STATE ESTIMATION AND OBSERVABILITY OF SYSTEMS OVER FINITE ALPHABETS

by

Donglei Fan

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March, 2016

Abstract

In this dissertation, the state estimation problem for systems over finite alphabets is studied, focusing in particular on a significant special instance of such systems consisting of an LTI system with a finite input set and an output quantizer. The need for new notions of observability is motivated, and a set of new notions of observability are formulated quantifying the degree to which the output of such systems can be predicted by an observer. The characterization of observability is investigated, with both necessary and sufficient conditions derived in terms of the dynamics of the system, the properties of the quantizer, and the finite alphabet sets. The use of deterministic finite state machine as observers is also explored, with a view towards understanding their advantages and limitations. Building on the notion of finite memory observability, a control design problem is formulated. Lastly, an idea inspired by the characterization of observability is applied to solve some remaining open questions in the theory of bisimulation.

Thesis Advisor: Dr. Danielle C. Tarraf, Assistant Professor of ECE

Second Reader: Dr. Mounya Elhilali, Associate Professor of ECE

ACKNOWLEDGMENTS

I thank my advisor Dr. Danielle C. Tarraf for her inspirational guidance and her instrumental advices. I also thank my committee members, Dr. Mounya Elhilali and Dr. Pablo Iglesias, for their thoughtful feedback. This dissertation is dedicated to my family. I thank my parents for their unconditional support during my student years. I thank my daughter Olivia for bringing such great joy into our lives. Lastly but most importantly, I would not be able to achieve this work without my wife Ying's company.

This research was supported by AFOSR grant FA9550-16-1-0132, NSF CAREER grant 0954601 and AFOSR Young Investigator grant FA9550-11-1-0118.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Overview

In many modern control applications, the dynamics of the physical system being controlled is continuous, while the digital system that controls the physical system is discrete. A *finite quantizer* is an interface between the physical system and the digital controller that maps real numbers to "words" from a finite alphabet. Research on hybrid systems, which seeks an overarching theory regarding both continuous and discrete systems, has been very extensive in the recent years [1–6].

In this dissertation, we study a particular class of hybrid systems, namely *systems over finite alphabets*, which were first introduced in [7]. Specifically, a system over finite alphabet is a discrete-time system P described as

$$P \subset \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}, \text{ with } |\mathcal{U}| < \infty, |\mathcal{Y}| < \infty. \tag{1}$$

Here $|\mathcal{U}|$ is the cardinality of $\mathcal{U}$. $|\mathcal{U}| < \infty, |\mathcal{Y}| < \infty$ indicates that $\mathcal{U}, \mathcal{Y}$ are finite sets. Essentially, P is a set of pairs of signals over some finite alphabets. Although not required in the definition, we assume P has some underlying continuous dynamics, as we shall see next.

While the definition of P given in (1) is quite general, in many parts of this dissertation, we focus on a specific class of systems P whose internal dynamics are described by:

$$x_{t+1} = Ax_t + Bu_t \tag{2a}$$

$$y_t = Cx_t + Du_t \tag{2b}$$

$$\tilde{y}_t = Q(y_t) \tag{2c}$$

where $t \in \mathbb{N}$ is the time index, $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathcal{U}$ is the input, $y_t \in \mathbb{R}^p$ is the

output of the underlying physical system, and $\tilde{y}_t \in \mathcal{Y}$ is the quantized output. We assume that $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$ are given. Finite sets $\mathcal{U} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}^p$ represent the possible values of the finite-valued input and the quantized output respectively. The quantizer $Q : \mathbb{R}^p \to \mathcal{Y}$ is a piecewise-constant function. Note that in this case, the physical state x_t of the system is continuous, while the input and output of the system are discrete, and take values from finite alphabets. We think systems modeled as in (2) could represent significant applications in practice.

An omnipresent problem in control theory is to design the control input of a system such that the system behavior satisfies certain prescribed objectives. In most cases, knowledge of the system state is crucial for designing a feedback control law that solves this problem. However, the system state may not always be available: Rather, only the measurement output, which contains partial information of the state, is available. For instance, considering system (2), this situation corresponds to x_t not being available but $\tilde{y}_t$ being available. Therefore, in circumstances where the system state is not available for control design, it is necessary for an estimate of the state to be generated based on the partial information obtained from the measurement output. This brings us to the problem of *state estimation*. Another closely related concept in control theory is *observability*. Generally, observability refers to a property of the system whereby the initial state of the system can be uniquely determined from a single observation of its input and output over some finite time interval. In this dissertation, we study state estimation and observability of systems in the forms of (1) and (2).

At this point, we think it is a good idea to first briefly review observability and state estimation in the "traditional" setting where linear time-invariant (LTI) systems are considered. An LTI system described by (2a) and (2b) is said to be observable if different initial states x_0 produce different outputs under zero input (pp. 137, [8]). Observability is characterized by the pair (C, A) satisfying certain rank condition (pp. 144, [8]). If an LTI system is observable (the rank conditions is satisfied), then there is a state estimate generated by a Luenberger observer (pp. 141, [9]) that converges to the actual system state exponentially

fast.

As we shall see in Chapter 3, the traditional concept of observability does not generalize well to systems in the form of (2). Therefore, the question we wish to answer in this dissertation can be phrased in this most basic form: Under what circumstances and in what sense can we estimate the states of systems (1) and (2)? Moreover, for system (2), how does the answer to this question depend on properties of the underlying LTI system and on properties of the quantizer?

As a final point of interest, we study deterministic finite state machines (DFM) as observers for systems over finite alphabets. As inferred from its name, an *observer* for a system generates an estimate of the system state based on observations of the input and output of the system. Since the input and output values of system (1) are drawn from finite alphabets, we are curious about the question: What generality, if any, is lost when we restrict the class of observers of system (1) to the set of DFM?

1.2 Summary of Thesis Contributions

We summarize the contributions of this dissertation as follows.

- We propose notions of (output) observability for systems over finite alphabets, and then characterize conditions for the proposed notions for both systems (1) and (2). This work is presented in Chapter 3.
- We discuss the construction of DFM observers in Chapter 4.
- We formulate and solve a control design based on the proposed notion of finite memory observability in Chapter 5.
- We study the initialization process of an existing construction of finite state approximation, with the goal of characterizing instances in which it is possible to reduce its complexity. This work is presented in Chapter 6.

- We apply our technical results to address an open problem on the theory of bisimulation in Chapter 7, presenting a new topological approach.

1.3 Notation

We introduce the notation used in this dissertation in this section. We use $\mathbb{N}$ to denote the nonnegative integers, $\mathbb{Z}_+$ to denote the positive integers, $\mathbb{R}$ to denote the reals, and $\mathbb{C}$ to denote the complex numbers. For $\alpha \in \mathbb{R}^n$, we use $|\alpha|$ to denote the Euclidean norm of α. For $v \in \mathbb{C}^n$, use $[v]_i$ to denote its i-th component. For $w \in \mathbb{C}^n$, we use $Re(w)$ to denote the real part of w. We use the notation $[\alpha, \beta)$ to denote the interval $\{x \in \mathbb{R} : \alpha \leq x < \beta\}$ for $\alpha, \beta \in \mathbb{R}$. For two positive integers a, b, we use $a \mod b$ to denote the remainder of the division of a by b. For sets $\mathcal{A}$, $\mathcal{B}$ in $\mathbb{R}^n$, we use $|\mathcal{A}|$ to denote the cardinality of set $\mathcal{A}$ and $d(\mathcal{A}, \mathcal{B}) = \inf\{|\alpha - \beta| : \alpha \in \mathcal{A}, \beta \in \mathcal{B}\}$ to denote the distance between sets $\mathcal{A}$ and $\mathcal{B}$. We use $\mathcal{A}^{\mathbb{N}}$ to denote the collection of infinite sequences over $\mathcal{A}$: $\mathcal{A}^{\mathbb{N}} = \{f : \mathbb{N} \to \mathcal{A}\}$, and use the bold font $\mathbf{a}$ to denote elements in $\mathcal{A}^{\mathbb{N}}$: $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$. For $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$, we use a_t to denote its t^{th} component. Given a square matrix A, we use $\rho(A)$ to denote its spectral radius. We use $\mathbf{0}$ to denote the zero matrix of appropriate dimensions. For two functions f and g, we use $f \circ g$ to denote the composition of f and g.

Relevant mathematical concepts are reviewed in Section 2.2.2.

2 Background

Previous works on observability, state estimation, and observer design, especially when the system input involves switching or the system output is quantized, are related to the current research. Previous works on systems over finite alphabets are also related to the proposed research. We review the relevant literature on both topics in this chapter. We review relevant mathematical concepts and results in this chapter as well.

2.1 Related Research

The thesis work is based on and closely related to the robust control inspired design framework for systems over finite alphabets developed over the past few years [7] [10] [11] [12]. Specifically, Tarraf et al. developed a constructive procedure to synthesize finite memory controllers for systems over finite alphabets. The idea is to first construct a sequence of deterministic finite state machines (DFM) to approximate the original system over finite alphabets. Such finite state approximations should satisfy a set of well-defined properties. Next, these approximate models are used as the basis for certified-by-design control synthesis [11]: A full state feedback control law is first designed for the approximate model, to achieve a suitably defined auxiliary performance objective. This control law is then used, together with a copy of the approximate model serving as a finite memory observer of the plant, to certifiably close the loop around the system. Note that DFM's are used in this setting as common models of both dynamics and computing processes. The proposed research is also closely related to [12]. Specifically, we can construct a finite state approximation by associating each of its states with a sequence of feasible input and output signals of the original system. Therefore each state of the DFM approximation corresponds to a set-valued estimate of the state of the original system. If for certain systems, this set-valued estimate is good enough in the sense that the corresponding error system is gain stable, then this DFM approximation satisfies one of the desired conditions in [10].

The literature on observability of dynamical systems is also closely related to the pro-

posed research, since the term “observability" refers to the ability to determine the system state from the input and output signals measured over some time interval [8]. Recall, in particular, that an LTI system is observable if and only if different initial states produce different outputs under zero input. Similarly, a nonlinear system is locally observable at x_o if there is some neighborhood of x_o such that x_o and any other state in the neighborhood of x_o, as initial states of the system, can be distinguished [13].

A more closely related line of research is on observability of switched systems, because the action of control inputs within a finite set here can be understood as switching between a finite number of given systems. Observability of switched linear systems has been studied in recent years [14] [5] [15]. For instance in [5], the authors study switched linear systems with state jumps and known switching signals. They define such systems to be observable when identical input and output signals over a time interval imply identical initial states, and present a necessary and sufficient condition for observability. Under mild additional assumptions, they design an observer for this class of switched linear systems and show that the resulting state estimation error decays exponentially.

Observability of quantized-output systems has also been studied [16] [17] [18]. For instance in [16], the author views the quantizer as providing a limited amount of information, and poses the question: How much information about the system state can be extracted given the past output sequence? Specifically, the author studies observability for one-dimensional discrete-time LTI systems with quantized outputs, uses differential entropy to measure the uncertainty in the current state estimate given the observation record, and presents control laws that minimize this uncertainty showing that the differential entropy can tend to negative infinity as the length of past input/output record grows under certain assumptions on the distribution of the initial state. In [17], the authors design an impulsive Luenberger observer based on the idea that the continuous time system output is known exactly at quantizer transition values.

Control design with quantized state feedback has also been studied [1] [2]. For instance, in [1] the author first shows that for any control law depending on quantized state

feedback, for an unstable linear system, the set of all initial conditions whose closed-loop trajectories tend to the origin has measure zero. For this reason, the author proposes a new notion of “stabilization" as controlling the trajectory within an arbitrarily small ball around the origin for an arbitrarily long time, and proposes a control strategy to achieve such stabilization. The idea of such strategy can be described as follows: Assuming reachability of the linear system and a rectilinear quantizer, the control input drives the center of the set of states associated with the quantized measurement to the corner of a quantization block, and therefore achieves sharper knowledge of the system state as the time index increases.

Control design with discrete state estimators have also been studied. In [19], the authors use discrete state estimators to estimate the discrete variables in hybrid systems where the continuous variables are available for measurement. In [20], the authors formalize a notion of finite-state estimators for controller synthesis given temporal logic specifications. In [21], the authors propose to use locally-affine observers to estimate the system state for a class of hybrid systems where control specifications are expressed in linear temporal logic.

At this time, we are not aware of any work on observability of discrete-time systems that involve both finite input alphabet (or switching control) and output quantization. In addition, we are not aware of any work addressing the state estimation problem with DFM as observers. The present thesis work therefore aims to pave the way in understanding the question of state estimation and observability in these circumstances.

2.2 Mathematics Background

We conclude this chapter by briefly reviewing some relevant concepts and results in mathematics.

2.2.1 Review of Relevant Concepts

For the sake of completeness, we review here relevant mathematical concepts and notation, beginning with the concept of equivalence relations [22]. Given a set $\mathcal{A}$, a subset $\sim$ of $\mathcal{A} \times \mathcal{A}$ is called a relation on $\mathcal{A}$. With some slight abuse of notation, we write $a \sim b$,

read a is equivalent to b, to mean that (a, b) is an element of the relation $\sim$. A relation $\sim$ on $\mathcal{A}$ is an equivalence relation if for any $a, b, c \in \mathcal{A}$, we have:

$$
\begin{aligned}
&(i) \quad a \sim a \quad \text{(reflexive)},\\
&(ii) \quad \text{If } a \sim b, \text{then } b \sim a \quad \text{(symmetric)},\\
&(iii) \quad \text{If } a \sim b \text{ and } b \sim c, \text{then } a \sim c \quad \text{(transitive)}.
\end{aligned}
$$

An equivalence relation $\sim$ on a set $\mathcal{A}$ can be used to partition $\mathcal{A}$ into equivalence classes. We use $[x]$ to denote the equivalence class of x, defined as $[x] = \{y \in \mathcal{A} | y \sim x\}$. Note that this indeed defines a partition as the following properties are satisfied:

$$
\begin{aligned}
&(i) \quad [x] \neq \varnothing, \forall x \in \mathcal{A},\\
&(ii) \quad [x] \neq [y] \Rightarrow [x] \cap [y] = \varnothing,\\
&(iii) \quad \bigcup_{x \in \mathcal{A}} [x] = \mathcal{A}.
\end{aligned}
$$

Let a f be a function: $f : \mathcal{A} \to \mathcal{B}$. f is injective if for all a and b in $\mathcal{A}$, $f(a) = f(b)$ implies $a = b$. f is surjective if for any $c \in \mathcal{B}$, there is $d \in \mathcal{A}$ such that $f(d) = c$. f is bijective if it is both injective and surjective.

A point $x \in \mathbb{R}^n$ consists of an $n-$tuple of real numbers $x = (x_1, x_2, \ldots, x_n)$. Given a positive integer p, the p-norm of x is denoted by $\|x\|_p$ and is defined as $\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$, and the ∞-norm of x is denoted by $\|x\|_\infty$ and is defined as $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$.

Next, we review relevant concepts in analysis. For the purpose of illustration, we use the 1-norm to review relevant concepts, though arbitrary p-norms are equally viable alternatives. The distance between two points x and y is then simply $\|x - y\|_1$. Given a set $\mathcal{A}$ in $\mathbb{R}^n$, the diameter of $\mathcal{A}$ is defined as $diam(\mathcal{A}) = \sup\{\|y - x\|_1 : x \in \mathcal{A}, y \in \mathcal{A}\}$. The open ball in $\mathbb{R}^n$ centered at x and of radius r is defined by $B_r(x) = \{y \in \mathbb{R}^n : \|y - x\|_1 < r\}$. Given a set $\mathcal{A}$ in $\mathbb{R}^n$, a point x is a closure point of $\mathcal{A}$ if for every $r > 0$, the ball $B_r(x)$ contains a point of $\mathcal{A}$. Similarly, a point x is a limit point of $\mathcal{A}$ if for every $r > 0$, the ball

$B_r(x)$ contains a point of $\mathcal{A}$ that is distinct from x. The closure of $\mathcal{A}$, $cl(\mathcal{A})$, consists of all closure points of $\mathcal{A}$. A point $x \in \mathcal{A}$ is an interior point of $\mathcal{A}$ if there exists $r > 0$ such that $B_r(x) \subset \mathcal{A}$. The interior of $\mathcal{A}$, $int(\mathcal{A})$, consists of all interior points of $\mathcal{A}$. A boundary point of $\mathcal{A}$ is a point which is in $cl(\mathcal{A})$ but not in $int(\mathcal{A})$. The boundary of $\mathcal{A}$, $\partial\mathcal{A}$, consists of all boundary points of $\mathcal{A}$.

Lastly, we review the notion of spectral radius of a square matrix. Given a square matrix A, the spectral radius of A is the nonnegative real number $\rho(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$. If $\rho(A) < 1$, we say that matrix A is Schur-stable. Given a square matrix A, the p-induced norm of A is defined as $\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p$. Recall that the induced norms satisfy the submultiplicative property, namely: $\|AB\|_p \leq \|A\|_p \|B\|_p$.

2.2.2 Review of Relevant Results

We next present a set of mathematical results that will be useful to us in deriving our main results in chapters that follows.

Proposition 1. Given two sets $\mathcal{A}$ and $\mathcal{B}$, and $\mathcal{A}$ is countable. If $f : \mathcal{A} \to \mathcal{B}$ is onto, then $\mathcal{B}$ is countable.

Proof. Recall that a countable set is one that is equivalent to some subset of $\mathbb{Z}_+$ (pp. 19, [23]), and two sets are equivalent if there is a one-to-one correspondence between them. Since $\mathcal{A}$ is countable, there is a function $g : \mathcal{S} \to \mathcal{A}$, where $\mathcal{S} \subset \mathbb{Z}_+$ and g is a bijection.

For any $b \in \mathcal{B}$, define a function k as:

$$k(b) = \min\{n \in \mathcal{S} | f(g(n)) = b\}$$

Because f is onto, g is bijective and the well ordered principle of natural numbers, k is well defined. Let the image of $\mathcal{B}$ under k be $\mathcal{K}$, clearly $\mathcal{K} \subset \mathcal{S} \subset \mathbb{Z}_+$. The only thing left to show is that $k : \mathcal{B} \mapsto \mathcal{K}$ is one-to-one. Let b_1 and b_2 be two distinct elements in $\mathcal{B}$. Assume $k(b_1) = k(b_2) = n_0$, then $b_1 = f(g(n_0)) = b_2$, this contradicts with $b_1 \neq b_2$. So for any two elements in $\mathcal{B}$, $b_1 \neq b_2$ implies $k(b_1) \neq k(b_2)$. Therefore $k : \mathcal{B} \mapsto \mathcal{K}$ is one-to-one.

□

Proposition 2. If $A \in \mathbb{R}^{n\times n}$ has all eigenvalues within the unit disc, then $\sum_{\tau=0}^{\infty} \|A^\tau\|_\infty$ converges.

Proof. Since A has all eigenvalues within the unit disc, $\lim_{n\to\infty} A^n = \mathbf{0}$ (pp.298, [24]). Then there exists $\tau_0 \in \mathbb{N}$, such that $\|A^{\tau_0}\|_\infty < 1$. Then

$$\begin{aligned}
&\sum_{\tau=0}^{\infty} \|A^\tau\|_\infty \\
&\leq 1 + (\|A\|_\infty + \cdots + \|A^{\tau_0}\|_\infty)(1 + \|A^{\tau_0}\|_\infty + \\
&\qquad + \|A^{2\tau_0}\|_\infty + \|A^{3\tau_0}\|_\infty + \cdots) \\
&\leq 1 + (\|A\|_\infty + \cdots + \|A^{\tau_0}\|_\infty)(1 + \|A^{\tau_0}\|_\infty + \\
&\qquad + \|A^{\tau_0}\|_\infty^2 + \|A^{\tau_0}\|_\infty^3 + \cdots) \\
&= 1 + (\|A\|_\infty + \|A^2\|_\infty + \cdots + \|A^{\tau_0}\|_\infty)\frac{1}{1 - \|A^{\tau_0}\|_\infty}
\end{aligned}$$

where the first and second inequalities follow from the sub multiplicative property of induced norms. We have found an upper bound for the infinite series, and we conclude that $\sum_{\tau=0}^{\infty} \|A^\tau\|_\infty$ converges. □

Proposition 3. For $C \in \mathbb{R}^{1\times n}$, $A \in \mathbb{R}^{n\times n}$. If $CA^m \neq \mathbf{0}$ for all $m \in \mathbb{Z}_+$, then there is a $v \in \mathbb{R}^n$ such that $CA^m v \neq 0$ for infinitely many values of $m \in \mathbb{Z}_+$.

Proof. Assume contrary, then for any $v \in \mathbb{R}^n$, $CA^m v \neq 0$ for finitely many $m \in \mathbb{Z}_+$. Equivalently, there is $M(v) \in \mathbb{Z}_+$ such that $CA^m v = 0$ for all $m \geq M(v)$. In particular, this is true for any element in the standard basis of $\mathbb{R}^n$.

Let $\{e_i : 1 \leq i \leq n\}$ denote the standard basis for $\mathbb{R}^n$, where e_i denotes the vector with a 1 in the i^{th} coordinate and 0's elsewhere. Then there exist $M(e_i) \in \mathbb{Z}_+$ such that $CA^m e_i = 0$ for all $m \geq M(e_i)$, for all $i \in \{1, 2, \cdots, n\}$.

Choose $M = \max\{M(e_1), M(e_2), \cdots, M(e_n)\}$, then $CA^M e_i = 0, \forall\ 1 \leq i \leq n$. This implies $CA^M = \mathbf{0}$, which contradicts with the assumption that $CA^m \neq \mathbf{0}$ for all $m \in \mathbb{Z}_+$,

leading to a proof by contradiction. □

Intended to be blank.

3 Output Observability of Systems Over Finite Alphabets

In this chapter, we study the problem of observability of systems over finite alphabets. We first show that a new notion of observability is needed for the systems of interest, and we then propose a new set of notions of output observability. Finally, we characterize both necessary and sufficient conditions for observability, and we illustrate these conditions with a set of examples.

3.1 Class of Systems of Interest

Before we study the observability of systems over finite alphabets, we first take a closer look at system (2).

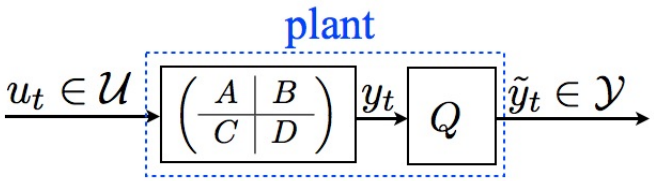


Figure 1: A specific class of systems.

The above figure illustrates the structure of systems (2). Recall that the quantizer $Q : \mathbb{R}^p \to \mathcal{Y}$ is a *piecewise-constant* function. To make this notion precise, we describe such functions as follows.

Definition 1. Given set $\mathcal{Y}$, we say a function $Q : \mathbb{R}^p \to \mathcal{Y}$ is piecewise-constant if for any $y \in \mathbb{R}^p$, if Q is continuous at y, then there is $\delta > 0$ such that $Q(z) = Q(y)$ for all $\|z - y\| < \delta$, $z \in \mathbb{R}^p$.

The above definition is in accordance with the definition of a function being continuous at a point (pp. 63, [23]). An example of such a quantizer with $p = 1$ is:

$$Q(y) = \begin{cases} i, & y \in [i - 0.5, i + 0.5) \text{ for } i \in \mathbb{Z} \text{ and } |i| \leq R \\ \lfloor R \rfloor, & y \geq \lfloor R \rfloor + 0.5 \\ -\lfloor R \rfloor, & y < -\lfloor R \rfloor - 0.5 \end{cases} \tag{3}$$

Here $R \in \mathbb{R}_+$ is a given parameter defining the quantizer.

If we assume the one-dimensional quantizer Q to be right-continuous, a more general form of Q is described by:

$$Q([\beta_i, \beta_{i+1})) = \tilde{y}_i,\ \tilde{y}_i \in \mathcal{Y}, \tag{4}$$

where $\{\beta_i\}_{i=0}^N$ contains the discontinuous points of Q, and $\beta_0 = -\infty$, $\beta_N = \infty$. Recall that we use the notation $[\alpha, \beta)$ to denote the interval $\{x \in \mathbb{R} : \alpha \leq x < \beta\}$ for $\alpha, \beta \in \mathbb{R}$.

The system (2) shown in Figure 1 with input u_t and output $\tilde{y}_t$ is nonlinear, takes on finite input values, produces finite output values, and is thus an instance of a system over finite alphabets defined in (1). In the following, we investigate the problem of observability for systems (1) and (2). The work presented in this section consists of the previous work reported in [25], as well as some new observations.

3.2 Motivation for a New Notion of Observability

A natural starting point in our study is to attempt to apply the definition of LTI system observability to system (2). Unsurprisingly, we quickly discover that no system in the class of systems under consideration is observable under the definition of observability of discrete-time LTI systems.

Lemma 1. The initial state of system (2) cannot be uniquely determined by knowledge of $(u_t, \tilde{y}_t)$ over any finite time interval.

Proof. All possible initial states of system (2) are in the set $\mathcal{X}_0 = \mathbb{R}^n$. Clearly $\mathcal{X}_0$ is uncountable.

Now assume that we can uniquely determine any initial condition from the input u_t and output $\tilde{y}_t$ over some time interval, say $t \in \{0, \ldots, T\}$ for some $T \in \mathbb{Z}_+$. Let $\mathcal{O}$ be the set of all such possible sequences: $\mathcal{O} \subseteq \mathcal{U}^T \times \mathcal{Y}^T$. Since $|\mathcal{U}| < \infty$ and $|\mathcal{Y}| < \infty$, $\mathcal{U}^T \times \mathcal{Y}^T$ is countable and so is $\mathcal{O}$.

By assumption, any initial condition in $\mathcal{X}_0$ can be uniquely determined by an element in $\mathcal{O}$. So there exists a map $f : \mathcal{O} \to \mathcal{X}_0$, and f is onto. This indicates that $\mathcal{X}_0$ is countable

(Proposition 1 in Section 2.2.2), leading to a contradiction.

□

Remark. Lemma 1 still holds when the initial state of system (2) is bounded. Specifically, if $\mathcal{X}_0 = \{x \in \mathbb{R}^n : \|x\|_\infty \leq b\}$ for some $b \in \mathbb{R}_+$, then $\mathcal{X}_0$ is still uncountable and the proof follows unchanged.

From Lemma 1, we conclude that any effort to uniquely determine the system state x_t from the observation of input and output sequences over any time interval will fail. We thus need to think of observability differently.

Specifically, we propose to shift our attention from state estimation, to state estimation for the purpose of output prediction.

Before we proceed, we point out a distinction of our problem. In the traditional LTI setting, the effects of the initial state of a stable LTI system will die down eventually. Consequently, the question of observability is only interesting for unstable systems. This is not the case for our class of systems of interest, as the following example demonstrates:

Example 1. For system (2) with parameters: $A = 0.5, B = 1, C = 1, D = 1, \mathcal{U} = \{0, \pm 1\}$, Q defined in (3) with $R = 1$, and (consequently) $\mathcal{Y} = \{0, \pm 1\}$, consider the following question: Given u_t and $\tilde{y}_t$ for all $0 \leq t \leq T$ for some $T \in \mathbb{Z}_+$, and an arbitrary $u_{T+1} \in \mathcal{U}$, is it possible to uniquely determine $\tilde{y}_{T+1}$?

We show that the answer to this question is negative. Assume that $\tilde{y}_{T+1}$ can be uniquely determined for some $T \in \mathbb{Z}_+$. Let $u_t = 0$ for $0 \leq t \leq T-2$, $u_{T-1} = 1$ and $u_T = u_{T+1} = 0$. For two distinct initial states $x_0^1 = 0.1$ and $x_0^2 = -0.1$, we use $\tilde{y}_t^1$ and $\tilde{y}_t^2$ to denote the quantized outputs respectively. Then $\tilde{y}_t^1 = \tilde{y}_t^2 = 0$ for $0 \leq t \leq T-2$, and $\tilde{y}_t^1 = \tilde{y}_t^2 = 1$ for $T-1 \leq t \leq T$. By assumption we can uniquely determine $\tilde{y}_{T+1}$, which contradicts with $\tilde{y}_{T+1}^1 = 1$ and $\tilde{y}_{T+1}^2 = 0$.

As shown in Example 1, the initial state of system (2) impacts the quantized output at arbitrarily large times, even though the underlying LTI system is stable. Consequently, the question of observability remains relevant even when the internal dynamics are stable.

3.3 Definitions of Observability

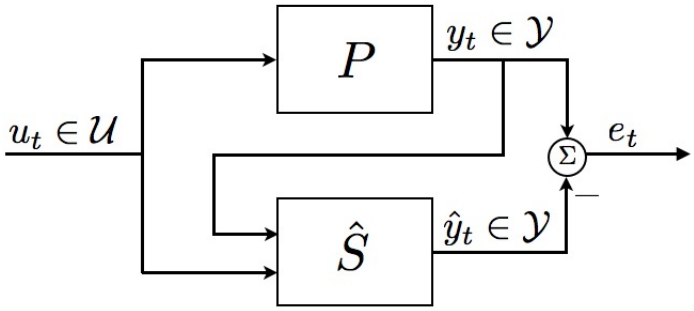


Figure 2: Interconnection of plant and observer.

In the above figure, an observer $\hat{S}$ for P (1) is described by:

$$q_{t+1} = f(q_t, u_t, y_t), \tag{5a}$$

$$\hat{y}_t = g(q_t, u_t), \tag{5b}$$

where $t \in \mathbb{N}$, $q_t \in \mathcal{Q}$ for some set $\mathcal{Q}$, $u_t \in \mathcal{U}$, $y_t \in \mathcal{Y}$, $\hat{y}_t \in \mathcal{Y}$ and functions $f : \mathcal{Q}\times\mathcal{U}\times\mathcal{Y} \to \mathcal{Q}$ and $g : \mathcal{Q} \times \mathcal{U} \to \mathcal{Y}$. Here u_t and y_t are the input and output of P respectively. In this work, we assume that the observer $\hat{S}$ (5) initialize at a fixed state: $q_0 = q_o$, for some $q_o \in \mathcal{Q}$.

Definition 2. Consider a system over finite alphabets P, and the setup in Figure 2. $\gamma \in \mathbb{R}_{\geq 0}$ is an *observation gain* of P if there exists an observer $\hat{S}$ (5) such that for any $(\mathbf{u}, \mathbf{y}) \in P$,

$$\sup_{T\geq 0} \sum_{t=0}^{T} \|y_t - \hat{y}_t\| - \gamma\|u_t\| < \infty. \tag{6}$$

Here $\|\cdot\|$ is assumed to be some well defined norm on $\mathcal{U}$ and $\mathcal{Y}$ (pp. 39, [23]). When system (2) is concerned, $\|\cdot\|$ is assumed to be the Euclidean norm corresponding with the appropriate dimensions.

Given a system P, define the *$\mathcal{O}$-gain* γ^* of P as:

$$\gamma^* = \inf\{\gamma \in \mathbb{R}_{\geq 0} : \gamma \text{ is an observation gain of } P\}. \tag{7}$$

Next, propose a set of new notions of output observability for P.

Definition 3. Consider a system over finite alphabets P (1), and the setup in Figure 2. Define the following:

- P is *finite memory observable* (C1) if there exists an observer $\hat{S}$ (5) and $T \in \mathbb{Z}_+$ such that for any $(\mathbf{u}, \mathbf{y}) \in P$, $\hat{y}_t = y_t$ for all $t \geq T$.
- P is *weakly observable* (C2) if $\gamma = 0$ is an observation gain of P.
- P is *asymptotically observable* (C3) if the $\mathcal{O}$-gain γ^* of P is 0.

3.4 Conditions for Finite Memory Observable

In this section, we characterize finite memory observability (C1) by proposing a set of conditions for (C1). First, we propose a sufficient condition for (C1). Given system (2), define sets $\mathcal{A}$ and $\mathcal{B}$ as

$$\mathcal{A} = \{\alpha \in \mathbb{R}^p : \alpha = \sum_{\tau=0}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t, u_{(\cdot)} \in \mathcal{U}, t \in \mathbb{N}\}, \tag{8}$$

$$\mathcal{B} = \{\beta \in \mathbb{R}^p : Q(y) \text{ is discontinuous at } y = \beta\}. \tag{9}$$

Now we are ready to propose a sufficient condition for (C1).

Theorem 1. Consider system (2), assume that A has all eigenvalues within the unit disc, and the initial state x_0 is bounded. If $d(\mathcal{A}, \mathcal{B}) \neq 0$, then system (2) is finite memory observable (C1).

Remark. Theorem 1 is an extension of Theorem 1 in [25], in the sense that Theorem 1 addresses multi-input multi-output systems.

Proof. First note that if C is the zero matrix, then $\tilde{y}_t = Q(Du_t)$. Since $\tilde{y}_t$ can be determined by the knowledge of u_t, system (2) is (C1). Therefore in the following derivation, we only consider the case $C \neq \mathbf{0}$.

Since the initial state x_0 is bounded, we have $\|x_0\|_\infty \leq b$ for some $b \in \mathbb{R}$. Next we find a uniform bound on the state x_t. Given system (2), the solution of x_t is given by:

$$x_t = A^t x_0 + \sum_{\tau=0}^{t-1} A^{t-1-\tau} B u_\tau.$$

Then for any $t \in \mathbb{N}$,

$$\begin{aligned} \|x_t\|_\infty &= \|A^t x_0 + \sum_{\tau=0}^{t-1} A^{t-1-\tau} B u_\tau\|_\infty \\ &\leq \max\{\|x_0\|_\infty, \|Bu_t\|_\infty\} \sum_{\tau=0}^{t} \|A^\tau\|_\infty. \end{aligned}$$

Since $\sum_{\tau=0}^{\infty} \|A^\tau\|_\infty$ converges (Proposition 2 in the Section 2.2.2), we can find an upper bound $b_1 \in \mathbb{R}_+$ such that $\sum_{\tau=0}^{\infty} \|A^\tau\|_\infty \leq b_1$. Since $u_t \in \mathcal{U}$ and $\mathcal{U}$ is finite, $\|Bu_t\|_\infty$ is also bounded. Let $b_2 = \max\{b, \max\{\|Bu\|_\infty : u \in \mathcal{U}\}\}$, we have $\|x_t\|_\infty \leq b_1 b_2$ for all $t \in \mathbb{N}$.

Next, choose $T \in \mathbb{Z}_+$ such that $\|CA^T x_t\| < d(\mathcal{A}, \mathcal{B})/2$. Since

$$\|CA^T x_t\|_\infty \leq \|C\|_\infty \|A^T\|_\infty \|x_t\|_\infty,$$

and $\lim_{\tau \to \infty} A^\tau = 0$ (pp. 298, [24]), recall the assumption that $C \neq \mathbf{0}$, we can choose $T \in \mathbb{Z}_+$ such that

$$\|A^T\|_\infty < \frac{d(\mathcal{A}, \mathcal{B})}{2 b_1 b_2 \|C\|_\infty}. \tag{10}$$

Then for all $t \geq T$:

$$\begin{aligned} y_t &= CA^T x_{t-T} + \sum_{\tau=t-T}^{t-1} CA^{t-1-\tau} B u_\tau + D u_t \\ &\in B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha), \end{aligned} \tag{11}$$

where $\alpha = \sum_{\tau=t-T}^{t-1} CA^{t-1-\tau} B u_\tau + D u_t$, and $\alpha \in \mathcal{A}$.

