-free binary measurements. In fact, under the condition $\frac{\underline{a}_0}{\overline{a}_0} \geq \frac{C_P - Q}{C_P + Q}$, only one threshold at a time can be used to reduce the parameter uncertainty leading to the same uncertainty reduction as in the binary case. Conversely, the solution provided by Theorem 6 fully exploits all the available quantization levels by choosing appropriately the first input sample $u^*(1)$.

The N-step optimal input design when $\frac{\underline{a}_0}{\overline{a}_0} < \frac{C_1 - Q}{C_P + Q}$ remains an open problem. In general, the N-step optimal input sequence is different from the 1-step optimal one given by Theorem 3, and it is possible to show that there exist examples in which the binary design condition is never reached. Nevertheless, Theorem 5 and 6 cover all the cases when $\frac{\underline{a}_0}{\overline{a}_0} \geq \frac{C_1 - Q}{C_P + Q}$, which represents a condition usually satisfied in practice. For instance, such a condition holds whenever the sensor range is large enough. In fact, for any initial feasible set, there exists a sufficiently large C_P such that the condition is satisfied. Moreover, the condition is always satisfied if $C_1 \leq Q$: this occurs, e.g., whenever the lowest threshold of the quantized sensor is zero or negative.

6 Time complexity

The next result concerns the evaluation of the time complexity for problem (37), in the case when the thresholds are equispaced and measurements are noise free.

Theorem 7 *Let $\mathcal{F}_0 = [\underline{a}_0, \overline{a}_0]$. The time complexity $N(D_N)$ required to reduce the diameter of the feasible set from D_0 to $D_N > 0$ satisfies*

$$N(D_N) = \begin{cases} \left\lceil \log_2\left(\frac{D_0}{D_N}\right) \right\rceil & \text{if } \frac{\underline{a}_0}{\overline{a}_0} \geq \frac{C_P - Q}{C_P + Q} \\ \left\lceil \log_2\left(\frac{2\, Q\, \overline{a}_0}{(C_P + Q) D_N}\right) \right\rceil & \text{if } \frac{C_1 - Q}{C_P + Q} \leq \frac{\underline{a}_0}{\overline{a}_0} \leq \frac{C_P - Q}{C_P + Q} \end{cases} \tag{41}$$

$$N(D_N) \leq \left\lceil \log_2\left(\frac{2(C_1 \overline{a}_0 - C_P \underline{a}_0)}{(C_1 + C_P) D_N}\right) \right\rceil \text{ if } \frac{\underline{a}_0}{\overline{a}_0} \leq \frac{C_1 - Q}{C_P + Q}.$$

Notice that the time complexity is computed exactly under the assumptions of Theorems 5 and 6, when the solution of the N-step input design problem is available. When these assumptions do not hold, Theorem 7 gives an upper bound on $N(D_N)$.

According to condition (26) in Theorem 2, let us choose the thresholds so that they satisfy

$$C_i = \left[\frac{P + 1 - i}{P + 1}\, \underline{a} + \frac{i}{P + 1}\, \overline{a} \right] \alpha \quad i = 1, \ldots, P \tag{42}$$

for some $\alpha > 0$. It is easy to check that the thresholds (42) are equispaced, with $Q = \frac{\overline{a}_0 - \underline{a}_0}{P+1} \alpha$. Moreover, one has

$$\frac{C_1 - Q}{C_P + Q} = \frac{\underline{a}_0}{\overline{a}_0}$$

and by (41) one gets, after simple manipulations

$$N(D_N) = \left\lceil \log_2\left(\frac{2\, Q\, \overline{a}_0}{(C_P + Q) D_N}\right) \right\rceil = \left\lceil 1 + \log_2\left(\frac{D_0}{(P+1) D_N}\right) \right\rceil. \tag{43}$$

From (43) it can be concluded that, for the thresholds chosen as in (42), $N(D_N) = 1$ whenever $\frac{D_0}{D_N} \leq P+1$, as predicted by Theorem 2. Conversely, if $\frac{D_0}{D_N} > P+1$, the time complexity is reduced by $\log_2(P+1)-1$, with respect to the binary case, in which the time complexity is equal to $\log_2\left(\frac{D_0}{D_N}\right)$.

The following theorem, which follows directly by combining the result in Theorem 7 with (7), provides upper bounds on the time complexity for a generic FIR model of order n.

