2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Parsimonious System Identification from Quantized Observations

Omar M. Sleem and Constantino M. Lagoa

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683192

***Abstract*— Quantization plays an important role as an interface between analog and digital environments. Since quantization is a many to few mapping, it is a non-linear irreversible process. This made, in addition of the quantization noise signal dependency, the traditional methods of system identification no longer applicable. In this work, we propose a method for parsimonious system identification when only quantized measurements of the output are observable. More precisely, we develop an algorithm that aims at identifying a low order system that is compatible with a priori information on the system and the collected quantized output information. Moreover, the proposed approach can be used even if only fragmented information on the quantized output is available. The proposed algorithm relies on an ADMM approach to ℓ_p quasi-norm optimization. Numerical results highlight the performance of the proposed approach when compared to the ℓ_1 minimization in terms of the sparsity of the induced solution.**

I. INTRODUCTION

A. Motivation

In mathematics and signal processing, quantization is the process of mapping an input signal from an infinite continuous set to a countable set with a finite number of elements [1]. Since most of the information mounting signals, i.e., speech and image, exhibit a continuous and analog nature while their processing requires a digital environment, quantization is considered an important mediator between analog and digital worlds. As a result, traditional system identification techniques are no more applicable when the input signal is subject to quantization [2], [3]. In [4]–[6] and references therein, the authors suggested that the traditional theory of system identification needs to be extended to tackle the fact that the measurements are quantized. This is because the quantization noise can no longer be modeled as a filtered white (zero mean and independent over time) noise in addition to being signal dependent. Moreover, from [7] (section 10.1), the classical identification procedures are not suitable for robust identification because they identify a set of parameters of a fixed mathematical structure, where a fixed system order must be assumed.

Various works –which will be discussed in the next section in more detail– explored the problem of system identification given quantized realizations. However, we aim to study the problem of parsimonious system identification given only quantized realizations from a multi-threshold sensor.

This work was partially supported by National Institutes of Health (NIH) Grant R01 HL142732, and National Science Foundation (NSF) Grant #1808266.

Omar M. Sleem and Constantino M. Lagoa are with the Department of Electrical Engineering, Pennsylvania State University, State College, PA 16801, USA oms46@psu.edu, cml18@psu.edu

B. Related work

The authors in [8] studied the problem of system identification using uniformally quantized realizations. The proposed formulation is a least square minimization of the difference equation errors over all time samples with the system parameters as optimization variables. Despite the prominence of the proposed method in estimating the unknown information in the I/O data, it still suffers the drawback of high computational complexity and noise neglection. The work in [9] accosted these drawbacks by exploiting statistical properties instead of deterministic treatment. In particular, an identification method for a linear system based on quantized measurements was derived. Using traditional equi-spaced quantizer, an instrumental level identification approach was proposed to enhance the estimation accuracy. A variation for the equispaced quantizer was considered in [10], where the authors showed the out-performance of the adoption of a generalized noise shaping coder in terms of the estimation accuracy.

Another line of research includes the identification using a general class of quantized observations that allows the segmentation of the output range into a collection of subsets that may have unequal, fixed lengths or even design variables such as quantization design in communication systems [11]. This serves in favor of understanding the potency of systems with limited sensor information, which in turns rapports the gap between resource limitations and identification complexity in sensor and communication networks. In particular, the work in [12] considered the identification of a gain system by exploiting the information from multiple threshold sensor and the convex combination of these thresholds. The results were extended to the case of a noisy communication channel through which the sensor output information is transmitted. The authors prove that their estimator is asymptotically efficient achieving the Cramer-Rao lower bound. Furthermore, the results were extended to a finite impulse response and transfer function models for periodic bounded input signals.

C. Contributions

Despite the outstanding performance of the different methods proposed for system identification with quantized outputs, none of them addressed the problem of identifying the system of least order that is compatible with collected information. The notions of system complexity and analysis are closely related to its order, i.e., systems with less order are easier to analyze and more amenable for controller design. Moreover, previous work cannot take into account prior information on the system and cannot use fragmented measures of the quantized output.

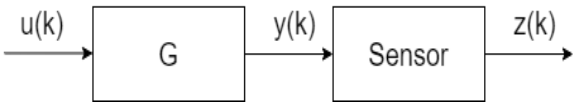

Fig. 1. System model.

Inspired by the results in [13], [14], in this paper we propose an algorithm that addresses the shortcomings mentioned above. More precisely, we assume that the a priori information on the system can be described as constraining the poles of the system to a known compact set. Then, by exploiting "simple representations" of transfer functions, we develop an efficient algorithm that aims at finding the lowest order system that is compatible with fragmented quantized output measurements. This algorithm is based on an ADMM approach to the problem of ℓ_p quasi-norm optimization.