Next, we observe that the quantized output $\tilde{y}_t$ can be determined by the knowledge of

α. Particularly, for any $\alpha \in \mathcal{A}$,

$$y \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha)) \;\Rightarrow\; Q(y) = Q(\alpha). \tag{12}$$

To show this, Q is continuous at any point $y \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$, otherwise $d(\mathcal{A},\mathcal{B}) \leq d(\mathcal{A},\mathcal{B})/2$, which contradicts with $d(\mathcal{A},\mathcal{B}) > 0$. Next, assume there is a $y \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$ such that $Q(y) \neq Q(\alpha)$. Define two sequences $\{w_n\}_{n=1}^{\infty}$, $\{v_n\}_{n=1}^{\infty}$ as follows: Let $w_1 = \alpha$, $v_1 = y$. For any $n \geq 2$, let $z = (w_{n-1} + v_{n-1})/2$, if $Q(z) \neq Q(w_{n-1})$, let $w_n = w_{n-1}, v_n = z$; otherwise, let $w_n = z, v_n = v_{n-1}$. By this definition, we see that $Q(w_n) \neq Q(v_n)$ implies $Q(w_{n+1}) \neq Q(v_{n+1})$. Since $Q(w_1) \neq Q(v_1)$, by induction, we have: $Q(w_n) \neq Q(v_n)$ for all $n \in \mathbb{Z}_+$. At the same time, it is clear that $\|w_n - v_n\| = (1/2)^n \|w_1 - v_1\|$.

Note that $\{w_n\}_{n=1}^{\infty} \subset cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$, and $cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$ is a compact set in $\mathbb{R}^p$, therefore there is a $\{w_{n_p}\}_{p=1}^{\infty}$ such that $\lim_{p\to\infty} w_{n_p} = w$ for some $w \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$. Similarly, since $\{v_n\}_{n=1}^{\infty} \subset cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$, and consequently $\{v_{n_p}\}_{p=1}^{\infty} \subset cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$, there is a subsequence $\{v_{n_{p_q}}\}_{q=1}^{\infty}$ such that $\lim_{p\to\infty} v_{n_{p_q}} = v$ for some $v \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$. Rename $\{w_{n_{p_q}}\}$ and $\{v_{n_{p_q}}\}$ as $\{w'_n\}$ and $\{v'_n\}$ respectively, we have: $Q(w'_n) \neq Q(v'_n)$, $\|w'_n - v'_n\| \leq (1/2)^n \|w_1 - v_1\|$, for all $n \in \mathbb{Z}_+$, and $\lim_{n\to\infty} w'_n = w$, $\lim_{n\to\infty} v'_n = v$ for some w, v in $cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$. Since $\|w - v\| \leq \|w - w'_n\| + \|w'_n - v'_n\| + \|v'_n - v\|$, we see that for any $\epsilon > 0$, $\|w - v\| < \epsilon$, and consequently $w = v$. Since $w \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$, Q is continuous at w. Recall Definition 1, there is $\delta > 0$ such that $Q(z) = Q(w)$ for all $\|z - w\| < \delta$, $z \in \mathbb{R}^p$. Since $\lim_{n\to\infty} w'_n = w$, there is N_1 such that $Q(w'_n) = Q(w)$, for all $n \geq N_1$. Similarly, there is N_2 such that $Q(v'_n) = Q(v) = Q(w)$ for all $n \geq N_2$. Let $n = \max\{N_1, N_2\}$, then $Q(w'_n) = Q(w) = Q(v'_n)$, which contradicts with $Q(w'_n) \neq Q(v'_n)$ for all $n \in \mathbb{Z}_+$. Therefore, assumption is false, and we conclude that for any $\alpha \in \mathcal{A}$, and any $y \in cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$,

$$Q(y) = Q(\alpha). \tag{13}$$

Based on the observations made in (11), (13), design an observer for systems (2). Similar to the construction in the derivation of Lemma 2, consider an observer $\hat{S}$ described

by:

$$\begin{aligned} q_{t+1} &= \phi(q_t, u_t), \\ \hat{y}_t &= \theta(q_t, u_t), \end{aligned} \tag{14}$$

where $q_t \in q_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i)$ is the state of $\hat{S}$, $u_t \in \mathcal{U}$ is the input of (2). Here T is determined by (10). Function $\phi : (q_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i)) \times \mathcal{U} \to \bigcup_{i=1}^{T} \mathcal{U}^i$ is described by: For any $q \in q_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i)$, any $u \in \mathcal{U}$,

$\triangleright$ If $q = q_o$, then

$$\phi(q, u) = u.$$

$\triangleright$ If $q \in \bigcup_{i=1}^{T-1} \mathcal{U}^i$, write $q = (u_1, u_2 \dots u_i)$ for some $i \in \{1, \dots, T-1\}$, then

$$\phi(q, u) = (u, u_1, u_2 \dots u_i).$$

$\triangleright$ If $q \in \mathcal{U}^T$, write $q = (u_1, u_2 \dots u_T)$, then

$$\phi(q, u) = (u, u_1, u_2 \dots u_{T-1}).$$

Let $q_0 = q_o$, then by the same argument in the derivation of Lemma 2, we have

$$q_t = (u_{t-1}, u_{t-2}, \dots, u_{t-T}), \ \forall\, t \geq T. \tag{15}$$

where $\{u_t\}$ is the input of system (2).

Next, define function $\theta : (q_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i)) \times \mathcal{U} \to \mathcal{Y}$ as: For any $q \in q_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i)$, any $u \in \mathcal{U}$, if $q \in \mathcal{U}^T$, write $q = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_T)$, then

$$\theta(q, u) = Q(\sum_{\tau=1}^{T} CA^{\tau-1}B\bar{u}_\tau + Du). \tag{16}$$

If $q \notin \mathcal{U}^T$, then let $\theta(q, u) = y_\varnothing$ for some $y_\varnothing \in \mathcal{Y}$.

For any $t \geq T$, recall (14), (15), (16), we have

$$\begin{aligned}\hat{y}_t &= \theta(q_t, u_t) \\ &= Q(\sum_{\tau=1}^{T} CA^{\tau-1}Bu_{t-\tau} + Du_t) \\ &= Q(\sum_{\tau=t-T}^{t-1} CA^{t-\tau-1}Bu_\tau + Du_t) \\ &= Q(\alpha),\end{aligned}$$

where $\alpha = \sum_{\tau=t-T}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t$. Recall (11), we have $y_t \in B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha) \subset cl(B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha))$. By (13), we have $Q(y_t) = Q(\alpha)$. We conclude that $\hat{y}_t = Q(\alpha) = Q(y_t) = \tilde{y}_t$ for all $t \geq T$, and consequently system (2) is finite memory observable. □

Next, we propose another sufficient condition for (C1).

Theorem 2. Consider system (2), if $CA^l = \mathbf{0}$ for some $l \in \mathbb{Z}_+$, then system (2) is finite memory observable (C1).

Remark. Again, Theorem 2 is an extension of Theorem 2 in [25], in the sense that Theorem 2 addresses multi-input multi-output systems.

Proof. Given system (2), the solution of y_t is given by:

$$y_t = CA^t x_0 + \sum_{\tau=0}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t.$$

Since $CA^l = \mathbf{0}$, then for all $t \geq l$:

$$y_t = \sum_{\tau=t-l}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t, \quad \forall \quad t \geq l$$

Use the same observer described in the derivation of Theorem 1, particularly from (14)

to (16), except changing "T" to "l". Then

$$\hat{y}_t = Q(\sum_{\tau=t-l}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t) = Q(y_t) = \tilde{y}_t,\ \forall\, t \geq l.$$

We conclude that system (2) is (C1). □

Remark. The condition stated in Theorem 2 is related to the "traditional" observability of linear time-invariant systems. In particular, we make the following observation:

Observation 1. Given system (2) with "A" not being nilpotent, if $CA^l = \mathbf{0}$ for some $l \in \mathbb{Z}_+$, then the pair (C, A) is not observable.

To see this, let λ be a nonzero eigenvalue of A. Since A is not nilpotent, such a λ always exists. Let v be an eigenvector associated with λ, then $CA^lv = \lambda^l Cv = 0$. Since $\lambda \neq 0$, we have $\lambda^l \neq 0$, and consequently $Cv = 0$. Since v is in the kernel of C, and v is an eigenvector of A, the pair (C, A) is not observable (pp. 145, [8]).

We point out that the requirement of A not being nilpotent in Observation 1 can not be dropped. In particular, consider a system with $C = [1\ 0]$, $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, then $CA^2 = \mathbf{0}$. But the corresponding observability matrix is the identity matrix, and therefore (C, A) is observable.

Next, we present a necessary condition for (C1).

Theorem 3. Consider system (2), assume that A has all eigenvalues within the unit disc, $0 \in \mathcal{U}$, and $0 \notin \mathcal{B}$. If $rank(CA^l) = p$ for all $l \in \mathbb{Z}_+$, and $\mathcal{A} \cap \mathcal{B} \neq \varnothing$, then system (2) is not finite memory observable (C1).

Remark. Compared to sufficient conditions in Theorems 1 and 2, both the assumptions "$\mathcal{A} \cap \mathcal{B} \neq \varnothing$" and "$rank(CA^l) = p$ for all $l \in \mathbb{Z}_+$" here are stronger than "$d(\mathcal{A}, \mathcal{B}) = 0$", and "$CA^l \neq \mathbf{0}$ for all $l \in \mathbb{Z}_+$" respectively. However, when $p = 1$, $rank(CA^l) = p$ and $CA^l \neq \mathbf{0}$ are equivalent, and a corresponding version of Theorem 3 is reported in [25].

Remark. Again, Theorem 3 is an extension of Theorem 2 in [25], in the sense that Theorem 3 addresses multi-input multi-output systems. However, the proof of Theorem 2 in [25]

contains additional analysis regarding the construction of DFM observers, which is not necessary for the derivation of Theorem 3 here. That particular part of analysis is presented in Section 4 instead.

Proof. We show Theorem 3 by contradiction. For simplicity of notation, for any input sequence $\mathbf{u}$, write $F(u,t) = \sum_{\tau=0}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t$. Essentially $F(u,t)$ is the forced response of the underlying LTI system at time t under the input $\mathbf{u}$.

Since $\mathcal{A} \cap \mathcal{B} \neq \varnothing$, there exist $t_1 \in \mathbb{N}$ and $\mathbf{u^1}$ such that $t_1 = \min\{t : F(u,t) \in \mathcal{A} \cap \mathcal{B}\}$ and $F(u^1, t_1) \in \mathcal{B}$. The existence of the minimum is guaranteed by the well-ordering principle of nonnegative integers (pp. 28, [22]). t_1 being a minimum indicates that $F(u^1, t)$ is not in $\mathcal{B}$ for any $t < t_1$. So we can define the following distance:

$$d_1 = \begin{cases} d(\{0\}, \mathcal{B}), & \text{if } t_1 = 0 \\ d(\{0\} \cup \{F(u^1,t) : 0 \leq t \leq t_1 - 1\}, \mathcal{B}), & \text{if } t_1 \geq 1 \end{cases} \tag{17}$$

The definition of t_1 and $0 \notin \mathcal{B}$ imply $d_1 > 0$.

Assume that system (2) is finite memory observable, than there exists an observer $\hat{S}$ (5) and T such that for any $x_0 \in \mathbb{R}^n$, any $\mathbf{u} \in \mathcal{U}^{\mathbb{N}}$, $\hat{y}_t = \tilde{y}_t$ for all $t \geq T$. Without loss of generality, we assume that $T \geq t_1$ (if $T < t_1$, just let $T = t_1$, then $\hat{y}_t = \tilde{y}_t$ for all $t \geq T$ still holds).

Next, construct an input sequence $\mathbf{u}$ of system (2). Given $\mathbf{u^1}$, use the truncated sequence of $\mathbf{u^1}$: $\{u_t^1 : 0 \leq t \leq t_1\}$, the input sequence $\mathbf{u}$ is described as follows:

$$u_t = \begin{cases} 0, & 0 \leq t \leq T - t_1 - 1 \\ u^1_{t-(T-t_1)}, & T - t_1 \leq t \leq T \\ 0, & t > T \end{cases} \tag{18}$$

Basically we insert the truncated sequence of $\{u_t^1 : 0 \leq t \leq t_1\}$ into a zero input.

Next, if distinct initial states x_0^1 and x_0^2 satisfy:

$$\|CA^t x_0^i\| < d_1/2, \qquad i = 1, 2 \tag{19}$$

for $t = 0, 1 \cdots T-1$, then under input $\mathbf{u}$ (18), the corresponding outputs of the underlying LTI system, y_t^1 and y_t^2, satisfy:

$$y_t^i \in B_{d_1/2}(\alpha), \qquad i = 1, 2$$

for some $\alpha \in \{0\} \cup \{F(u^1, t) : 0 \leq t \leq t_1 - 1\}$, for $t = 0, 1 \cdots T-1$. Recall the definition of d_1, and the property (12) of the quantizer Q, we have $Q(y_t^1) = Q(\alpha) = Q(y_t^2)$. Consequently, we have

$$\tilde{y}_t^1 = \tilde{y}_t^2, \ t = 0, 1 \cdots T-1,$$

where $\tilde{y}_t^i$ is the output of system (2) when the initial state is x_0^i and the input is $\mathbf{u}$ (18).

In addition, since Q is not continuous at $F(u, T) = F(u^1, t_1)$, for any $\delta > 0$, there is $z \in \mathbb{R}^p$ such that $Q(z + F(u, T)) \neq Q(F(u, T))$, and $\|z\| < \delta$. Since $rank(CA^T) = p$ by assumption, write $CA^T = [v_1 \ v_2 \ \cdots \ v_n]$, where $v_1, \ldots, v_n \in \mathbb{R}^p$, then there is $\{i_1, i_2, \ldots, i_p\} \subset \{1, 2, \ldots, n\}$ such that $[v_{i_1} \ v_{i_2} \ \cdots \ v_{i_p}]$ is invertible.

Let $V = [v_{i_1} \ v_{i_2} \ \cdots \ v_{i_p}]$, and let $K_A = \sup\{\|A^t\| : t = 0, 1, 2, \cdots\}$, choose $\delta > 0$ as follows:

$$\delta = \frac{d_1}{2\|V^{-1}\|\|C\|K_A}.$$

Then there is $z \in \mathbb{R}^p$ such that $Q(z+F(u,T)) \neq Q(F(u,T))$, and $\|z\| < \delta$. Let $w = V^{-1}z$, and write $w = [w_1 \ w_2 \ \cdots \ w_p]^T \in \mathbb{R}^p$, define a vector $x^* = [x_1^* \ x_2^* \ \cdots \ x_n^*]^T \in \mathbb{R}^n$ as: For all $1 \leq j \leq p$, $x_{i_j}^* = w_j$; for all $1 \leq l \leq n$ and $l \notin \{i_1, i_2, \ldots, i_p\}$, $x_l^* = 0$. Then we have $CA^T x^* = Vw$, and $\|x^*\| = \|w\|$, where $\|\cdot\|$ is any vector p-norm.

Consider two distinct initial conditions: $x_0^1 = 0, x_0^2 = x^*$. Then for all $t = 0, 1 \cdots T-1$, $\|CA^t x_0^2\| \leq \|C\|\|A^t\|\|x^*\| \leq \|C\|K_A\|w\| \leq \|C\|K_A\|V^{-1}\|\|z\| < \|C\|K_A\|V^{-1}\| \cdot \frac{d_1}{2\|V^{-1}\|\|C\|K_A} = d_1/2$. Therefore (19) holds, and consequently $\tilde{y}_t^1 = \tilde{y}_t^2, \ t = 0, 1 \cdots T-1$.

At $t = T$, $\tilde{y}_T^1 = Q(F(u,T))$, and $\tilde{y}_T^2 = Q(CA^T x^* + F(u,T)) = Q(z + F(u,T)) \neq Q(F(u,T))$, therefore $\tilde{y}_T^1 \neq \tilde{y}_T^2$.

Since system (2) is assumed to be (C1), let $\hat{y}_t^1$ and $\hat{y}_t^2$ be the output of the corresponding $\hat{S}$ when the input is $\mathbf{u}$ (18), and initial conditions are x_0^1, x_0^2 respectively. Then at $t = T$, recall (5), we have $\hat{y}_T^1 = g(f(\dots f(f(q_o, u_0, \tilde{y}_0^1), u_1, \tilde{y}_1^1) \dots, u_{T-1}, \tilde{y}_{T-1}^1), u_T)$, and $\hat{y}_T^2 = g(f(\dots f(f(q_o, u_0, \tilde{y}_0^2), u_1, \tilde{y}_1^2) \dots, u_{T-1}, \tilde{y}_{T-1}^2), u_T)$. Recall that $\tilde{y}_t^1 = \tilde{y}_t^2,\ t = 0, 1 \cdots T-1$, we have $\hat{y}_T^1 = \hat{y}_T^2$. Since $\tilde{y}_T^1 \neq \tilde{y}_T^2$, there is $i \in \{1, 2\}$ such that $\hat{y}_T^i \neq \tilde{y}_T^i$. This is a contradiction with system (2) being (C1), therefore the assumption is false, and system (2) is not finite memory observable (C1).

□

In light of Theorems 1 and 3, we continue to study conditions for finite memory observability particularly when the matrix A is not Schur-stable. In the following, we propose another sufficient condition for (C1).

Theorem 4. Consider system (2), assume that $\rho(A) \geq 1$, and the initial state x_0 is bounded. Given the Jordan canonical form of matrix A, define a set $\mathcal{E}^U$ as the collection of generalized eigenvectors of A corresponding with eigenvalues whose magnitudes are greater than or equal to 1. If $d(\mathcal{A}, \mathcal{B}) \neq 0$, and $\mathcal{E}^U$ is in the kernel of C, then system (2) is finite memory observable.

Remark. We remind ourselves that v is a generalized eigenvector corresponding to the matrix A and the eigenvalue λ if $(A - \lambda I)^l v = 0$, but $(A - \lambda I)^{l-1} v \neq 0$ for some $l \in \mathbb{Z}_+$ (pp. 189, [26]). The condition "$\mathcal{E}^U$ is in the kernel of C" is similar to the eigenvector test for observability of LTI systems (pp. 145, [8]).

Proof. Consider the Jordan canonical form of the matrix A,

$$A = MJM^{-1},$$

where matrix J is in partitioned diagonal form, and matrix M is a generalized modal matrix for A (pp. 205, [26]). Write $M = [v_1\ v_2\ \cdots\ v_n]$, where $v_i \in \mathbb{C}^n$ for $1 \leq i \leq n$, then each

v_i is a generalized eigenvector of A, and $\{v_i\}_{i=1}^n$ form a basis of $\mathbb{R}^n$. For each v_i, use λ_i to denote the eigenvalue of A corresponding to v_i.

Next, we decompose the state vector x_t using $\{v_i\}_{i=1}^n$. Recall the notation: For any $v \in \mathbb{C}^n$, use $[v]_i$ to denote its i-th element. For all $t \in \mathbb{N}$, write x_t as a linear combination of $\{v_i\}_{i=1}^n$,

$$x_t = \sum_{i=1}^{n} [\alpha_t]_i v_i, \tag{20}$$

where $\alpha_t \in \mathbb{C}^n$ is the coordinates of x_t corresponding to the basis $\{v_i\}_{i=1}^n$.

We claim that $\sum_{i:v_i \notin \mathcal{E}^U} |[\alpha_t]_i|$ is uniformly bounded for all t. Here $[\alpha_t]_i$ is the coordinates of x_t with respect to the basis $\{v_i\}_{i=1}^n$, and $[\alpha_t]_i$ with $v_i \notin \mathcal{E}^U$ are the coordinates corresponding with the stable generalized eigenvectors.

In order to show the preceding claim, recall that $M = [v_1\ v_2\ \cdots\ v_n]$, we have $M^{-1}x_t = \alpha_t$. Define $\alpha_t^U \in \mathbb{C}^n$ as

$$[\alpha_t^U]_i = \begin{cases} [\alpha_t]_i, & \text{if } v_i \in \mathcal{E}^U, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, define $\alpha_t^S \in \mathbb{C}^n$ as

$$[\alpha_t^S]_i = \begin{cases} [\alpha_t]_i, & \text{if } v_i \notin \mathcal{E}^U, \\ 0, & \text{otherwise.} \end{cases}$$

We have $\alpha_t = \alpha_t^U + \alpha_t^S$. Essentially, α_t^U are the coordinates corresponding with the unstable or marginally stable generalized eigenvectors, while α_t^S are the coordinates corresponding with the stable generalized eigenvectors.

Similarly, we decompose the "Bu_t" term of system (2). For all $t \in \mathbb{N}$, write Bu_t as a linear combination of $\{v_i\}_{i=1}^n$,

$$Bu_t = \sum_{i=1}^{n} [\beta_t]_i v_i,$$

where $\beta_t \in \mathbb{C}^n$ is the coordinates of Bu_t corresponding to the basis $\{v_i\}_{i=1}^n$.

Then $M^{-1}Bu_t = \beta_t$. Define $\beta_t^U \in \mathbb{C}^n$ as

$$[\beta_t^U]_i = \begin{cases} [\beta_t]_i, & \text{if } v_i \in \mathcal{E}^U, \\ 0, & \text{otherwise,} \end{cases}$$

and define $\beta_t^S \in \mathbb{C}^n$ as

$$[\beta_t^S]_i = \begin{cases} [\beta_t]_i, & \text{if } v_i \notin \mathcal{E}^U, \\ 0, & \text{otherwise.} \end{cases}$$

We have $\beta_t = \beta_t^U + \beta_t^S$.

Since for all $t \in \mathbb{N}$,

$$x_{t+1} = Ax_t + Bu_t,$$

recall $A = MJM^{-1}$, we have

$$MM^{-1}x_{t+1} = MJM^{-1}x_t + MM^{-1}Bu_t.$$

Left multiply by M^{-1}, we have

$$M^{-1}x_{t+1} = JM^{-1}x_t + M^{-1}Bu_t.$$

Recall $M^{-1}x_t = \alpha_t$, and $M^{-1}Bu_t = \beta_t$, we have

$$\alpha_{t+1} = J\alpha_t + \beta_t.$$

Recall $\alpha_t = \alpha_t^U + \alpha_t^S$ and $\beta_t = \beta_t^U + \beta_t^S$, we have

$$\begin{aligned} \alpha_{t+1}^U + \alpha_{t+1}^S &= J(\alpha_t^U + \alpha_t^S) + (\beta_t^U + \beta_t^S) \\ &= (J\alpha_t^U + \beta_t^U) + (J\alpha_t^S + \beta_t^S). \end{aligned} \tag{21}$$

Consider the term $J\alpha_t^U$. Write $J = [w_1 \cdots w_n]$, where $w_1, \ldots, w_n \in \mathbb{C}^n$. Then

$$J\alpha_t^U = \sum_{i=1}^{n} [\alpha_t^U]_i w_i.$$

Since $[\alpha_t^U]_i = 0$ for all i such that $v_i \notin \mathcal{E}^U$, we have

$$J\alpha_t^U = \sum_{i:v_i \in \mathcal{E}^U} [\alpha_t^U]_i w_i. \tag{22}$$

Recall the definition of λ_i for $i = 1, \ldots, n$, and the form of J, we see that if $\lambda_j \neq \lambda_i$, then $[w_i]_j = 0$, for all $1 \leq i, j \leq n$. For any i such that $v_i \in \mathcal{E}^U$, and any j such that $v_j \notin \mathcal{E}^U$, we have $|\lambda_j| < 1$ and $|\lambda_i| \geq 1$, therefore $\lambda_j \neq \lambda_i$, and consequently $[w_i]_j = 0$. Recall the preceding equation, we have: For all j such that $v_j \notin \mathcal{E}^U$,

$$[J\alpha_t^U]_j = 0. \tag{23}$$

Similarly, we can show that: For all j such that $v_j \in \mathcal{E}^U$,

$$[J\alpha_t^S]_j = 0. \tag{24}$$

Recall (21), we see that

$$\alpha_{t+1}^S = (J\alpha_t^S + \beta_t^S), \tag{25}$$

otherwise assume $[\alpha_{t+1}^S - (J\alpha_t^S + \beta_t^S)]_j \neq 0$ for some $j \in \{1, \ldots, n\}$, then by the definitions of α_{t+1}^S, β_t^S, and (24), we see that j satisfies $v_j \notin \mathcal{E}^U$. By (21), we have $\alpha_{t+1}^S - (J\alpha_t^S + \beta_t^S) = (J\alpha_t^U + \beta_t^U) - \alpha_{t+1}^U$, and note that $[(J\alpha_t^U + \beta_t^U) - \alpha_{t+1}^U]_j = 0$ for any j such that $v_j \notin \mathcal{E}^U$. Therefore $0 \neq [\alpha_{t+1}^S - (J\alpha_t^S + \beta_t^S)]_j = [(J\alpha_t^U + \beta_t^U) - \alpha_{t+1}^U]_j = 0$, which is a contradiction.

Consider the term $J\alpha_t^S$, we have

$$J\alpha_t^S = \sum_{i:v_i \notin \mathcal{E}^U} [\alpha_t^S]_i w_i. \tag{26}$$

Define a square matrix J^S to be $J^S = [w_1^S \; w_2^S \cdots \; w_n^S]$, where

$$w_i^S = \begin{cases} w_i, & \text{if } v_i \notin \mathcal{E}^U, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Note that J^S is Schur-stable. Then we have

$$J\alpha_t^S = \sum_{i:v_i \notin \mathcal{E}^U} [\alpha_t^S]_i w_i = \sum_{i=1}^{n} [\alpha_t^S]_i w_i^S = J^S \alpha_t^S. \tag{28}$$

And consequently, we have

$$\alpha_{t+1}^S = J^S \alpha_t^S + \beta_t^S, \tag{29}$$

where J^S is Schur-stable.

Consider system (29), since $\|\alpha_0^S\|_1 \leq \|\alpha_0\|_1 = \|M^{-1}x_0\|_1 \leq \|M^{-1}\|_1 \|x_0\|_1$, and x_0 is bounded, we see that $\|\alpha_0^S\|_1$ is bounded. Similarly, for all $t \in \mathbb{N}$, $\|\beta_t^S\|_1 \leq \|\beta_t\|_1 = \|M^{-1}Bu_t\|_1 \leq \|M^{-1}B\|_1 \|u_t\|_1 \leq \|M^{-1}B\|_1 \max\{\|u\|_1 : u \in \mathcal{U}\}$. Since $\mathcal{U}$ is a finite set in $\mathbb{R}^m$, $\max\{\|u\|_1 : u \in \mathcal{U}\}$ is finite, and consequently $\|\beta_t^S\|_1$ is uniformly bounded. Given system (29), since J^S is Schur-stable, $\|\alpha_0^S\|_1$ is bounded, and $\|\beta_t^S\|_1$ is uniformly bounded, we can show that $\|\alpha_t^S\|_1 < b_1$ for some $b_1 \in \mathbb{R}_+$ for all $t \in \mathbb{N}$ (an explicit derivation is presented in the proof of Theorem 1). Note that $\sum_{i:v_i \notin \mathcal{E}^U} |[\alpha_t]_i| = \|\alpha_t^S\|_1$, we have

$$\sum_{i:v_i \notin \mathcal{E}^U} |[\alpha_t]_i| < b_1, \; b_1 \in \mathbb{R}_+, \tag{30}$$

for all $t \in \mathbb{N}$.

Now we are ready to show Theorem 4. Use $\{e_i\}_{i=1}^n$ to denote the standard basis of $\mathbb{R}^n$,

and recall the computation of powers of a Jordan block (pp. 57, [8]), then for all $t \geq n$,

$$\begin{aligned} A^t v_i &= M J^t M^{-1} v_i \\ &= M J^t e_i \\ &= \sum_{j:\lambda_j=\lambda_i} \lambda_i^t \, p_{(j,i)}(t) \, v_j, \end{aligned}$$

where $p_{(j,i)}(t)$ is some polynomial in t, which depends on the pair (j,i). Recall the particular form of J_i^t, the upper triangular elements of J_i^t has the form $\begin{pmatrix} t \\ k \end{pmatrix} \lambda_i^{t-k}$, where $0 \leq k \leq n_i - 1$, and n_i corresponds to the size of J_i. Note that $\begin{pmatrix} t \\ k \end{pmatrix} \leq t^n$, and that $|\lambda_i|^{t-k} \leq \kappa |\lambda_i|^t$ for some $\kappa \in \mathbb{R}_+$: If $|\lambda_i| \geq 1$ or $|\lambda_i| = 0$, let $\kappa_1 = 1$; if $1 > |\lambda_i| > 0$, let $\kappa_2 = (\max\{1/|\lambda_j| : 0 < |\lambda_j| < 1, 1 \leq j \leq n\})^n$; take $\kappa = \max\{\kappa_1, \kappa_2\}$. Combine these observations, we conclude that for any $i \in \{1, \ldots, n\}$, and any $t \geq n$,

$$A^t v_i = \sum_{j:\lambda_j=\lambda_i} \lambda_i^t \, p_{(j,i)}(t) \, v_j, \tag{31}$$

where

$$|\lambda_i^t \, p_{(j,i)}(t)| \leq \kappa \cdot t^n |\lambda_i|^t, \tag{32}$$

for some $\kappa \in \mathbb{R}_+$.

For any $T \geq n$, and any $t \in \mathbb{N}$, recall (20), (31), we have

$$\begin{aligned} CA^T x_t &= CA^T(\sum_{i=1}^{n}[\alpha_t]_i v_i) \\ &= \sum_{i=1}^{n}[\alpha_t]_i CA^T v_i \\ &= \sum_{i=1}^{n}[\alpha_t]_i C \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) v_j \\ &= \sum_{i=1}^{n}[\alpha_t]_i \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) C v_j \\ &= \sum_{i:v_i\in\mathcal{E}^U}[\alpha_t]_i \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) C v_j + \sum_{i:v_i\notin\mathcal{E}^U}[\alpha_t]_i \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) C v_j \end{aligned} \tag{33}$$

If $v_i \in \mathcal{E}^U$, and $\lambda_j = \lambda_i$, then $\lambda_i \geq 1$, and therefore $\lambda_j \geq 1$, and $v_j \in \mathcal{E}^U$. Since $\mathcal{E}^U$ is in the kernel of C, for any i such that $v_i \in \mathcal{E}^U$, $Cv_j = 0$ for all j such that $\lambda_j = \lambda_i$. Therefore

$$\sum_{i:v_i\in\mathcal{E}^U}[\alpha_t]_i \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) C v_j = 0.$$

Continued from (33), we have

$$CA^T x_t = \sum_{i:v_i\notin\mathcal{E}^U}[\alpha_t]_i \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) C v_j.$$

Recall (32), we have

$$\begin{aligned} \|CA^T x_t\| &\leq \sum_{i:v_i\notin\mathcal{E}^U} |[\alpha_t]_i| \| \sum_{j:\lambda_j=\lambda_i} \lambda_i^T p_{(j,i)}(T) C v_j\| \\ &\leq \sum_{i:v_i\notin\mathcal{E}^U} |[\alpha_t]_i| \sum_{j:\lambda_j=\lambda_i} |\lambda_i^T p_{(j,i)}(T)| \|C v_j\| \\ &\leq \sum_{i:v_i\notin\mathcal{E}^U} |[\alpha_t]_i| \sum_{j:\lambda_j=\lambda_i} \kappa \cdot T^n |\lambda_i|^T \|C v_j\| \\ &= \sum_{i:v_i\notin\mathcal{E}^U} \kappa n \eta \cdot |[\alpha_t]_i| \; T^n \; |\lambda_i|^T, \end{aligned} \tag{34}$$

where $\eta = \max\{\|Cv_j\| : 1 \leq j \leq n\}$. For any i such that $v_i \notin \mathcal{E}^U$, $|\lambda_i| < 1$. Let $\rho = \max\{|\lambda_i| : v_i \notin \mathcal{E}^U\}$, then $0 \leq \rho < 1$. Consequently,

$$\begin{aligned}\|CA^T x_t\| &\leq \sum_{i:v_i \notin \mathcal{E}^U} \kappa n \eta \cdot |[\alpha_t]_i| T^n \rho^T \\ &\leq \kappa n \eta \cdot T^n \rho^T \sum_{i:v_i \notin \mathcal{E}^U} |[\alpha_t]_i|.\end{aligned}$$

Recall (30), we have

$$\|CA^T x_t\| \leq b_1 \kappa n \eta \cdot T^n \rho^T, \tag{35}$$

for any $T \geq n$, and any $t \in \mathbb{N}$.

Note that $0 \leq \rho < 1$, therefore $\lim_{T\to\infty} T^n \rho^T = 0$. Choose $T \in \mathbb{Z}_+$ such that

$$T^n \rho^T < \frac{d(\mathcal{A}, \mathcal{B})}{2b_1 \kappa n \eta}.$$

Then we have

$$\|CA^T x_t\| < d(\mathcal{A}, \mathcal{B})/2,$$

for all $t \in \mathbb{N}$. Then for all $t \geq T$:

$$y_t = CA^T x_{t-T} + \sum_{\tau=t-T}^{t-1} CA^{t-1-\tau} B u_\tau + D u_t \in B_{\frac{d(\mathcal{A},\mathcal{B})}{2}}(\alpha), \tag{36}$$

where $\alpha = \sum_{\tau=t-T}^{t-1} CA^{t-1-\tau} B u_\tau + D u_t$, and $\alpha \in \mathcal{A}$.

The rest of this derivation follows the exact same lines of the proof of Theorem 1 starting from equation (11). □

Next, we propose a necessary condition for (C1) when the A matrix is Schur-unstable. First, we introduce some notations. For any $\tilde{y} \in \mathcal{Q}$, we use $Q^{-1}(\tilde{y})$ to denote the set $\{x \in \mathbb{R}^p | Q(x) = \tilde{y}\}$.

Theorem 5. Consider system (2), assume that $\rho(A) \geq 1$, $0 \in \mathcal{U}$, and $0 \notin \mathcal{B}$. Define $\mathcal{V}^U = \{v \in \mathbb{C}^n \setminus 0 : Av = \lambda v, \text{ for some } |\lambda| > 1\}$. If $\mathcal{V}^U$ is not in the kernel of C, and

$Q^{-1}(Q(0))$ is bounded, then system (2) is not finite memory observable.

Remark. Compared with Theorem 2, the hypotheses in Theorem 5 implies $CA^l \neq \mathbf{0}$ for all $l \in \mathbb{Z}_+$. Otherwise, assume $CA^l = \mathbf{0}$ for some l. Since $\mathcal{V}^U$ is not in the kernel of C, there is $\lambda \in \mathbb{C}, v \in \mathbb{C}^n$ such that $Av = \lambda v$, $|\lambda| > 1$, and $Cv \neq 0$. Then $\|CA^l v\| = |\lambda|^l \|Cv\| \neq 0$, which draws a contradiction.

Proof. Since $\mathcal{V}^U$ is not in the kernel of C, there is $v \in \mathcal{V}^U$ such that $Cv \neq 0$. Without loss of generality, let $\|Cv\|_1 = 1$. Since $v \in \mathcal{V}^U$, we have $Av = \lambda v$ for some $|\lambda| > 1$, $\lambda \in \mathbb{C}$. Next, we define a set $\mathcal{O}$ as

$$\mathcal{O} = \{\alpha \in \mathbb{R}_+ | Re(\gamma Cv) \in Q^{-1}(Q(0)), \text{ for all } |\gamma| \leq \alpha, \gamma \in \mathbb{C}\}. \tag{37}$$

Next, we show that $\mathcal{O}$ is non-empty and bounded. Write $Cv = [v_1\ v_2\ \dots\ v_n]^T$, where $v_1, \dots, v_p \in \mathbb{C}$ and $|v_1| + \cdots + |v_p| = 1$. For any $\gamma \in \mathbb{C}$, we have $|Re(\gamma v_i)| \leq |\gamma||v_i|$, therefore

$$\|Re(\gamma Cv)\|_1 = \sum_{i=1}^{p} |Re(\gamma v_i)| \leq |\gamma| \sum_{i=1}^{p} |v_i| = |\gamma|.$$

Since Q is a piecewise-constant function, and $0 \notin \mathcal{B}$, there is $r > 0$ such that $B_r(0) \subset Q^{-1}(Q(0))$, where $B_r(0) = \{x \in \mathbb{R}^p : \|x\|_1 < r\}$. Therefore, for all γ with $|\gamma| \leq r/2$, $Re(\gamma Cv) \in B_r(0)$. Consequently $r/2 \in \mathcal{O}$, and $\mathcal{O}$ is nonempty.