Theorem 8 *Consider a FIR model of order n and let*

$$\Theta_0 = B_\infty\left(c_0, \frac{D_0}{2}\right) = [\underline{\theta}_{1,0}, \overline{\theta}_{1,0}] \times \times \ldots \times [\underline{\theta}_{n,0}, \overline{\theta}_{n,0}]$$

represent the prior information available on the FIR parameter vector. The time complexity $N(D_N)$ required to reduce the ℓ_∞ diameter of the feasible set from D_0 to $D_N > 0$ satisfies

$$N(D_N) \leq \frac{n(n+1)}{2} \max_{i=1,\ldots,n}\{N_i(D_N)\}$$

where

$$N_i(D_N) = \begin{cases} \left\lceil \log_2\left(\frac{\overline{\theta}_{i,0}-\underline{\theta}_{i,0}}{D_N}\right)\right\rceil & \text{if } \frac{\underline{\theta}_{i,0}}{\overline{\theta}_{i,0}} \geq \frac{C_P-Q}{C_P+Q} \\ \left\lceil \log_2\left(\frac{2\,Q\,\overline{\theta}_{i,0}}{(C_P+Q)D_N}\right)\right\rceil & \text{if } \frac{C_1-Q}{C_P+Q} \leq \frac{\underline{\theta}_{i,0}}{\overline{\theta}_{i,0}} \leq \frac{C_P-Q}{C_P+Q} \\ \left\lceil \log_2\left(\frac{2(C_1\overline{\theta}_{i,0}-C_P\underline{\theta}_{i,0})}{(C_1+C_P)D_N}\right)\right\rceil & \text{if } \frac{\underline{\theta}_{i,0}}{\overline{\theta}_{i,0}} \leq \frac{C_1-Q}{C_P+Q}. \end{cases}$$

7 Examples

In this section, two numerical examples are presented to illustrate the benefits of the proposed input design technique.

7.1 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 3.

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.

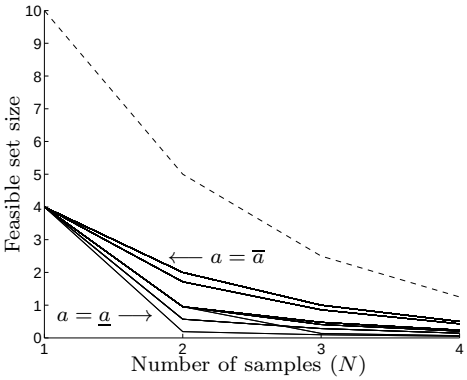

Fig. 2. Example 1: Feasible set size of a 4-thresholds sensor (solid) compared with a binary one (dashed) for different values of the true parameters a.

7.2 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 $P = 5$ equispaced thresholds, namely $C_j = 10\,j$, $j = 1, \ldots, 5$.

Let us suppose we want to independently excite each FIR coefficient 4 times. By applying the input strategy provided in Appendix A, one needs $N = 4\,n(n+1)/2 = 220$ samples according to (7). Then, each FIR coefficient is excited by the optimal input derived in Theorem 4. The true parameters and the noise signal $d(t)$ are chosen to maximize the size of the final feasible sets, according to the worst-case setting.

In Fig. 3, the feasible set bounds for each parameter after applying the input signal of Theorem 4, are reported. Such bounds are compared to that obtained by assuming a uniformly distributed input in $[0, 50]$. Notice that the true values of the parameters are chosen in a worst-case setting, and hence they may vary between the two input strategies applied.

Fig. 3 shows how the input choice reported in Theorem 4 outperforms that given by a uniformly distributed input. To better emphasize such a behavior, in Table 1 the final feasible set size obtained by applying the input signal reported in Theorem 4 is compared with the mean value obtained by a uniformly distributed input over 100000 realizations.

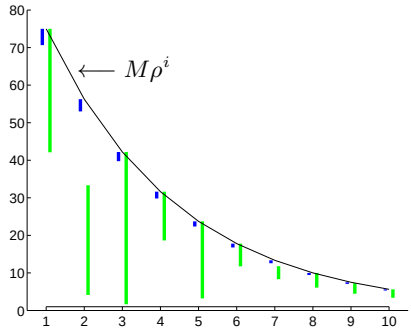

Fig. 3. Example 2: Feasible set bounds for an input signal independently exciting each parameter 4 times. Blue intervals denote feasible sets related to the input provided by Theorem 4, while green intervals are related to a uniformly distributed input.

Table 1
Comparison between the feasible set size obtained by applying the input provided in Theorem 4 ($D_{Th.\ 4}$) and the mean value obtained by a uniformly distributed input over 100000 realizations ($Mean(D_{uniform})$). The last column reports the standard deviation of $D_{uniform}$.

i	$D_{Th.\ 4}$	$Mean(D_{uniform})$	$STD(D_{uniform})$
1	4.33	63.32	13.49
2	3.25	45.60	10.90
3	2.44	32.60	8.68
4	1.83	23.04	6.83
5	1.37	16.13	5.25
6	1.03	11.12	3.96
7	0.77	7.56	2.90
8	0.58	5.07	2.06
9	0.43	3.35	1.40
10	0.32	2.18	0.91

8 Conclusions

System identification with quantized measurements is a challenging problem even for very simple models. In this paper, the input design problem has been tackled in a worst-case setting. Considering static gains, a complete solution has been given for the one-step optimal design, while the N-step optimal problem has been solved under suitable technical assumptions. These results can be used to devise suboptimal input sequences for generic FIR models, by exploiting input signals which allow one to excite individually each FIR parameter. Moreover, the developed theory has led to the construction of an upper bound to the time-complexity for the identification of FIR models.