D. Notation

Throughout the paper, bold face letters denote vectors, $\mathbb{R}$ and $\mathbb{C}$ are the sets of real and complex numbers respectively. For any vector $\mathbf{x}$, $\mathbf{x}^{\{.\}}$ is an element-wise power of the elements of $\mathbf{x}$. We use $\preceq$ and $\odot$ for element wise inequality and multiplication of vectors. The $|.|$ operator stands for the absolute value. The p-th norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined such that

$$\|\mathbf{x}\|_p \triangleq (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}. \tag{1}$$

It is important to note that when $0 < p < 1$, the expression in (1) is termed as the quasi-norm satisfying the same axioms of the norm except the triangular inequality making it a non-convex function. For a complex number x, we use $\bar{x}$ to denote the complex conjugate of that number. Finally, let the unit circle be denoted by $\mathbb{D}$.

II. Problem Statement

We consider the system shown in figure 1, where discrete input samples, $u(k)$, on a finite time horizon of length $\mathrm{N} \in \mathbb{Z}_+$ are applied to a Linear Time Invariant (LTI) system G. The output $y(k)$ is measured by an m-levels sensor with thresholds vector $\mathbf{C} = [c_1, \ldots, c_m]^\top \in \mathbb{R}^m$. The sensor acts as a quantizer where the output $\mathbf{z}(k) \in \{0,1\}^m$ is defined such that $z_i(k) = \mathcal{I}(y(k) \geq c_i)$ for $i \in [m]$ and $\mathcal{I}(.)$ is an indicator function equal to 1 if the applied argument is true and 0 otherwise. Given the realizations $\mathbf{z}(k)$, we aim to reconstruct the least order system that is compatible with the input output information.

More formally, the problem we aim to address can be stated as follows

Problem. *Given*

- *Set $\mathbb{D}$ that contains the poles of the LTI system G.*
- *An input sequence $u(k)$, $k \in \{0, 1, \ldots, N-1\}$, which is applied to the system G*
- *Measurements of the binary (quantized) output realizations $z(k)$ for $k \in \mathcal{K} \subseteq \{0, 1, \ldots, N-1\}$.*

find the most parsimonious system that is compatible with the a priori assumptions and a posteriori data mentioned above.

Remark 1. *The formulation above assumes that the only a priori information available on the system is that it is stable. However, any a priori information on the system G that can be represented as constraints on the position of the poles (such as settling time) is compatible with the approach presented in this paper.*

Remark 2. *The formulation in this paper can also be extended to the cases where the intermediate signal $y(k)$ is corrupted by norm bounded noise. For ease of presentation, this is not addressed in this paper.*

A. Parsimonious Identification as a Block Sparsification Problem

Note that the transfer function of any LTI system with poles in $\mathbb{D}$ can represented as

$$G(\mathrm{z}) = r + \sum_{q \in \mathbb{D}} \frac{a_q \mathrm{z}^{-1}}{1 - q\mathrm{z}^{-1}}, \tag{2}$$

with $r \in \mathbb{R}$ and $a_q \in \mathbb{C}$ being the coefficient that is associated with pole q. If the system has repeated poles, then it can be approximated by a system of the form above to an arbitrarily small level of precision.

From the definition of linear systems, the output $y(k)$ can be decomposed as,

$$y(k) = y_{\mathrm{zi}}(k) + y_{\mathrm{zs}}(k), \tag{3}$$

where, $y_{\mathrm{zi}}(k)$ is the zero input response, i.e., the response due to the initial conditions of the system before the input is applied, while $y_{\mathrm{zs}}(k)$ is the zero state response. From [15], the zero input response can be written as,

$$y_{\mathrm{zi}}(k) = \sum_{q \in \mathbb{D}} b_q q^{k-1}, \quad \forall k \in [\mathrm{N}-1], \tag{4}$$

such that, similar to (2), $b_q \in \mathbb{C}$ is the coefficient that is associated to pole q and $y_{\mathrm{zi}}(0) = 0$. The zero state response is obtained by convolving the input sequence with the system's impulse response,

$$y_{\mathrm{zs}}(k) = \sum_{j=0}^{k} u(j)h(k-j), \tag{5}$$

where $h(k)$ is the impulse response of the system. By taking the inverse z-transform of (2), the impulse response can be easily found to be

$$h(k) = \delta(k)r + \sum_{n=1}^{N-1} \delta(k-n) \sum_{q \in \mathbb{D}} a_q q^{n-1} \tag{6}$$

where $\delta(k)$ is the dirac delta functional. In other words, the expression in (6) states that $h(0) = r$ and $h(k) = \sum_{q \in \mathbb{D}} a_q q^{k-1}$ for any other k.