Next, we show that $\mathcal{O}$ is bounded. Since $Q^{-1}(Q(0))$ is bounded by assumption, let $Q^{-1}(Q(0)) \subset B_\sigma(0)$ for some $\sigma > 0$. Since $Cv = [v_1\ v_2\ \dots\ v_n]^T \neq 0$, let $|v_k| > 0$ for some $1 \leq k \leq n$. Write v_k as $v_k = |v_k|e^{i\phi}$ for some $\phi \in [0, 2\pi)$. Assume $\mathcal{O}$ is unbounded, then there exist $\alpha \in \mathcal{O}$ with $\alpha > 2\sigma/|v_k|$. Let $\gamma = (2\sigma/|v_k|)e^{i(-\phi)}$, then $|\gamma| < \alpha$. By the definition of $\mathcal{O}$ (235), we have $Re(\gamma Cv) \in Q^{-1}(Q(0))$. Observe that

$$\|Re(\gamma Cv)\|_1 \geq |Re(\gamma v_k)| = |Re(\frac{2\sigma}{|v_k|}e^{i(-\phi)}|v_k|e^{i\phi})| = |Re(2\sigma)| = 2\sigma.$$

Therefore $Re(\gamma Cv) \notin B_\sigma(0)$, and consequently $Re(\gamma Cv) \notin Q^{-1}(Q(0))$, which draws a contradiction. Therefore $\mathcal{O}$ is bounded.

Next, we define $\beta = \sup \mathcal{O}$. Since $\mathcal{O}$ is non-empty and bounded, we have $0 < \beta < \infty$. Then for any $\epsilon > 0$, there is $\kappa \in \mathbb{C}$ such that

$$Re(\kappa C v) \notin Q^{-1}(Q(0)), \text{ and } \beta \leq |\kappa| < \beta + \epsilon. \tag{38}$$

and we will apply this observation to prove Theorem 5 by contradiction.

Assume system (2) is finite memory observable, than there exists an observer $\hat{S}$ (5) and $T \in \mathbb{Z}_+$ such that $\tilde{y}_t = \hat{y}_t$ for all $t \geq T$. Consider the input $u_t \equiv 0$, for two initial states x_0^1, $x_0^2 \in \mathbb{R}^n$, we use $\tilde{y}_t^1$, $\tilde{y}_t^2$ to denote the outputs of system (2) respectively. Choose $x_0^1 = 0$, then $\tilde{y}_t^1 = Q(0)$ for all $t \in \mathbb{N}$. In (38), let $\epsilon = \beta(|\lambda| - 1)$, and choose

$$x_0^2 = Re(\frac{\kappa}{\lambda^T} v).$$

Then for all $0 \leq t \leq T - 1$,

$$CA^t x_0^2 = Re(\frac{\kappa}{\lambda^T} CA^t v) = Re(\frac{\kappa \lambda^t}{\lambda^T} Cv).$$

Since $|\kappa \lambda^t / \lambda^T| < \beta$, by (235), we see that $CA^t x_0^2 \in Q^{-1}(Q(0))$, and consequently $\tilde{y}_t^2 = Q(0)$ for all $0 \leq t \leq T - 1$. At $t = T$, $CA^T x_0^2 = Re(\kappa C v) \notin Q^{-1}(Q(0))$, therefore $\tilde{y}_T^2 \neq Q(0)$. Now we see that $\tilde{y}_t^1 = \tilde{y}_t^2$ for $0 \leq t \leq T - 1$, and $\tilde{y}_T^1 \neq \tilde{y}_T^2$. Similar to the proof of Theorem 3, we can show that $\hat{y}_T^1 = \hat{y}_T^2$, and therefore either $\hat{y}_T^1 \neq \tilde{y}_T^1$ or $\hat{y}_T^2 \neq \tilde{y}_T^2$ or both, and we conclude that system (2) is not finite memory observable. □

3.5 Conditions for Weakly Observable and Asymptoticly Observable

In this section, we characterize conditions for (C2) and (C3). Since (C1) implies (C2), and (C2) implies (C3), the sufficient conditions for (C1) stated in the previous section are automatically sufficient conditions for (C2), (C3) as well. Therefore, we will focus on necessary conditions here. In particular, we propose a necessary condition for (C2) for the general class of systems (1), and then apply this result to characterize the necessary

conditions for (C2) and (C3) for the specific class of systems (2).

Before presenting the technical results, we first introduce some notations: For any $(\mathbf{u}, \mathbf{y}) \in \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$, and any $T \in \mathbb{Z}_+$, we use $(\{u_t\}_{t=0}^{T}, \{y_t\}_{t=0}^{T})$ to denote the segment of $(\mathbf{u}, \mathbf{y})$ from time 0 to time T. For any finite set $\mathcal{S} = \{s_1, s_2, \ldots, s_l\}$, where $l \in \mathbb{Z}_+$, we use $\{s_j\}_{j=1}^{l}$ as an abbreviation of $\{s_1, s_2, \ldots, s_l\}$. Now we are ready to propose a necessary condition for (C2) for the general class of systems (1).

Theorem 6. Given a system P (1) with $|\mathcal{U}| < \infty, 1 < |\mathcal{Y}| < \infty$, consider a family Ψ of input and output segments of P, where Ψ is described as

$$\Psi = \{\{(\{u_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}, \{y_t^{(k,j)}\}_{t=0}^{T_{(k,j)}})\}_{j=1}^{2^k}\}_{k=1}^{\infty}, \tag{39}$$

and the following items:

i. For any $k \in \mathbb{Z}_+$, and any $j \in \{1, \ldots, 2^k\}$, $T_{(k,j)} \in \mathbb{Z}_+$, and there is $(\mathbf{u}, \mathbf{y}) \in P$ such that $(\{u_t\}_{t=0}^{T_{(k,j)}}, \{y_t\}_{t=0}^{T_{(k,j)}}) = (\{u_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}, \{y_t^{(k,j)}\}_{t=0}^{T_{(k,j)}})$.

ii. For any $k \geq 1$, and any $j \in \{1, 2, \ldots, 2^{k-1}\}$,

$$T_{(k,2j-1)} = T_{(k,2j)}, \tag{40a}$$

$$u_t^{(k,2j-1)} = u_t^{(k,2j)},\ y_t^{(k,2j-1)} = y_t^{(k,2j)},\ t = 0, 1, \ldots, T_{(k,2j-1)} - 1, \tag{40b}$$

$$u_t^{(k,2j-1)} = u_t^{(k,2j)},\ y_t^{(k,2j-1)} \neq y_t^{(k,2j)},\ t = T_{(k,2j-1)}. \tag{40c}$$

iii. For any $k \geq 2$, and any $j \in \{1, 2, \ldots, 2^{k-1}\}$,

$$T_{(k,2j)} > T_{(k-1,j)}, \tag{41a}$$

$$u_t^{(k,2j)} = u_t^{(k-1,j)},\ y_t^{(k,2j)} = y_t^{(k-1,j)},\ t = 0, 1, \ldots, T_{(k-1,j)}. \tag{41b}$$

iv. For any sequence $\{j(k)\}_{k=1}^{\infty}$ that satisfies $j(k) \in \{1, \ldots, 2^k\}$ and $j(k+1) \in \{2j(k)-$

$1, 2j(k)\}$, define $(\mathbf{u}, \mathbf{y}) \in \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ as

$$\begin{aligned} u_t &= u_t^{(1,j(1))}, \quad y_t = y_t^{(1,j(1))}, \quad 0 \le t \le T_{(1,j(1))}, \\ u_t &= u_t^{(k,j(k))}, \quad y_t = y_t^{(k,j(k))}, \quad T_{(k-1,j(k-1))} < t \le T_{(k,j(k))}, \ \forall\, k \ge 2, \end{aligned} \tag{42}$$

then $(\mathbf{u}, \mathbf{y})$ satisfies

$$(\mathbf{u}, \mathbf{y}) \in P. \tag{43}$$

If there is a family Ψ (39) that satisfies items i) through iv), then system P is not weakly observable.

Proof. First, we make an observation regarding a system P (1) not being (C2). By Definition 3, P is not (C2) if and only if $\gamma = 0$ is not an observation gain of P. Then, by Definition 6, for any observer $\hat{S}$ (5), there is $(\mathbf{u}, \mathbf{y}) \in P$ such that

$$\sup_{T \ge 0} \sum_{t=0}^{T} \|y_t - \hat{y}_t\| = \infty.$$

Recall that $e_t = y_t - \hat{y}_t$, we see that P is not (C2) if and only if for any observer $\hat{S}$ (5), there is $(\mathbf{u}, \mathbf{y}) \in P$ such that

$$\sup_{T \ge 0} \sum_{t=0}^{T} \|e_t\| = \infty. \tag{44}$$

Assume the hypotheses in Theorem 6 is satisfied, then system P (1) and Ψ (39) are given. For *any* observer $\hat{S}$ (5), we define a family of its output segments $\hat{\Theta}$ as:

$$\hat{\Theta} = \{\{\{\hat{y}_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}\}_{j=1}^{2^k}\}_{k=1}^{\infty}, \tag{45}$$

where for all $k \in \mathbb{Z}_+$, $1 \le j \le 2^k$, $\{\hat{y}_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}$ is the output of $\hat{S}$ when u_t and y_t in equation (5) satisfy

$$\begin{aligned} u_t &= u_t^{(k,j)}, \text{ for } t = 0, 1, \ldots, T_{(k,j)}, \\ y_t &= y_t^{(k,j)}, \text{ for } t = 0, 1, \ldots, T_{(k,j)} - 1. \end{aligned} \tag{46}$$

Here $u_t^{(k,j)}, y_t^{(k,j)}$ are given by Ψ. Essentially, $\hat{y}_t^{(k,j)}$ is the output of $\hat{S}$ when $(u_t^{(k,j)}, y_t^{(k,j)})$ are applied to its input.

Given $\hat{S}$, and $\hat{\Theta}$ defined as in (45) (46), we claim that there is a sequence $\{j(k)\}_{k=1}^{\infty}$ such that for all $k \in \mathbb{Z}_+$, the following are satisfied:

$$j(k) \in \{1, \ldots, 2^k\}, \tag{47a}$$

$$\hat{y}_t^{(k,j(k))} \neq y_t^{(k,j(k))}, \text{ for all } t \in \{T_{(i,j(i))}\}_{i=1}^{k}, \tag{47b}$$

$$u_t^{(k,j(k))} = u_t^{(k-1,j(k-1))}, \ 0 \leq t \leq T_{(k-1,j(k-1))}, \tag{47c}$$

$$j(k) \in \{2j(k-1) - 1, \ 2j(k-1)\}. \tag{47d}$$

where (47c), (47d) are only required for $k \geq 2$.

We use induction to show this claim. For $k = 1$, first make an observation of the output $\hat{y}_t$ of the observer. By the dynamics of $\hat{S}$ (5), for any $t \in \mathbb{Z}_+$,

$$\hat{y}_t = g(f(\ldots f(f(q_o, u_0, y_0), u_1, y_1) \ldots, u_{t-1}, y_{t-1}), u_t). \tag{48}$$

Recall (40), we have $u_t^{(1,1)} = u_t^{(1,2)}$, for $t = 0, 1, \ldots, T_{(1,1)}$, and $y_t^{(1,1)} = y_t^{(1,2)}$, for $t = 0, 1, \ldots, T_{(1,1)} - 1$. Let $t = T_{(1,1)}$ in (48), and recall (46), we have $\hat{y}_{T_{(1,1)}}^{(1,1)} = \hat{y}_{T_{(1,1)}}^{(1,2)}$. Recall (40c), $y_{T_{(1,1)}}^{(1,1)} \neq y_{T_{(1,1)}}^{(1,2)}$. Consequently, there is $j^* \in \{1, 2\}$ such that $\hat{y}_{T_{(1,1)}}^{(1,j^*)} \neq y_{T_{(1,1)}}^{(1,j^*)}$. Let $j(1) = j^*$, then $\hat{y}_t^{(1,j(1))} \neq y_t^{(1,j(1))}$, for $t = T_{(1,j(1))}$, therefore (47) holds at $k = 1$.

For $k = 2$, recall (41a), $T_{(2,2)} > T_{(1,1)}$, and $T_{(2,4)} > T_{(1,2)}$. Recall (40a), $T_{(2,j)} > T_{(1,1)}$ for all $j \in \{1, 2, 3, 4\}$. Recall (41b), $u_t^{(2,2)} = u_t^{(1,1)}, u_t^{(2,4)} = u_t^{(1,2)}, t = 0, 1, \ldots, T_{(1,1)}$. Recall (40b), (40c), $u_t^{(2,1)} = u_t^{(2,2)} = u_t^{(1,1)}, u_t^{(2,3)} = u_t^{(2,4)} = u_t^{(1,2)}, \ t = 0, 1, \ldots, T_{(1,1)}$. Similarly, by (41b), (40b), (40c), $y_t^{(2,1)} = y_t^{(2,2)} = y_t^{(1,1)}, y_t^{(2,3)} = y_t^{(2,4)} = y_t^{(1,2)}, t = 0, 1, \ldots, T_{(1,1)}$. Recall (48),

$$\hat{y}_t^{(2,1)} = \hat{y}_t^{(2,2)} = \hat{y}_t^{(1,1)}, \hat{y}_t^{(2,3)} = \hat{y}_t^{(2,4)} = \hat{y}_t^{(1,2)}, t = T_{(1,1)}. \tag{49}$$

Since (47) holds at $k = 1$, $\hat{y}_t^{(1,j(1))} \neq y_t^{(1,j(1))}$, for $t = T_{(1,j(1))} = T_{(1,1)}$. Recall $j(1) \in \{1, 2\}$,

by (49), $\hat{y}_t^{(2,2j(1)-1)} = \hat{y}_t^{(2,2j(1))} = \hat{y}_t^{(1,j(1))}, t = T_{(1,1)}$. Recall (41b), $y_t^{(2,2j(1)-1)} = y_t^{(2,2j(1))} = y_t^{(1,j(1))}, t = T_{(1,1)}$. Therefore,

$$\hat{y}_t^{(2,2j(1)-1)} \neq y_t^{(2,2j(1)-1)},\ \hat{y}_t^{(2,2j(1))} \neq y_t^{(2,2j(1))}, t = T_{(1,1)}. \tag{50}$$

Next, recall (40), (48), we have $\hat{y}_t^{(2,2j(1)-1)} = \hat{y}_t^{(2,2j(1))}, t = T_{(2,2j(1))}$. Recall (40c),

$$y_t^{(2,2j(1)-1)} \neq y_t^{(2,2j(1))}, t = T_{(2,2j(1)-1)} = T_{(2,2j(1))}.$$

Consequently, there is $j^* \in \{2j(1)-1, 2j(1)\} \subset \{1,2,3,4\}$ such that $\hat{y}_t^{(2,j^*)} \neq y_t^{(2,j^*)}$, $t = T_{(2,j^*)}$. Let $j(2) = j^*$, then (47a), (47d) hold. By (50), and $\hat{y}_t^{(2,j^*)} \neq y_t^{(2,j^*)}$, $t = T_{(2,j^*)}$, (47b) holds. By (40), (41b), we see (47c) holds. Therefore (47) holds at $k = 2$.

Assume that (47) holds for some $k \geq 2$. Recall (40b), (40c), (41b), $u_t^{(k+1,2j(k)-1)} = u_t^{(k+1,2j(k))} = u_t^{(k,j(k))}$, and $y_t^{(k+1,2j(k)-1)} = y_t^{(k+1,2j(k))} = y_t^{(k,j(k))}$, $t = 0, 1, \ldots, T_{(k,j(k))}$. Recall (46), (48), we see that

$$\hat{y}_t^{(k+1,2j(k)-1)} = \hat{y}_t^{(k+1,2j(k))} = \hat{y}_t^{(k,j(k))}, t = 0, 1, \ldots, T_{(k,j(k))}.$$

Since (47b) holds for k, we see that

$$\hat{y}_t^{(k+1,2j(k)-1)} \neq y_t^{(k+1,2j(k)-1)}, \quad \hat{y}_t^{(k+1,2j(k))} \neq y_t^{(k+1,2j(k))}, \text{ for all } t \in \{T_{(i,j(i))}\}_{i=1}^k. \tag{51}$$

At $t = T_{(k+1,2j(k))}$, by (40), (48),

$$\hat{y}_t^{(k+1,2j(k)-1)} = \hat{y}_t^{(k+1,2j(k))}, t = T_{(k+1,2j(k))}.$$

By (40a), (40c), $y_t^{(k+1,2j(k)-1)} \neq y_t^{(k+1,2j(k))}, t = T_{(k+1,2j(k))}$. Therefore there is $j^* \in \{2j(k)-1, 2j(k)\}$ such that $\hat{y}_t^{(k+1,j^*)} \neq y_t^{(k+1,j^*)}$ at $t = T_{(k+1,2j(k))} = T_{(k+1,j^*)}$. Let $j(k+1) = j^*$, and recall (51), we see that (47b) holds for $k+1$. Since $j(k+1) = j^* \in \{2j(k)-1, 2j(k)\}$, and (47a) holds for k, we see that (47a), (47d) holds for $k+1$. By (40),

(41b), we see (47c) holds for $k+1$. We see that (47) holds for $k+1$. This completes the derivation of the existence of $\{j(k)\}_{k=1}^{\infty}$ such that (47) holds for all $k \in \mathbb{Z}_+$.

Given $\hat{S}$, let $\{j(k)\}_{k=1}^{\infty}$ be such that (47) holds, define an input and output pair $(\mathbf{u}, \mathbf{y}) \in \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ as

$$\begin{aligned} &u_t = u_t^{(1,j(1))}, \quad y_t = y_t^{(1,j(1))}, \quad 0 \le t \le T_{(1,j(1))}, \\ &u_t = u_t^{(k,j(k))}, \quad y_t = y_t^{(k,j(k))}, \quad T_{(k-1,j(k-1))} < t \le T_{(k,j(k))}, \ \forall\, k \ge 2. \end{aligned} \tag{52}$$

First we show that $(\mathbf{u}, \mathbf{y})$ (52) is well-defined.

Let $L(k) = \{T_{(k-1,j(k-1))} + 1, \ldots, T_{(k,j(k))}\} \subset \mathbb{Z}_+$ for all $k \ge 2$. We observe that

$$\{0, 1, \ldots, T_{(1,j(1))}\} \cup (\bigcup_{k=2}^{\infty} L(k)) = \mathbb{N}, \tag{53a}$$

$$L(k_1) \cap L(k_2) = \varnothing, \text{ if } k_1 \ne k_2. \tag{53b}$$

To see this, by (47d), (40a), (41a),

$$T_{(k,j(k))} > T_{(k-1,j(k-1))}, \tag{54}$$

or equivalently $T_{(k,j(k))} \ge T_{(k-1,j(k-1))} + 1$, for all $k \ge 2$. Consequently, $T_{(k,j(k))} \ge T_{(1,j(1))} + (k-1)$, for all $k \ge 2$, and

$$\sup_{k \in \mathbb{Z}_+} T_{(k,j(k))} = \infty. \tag{55}$$

For any $t \in \mathbb{N}$, if $t > T_{(1,1)}$, let $k(t) = \min\{k \in \mathbb{Z}_+ : T_{(k,j(k))} \ge t\}$. Since $\{k \in \mathbb{Z}_+ : T_{(k,j(k))} \ge t\}$ is none-empty, $k(t)$ is well-defined (pp. 28, [22]), and $t \in L(k(t))$. Therefore (53a) holds. For any $k_1, k_2 \ge 2$, and $k_1 \ne k_2$, without loss of generality, let $k_1 < k_2$. Assume $t^* \in L(k_1) \cap L(k_2)$, then $t^* > T_{(k_2-1,j(k_2-1))}$, and $t^* \le T_{(k_1,j(k_1))}$. Since $k_2 - 1 \ge k_1$, by (54), $T_{(k_2-1,j(k_2-1))} \ge T_{(k_1,j(k_1))}$, but $T_{(k_2-1,j(k_2-1))} < t^* \le T_{(k_1,j(k_1))}$, which draws a contradiction. Therefore (53b) holds. Consequently, $(\mathbf{u}, \mathbf{y})$ (52) is well-defined.

By (47a), (47d), and item iv) of Ψ (see (43)), $(\mathbf{u}, \mathbf{y})$ (52) satisfies $(\mathbf{u}, \mathbf{y}) \in P$.

Next, we apply the observation made in (44) to show P is not (C2). For $\hat{S}$, consider $(\mathbf{u}, \mathbf{y}) \in P$ in (52). As in the setup in Figure 2, let $\{\hat{y}_t\}_{t=0}^{\infty}$ be the output of $\hat{S}$ corresponding with $(\mathbf{u}, \mathbf{y})$ (52). Recall (52), we see that $u_t = u_t^{(1,j(1))}$, $0 \leq t \leq T_{(1,j(1))}$. Recall (41b), (40c), (47d), we see that $u_t^{(1,j(1))} = u_t^{(2,2j(1))} = u_t^{(2,2j(1)-1)} = u_t^{(2,j(2))}$, $0 \leq t \leq T_{(1,j(1))}$. Repeat this argument, for any $k \in \mathbb{Z}_+$, we have $u_t = u_t^{(k,j(k))}$, $0 \leq t \leq T_{(1,j(1))}$. Similarly, we can show that for any $k \geq 2$, $u_t = u_t^{(k,j(k))}$, $T_{(1,j(1))} + 1 \leq t \leq T_{(2,j(2))}$. Repeat this argument, then for any $k \in \mathbb{Z}_+$, we have $u_t = u_t^{(k,j(k))}$, $0 \leq t \leq T_{(k,j(k))}$. Similarly, we can show that for any $k \in \mathbb{Z}_+$, $y_t = y_t^{(k,j(k))}$, $0 \leq t \leq T_{(k,j(k))} - 1$.

Recall the definition of $\hat{\Theta}$ (see (46)), and (48), we have

$$\hat{y}_t = \hat{y}_t^{(k,j(k))}, t = T_{(k,j(k))}, \text{ for all } \ k \in \mathbb{Z}_+.$$

Recall (52), $y_t = y_t^{(k,j(k))}, t = T_{(k,j(k))}, \text{ for all } k \in \mathbb{Z}_+$.

Recall (47b), $\hat{y}_t^{(k,j(k))} \neq y_t^{(k,j(k))}, t = T_{(k,j(k))}, \text{ for all } k \in \mathbb{Z}_+$. Therefore, for all $k \in \mathbb{Z}_+$,

$$\hat{y}_t \neq y_t, \quad t = T_{(k,j(k))}. \tag{56}$$

Let $\delta = \min\{\|y_1 - y_2\| : y_1 \neq y_2, y_1, y_2 \in \mathcal{Y}\}$, since $|\mathcal{Y}| < \infty$, the minimum is well-defined and $\delta > 0$. By (56), $\|e_t\| \geq \delta$ for $t = T_{(k,j(k))}$, $k \in \mathbb{Z}_+$. Consequently, $\sum_{t=0}^{T_{(k,j(k))}} \|e_t\| \geq k\delta$. Recall (55), we see that $\sup_{T \geq 0} \sum_{t=0}^{T} \|e_t\| = \infty$. By (44), P is not (C2). □

Next, we study the necessary conditions for (C2) for systems (2). As shown in the preceding, the hypotheses in Theorem 6 might seem a bit abstract, and one might ask whether systems satisfying such hypotheses exist in practice. Therefore, in the following, we will present a concrete example of Theorem 6 when it is applied to systems (2).

For the purpose of exposition, we consider systems (2) with the underlying LTI system being a scalar system.

Theorem 7. Consider system (2) with $n = m = p = 1$, $A > 1$, $0 \in \mathcal{U}$, and Q is of the form (4). Assume $\mathcal{B} \cap \mathbb{R}_+ \neq \varnothing$, and let $\beta = \operatorname{argmin}\{|b| : b \in \mathcal{B}, b > 0\}$. If there is $u^* \in \mathcal{U}$ such

that $CBu^* + A^2\beta = 0$, then the following hold:

i. System (2) is not (C2).

ii. Moreover, system (2) is not (C3).

Remark. By Theorem 5, systems that satisfy the hypotheses in Theorem 7 are not (C1). And compared to the necessary condition of (C1), the requirement "$CBu^* + A^2\beta = 0$" is new.

Proof. The proof idea is to show there is Ψ (39) such that the hypotheses in Theorem 6 is satisfied. With out loss of generality, we consider the case where C is positive. The other case could be shown in a similar fashion. Note that $CBu^* + A^2\beta = 0$ implies $C \neq 0$.

Within the scope of this derivation, we use "s" to denote the initial state x_0 of system (2). Given system (2), define a quantity s_o as

$$s_o = \frac{\beta}{CA^T}, \tag{57}$$

where $T \in \mathbb{Z}_+$, and $T \geq 2$ is to be determined. Define a family $\mathcal{S}$ of initial states of system (2) as

$$\mathcal{S} = \{\{s_{(k,j)}\}_{j=1}^{2^k}\}_{k=1}^{\infty}, \tag{58}$$

where

$$s_{(1,1)} = 0, \quad s_{(1,2)} = s_o, \tag{59}$$

and for all $k \geq 2$, all $j = 1, 2, \ldots, 2^{k-1}$,

$$s_{(k,2j-1)} = s_{(k-1,j)}, \quad s_{(k,2j)} = s_{(k-1,j)} + (A^{-T})^{k-1}s_o. \tag{60}$$

Then $s_{(k,j)}$ is defined for all $k \in \mathbb{Z}_+$ and $j \in \{1, \ldots, 2^k\}$.

Next, define a family $\mathcal{I}$ of input segments of system (2) as

$$\mathcal{I} = \{\{\{u_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}\}_{j=1}^{2^k}\}_{k=1}^{\infty}, \tag{61}$$

where for all $k \in \mathbb{Z}_+$, all $j = 1, 2, \ldots, 2^k$,

$$T_{(k,j)} = k \cdot T, \tag{62}$$

and

$$u_t^{(1,1)} = u_t^{(1,2)} = 0,\ t = 0, 1, \ldots, T, \tag{63}$$

and for all $k \geq 2$, all $j = 1, 2, \ldots, 2^{k-2}$,

$$\begin{aligned} u_t^{(k,4j)} = u_t^{(k,4j-1)} = u_t^{(k-1,2j)},\ u_t^{(k,4j-2)} = u_t^{(k,4j-3)} = u_t^{(k-1,2j-1)}, \quad 0 \leq t \leq (k-1)T, \\ u_t^{(k,4j)} = u_t^{(k,4j-1)} = u^*, \quad u_t^{(k,4j-2)} = u_t^{(k,4j-3)} = 0, \quad t = (k-1)T + 1, \\ u_t^{(k,4j)} = u_t^{(k,4j-1)} = u_t^{(k,4j-2)} = u_t^{(k,4j-3)} = 0, \quad (k-1)T + 2 \leq t \leq kT, \end{aligned} \tag{64}$$

Then $u_t^{(k,j)}$ is defined for all $k \in \mathbb{Z}_+$, $j \in \{1, \ldots, 2^k\}$, and $t \in \{0, \ldots, kT\}$.

Given $\mathcal{S}$ and $\mathcal{I}$ defined in the preceding, define a family $\tilde{\Theta}$ of output segments of system (2) as:

$$\tilde{\Theta} = \{\{\{\tilde{y}_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}\}_{j=1}^{2^k}\}_{k=1}^{\infty}, \tag{65}$$

where for all $k \in \mathbb{Z}_+$, $1 \leq j \leq 2^k$, $\{\tilde{y}_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}$ is the quantized output $\tilde{y}_t$ (2c) of system (2), when u_t and x_t in equation (2) satisfy

$$\begin{aligned} u_t &= u_t^{(k,j)},\ \text{for } t = 0, 1, \ldots, T_{(k,j)}, \\ x_t &= s_{(k,j)},\ \text{for } t = 0. \end{aligned} \tag{66}$$

Essentially, $\tilde{y}_t^{(k,j)}$ is the quantized output of system (2) when $u_t^{(k,j)}$ is applied to its input, and its initial state is $s_{(k,j)}$. In the following, we also use $x_t^{(k,j)}$ to denote the state x_t of system (2) corresponding with (66).

Given $\mathcal{I}$ (61), and $\tilde{\Theta}$ (65), define Ψ as

$$\Psi = \{\{(\{u_t^{(k,j)}\}_{t=0}^{T_{(k,j)}}, \{\tilde{y}_t^{(k,j)}\}_{t=0}^{T_{(k,j)}})\}_{j=1}^{2^k}\}_{k=1}^{\infty}, \tag{67}$$

where $u_t^{(k,j)}$, $\tilde{y}_t^{(k,j)}$ are given by $\mathcal{I}$ and $\tilde{\Theta}$ respectively. In the following, we will show that Ψ (67) satisfies items i) to iv) in Theorem 6, and therefore the system is not (C2).

Regarding item i), for any $k \in \mathbb{Z}_+$, $1 \leq j \leq 2^k$, let $\mathbf{u}$ and x_0 of system (2) be such that (66) is satisfied, and $u_t = 0$, for $t > T_{(k,j)}$, and let $\tilde{\mathbf{y}}$ be the corresponding output of system (2). By the definition of $\tilde{\Theta}$, we see that item i) is satisfied.

Next, we use induction to show Ψ (67) satisfies items ii) and iii). For $k = 1$, recall (62), $T_{(1,1)} = T_{(1,2)} = T$. Recall (63), $u_t^{(1,1)} = u_t^{(1,2)} = 0, t = 0, 1, \ldots, T$. By the definition of $\tilde{\Theta}$ (66), and $s_{(1,1)} = 0$, $\tilde{y}_t^{(1,1)} = Q(0)$ for $t = 0, 1, \ldots, T$. Recall $s_{(1,2)} = s_o$, $A > 1$, $C > 0$, and (2), then for all $t = 0, 1, \ldots, T-1$, $y_t^{(1,2)} = CA^t s_o < CA^T \frac{\beta}{CA^T} = \beta$. Recall (4), and $\beta = \operatorname{argmin}\{|b| : b \in \mathcal{B}, b > 0\}$, $\tilde{y}_t^{(1,2)} = Q(0)$ for $t = 0, 1, \ldots, T-1$. At $t = T$, $\tilde{y}_t^{(1,2)} = Q(CA^T \frac{\beta}{CA^T}) = Q(\beta) \neq Q(0)$. Therefore, $\tilde{y}_t^{(1,1)} = \tilde{y}_t^{(1,2)}$, $t = 0, 1, \ldots, T-1$, and $\tilde{y}_t^{(1,1)} \neq \tilde{y}_t^{(1,2)}$, $t = T$, and item ii) is satisfied for $k = 1$.

For $k = 2$, recall (62), for any $1 \leq j \leq 4$, $T_{(2,j)} = 2T$. Recall (59), (60), $s_{(2,1)} = 0, s_{(2,2)} = A^{-T} s_o, s_{(2,3)} = s_o, s_{(2,4)} = s_o + A^{-T} s_o$. By (63), (64), $u_t^{(2,1)} = u_t^{(2,2)} = 0$ for $0 \leq t \leq 2T$, and $u_t^{(2,3)} = u_t^{(2,4)} = 0$ for $0 \leq t \leq 2T$ and $t \neq T+1$, $u_t^{(2,3)} = u_t^{(2,4)} = u^*$ for $t = T+1$. Therefore $\tilde{y}_t^{(2,1)} = Q(0)$ for $t = 0, 1, \ldots, 2T$. For all $t = 0, 1, \ldots, 2T-1$, $CA^t(A^{-T} s_o) < CA^{2T}(A^{-T} \frac{\beta}{CA^T}) = \beta$, and therefore $\tilde{y}_t^{(2,2)} = Q(CA^t(A^{-T} s_o)) = Q(0)$ for $t = 0, 1, \ldots, 2T-1$. At $t = 2T$, $\tilde{y}_t^{(2,2)} = Q(CA^{2T} A^{-T} \frac{\beta}{CA^T}) = Q(\beta) \neq Q(0)$. Consequently, items ii) and iii) are satisfied when $k = 2, j = 1$.

For $k = 2$ and $j = 2$, by the derivation when $k = 1$, and note that $s_{(2,3)} = s_{(1,2)}$, $u_t^{(2,3)} = u_t^{(1,2)}$ for $0 \leq t \leq T$, we see that $\tilde{y}_t^{(2,3)} = Q(0)$ for $t = 0, 1, \ldots, T-1$, and $\tilde{y}_t^{(2,3)} = Q(\beta)$ for $t = T$. At $t = T+1$, $\tilde{y}_t^{(2,3)} = Q(CA^{T+1} \frac{\beta}{CA^T} + Du^*) = Q(A\beta + Du^*)$. At $t = T+2$, $\tilde{y}_t^{(2,3)} = Q(CA^{T+2} s_o + CBu^*) = Q(CA^{T+2} \frac{\beta}{CA^T} + CBu^*) = Q(A^2\beta + CBu^*) = Q(0)$ by assumption. Also note that for the system state, $x_t^{(2,3)} = A^{T+2} \frac{\beta}{CA^T} + Bu^* = 0$ for $t = T+2$. Therefore, $x_t^{(2,3)} = 0$, and $\tilde{y}_t^{(2,3)} = Q(0)$ for $t = T+2, \ldots, 2T$. We summarize

the preceding as

$$\tilde{y}_t^{(2,3)} = \begin{cases} Q(0), & t = 0, 1, \ldots, T-1, \\ Q(\beta), & t = T, \\ Q(A\beta + Du^*), & t = T+1, \\ Q(0), & t = T+2, \ldots, 2T. \end{cases} \tag{68}$$

For $\tilde{y}_t^{(2,4)}$, at $t = T-1$, $\tilde{y}_t^{(2,4)} = Q(CA^{T-1}(s_o + A^{-T}s_o)) = Q(A^{-1}(1 + A^{-T})\beta)$. Choose $T \in \mathbb{Z}_+$ such that

$$A^T > \frac{A}{A-1}. \tag{69}$$

Since $A > 1$, such a choice of T always exist, for example let $T > log_A(\frac{A}{A-1})$. Then $A^{-T} < 1 - \frac{1}{A}$, and $\frac{1}{A} < 1 - A^{-T}$, or equivalently $A^{-1}\frac{1}{1-A^{-T}} < 1$. Since $A^{-1}(1+A^{-T})\beta < A^{-1}(1 + A^{-T} + A^{-2T} + \cdots)\beta = A^{-1}\frac{1}{1-A^{-T}}\beta < \beta$, and recall $\tilde{y}_t^{(2,4)} = Q(A^{-1}(1+A^{-T})\beta)$ for $t = T-1$, we have $\tilde{y}_t^{(2,4)} = Q(0)$ for $t = T-1$. Since $A > 1$, $C > 0$, we see that $\tilde{y}_t^{(2,4)} = Q(0)$ for all $0 \leq t \leq T-1$. At $t = T$, $\tilde{y}_t^{(2,4)} = Q(CA^T(s_o + A^{-T}s_o)) = Q((1 + A^{-T})\beta)$. Recall Q (4) is right-continuous and piecewise-constant, for $\beta \in \mathbb{R}$, there is $\delta_1 > 0$ such that

$$Q(y) = Q(\beta), \ \forall\, y \in [\beta, \beta + \delta_1).$$

Choose $T \in \mathbb{Z}_+$ such that

$$\frac{1}{1 - A^{-T}}\beta < \beta + \delta_1. \tag{70}$$

Again, since $A > 1$, (70) is satisfied for all T sufficiently large. Then $\beta < (1 + A^{-T})\beta < \frac{1}{1-A^{-T}}\beta < \beta + \delta_1$, and therefore $\tilde{y}_t^{(2,4)} = Q(\beta)$ for $t = T$. At $t = T+1$, $\tilde{y}_t^{(2,4)} = Q(CA^{T+1}(s_o + A^{-T}s_o) + Du^*) = Q((1 + A^{-T})A\beta + Du^*)$. Again, by the assumptions of Q, for $A\beta + Du^* \in \mathbb{R}$, there is $\delta_2 > 0$ such that

$$Q(y) = Q(A\beta + Du^*), \ \forall\, y \in [A\beta + Du^*, A\beta + Du^* + \delta_2).$$

Choose $T \in \mathbb{Z}_+$ such that

$$\frac{1}{1-A^{-T}} A\beta < A\beta + \delta_2. \tag{71}$$

Then $(1+A^{-T})A\beta + Du^* < A\beta + Du^* + \delta_2$, and therefore $\tilde{y}_t^{(2,4)} = Q(A\beta + Du^*)$, $t = T+1$. At $t = T+2$, the system state $x_t^{(2,4)} = A^{T+2}(s_o + A^{-T}s_o) + Bu^* = A^2 s_o + A^{T+2}\frac{\beta}{CA^T} + Bu^* = A^2 s_o$. Recall $u_t^{(2,4)} = 0$ for $T+2 \le t \le 2T$, therefore $x_t^{(2,4)} = A^{t-T}s_o$ for all $T+2 \le t \le 2T$. For all $T+2 \le t \le 2T-1$, $Cx_t^{(2,4)} = CA^{t-T}s_o < CA^T(\frac{\beta}{CA^T})$ = β, and therefore $\tilde{y}_t^{(2,4)} = Q(Cx_t^{(2,4)}) = Q(0)$ for $T+2 \le t \le 2T-1$. At $t = 2T$, $\tilde{y}_t^{(2,4)} = Q(CA^{2T-T}s_o) = Q(CA^T\frac{\beta}{CA^T}) = Q(\beta)$. Again, we summarize the preceding as

$$\tilde{y}_t^{(2,4)} = \begin{cases} Q(0), & t = 0, 1, \ldots, T-1, \\ Q(\beta), & t = T, \\ Q(A\beta + Du^*), & t = T+1, \\ Q(0), & t = T+2, \ldots, 2T-1, \\ Q(\beta), & t = 2T. \end{cases}$$

Recall (68), and the explicit form of $\tilde{y}_t^{(1,2)}$, we see that items ii) and iii) are satisfied when $k = 2, j = 2$.