The worst-case approach adopted in the paper seems the most natural choice when dealing with quantization errors, which are unknown-but-bounded in nature. This brings in a remarkable simplification, as the measurement noise and the quantization error can be treated in the same way. On the other hand, the classic stochastic setting allows one to prove strong results such as asymptotic efficiency of the estimates (see e.g. [16]). From a practical point of view, one should see the two approaches as complementary and choose the most appropriate one depending on the a priori information on the measurement noise and on the required level of confidence on the parameter bounds during the transient (which are usually more conservative in the worst-case setting).

There are several open issues in worst-case system identification with quantized measurements, which deserve future investigations. The N-step optimal design problem with noisy data has not been fully solved yet. Such a result would allow one to devise upper bounds to the time-complexity also in the noisy scenario. Another open problem concerns the optimal selection of the sensor thresholds. Choosing the quantization mechanism in order to maximize the resulting model quality is an intriguing problem which has been addressed in the stochastic setting, but not yet in the worst-case one. Finally, identification of models with different structure, such as ARX or OE, is the subject of ongoing research.

References

[1] J. C. Aguero, G. C. Goodwin, and J. I. Yuz. System identification using quantized data. In *46th IEEE Conference on Decision and Control*, pages 4263 –4268, December 2007.

[2] 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*, pages 5528–5533, New Orleans (USA), December 2007.

[3] M. Casini, A. Garulli, and A. Vicino. Optimal input design for identification of systems with quantized measurements. In *Proc. 47th IEEE Conf. on Decision and Control*, pages 5506–5512, Cancun (Mexico), December 2008.

[4] M. Casini, A. Garulli, and A. Vicino. Input design for worst-case system identification with uniformly quantized measurements. In *15th IFAC Symposium on System Identification*, pages 54–59, Saint-Malo (France), July 2009.

[5] M. Casini, A. Garulli, and A. Vicino. Input design in worst-case system identification using binary sensors. *IEEE Transactions on Automatic Control*, 56(5):1186–1191, 2011.

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

[7] B. I. Godoy, G. C. Goodwin, J. C. Aguero, D. Marelli, and T. Wigren. On identification of FIR systems having quantized output data. *Automatica*, 47(9):1905–1915, 2011.

[8] F. Gustafsson and R. Karlsson. Estimation based on quantized information. In *15th IFAC Symposium on System Identification*, pages 78–83, Saint-Malo, France, July 2009.

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

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

[11] K. Poolla and A. Tikku. On the time complexity of worst-case system identification. *IEEE Transactions on Automatic Control*, 39(5):944–950, 1994.

[12] H. Suzuki and T. Sugie. System identification based on quantized I/O data corrupted with noises and its performance improvement. In *45th IEEE Conference on Decision and Control*, pages 3684–3689, 2006.

[13] 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.

[14] K. Tsumura. Optimal quantizer for mixed probabilistic/ deterministic parameter estimation. In *45th IEEE Conference on Decision and Control*, pages 6259 –6264, December 2006.

[15] K. Tsumura and J. Maciejowski. Optimal quantization of signals for system identification. In *Proceedings of the European Control Conference*, Cambridge, UK, 2003.

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

[17] L. Y. Wang and G. G. Yin. Quantized identification with dependent noise and fisher information ratio of communication channels. *IEEE Transactions on Automatic Control*, 55(3):674 –690, 2010.

[18] L. Y. Wang, G. G. Yin, J. F. Zhang, and Y. Zhao. Space and time complexities and sensor threshold selection in quantized identification. *Automatica*, 44(12):3014 – 3024, 2008.

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

[20] 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.

[21] E. Weyer, S. Ko, and M. C. Campi. Finite sample properties of system identification with quantized output data. In *Proceedings of the 48th IEEE Conference onDecision and Control, held jointly with the 28th Chinese Control Conference*, pages 1532 –1537, December 2009.

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

A Input sequence construction

Consider the FIR model (1). We want to excite each FIR parameter individually, with a predefined value of the input. In [2], a procedure for constructing an input sequence $u(t), t = 1, \ldots, k(n(n+1)/2)$ which allows one to individually excite the n FIR parameters k times is reported. A pseudo-code implementing such a procedure is reported in Algorithm 1 for the case of n even. Here, u_i^j denotes the input sample exciting the i-th coefficient for the j-th time ($i = 1, \ldots, n,\ j = 1, \ldots, k$), and function `mod` denotes the remainder after division. Algorithm 1 can be easily modified to consider the case of n odd.