The notions of system complexity and order are always related to the number of poles used to describe the system. The larger number of poles, the higher order and more complex the system is. Hence, in this work, and given only binary realizations of $z(k)$, we aim to reconstruct the corresponding system with the least number of associated

poles. First we define the mapping $\Upsilon : \mathbb{D} \to \mathbb{C}^2$ which maps every pole q to the corresponding coefficients a_q and b_q, i.e., $\Upsilon(q) = \begin{bmatrix} a_q & b_q \end{bmatrix}^\top$. Hence, the problem described above can then be formulated as follows,

$$\min_{\mathbf{a},\mathbf{b},r} \quad \text{Cardinality}\{q \in \mathbb{D} : \Upsilon(q) \neq \mathbf{0}\}, \tag{7a}$$

$$\text{s.t.} \quad y(k) = y_{\text{zi}}(k) + y_{\text{zs}}(k), \tag{7b}$$

$$y_{\text{zi}}(k) = \sum_{q\in\mathbb{D}} b_q q^{k-1}, \tag{7c}$$

$$y_{\text{zs}}(k) = \sum_{j=0}^{k} u(j)h(k-j), \tag{7d}$$

$$h(k) = \delta(k)r + \sum_{n=1}^{N-1} \delta(k-n) \sum_{q\in\mathbb{D}} a_q q^{n-1}, \tag{7e}$$

$$z_i(k) = \mathcal{I}(y(k) \geq c_i), \quad \forall i \in [m],\ \forall k \in \mathcal{K}, \tag{7f}$$

$$a_q = \bar{a}_{\bar{q}}, \quad b_q = \bar{a}_{\bar{q}} \quad \forall q \in \mathbb{D}. \tag{7g}$$

Constraint (7g) implies that the coefficients that are associated with complex conjugate poles have to be complex conjugate as well and $\mathbf{0}$ in (7a) is a vector of zeros in $\mathbb{R}^2$.

III. Proposed Solution

Ideally, we aim to solve the problem in (7), however, this could be practically impossible because the unit circle has infinite number of poles so the computational complexity of the problem would be very high. Therefore, we implement an approximation for the above problem. This approximation is based on gridding the unit circle to a finite number of poles n. It is important to note that the denser the unit circle is grided, the more accurate the approximation is to the original problem. However, this increases the problem's computational complexity and hence, a trade-off exists. First, we define the vector $\mathbf{q}^\top = [q_1, \dots q_n]$, which is composed of complex conjugates and real poles resulted from the gridding effect, and the vectors of associated coefficients $\mathbf{a}^\top = [a_{q_1}, \dots a_{q_n}]$ and $\mathbf{b}^\top = [b_{q_1}, \dots b_{q_n}]$. Second, we define the scaling factor $\boldsymbol{\alpha} \in \mathbb{R}^n$ where,

$$\alpha_i = \frac{1 - |q_i|^2}{1 - |q_i|^{2\text{N}+2}} \quad \forall i \in [n]. \tag{8}$$

For more information on $\boldsymbol{\alpha}$ and its proper choice, the interested reader is recommended to visit [14]. Then, a proper approximation for the original problem in (7) can be defined such that we solve;

$$\min_{\mathbf{a},\mathbf{b},r,\mathbf{d}} \quad \|\mathbf{d}\|_0 \tag{9a}$$

$$\text{s.t.} \quad y(k) = (\boldsymbol{\alpha} \odot \mathbf{b})^\top \mathbf{q}^{\{k-1\}} + \sum_{j=0}^{k} u(j)h(k-j) \tag{9b}$$

$$h(k) = \delta(k)r + \sum_{n=1}^{N-1} \delta(k-n)(\boldsymbol{\alpha} \odot \mathbf{a})^\top \mathbf{q}^{\{n-1\}}, \tag{9c}$$

$$z_i(k) = \mathcal{I}(y(k) \geq c_i), \quad \forall i \in [m],\ \forall k \in \mathcal{K}, \tag{9d}$$

$$a_{q_i} = \bar{a}_{\bar{q}_i}, \quad b_{q_i} = \bar{b}_{\bar{q}_i} \quad \forall i \in [n], \tag{9e}$$

$$|\mathbf{a}| \preceq \mathbf{d}, \quad |\mathbf{b}| \preceq \mathbf{d}, \tag{9f}$$

where $\|.\|_0$ is the ℓ_0 pseudo-norm which counts the number of non-zero elements of the applied argument while $\mathbf{d}$ ensures block sparsity of the zero state and zero input coefficients, i.e., $\mathbf{a}$ and $\mathbf{b}$. A proper choice of the vector $\boldsymbol{\alpha}$, defined in (8), and the use of (9a) and (9f) allow the identification of the system with the least number of poles, i.e., least order system. However, the ℓ_0 pseudo-norm is an NP hard problem and hence, using notions of sparsity [13], the objective function is relaxed using the $\ell_p (0 < p < 1)$ quasi-norm, i.e., $\|\mathbf{d}\|_0$ in (9a) is replaced with $\|\mathbf{d}\|_p^p$ defined as in (1).