We conclude that Ψ (67) satisfies items ii) and iii) in Theorem 6 when $k = 2$.

Before invoking induction to show items ii) and iii) are satisfied for all k, we first make some observations about $\mathcal{S}$ and $\mathcal{I}$. Recall the family of initial states $\mathcal{S}$ defined in (58), (59), (60), we observe that for any $k \in \mathbb{Z}_+$ and $j \in \{1, \ldots, 2^k\}$,

$$s_{(k,j)} = (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1}) \cdot s_o, \tag{72}$$

where $q = A^{-T}$, and $\alpha_l^{(k,j)}$, $l \in \{1, \ldots, k\}$, is defined to be

$$\alpha_l^{(k,j)} = \begin{cases} 0, & \text{if } 0 \le (j-1) \mod (2^{k-l+1}) < \frac{2^{k-l+1}}{2}, \\ 1, & \text{if } \frac{2^{k-l+1}}{2} \le (j-1) \mod (2^{k-l+1}) < 2^{k-l+1}. \end{cases} \tag{73}$$

Essentially, the above is the explicit form of $s_{(k,j)}$.

We use induction to show the preceding is valid. For $k = 1$, by (73), $\alpha_1^{(1,1)} = 0$, $\alpha_1^{(1,2)} = 1$. Recall (59), we see that (72) hold for $k = 1$.

Next, assume (72) hold for some $k \geq 1$, about $\alpha_l^{(k,j)}$ (73), observe that for any $l \in \{1, \ldots, k\}$

$$\alpha_l^{(k,j)} = \alpha_l^{(k+1,2j-1)} = \alpha_l^{(k+1,2j)}. \tag{74}$$

To see this, let $(j-1) \mod 2^{k-l+1} = b$, for some $0 \leq b < 2^{k-l+1}$. Then $j - 1 = 2^{k-l+1} \cdot a + b$, for some unique $a \in \mathbb{N}$ (pp. 32, [22]). Therefore $2j - 2 = 2^{(k+1)-l+1} \cdot a + 2b$, and $2j-2 = 2^{(k+1)-l+1} \cdot a+2b+1$. Therefore $((2j-1)-1) \mod 2^{(k+1)-l+1} = 2b$, and $(2j-1) \mod 2^{(k+1)-l+1} = 2b+1$. If $\alpha_l^{(k,j)} = 0$, then $0 \leq b < 2^{k-l+1}/2$. Then $0 \leq b \leq 2^{k-l+1}/2-1$, and $0 \leq 2b \leq 2^{(k+1)-l+1}/2 - 2$. Therefore, $0 \leq 2b < 2b + 1 < 2^{(k+1)-l+1}/2$, by (73), $\alpha_l^{(k+1,2j-1)} = \alpha_l^{(k+1,2j)} = 0$. Similarly, we can show that if $\alpha_l^{(k,j)} = 1$, then $\alpha_l^{(k+1,2j-1)} = \alpha_l^{(k+1,2j)} = 1$. Therefore (74) holds. Also observe that for all $j = 1, 2, \ldots, 2^k$, by (73), $\alpha_{k+1}^{(k+1,2j-1)} = 0$, and $\alpha_{k+1}^{(k+1,2j)} = 1$. Recall (60), for any $1 \leq j \leq 2^k$

$$\begin{aligned}
s_{(k+1,2j-1)} &= s_{(k,j)} \\
&= (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1}) \cdot s_o \\
&= (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1}) \cdot s_o + \alpha_{k+1}^{(k+1,2j-1)} q^k s_o \\
&= (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1} + \alpha_{k+1}^{(k+1,2j-1)} q^k) \cdot s_o.
\end{aligned}$$

Similarly,

$$\begin{aligned}
s_{(k+1,2j)} &= s_{(k,j)} + q^k \cdot s_o \\
&= (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1}) \cdot s_o + 1 \cdot q^k \cdot s_o \\
&= (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1}) \cdot s_o + \alpha_{k+1}^{(k+1,2j)} q^k s_o \\
&= (\alpha_1^{(k,j)} + \alpha_2^{(k,j)} \cdot q + \cdots + \alpha_k^{(k,j)} \cdot q^{k-1} + \alpha_{k+1}^{(k+1,2j-1)} \cdot q^k) \cdot s_o.
\end{aligned}$$

Therefore (72) hold for $k + 1$. By induction, (72) hold for all $k \in \mathbb{Z}_+$.

Next, we also make some observations on $\mathcal{I}$ (61). For any $k \in \mathbb{Z}_+$, and any $j \in \mathbb{Z}_+$, define a function $h(k,j)$ as

$$h(k,j) = ((j-1) \mod (2^{k-1})) + 1. \tag{75}$$

Then for all $k \geq 2$, $j \in \{1, 2, \ldots, 2^k\}$, $\mathcal{I}$ (61) satisfies:

$$u_t^{(k,j)} = 0, \quad t = 0, 1, \ldots, T, \tag{76a}$$

$$u_t^{(k,j)} = \alpha_1^{(k,j)} u^*, \quad t = T+1, \tag{76b}$$

$$u_t^{(k,j)} = u_{t-T}^{(k-1,h(k,j))}, \quad t = T+2, T+3, \ldots, kT. \tag{76c}$$

We use induction to show (76) holds. For $k = 2$, recall $u_t^{(2,1)} = u_t^{(2,2)} = 0$ for $0 \leq t \leq 2T$, $u_t^{(2,3)} = u_t^{(2,4)} = 0$ for $0 \leq t \leq 2T$ and $t \neq T+1$, $u_t^{(2,3)} = u_t^{(2,4)} = u^*$ for $t = T+1$, and $u_t^{(1,1)} = u_t^{(1,2)} = 0$, for $0 \leq t \leq T$. By (73), $\alpha_1^{(2,1)} = \alpha_1^{(2,2)} = 0$, $\alpha_1^{(2,3)} = \alpha_1^{(2,4)} = 1$. By (75), $h(2,1) = 1, h(2,2) = 2, h(2,3) = 1, h(2,4) = 2$. We see that (76) holds for $k = 2$.

Assume (76) holds for some $k \geq 2$. Recall (64),

$$u_t^{(k+1,2j-1)} = u_t^{(k+1,2j)} = u_t^{(k,j)}, t = 0, 1, \ldots, kT, \tag{77}$$

for all $j \in \{1, \ldots, 2^k\}$. By assumption, $u_t^{(k,j)} = 0, t = 0, 1, \ldots, T$, $j \in \{1, 2, \ldots, 2^k\}$, therefore $u_t^{(k+1,2j-1)} = u_t^{(k+1,2j)} = 0$, $t = 0, 1, \ldots, T$, $j \in \{1, 2, \ldots, 2^k\}$, and (76a) holds for $k+1$. At $t = T+1$, by (77) and (76b), $u_t^{(k+1,2j-1)} = u_t^{(k+1,2j)} = u_t^{(k,j)} = \alpha_1^{(k,j)} u^*$. Recall (74), $\alpha_1^{(k,j)} = \alpha_1^{(k+1,2j-1)} = \alpha_1^{(k+1,2j)}$. Therefore $u_t^{(k+1,2j-1)} = \alpha_1^{(k+1,2j-1)} u^*$, and $u_t^{(k+1,2j)} = \alpha_1^{(k+1,2j)} u^*$ for $t = T+1$. Consequently, (76b) holds for $k+1$.

Next, we show (76c) holds for $k+1$. Recall (75), we observe that for all $k \geq 2$, and $j \in \{1, 2, \ldots, 2^{k-1}\}$,

$$h(k, 2j-1) = 2h(k-1, j) - 1, \quad h(k, 2j) = 2h(k-1, j). \tag{78}$$

To see this, let $h(k, 2j-1) = a$, then $2j - 2 = 2^{k-1}z + a - 1$, for some $z \in \mathbb{N}$, and $(a-1) \in \{0, 1, \ldots, 2^{k-1} - 1\}$. Note that $a-1$ is even, divide the preceding equation by 2, we have $j - 1 = 2^{k-2}z + \frac{a-1}{2}$. Therefore $(j-1) \mod (2^{k-2}) = \frac{a-1}{2}$, and $h(k-1, j) = \frac{a-1}{2} + 1 = \frac{a+1}{2}$. Therefore $h(k, 2j-1) = 2h(k-1, j) - 1$. Similarly, $h(k, 2j) = b$, then $2j - 1 = 2^{k-1}z + b - 1$, for some $z \in \mathbb{N}$, and $(b-1) \in \{0, 1, \ldots, 2^{k-1} - 1\}$. Note that $b-1$ is odd, and $j - 1 = 2^{k-2}z + \frac{b-2}{2}$, we have $(j-1) \mod (2^{k-2}) = \frac{b}{2} - 1$. Therefore $h(k-1, j) = \frac{b}{2}$, and $h(k, 2j) = 2h(k-1, j)$. We conclude that (78) holds.

Recall (77), for all $j \in \{1, \ldots, 2^k\}$, $u_t^{(k+1,2j-1)} = u_t^{(k+1,2j)} = u_t^{(k,j)}, T+2 \le t \le kT$. Since (76c) holds for k by assumption, we have $u_t^{(k,j)} = u_{t-T}^{(k-1,h(k,j))}, T+2 \le t \le kT$. Recall (77), $u_t^{(k-1,h(k,j))} = u_t^{(k,2h(k,j)-1)} = u_t^{(k,2h(k,j))}$, $t = 0, 1, \ldots, (k-1)T$. Therefore $u_{t-T}^{(k-1,h(k,j))} = u_{t-T}^{(k,2h(k,j))}$, $t = T, T+1, \ldots, kT$. Combine the preceding, we have

$$u_t^{(k+1,2j-1)} = u_{t-T}^{(k,2h(k,j)-1)},\ u_t^{(k+1,2j)} = u_{t-T}^{(k,2h(k,j))}, \quad T+2 \le t \le kT.$$

Recall (78), $2h(k,j) - 1 = h(k+1, 2j-1)$, $2h(k,j) = h(k+1, 2j)$, and we have

$$u_t^{(k+1,2j-1)} = u_{t-T}^{(k,h(k+1,2j-1))},\ u_t^{(k+1,2j)} = u_{t-T}^{(k,h(k+1,2j))}, \quad T+2 \le t \le kT.$$

Consequently, for all $j \in \{1, \ldots, 2^{k+1}\}$,

$$u_t^{(k+1,j)} = u_{t-T}^{(k,h(k+1,j))}, \quad T+2 \le t \le kT. \tag{79}$$

Recall (64), we see that for any $k \ge 2$, any $j = 1, 2, \ldots, 2^k$, $u_t^{(k,j)} = 0, (k-1)T+2 \le t \le kT$. Therefore for all $j \in \{1, \ldots, 2^{k+1}\}$,

$$u_t^{(k+1,j)} = 0 = u_{t-T}^{(k,h(k+1,j))}, \quad kT+2 \le t \le (k+1)T. \tag{80}$$

Recall (64), for all $k \ge 2$, all $j = 1, 2, \ldots, 2^k$, if $j \mod 4 \in \{1, 2\}$, then $u_t^{(k,j)} = 0$, $t = (k-1)T + 1$; if $j \mod 4 \in \{3, 0\}$, then $u_t^{(k,j)} = u^*$, $t = (k-1)T + 1$. Observe that

for any $k \geq 2$, $1 \leq j \leq 2^{k+1}$, let $h(k+1,j) = a$, then $((j-1) \mod 2^k) + 1 = a$, or $j = a + (2^{k-2}b) \cdot 4$, for some $b \in \mathbb{N}$. Therefore $j \mod 4 = h(k+1,j) \mod 4$. Therefore, $u_{kT+1}^{(k+1,j)} = u_{(k-1)T+1}^{(k,h(k+1,j))}$. Consequently,

$$u_t^{(k+1,j)} = u_{t-T}^{(k,h(k+1,j))}, \quad t = kT+1. \tag{81}$$

By (79), (80), (81), we see that (76c) holds for $k+1$. We conclude that (76) holds for all $k \geq 2$.

Now we are ready to show items ii) and iii) are satisfied for $k+1$, assuming that they are satisfied for some $k \geq 2$. Recall (64), we see that

$$u_t^{(k+1,2j-1)} = u_t^{(k+1,2j)}, 0 \leq t \leq (k+1)T, \tag{82}$$

for all $1 \leq j \leq 2^k$. Recall (72), $s_{(k+1,2j-1)} \leq (1+q+\cdots+q^k)s_o < \frac{1}{1-q}s_o$. Similarly, $s_{(k+1,2j)} < \frac{1}{1-q}s_o$. Recall (76a), $A > 1$, $C > 0$, for all $0 \leq t \leq T-1$, $Cx_t^{(k+1,2j-1)} = CA^t s_{(k+1,2j-1)} < CA^{T-1}\frac{1}{1-q}s_o = A^{-1}\frac{1}{1-A^{-T}}\beta$. Recall (69), $A^{-1}\frac{1}{1-A^{-T}} < 1$, and $\tilde{y}_t^{(k+1,2j-1)} = Q(Cx_t^{(k+1,2j-1)}) = Q(0)$ for all $0 \leq t \leq T-1$. Similarly, $\tilde{y}_t^{(k+1,2j)} = Q(0)$ for all $0 \leq t \leq T-1$. Therefore,

$$\tilde{y}_t^{(k+1,2j-1)} = \tilde{y}_t^{(k+1,2j)}, \quad t = 0,1,\ldots,T-1. \tag{83}$$

At $t = T$, $\tilde{y}_T^{(k+1,2j-1)} = Q(CA^T s_{(k+1,2j-1)})$, and $\tilde{y}_T^{(k+1,2j)} = Q(CA^T s_{(k+1,2j)})$. Recall (74), $\alpha_1^{(k+1,2j-1)} = \alpha_1^{(k+1,2j)}$. Recall (72), if $\alpha_1^{(k+1,2j-1)} = \alpha_1^{(k+1,2j)} = 0$, then $CA^T s_{(k+1,2j-1)} < CA^T(q+q^2+\cdots)s_o = CA^T\frac{q}{1-q}\frac{\beta}{CA^T} = \frac{1}{A^T-1}\beta$. Choose $T \in \mathbb{Z}_+$ such that

$$A^T - 1 > 1. \tag{84}$$

Then $CA^T s_{(k+1,2j-1)} < \beta$, and therefore $\tilde{y}_T^{(k+1,2j-1)} = Q(0)$. Similarly, we can show that $\tilde{y}_T^{(k+1,2j)} = Q(0)$ in this case. If $\alpha_1^{(k+1,2j-1)} = \alpha_1^{(k+1,2j)} = 1$, then $CA^T s_o \leq CA^T s_{(k+1,2j-1)} < CA^T(1+q+q^2+\cdots)s_o = \beta\frac{1}{1-A^{-T}}$.

Recall (70), we see that $CA^T s_{(k+1,2j-1)} \in [\beta, \beta+\delta_1)$, and therefore $\tilde{y}_T^{(k+1,2j-1)} = Q(CA^T s_{(k+1,2j-1)}) = Q(\beta)$. Similarly, we can show that $\tilde{y}_T^{(k+1,2j)} = Q(\beta)$ in this case. We summarize the preceding as

$$\tilde{y}_t^{(k+1,2j-1)} = \tilde{y}_t^{(k+1,2j)}, \quad t = T. \tag{85}$$

At $t = T+1$, recall (76b), $u_{T+1}^{(k+1,2j-1)} = \alpha_1^{(k+1,2j-1)} u^*, u_{T+1}^{(k+1,2j)} = \alpha_1^{(k+1,2j)} u^*$. Recall (74), $\alpha_1^{(k+1,2j-1)} = \alpha_1^{(k+1,2j)} = \alpha_1^{(k,j)}$, and $u_{T+1}^{(k+1,2j-1)} = u_{T+1}^{(k+1,2j)} \in \{0, u^*\}$. If $\alpha_1^{(k,j)} = 0$, and therefore $u_{T+1}^{(k+1,2j-1)} = 0$, recall (72), then $CA^{T+1} s_{(k+1,2j-1)} < CA^{T+1}(q+q^2+\cdots)s_o = \frac{A}{A^T-1}\beta$. Choose $T \in \mathbb{Z}_+$ such that

$$A^T - 1 > A. \tag{86}$$

Then $CA^{T+1} s_{(k+1,2j-1)} < \beta$, and therefore $\tilde{y}_{T+1}^{(k+1,2j-1)} = Q(0)$. Similarly, we can show that $\tilde{y}_{T+1}^{(k+1,2j)} = Q(0)$ in this case. If $\alpha_1^{(k,j)} = 1$, and therefore $u_{T+1}^{(k+1,2j-1)} = u^*$, then $CA^{T+1}s_o + Du^* \le CA^{T+1}s_{(k+1,2j-1)} + Du^* < CA^{T+1}(1+q+\cdots)s_o + Du^* = A\beta\frac{1}{1-A^{-T}} + Du^*$. Recall (71), we see that $CA^{T+1}s_{(k+1,2j-1)} + Du^* \in [A\beta + Du^*, A\beta + Du^* + \delta_2)$, and therefore $\tilde{y}_{T+1}^{(k+1,2j-1)} = Q(CA^{T+1}s_{(k+1,2j-1)} + Du^*) = Q(A\beta + Du^*)$. Similarly, $\tilde{y}_{T+1}^{(k+1,2j)} = Q(A\beta + Du^*)$. We conclude that

$$\tilde{y}_t^{(k+1,2j-1)} = \tilde{y}_t^{(k+1,2j)}, \quad t = T+1. \tag{87}$$

At $t = T+2$, for any $1 \le j \le 2^k$, recall (72), (76b), and $q = A^{-T}$

$$\begin{aligned} x_{T+2}^{(k+1,2j-1)} &= A^{T+2} s_{(k+1,2j-1)} + B u_{T+1}^{(k+1,2j-1)} \\ &= A^{T+2}(\alpha_1^{(k+1,2j-1)} + \alpha_2^{(k+1,2j-1)} q + \cdots + \alpha_{k+1}^{(k+1,2j-1)} q^k) s_o + B\alpha_1^{(k+1,2j-1)} u^* \\ &= A^2(\alpha_2^{(k+1,2j-1)} + \cdots + \alpha_{k+1}^{(k+1,2j-1)} q^{k-1}) s_o + \alpha_1^{(k+1,2j-1)}(A^{T+2} s_o + B u^*). \end{aligned}$$

Note that

$$A^{T+2}s_o + Bu^* = A^{T+2}\frac{\beta}{CA^T} + Bu^* = \frac{1}{C}(A^2\beta + CBu^*) = 0, \tag{88}$$

therefore

$$x_{T+2}^{(k+1,2j-1)} = A^2(\alpha_2^{(k+1,2j-1)} + \cdots + \alpha_{k+1}^{(k+1,2j-1)}q^{k-1})s_o. \tag{89}$$

Consider $x_2^{(k,h(k+1,2j-1))}$, recall (72), (76a),

$$x_2^{(k,h(k+1,2j-1))} = A^2(\alpha_1^{(k,h(k+1,2j-1))} + \cdots + \alpha_k^{(k,h(k+1,2j-1))}q^{k-1})s_o. \tag{90}$$

In the following, we show that $x_{T+2}^{(k+1,2j-1)} = x_2^{(k,h(k+1,2j-1))}$.

Recall (73), (75), note that $(j-1) \mod (2^{k-l+1}) = h(k-l+2,j)-1$, we see that

$$\alpha_l^{(k,j)} = \begin{cases} 0, & \text{if } 0 \le h(k-l+2,j)-1 < \frac{2^{k-l+1}}{2}, \\ 1, & \text{if } \frac{2^{k-l+1}}{2} \le h(k-l+2,j)-1 < 2^{k-l+1}, \end{cases} \tag{91}$$

for any $k \in \mathbb{Z}_+$, $j \in \{1,\ldots,2^k\}$, and $l \in \{1,\ldots,k\}$.

Recall (89), (78), for any $l \in \{2,3,\ldots,k+1\}$, (78),

$$h((k+1)-l+2,(2j-1)) = 2h(k-l+2,j)-1. \tag{92}$$

Similarly, recall (90), (78), for any $l \in \{2,3,\ldots,k+1\}$,

$$\begin{aligned} h(k-(l-1)+2,h(k+1,2j-1)) &= h(k-l+3,h(k+1,2j-1)) \\ &= h(k-l+3,2h(k,j)-1) \\ &= 2h(k-l+2,h(k,j))-1. \end{aligned} \tag{93}$$

We observe that for any $l \in \{2,3,\ldots,k+1\}$,

$$h(k-l+2,j) = h(k-l+2,h(k,j)). \tag{94}$$

Indeed, let $h(k,j) = a$, and $h(k-l+2,a) = b$. Then $j-1 = (a-1) + 2^{k-1}z$, and $a-1 = (b-1)+2^{k-l+1}z'$, for some $z, z' \in \mathbb{N}$. Then $j-1 = (b-1)+2^{k-l+1}(z'+2^{l-2}z)$. Since $l \geq 2$, $z'+2^{l-2}z \in \mathbb{N}$. Recall (75), $h(k-l+2,j) = b$. Therefore $h(k-l+2,j) = h(k-l+2, h(k,j))$. Recall (92), (93), $h((k+1)-l+2,(2j-1)) = h(k-(l-1)+2, h(k+1,2j-1))$. Recall (91), for any $l \in \{2,3,\ldots,k+1\}$,

$$\alpha_l^{(k+1,2j-1)} = \alpha_{l-1}^{(k,h(k+1,2j-1))}.$$

Recall (89), (90), for any $1 \leq j \leq 2^k$,

$$x_{T+2}^{(k+1,2j-1)} = x_2^{(k,h(k+1,2j-1))}.$$

Similarly, we can show that

$$x_{T+2}^{(k+1,2j)} = x_2^{(k,h(k+1,2j))}.$$

Therefore, for any $1 \leq j \leq 2^{k+1}$,

$$x_{T+2}^{(k+1,j)} = x_2^{(k,h(k+1,j))}. \tag{95}$$

Recall (76c), $u_t^{(k+1,j)} = u_{t-T}^{(k,h(k+1,j))}, t = T+2, T+3, \ldots, (k+1)T$, or equivalently,

$$u_{T+t}^{(k+1,j)} = u_t^{(k,h(k+1,j))}, t = 2,3,\ldots,kT. \tag{96}$$

By (95), (96), and the time-invariance of system (2), we see that for any $1 \leq j \leq 2^{k+1}$, $\tilde{y}_{t+T}^{(k+1,j)} = \tilde{y}_t^{(k,h(k+1,j))}, t = 2,3,\ldots,kT$. Recall (78), for any $1 \leq j \leq 2^k$,

$$\tilde{y}_{t+T}^{(k+1,2j-1)} = \tilde{y}_t^{(k,2h(k,j)-1)},\ \tilde{y}_{t+T}^{(k+1,2j)} = \tilde{y}_t^{(k,2h(k,j))}, \quad t = 2,3,\ldots,kT. \tag{97}$$

By assumption, item ii) is satisfied for k. Therefore, by (40), note that $1 \leq h(k,j) \leq$

2^{k-1}, we see that

$$\begin{aligned}
&\tilde{y}_t^{(k,2h(k,j)-1)} = \tilde{y}_t^{(k,2h(k,j))}, \quad t = 2, 3, \ldots, kT-1, \\
&\tilde{y}_t^{(k,2h(k,j)-1)} \neq \tilde{y}_t^{(k,2h(k,j))}, \quad t = kT.
\end{aligned}$$

Recall (97), we see that for any $1 \leq j \leq 2^k$,

$$\begin{aligned}
&\tilde{y}_t^{(k+1,2j-1)} = \tilde{y}_t^{(k+1,2j)}, \quad t = T+2, T+3, \ldots, (k+1)T-1, \\
&\tilde{y}_t^{(k+1,2j-1)} \neq \tilde{y}_t^{(k+1,2j)}, \quad t = (k+1)T.
\end{aligned} \tag{98}$$

Recall (62), (82), (83), (85), (87), (98), we see that item ii) is satisfied for $k+1$. For item iii), recall (64), we see that $u_t^{(k+1,2j-1)} = u_t^{(k,j)}, 0 \leq t \leq kT$, for any $1 \leq j \leq 2^k$. Recall (60), $s_{(k+1,2j-1)} = s_{(k,j)}$. Therefore $\tilde{y}_t^{(k+1,2j-1)} = \tilde{y}_t^{(k,j)}, 0 \leq t \leq kT$, for any $1 \leq j \leq 2^k$. Since item ii) is satisfied for $k+1$, $\tilde{y}_t^{(k+1,2j-1)} = \tilde{y}_t^{(k+1,2j)}, 0 \leq t \leq kT$, and therefore item iii) is satisfied for $k+1$.

By induction, we conclude that Ψ (67) satisfies items ii) and iii) in Theorem 6.

Next, we show that Ψ (67) satisfies item iv) in Theorem 6.

Assume a sequence $\{j(k)\}_{k=1}^{\infty}$ is given, and

$$j(k) \in \{1, \ldots, 2^k\}, \text{ and } j(k+1) \in \{2j(k)-1, 2j(k)\}, \quad \forall\, k \in \mathbb{Z}_+. \tag{99}$$

Recall (72), we observe that $\lim_{k\to\infty} s_{(k,j(k))}$ exists. To see this, for any $k \in \mathbb{Z}_+$, recall (74),

$$\begin{aligned}
&s_{(k+1,j(k+1))} - s_{(k,j(k))} \\
&= (\alpha_1^{(k+1,j(k+1))} + \alpha_2^{(k+1,j(k+1))} q + \cdots + \alpha_k^{(k+1,j(k+1))} q^{k-1} + \alpha_{k+1}^{(k+1,j(k+1))} q^k) s_o \\
&\quad - (\alpha_1^{(k,j(k))} + \alpha_2^{(k,j(k))} q + \cdots + \alpha_k^{(k,j(k))} q^{k-1}) s_o \\
&= (\alpha_1^{(k+1,j(k+1))} - \alpha_1^{(k,j(k))}) s_o + (\alpha_2^{(k+1,j(k+1))} - \alpha_2^{(k,j(k))}) q s_o + \cdots \\
&\quad + (\alpha_k^{(k+1,j(k+1))} - \alpha_k^{(k,j(k))}) q^{k-1} s_o + \alpha_{k+1}^{(k+1,j(k+1))} q^k s_o \\
&= \alpha_{k+1}^{(k+1,j(k+1))} q^k s_o.
\end{aligned}$$

Since $\alpha_{k+1}^{(k+1,j(k+1))} \geq 0$, $q > 0$, $s_o > 0$, we see that $s_{(k+1,j(k+1))} - s_{(k,j(k))} \geq 0$, and $\{s_{(k,j(k))}\}_{k=1}^{\infty}$ is a monotone sequence. Recall (72), we see that $s_{(k,j(k))} < \frac{1}{1-q} s_o < \infty$, therefore $\{s_{(k,j(k))}\}_{k=1}^{\infty}$ is bounded from above. Consequently, $\{s_{(k,j(k))}\}_{k=1}^{\infty}$ converges (pp. 6, [23]).

Given $\{j(k)\}_{k=1}^{\infty}$, define an initial state s of system (2) as

$$s = \lim_{k \to \infty} s_{(k,j(k))}. \tag{100}$$

Also, recall (61), define an input sequence $\mathbf{u}$ as

$$\begin{aligned} u_t &= u_t^{(1,j(1))}, \quad 0 \leq t \leq T, \\ u_t &= u_t^{(k,j(k))}, \quad (k-1)T < t \leq kT, \ \forall\, k \geq 2. \end{aligned} \tag{101}$$

In the following, let $\tilde{\mathbf{y}} = \{\tilde{y}_t\}_{t=0}^{\infty}$ be the output of system (2) when its initial state is $x_0 = s$ (100), and its input is $\mathbf{u}$ (101). And we also use x_t to denote the state of (2) corresponding with $x_0 = s$ (100) and $\mathbf{u}$ (101).

For $\tilde{\mathbf{y}}$ defined in the preceding, we observe that for any $k \in \mathbb{Z}_+$,

$$\tilde{y}_t = \tilde{y}_t^{(k,j(k))}, \quad 0 \leq t \leq kT. \tag{102}$$

We use induction to show this observation.

At $k = 1$, by the previous derivation showing Ψ (67) satisfies item ii) for $k = 1$, we have $\tilde{y}_t^{(1,1)} = Q(0)$ for $t = 0, 1, \ldots, T$, and $\tilde{y}_t^{(1,2)} = Q(0)$ for $t = 0, 1, \ldots, T-1$, $\tilde{y}_T^{(1,2)} = Q(\beta)$. If $j(1) = 1$, then $\alpha_1^{(1,j(1))} = 0$, recall (72), we see that $s_{(k,j(k))} \leq \frac{q}{1-q} s_o$ for all $k \in \mathbb{Z}_+$. Recall (100), $s \leq \frac{q}{1-q} s_o$. Then for all $0 \leq t \leq T$, $u_t = 0$ by (101), therefore $Cx_t = CA^t s \leq CA^T \frac{q}{1-q} s_o = \frac{1}{A^T - 1}\beta$. Recall (84), $Cx_t < \beta$, and $\tilde{y}_t = Q(0) = \tilde{y}_t^{(1,j(1))}$ for $0 \leq t \leq T$. Similarly, if $j(1) = 2$, then $\alpha_1^{(1,j(1))} = 1$, recall (72), we see that $s_{(k,j(k))} \leq \frac{1}{1-q} s_o$ for all $k \in \mathbb{Z}_+$, $s \leq \frac{1}{1-q} s_o$. Then for all $0 \leq t \leq T-1$, $u_t = 0$ by (101), therefore $Cx_t = CA^t s \leq CA^{T-1} \frac{1}{1-q} s_o = A^{-1} \frac{1}{1 - A^{-T}} \beta$. Recall (69), $A^{-1} \frac{1}{1-A^{-T}} < 1$, and $Cx_t < \beta$,

and $\tilde{y}_t = Q(0) = \tilde{y}_t^{(1,j(1))}$ for $0 \le t \le T-1$. At $t = T$, since $\alpha_1^{(1,j(1))} = 1$, $s - s_o \ge 0$, write $Cx_T = CA^T(s - s_o + s_o)$, then $\beta \le Cx_T = \beta + CA^T(s - s_o) \le \beta + CA^T \frac{q}{1-q} s_o = \beta \frac{1}{1-A^{-T}}$. Recall (70), we see that $\tilde{y}_T = Q(\beta) = \tilde{y}_T^{(1,j(1))}$. We conclude that (102) holds for $k = 1$.

Assume (102) holds for some $k \ge 1$, since Ψ (67) satisfies items ii) and iii) in Theorem 6, recall (40), (41b), we see that $\tilde{y}_t^{(k+1,2j(k)-1)} = \tilde{y}_t^{(k+1,2j(k))} = \tilde{y}_t^{(k,j(k))}, t = 0, 1, \ldots, kT$. Recall $j(k+1) \in \{2j(k) - 1, 2j(k)\}$, $\tilde{y}_t^{(k+1,j(k+1))} = \tilde{y}_t^{(k,j(k))}, t = 0, 1, \ldots, kT$. By assumption, (102) holds for k, therefore

$$\tilde{y}_t = \tilde{y}_t^{(k,j(k))} = \tilde{y}_t^{(k+1,j(k+1))}, t = 0, 1, \ldots, kT.$$

Consequently, to show (102) holds for some $k+1$, we only need to consider $kT + 1 \le t \le (k+1)T$.