Algorithm 1 Input sequence construction (n even)

```
Set n and k
for t ← 1 : k(n(n+1)/2) do
    i ← ⌈(t−1)/(k(n+1))⌉ + 1
    j ← mod(⌈(t−1)/(n+1)⌉, k) + 1
    if mod(t, n+1) == 1 then
        u(t) ← u_i^j
    else if mod(t, n+1) == i+1 then
        u(t) ← u_{n+1−i}^j
    else
        u(t) ← 0
    end if
end for
```

B Proofs of lemmas and theorems

Proof of Lemma 1: By contradiction, assume that $v^* \notin V^*$, e.g., $v^* < \underline{v}_P$. One has

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

So, the sensor output is $s = P$ independently of the true value of a, and being $C_P v^* < C_P \underline{v}_P = \underline{a}$, one gets

$$\min\{\overline{a}, C_{P+1} v^*\} - \max\{\underline{a}, C_P v^*\} = \overline{a} - \underline{a}.$$

Therefore, v^* does not cause any reduction of the feasible set, which means that v^* is not optimal, leading to a contradiction. A similar reasoning can be repeated for the case $v^* > \overline{v}_1$. □

Proof of Lemma 2: 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, for any $v \in [v_L, v_R]$,

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

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

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

Proof of Lemma 3: Let

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

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

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

Being $F_i(v)$ linear in v, the $v^{(j)}$ at which the infimum in (B.1) 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 one would have

$$\max\{F_{m-1}(\widetilde{v}-\varepsilon),\, F_q(\widetilde{v}-\varepsilon),\, F_M(\widetilde{v}-\varepsilon)\} < \max\{F_{m-1}(\widetilde{v}),\, F_q(\widetilde{v}),\, F_M(\widetilde{v})\}.$$

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 $F_M(v_R) = \overline{a} - C_M\, v_R$. □

Proof of Theorem 2: Let us assume that (26) holds and let $\hat{v} = 1/\hat{u}$. Since

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

and

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

one has $\underline{v}_1 \le \overline{v}_P$. By (17), 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

$$\begin{aligned} F_0(\hat{v}) &= C_1\,\hat{v} - \underline{a} = \frac{\overline{a}-\underline{a}}{P+1} \\ F_i(\hat{v}) &= (C_{i+1} - C_i)\,\hat{v} = \frac{\overline{a}-\underline{a}}{P+1},\ i = 1, \ldots, P-1 \quad \text{(B.2)} \\ F_P(\hat{v}) &= \overline{a} - C_P\,\hat{v} = \frac{\overline{a}-\underline{a}}{P+1}. \end{aligned}$$

Since the maximum possible reduction of the feasible set is by a factor $P+1$, one has $D^* = \frac{\overline{a}-\underline{a}}{P+1}$ and $v^* = \hat{v}$.
Conversely, assume that one can reduce the feasible set size by a factor $P+1$ with some $\hat{u} = \frac{1}{\hat{v}}$. Then, the relationships (B.2) must hold and (26) easily follows. □

Proof of Lemma 5: Define the intervals W_j as in (18) and $D^{(j)}$ as in (19). From (16), it turns out that $v^* = v^{(j^*)}$, where $j^* = \arg\ \min_{j=1,\ldots,2P-1} D^{(j)}$ and W_j^* is the interval where the optimum lies. Let us denote by v_L and v_R the left and right bounds of such an interval, i.e. $W_{j^*} = [v_L, v_R]$. According to (B.1) and the definition of m and M in (31)-(32), one has

$$D^{(j^*)} = \inf_{v \in [v_L,\, v_R]} \max\{C_m\, v - \underline{a},\, Q\, v,\, \overline{a} - C_M\, v\} \quad \text{(B.3)}$$

Moreover, it can be observed that m and M defined in (31)-(32) are the same as those in (20)-(21) for the specific optimal interval W_{j^*}. Therefore, Lemma 3 holds within the interval W_{j^*}, with $D^{(j^*)}$ and $v^{(j^*)}$ given by (24) and (25), respectively.
Let us first consider the case $m = M$. Then (B.3) simplifies to

$$D^{(j^*)} = \inf_{v \in [v_L,\, v_R]} \max\{C_M\, v - \underline{a},\, \overline{a} - C_M\, v\}.$$

and hence the solution satisfies $C_M\, v - \underline{a} = \overline{a} - C_M\, v$, thus giving (33).