For notation simplicity, we define the vector $\mathbf{w} \in \mathbb{R}^{2n+1}$, which is the concatenation of the variables $\mathbf{a}, \mathbf{b}$ and r. Let the set $\mathcal{D} \subseteq \mathbb{R}^{2n+1} \times \mathbb{R}^n$ as the set of doubles $(\mathbf{w}, \mathbf{d})$ where constraints (9b) to (9f) are satisfied. Hence, the problem in (9), after the objective function relaxation, will have the compact representation in the form;

$$\min_{\mathbf{w},\mathbf{d}} \quad \|\mathbf{d}\|_p^p, \tag{10a}$$

$$\text{s.t.} \quad \mathbf{w}, \mathbf{d} \in \mathcal{D}. \tag{10b}$$

Since we are interested in the recovery of the lowest order system, we consider the case when $0 < p < 1$, i.e., specifically, $p = 0.5$, where the objective function (10a) is a non-convex one. In our solution approach, we consider an ADMM algorithm which exploits the structure of the problem to split the optimization over the variables via iteratively solving fairly simple sub-problems. We first start with the epi-graph form of (10) by introducing the variable $\mathbf{t} \in \mathbb{R}^n$, where,

$$\begin{aligned} \min_{\mathbf{w},\mathbf{d},\mathbf{t}} \quad & \mathbf{1}^\top \mathbf{t}, \\ \text{s.t.} \quad & t_i \geq |d_i|^p, \quad i \in [n], \\ & \mathbf{w}, \mathbf{d} \in \mathcal{D}, \end{aligned} \tag{11}$$

where $\mathbf{1}$ is a vector of all ones. Let the non-convex set $\mathcal{X} \subset \mathbb{R}^2$ be the epigraph of the scalar function $|d|^p$, i.e., $\mathcal{X} = \{(d,t) \in \mathbb{R}^2 : t \geq |d|^p\}$. Then, (11) can be cast as

$$\begin{aligned} \min_{\mathbf{w},\mathbf{d},\mathbf{t}} \quad & \sum_{i\in[n]} g_{\mathcal{X}}(d_i, t_i) + \mathbf{1}^\top \mathbf{t}, \\ \text{s.t.} \quad & \mathbf{w}, \mathbf{d} \in \mathcal{D} \end{aligned} \tag{12}$$

where $g_{\mathcal{X}}(.)$ is the indicator function to the set $\mathcal{X}$, i.e., it evaluates to zero if its argument belongs to the set $\mathcal{X}$ and is $+\infty$ otherwise. In particular, we introduce the auxiliary variables $\mathbf{s} \in \mathbb{R}^{2n+1}$, $\mathbf{f}$ and $\mathbf{z} \in \mathbb{R}^n$. An equivalent ADMM formulation of (12) can be then given by:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{d},\mathbf{t},\mathbf{s},\mathbf{f},\mathbf{z}} \quad & \sum_{i\in[n]} g_{\mathcal{X}}(d_i, t_i) + g_{\mathcal{D}}(\mathbf{s}, \mathbf{f}) + \mathbf{1}^\top \mathbf{z}, \\ \text{s.t.} \quad & \mathbf{w} = \mathbf{s} : \quad \boldsymbol{\lambda}_1, \\ & \mathbf{d} = \mathbf{f} : \quad \boldsymbol{\lambda}_2, \\ & \mathbf{t} = \mathbf{z} : \quad \boldsymbol{\theta}. \end{aligned} \tag{13}$$

The dual variables associated with the constraints $\mathbf{w} = \mathbf{s}$, $\mathbf{d} = \mathbf{f}$ and $\mathbf{t} = \mathbf{z}$ are $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ and $\boldsymbol{\theta}$, respectively. Hence, the Lagrangian function corresponding to (13) augmented with a quadratic penalty on the violation of the equality constraints

with penalty parameter $\rho > 0$, is given by:

$$\mathcal{L}_\rho(\mathbf{d},\mathbf{t},\mathbf{s},\mathbf{f},\mathbf{w},\mathbf{z},\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\boldsymbol{\theta}) = \sum_{i\in[n]} g_{\mathcal{X}}(d_i,t_i) + g_{\mathcal{D}}(\mathbf{s},\mathbf{f}) + \mathbf{1}^\top\mathbf{z} + \boldsymbol{\lambda}_1^\top(\mathbf{w}-\mathbf{s}) + \boldsymbol{\lambda}_2^\top(\mathbf{d}-\mathbf{f}) + \boldsymbol{\theta}^\top(\mathbf{t}-\mathbf{z}) + \frac{\rho}{2}(\|\mathbf{w}-\mathbf{s}\|_2^2 + \|\mathbf{d}-\mathbf{f}\|_2^2 + \|\mathbf{t}-\mathbf{z}\|_2^2). \quad (14)$$