Recall (76), for any $l \in \{2, \ldots, k-1\}$

$$\begin{aligned} u_{lT+1}^{(k,j)} &= u_{(l-1)T+1}^{(k-1,h(k,j))} \\ &= u_{(l-2)T+1}^{(k-2,h(k-1,h(k,j)))} \\ &= \cdots \\ &= u_{T+1}^{(k-(l-1),h(k-(l-2),h(\ldots,h(k-1,h(k,j)),\ldots)))}. \end{aligned}$$

Recall (94), $h(k-(l-2), h(\ldots, h(k-1, h(k,j)), \ldots)) = h(k-(l-2), h(\ldots, h(k-1, j), \ldots)) = h(k-(l-2), h(\ldots, h(k-2, j), \ldots)) = \cdots = h(k-(l-2), j)$. Recall (76b), (64), and note that $h(k-(1-2), j) = j$ when $1 \le j \le 2^k$, we see that for all $k \ge 2$, $j \in \{1, 2, \ldots, 2^k\}$, $\mathcal{I}$ (61) satisfies:

$$u_t^{(k,j)} = \begin{cases} \alpha_1^{(k-(l-1),h(k-(l-2),j))} u^*, & t = lT + 1 \text{ for some } l \in 1, \ldots, k-1, \\ 0, & \text{otherwise.} \end{cases} \tag{103}$$

Next, consider $x_{kT}^{(k+1,j(k+1))}$: By (72), (103)

$$\begin{aligned}
x_{kT}^{(k+1,j(k+1))} &= A^{kT} s_{(k+1,j(k+1))} + \sum_{\tau=0}^{kT-1} A^{kT-1-\tau} B u_\tau^{(k+1,j(k+1))} \\
&= q^{-k}(\sum_{l=1}^{k-1} \alpha_l^{(k+1,j(k+1))} q^{l-1} + \alpha_k^{(k+1,j(k+1))} q^{k-1} + \alpha_{k+1}^{(k+1,j(k+1))} q^k) s_o + \\
&\quad \sum_{l=1}^{k-1} A^{(k-l)T-2} B \alpha_1^{((k+1)-(l-1),h((k+1)-(l-2),j(k+1)))} u^* \\
&= \sum_{l=1}^{k-1} q^{-k+l} (\alpha_l^{(k+1,j(k+1))} A^T s_o + \alpha_1^{(k-l+2,h(k-l+3,j(k+1)))} A^{-2} B u^*) + \\
&\quad A^T \alpha_k^{(k+1,j(k+1))} s_o + \alpha_{k+1}^{(k+1,j(k+1))} s_o.
\end{aligned}$$

Note that by (73), $\alpha_l^{(k+1,j(k+1))} = \alpha_1^{(k-l+2,j(k+1))}$. Recall (91), $\alpha_1^{(k-l+2,j(k+1))}$ is determined by $h(k-l+2-1+2, j(k+1)) = h(k-l+3, j(k+1))$. Similarly, $\alpha_1^{(k-l+2,h(k-l+3,j(k+1)))}$ is determined by $h(k-l+3, h(k-l+3, j(k+1)))$. Recall (75), we see that $h(k-l+3, h(k-l+3, j(k+1))) = h(k-l+3, j(k+1))$, therefore

$$\alpha_l^{(k+1,j(k+1))} = \alpha_1^{(k-l+2,h(k-l+3,j(k+1)))}. \tag{104}$$

Consequently,

$$\begin{aligned}
x_{kT}^{(k+1,j(k+1))} = \sum_{l=1}^{k-1} q^{-k+l} \alpha_l^{(k+1,j(k+1))} &(A^T s_o + A^{-2} B u^*) \\
&+ A^T \alpha_k^{(k+1,j(k+1))} s_o + \alpha_{k+1}^{(k+1,j(k+1))} s_o.
\end{aligned}$$

Note that $A^T s_o + A^{-2} B u^* = \frac{1}{C} A^{-2} (A^2 \beta + C B u^*) = 0$. Therefore,

$$x_{kT}^{(k+1,j(k+1))} = A^T \alpha_k^{(k+1,j(k+1))} s_o + \alpha_{k+1}^{(k+1,j(k+1))} s_o. \tag{105}$$

Next, consider x_t at $t = kT$. Recall x_t is the system state corresponding with $x_0 = s$ (100) and $\mathbf{u}$ (101). Recall (40b), (41b), (99), (101), we see that $u_t = u_t^{(z,j(z))}, 0 \le t \le zT$,

for any $z \in \mathbb{Z}_+$. Consequently,

$$\begin{aligned} x_{kT} &= A^{kT}(s - s_{(k+1,j(k+1))} + s_{(k+1,j(k+1))}) + \sum_{\tau=0}^{kT-1} A^{kT-1-\tau} B u_\tau \\ &= A^{kT}(s - s_{(k+1,j(k+1))}) + (A^{kT} s_{(k+1,j(k+1))} + \sum_{\tau=0}^{kT-1} A^{kT-1-\tau} B u_\tau^{(k+1,j(k+1))}) \\ &= A^{kT}(s - s_{(k+1,j(k+1))}) + x_{kT}^{(k+1,j(k+1))}. \end{aligned}$$

Recall (72), (100), we see that $0 \leq s - s_{(k+1,j(k+1))} = \sum_{i=1}^{\infty} \alpha_{k+1+i}^{(k+1+i,j(k+1+i))} q^{k+i} s_o$. Therefore $0 \leq A^{kT}(s - s_{(k+1,j(k+1))}) \leq \frac{q}{1-q} s_o$, and

$$x_{kT}^{(k+1,j(k+1))} \leq x_{kT} \leq x_{kT}^{(k+1,j(k+1))} + \frac{q}{1-q} s_o. \tag{106}$$

At $t = kT + 1$, recall (103), (104), $u_{kT+1} = u_{kT+1}^{(k+1,j(k+1))} = \alpha_k^{(k+1,j(k+1))} u^*$. Recall (105), then $Cx_{kT+1}^{(k+1,j(k+1))} + Du_{kT+1}^{(k+1,j(k+1))} = CA(A^T \alpha_k^{(k+1,j(k+1))} s_o + \alpha_{k+1}^{(k+1,j(k+1))} s_o) + D\alpha_k^{(k+1,j(k+1))} u^* = \alpha_k^{(k+1,j(k+1))}(CA^{T+1} s_o + Du^*) + CA\alpha_{k+1}^{(k+1,j(k+1))} s_o$. If $\alpha_k^{(k+1,j(k+1))} = 0$, then $Cx_{kT+1}^{(k+1,j(k+1))} + Du_{kT+1}^{(k+1,j(k+1))} \leq CAs_o < CA^T s_o = \beta$, and $\tilde{y}_{kT+1}^{(k+1,j(k+1))} = Q(0)$. Recall (106), then $Cx_{kT+1} + Du_{kT+1} \leq CAs_o + CA\frac{q}{1-q} s_o = CA\frac{1}{1-q} s_o = A\frac{1}{A^T-1}\beta$. Recall (86), $A\frac{1}{A^T-1}\beta < \beta$, and therefore $\tilde{y}_{kT+1} = Q(0) = \tilde{y}_{kT+1}^{(k+1,j(k+1))}$. If $\alpha_k^{(k+1,j(k+1))} = 1$, then by the preceding, $Cx_{kT+1}^{(k+1,j(k+1))} + Du_{kT+1}^{(k+1,j(k+1))} \in [A\beta + Du^*, A\beta + Du^* + A\frac{1}{A^T-1}\beta)$, and $Cx_{kT+1} + Du_{kT+1} \in [A\beta + Du^*, A\beta + Du^* + A\frac{1}{A^T-1}\beta]$. Recall (71), $A\frac{1}{A^T-1}\beta < \delta_2$, and therefore $\tilde{y}_{kT+1} = \tilde{y}_{kT+1}^{(k+1,j(k+1))} = Q(A\beta + Du^*)$. We conclude that $\tilde{y}_t = \tilde{y}_t^{(k+1,j(k+1))}$, $t = kT + 1$.

For $kT + 2 \leq t \leq (k+1)T - 1$, by (103), (104), (105), we have

$$\begin{aligned} x_{kT+2}^{(k+1,j(k+1))} &= A^2(A^T \alpha_k^{(k+1,j(k+1))} s_o + \alpha_{k+1}^{(k+1,j(k+1))} s_o) + B\alpha_k^{(k+1,j(k+1))} u^* \\ &= A^2 \alpha_{k+1}^{(k+1,j(k+1))} s_o + \alpha_k^{(k+1,j(k+1))}(A^{T+2} s_o + Bu^*) = A^2 \alpha_{k+1}^{(k+1,j(k+1))} s_o. \end{aligned}$$

Recall (103), $u_t^{(k+1,j(k+1))} = 0$, $kT + 2 \leq t \leq (k+1)T - 1$. Therefore $Cx_t^{(k+1,j(k+1))} +$

$Du_t^{(k+1,j(k+1))} = CA^{t-(kT+2)}x_{kT+2}^{(k+1,j(k+1))} \leq CA^{T-3}A^2\alpha_{k+1}^{(k+1,j(k+1))}s_o = A^{-1}\beta < \beta$. Therefore $\tilde{y}_t^{(k+1,j(k+1))} = Q(0), kT+2 \leq t \leq (k+1)T-1$. Recall (101), (106), we see that $Cx_t + Du_t \leq A^{-1}\beta + CA^{T-1}\frac{q}{1-q}s_o = A^{-1}\beta\frac{1}{1-q}$. Recall (69), $A^{-1}\beta\frac{1}{1-q} < \beta$, therefore $\tilde{y}_t = Q(0) = \tilde{y}_t^{(k+1,j(k+1))} = Q(0), kT+2 \leq t \leq (k+1)T-1$.

At $t = (k+1)T$, recall (103), $u_{(k+1)T}^{(k+1,j(k+1))} = u_{(k+1)T} = 0$. Recall $x_{kT+2}^{(k+1,j(k+1))} = A^2\alpha_{k+1}^{(k+1,j(k+1))}s_o$, we see that $x_{kT+T}^{(k+1,j(k+1))} = A^T\alpha_{k+1}^{(k+1,j(k+1))}s_o$. Recall (106), and note that $x_{(k+1)T} = A^T(x_{kT} - x_{kT}^{(k+1,j(k+1))}) + x_{kT+T}^{(k+1,j(k+1))}$, we have $A^T\alpha_{k+1}^{(k+1,j(k+1))}s_o \leq x_{(k+1)T} \leq A^T\alpha_{k+1}^{(k+1,j(k+1))}s_o + \frac{q}{1-q}s_o$. If $\alpha_{k+1}^{(k+1,j(k+1))} = 0$, then $\tilde{y}_{(k+1)T}^{(k+1,j(k+1))} = Q(Cx_{kT+T}^{(k+1,j(k+1))}) = Q(0)$. And $Cx_{(k+1)T} \leq \frac{1}{A^T-1}\beta$. Recall (84), $\frac{1}{A^T-1}\beta < \beta$, and therefore $\tilde{y}_{(k+1)T} = Q(0) = \tilde{y}_{(k+1)T}^{(k+1,j(k+1))}$. If $\alpha_{k+1}^{(k+1,j(k+1))} = 1$, then $\tilde{y}_{(k+1)T}^{(k+1,j(k+1))} = Q(Cx_{kT+T}^{(k+1,j(k+1))}) = Q(\beta)$. And $\beta \leq Cx_{(k+1)T} \leq \frac{1}{1-q}\beta$, by (70), $\tilde{y}_{(k+1)T} = Q(\beta) = \tilde{y}_{(k+1)T}^{(k+1,j(k+1))}$. We conclude that $\tilde{y}_t = \tilde{y}_t^{(k+1,j(k+1))}, t = kT+T$.

So far, we have shown that (102) holds for $k+1$. By induction, we conclude that (102) holds for all $k \in \mathbb{Z}_+$.

Next, we show that the pair $(\mathbf{u}, \tilde{\mathbf{y}})$, which corresponds with the initial state s (100) and the input (101), satisfies (42). Recall (62), $T_{(k,j(k))} = k \cdot T$, for all $k \in \mathbb{Z}_+$. For $k = 1$, by (101), $u_t = u_t^{(1,j(1))}, 0 \leq t \leq T$. And by (102), $\tilde{y}_t = \tilde{y}_t^{(1,j(1))}, 0 \leq t \leq T$. For any $k \geq 2$, by (101), $u_t = u_t^{(k,j(k))}, (k-1)T < t \leq kT$. By (102), $\tilde{y}_t = \tilde{y}_t^{(k,j(k))}, 0 \leq t \leq kT$, and consequently $\tilde{y}_t = \tilde{y}_t^{(k,j(k))}, (k-1)T < t \leq kT$. Therefore $(\mathbf{u}, \tilde{\mathbf{y}})$ satisfies (42). Note that $(\mathbf{u}, \tilde{\mathbf{y}}) \in P$, where P is system (2), we conclude that Ψ (67) satisfies item iv) in Theorem 6.

Since Ψ (67) satisfies the hypotheses in Theorem 6, we see that system (2) is not weakly observable (C2). This completes the first part of this derivation of Theorem 7.

Next, we show that system (2) is not asymptotically observable (C3). Since system (2) is not weakly observable (C2), we apply the derivation of Theorem 6, particularly (56) in the following. Then for any observer $\hat{S}$, there is $(\mathbf{u}, \tilde{\mathbf{y}})$, which corresponds with the initial

state s (100) and the input (101), such that for all $k \in \mathbb{Z}_+$,

$$\hat{y}_t \neq \tilde{y}_t, \quad t = kT.$$

Let $\delta = \min\{\|y_1 - y_2\| : y_1 \neq y_2, y_1, y_2 \in \mathcal{Y}\}$, and define $\gamma = \frac{\delta}{2\|u^*\|} > 0$. For any $N \in \mathbb{Z}_+$, and $N \geq 2$,

$$\sum_{t=T+1}^{NT} \|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\| = \sum_{k=2}^{N}(\sum_{t=(k-1)T+1}^{kT} (\|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\|)).$$

For any $k \in \{2, \ldots, N\}$, recall (101), (103), $\sum_{t=(k-1)T+1}^{kT} \|u_t\| = \|u^*\|$. Since $\hat{y}_t \neq \tilde{y}_t, t = kT$, $(\sum_{t=(k-1)T+1}^{kT} \|\tilde{y}_t - \hat{y}_t\|) \geq \|\tilde{y}_{kT} - \hat{y}_{kT}\| \geq \delta$. Therefore

$$\sum_{t=(k-1)T+1}^{kT} (\|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\|) \geq \delta - \frac{\delta}{2\|u^*\|}\|u^*\| = \frac{\delta}{2}.$$

Consequently,

$$\sum_{t=T+1}^{NT} \|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\| \geq (N-1)\frac{\delta}{2}, \quad \forall\, N \geq 2.$$

Therefore, $\sup_{N\geq 2} \sum_{t=T+1}^{NT} \|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\| = \infty$. Since $\mathcal{U}, \mathcal{Y}$ are finite sets, and each of their elements are of finite norm, $\sum_{t=0}^{T} \|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\|$ is finite, and consequently $\sup_{T\geq 0} \sum_{t=0}^{T} \|\tilde{y}_t - \hat{y}_t\| - \gamma\|u_t\| = \infty$. By definition, $\gamma = \frac{\delta}{2\|u^*\|} > 0$ is not an observation gain of system (2). Recall Definition 2, we see that for any $\gamma' \in \mathbb{R}_{\geq 0}$, if $\gamma' < \gamma$, then γ' is not an observation gain. Recall (7), we see that the $\mathcal{O}$-gain γ^* of system (2) satisfies $\gamma^* \geq \gamma > 0$. Recall Definition 3, system (2) is not asymptotically observable (C3). □

Remark. Alternatively, in the hypotheses of Theorem 7, we can require (88) in terms of the state-space of system (2).

3.6 Illustrative Examples

In this section, we first present two examples to demonstrate the concept of finite memory observability. The first example corresponds to Theorem 1, and the second example corresponds to Theorem 2. We also use a third example to illustrate the conditions of weakly observable (C2) and asymptotically observable (C3).

Example 2. We present a second order system (2) which is finite memory observable. The parameters of the LTI system in (2) are:

$$A = \begin{bmatrix} 0.25 & -0.05 \\ 0 & 0.2 \end{bmatrix} B = \begin{bmatrix} 2 \\ 1 \end{bmatrix} C = \begin{bmatrix} 0.5 & 0 \end{bmatrix} D = 1$$

The above parameters are *minimal*. $\mathcal{U} = \{0, 1, -1\}$. The quantizer $Q(\cdot)$ is defined in (3), and $R = 5$. Next we will show this system satisfies the condition in theorem 1.

We assume that the initial state $x(0)$ of the LTI system is bounded, particularly:

$$\|x_0\|_\infty < b$$

for some $b \in \mathbb{R}_+$.

First we find the distance $d(\mathcal{A}, \mathcal{B})$ between the two sets $\mathcal{A}$ and $\mathcal{B}$ defined in (8) and (9).

Since

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1/4 & 0 \\ 0 & 1/5 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

We have

$$A^n = \begin{bmatrix} (1/4)^n & (1/5)^n - (1/4)^n \\ 0 & (1/5)^n \end{bmatrix}, n = 0, 1, 2, 3 \cdots$$

and

$$CA^nB = 1/2((1/5)^n + (1/4)^n), n = 0, 1, 2, 3 \cdots$$

Then $\mathcal{A} \subset (-55/24, -41/24) \cup (-31/24, -17/24) \cup (-7/24, 7/24) \cup (17/24, 31/24) \cup$

$(41/24, 55/24)$. And the set $\mathcal{B}$ is $\{-4.5, -3.5, -2.5 \cdots 3.5, 4.5\}$, we arrive at the result:

$$d(\mathcal{A}, \mathcal{B}) = 5/24$$

This means that the forced response of the underlying LTI system is at least $5/24$ away from any discontinuous point of the quantizer.

Next, we try to find a uniform bound on the state x_t. Refer to the proof of theorem 1, we arrive at:

$$x_t \leq \max\{\|x_0\|_\infty, \|Bu_t\|_\infty\} \sum_{\tau=0}^{t} \|A\|_\infty^\tau$$

Since $\|x_0\|_\infty \leq b$ and $u_t \in \{0, 1, -1\}$, we have $\max\{\|x_0\|_\infty, \|Bu_0\|_\infty\} \leq \max\{b, 2\}$. And $\|A\|_\infty$ is the greatest row sum of the matrix A, we have $\|A\|_\infty = 0.3$. Combined with the above upper bound of $\|x_t\|_\infty$, we have:

$$\|x_t\|_\infty \leq \frac{10}{7} \max\{b, 2\}, \ \forall t \in \mathbb{N}$$

Next we find a memory length l for the finite memory observer. Let $x_t = [x_t^1 \ x_t^2]^T$, and choose l such that:

$$|\frac{1}{2}(1/4)^l x_{t-l}^1 + \frac{1}{2}[(1/5)^l - (1/4)^l] x_{t-l}^2| \leq \frac{5}{24}$$

Notice that $\|x_t\|_\infty \leq \frac{10}{7} \max\{b, 2\}$, it is easy to come up with a choice of l, say: $l = \lceil \log_4 \frac{72}{7} \max\{b, 2\} \rceil + 1$. In this case, all possible values of y_t lies within the continuous part of the quantizer $Q(\cdot)$. Then the DFM observer which stores past l steps of input and output of system (2) achieves (C1).

Example 3. We present another second order system (2) that is also observable. The parameters of the LTI system in (2) are:

$$A = \begin{bmatrix} 2 & 2 \\ 0 & 0 \end{bmatrix} B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} C = \begin{bmatrix} 0 & 1 \end{bmatrix} D = 1$$

$\mathcal{U} = \{0, 1, -1\}$. The quantizer $Q(\cdot)$ is defined in (3), and $R = 5$.

Clearly $CA = \mathbf{0}$, so this system satisfies the condition in Theorem 2. Notice that the solution of y_t is: $y_t = u_{t-1} + u_t, \forall\, t \geq 2$. Then it is easy to come up with a finite memory observer that achieves the requirement in (C1), described as follows:

$$q_{t+1} = u_t$$
$$\hat{y}_t = Q(q_t + u_t)$$

where $t \in \mathbb{N}$, $q_t \in \mathcal{Q}$ and $\mathcal{Q} = \mathcal{U}$. q_0 can be arbitrary, say 0.

Remark. Both examples 2 and 3 are finite memory observable. If we consider the observability of the underlying LTI system, then example 2 is observable but example 3 is not observable. So for system (2), there is no direct link between finite memory observability and observability of the underlying LTI system.

Example 4. We present a one-dimensional system (2) which is not weakly observable (C3). The parameters of the LTI system in (2) are: $A = 2, B = 1, C = 1, D = 0$. The quantizer Q is described by: $Q((-\infty, 0.5)) = 0, Q([0.5, \infty)) = 1$. The input set is $\mathcal{U} = \{0, 2, -2\}$. Let $u^* = -2$, and note that the discontinuous point of Q is $\beta = 0.5$, then $CBu^* + A^2\beta = -2 + 2^2 0.5 = 0$, therefore the hypotheses in Theorem 7 is satisfied, and consequently this system is not (C3).

3.7 Summary

In this section we propose a set of notions of observability for systems over finite alphabets with quantized output based on how well an observer can predict the output of such systems. We characterized this notion by deriving both necessary and sufficient conditions for observability.

4 DFM Observers and Their Construction

Following our study on observability of systems over finite alphabets, in this chapter, we discuss the construction of DFM observers. In particular, we study the limitations of an existing construction, and we propose a new construction that has better performance in certain cases. Some of the work presented in this chapter consists of the previous work reported in [25], as well as some new observations.

In the setup shown in Figure 2, the observer $\hat{S}$ generate $\hat{y}_t$ to approximate the output y_t of system P, based on the input and output history of P. Therefore, $\hat{S}$ can also be viewed as an *approximation model* for P. And the observation gain γ (see Definition 2) serves as a measure of the quality of such approximations. In the following, we are particularly interested in constructing DFM observers as approximation models for systems (2). In this case, the approximation model $\hat{S}$ has finite cardinality, while the original system P has infinite cardinality. In this chapter, we show that in certain cases we do not lose any generality in requiring the observer $\hat{S}$ be a DFM.

From the derivation of Theorem 1, in terms of constructing DFM observers, it seems a reasonable idea to associate finite length of input and output sequences of system (2) to the states of DFM observers. However, this approach has its limitations, especially when (C2) or (C3) is concerned. As we shall see next, for certain systems, if the DFM observer is constructed using the “associating state" approach, the observation gain γ is greater than zero, but there exist other DFM observers that achieve $\gamma = 0$.

4.1 Connections between Finite Memory Observable and DFM Observers

Note that the definition of (C1) in the previous section is different from that defined in [25] (namely whether the set $\mathcal{Q}$ is finite). Recall that if $|\mathcal{Q}| < \infty$ for an observer $\hat{S}$ (5), then $\hat{S}$ is indeed a deterministic finite state machine. Interestingly, as we shall see next, the two definitions of (C1) are equivalent in certain cases.

First, we introduce some notations. Given a pair of signals $(\mathbf{u}, \mathbf{y}) = (\{u_t\}_{t=0}^{\infty}, \{y_t\}_{t=0}^{\infty})$, and an integer $\tau \in \mathbb{N}$, use $(\mathbf{u}^\tau, \mathbf{y}^\tau)$ to denote the shifted pair of signals: $(\mathbf{u}^\tau, \mathbf{y}^\tau) = (\{u_{t+\tau}\}_{t=0}^{\infty}, \{y_{t+\tau}\}_{t=0}^{\infty})$.

Definition 4. Given a system over finite alphabet $P \subset \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$ (1), we say P is *time-invariant* if for any $(\mathbf{u}, \mathbf{y}) \in P$, and any $\tau \in \mathbb{N}$, $(\mathbf{u}^\tau, \mathbf{y}^\tau) \in P$.

Now, we are ready to present the equivalence between the two definitions of (C1).

Lemma 2. Given any time-invariant system P (1), P is finite memory observable if and only if there is a DFM observer $\hat{S}$ (5) with $|\mathcal{Q}| < \infty$, and $T \in \mathbb{Z}_+$ such that for any $(\mathbf{u}, \mathbf{y}) \in P$, $\hat{y}_t = y_t$ for all $t \geq T$.

Remark. Clearly, system (2) is time-invariant according to Definition 4. Consequently, as stated in Lemma 2, we only need to consider DFM observers for systems (2) as long as (C1) is concerned. And this is the reason why we adopt the name “finite memory observable" for (C1), since a DFM has a finite amount of memory.

Proof. Assume a time-invariant P is (C1) as stated in Definition 3. Then there is an observer $\hat{S}$ (5) and $T \in \mathbb{Z}_+$ such that for any $(\mathbf{u}, \mathbf{y}) \in P$, $\hat{y}_t = y_t$ for all $t \geq T$. Next, define a truncation operator $\psi(\cdot) : \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}} \to \mathcal{U}^{T+1} \times \mathcal{Y}^{T}$ as: For any $(\mathbf{u}, \mathbf{y}) \in \mathcal{U}^{\mathbb{N}} \times \mathcal{Y}^{\mathbb{N}}$, where $u = \{u_t\}_{t=0}^{\infty}$ and $y = \{y_t\}_{t=0}^{\infty}$, $\psi(\mathbf{u}, \mathbf{y})$ is defined as

$$\psi(\mathbf{u}, \mathbf{y}) = (u_T, u_{T-1}, \ldots, u_0, y_{T-1}, \ldots, y_0). \tag{107}$$

Given P, define a set $\mathcal{Q}_F$ as

$$\mathcal{Q}_F = \{\psi(\mathbf{u}, \mathbf{y}) : (\mathbf{u}, \mathbf{y}) \in P\}. \tag{108}$$

Essentially, $\mathcal{Q}_F$ is the collection of all feasible input and output segments of P with appropriate lengths. Clearly, $\mathcal{Q}_F \subset \mathcal{U}^{T+1} \times \mathcal{Y}^{T}$.

Given $\hat{S}$, recall that $\hat{S}$ is of the form (5) with initial state q_o. Define a function $\theta(\cdot)$:

$\mathcal{Q}_F \to \mathcal{Y}$ as: For any $q \in \mathcal{Q}_F$, write $q = (u_T, u_{T-1}, \ldots, u_0, y_{T-1}, \ldots, y_0)$, then

$$\theta(q) = g(f(\ldots f(f(q_o, u_0, y_0), u_1, y_1) \ldots, u_{T-1}, y_{T-1}), u_T). \tag{109}$$

Essentially, $\theta(q)$ is the output $\hat{y}_T$ of $\hat{S}$ at time T, when the input signals of $\hat{S}$ for $0 \leq t \leq T$ are in accordance with q. For the completeness of the construction of DFM (as we shall see next), define: If $q \notin \mathcal{Q}_F$, then let

$$\theta(q) = y_\varnothing, \text{ for some } y_\varnothing \in \mathcal{Y}. \tag{110}$$

Next, define a transition function $\phi(\cdot)$. Let q'_o be a symbolic state, then $\phi(q, u, y)$ is a mapping: $(q'_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i \times \mathcal{Y}^i)) \times \mathcal{U} \times \mathcal{Y} \to \bigcup_{i=1}^{T} \mathcal{U}^i \times \mathcal{Y}^i$, described by:

For any $q \in q'_o \cup (\bigcup_{i=1}^{T} \mathcal{U}^i \times \mathcal{Y}^i), y \in \mathcal{Y}, u \in \mathcal{U}$,

$\triangleright$ If $q = q'_o$, then

$$\phi(q, u, y) = (u, y). \tag{111}$$

$\triangleright$ If $q \in \bigcup_{i=1}^{T-1} \mathcal{U}^i \times \mathcal{Y}^i$, write $q = (u_1, u_2 \ldots u_i, y_1, y_2 \ldots y_i)$ for some $i \in \{1, \ldots, T-1\}$, then

$$\phi(q, u, y) = (u, u_1, u_2 \ldots u_i, y, y_1, y_2 \ldots y_i). \tag{112}$$

$\triangleright$ If $q \in \mathcal{U}^T \times \mathcal{Y}^T$, write $q = (u_1, u_2 \ldots u_T, y_1, y_2 \ldots y_T)$, then

$$\phi(q, u, y) = (u, u_1, u_2 \ldots u_{T-1}, y, y_1, y_2 \ldots y_{T-1}). \tag{113}$$

Now, consider a system $\hat{S}'$ described by:

$$\begin{aligned} q'_{t+1} &= \phi(q'_t, u_t, y_t), \\ \hat{y}'_t &= \theta(u_t, q'_t), \end{aligned} \tag{114}$$

where functions θ and ϕ are defined in equations (109) through (113). We enforce that $\hat{S}'$ starts at a fixed initial state: $q'_0 = q'_o$.

Use $\hat{S}'$ as an observer for P as in shown in Figure 2, then u_t and y_t in (114) corresponds to the input and output of P respectively. Then, for any $(\mathbf{u}, \mathbf{y}) \in P$, recall (111), we have $q'_1 = (u_0, y_0)$. Recall (112), we have $q'_2 = (u_1, u_0, y_1, y_0)$, and consequently $q'_3 = (u_2, u_1, u_0, y_2, y_1, y_0)$. Repeat this argument, we have $q'_T = (u_{T-1}, \ldots, u_0, y_{T-1}, \ldots, y_0)$. At $t = T$, recall (113), we have $q'_{T+1} = (u_T, \ldots, u_1, y_T, \ldots, y_1)$. Consequently, we have

$$q'_t = (u_{t-1}, \ldots, u_{t-T}, y_{t-1}, \ldots, y_{t-T}), \forall\, t \geq T. \tag{115}$$

Essentially, q'_t contains the past T steps of the input and output of P. Recall (114), we have

$$\hat{y}'_t = \theta(u_t, u_{t-1}, \ldots, u_{t-T}, y_{t-1}, \ldots, y_{t-T}),\ \forall\, t \geq T. \tag{116}$$

For any $t \geq T$, let $\tau = t - T$. As stated previously, use $(\mathbf{u}^\tau, \mathbf{y}^\tau)$ to denote the shifted pair of signals $(\mathbf{u}, \mathbf{y})$: $u_t^\tau = u_{t+\tau}$, and $y_t^\tau = y_{t+\tau}$. Since P is time-invariant, we have $(\mathbf{u}^\tau, \mathbf{y}^\tau) \in P$. Recall (108), we have $\psi(u^\tau, y^\tau) \in \mathcal{Q}_F$. Since

$$\begin{aligned}
\psi(\mathbf{u}^\tau, \mathbf{y}^\tau) &= (u_T^\tau, u_{T-1}^\tau, \ldots, u_0^\tau, y_{T-1}^\tau, \ldots, y_0^\tau) \\
&= (u_{T+\tau}, u_{T-1+\tau}, \ldots, u_\tau, y_{T-1+\tau}, \ldots, y_\tau) \\
&= (u_t, u_{t-1}, \ldots, u_{t-T}, y_{t-1}, \ldots, y_{t-T}),
\end{aligned}$$

and therefore $(u_t, u_{t-1}, \ldots, u_{t-T}, y_{t-1}, \ldots, y_{t-T}) \in \mathcal{Q}_F$.

Next, use $\hat{y}_t^s$ to denote the output of $\hat{S}$ when the pair $(\mathbf{u}^\tau, \mathbf{y}^\tau)$ is applied to $\hat{S}$ as its input. Recall (109), (116), we have

$$\begin{aligned}
\hat{y}'_t &= \theta(u_t, u_{t-1}, \ldots, u_{t-T}, y_{t-1}, \ldots, y_{t-T}) \\
&= \theta(u_T^\tau, u_{T-1}^\tau, \ldots, u_0^\tau, y_{T-1}^\tau, \ldots, y_0^\tau) \\
&= \hat{y}_T^s.
\end{aligned}$$

Since P is (C1), by definition, we have $\hat{y}_T^s = y_T^\tau$. Note that $y_T^\tau = y_{T+\tau} = y_t$, we have

$$\hat{y}'_t = \hat{y}_T^s = y_T^\tau = y_t. \tag{117}$$

Note that the choice of $t \geq T$ is arbitrary, we conclude that for any $(\mathbf{u}, \mathbf{y}) \in P$, $\hat{y}'_t = y_t$ for all $t \geq T$. Recall (114), we have $q'_t \in q'_o \cup (\bigcup_{i=1}^T \mathcal{U}^i \times \mathcal{Y}^i)$, and $|q'_o \cup (\bigcup_{i=1}^T \mathcal{U}^i \times \mathcal{Y}^i)| < \infty$, therefore $\hat{S}'$ is a DFM. This completes the proof of the forward implication of Lemma 2.

Recall the definition of (C1), then the backward implication of Lemma 2 follows. This completes the proof. □

4.2 Limitations of an Existing Construction

We present a technical result to illustrate the limitations of the "associating state" construction. For the purpose of exposition, we restrict our attention to systems (2) with $m = p = 1$, and the quantizer Q is in the form of (4). Now we are ready to state our result.

Theorem 8. Consider system (2) with $m = p = 1$, and Q is of the form (4). Assume that $\rho(A) < 1, \mathcal{U}$ contains positive, negative and zero elements, $0 \notin \mathcal{B}$, $CA^l \neq 0$ for all $l \in \mathbb{Z}_+$, and $\mathcal{A} \cap \mathcal{B} \neq \varnothing$. Then given a DFM observer $\hat{S}$, if $\hat{y}_t$ is uniquely determined by $(\tilde{y}_{t-1}, \ldots, \tilde{y}_{t-T}, u_t, u_{t-1}, \ldots, u_{t-T})$ for some $T \in \mathbb{Z}_+$, then for this $\hat{S}$, the observation gain $\gamma > 0$.

Remark. The hypotheses on $\hat{S}$ in Theorem 8 corresponds to the "associating state" construction of DFM observers.

Proof. For any observer $\hat{S}$ that has the following property:

$$(\tilde{y}_{t-1}, \ldots, \tilde{y}_{t-T}, u_t, u_{t-1}, \ldots, u_{t-T}) \xrightarrow{deterministic} \hat{y}_t \tag{118}$$

for some $T \in \mathbb{Z}_+$, we find an input $\mathbf{u}$ of system (2) such that prediction error occurs ($e_t \neq 0$) infinitely often. Similar to the derivation of Theorem 3, let $t_1 \in \mathbb{N}$, $\mathbf{u}^1 \in \mathcal{U}^{\mathbb{N}}$ be such that

$t_1 = \min\{t : F(u,t) \in \mathcal{A} \cap \mathcal{B}\}$, and $F(u^1, t_1) \in \mathcal{B}$. Given $\mathbf{u^1}$, we use $\mathbf{u_\tau^1}$ to denote the truncated sequence of $\mathbf{u^1}$: $\mathbf{u_\tau^1} = \{u_t^1 : 0 \le t \le t_1\}$. Without loss of generality, we assume that $T > t_1$ (otherwise just take $T = t_1 + 1$). Next, we divide this problem into two cases:

Case 1. $CA^rB \neq 0$ for infinitely many $r \in \mathbb{Z}_+$.

We start with constructing an input sequence $\mathbf{u}$ of system (2) as follows:

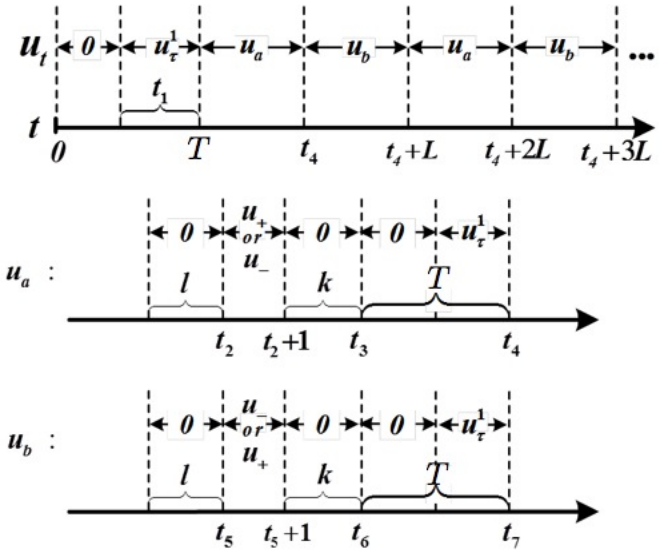


Figure 3: Input sequence construction for case 1.

As shown in the above figure, the idea to construct $\mathbf{u}$ is to drive the state x_t such that at a particular time, t_3 in this case, x_t satisfies

$$\|x_t\|_\infty < \frac{d_1}{K_A \|C\|_1} \quad \text{, and} \quad CA^T x_t \neq 0, \tag{119}$$

where $K_A = \sup\{\|A^t\| : t = 0, 1, 2, \cdots\}$, and d_1 is defined in (17). Then by repeating the input between time 0 and time T, we arrive at a prediction error. Shown in Figure 3 as u_a, the input u_t for $t > T$ consists of a sequence of zero input of length l, a one-step nonzero input, another sequence of zero input of length k, and a repetition of input between time 0 and time T (essentially a zero sequence followed by $\mathbf{u_\tau^1}$). By assumption, there exist two nonzero input $u_+, u_- \in \mathcal{U}$ and $u_+ > 0, u_- < 0$. By choosing the length of zero inputs l and k properly, x_{t_3} will satisfy (119).

We want to choose a k large enough such that $\|x_{t_3}\|$ is sufficiently small to guarantee that the outputs $\tilde{y}_t$ under u_+ or u_- are identical from t_3 to $t_4 - 1$, but different at t_4. And we choose an l large enough such that $\|x_{t_2}\|$ is sufficiently small, hence the terms associated with u_+ and u_- are dominant in x_{t_2+1}.

First we choose k. Let $K_{u_{max}} = \max\{|u_+|, |u_-|\}$ and $K_{u_{min}} = \min\{|u_+|, |u_-|\}$, then there exists $k_0 \in \mathbb{Z}_+$ such that

$$\|A^k\|_\infty < \frac{d_1}{2K_A K_{u_{max}} \|C\|_1 \|B\|_\infty}, \quad \forall k \geq k_0. \tag{120}$$

Since $CA^rB \neq 0$ for infinitely many $r \in \mathbb{Z}_+$, choose k such that $k \geq k_0$ and $CA^{T+k}B \neq 0$.

Next we choose l. By the proof of Theorem 1, $\|x_t\|_\infty$ is uniformly bounded. Let $\|x_t\|_\infty \leq b_3, b_3 \in \mathbb{R}$. Choose $l \in \mathbb{Z}_+$ such that

$$\|A^l\|_\infty < \epsilon \frac{K_{u_{min}} \|B\|_\infty}{b_3}. \tag{121}$$

where ϵ is to be determined. Then $x_{t_2} = A^l x_T$ implies $\|x_{t_2}\|_\infty < \epsilon K_{u_{min}} \|B\|_\infty$.