Now, let us consider the case $m < M$. First, we want to prove that $v^* = v_c \triangleq \min\left\{\frac{\underline{a}+\overline{a}}{C_m+C_M}, \frac{\overline{a}}{Q+C_M}\right\}$. It follows that $v^* \ne v_c$ may occur only when $v^* = v_L$ or $v^* = v_R$. By construction, both v_L and v_R are equal either to $\underline{v}_h$ or $\overline{v}_h$, for some h.
Let us consider the case $v^* = \overline{v}_h$, which implies $M = h$. By (B.3), $D^* = \max\{Q\, v^*,\, C_m\, v^* - \underline{a}\}$ since $\overline{a} - C_M\, v^* = \overline{a} - C_h\, v^* = 0$. Thus, it is easy to see that $\exists \varepsilon > 0$: $\widetilde{v} = v^* - \varepsilon$ and $\widetilde{D} = \max\{Q\,\widetilde{v},\, C_m\,\widetilde{v} - \underline{a}, \overline{a} - C_i\,\widetilde{v}\} < D^*$, and so v^* is not optimal.
Let us now consider the case $v^* = \underline{v}_h$, which implies $m = h$. Then, $D^* = \max\{Q\, v^*,\, \overline{a} - C_M\, v^*\}$ since $C_h\, v^* - \underline{a} = C_m\, v^* - \underline{a} = 0$. If $Q\, v^* = \overline{a} - C_M\, v^*$, it follows that $v^* = \frac{\overline{a}}{Q+C_M} \le \frac{\underline{a}+\overline{a}}{C_m+C_M}$ and hence $v^* = v_c$. Then, let us assume $Q\, v^* \ne \overline{a} - C_M\, v^*$. If $Q\, v^* < \overline{a} - C_M\, v^*$, there exists $\varepsilon > 0$: $\widetilde{v} = v^* + \varepsilon$ and $\widetilde{D} = \max\{Q\,\widetilde{v},\, C_i\,\widetilde{v} - \underline{a}, \overline{a} - C_M\,\widetilde{v}\} < D^*$. A similar reasoning can be repeated

if $Q\,v^* > \overline{a} - C_M\,v^*$.
Therefore $v^* = v_c$, and according to Lemma 3

$$D^{(j^*)} = D^* = \max\left\{\frac{\overline{a}\,C_m - \underline{a}\,C_M}{C_m + C_M},\ \frac{Q\,\overline{a}}{Q + C_M}\right\}$$

and (34)-(35) follow by simple algebraic manipulations. □

Proof of Lemma 6: By contradiction, let us assume $v^* < \frac{\underline{a}+\overline{a}}{2\,C_P}$. Let $s = P$ which leads to a feasible set $\mathcal{F} = [C_P\,v^*, \overline{a}]$. The diameter of such set is

$$D^* = \overline{a} - C_P\,v^* > \overline{a} - \frac{\underline{a}+\overline{a}}{2} = \frac{\overline{a}-\underline{a}}{2}.$$

Since the optimal worst-case diameter reduction rate is at least 1/2, then v^* cannot be optimal.
Let us now assume $v^* > \overline{v}_P$, i.e., $M < P$. If $m > M$ no reduction can be achieved and hence v^* cannot be optimal. If $m = M$, then only one threshold contributes to reducing the feasible set and the maximum reduction rate is 1/2. By choosing $\widetilde{v} = \frac{\underline{a}+\overline{a}}{2\,C_P} \in V^*$, one gets the worst-case diameter $\widetilde{D} = \overline{a} - C_P\widetilde{v} = \frac{\overline{a}-\underline{a}}{2}$ (for $s = P$) and hence $\widetilde{v}$ is optimal. Finally, let $m < M < P$. By definition of M one has $v^* > \overline{v}_{M+1}$. By (35), it follows that

$$D^* = \overline{a} - C_M\,v^* \geq Q\,v^*.$$

Since $M < P$, C_{M+1} is finite and then

$$D^* = \overline{a} - C_M\,v^* \geq Q\,v^* = (C_{M+1} - C_M)\,v^*.$$

So, $v^* \leq \frac{\overline{a}}{C_{M+1}} = \overline{v}_{M+1}$ which leads to a contradiction. □

Proof of Theorem 3: By Lemma 6, since $v^* \in V^* = \left[\frac{\underline{v}_P+\overline{v}_P}{2}, \overline{v}_P\right]$ one has $M = P$.
First, let us consider the case $m = P$. By Lemma 5 it follows that $v^* = \frac{\underline{a}+\overline{a}}{2\,C_P}$ and $D^* = \frac{\overline{a}-\underline{a}}{2}$. Moreover, by (31) one has $v^* < \underline{v}_{P-1}$ which is equivalent to $\frac{\underline{a}+\overline{a}}{2\,C_P} < \frac{\underline{a}}{C_P-Q}$, and hence $\frac{\underline{a}}{\overline{a}} > \frac{C_P-Q}{C_P+Q}$.
Let us now consider the case $m < P$. By Lemma 5, one has $D^* = \max\{Q\,v^*, C_m\,v^* - \underline{a}\}$. Let us suppose $D^* = C_m\,v^* - \underline{a} \geq Q\,v^*$. Then, if $m > 1$ one can write

$$v^* \geq \frac{\underline{a}}{C_m - Q} = \frac{\underline{a}}{C_{m-1}} = \underline{v}_{m-1}.$$