Considering the three block variables $\mathbf{Q}_1 = (\mathbf{d},\mathbf{t})$, $\mathbf{Q}_2 = (\mathbf{s},\mathbf{f})$ and $\mathbf{Q}_3 = (\mathbf{w},\mathbf{z})$, ADMM [16] consists of the following iterations, where l is the iteration number:

$$\mathbf{Q}_1^{(l+1)} = \underset{\mathbf{d},\mathbf{t}}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{Q}_1,\mathbf{Q}_2^{(l)},\mathbf{Q}_3^{(l)},\boldsymbol{\lambda}_1^{(l)},\boldsymbol{\lambda}_2^{(l)},\boldsymbol{\theta}^{(l)}), \quad (15)$$

$$\mathbf{Q}_2^{(l+1)} = \underset{\mathbf{s},\mathbf{f}}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{Q}_1^{(l+1)},\mathbf{Q}_2,\mathbf{Q}_3^{(l)},\boldsymbol{\lambda}_1^{(l)},\boldsymbol{\lambda}_2^{(l)},\boldsymbol{\theta}^{(l)}), \quad (16)$$

$$\mathbf{Q}_3^{(l+1)} = \underset{\mathbf{w},\mathbf{z}}{\operatorname{argmin}}\ \mathcal{L}_\rho(\mathbf{Q}_1^{(l+1)},\mathbf{Q}_2^{(l+1)},\mathbf{Q}_3,\boldsymbol{\lambda}_1^{(l)},\boldsymbol{\lambda}_2^{(l)},\boldsymbol{\theta}^{(l)}), \quad (17)$$

$$\boldsymbol{\lambda}_1^{(l+1)} = \boldsymbol{\lambda}_1^{(l)} + \rho(\mathbf{w}^{(l+1)} - \mathbf{s}^{(l+1)}), \quad (18)$$

$$\boldsymbol{\lambda}_2^{(l+1)} = \boldsymbol{\lambda}_2^{(l)} + \rho(\mathbf{d}^{(l+1)} - \mathbf{f}^{(l+1)}), \quad (19)$$

$$\boldsymbol{\theta}_1^{(l+1)} = \boldsymbol{\theta}_1^{(l)} + \rho(\mathbf{t}^{(l+1)} - \mathbf{z}^{(l+1)}). \quad (20)$$

A. ($\mathbf{d}$, $\mathbf{t}$) *update*

From the expression of the augmented Lagrangian in (14) and by completing the square, the update of $\mathbf{d}$ and $\mathbf{t}$ in (15) can be found by solving the following optimization,

$$\begin{aligned} \min_{\mathbf{d},\mathbf{t}} \quad & \|\mathbf{d} - (\mathbf{f}^{(l)} - \frac{\boldsymbol{\lambda}_2^{(l)}}{\rho})\|_2^2 + \|\mathbf{t} - (\mathbf{z}^{(l)} - \frac{\boldsymbol{\theta}^{(l)}}{\rho})\|_2^2, \\ \text{s.t.} \quad & (d_i,t_i) \in \mathcal{X} \quad \forall i \in [n]. \end{aligned} \quad (21)$$

It can be realized that the problem in (21) enjoys a separable structure and hence is amenable to decentralization. However, it is a non-convex problem due to the nature of the set $\mathcal{X}$. In [17], the authors considered a similar problem and it was shown that the element-wise optimization of (21) boils down to finding the roots, a_i^*, of the scalar $2v$ polynomial;

$$a_i^{2v} + \frac{u}{v}\left(a_i^{2u} - \tilde{t}_i a_i^u\right) - \tilde{x}_i a_i^v, \quad (22)$$

where $\tilde{x}_i = f_i^{(l)} - \frac{\lambda_{i,2}^{(l)}}{\rho}$, $\tilde{t}_i = z_i^{(l)} - \frac{\theta_i^{(l)}}{\rho}$ and $u, v \in \mathbb{Z}_+$ such that $p = u/v$. They showed that, in proposition 1, the entry-wise solution of (21) is given by $(d_i^*, t_i^*) = (a_i^{*^v}, a_i^{*^u})$ for all $i \in [n]$.

B. ($\mathbf{s}$, $\mathbf{f}$) *update*

By fixing all the remaining variables, the ($\mathbf{s}$, $\mathbf{f}$) update in (16) can be easily shown to be the solution of the following optimization problem;

$$\begin{aligned} \min_{\mathbf{s},\mathbf{f}} \quad & \|\mathbf{s} - (\mathbf{w}^{(l)} + \frac{\boldsymbol{\lambda}_1^{(l)}}{\rho})\|_2^2 + \|\mathbf{f} - (\mathbf{d}^{(l+1)} + \frac{\boldsymbol{\lambda}_2^{(l)}}{\rho})\|_2^2, \\ \text{s.t.} \quad & (\mathbf{s},\mathbf{f}) \in \mathcal{D}. \end{aligned} \quad (23)$$