At $t = t_2 + 1$, $x_{t_2+1} = Ax_{t_2} + Bu_+$ or $x_{t_2+1} = Ax_{t_2} + Bu_-$. Then at $t = t_3$,

$$\|x_{t_3}\|_\infty < \frac{d_1}{2K_A \|C\|_1} + \frac{d_1 \|A\|_\infty \epsilon K_{u_{min}}}{2K_A \|C\|_1 K_{u_{max}}}$$

Choose $\epsilon < K_{u_{max}} (\|A\|_\infty K_{u_{min}})^{-1}$, then

$$\|x_{t_3}\|_\infty < \frac{d_1}{K_A \|C\|_1}$$

In order to achieve $CA^T x_{t_3} = CA^{T+k} x_{t_2+1} \neq 0$, we require that:

$$|CA^{T+k} A x_{t_2}| < \frac{1}{3} |CA^{T+k} B| K_{u_{min}}$$

which can be achieved by:

$$\epsilon < \frac{|CA^{T+k}B|}{3\|A\|_\infty \|B\|_\infty \|CA^{T+k}\|_1}$$

then $CA^{T+k}(Ax_{t_2} + Bu_+) \neq 0$ and $CA^{T+k}(Ax_{t_2} + Bu_-) \neq 0$, therefore $CA^{T+k}x_{t_2+1} \neq 0$.

Let ϵ in (121) be:

$$\epsilon = \frac{1}{2}\min\{\frac{K_{u_{max}}}{\|A\|_\infty K_{u_{min}}}, \quad \frac{|CA^{T+k}B|}{3\|A\|_\infty \|B\|_\infty \|CA^{T+k}\|_1}\}$$

and choose $l \in \mathbb{Z}_+$ to satisfy (121), then x_{t_3} satisfies (119). The outputs $\tilde{y}_t^1$ and $\tilde{y}_t^2$, which correspond to $u_{t_2} = u_+$ and $u_{t_2} = u_-$, are identical for $t = t_3, t_3+1, \cdots, t_3+T-1$. Since the quantizer Q is of the form (4), we see that $\tilde{y}_{t_3+T}^1 \neq \tilde{y}_{t_3+T}^2$.

In Figure 3, repeat the sequence u_a and u_b infinitely many times, and choose u_+ or u_- different from the previous repetition, say u_+ for u_a and u_- for u_b. Under this input sequence $\mathbf{u}$, recall that $\hat{S}$ satisfies (118), we have $\hat{y}_{t_4} = \hat{y}_{t_7}$. But by the previous discussions, we see that $\tilde{y}_{t_4} \neq \tilde{y}_{t_7}$. Therefore $e_t \neq 0$, for some $t \in \{t_4, t_7\}$. Consequently, $e_t \neq 0$ for infinitely many $t \in \mathbb{N}$, and the observation gain $\gamma > 0$.

Case 2. $CA^rB \neq 0$ for finitely many $r \in \mathbb{Z}_+$.

In this case, there exists $N \in \mathbb{Z}_+$ such that $CA^rB = 0$ for all $r \geq N$. For $t \geq N$, the output of the LTI system y_t is :

$$\begin{aligned} y_t &= CA^t x_0 + \sum_{\tau=0}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t \\ &= CA^t x_0 + \sum_{\tau=t-N}^{t-1} CA^{t-1-\tau}Bu_\tau + Du_t \end{aligned}$$

This shows that y_t only depends on the initial condition x_0 and previous N steps of input u_t. We will show that under some initial condition, prediction error will occur infinitely often.

Since $CA^T \neq 0$ for all $T \in \mathbb{Z}_+$, there is $v \in \mathbb{R}^n$ such that $CA^t v \neq 0$ infinitely often (see Proposition 3 in the Section 2.2.2). And we enforce that:

$$\|v\|_\infty < \frac{d_1}{K_A \|C\|_1},$$

which can easily done by rescaling v. Then there is a sequence of time instances: $\{T_i\}_{i=1}^{\infty}$ such that for all $i \in \mathbb{Z}_+$:

$$CA^{T_i} v \neq 0,$$

$$T_{i+1} - T_i \geq \max\{N+1, T\}.$$

Given the T_i's, the input sequence $\mathbf{u}$ is shown below:

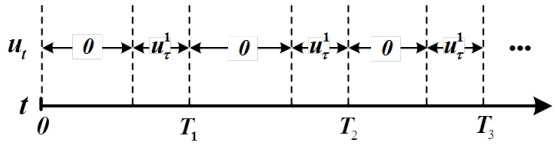


Figure 4: Input sequence construction for case 2.

Under this input $\mathbf{u}$, let $\tilde{y}_t^1$ and $\tilde{y}_t^2$ be the outputs of system (2) which correspond to the initial states $x_0^1 = v$ and $x_0^2 = -v$ respectively. Then $\tilde{y}_{T_i}^1 \neq \tilde{y}_{T_i}^2$ for all $i \in \mathbb{Z}_+$. But $\hat{y}_t^1$ and $\hat{y}_t^2$, which correspond to the two different initial states, are the same at $t = T_i$ for all $i \in \mathbb{Z}_+$. So we conclude prediction error occurs infinitely often for either x_0^1 or x_0^2. □

4.3 A New Construction

We propose a new construction of DFM observers to overcome the limitations of the "associate state" construction stated in the previous section.

Theorem 9. For certain instance of system (2) that satisfies the hypotheses in Theorem 8, there is a DFM observer that achieves $\gamma = 0$.

Remark. The system (2) in Theorem 9 is an instance that is (C2) but not (C1).

Proof. We consider the following example.

Example 5. Given system (2) with parameters $A = 0.5, B = C = 1, D = 0$, the input set is $\mathcal{U} = \{0, 1, -1\}$, and the initial state x_0 satisfies $|x_0| < 2$. The quantizer Q is described by: $Q((-\infty, 0.5)) = 0$, and $Q([0.5, \infty)) = 1$.

Clearly, Example 5 satisfies the hypotheses in Theorem 8. Next, consider a DFM observer described as:

$$q_{t+1} = f(q_t, u_t), \tag{122a}$$

$$\hat{y}_t = g(q_t), \tag{122b}$$

where $t \in \mathbb{N}$, $q_t \in \mathcal{Q}$, and $\mathcal{Q} = \{0, 1, -1\}^5$. Recall the notation of $[q_t]_i$ being the i-th element of q_t, the function $f : \mathcal{Q} \times \{0, 1, -1\} \to \mathcal{Q}$ in (122a) is described by:

For any $q \in \mathcal{Q}$, any $u \in \{0, 1, -1\}$,

$$\begin{aligned}
&[f(q,u)]_1 = [q]_2, && [f(q,u)]_2 = u,\\
&[f(q,u)]_3 = \begin{cases} [q]_3, & \text{if } u = 0,\\ [q]_4, & \text{if } u \neq 0, \end{cases} && [f(q,u)]_4 = \begin{cases} [q]_4, & \text{if } u = 0,\\ [q]_5, & \text{if } u \neq 0, \end{cases}\\
&[f(q,u)]_5 = \begin{cases} [q]_5, & \text{if } u = 0,\\ u, & \text{if } u \neq 0. \end{cases}
\end{aligned} \tag{123}$$

Essentially, as we shall see in the following, q_t stores the last two steps of input as well as the last three nonzero inputs.

The output function $g : \mathcal{Q} \to \{0, 1\}$ is described by the following table:

For any $q \in \mathcal{Q}$,

$[q]_1$	$[q]_2$	$[q]_3$	$[q]_4$	$[q]_5$	$g(q)$
0	-1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	0
0	1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	1
1	-1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	0
-1	0	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	0
1	1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	1
-1	-1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	0
0	0	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	0
1	0	$0, 1, \text{or} -1$	1	$0, 1, \text{or} -1$	1
1	0	$0, 1, \text{or} -1$	-1	$0, 1, \text{or} -1$	0
-1	1	1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	1
-1	1	-1	$0, 1, \text{or} -1$	$0, 1, \text{or} -1$	0
All q not listed in the above					0

Table 1: Look-up table of the function g.

Lastly, we let the initial state of the observer (122) be $q_0 = (0, 0, 0, 0, 0)$. This completes the construction of the DFM observer (122). Next, we will show that the observer (122) achieves $\gamma = 0$ for Example 5.

For any $\mathbf{u} \in \{0, 1, -1\}^{\mathbb{N}}$, and any $|x_0| < 2$, we first consider the case when $u_t \neq 0$ for finitely many $t \in \mathbb{N}$. Then there is $T_u \in \mathbb{Z}_+$ such that $u_t = 0$ for all $t \geq T_u$. Since

$$x_t = 0.5^2 x_{t-2} + u_{t-1} + 0.5u_{t-2},$$

we have $x_t = 0.25x_{t-2}$ for all $t \geq T_u + 2$. We also observe that $|x_t| < 2$ for all $t \in \mathbb{N}$: At $t = 0$, $|x_0| < 2$; assume $|x_t| < 2$, at $t + 1$, $|x_{t+1}| = |0.5x_t + u_t| \leq |0.5x_t| + |u_t| < 0.5 \cdot 2 + 1 = 2$, therefore by induction $|x_t| < 2$ for all $t \in \mathbb{N}$. Therefore $|x_t| < 0.25 \cdot 2 = 0.5$ for all $t \geq T_u + 2$, and consequently $\tilde{y}_t = 0$ for all $t \geq T_u + 2$. Recall the definition of f (123), we see that

$$[q_t]_1 = u_{t-2}, \quad [q_t]_2 = u_{t-1}, \quad \text{for all } t \geq 2. \tag{124}$$

Indeed, $[q_2]_1 = [q_1]_2 = u_0$, and $[q_2]_2 = u_1$, therefore (124) holds at $t = 2$. Assume

(124) holds at t, then $[q_{t+1}]_1 = [q_t]_2 = u_{t-1}$, and $[q_{t+1}]_2 = u_t$, therefore (124) holds at $t+1$. By induction, (124) holds. Since $u_t = 0$ for all $t \geq T_u$ by assumption, we have $[q_t]_1 = [q_t]_2 = 0$ for all $t \geq T_u + 2$. By the definition of g shown in Table 1, we have $\hat{y}_t = 0$ for all $t \geq T_u + 2$. Therefore $\tilde{y}_t = \hat{y}_t$ for all $t \geq T_u + 2$.

Next, we consider the case when $u_t \neq 0$ for infinitely many $t \in \mathbb{N}$. For all $t \geq 2$, $x_t = 0.5^2 x_{t-2} + u_{t-1} + 0.5u_{t-2}$. If $(u_{t-2}, u_{t-1}) = (0, -1)$, then $x_t = 0.25x_{t-2} - 1$. Recall that $|x_{t-2}| < 2$, we have $x_t \in (-1.5, -0.5)$, and therefore $\tilde{y}_t = 0$. Recall (124), we have $([q_t]_1, [q_t]_2) = (0, -1)$, and by Figure 1, $\hat{y}_t = 0$. Therefore we have: For all $t \geq 2$, if $(u_{t-2}, u_{t-1}) = (0, -1)$, then $\hat{y}_t = \tilde{y}_t$.

Similarly, for all $(u_{t-2}, u_{t-1}) \in \mathcal{U}^2$ such that $u_{t-1} + 0.5u_{t-2} \leq 0$ or $u_{t-1} + 0.5u_{t-2} \geq 1$, apply the preceding argument, we conclude that: For all $t \geq 2$, if $(u_{t-2}, u_{t-1}) \in \{(0, -1), (0, 1), (1, -1),$
$(-1, 0), (1, 1), (-1, -1), (0, 0)\}$, then $\hat{y}_t = \tilde{y}_t$.

In the following, we only need to consider the cases when $(u_{t-2}, u_{t-1}) = (1, 0)$ or $(u_{t-2}, u_{t-1}) = (-1, 1)$. In both cases, $x_t = 0.25x_{t-2} + 0.5$.

Since $u_t \neq 0$ infinitely many times, there is $T'_u \in \mathbb{Z}_+$ such that $u_{T'_u} \neq 0$. For any $t \geq T'_u + 3$, define τ^* as

$$\tau^* = \max\{\tau \in \mathbb{N} : \tau \leq t - 3, \text{ and } u_\tau \neq 0\}.$$

Since $u_{T'_u} \neq 0$ and $T'_u \leq t - 3$, $\{\tau \in \mathbb{N} : \tau \leq t - 3, \text{ and } u_\tau \neq 0\}$ is nonempty and τ^* is well-defined. We claim that if $(u_{t-2}, u_{t-1}) = (1, 0)$, then $[q_t]_4 = u_{\tau^*}$. To see this, by (123), $[q_{\tau^*+1}]_5 = u_{\tau^*}$. If $\tau^* = t - 3$, then by (123), $[q_t]_4 = [q_{t-1}]_4 = [q_{t-2}]_5 = u_{\tau^*}$, and the claim holds. If $\tau^* < t - 3$, then by the definition of τ^*, $u_{\tau^*+1} = u_{\tau^*+2} = \cdots = u_{t-3} = 0$. By (123), $[q_{\tau^*+2}]_5 = [q_{\tau^*+3}]_5 = \cdots = [q_{t-2}]_5 = u_{\tau^*}$. Recall $(u_{t-2}, u_{t-1}) = (1, 0)$, we see $[q_t]_4 = [q_{t-1}]_4 = [q_{t-2}]_5 = u_{\tau^*}$, and the claim also holds. Similarly, if $(u_{t-2}, u_{t-1}) = (-1, 1)$, then $[q_t]_3 = [q_{t-1}]_4 = [q_{t-2}]_5$. As stated previously, $[q_{t-2}]_5 = u_{\tau^*}$, therefore

$[q_t]_3 = u_{\tau^*}$. We conclude that for any $t \geq T'_u + 3$,

$$(u_{t-2}, u_{t-1}) = (1, 0) \Rightarrow [q_t]_4 = u_{\tau^*}, \quad \text{and} \quad (u_{t-2}, u_{t-1}) = (-1, 1) \Rightarrow [q_t]_3 = u_{\tau^*}. \tag{125}$$

Since $x_{\tau^*+1} = 0.5x_{\tau^*} + u_{\tau^*}$, and $|x_{\tau^*}| < 2$, we see that $u_{\tau^*} = 1$ implies $x_{\tau^*+1} > 0$, while $u_{\tau^*} = -1$ implies $x_{\tau^*+1} < 0$. By the definition of τ^*, $x_{t-2} = 0.5^{t-3-\tau^*} x_{\tau^*+1}$. If $(u_{t-2}, u_{t-1}) = (1, 0)$ or $(u_{t-2}, u_{t-1}) = (-1, 1)$, then $x_t = 0.25x_{t-2} + 0.5$. Consequently, $u_{\tau^*} = 1$ implies $\tilde{y}_t = 1$, while $u_{\tau^*} = -1$ implies $\tilde{y}_t = 0$. Recall Table 1, (125), assume $(u_{t-2}, u_{t-1}) = (1, 0)$ or $(u_{t-2}, u_{t-1}) = (-1, 1)$, if $u_{\tau^*} = 1$, then $\hat{y}_t = 1$, while if $u_{\tau^*} = -1$, then $\hat{y}_t = 0$. Therefore for any u_{τ^*} (note that $u_{\tau^*} \neq 0$), $\tilde{y}_t = \hat{y}_t$. Combined with previous observations, we see that when $u_t \neq 0$ for infinitely many $t \in \mathbb{N}$, there is $T'_u \in \mathbb{Z}_+$ for any $t \geq T'_u + 3$, $\tilde{y}_t = \hat{y}_t$.

We conclude that for any $\mathbf{u} \in \{0, 1, -1\}^{\mathbb{N}}$, and any $|x_0| < 2$, $\tilde{y}_t \neq \hat{y}_t$ for finitely many t, and therefore $\gamma = 0$ is an observation gain. This completes the proof. $\square$

Intended to be blank.

5 Control Design based on Finite Memory Observability

In this chapter, we motivate and formulate a control design problem of system (2), and then propose a procedure to synthesis DFM controllers based on finite memory observability.

5.1 Background and Motivation

A natural next step is to look at the controller synthesis of systems (2) that are observable in the sense of (C1), (C2), or (C3). And a common theme in control theory is to design a control input to stabilize an unstable system around its equilibrium.

However, along this line of reasoning, we quickly encounter some technical challenges. One of the difficulties associated with finitely quantized inputs is that objectives that are feasible under analog control become infeasible in this setting. For example, the origin of system (2) is not stablizable in the traditional sense of Lyapunov under any control input sequence when matrix A is Schur unstable. This impossibility of stabilizing an unstable system in the sense of Lyapunov using only quantized state feedback is clearly formulated in [1]. In particular, the author of [1] show that for any system with an Schur-unstable A matrix, and any feedback control strategy that is determined by quantized state information, the set of all initial states whose closed-loop trajectories tend to the origin as time tends to infinity has Lebesgue measure zero.

Under these circumstances, we shift our attention from the stabilization problem to the problem of constraint satisfaction and cost minimization. In particular, we consider system (2) with a Schur-stable matrix A, but the state x_t is required to stay within some constraint set. We formulate a control objective as controlling the system state to stay within the prescribed constraint set, and optimizing some cost function of the system state and input.

5.2 Setup and Problem Statement

In the following, we apply our analysis of finite memory observability (C1) to synthesize controllers for a class of systems (2) with state constraints. In particular, we consider

$$x_{t+1} = Ax_t + Bu_t, \tag{126a}$$

$$y_t = Cx_t, \tag{126b}$$

$$\tilde{y}_t = Q(y_t), \tag{126c}$$

$$x_t \in \mathcal{X}, u_t \in \mathcal{U}, \tilde{y}_t \in \mathcal{Y}, \tag{126d}$$

where equations (126a) to (126c) represent system (2) with D being the zero matrix. (126d) describes the constraints on the system: $\mathcal{X} \in \mathbb{R}^n$ is the constraint on system states, and $\mathcal{U}, \mathcal{Y}$ are the finite input and output sets of the system as stated previously.

In the following, we formulate a first case of control design of system (126). Let $x_e \in \mathbb{R}^n$ be an equilibrium point of (126):

$$Ax_e + Bu_e = x_e, \quad \text{for some } u_e \in \mathcal{U}. \tag{127}$$

Remark. Note that system (126) is *not* "translation invariant", as opposed to LTI systems. For example, the origin could be continuous for Q, but be discontinuous for Q', where Q' is a translation of Q. Due to this reason, we define the equilibrium point of system (126) as in (127).

Next, we consider the following problem.

Problem 1. Given system (126) and its equilibrium x_e, assume that only $\tilde{y}_t$ is available at each $t \in \mathbb{N}$. Design $\{u_t\}_{t=0}^{\infty} \in \mathcal{U}^{\mathbb{N}}$ such that the following are satisfied:

i. (Constraint Satisfaction) $x_t \in \mathcal{X}$, for all $t \in \mathbb{N}$.

ii. (Attractivity) $\lim_{t\to\infty} x_t = x_e$.

iii. (Cost Optimization) $\sum_{t=0}^{\infty} l(x_t, u_t)$ is minimized.

Here $l : \mathcal{X} \times \mathcal{U} \to \mathbb{R}_{\geq 0}$ is a given cost function, and $l(x_e, u_e) = 0$.

5.3 A Control Design Procedure

In the following, we make efforts to solve Problem 1. In particular, we propose a control design, and we also study the conditions under which the proposed design will work. Such conditions will be stated in terms of the parameters of system (126) and the equilibrium x_e.

First, we "translate" the control objectives i) and ii) in Problem 1 into objectives in terms of $\tilde{y}_t$. Let $\tilde{y}_e$ be

$$\tilde{y}_e = Q(Cx_e). \tag{128}$$

And given system (126), for any $\tilde{y} \in \mathcal{Y}$, define

$$Q_C^{-1}(\tilde{y}) = \{x \in \mathbb{R}^n : Q(Cx) = \tilde{y}\}. \tag{129}$$

Essentially, $Q_C^{-1}(\tilde{y})$ is the inverse image of $\tilde{y}$ under the mapping $Q \circ C$ (pp. 49, [23]).

Next, we make an assumption on system (126) and the equilibrium x_e.

Assumption 1. (Local Attractivity) For any $t \in \mathbb{N}$, if $x_t \in Q_C^{-1}(\tilde{y}_e)$, and $u_\tau = u_e$ for all $\tau \geq t$, then

$$x_\tau \in \mathcal{X}, \quad \text{for all } \tau \geq t, \tag{130}$$

$$\lim_{\tau\to\infty} x_\tau = x_e. \tag{131}$$

Now, we are ready to formulate a sub-problem of Problem 1.

Problem 2. Given system (126) and its equilibrium x_e, assume Assumption 1 holds. Find $N \in \mathbb{Z}_+$ and $\{u_t\}_{t=0}^N \in \mathcal{U}^N$ such that $u_N = u_e$, $\tilde{y}_N = \tilde{y}_e$, and $x_t \in \mathcal{X}$, for all $0 \leq t \leq N$.

We observe that if $\{u_t\}_{t=0}^N$ is a solution to Problem 2, then the control input sequence $(\{u_t\}_{t=0}^N, u_e, u_e, \dots)$ solves items i) and ii) in Problem 1. To see this, since $\tilde{y}_N = \tilde{y}_e$ by assumption, $x_N \in Q_C^{-1}(\tilde{y}_e)$. By Assumption 1, and note that $u_t = u_e$ for all $t \geq N$, we have $x_t \in \mathcal{X}$ for all $t \geq N$, and $\lim_{t\to\infty} x_t = x_e$. Consequently, items i) and ii) in Problem 1 are satisfied. Based on this discussion, to solve the first parts of Problem 1, it suffices to solve Problem 2.

Next, we study the conditions under which Assumption 1 holds. The following approach and derivation are based on the works on "maximal output admissible sets" [27] and on backward reachability iterations (pp. 153, [28]).

Given system (126) and its equilibrium x_e, and write $f_e(x) = Ax + Bu_e$, we define a set $\mathcal{X}^\infty$ as

$$\mathcal{X}^\infty = \{x \in \mathbb{R}^n : x \in \mathcal{X}, f_e(x) \in \mathcal{X}, f_e \circ f_e(x) \in \mathcal{X}, f_e \circ f_e \circ f_e(x) \in \mathcal{X}, \dots\}. \quad (132)$$

Essentially, $\mathcal{X}^\infty$ is the maximal positive invariant set corresponding with $\mathcal{X}$. Then, we propose a result on the conditions under which Assumption 1 holds.

Lemma 3. Given system (126) and its equilibrium x_e, assume $\rho(A) < 1$. If $Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}^\infty$, then Assumption 1 holds. Moreover, if in addition $\mathcal{X}$ is bounded, and $x_e \in int(\mathcal{X})$, then $\mathcal{X}^\infty$ is finitely determined and can be computed in finite steps.

Remark. In the second part of this Lemma, we characterize the situations under which we can computationally determine the condition "$Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}^\infty$". In other words, the conditions "$\mathcal{X}$ is bounded and $x_e \in int(\mathcal{X})$" are not necessary for Assumption 1 to hold.

Proof. Since $Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}^\infty$, for any $x_t \in Q_C^{-1}(\tilde{y}_e)$, $x_t \in \mathcal{X}^\infty$. If $u_\tau = u_e$ for all $\tau \geq t$, recall (132), we see that $x_{t+1} \in \mathcal{X}, x_{t+2} \in \mathcal{X}, \dots$. Therefore (130) is satisfied. Next, we observe that for any $h \in \mathbb{Z}_+$, $x_{t+h} - x_e = A^h(x_t - x_e)$. To see this, for $h = 1$, $x_{t+1} - x_e = Ax_t + Bu_e - x_e = A(x_t - x_e + x_e) + Bu_e - x_e = A(x_t - x_e)$. Assume $x_{t+h} - x_e = A^h(x_t - x_e)$ for some $h \geq 1$, $x_{t+h+1} - x_e = Ax_{t+h} + Bu_e - x_e = A(A^h(x_t - x_e) + x_e) + Bu_e - x_e = A^{h+1}(x_t - x_e)$. Therefore $x_{t+h} - x_e = A^h(x_t - x_e)$ for all $h \in \mathbb{Z}_+$.

Since $\rho(A) < 1$, $\lim_{h\to\infty} A^h = \mathbf{0}$, and consequently $\lim_{h\to\infty} x_{t+h} - x_e = 0$. Therefore (131) holds. We see that Assumption 1 holds.

For the second part of Lemma 3, first we review the term "finitely determined" introduced in [27]. Given $t \in \mathbb{Z}_+$, we use the notation $f_e^{(t)}$ to denote the t-time composition of the function $f_e : f_e^{(t)}(x) = f_e \circ \cdots \circ f_e(x)$. Next, define set

$$\mathcal{X}^t = \{x \in \mathbb{R}^n : x \in \mathcal{X}, f_e^{(k)}(x) \in \mathcal{X} \text{ for } k = 1, \ldots, t\}.$$

Then $\mathcal{X}^\infty$ is said to be finitely determined if for some $t \in \mathbb{Z}_+$, $\mathcal{X}^\infty = \mathcal{X}^t$.

For any $x \in \mathbb{R}^n$, let $z = x - x_e$. Let $\bar{\mathcal{X}} = \mathcal{X} - x_e$, then $x \in \mathcal{X}$ if and only if $z \in \bar{\mathcal{X}}$. Next, we observe that for any $k \in \mathbb{Z}_+$,

$$f_e^{(k)}(x) \in \mathcal{X} \quad \text{if and only if} \quad A^k z \in \bar{\mathcal{X}}. \tag{133}$$

Indeed, for $k = 1$, $f_e^{(1)}(x) \in \mathcal{X} \Leftrightarrow Ax + Bu_e \in \mathcal{X}$. Recall (127), we note that $Ax + Bu_e - x_e = A(z + x_e) + Bu_e - x_e = Az$. Since $Ax + Bu_e \in \mathcal{X} \Leftrightarrow Ax + Bu_e - x_e \in \bar{\mathcal{X}}$, we have $f_e^{(1)}(x) \in \mathcal{X} \Leftrightarrow Az \in \bar{\mathcal{X}}$. More generally, for any $k \in \mathbb{Z}_+$, $f_e^{(k+1)}(x) \in \mathcal{X} \Leftrightarrow Af_e^{(k)}(x) + Bu_e \in \mathcal{X}$. By the derivation for $k = 1$, $Af_e^{(k)}(x) + Bu_e \in \mathcal{X} \Leftrightarrow A(f_e^{(k)}(x) - x_e) \in \bar{\mathcal{X}}$. We observe that $f_e^{(k)}(x) - x_e = A^k(x - x_e)$: For $k = 1$, $Ax + Bu_e - x_e = A(x - x_e)$; assume the equation holds for some $k \geq 1$, then $f_e^{(k+1)}(x) - x_e = A(f_e^{(k)}(x)) + Bu_e - x_e = A(f_e^{(k)}(x)) - Ax_e = A(f_e^{(k)}(x) - x_e) = A^{k+1}(x - x_e)$. Consequently, $f_e^{(k)}(x) - x_e = A^k(x - x_e)$ holds for all $k \in \mathbb{Z}_+$. Therefore, $A(f_e^{(k)}(x) - x_e) \in \bar{\mathcal{X}} \Leftrightarrow A^{k+1}(x - x_e) \in \bar{\mathcal{X}} \Leftrightarrow A^{k+1}z \in \bar{\mathcal{X}}$. Therefore for any $k \in \mathbb{Z}_+$, $f_e^{(k+1)}(x) \in \mathcal{X} \Leftrightarrow A^{k+1}z \in \bar{\mathcal{X}}$, and (133) holds.

Next, define set

$$\bar{\mathcal{X}}^t = \{z \in \mathbb{R}^n : z \in \bar{\mathcal{X}}, A^k z \in \bar{\mathcal{X}} \text{ for } k = 1, \ldots, t\},$$

then we observe that

$$\mathcal{X}^t = \bar{\mathcal{X}}^t + x_e. \tag{134}$$

Indeed, recall (133), $x \in \mathcal{X}^t \Leftrightarrow (x - x_e) \in \bar{\mathcal{X}}^t$. Consequently, $\mathcal{X}^t = \bar{\mathcal{X}}^t + x_e$. Similarly, let $\bar{\mathcal{X}}^\infty = \{z \in \mathbb{R}^n : z \in \bar{\mathcal{X}}, A^k z \in \bar{\mathcal{X}} \text{ for } k = 1, 2, \dots\}$, then

$$\mathcal{X}^\infty = \bar{\mathcal{X}}^\infty + x_e.$$

Since $x_e \in int(\mathcal{X})$, and $\bar{\mathcal{X}} = \mathcal{X} - x_e$, we see that $0 \in int(\bar{\mathcal{X}})$. Then there is $r > 0$ such that $B_r(0) \subset \bar{\mathcal{X}}$. Since $\rho(A) < 1$, $\lim_{k\to\infty} \|A^k\| = 0$. Since $\mathcal{X}$ is bounded, $\bar{\mathcal{X}}$ is also bounded. Therefore, there is $\tau^* \in \mathbb{Z}_+$ such that $A^{\tau^*+h} x \in B_r(0) \subset \bar{\mathcal{X}}$, for all $h \in \mathbb{N}$, and all $x \in \bar{\mathcal{X}}$. Therefore, $\bar{\mathcal{X}}^\infty = \bar{\mathcal{X}}^{\tau^*}$. Consequently, $\mathcal{X}^\infty = \mathcal{X}^{\tau^*}$, and $\mathcal{X}$ is finitely determined.

Next, we introduce a procedure to compute $\mathcal{X}$. Let $\mathcal{O}_0 = \bar{\mathcal{X}}$, define

$$\mathcal{O}_{k+1} = \{z \in \mathbb{R}^n : Az \in \mathcal{O}_k\} \bigcap \mathcal{O}_k, k = 0, 1, 2, \dots \tag{135}$$

Then we observe that

$$\mathcal{O}_k = \bar{\mathcal{X}}^k, \quad \forall\, k \in \mathbb{Z}_+.$$

To see this, for $k = 1$, $z \in \mathcal{O}_1 \Leftrightarrow Az \in \mathcal{O}_0$, and $z \in \mathcal{O}_0$. Recall $\mathcal{O}_0 = \bar{\mathcal{X}}$, we see that $\mathcal{O}_1 = \bar{\mathcal{X}}^1$. Assume $\mathcal{O}_k = \bar{\mathcal{X}}^k$ for some $k \geq 1$. If $z \in \mathcal{O}_{k+1}$, then $z \in \mathcal{O}_k = \bar{\mathcal{X}}^k$. Also $Az \in \bar{\mathcal{X}}^k$, therefore $A^k(Az) \in \bar{\mathcal{X}}$. Therefore $z \in \bar{\mathcal{X}}^{k+1}$. If $z \in \bar{\mathcal{X}}^{k+1}$, then $Az \in \bar{\mathcal{X}}$, $A^2 z \in \bar{\mathcal{X}}, \cdots, A^{k+1} z \in \bar{\mathcal{X}}$. Therefore $Az \in \bar{\mathcal{X}}^k = \mathcal{O}_k$, and $z \in \{z \in \mathbb{R}^n : Az \in \mathcal{O}_k\}$. Note that $z \in \bar{\mathcal{X}}^{k+1} \subset \bar{\mathcal{X}}^k = \mathcal{O}_k$, we see that $z \in \mathcal{O}_{k+1}$. We see that $\mathcal{O}_{k+1} = \bar{\mathcal{X}}^{k+1}$.

We conclude that $\mathcal{O}_k = \bar{\mathcal{X}}^k$, for all $k \in \mathbb{Z}_+$.

The sets $\mathcal{O}_k$ (135) can be computed iteratively, and such computations are relatively straightforward especially when $\mathcal{X}$ is a polytope. When $\mathcal{O}_{k+1} = \mathcal{O}_k$ for some k, terminate the iteration, and we have $\mathcal{X}^\infty = \mathcal{O}_k + x_e$. This completes the procedure of computing $\mathcal{X}^\infty$. □

So far, we have transformed Problem 1 to a sub-problem Problem 2. Next, we propose a first solution to Problem 2. We want to point out that this solution is rather basic and relatively conservative. Other approaches could be designed later.

The idea of this solution is: Given a system (126) that is (C1), we first make sure x_t stays in $\mathcal{X}$ for $0 \leq t \leq T$. And after T steps, design u_t based on the output of the observer $\hat{S}$, which is identical to $\tilde{y}_t$ for $t \geq T$. We will identify the initial states under which this approach will work.

Now we start formulation this solution. Recall $f_e(x) = Ax + Bu_e$, define

$$\mathcal{R}^{(0)} = \{x \in \mathbb{R}^n : x \in Q_C^{-1}(\tilde{y}_e), f_e(x) \in Q_C^{-1}(\tilde{y}_e), f_e \circ f_e(x) \in Q_C^{-1}(\tilde{y}_e), \dots\}. \quad (136)$$

Essentially, $\mathcal{R}^{(0)}$ is the maximal positive invariant set within $Q_C^{-1}(\tilde{y}_e)$. Similar to $\mathcal{X}^\infty$, if $\rho(A) < 1$, $Q_C^{-1}(\tilde{y}_e)$ is bounded, and $x_e \in int(Q_C^{-1}(\tilde{y}_e))$, then $\mathcal{R}^{(0)}$ is finitely determined and could be computed.

For any $k \in \mathbb{Z}_+$, we define the backward reachability sets as

$$\begin{aligned}
\mathcal{R}^{(1)} &= \{x \in \mathcal{X} : Ax + Bu \in \mathcal{R}^{(0)}, \text{ for some } u \in \mathcal{U}\}, \\
\mathcal{R}^{(2)} &= \{x \in \mathcal{X} : Ax + Bu \in \mathcal{R}^{(1)}, \text{ for some } u \in \mathcal{U}\}, \\
&\vdots \\
\mathcal{R}^{(k)} &= \{x \in \mathcal{X} : Ax + Bu \in \mathcal{R}^{(k-1)}, \text{ for some } u \in \mathcal{U}\}.
\end{aligned} \quad (137)$$

If Assumption 1 holds, by (130), we have $\mathcal{R}^{(0)} \in \mathcal{X}$. For any $x \in \mathcal{R}^{(0)}$, $Ax + Bu_e \in Q_C^{-1}(\tilde{y}_e)$, $f_e(Ax + Bu_e) \in Q_C^{-1}(\tilde{y}_e)$, $f_e \circ f_e(Ax + Bu_e) \in Q_C^{-1}(\tilde{y}_e)$, …, therefore $Ax + Bu_e \in \mathcal{R}^{(0)}$. Consequently, $\mathcal{R}^{(0)} \subset \mathcal{R}^{(1)}$.

More generally, we observe that

$$\mathcal{R}^{(k)} \subset \mathcal{R}^{(k+1)}, \quad \forall\, k \in \mathbb{N}. \quad (138)$$

To see this, we have shown the case $k = 0$ previously. For $k = 1$, for any $x \in \mathcal{R}^{(1)}$,

$Ax + Bu \in \mathcal{R}^{(0)} \subset \mathcal{R}^{(1)}$ for some $u \in \mathcal{U}$, therefore $x \in \mathcal{R}^{(2)}$, and $\mathcal{R}^{(1)} \subset \mathcal{R}^{(2)}$. Assume $\mathcal{R}^{(k-1)} \subset \mathcal{R}^{(k)}$ for some $k \geq 2$, then for any $x \in \mathcal{R}^{(k)}$, $Ax + Bu \in \mathcal{R}^{(k-1)} \subset \mathcal{R}^{(k)}$ for some $u \in \mathcal{U}$, therefore $x \in \mathcal{R}^{(k+1)}$, and $\mathcal{R}^{(k)} \subset \mathcal{R}^{(k+1)}$. By induction, (138) holds.