But this is in contradiction with the definition of m in (31). Hence, it must be $m = 1$. From Lemma 5, by setting $\overline{a} - C_P\,v^* = C_1\,v^* - \underline{a}$ one gets $v^* = \frac{\underline{a}+\overline{a}}{C_1+C_P}$. Moreover, being $v^* \geq \frac{\underline{a}}{C_1-Q}$, one has $\frac{\underline{a}}{\overline{a}} \leq \frac{C_1-Q}{C_P+Q}$.
The remaining case $D^* = Q\,v^* \geq C_m\,v^* - \underline{a}$, gives, by Lemma 5, $\overline{a} - C_P\,v^* = Q\,v^*$, which returns $v^* = \frac{\overline{a}}{C_P+Q}$.
Moreover, one has $\frac{\underline{a}}{\overline{a}} \geq \frac{C_m-Q}{C_P+Q} \geq \frac{C_1-Q}{C_P+Q}$, and, since $v^* \geq \underline{v}_{P-1}$, $\frac{\underline{a}}{\overline{a}} \leq \frac{C_P-Q}{C_P+Q}$. □

Proof of Theorem 4: The expressions of v^* and D^* can be derived by following the same reasoning as in the proof of Theorem 3. Finally, in order to guarantee that the feasible set size does not increase, one has to impose that $D^* \leq \overline{a} - \underline{a}$. Long but straightforward calculations show that this is equivalent to $\frac{\underline{a}}{\overline{a}} \leq \frac{C_P-\delta}{C_P+\delta}$. □

Proof of Lemma 7: For $i = 2, \ldots, P$ one has

$$\frac{\underline{v}_{i-1}}{\overline{v}_i} = \frac{\underline{a}}{\overline{a}}\frac{C_{i-1}+Q}{C_{i-1}} = \frac{\underline{a}}{\overline{a}}\left(1 + \frac{Q}{C_{i-1}}\right) \geq \frac{\underline{a}}{\overline{a}}\left(1 + \frac{Q}{C_{P-1}}\right) = \frac{\underline{v}_{P-1}}{\overline{v}_P}.$$

Therefore, $\overline{v}_P \leq \underline{v}_{P-1}$ implies $\overline{v}_i \leq \underline{v}_{i-1}$ for any $i = 2, \ldots, P$. For any $v \in \mathbb{R}$ and any i, $1 \leq i \leq P$, such that $\underline{a} \leq C_i\,v \leq \overline{a}$ one has:

$$C_{i-1}\,v = \frac{\underline{a}}{\underline{v}_{i-1}}\,v \leq \frac{\underline{a}}{\overline{v}_i}\,v = \frac{\underline{a}}{\overline{a}}\,C_i\,v \leq \underline{a}$$

$$C_{i+1}\,v = \frac{\overline{a}}{\overline{v}_{i+1}}\,v \geq \frac{\overline{a}}{\underline{v}_i}\,v = \frac{\underline{a}}{\overline{a}}\,C_i\,v \geq \overline{a}$$

which proves that the problem is binary design.
To prove necessity, let us assume that $\overline{v}_P > \underline{v}_{P-1}$. Let us apply an input $u = 1/v$ with $v = \frac{\overline{v}_P+\underline{v}_{P-1}}{2}$. One has

$$\overline{a} = C_P\,\overline{v}_P > C_P\,v > C_{P-1}\,v > C_{P-1}\,\underline{v}_{P-1} = \underline{a}.$$

Therefore, this is not a binary design problem. □

Proof of Theorem 5: Since $\frac{\underline{a}_t}{\overline{a}_t}$ is a non decreasing function, if $\frac{\underline{a}_0}{\overline{a}_0} \geq \frac{C_P-Q}{C_P+Q}$ then also $\frac{\underline{a}_t}{\overline{a}_t} \geq \frac{C_P-Q}{C_P+Q}$ for any $t \geq 1$. From Theorem 3, one has that such a condition implies $D_t^* = \frac{\overline{a}_{t-1}-\underline{a}_{t-1}}{2}$, i.e. the 1-step optimal input is the same as in binary design problems. Since this holds for all $t \geq 1$, it follows immediately that (38) holds at any time t. The expression of D_N^* follows directly. □

Proof of Theorem 6: In order to prove the theorem, we first need to introduce two technical lemmas.