Algorithm 1 ADMM algorithm

1: Initialize: $\mathbf{w}$, $\mathbf{z}$, $\mathbf{s}$, $\mathbf{f}$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\boldsymbol{\theta}$, ρ, $k = 0$, $v = 1$, $u = 2$.
2: **repeat**
3: **for** $i \in [n]$ **do**
4: solve $a_i^{2v} + \frac{u}{v}\left(a_i^{2u} - \tilde{t}_i a_i^u\right) - \tilde{x}_i a_i^v = 0$
5: $(d_i^{(l+1)}, t_i^{(l+1)}) = (a_i^{*^v}, a_i^{*^u})$
6: $\hat{\mathbf{d}} = \mathbf{d}^{(l+1)} + \frac{\boldsymbol{\lambda}_2^{(l)}}{\rho}$, $\hat{\mathbf{w}} = \mathbf{w}^{(l)} + \frac{\boldsymbol{\lambda}_1^{(l)}}{\rho}$
7: $(\mathbf{s}^{(l+1)}, \mathbf{f}^{(l+1)}) = \underset{\mathbf{s},\mathbf{f}\in\mathcal{D}}{\operatorname{argmin}} \|\mathbf{s} - \hat{\mathbf{w}}\|_2^2 + \|\mathbf{f} - \hat{\mathbf{d}}\|_2^2$
8: $\mathbf{w}^{(l+1)} = \mathbf{s}^{(l+1)} - \frac{\boldsymbol{\lambda}_1^{(l)}}{\rho}$
9: $\mathbf{z}^{(l+1)} = \mathbf{t}^{(l+1)} + \frac{\boldsymbol{\theta}^{(l)} - \mathbf{1}}{\rho}$
10: $\boldsymbol{\lambda}_1^{(l+1)} = \boldsymbol{\lambda}_1^{(l)} + \rho(\mathbf{w}^{(l+1)} - \mathbf{s}^{(l+1)})$
11: $\boldsymbol{\lambda}_2^{(l+1)} = \boldsymbol{\lambda}_2^{(l)} + \rho(\mathbf{d}^{(l+1)} - \mathbf{f}^{(l+1)})$
12: $\boldsymbol{\theta}_1^{(l+1)} = \boldsymbol{\theta}_1^{(l)} + \rho(\mathbf{t}^{(l+1)} - \mathbf{z}^{(l+1)})$
13: $l = l + 1$
14: **until** convergence

The problem in (23) is clearly a convex optimization one that can be solved by various methods including sub-gradient projection [18], interior point and ellipsoid methods [19], [20].

C. ($\mathbf{w}$, $\mathbf{z}$) *update*

From the Lagrangian expression in (14), the $\mathbf{w}$ update can be found by solving;

$$\begin{aligned} \mathbf{w}^{(l+1)} &= \underset{\mathbf{w}}{\operatorname{argmin}} \|\mathbf{w} - (\mathbf{s}^{(l+1)} - \frac{\boldsymbol{\lambda}_1^{(l)}}{\rho})\|_2^2 \\ &= \mathbf{s}^{(l+1)} - \frac{\boldsymbol{\lambda}_1^{(l)}}{\rho}, \end{aligned} \quad (24)$$

while that of $\mathbf{z}$ is given by;

$$\begin{aligned} \mathbf{z}^{(l+1)} &= \underset{\mathbf{z}}{\operatorname{argmin}}\ \mathbf{1}^\top\mathbf{z} + \boldsymbol{\theta}^{(l)^\top}(\mathbf{t}^{(l+1)} - \mathbf{z}) + \frac{\rho}{2}\|\mathbf{t}^{(l+1)} - \mathbf{z}\|_2^2 \\ &= \mathbf{t}^{(l+1)} + \frac{\boldsymbol{\theta}^{(l)} - \mathbf{1}}{\rho}. \end{aligned} \quad (25)$$

The steps of the ADMM algorithm described in the previous sections can then be summarized as in algorithm 1.

IV. Numerical Results

In this section, we test the validity of the algorithm 1 for solving the problem in (10). For comparison purpose, we use the solution of (10) with an ℓ_1 relaxation in the objective function as a baseline. Different algorithms discussed I-B were not used for comparison as none of them can handle stability and data fragmentation constraints. We perform two different experiments: 1) A single system is considered and

different properties from ℓ_1 and ℓ_p relaxations are compared. 2) Multiple systems with same original order are generated and the different statistical properties are studied. In the next parts, we will use the notions of ℓ_p and $\ell_{0.5}$ interchangeably.