Given $\mathcal{R}^{(k)}$ (137) for some $k \in \mathbb{Z}_+$, and given $T \in \mathbb{Z}_+$, define set $\mathcal{Y}_c \subset \mathcal{Y}$ as: $\tilde{y} \in \mathcal{Y}_c$ if there is $(u_0, u_1, \ldots, u_{T-1}) \in \mathcal{U}^T$ such that

$$\begin{aligned}
&Q_C^{-1}(\tilde{y}) \subset \mathcal{X}, \\
&AQ_C^{-1}(\tilde{y}) + Bu_0 \subset \mathcal{X}, \\
&A^2Q_C^{-1}(\tilde{y}) + Bu_1 + ABu_0 \subset \mathcal{X}, \\
&\quad \vdots \\
&A^{T-1}Q_C^{-1}(\tilde{y}) + \sum_{\tau=0}^{T-2} A^{T-2-\tau}Bu_\tau \subset \mathcal{X}, \\
&A^{T}Q_C^{-1}(\tilde{y}) + \sum_{\tau=0}^{T-1} A^{T-1-\tau}Bu_\tau \subset \mathcal{R}^{(k)}.
\end{aligned} \tag{139}$$

And define

$$\mathcal{C} = \bigcup_{\tilde{y} \in \mathcal{Y}_c} Q_C^{-1}(\tilde{y}). \tag{140}$$

We propose a solution to Problem 2 when system (126) is finite memory observable, and its initial state x_0 satisfies $x_0 \in \mathcal{C}$.

Assume system (126) is finite memory observable with parameter T (see Definition 3), and let $\mathcal{Y}_c$ (139) and $\mathcal{C}$ (140) be defined corresponding with this T. If $x_0 \in \mathcal{C}$, then $x_0 \in Q_C^{-1}(\tilde{y})$ for some $\tilde{y} \in \mathcal{Y}_c$. Let $(\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_{T-1}) \in \mathcal{U}^T$ be such that (139) holds for $\tilde{y}$. Let the input u_t of system (126) be $u_t = \bar{u}_t$ for $0 \leq t \leq T-1$. By (139), we have

$$\begin{aligned}
&x_t \in \mathcal{X}, \quad t = 0, 1, \ldots, T-1, \\
&x_T \in \mathcal{R}^{(k)}.
\end{aligned} \tag{141}$$

Since $x_T \in \mathcal{R}^{(k)}$, by (137), there is $(\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_k) \in \mathcal{U}^k$ such that if the input u_t of system

(126) satisfies $u_{t+T} = \underline{u}_{t+1}$ for $0 \le t \le k-1$, then

$$\begin{aligned} x_{T+1} &\in \mathcal{R}^{(k-1)}, \\ x_{T+2} &\in \mathcal{R}^{(k-2)}, \\ &\vdots \\ x_{T+k} &\in \mathcal{R}^{(0)}. \end{aligned} \tag{142}$$

Note that $\mathcal{R}^{(0)} \subset Q_C^{-1}(\tilde{y}_e)$, and $\mathcal{R}^{(i)} \subset \mathcal{X}$ for $i = 0, \dots, k$, we see that

$$\begin{aligned} x_t \in \mathcal{X}, \quad t = T+1, T+2, \dots, T+k, \\ \tilde{y}_{T+k} = \tilde{y}_e. \end{aligned} \tag{143}$$

Next, we determine $(\underline{u}_1, \underline{u}_2, \dots, \underline{u}_k)$ based on the finite memory observability of system (126).

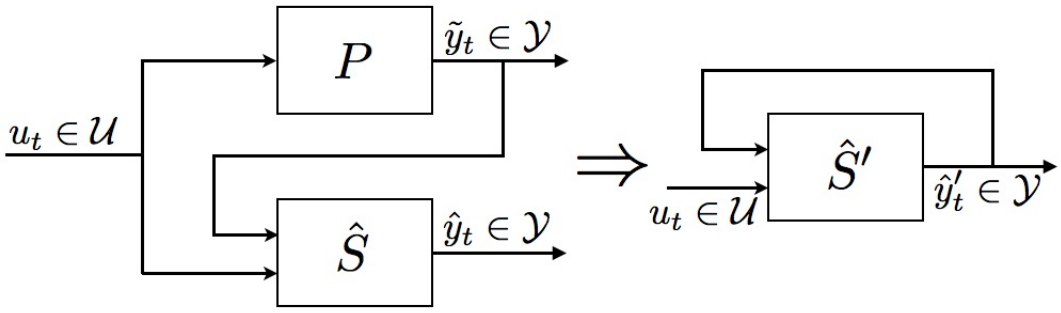


Figure 5: A copy of the observer for predictive control.

In the above figure, P represents system (126) which is finite memory observable. Then there is an observer $\hat{S}$ (5) such that for any $\mathbf{u} \in \mathcal{U}^{\mathbb{N}}$, $\hat{y}_t = \tilde{y}_t$ for all $t \ge T$. Note that for system (126), $\tilde{y}_t = Q(Cx_t)$, therefore we observe that the output $\hat{y}_t$ of $\hat{S}$ can be written as $\hat{y}_t = g(q_t)$ instead of $\hat{y}_t = g(q_t, u_t)$. Indeed, for any $t \ge T$, $\tilde{y}_t = \hat{y}_t = g(q_t, u_t)$ for any $u_t \in \mathcal{U}$ by assumption. Since $\tilde{y}_t$ does not depend on u_t, we see that $g(q_t, u) = g(q_t, u')$ for any $u, u' \in \mathcal{U}$. Therefore we let $\hat{y}_t = g'(q_t) = g(q_t, u)$ for some $u \in \mathcal{U}$, and this $\hat{y}_t$ also

satisfies $\tilde{y}_t = \hat{y}_t$ for all $t \geq T$. We conclude that there is an observer $\hat{S}$ described by

$$\begin{aligned} q_{t+1} &= f(q_t, u_t, \tilde{y}_t), \\ \hat{y}_t &= g(q_t), \end{aligned} \tag{144}$$

such that for any $\mathbf{u} \in \mathcal{U}^{\mathbb{N}}$, $\hat{y}_t = \tilde{y}_t$ for all $t \geq T$.

As shown in Figure 5, we define a "copy" $\hat{S}'$ of $\hat{S}$ described as

$$\begin{aligned} q'_{t+1} &= f(q'_t, u'_t, \hat{y}'_t),\ t \geq T \\ \hat{y}'_t &= g(q'_t),\ t \geq T, \\ q'_T &= q_T, \end{aligned} \tag{145}$$

where f, g are the same functions as in (144). We observe that

$$\text{If } u'_t = u_t \text{ for all } \ t \geq T, \text{then } q'_t = q_t,\ \hat{y}'_t = \tilde{y}_t, \text{ for all } \ t \geq T. \tag{146}$$

To see this, at $t = T$, $\hat{y}'_T = g(q'_T) = g(q_T) = \hat{y}_T$. By the definition of finite memory observability, $\hat{y}_T = \tilde{y}_T$. Therefore (146) holds for $t = T$. Assume (146) holds for some $t \geq T$, at $t+1$, $q'_{t+1} = f(q'_t, u'_t, \hat{y}'_t) = f(q_t, u_t, \tilde{y}_t) = f(q_t, u_t, \hat{y}_t) = q_{t+1}$, therefore $q'_{t+1} = q_{t+1}$. $\hat{y}'_{t+1} = g(q'_{t+1}) = g(q_{t+1}) = \hat{y}_{t+1} = \tilde{y}_{t+1}$, therefore (146) holds for $t+1$. By induction, (146) holds for all $t \geq T$.

Based on (146), we see that for any $k \in \mathbb{Z}_+$,

$$\text{If } u'_t = u_t \text{ for } \ T \leq t \leq T+k-1, \text{then } q'_t = q_t,\ \hat{y}'_t = \tilde{y}_t, \text{ for } \ T \leq t \leq T+k-1. \tag{147}$$

Recall (142), there is $(\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_k) \in \mathcal{U}^k$ such that if the input u_t of system (126) satisfies $u_{t+T} = \underline{u}_{t+1}$ for $0 \leq t \leq k-1$, then $\tilde{y}_{T+k} = \tilde{y}_e$. By (147), if $u'_{t+T} = u_{t+T} = \underline{u}_{t+1}$ for $0 \leq t \leq k-1$, then $q'_{t+T-1} = q_{t+T-1}, \hat{y}'_{t+T-1} = \tilde{y}_{t+T-1}$. Recall (144), (145), $q'_{T+k} =$

q_{T+k}. Consequently, we have

$$\hat{y}'_{T+k} = \hat{y}_{T+k} = \tilde{y}_{T+k} = \tilde{y}_e. \tag{148}$$

We conclude that there is $(\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_k) \in \mathcal{U}^k$ such that if the input u'_t of system (145) satisfies $u'_{t+T} = \underline{u}_{t+1}$ for $0 \leq t \leq k-1$, then $\hat{y}'_{T+k} = \tilde{y}_e$.

Next, we begin to determine the input u_t of system (126) for $T \leq t \leq T+k-1$. First, let u_t for $t = 0, 1, \ldots, T-1$ be chosen such that (141) is satisfied. At $t = T-1$, $q_T = f(q_{T-1}, u_{T-1}, \tilde{y}_{T-1})$ can be determined. Then q'_T is also determined. Given $\hat{S}'$ (145), calculate the output $\hat{y}'_t$ of $\hat{S}'$, when $\{u'_t\}_{t=T}^{T+k-1}$ assumes all possible values in $\mathcal{U}^k$. By the previous discussion, there is a set of input segments $\omega^1, \omega^2, \ldots, \omega^j \in \mathcal{U}^k$ ($1 \leq j \leq |\mathcal{U}|^k$) such that when $\{u'_t\}_{t=T}^{T+k-1} = \omega^i$ for any $1 \leq i \leq j$, $\hat{y}'_{T+k} = \tilde{y}_e$. Consequently, if $\{u_t\}_{t=T}^{T+k-1} = \omega^i$ for any $1 \leq i \leq j$, then $\tilde{y}_{T+k} = \tilde{y}_e$. Note that $\omega^1, \omega^2, \ldots, \omega^j$ can be computed at time $t = T-1$.

So far, we have formulated a method to determine $\{u_t\}_{t=T}^{T+k-1}$ at $t = T-1$ such that $\tilde{y}_{T+k} = \tilde{y}_e$.

However, for any ω^i, $1 \leq i \leq j$, when $\{u_t\}_{t=T}^{T+k-1} = \omega^i$, the requirement "$x_t \in \mathcal{X}$ for $t = T+1, \ldots, T+k$" may or may not be satisfied. Therefore, to address this requirement, we need to modify this method in the following.

Given system (126), recall (129), define

$$\mathcal{Y}_{\mathcal{X}} = \{\tilde{y} \in \mathcal{Y} : Q_C^{-1}(\tilde{y}) \subset \mathcal{X}\}. \tag{149}$$

If Assumption 1 holds, then $Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}$, and therefore $\mathcal{Y}_{\mathcal{X}}$ is non-empty.

For system (126), if $\tilde{y}_t \in \mathcal{Y}_{\mathcal{X}}$, then $x_t \in Q_C^{-1}(\tilde{y}_t) \subset \mathcal{X}$. Therefore, for any $t \in \mathbb{N}$,

$$\tilde{y}_t \in \mathcal{Y}_{\mathcal{X}} \Rightarrow x_t \in \mathcal{X}. \tag{150}$$

Next, define

$$\mathcal{X}' = \{x \in \mathcal{X} : Q_C^{-1}(Q(Cx)) \subset \mathcal{X}\}. \tag{151}$$

Then for any $x \in \mathcal{X}'$, $Q(Cx) \in \mathcal{Y}$, and $Q_C^{-1}(Q(Cx)) \subset \mathcal{X}$. Recall (149), we see that

$$x \in \mathcal{X}' \Rightarrow Q(Cx) \in \mathcal{Y}_{\mathcal{X}}. \tag{152}$$

If Assumption 1 holds, then $Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}$. For any $x \in Q_C^{-1}(\tilde{y}_e)$, $Q_C^{-1}(Q(Cx)) = Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}$. Therefore $x \in \mathcal{X}'$, and we see that $Q_C^{-1}(\tilde{y}_e) \in \mathcal{X}'$. Recall (136), we have

$$\mathcal{R}^{(0)} \subset \mathcal{X}'. \tag{153}$$

Similar to (137), define

$$\begin{aligned}
\mathcal{R}'^{(1)} &= \{x \in \mathcal{X}' : Ax + Bu \in \mathcal{R}^{(0)}, \text{ for some } u \in \mathcal{U}\}, \\
\mathcal{R}'^{(2)} &= \{x \in \mathcal{X}' : Ax + Bu \in \mathcal{R}'^{(1)}, \text{ for some } u \in \mathcal{U}\}, \\
&\vdots \\
\mathcal{R}'^{(k)} &= \{x \in \mathcal{X}' : Ax + Bu \in \mathcal{R}'^{(k-1)}, \text{ for some } u \in \mathcal{U}\}.
\end{aligned} \tag{154}$$

Essentially, in the above definition, we change "$\mathcal{X}$" in (137) to "$\mathcal{X}'$".

Correspondingly, define set $\mathcal{Y}_c' \subset \mathcal{Y}$ as: $\tilde{y} \in \mathcal{Y}_c'$ if there is $(u_0, u_1, \ldots, u_{T-1}) \in \mathcal{U}^T$ such that

$$\begin{aligned}
&Q_C^{-1}(\tilde{y}) \subset \mathcal{X}, \\
&AQ_C^{-1}(\tilde{y}) + Bu_0 \subset \mathcal{X}, \\
&A^2Q_C^{-1}(\tilde{y}) + Bu_1 + Bu_0 \subset \mathcal{X}, \\
&\quad\vdots \\
&A^{T-1}Q_C^{-1}(\tilde{y}) + \sum_{\tau=0}^{T-2} A^{T-2-\tau}Bu_\tau \subset \mathcal{X}, \\
&A^{T}Q_C^{-1}(\tilde{y}) + \sum_{\tau=0}^{T-1} A^{T-1-\tau}Bu_\tau \subset \mathcal{R}'^{(k)}.
\end{aligned} \tag{155}$$

And define

$$\mathcal{C}' = \bigcup_{\tilde{y} \in \mathcal{Y}'_c} Q_C^{-1}(\tilde{y}). \tag{156}$$

Recall (141), we see that if $x_0 \in \mathcal{C}'$, then there is $(\bar{u}_0, \bar{u}_1, \ldots, \bar{u}_{T-1}) \in \mathcal{U}^T$ such that when the input u_t of system (126) is $u_t = \bar{u}_t$ for $0 \leq t \leq T-1$, then

$$\begin{aligned} &x_t \in \mathcal{X}, \quad t = 0, 1, \ldots, T-1, \\ &x_T \in \mathcal{R}'^{(k)}. \end{aligned} \tag{157}$$

Recall (142), note that $\mathcal{R}'^{(j)} \subset \mathcal{X}'$ for $1 \leq j \leq k$, and recall (152), we see that there is $(\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_k) \in \mathcal{U}^k$ such that if the input u_t of system (126) satisfies $u_{t+T} = \underline{u}_{t+1}$ for $0 \leq t \leq k-1$, then

$$\begin{aligned} &\tilde{y}_t \in \mathcal{Y}_{\mathcal{X}}, \quad t = T+1, \ldots, T+k, \\ &\tilde{y}_{T+k} = \tilde{y}_e. \end{aligned} \tag{158}$$

Next, we determine the input u_t of system (126) for $T \leq t \leq T+k-1$. If $x_0 \in \mathcal{C}'$, let u_t for $t = 0, 1, \ldots, T-1$ be chosen such that (157) is satisfied. Consider $\hat{S}'$ (145), then at $t = T-1$, $q'_T = q_T$ is determined. By (158), (147), (148), there is a set of input segments $\omega^1, \omega^2, \ldots, \omega^{j'} \in \mathcal{U}^k$ ($1 \leq j' \leq |\mathcal{U}|^k$) such that if $\{u'_t\}_{t=T}^{T+k-1} = \omega^i$ for any $1 \leq i \leq j'$, then

$$\begin{aligned} &\hat{y}'_t \in \mathcal{Y}_{\mathcal{X}}, \quad t = T+1, \ldots, T+k, \\ &\hat{y}'_{T+k} = \tilde{y}_e. \end{aligned} \tag{159}$$

Again, note that $\omega^1, \omega^2, \ldots, \omega^{j'}$ can be computed at time $t = T-1$.

For any $i \in \{1, \ldots, j'\}$, let the input of system (126) be $\{u_t\}_{t=T}^{T+k-1} = \omega^i$, then by (147), (148), (159), we have

$$\begin{aligned} &\tilde{y}_t \in \mathcal{Y}_{\mathcal{X}}, \quad t = T+1, \ldots, T+k, \\ &\tilde{y}_{T+k} = \tilde{y}_e. \end{aligned}$$

Recall (150), we see that

$$\begin{aligned} x_t &\in \mathcal{X}, \quad t = T+1, \ldots, T+k, \\ \tilde{y}_{T+k} &= \tilde{y}_e. \end{aligned} \tag{160}$$

So far, we have computed the input u_t of system (126) for $t = T, \ldots, T+k-1$. For $0 \leq t \leq T-1$, given $\tilde{y}_0 \in \mathcal{Y}'_c$, recall (155), there is a set of input segments $\nu^1, \nu^2, \ldots, \nu^l \in \mathcal{U}^T$ ($1 \leq l \leq |\mathcal{U}|^T$) such that

$$\begin{aligned} x_t &\in \mathcal{X}, \quad t = 0, \ldots, T, \\ x_T &\in \mathcal{R}'^{(k)}. \end{aligned} \tag{161}$$

Note that $\nu^1, \nu^2, \ldots, \nu^l$ can be computed at $t = 0$.

For Problem 2, let $N = T + k$, for any $i_1 \in \{1, \ldots, l\}, i_2 \in \{1, \ldots, j'\}$, let

$$\{u_t\}_{t=0}^{T-1} = \nu^{i_1}, \{u_t\}_{t=T}^{T+k-1} = \omega^{i_2}, u_{T+k} = u_e, \tag{162}$$

recall (161), (160), we see that Problem 2 is solved for all $x_0 \in \mathcal{C}'$.

In addition, assume $Q_C^{-1}(\tilde{y}_e)$ bounded, $Q_C^{-1}(\tilde{y}_e) \in \mathcal{X}^\infty$, $x_e \in int(Q_C^{-1}(\tilde{y}_e))$, and $\rho(A) < 1$. In (155), let $(u_0, u_1, \ldots, u_{T-1}) = (u_e, \ldots, u_e)$, choose T large enough (note that finite memory observability holds for all T larger than the parameter "T" of system (126)), and realize that $0 \in int(\mathcal{R}^{(0)}) \subset int(\mathcal{R}'^{(k)})$, we see that $\tilde{y}_e \in \mathcal{Y}'_c$.

We conclude the control design based on finite memory observability as follows:

Theorem 10. Given system (126) and its equilibrium x_e, assume $\rho(A) < 1$, $\mathcal{X}$ is bounded, $Q_C^{-1}(\tilde{y}_e) \subset \mathcal{X}^\infty$, $x_e \in int(Q_C^{-1}(\tilde{y}_e))$, and system (126) is (C1). Then for all $x_0 \in \mathcal{C}'$, where $\mathcal{C}' \neq \varnothing$, Problem 1 has a solution, and it can be computed according to the procedure stated in this section.

Remark. The input segments $\omega^1, \ldots \omega^{j'}$ and $\nu^1, \ldots, \nu^l$ can be computed offline: Compute $\nu^1, \ldots, \nu^l$ for all $\tilde{y} \in \mathcal{Y}$, and $\omega^1, \ldots \omega^{j'}$ for all $q \in \mathcal{Q}$ according to the dynamics of $\hat{S}'$,

where $\mathcal{Q}$ is the state space of $\hat{S}'$. Then the online controller can simply use a table-lookup.

The cost optimization of Problem 1 can be implemented by choosing an appropriate $(i_1, i_2) \in \{1, \ldots, l\} \times \{1, \ldots, j'\}$ that minimize a reformulated cost function, which corresponds to the quantized output $\tilde{y}_t$ instead of x_t.

Remark. A conservatism of this approach is that the designed control input u_t for $0 \leq t \leq T-1$ is determined only by the initial observation $\tilde{y}_0$. This corresponds to a static feedback law. A less conservative approach could be that for $0 \leq t \leq T-1$, u_t is determined by $(\tilde{y}_0, \tilde{y}_1, \ldots, \tilde{y}_t)$.

Intended to be blank.

6 On Initialization of DFM Approximations

In this chapter, we continue to study DFM approximations for systems over finite alphabets. In particular, we focus on improving an input-output construction of finite state ρ/μ approximation proposed in [12], by seeking to build such approximations using fewer states. We propose a necessary and sufficient condition for simplifying the initialization process of the input-output construction of ρ/μ approximations without incurring a loss in performance. We give an alternative characterization of this necessary and sufficient condition for a specific class of systems with linear internal dynamics. For instances where this necessary and sufficient condition is not satisfied, we present an alternate initialization process leading to approximations with fewer states than those resulting from the existing construction in [12]. The work presented in this section has been previously reported in [29].

6.1 Preliminaries: Finite State ρ/μ Approximations

6.1.1 Existing Input/Output Construction

Consider a discrete-time system P described by:

$$x_{t+1} = f(x_t, u_t), \tag{163a}$$

$$y_t = g(x_t), \tag{163b}$$

where $t \in \mathbb{N}$ is the time index, $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathcal{U}$ is the input, $y_t \in \mathcal{Y}$ is the output, and f, g are functions $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathcal{Y}$. We enforce that the input u_t and the output y_t take finitely many values in sets $\mathcal{U}$ and $\mathcal{Y}$ respectively. Therefore P is a "system over finite alphabets" as defined in [7].

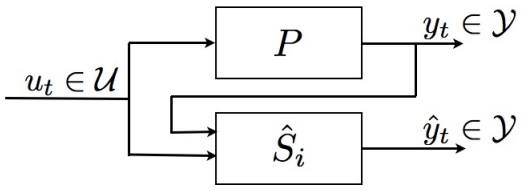


Figure 6: Finite state approximation.

In Figure 6, P represents the system to be approximated, and $\hat{S}_i$ represents the finite state ρ/μ approximation of P. Given a plant P, a sequence of approximations $\{\hat{S}_i\}_{i=1}^{\infty}$ can be constructed, and $\hat{S}_i$ is the i^{th} element of this sequence.

$\hat{S}_i$ is a finite state ρ/μ approximation of system (163), which is described by:

$$q_{t+1} = \phi(q_t, u_t, y_t) \tag{164a}$$

$$\hat{y}_t = \theta(q_t) \tag{164b}$$

where $t \in \mathbb{N}$, $q_t \in \mathcal{Q}$ for some finite set $\mathcal{Q}$, $u_t \in \mathcal{U}$, $y_t \in \mathcal{Y}$, $\hat{y}_t \in \mathcal{Y}$, and ϕ, θ are functions $\phi : \mathcal{Q} \times \mathcal{U} \times \mathcal{Y} \to \mathcal{Q}$ and $\theta : \mathcal{Q} \times \mathcal{U} \to \mathcal{Y}$. $i \in \mathbb{Z}_+$ is a parameter of $\hat{S}_i$. As shown in Figure 6, $\hat{S}_i$ has for its inputs both the input u_t and the output y_t of system (163), and generates $\hat{y}_t$ as an estimate of y_t. We enforce that there is no direct feedthrough from y_t to $\hat{y}_t$ (see (164b)), therefore $\hat{y}_t$ and y_t cannot be trivially matched. We require the state-space $\mathcal{Q}$ of $\hat{S}_i$ be of finite cardinality.

In this section, we will describe the finite state-space $\mathcal{Q}$ and the transition function ϕ. We refer to [12] for more details such as the construction of function θ.

In order to describe $\mathcal{Q}$, we first introduce the following definition:

Definition 5. Given system (163), use $f_u(x)$ to denote $f(x, u)$. For any $j \in \mathbb{Z}_+$, let $q = (y_1, y_2 \dots y_j, u_1, u_2 \dots u_j) \in \mathcal{Y}^j \times \mathcal{U}^j$. We say that q is *feasible* if there exists $x \in \mathbb{R}^n$ such

that

$$\begin{aligned} y_j &= g(x) \\ y_{j-1} &= g(f_{u_j}(x)) \\ y_{j-2} &= g(f_{u_{j-1}} \circ f_{u_j}(x)) \\ &\vdots \\ y_1 &= g(f_{u_2} \circ f_{u_3} \circ \cdots \circ f_{u_{j-1}} \circ f_{u_j}(x)) \end{aligned} \tag{165}$$

are satisfied.

Essentially, q is feasible if and only if q consists of segments of some input and output sequences of system (163), namely $q = (y_{t-1}, y_{t-2} \dots y_{t-j}, u_{t-1}, u_{t-2} \dots u_{t-j})$ for some input sequence $\{u_t\}$ and output sequence $\{y_t\}$ of (163).

Remark. For $j = 1$, every element q in $\mathcal{Y} \times \mathcal{U}$ is feasible.

Now we are ready to describe the state-space $\mathcal{Q}$ of $\hat{S}_i$.

State Set: For the setup in Figure 6, the finite state-space $\mathcal{Q}$ of $\hat{S}_i$ is defined as:

$$\mathcal{Q} = \mathcal{Q}_F \cup \mathcal{Q}_I \cup \{q_o, q_\varnothing\} \tag{166}$$

where

$$\mathcal{Q}_F = \{q \in \mathcal{Y}^i \times \mathcal{U}^i | q \text{ is feasible}\} \tag{167}$$

and

$$\mathcal{Q}_I = \bigcup_{j=1}^{i-1} \{q \in \mathcal{Y}^j \times \mathcal{U}^j | q \text{ is feasible}\} \tag{168}$$

and $q_o, q_\varnothing$ are two symbolic elements.

Next, we define the state transition function ϕ of $\hat{S}_i$.

State Transition Function: For the setup in Figure 6, the function $\phi : \mathcal{Q} \times \mathcal{Y} \times \mathcal{U} \to \mathcal{Q}$ is defined as:

For any $q \in \mathcal{Q}, y \in \mathcal{Y}, u \in \mathcal{U}$,

$\triangleright$ If $q = q_o$, then

$$\phi(q, y, u) = (y, u). \tag{169}$$

$\triangleright$ If $q = q_\varnothing$, then

$$\phi(q, y, u) = q_\varnothing. \tag{170}$$

$\triangleright$ If $q \in \mathcal{Q}_I$, write $q = (y_1, y_2 \dots y_j, u_1, u_2 \dots u_j)$, where $1 \leq j \leq i-1$, and let $\bar{q} = (y, y_1, y_2 \dots y_j, u, u_1, u_2 \dots u_j)$, then

$$\phi(q, y, u) = \begin{cases} \bar{q}, & \text{if } \bar{q} \in \mathcal{Q}_I \cup \mathcal{Q}_F \\ q_\varnothing, & \text{otherwise} \end{cases} \tag{171}$$

$\triangleright$ If $q \in \mathcal{Q}_F$, write $q = (y_1, y_2 \dots y_i, u_1, u_2 \dots u_i)$, and let $\bar{q} = (y, y_1, y_2 \dots y_{i-1}, u, u_1, u_2 \dots u_{i-1})$, then

$$\phi(q, y, u) = \begin{cases} \bar{q}, & \text{if } \bar{q} \in \mathcal{Q}_F \\ q_\varnothing, & \text{otherwise} \end{cases} \tag{172}$$

Lastly, we require that the initial state of $\hat{S}_i$ be fixed.

Fixed Initial Condition: For the setup in Figure 6, at $t = 0$ the initial state q_0 of $\hat{S}_i$ satisfies:

$$q_0 = q_o.$$

Up to this point, we have defined the state space and the state transitions of $\hat{S}_i$. As a result, given any input sequence $\{u_t\}$ and output sequence $\{y_t\}$ of system (163), we are able to determine the state q_t of $\hat{S}_i$ for all $t \in \mathbb{N}$.

6.1.2 Initialization

We next describe the initialization process of $\hat{S}_i$: At $t = 0$, $\hat{S}_i$ starts at a fixed initial state q_o; for $t = 1, 2, \dots, i-1$, $\hat{S}_i$ adds the input u_t and the output y_t into its state at each time increment. After i steps, the initialization process is complete and the state q_i of $\hat{S}_i$ corresponds to the previous i steps of the input and output of the plant P, namely $q_i = (y_{i-1}, y_{i-2} \dots y_0, u_{i-1}, u_{i-2} \dots u_0)$.

One issue with this initialization process is that it requires additional states be added to the state-space $\mathcal{Q}$ of $\hat{S}_i$. Specifically, $\mathcal{Q}_I$ in (166) will never be visited after the initialization process: $q_t \notin \mathcal{Q}_I$ for all $t \geq i$. In an application of ρ/μ approximation to a water tank model [30], the size of $\mathcal{Q}_I$ is roughly a third of the size of $\mathcal{Q}$, and is thus significant. Since the number of states, or the "memory" of the approximation, directly impacts the complexity of the associated control design procedure as well as the implementation of the resulting controller, it is imperative to keep the size of $\mathcal{Q}$ as small as possible.

6.2 Problem Statement

We notice that a straightforward approach to potentially reduce the size of $\mathcal{Q}$ is as follows: Initialize $\hat{S}_i$ at an arbitrary state q_0 in $\mathcal{Q}_F$. If after i steps, the state q_i corresponds to the sequence $(y_{i-1}, y_{i-2} \dots y_0, u_{i-1}, u_{i-2} \dots u_0)$, then we can remove $\mathcal{Q}_I$ from $\mathcal{Q}$ and still achieve the same approximation quality for $t \geq i$. We want to identify the instances of system (163) this straightforward approach will work.

Before posing the problem of interest, we make the following observation:

Observation 2. Given a system P (163) and its finite state ρ/μ approximation $\hat{S}_i$ as constructed in Section 6.1. For any input sequence $\mathbf{u}$, any $x_0 \in \mathbb{R}^n$ and any $q_0 \in \mathcal{Q}$, if $q_i \in \mathcal{Q}_F$, then $q_i = (y_{i-1}, y_{i-2} \dots y_0, u_{i-1}, u_{i-2} \dots u_0)$, where u_k and y_k are the input and output of P at time $t = k, 0 \leq k \leq i-1$.

To see that Observation 2 holds, first note that $q_i \in \mathcal{Q}_F$ implies that $q_{i-1} \in \mathcal{Q}_F \cup \mathcal{Q}_I$. Consequently $q_k \in \mathcal{Q}_F \cup \mathcal{Q}_I$ for all $1 \leq k \leq i$. From the definition of the transition function ϕ, we see that the pair (y_k, u_k) is stored in q_{k+1} for all $0 \leq k \leq i-1$. We also note that during the first i time increments, (y_k, u_k) get shifted at most $i-1$ times within $\{q_k\}_{k=1}^{i}$. Therefore $q_i = (y_{i-1}, y_{i-2} \dots y_0, u_{i-1}, u_{i-2} \dots u_0)$.

With the straightforward approach to reduce $\mathcal{Q}$ and Observation 2 in mind, we formulate the problem of interest as follows:

Problem 3. Given system (163) and its finite state ρ/μ approximation $\hat{S}_i$, constructed as

described in Section 6.1. Consider the statement:

$$q_0 \in \mathcal{Q}_F \Rightarrow q_i \in \mathcal{Q}_F, \tag{173}$$

under what conditions on system (163) does (173) hold for any input sequence $\mathbf{u} \in \mathcal{U}^{\mathbb{Z}_+}$ and any $x_0 \in \mathbb{R}^n$?

Remark. We will first discuss Problem 3 in the general context of system (163), and we will propose a set of specific results for system (2) with $m = p = 1$.

6.3 Conditions for Simplifying the Initialization Process

In this section we state our technical results regarding Problem 3. First, we propose a necessary and sufficient condition for system (163) such that (173) holds.

Lemma 4. Given a system (163) and its finite state ρ/μ approximation $\hat{S}_i$ as constructed in Section 6.1. (173) holds for any input sequence $\mathbf{u} \in \mathcal{U}^{\mathbb{Z}_+}$ and any $x_0 \in \mathbb{R}^n$ if and only if $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$.

Proof. For the forward implication, assume that $q_0 \in \mathcal{Q}_F$ implies $q_i \in \mathcal{Q}_F$ for any input sequence $\mathbf{u}$ and any x_0. Then particularly $q_1 \neq q_\varnothing$, otherwise $q_i = q_\varnothing$ and $q_\varnothing \notin \mathcal{Q}_F$. Since $q_1 = \phi(q_0, y_0, u_0)$. Let $y_0 = y$, $u_0 = u$, and write $q_0 = (y_1, y_2 \dots y_i, u_1, u_2 \dots u_i)$, and let $\bar{q} = (y, y_1, y_2 \dots y_{i-1}, u, u_1, u_2 \dots u_{i-1})$, then $\bar{q} \in \mathcal{Q}_F$ holds for any $q_0 \in \mathcal{Q}_F$, any $y \in \mathcal{Y}$ and $u \in \mathcal{U}$. Consider any $q \in \mathcal{Y}^i \times \mathcal{U}^i$, write $q = (\tilde{y}_1, \tilde{y}_2 \dots \tilde{y}_i, \tilde{u}_1, \tilde{u}_2 \dots \tilde{u}_i)$, then for any $q_0 = (y_1, y_2 \dots y_i, u_1, u_2 \dots u_i)$.

Let $y_0 = \tilde{y}_i$, $u_0 = \tilde{u}_i$, then $(\tilde{y}_i, y_1, y_2 \dots y_{i-1}, \tilde{u}_i, u_1, u_2 \dots u_{i-1}) \in \mathcal{Q}_F$. Next, let $y_0 = \tilde{y}_{i-1}$, $u_0 = \tilde{u}_{i-1}$, and let $q_0 = (\tilde{y}_i, y_1, y_2 \dots y_{i-1}, \tilde{u}_i, u_1, u_2 \dots u_{i-1}) \in \mathcal{Q}_F$, then $(\tilde{y}_{i-1}, \tilde{y}_i, y_1, y_2 \dots y_{i-2}, \tilde{u}_{i-1}, \tilde{u}_i, u_1, u_2 \dots u_{i-2}) \in \mathcal{Q}_F$. Repeat this argument i times, we have $q \in \mathcal{Q}_F$. Since q is arbitrary, $\mathcal{Y}^i \times \mathcal{U}^i \subset \mathcal{Q}_F$. By the definition of $\mathcal{Q}_F$ (167), we conclude $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$.

For the backward implication, assume that $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$. For any $q_0 \in \mathcal{Q}_F$, $y_0 \in \mathcal{Y}$ and $u_0 \in \mathcal{U}$, by (172) we see that $q_1 \in \mathcal{Q}_F$. Assume that $q_k \in \mathcal{Q}_F$ for some $1 \leq k \leq i-1$,

then for any $y_k \in \mathcal{Y}$ and $u_k \in \mathcal{U}$, $\bar{q} \in \mathcal{Y}^i \times \mathcal{U}^i = \mathcal{Q}_F$. Therefore $q_{k+1} \in \mathcal{Q}_F$. By mathematical induction, we have $q_i \in \mathcal{Q}_F$.

□

Next, we propose a sufficient condition, and a necessary condition stated in terms of the properties and parameters of system (2) with $m = p = 1$.

Theorem 11. Given a system (2) with $m = p = 1$ and its finite state ρ/μ approximation $\hat{S}_i$ as constructed in Section 6.1. If the pair (C, A) is observable and $i \leq n$, then (173) holds for any input sequence $\mathbf{u} \in \mathcal{U}^{\mathbb{Z}_+}$ and any $x_0 \in \mathbb{R}^n$.

Proof. We first consider the case $i = n$. We will show that every element in $\mathcal{Y}^n \times \mathcal{U}^n$ is feasible.

For any $q \in \mathcal{Y}^n \times \mathcal{U}^n$, write $q = (y_1, y_2 \dots y_n, u_1, u_2 \dots u_n)$. By the axiom of choice (pp.26, [31]), there exist $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n \in \mathbb{R}$ such that

$$Q(\bar{y}_k) = y_k, \text{ for all } 1 \leq k \leq n. \tag{174}$$

Let $\bar{q} = (\bar{y}_1, \bar{y}_2 \dots \bar{y}_n, u_1, u_2 \dots u_n) \in \mathbb{R}^{2n}$, define $x \in \mathbb{R}^n$ as:

$$x = \Theta^{-1}[I \ \ -M]\bar{q} \tag{175}$$

where the matrix Θ is defined as:

$$\Theta = \begin{bmatrix} CA^{n-1} \\ CA^{n-2} \\ \vdots \\ C \end{bmatrix} \tag{176}$$

and the upper-triangular matrix M is defined as:

$$M = \begin{bmatrix} D & CB & CAB & \cdots & CA^{n-2}B \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & CAB \\ & & & \ddots & CB \\ & & & & D \end{bmatrix} \tag{177}$$

We see that Θ is invertible by the assumption (C, A) being observable.