Lemma 8 *At a given time t, if $s = P-1$, then the choice of $u^*(t)$ is a binary design problem.*

Proof: Let $u = 1/v$ and let $s = P - 1$. By definition, the feasible set at time t becomes $\mathcal{F} \triangleq [\underline{a}_t, \overline{a}_t] = [\underline{a}_{t-1}, \overline{a}_{t-1}] \bigcap [C_{P-1}\,v, C_P\,v]$. Let us evaluate $\overline{v}_P$ and $\underline{v}_{P-1}$ at time t. One has

$$\overline{v}_P \triangleq \frac{\overline{a}_t}{C_P} = \min\left\{v, \frac{\overline{a}_{t-1}}{C_P}\right\}\ ,\ \underline{v}_{P-1} \triangleq \frac{\underline{a}}{C_{P-1}} = \max\left\{v, \frac{\underline{a}_{t-1}}{C_{P-1}}\right\}.$$

Therefore, one has $\overline{v}_P \leq v \leq \underline{v}_{P-1}$, and hence by Lemma 7 we are facing a binary design problem. □

Lemma 9 *At each time t, an N-step optimal input $u^*(t) = 1/v^*(t)$ satisfies $v^*(t) \leq \overline{v}_P$.*

Proof: Without loss of generality, let us assume $t = 1$ and drop the dependence on time. By contradiction, assume that the optimal input is $\widetilde{u} = 1/\widetilde{v}$, with $\widetilde{v} > \overline{v}_P$. Let $\widetilde{M} = \arg\max\{i : \overline{v}_i \geq \widetilde{v}\}$. By construction, $\widetilde{M} < P$. Now, define $\hat{v} = \dfrac{C_{\widetilde{M}}}{C_P}\,\widetilde{v}$. One has

$$C_P\,\hat{v} \;=\; C_{\widetilde{M}}\,\widetilde{v} \;\leq\; C_{\widetilde{M}}\,\overline{v}_{\widetilde{M}} = \overline{a}$$

i.e., $\hat{v} \leq \overline{v}_P$. We prove the lemma by showing that $\hat{v}$ is always a choice at least as good as $\widetilde{v}$, in the N-step optimal context.
If we set $\hat{m} = \arg\min\{i : \underline{v}_i \leq \hat{v}\}$, the admissible sensor outputs for the input $\hat{u} = 1/\hat{v}$ are $s \in \{\hat{m}-1, \ldots, P\}$. Let us first assume $\hat{m} > 1$. If $C_{\widetilde{M}-1}\,\widetilde{v} \leq \underline{a}$, it is easy to check that all the possible feasible sets resulting from choosing $\hat{v}$ are either equal to or strictly contained within feasible sets resulting from the choice $\widetilde{v}$ (and hence $\widetilde{v}$ cannot be optimal). So, let us assume $C_{\widetilde{M}-1}\,\widetilde{v} > \underline{a}$. The feasible set resulting from the choice $\hat{v}$ is

$$\hat{\mathcal{F}} = \begin{cases} [C_P\,\hat{v},\, \overline{a}] = [C_{\widetilde{M}}\,\widetilde{v},\, \overline{a}] & \text{if } s = P \\ [C_{P-1}\hat{v}, C_P\hat{v}] = [C_{P-1}\,\hat{v}, C_{\widetilde{M}}\,\widetilde{v}] & \text{if } s = P-1 \\ [\max\{\underline{a}, C_{s-1}\,\hat{v}\},\, C_s\,\hat{v}] & \text{if } \hat{m}-1 \leq s < P-1 \end{cases}$$

By Lemma 8, if the sensor returns $s = P-1$ the problem becomes a binary design. Since the size of the feasible set for $\hat{m} - 1 \leq s < P - 1$ is always less than or equal to that corresponding to $P-1$, it can be concluded that the worst case sensor output corresponding to $\hat{v}$ is either P or $P-1$. If the worst-case output for $\hat{v}$ is $s = P$, the feasible set is exactly the same one would obtain by choosing $\widetilde{v}$ with a sensor output $s = \widetilde{M}$. Therefore, $\hat{v}$ is a choice at least as good as $\widetilde{v}$, in the N-step optimal sense. Conversely, let $s = P-1$ be the worst-case output for $\hat{v}$. If by applying $\widetilde{v}$ the sensor would return $s = \widetilde{M} - 1$, the feasible set would be $\widetilde{\mathcal{F}} = [C_{\widetilde{M}-1}\,\widetilde{v},\, C_{\widetilde{M}}\,\widetilde{v}]$, which strictly contains $\hat{\mathcal{F}} = [C_{P-1}\,\hat{v},\, C_{\widetilde{M}}\,\widetilde{v}]$, being

$$C_{P-1}\,\hat{v} = C_{P-1}\,\frac{C_{\widetilde{M}}}{C_P}\,\widetilde{v} > C_{\widetilde{M}-1}\,\widetilde{v}.$$

Hence, $\widetilde{v}$ cannot be the optimal choice.
Finally, let us consider the case $\hat{m} = 1$. If the worst-case sensor output for $\hat{v}$ is $s \neq 0$, one can repeat the same reasoning adopted in the case $\hat{m} > 1$. Otherwise, if $s = 0$, the resulting feasible set is $\hat{\mathcal{F}} = [\underline{a},\, C_1\,\hat{v}]$. However, notice that $s = 0$ is a feasible output also for the choice $\widetilde{v}$, since $\widetilde{v} > \hat{v} \geq \underline{v}_1$ (being $\hat{m} = 1$). The resulting feasible set $[\underline{a},\, C_1\,\widetilde{v}]$ strictly contains $\hat{\mathcal{F}}$ and therefore $\widetilde{v}$ cannot be optimal. This concludes the proof. □
Now we are ready to proove Theorem 6.