A. Experimental setup

We let the input horizon length $\mathrm{N} = 40$ samples, where the input samples are drawn independently from a standard Gaussian distribution. For the sake of simplicity, we assume that the number of sensor threshold levels, $m = 3$ and the unit circle is grided uniformly into $n = 146$ points. As mentioned before in section III, the more dense the unit circle is, the better the system is represented but the more complex it will be. From [21], our choice seems to be a good approximation. The threshold values are chosen such that they partition the system output (sensor input) range into m-equally sized intervals. This is to ensure that no singular cases dominate the simulations, however, any other choice for the threshold values is applicable as well. Hence, for $m = 3$, we have $c_1 = \min_k\{y(k)\}+\tilde{\epsilon}$, $c_3 = \max_k\{y(k)\}-\tilde{\epsilon}$, where $\tilde{\epsilon}$ is a very small scalar, and $c_2 = c_1 + \frac{c_3-c_1}{2}$. All the other parameters in step 1 of algorithm 1 are initialized through samples from a Gaussian distribution of zero mean and 10^{-1} standard deviation. The value of ρ is set to 20. We define a threshold ϵ as the value below which, a vector entry is considered zero. The value of the threshold ϵ is chosen such that it is less than 0.5% of the maximum value of the optimal vector $\mathbf{d}$, which makes $\epsilon = 10^{-3}$ a good choice.

B. Single system experiment

In this subsection, we consider the experiment where an input is applied to a stable randomly generated system of order 10. The output of this system is applied to a 3-thresholds sensor with values; $-7.9, -1.9$ and 3.9. The values of these thresholds are calculated as discussed in the previous subsection. Given the binary sensor outputs, the problem is solved via ℓ_1 and ℓ_p relaxations and the detected system orders and outputs are compared.

Figure 2 plots the original system poles vs those that are associated with the non zero coefficients in the vectors $\mathbf{a}$ and $\mathbf{b}$ from the ℓ_1 and ℓ_p relaxations' solutions. From the figure, it can be concluded that the ℓ_p detected a system of order 7 which is less than that of the ℓ_1 of detected order 12. This outlines the out-performance of the ℓ_p quasi-norm when compared to the ℓ_1 convex relaxation.

In figure 3, we plot the system output $y(k)$, i.e. sensor input, vs the considered time horizon. It shows how accurate the considered relaxations, whether ℓ_1 or ℓ_p, can represent the original output. We define the representation error across

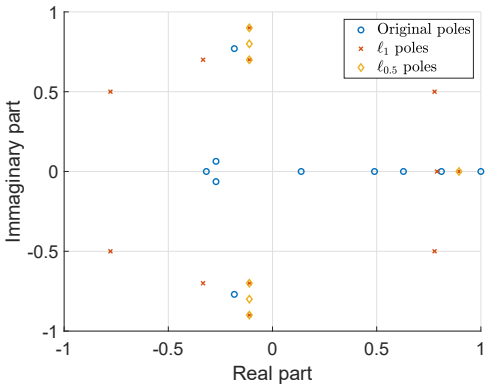

Fig. 2. System poles.

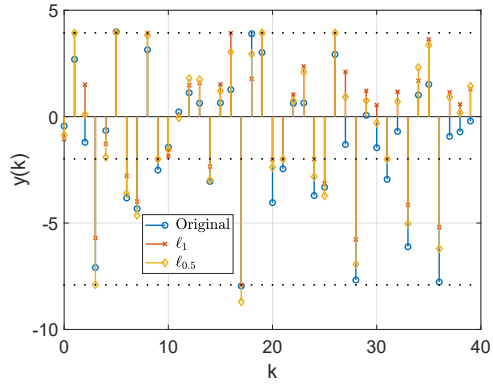

Fig. 3. System output. The dotted lines indicate the used sensor thresholds.

a time horizon of length N as, $\zeta_x, x \in \{\ell_1, \ell_p\}$ where;

$$\zeta_x = \sum_{i=0}^{\mathrm{N}-1} (y(i) - y_x(i))^2, \tag{26}$$

with $y(i)$ is the output from the original system. For the ℓ_p relaxation, the representation error ζ_{ℓ_p} was found to be equal 6.85 which is less than that of the ℓ_1 convex relaxation which had a value $\zeta_{\ell_1} = 9.4$. It is important to note that we are not interested in perfectly fitting the original system's output. However, we aim to fit the sensor's binary realizations. In figure 3, the sensor levels are indicated by the dotted horizontal lines. It can be clearly realized that at all time instances, the outputs from the original system, ℓ_1 and ℓ_p relaxations all lie in the same range of sensor thresholds which indicates that the binary senor outputs from both relaxations are the same as the original ones.

C. Multiple system experiment

Since the systems that we generate to validate our solution method are random, the main idea in this part is to study the

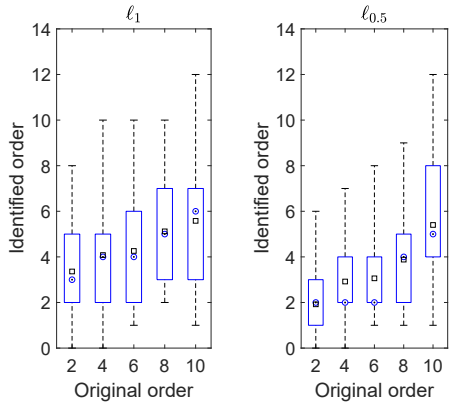

Fig. 4. Box plot for the system order statistics. Circles with dots and black squares indicate the median and mean values respectively. Bottom/top edges of the boxes are the 25th/75th quantile. The whiskers extend from the minimum (downwards) to the maximum (upwards) value.

statistical properties of the derived algorithm solution. We perform an experiment where for a given original order, 50 random systems are generated. For each system, the same input is applied and the identification problem in (10) is solved, using the ℓ_1 norm and $\ell_{0.5}$ quasi-norm relaxations, given the quantized realizations from the sensor output.