Let us first show that for the considered sequence $v^*(t)$, the resulting worst-case diameter at time N is given by (40). When applying $v^*(1)$ in (39), by using (15) it is easy to check that for both $s(1) = P$ and $s(1) = P-1$ one gets $D_1^* = \frac{Q\,\overline{a}_0}{C_P+Q}$ which corresponds to the 1-step optimal feasible set diameter according to Theorem 3. Moreover, the resulting feasible set is such that $\frac{\underline{a}_1}{\overline{a}_1} \geq \frac{C_P-Q}{C_P+Q}$. Hence, by using Theorem 5 for $t = 2, \ldots, N$, (40) holds.
Now, let us prove that $v^*(t)$ is N-step optimal. For input $v^*(1)$, the worst-case sensor output is either $s(1) = P$ or $s(1) = P-1$. In fact, any sensor output $s(1) < P-1$ returns a feasible set not larger than that corresponding to $s(1) = P-1$, which is known to lead to a binary design problem by Lemma 8. Let us denote by $\mathcal{F}^*_{(P)} = [C_P v^*(1), \overline{a}_0]$ and $\mathcal{F}^*_{(P-1)} = [C_{P-1}v^*(1), C_P v^*(1)]$ the feasible sets at time 1 for $s(1) = P$ and $s(1) = P-1$, respectively. Assume by contradiction that the optimal input is such that $\widetilde{v}(1) < v^*(1)$. If $s(1) = P$, one gets the feasible set $\widetilde{\mathcal{F}}_{(P)} = [C_P\,\widetilde{v}(1),\, \overline{a}_0]$. Since $\mathcal{F}^*_{(P)} \subset \widetilde{\mathcal{F}}_{(P)}$, then $\widetilde{v}(1)$ cannot be N-step optimal. Then, let us assume the optimal input is such that $\widetilde{v}(1) > v^*(1)$. If $s(1) = P-1$, the feasible set is $\widetilde{\mathcal{F}}_{(P-1)} = [\underline{a}_0, \overline{a}_0] \bigcap [C_{P-1}\,\widetilde{v}(1),\, C_P\,\widetilde{v}(1)]$. By exploiting the assumption on $\frac{\underline{a}_0}{\overline{a}_0}$, one has

$$C_{P-1}\,\widetilde{v}(1) = (C_P - Q)\,\widetilde{v}(1) > (C_P - Q)\,\frac{\overline{a}_0}{C_P+Q} \geq \underline{a}_0.$$

Moreover, by Lemma 9, $\widetilde{v}(1) \leq \overline{v}_P$ and hence $C_P\,\widetilde{v}(1) \leq \overline{a}_0$. Therefore, $\widetilde{\mathcal{F}}_{(P-1)} = [C_{P-1}\,\widetilde{v}(1),\, C_P\,\widetilde{v}(1)]$, whose diameter is greater than that of $\mathcal{F}^*_{(P-1)}$. By Lemma 8 the resulting problem is a binary design and therefore, for any $t > 1$, the maximum diameter reduction provided by any $\widetilde{v}(t)$ will not exceed that given by $v^*(t)$. Hence $\widetilde{v}(1)$ is not N-step optimal. Then, $v^*(1)$ is N-step optimal, and $v^*(t)$, $t \geq 2$, is N-step optimal by Theorem 5. □

Proof of Theorem 7: Equation (41) follows directly from Theorems 5 and 6. Let $\frac{\underline{a}_0}{\overline{a}_0} \leq \frac{C_1-Q}{C_P+Q}$. Suppose to apply at time 1 the 1-step optimal input reported in Theorem 3, thus obtaining $D_1 = \frac{C_1\overline{a}_0 - C_P\underline{a}_0}{C_1+C_P}$. The worst-case corresponds to a binary design problem for $t \geq 2$ which gives $D_N = \frac{C_1\overline{a}_0 - C_P\underline{a}_0}{C_1+C_P}\left(\frac{1}{2}\right)^{N-1} = \frac{2(C_1\overline{a}_0 - C_P\underline{a}_0)}{C_1+C_P}\left(\frac{1}{2}\right)^{N}$ and then $N = \left\lceil \log_2\left(\frac{2(C_1\overline{a}_0 - C_P\underline{a}_0)}{(C_1+C_P)D_N}\right)\right\rceil$. This is an upper bound, because it is not guaranteed that at time 1 one gets a binary design problem. □