Figure 4 outlines the different statistical properties from the ℓ_1 and $\ell_{0.5}$ relaxations. It can be realized that for all original system orders, the $\ell_{0.5}$ relaxation solution enjoys less mean and median values than its counterpart, i.e., ℓ_1 relaxation. Moreover, the $\ell_{0.5}$ relaxation has a maximum value for each original order that is less than that of the ℓ_1, expect for an order of 10 when the maximum values for both are the same. It can be realized that in either cases, some systems have a detected order of zero, i.e., the minimum value of the whisker is zero, which means that the estimation of the constant r in (2) is enough to describe the I/O relationship. Finally, some systems are detected with higher order than the original, this because the $\ell_{0.5}$ minimization is a non convex problem and hence algorithm 1 converges to a local minimum. Moreover, it motivates that the unit circle should be gridded into more points to increase precision, i.e., $n > 146$ mentioned in IV-A, in expense of computational complexity.

V. Conclusion

In this paper we have presented a new approach to the problem of LTI system identification from quantized outputs. This approach allows for the use of a priori information on the system and fragmented measurements of the output. The algorithm described uses an ADMM approach to the problem of ℓ_p quasi-norm minimization. Numerical results presented show that the algorithm is very effective in obtaining low complexity explanations of the data collected. Effort is being put on the improvement of the numerical performance of the algorithm and its extension to continuous-time systems.

References

[1] A. Gersho, "Quantization," *IEEE Communications Society Magazine*, vol. 15, no. 5, pp. 16–16, 1977.

[2] K. Tsumura and J. Maciejowski, "Optimal quantization of signals for system identification," in *2003 European Control Conference (ECC)*, pp. 785–790, 2003.

[3] H. Suzuki and T. Sugie, "System identification based on quantized i/o data corrupted with noises," in *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, 2006.

[4] F. Gustafsson and R. Karlsson, "Statistical results for system identification based on quantized observations," *Automatica*, vol. 45, no. 12, pp. 2794–2801, 2009.

[5] G. G. Yin, J.-F. Zhang, *et al.*, "System identification using quantized data," *IFAC Proceedings Volumes*, vol. 39, no. 1, pp. 255–260, 2006.

[6] Y. Zhao, G. G. Yin, J.-F. Zhang, *et al.*, "Identification of wiener systems with binary-valued output observations," *Automatica*, vol. 43, no. 10, pp. 1752–1765, 2007.

[7] R. S. Sanchez-Pena and M. Sznaier, *Robust systems theory and applications*. John Wiley & Sons, Inc., 1998.

[8] A. Okao, M. Ikeda, and R. Takahashi, "System identification for nanocontrol: A finite wordlength problem," in *Proceedings of 2003 IEEE Conference on Control Applications, 2003. CCA 2003.*, vol. 1, pp. 49–53 vol.1, 2003.

[9] H. Suzuki and T. Sugie, "System identification based on quantized i/o data corrupted with noises and its performance improvement," in *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 3684–3689, 2006.

[10] J. C. Aguero, G. C. Goodwin, and J. I. Yuz, "System identification using quantized data," in *2007 46th IEEE Conference on Decision and Control*, pp. 4263–4268, 2007.

[11] L. Y. Wang, G. G. Yin, J.-F. Zhang, and Y. Zhao, *System identification with quantized observations*. Springer, 2010.

[12] G. G. Yin *et al.*, "Asymptotically efficient parameter estimation using quantized output observations," *Automatica*, vol. 43, no. 7, pp. 1178–1191, 2007.

[13] P. Shah, B. N. Bhaskar, G. Tang, and B. Recht, "Linear system identification via atomic norm regularization," in *2012 IEEE 51st IEEE conference on decision and control (CDC)*, pp. 6265–6270, IEEE, 2012.

[14] B. Yılmaz, K. Bekiroglu, C. Lagoa, and M. Sznaier, "A randomized algorithm for parsimonious model identification," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 532–539, 2018.

[15] B. P. Lathi and R. A. Green, *Linear systems and signals*, vol. 2. Oxford University Press New York, 2005.

[16] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[17] M. E. Ashour, C. M. Lagoa, and N. S. Aybat, "Lp quasi-norm minimization," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 726–730, 2019.

[18] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[19] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[20] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[21] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, vol. 6, pp. 4734–4739, IEEE, 2001.