Within the scope of this derivation of Theorem 11, we use "u_t'" to denote "u_t" , "y_t'" to denote "y_t", and "$\tilde{y}_t'$" to denote "$\tilde{y}_t$" for system (2) in order to avoid overwriting notations with that of q used.

From the solution of LTI systems, we have:

$$y_t' = CA^t x_0 + \sum_{\tau=0}^{t-1} CA^{t-1-\tau} B u_\tau' + D u_t' \tag{178}$$

Let $t = 0, 1, \ldots, n-1$, and stack the above equation, we have:

$$\begin{bmatrix} y_{n-1}' \\ \vdots \\ y_0' \end{bmatrix} = \Theta x_0 + M \begin{bmatrix} u_{n-1}' \\ \vdots \\ u_0' \end{bmatrix} \tag{179}$$

Let $x_0 = x$, where x is defined in (175), and let $(u_{n-1}', u_{n-2}' \ldots u_0') = (u_1, u_2, \ldots u_n)$. Then the preceding equation becomes:

$$\begin{aligned} \begin{bmatrix} y_{n-1}' \\ \vdots \\ y_0' \end{bmatrix} &= \Theta\Theta^{-1}[I \;\; -M]\bar{q} + M[\mathbf{0}_{n\times n} \;\; I]\bar{q} \\ &= [I \;\; \mathbf{0}_{n\times n}]\bar{q} \end{aligned} \tag{180}$$

Or equivalently, $(y'_{n-1}, y'_{n-2} \dots y'_0) = (\bar{y}_1, \bar{y}_2 \dots \bar{y}_n)$.

By (174) and (2c), we have $(\tilde{y}'_{n-1}, \tilde{y}'_{n-2} \dots \tilde{y}'_0) = (y_1, y_2 \dots y_n)$. Comparing this with the definition of feasibility in (165), we see that $q \in \mathcal{Y}^n \times \mathcal{U}^n$ is feasible with the choice of x defined in (175). Since q is chosen arbitrarily, we conclude that $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$.

Next we show that for $i < n$, $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$ also holds. For any $q \in \mathcal{Y}^i \times \mathcal{U}^i$, write $q = (y_1, y_2 \dots y_i, u_1, u_2 \dots u_i)$, and choose any $y \in \mathcal{Y}$ and $u \in \mathcal{U}$. Let q' be an element in $\mathcal{Y}^n \times \mathcal{U}^n$ and $q' = (y, y \dots y, y_1, y_2 \dots y_i, u, u \dots u, u_1, u_2 \dots u_i)$. Then q' is feasible since $\mathcal{Q}_F = \mathcal{Y}^n \times \mathcal{U}^n$. Particularly, the first i equations in (165) holds, therefore q is feasible.

Finally, we conclude that $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$ for all $i \leq n$. By Lemma 4, $q_0 \in \mathcal{Q}_F$ implies $q_i \in \mathcal{Q}_F$ for any input sequence $\mathbf{u}$ and any $x_0 \in \mathbb{R}^n$. □

Next, we propose a necessary condition for simplifying the initialization process for system (2).

Theorem 12. Given a system (2) with $m = p = 1$, $0 \in \mathcal{U}$, $|\mathcal{Y}| > 1$, and its finite state ρ/μ approximation $\hat{S}_i$ as constructed in Section 6.1. If (173) holds for any input sequence $\mathbf{u} \in \mathcal{U}^{\mathbb{Z}_+}$ and any $x_0 \in \mathbb{R}^n$, then $i \leq n$.

Remark. In Lemma 4, $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$ requires that *every* sequence of length $2i$ in $\mathcal{Y}^i \times \mathcal{U}^i$ be feasible, which can be quite restrictive in general, particularly when i is large.

In order to prove Theorem 12, we first derive the following Lemma, which will be instrumental in the derivation of Theorem 12.

Lemma 5. For any collection of vectors in $\mathbb{R}^n$: $\{v_i\}_{i=1}^{2^n}, v_i \in \mathbb{R}^n$, if for any $(s_1, s_2, \dots, s_n) \in \{-1, 1\}^n$, there exists $v_i = [v_i^1 \;\; v_i^2 \;\; \dots \;\; v_i^n]^T$, such that for all $1 \leq k \leq n$,

$$\begin{cases} v_i^k < 0, & \text{if } \; s_k = -1, \\ v_i^k > 0, & \text{if } \; s_k = 1, \end{cases} \tag{181}$$

then $span(\{v_i\}_{i=1}^{2^n}) = \mathbb{R}^n$.

Proof. (Lemma 5)

Clearly $span(\{v_i\}_{i=1}^{2^n}) \subset \mathbb{R}^n$, therefore we only need to show $\mathbb{R}^n \subset span(\{v_i\}_{i=1}^{2^n})$.

We first show that $e_1 = [1 \ 0 \ \ldots \ 0]^T$ is in $span(\{v_i\}_{i=1}^{2^n})$. By assumption, there exists $v_i = [v_i^1 \ v_i^2 \ \ldots \ v_i^n]^T$ and $v_j = [v_j^1 \ v_j^2 \ \ldots \ v_j^n]^T$ such that

$$v_i^k > 0, \ \text{for all} \ 1 \leq k \leq n$$

and

$$v_j^k > 0, \ \text{for all} \ 1 \leq k \leq n-1$$
$$v_j^n < 0$$

Let $w_1 = v_i^n \cdot v_j + (-v_j^n) \cdot v_i$, then $w_1 \in span(\{v_i\}_{i=1}^{2^n})$. Write $w_1 = [w_1^1 \ w_1^2 \ \ldots \ w_1^n]^T$, then

$$w_1^k > 0, \ \text{for all} \ 1 \leq k \leq n-1$$
$$w_1^n = 0$$

By a similar argument, there exists $w_2 \in span(\{v_i\}_{i=1}^{2^n})$ that satisfies:

$$w_2^k > 0, \ \text{for all} \ 1 \leq k \leq n-2$$
$$w_2^{n-1} < 0,$$
$$w_2^n = 0$$

where $w_2 = [w_2^1 \ w_2^2 \ \ldots \ w_2^n]^T$. Next we define $w_3 = (-w_2^{n-1}) \cdot w_1 + w_1^{n-1} \cdot w_2$. Write $w_3 = [w_3^1 \ w_3^2 \ \ldots \ w_3^n]^T$, then

$$w_3^k > 0, \ \text{for all} \ 1 \leq k \leq n-2$$
$$w_3^{n-1} = w_3^n = 0$$

Repeat this process until we get $[w \ \ 0 \ldots \ \ 0] \in span(\{v_i\}_{i=1}^{2^n})$ with $w > 0$. Therefore $e_1 = [1 \ 0 \ \ldots \ 0]^T \in span(\{v_i\}_{i=1}^{2^n})$.

By re-ordering, we can repeat the above process for any $e_i = [0 \ \ldots 0 \ 1 \ \ldots 0]^T$, where $\{e_i\}_{i=1}^n$ is the standard basis of $\mathbb{R}^n$. We conclude that $e_i \in span(\{v_i\}_{i=1}^{2^n})$ for all $1 \leq i \leq n$, and $\mathbb{R}^n = span(\{v_i\}_{i=1}^{2^n})$. $\square$

Now we are ready to show Theorem 12.

Proof. (of Theorem 12)

We will prove by contradiction. Particularly, if we can show that $\mathcal{Q}_F \neq \mathcal{Y}^i \times \mathcal{U}^i$ for all $i > n$, then by Lemma 4, we see that Theorem 12 holds.

Since $q_0 \in \mathcal{Q}_F$ implies $q_i \in \mathcal{Q}_F$ for any input sequence $\mathbf{u}$ and any $x_0 \in \mathbb{R}^n$, by Lemma 4, $\mathcal{Q}_F = \mathcal{Y}^i \times \mathcal{U}^i$. We assume $i \geq n+1$, any $q \in \mathcal{Y}^n \times \mathcal{U}^n$ and any $q' \in \mathcal{Y}^{n+1} \times \mathcal{U}^{n+1}$ are feasible (this indication is discussed in the proof of Theorem 11).

First, if $CA^n = 0_{1\times n}$, then there exist $q \in \mathcal{Y}^{n+1} \times \mathcal{U}^{n+1}$ is not feasible (consider the input is identically zero for $t \in \mathbb{N}$ and investigate the output at $t = n$). Therefore, we only consider the case where $CA^n \neq 0_{1\times n}$ for the remainder of the derivation of Theorem 12.

Let $b \in \mathbb{R}$ be a point where the quantizer $Q(\cdot)$ is discontinous. The existence of b is guaranteed by $|\mathcal{Y}| > 1$. Next, choose $\alpha, \beta \in \mathcal{Y}$ such that $\sup\{Q^{-1}(\alpha)\} \leq b$ and $\inf\{Q^{-1}(\beta)\} \geq b$. Since any $q \in \mathcal{Y}^n \times \mathcal{U}^n$ is feasible, then any $q \in \{\alpha, \beta\}^n \times \{0\}^n$ is also feasible. By the definition of feasibility (165), the dynamics of system (2) and the input and output relations expressed in (179), we conclude that for any $(s_1, s_2, \ldots, s_n) \in \{-1, 1\}^n$, there exists $x \in \mathbb{R}^n$ such that for all $1 \leq k \leq n$,

$$\begin{cases} CA^{k-1}x < b, & \text{if } s_k = -1, \\ CA^{k-1}x \geq b, & \text{if } s_k = 1. \end{cases} \tag{182}$$

In the above equation, we use "$\geq$" because $Q(\cdot)$ is right-continuous. In order to invoke Lemma 5, we need to replace "$\geq$" with "$>$". Given any $(s_1, s_2, \ldots, s_n) \in \{-1, 1\}^n$, let $x \in \mathbb{R}^n$ be the vector that satisfies (182), then define index set $\mathcal{I}$ as:

$$CA^{k-1}x = b, \quad \text{for all } k \in \mathcal{I} \tag{183}$$

If $\mathcal{I}$ is nonempty and $b \neq 0$, let

$$x' = (1 + \epsilon\frac{b}{|b|})x \tag{184}$$

Then we can choose $\epsilon > 0$ and ϵ sufficiently small, such that:

$$CA^{k-1}x' > b, \text{ for all } k \in \mathcal{I}, \tag{185}$$

and

$$\begin{cases} CA^{k-1}x < b, & \text{if } s_k = -1, \\ CA^{k-1}x > b, & \text{if } s_k = 1, \end{cases} \quad \text{for all } k \in \{1, \ldots, n\} \setminus \mathcal{I}. \tag{186}$$

If $\mathcal{I}$ is nonempty and $b = 0$, by (182), there exist $\bar{x} \in \mathbb{R}^n$ such that

$$CA^{k-1}\bar{x} < b, \text{ for all } k \in \mathcal{I}$$

Let

$$x' = x + (-\epsilon)\bar{x},$$

then we can choose $\epsilon > 0$ and ϵ sufficiently small, such that (185) and (186) are also satisfied.

We conclude that: For any $(s_1, s_2, \ldots, s_n) \in \{-1, 1\}^n$, there exists $x \in \mathbb{R}^n$ such that

$$\begin{cases} CA^{k-1}x < b, & \text{if } s_k = -1, \\ CA^{k-1}x > b, & \text{if } s_k = 1. \end{cases} \tag{187}$$

Next we divide this derivation into two cases.

Case A *(C,A) Observable*

If the pair (C,A) is observable, define $x^* \in \mathbb{R}^n$ as:

$$x^* = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}^{-1} \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix} \tag{188}$$

then

$$CA^{k-1}x^* = b, \text{ for all } 1 \leq k \leq n \tag{189}$$

Since $|\{<, \geq\}^n| = |\{1, \ldots, 2^n\}|$, let function

$$h : \{1, \ldots, 2^n\} \to \{<, \geq\}^n \tag{190}$$

be a bijection. Next define a collection of sets $\{S_j\}_{j=1}^{2^n}$ as: For any $j \in \{1, \ldots, 2^n\}$, write $h(j) = (s_1, s_2, \ldots, s_n)$, then

$$S_j = \{x \in \mathbb{R}^n : CA^{k-1}x \quad s_k \quad b, \text{for all } 1 \leq k \leq n\} \tag{191}$$

We claim that

$$S_j^\circ \neq \varnothing, \text{ for all } 1 \leq j \leq 2^n. \tag{192}$$

To see this, by (187), it suffice to show that x (187) is in the interior of the corresponding set S_j. Particularly, for any S_j, there exist $x \in S_j$ and x satisfy (187). Then either $CA^k x < b$ or $CA^k x > b$ for any $0 \leq k \leq n-1$. Let $\delta_k = |CA^k x - b|$, then $\delta_k > 0$. Let

$$\epsilon_k = \frac{\delta_k}{2\|CA^k\|} \tag{193}$$

Then for any $y \in B_{\epsilon^k}(x)$,

$$\begin{aligned} CA^k y - b &= CA^k(y - x + x) - b \\ &= CA^k(y - x) + CA^k x - b \end{aligned} \tag{194}$$

Since

$$\begin{aligned} |CA^k(y - x)| &\leq \|CA^k\|\|y - x\| \\ &< \|CA^k\|\epsilon_k \\ &= \delta_k/2 \end{aligned} \tag{195}$$

By (194) and (195), we see that $CA^k y - b$ and $CA^k x - b$ have the same sign. Let

$$\epsilon = \min\{\epsilon_k : 0 \leq k \leq n-1\} \tag{196}$$

then $B_\epsilon(x) \subset S_j$, therefore $x \in S_j^\circ$ and $S_j^\circ \neq \varnothing$.

Next, we claim that there exist $j_+, j_- \in \{1, \ldots, 2^n\}$ such that

$$\begin{aligned} CA^n x > CA^n x^*, \ \text{ for all } \ x \in S_{j_+}^\circ \\ \text{and } \ CA^n x < CA^n x^*, \ \text{ for all } \ x \in S_{j_-}^\circ \end{aligned} \tag{197}$$

We show this by contradiction: Assume that $\forall j \in \{1, \ldots, 2^n\}$, there exist $z_j \in S_j^\circ$ such that

$$CA^n z_j = CA^n x^* \tag{198}$$

Notice that for all j, $CA^k z_j \neq b$ for all $0 \leq k \leq n-1$. Otherwise, for any $\epsilon > 0$, we can find two points $u, v \in B_\epsilon(z_j)$ such that:

$$CA^k u < b \ \text{ and } \ CA^k v > b \tag{199}$$

therefore z_j is a boundary point of S_j, which contradicts with $z_j \in S_j^\circ$.

Next, we define a collection of vectors $\{v_j\}_{j=1}^{2^n}$:

$$v_j = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} (z_j - x^*) \tag{200}$$

then $\{v_j\}_{j=1}^{2^n}$ satisfy (181), by Lemma 5, $span(\{v_j\}_{j=1}^{2^n}) = \mathbb{R}^n$. For any $z \in \mathbb{R}^n$, then exist

coefficients $(\alpha_1, \ldots, \alpha_{2^n}) \in \mathbb{R}^{2^n}$ such that:

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} z = \sum_{j=1}^{2^n} \alpha_j v_j$$

this indicates

$$z = \sum_{j=1}^{2^n} \alpha_j (z_j - x^*)$$

therefore $span(\{z_j - x^*\}_{j=1}^{2^n}) = \mathbb{R}^n$. For any $z \in \mathbb{R}^n$, from the above equation and (198)

$$CA^n z = \sum_{j=1}^{2^n} \alpha_j (CA^n z_j - CA^n x^*) = 0$$

This contradicts with the condition that $CA^n \neq 0_{1\times n}$ as stated in the beginning of this derivation. Therefore the assumption (198) is false, and we conclude: There exist $j \in \{1, \ldots, 2^n\}$, such that for all $x \in S_j^\circ$

$$CA^n x \neq CA^n x^* \tag{201}$$

We observe that if $CA^n x > CA^n x^*$ for some $x \in S_j^\circ$, then $CA^n x > CA^n x^*$ for all $x \in S_j^\circ$. To see this, assume $CA^n y < CA^n x^*$ for some $y \in S_j^\circ$, then we can show that S_j° is a convex set, therefore the line segment connecting x and y lies in S_j°. Since the line segment is compact, and the function $l(x) = CA^n x$ is continuous, there exist $z \in S_j^\circ$ such that $CA^n z = CA^n x^*$, which contradicts with (201).

From (201), for the particular j, we see that there exist $\bar{j} \in \{1, \ldots, 2^n\}$ such that:

$$S_{\bar{j}}^\circ = \{x \in \mathbb{R}^n : x = 2x^* - y, y \in S_j^\circ\} \tag{202}$$

Finally we observe that $\{j, \bar{j}\} = \{j_+, j_-\}$, where j_+, j_- are defined in (197), and conse-

quently the claim (197) holds.

Next, we claim that there exist $j_+, j_- \in \{1, \ldots, 2^n\}$ such that

$$\begin{aligned} CA^n x \geq CA^n x^*, \ \text{ for all } \ x \in S_{j_+} \\ \text{and } \ CA^n x \leq CA^n x^*, \ \text{ for all } \ x \in S_{j_-} \end{aligned} \tag{203}$$

Comparing (197) and (203), we see that it suffice to show: For any $j \in \{1, \ldots, 2^n\}$, any point in $x \in S_j$ can be approached by a sequence of points in S_j°, namely there exist $\{x_k\}_{k=1}^\infty$ such that

$$\lim_{k \to \infty} x_k = x, \text{and } \ \{x_k\}_{k=1}^\infty \subset S_j^\circ \tag{204}$$

If $x \in S_j^\circ$, then the above is evident. Therefore we only consider the case when x is a boundary point of S_j. From (196) and (199), x is a boundary point of S_j if and only if

$$CA^k x = b, \text{for some } \ k \in \{0, \ldots, n-1\} \tag{205}$$

Then from equations (183) through (187), we can construct a sequence of points that satisfy (204). Consequently the claim regarding (203) holds.

Recall that $b \in \mathbb{R}$ is a point where the quantizer $Q(\cdot)$ is discontinous. $\alpha, \beta \in \mathcal{Y}$ satisfy $\sup\{Q^{-1}(\alpha)\} \leq b$ and $\inf\{Q^{-1}(\beta)\} \geq b$. And by assumption any $q \in \{\alpha, \beta\}^{n+1} \times \{0\}^{n+1}$ is feasible. Next we define a correspondence between the sets $\{S_j\}$ and $\{\alpha, \beta\}^n \times \{0\}^n$, where $\{\alpha, \beta\}^n \times \{0\}^n \subset \mathcal{Y}^n \times \mathcal{U}^n$.

Given any $j \in \{1, \ldots, 2^n\}$, let $h(j) = (s_1, \ldots, s_n)$, where function h is defined in (190). Let a function $q(j) = (q_1, \ldots, q_n, 0, \ldots, 0) \in \{\alpha, \beta\}^n \times \{0\}^n$ satisfy:

$$q_k = \begin{cases} \alpha, & \text{if } \ s_k \ \text{ is } \ < \\ \beta, & \text{if } \ s_k \ \text{ is } \ \geq \end{cases} \tag{206}$$

Next, if $b \leq CA^n x^*$, then let $q(j_+) = (q_1, \ldots, q_n, 0, \ldots, 0)$ where j_+, $q(\cdot)$ are defined in (203), (206) respectively. Consider $q = (\alpha, q_1, \ldots, q_n, 0, \ldots, 0) \in \mathcal{Y}^{n+1} \times \mathcal{U}^{n+1}$. If

q is feasible, then the corresponding state x (165) is in the set S_{j_+} and $CA^n x < b$. This implies that there exist $x \in S_{j_+}$ such that $CA^n x < CA^n x^*$, which contradicts with (203). If $b > CA^n x^*$, let $q(j_-) = (q_1, \ldots, q_n, 0, \ldots, 0)$ where j_- is defined in (203). Let $q = (\beta, q_1, \ldots, q_n, 0, \ldots, 0) \in \mathcal{Y}^{n+1} \times \mathcal{U}^{n+1}$, then there exist $x \in S_{j_-}$ such that $CA^n x > CA^n x^*$, which contradicts with (203). In both cases, we can find an element q in $\mathcal{Y}^{n+1} \times \mathcal{U}^{n+1}$ that is not feasible. This contradicts with the assumption at the beginning of this derivation that any $q \in \mathcal{Y}^{n+1} \times \mathcal{U}^{n+1}$ is feasible.

Case B *(C,A) Unobservable*

First we claim that there exist $x^* \in \mathbb{R}^n$ such that:

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x^* = \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix} \tag{207}$$

To see this, by (187), there exist $x_1, x_2 \in \mathbb{R}^n$ such that:

$$CA^k x_1 > b, \text{ for all } 0 \leq k \leq n-1$$

and

$$CA^k x_2 > b, \text{ for all } 1 \leq k \leq n-1$$

and

$$Cx_2 < b$$

Let $l(\gamma) = C(\gamma x_1 + (1-\gamma)x_2), \gamma \in [0,1]$. Since $[0,1]$ is compact, $l(\cdot)$ is continuous, and $l(0) < b$ and $l(1) > b$, there exist $\gamma^* \in (0,1)$ such that $l(\gamma^*) = b$. Let $x_3 = \gamma^* x_1 + (1-\gamma^*)x_2$, then:

$$CA^k x_3 > b, \text{ for all } 1 \leq k \leq n-1$$

and

$$Cx_3 = b$$

Similarly, we can find x_4 such that

$$CA^k x_4 > b, \text{ for all } 2 \leq k \leq n-1$$

and

$$Cx_4 = b \text{ and } CAx_4 < b$$

Then there exist a convex combination x_5 of x_4 and x_3 such that

$$CA^k x_5 > b, \text{ for all } 2 \leq k \leq n-1$$

and

$$Cx_5 = b \text{ and } CAx_5 = b$$

Repeat this process, we can find a x^* such that (207) holds.

From (187), define $\{z_j\}_{j=1}^{2^n} \subset \mathbb{R}^n$ such that for all $(s_1, s_2, \ldots, s_n) \in \{<, >\}^n$, there exist j such that

$$CA^{k-1} z_j \quad s_k \quad b, \text{ for all } 1 \leq k \leq n \tag{208}$$

Similar to Case A, define another collection of vectors $\{v_j\}_{j=1}^{2^n}$:

$$v_j = \Theta(z_j - x^*) \tag{209}$$

where Θ is defined in (176). We see that $\{v_j\}_{j=1}^{2^n}$ satisfy (181).

By Lemma 5, $span(\{v_j\}_{j=1}^{2^n}) = \mathbb{R}^n$. By (209), we see that v_j is in the column span of Θ. Therefore the column span of Θ equals $\mathbb{R}^n$, which contradicts with (C, A) being unobservable. This completes the derivation of Theorem 12. □

6.4 Alternate Initialization Scheme

According to Theorem 12, if the memory length i of the finite state approximation $\hat{S}_i$ is greater than the dimension n of the state space of system (2) with $m = p = 1$, then $\hat{S}_i$ cannot start arbitrarily within $\mathcal{Q}_F$ and achieve the same approximation quality as when it is initialized at q_o. Consequently, in order to reduce the size of $\mathcal{Q}$, we need to design an alternate initialization process for $\hat{S}_i$ when the conditions in Lemma 4 or Theorem 12 are not satisfied. We start constructing the alternate initialization scheme by making the following observation.

Observation 3. Given system (163) and its finite state ρ/μ approximation $\hat{S}_i$, $\mathcal{Q}_F$ and $\mathcal{Q}_I$ are defined in (167) and (168) respectively. There exists a function $\psi : \mathcal{Q}_I \to \mathcal{Q}_F$ such that for any $q = (y_1, y_2 \dots y_j, u_1, u_2 \dots u_j) \in \mathcal{Q}_I$, let $\psi(q) = (\tilde{y}_1, \tilde{y}_2 \dots \tilde{y}_i, \tilde{u}_1, \tilde{u}_2 \dots \tilde{u}_i)$, the following relations hold:

$$\begin{array}{rcl} \tilde{y}_{i-(j-1)} &=& y_1 \\ \tilde{y}_{i-(j-2)} &=& y_2 \\ &\vdots& \\ \tilde{y}_i &=& y_j \end{array}, \quad \text{and} \quad \begin{array}{rcl} \tilde{u}_{i-(j-1)} &=& u_1 \\ \tilde{u}_{i-(j-2)} &=& u_2 \\ &\vdots& \\ \tilde{u}_i &=& u_j \end{array}. \tag{210}$$

To see this observation, we construct a function $\psi(\cdot)$. Given $q \in \mathcal{Q}_I$, write $q = (y_1, y_2 \dots y_j, u_1, u_2 \dots u_j)$, then there exist $x \in \mathbb{R}^n$ such that (165) is satisfied. Choose any $u \in \mathcal{U}$, and let

$$\begin{aligned} \tilde{y}_{i-j} &= g(f_{u_1} \circ \cdots \circ f_{u_j}(x)) \\ \tilde{y}_{i-(j+1)} &= g(f_u \circ f_{u_1} \circ \cdots \circ f_{u_j}(x)) \\ \tilde{y}_{i-(j+2)} &= g(f_u \circ f_u \circ f_{u_1} \circ \cdots \circ f_{u_j}(x)) \\ &\vdots \\ \tilde{y}_1 &= g(f_u \circ \cdots \circ f_u \circ f_{u_1} \circ \cdots \circ f_{u_j}(x)) \end{aligned} \tag{211}$$

and let

$$\psi(q) = (\tilde{y}_1, \ldots, \tilde{y}_{i-j}, y_1, \ldots y_j, u, \ldots, u, u_1, \ldots u_j) \tag{212}$$

then $\psi(q)$ is in $\mathcal{Q}_F$.

Now we are ready to present an alternate initialization scheme which reduces the size of $\mathcal{Q}$.

Consider $\hat{S}'_i$ described by:

$$q_{t+1} = \phi'(q_t, u_t, y_t, t) \tag{213a}$$

$$\hat{y}_t = \theta(q_t) \tag{213b}$$

where $q_t \in \mathcal{Q}'$ for some finite set $\mathcal{Q}'$, $u_t \in \mathcal{U}$, $y_t \in \mathcal{Y}$, $\hat{y}_t \in \mathcal{Y}$ and functions $\phi' : \mathcal{Q} \times \mathcal{U} \times \mathcal{Y} \times \mathbb{N} \to \mathcal{Q}$ and $\theta : \mathcal{Q} \times \mathcal{U} \to \mathcal{Y}$. Next we describe the new state set $\mathcal{Q}'$ and the transition function ϕ'.

State Set $\mathcal{Q}'$: Given a system (163), the finite state-space $\mathcal{Q}'$ of $\hat{S}'_i$ is defined as:

$$\mathcal{Q}' = \{q \in \mathcal{Y}^i \times \mathcal{U}^i | q \text{ is feasible}\} \cup \{q_\varnothing\} \tag{214}$$

Essentially, $\mathcal{Q}' = \mathcal{Q}_F \cup \{q_\varnothing\}$ where $\mathcal{Q}_F$ is defined in (167).

Transition Function ϕ': For any $q \in \mathcal{Q}', y \in \mathcal{Y}, u \in \mathcal{U}$, and $t \in \mathbb{N}$

$\triangleright$ If $q = q_\varnothing$, then

$$\phi'(q, y, u, t) = q_\varnothing, \text{ for all } t \in \mathbb{N}$$

$\triangleright$ If $q \neq q_\varnothing$ and $t = 0$, then

$$\phi'(q, y, u, t) = \psi((y, u))$$

where function ψ satisfies (210).

$\triangleright$ If $q \neq q_\varnothing$ and $1 \leq t \leq i-1$, write $q = (y_1, y_2 \ldots y_i, u_1, u_2 \ldots u_i)$, and let

$$\bar{q} = (y, y_{i-(t-1)}, \ldots, y_i, u, u_{i-(t-1)}, \ldots, u_i), \tag{215}$$

then

$$\phi'(q, y, u, t) = \psi(\bar{q})$$

where function ψ satisfies (210).

$\triangleright$ If $q \neq q_\varnothing$ and $t \geq i$, write $q = (y_1, y_2 \dots y_i, u_1, u_2 \dots u_i)$, and let

$\bar{q} = (y, y_1, y_2 \dots y_{i-1}, u, u_1, u_2 \dots u_{i-1})$, then

$$\phi'(q, y, u, t) = \begin{cases} \bar{q}, & \text{if } \bar{q} \text{ is feasible} \\ q_\varnothing, & \text{otherwise} \end{cases} .$$

We claim that the above construction of $\hat{S}'_i$ achieve the same approximation quality as $\hat{S}_i$ for $t \geq i$.

Observation 4. Given a system (163) and its modified finite state approximation $\hat{S}'_i$ as constructed in Section 6.4, for any input sequence $\mathbf{u}$, any $x_0 \in \mathbb{R}^n$ and any $q_0 \in \mathcal{Q}'$, $q_i = (y_{i-1}, y_{i-2} \dots y_0, u_{i-1}, u_{i-2} \dots u_0)$ where u_k and y_k are the input and output of system (163) at time $t = k, 0 \leq k \leq i-1$.

We verify Observation 4 in a straightforward manner. Indeed, at $t = 0$, write $q_1 = \phi'(q_0, y_0, u_0, 0) = (\tilde{y}_1^1, \dots, \tilde{y}_i^1, \tilde{u}_1^1, \dots, \tilde{u}_i^1)$, then $(\tilde{y}_i^1, \tilde{u}_i^1) = (y_0, u_0)$. Next, at $t = 1$, write $q_2 = \phi'(q_1, y_1, u_1, 1) = (\tilde{y}_1^2, \dots, \tilde{y}_i^2, \tilde{u}_1^2, \dots, \tilde{u}_i^2)$, then $(\tilde{y}_{i-1}^2, \tilde{y}_i^2, \tilde{u}_{i-1}^2, \tilde{u}_i^2) = (y_1, y_0, u_1, u_0)$. Repeat this argument until $t = i-1$, then we have $q_i = (y_{i-1}, \dots, y_0, u_{i-1}, \dots, u_0)$.

Remark. We comment on the significance of $\hat{S}'_i$ here. Particularly, $\hat{S}'_i$ recovers the original construction $\hat{S}_i$ for time $t \geq i$, and $\hat{S}'_i$ has a strictly smaller state-space than $\hat{S}_i$. The tradeoff is that $\hat{S}'_i$ is a time-variant system, while $\hat{S}_i$ is time-invariant.

6.5 Summary

In this section, we derived conditions for systems over finite alphats such that the initialization process of their finite state ρ/μ approximations may be simplified. We characterized such conditions for the general case, as well as for a particular class of systems with linear internal dynamics. We also proposed a time-variant initialization scheme to reduce the

number of states used in the approximation, when these necessary and sufficient conditions are not met.

7 Existence of Finite Uniform Bisimulations

Along the way of our research, we realized that the results derived in [25] can be extended to address some relevant open problems in the theory of bisimulation. In this chapter, we begin by proposing a refined notion of finite bisimulation that we refer to as a 'finite uniform bisimulation'. We then derive a sufficient condition for the existence of such finite uniform bisimulations, and we investigated necessary conditions. We constructed an algorithm to compute finite uniform bisimulations when the sufficient condition is satisfied. We concluded with an illustrative example showing how to construct finite state machine models of the underlying system when these finite uniform bisimulations exist.

Some of the results presented in this chapter has been reported in [32]. A preprint version of this work can be found at [33].

7.1 Finite Uniform Bisimulations

7.1.1 Proposed Notions

We begin by defining the notion of *finite uniform bisimulation*, which is simply an equivalence relation that satisfies certain desired properties:

Definition 6. Consider a discrete-time system

$$x_{t+1} = f(x_t, u_t) \tag{216}$$

where $t \in \mathbb{N}$ is the time index, $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathcal{U}$ is the input, $f : \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}^n$ is given, and input alphabet $\mathcal{U}$ represents the collection of possible values of the input. Given a set $\mathcal{S} \subset \mathbb{R}^n$, we say an equivalence relation $\sim \subset \mathcal{S} \times \mathcal{S}$ is a *finite uniform bisimulation on* $\mathcal{S}$ if the following two conditions are satisfied:

(i) For any $x, x' \in \mathcal{S}$ and any $u \in \mathcal{U}$, if $x \sim x'$, then

$$f(x, u) \sim f(x', u) \tag{217}$$

(ii) For $x \in \mathcal{S}$ with $[x] = \{y \in \mathcal{S} | y \sim x\}$, we have

$$1 < |\{[x] | x \in \mathcal{S}\}| < \infty \tag{218}$$

Essentially (217) requires that each equivalence class transition into another equivalence class under any input, and (218) requires that there be a finite number of equivalence classes while avoiding the trivial instance of a single equivalence class.

We define a finite uniform bisimulation to be *regular* if the equivalence classes have a specific topological structure:

Definition 7. Given a finite uniform bisimulation $\sim$ on $\mathcal{S}$ of system (216), we say $\sim$ is *regular* if for all $x \in \mathcal{S}$, $[x] = \{y \in \mathcal{S} | y \sim x\}$ consists of open sets in $\mathbb{R}^n$ and possibly their boundary points.

We are interested in regular finite uniform bisimulations because we wish to avoid certain "pathological" finite uniform bisimulations, as will become clear when we discuss the necessary conditions for the existence of finite uniform bisimulations in Section 7.3.2.

7.1.2 Deterministic Finite State Bisimulation Models

Given a finite uniform bisimulation $\sim$ on $\mathcal{S}$ of system (216), it is straightforward to construct a deterministic finite state machine (DFM) that is bisimilar to the original system when the latter is restricted to evolve on $\mathcal{S}$. Indeed:

Definition 8. Given a system (216) denoted by P and a finite uniform bisimulation $\sim$ on $\mathcal{S}$ of P, consider the DFM $\hat{P}$ defined by

$$q_{t+1} = f_{\sim}(q_t, u_t), \tag{219}$$

where $t \in \mathbb{N}$ is the time index, $q_t \in \mathcal{Q}$ is the state, $u_t \in \mathcal{U}$ is the input, $\mathcal{Q} = \{[x] | x \in \mathcal{S}\}$ (essentially $\mathcal{Q}$ is the finite quotient set of $\mathcal{S}$ under equivalence relation $\sim$), $\mathcal{U}$ is the input

alphabet of system (216), and state transition function $f_\sim : \mathcal{Q} \times \mathcal{U} \to \mathcal{Q}$ is defined as

$$f_\sim(q, u) = [f(x, u)], \ \forall \ x \in q. \tag{220}$$

We say that $\hat{P}$ is *uniformly bisimilar* to P.

Note that since $\sim$ is a finite uniform bisimulation, it follows from (217) that $f_\sim$ is well-defined.

7.2 Problem Setup and Formulation

7.2.1 Systems of Interest and Problem Statement

We first introduce the specific class of systems (216) that we will study in this section. Consider a discrete-time dynamical system described by

$$x_{t+1} = Ax_t + Bu_t, \tag{221}$$

where $t \in \mathbb{N}$ is the time index, $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathcal{U}$ is the input, and $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$ are given. We enforce that the input u_t can only take *finitely* many values in $\mathcal{U} \subset \mathbb{R}^m$ (that is, $|\mathcal{U}| < \infty$).

For this class of systems, we are interested in questions of existence and construction of finite uniform bisimulations on a subset $\mathcal{S}$ of the state space $\mathbb{R}^n$. Particularly, in order for the bisimulation relation to yield a meaningful "equivalent" DFM, we require the set $\mathcal{S}$ be an invariant set of the system:

Definition 9. A set $\mathcal{S} \subset \mathbb{R}^n$ is an *invariant set* of system (221) if for any input sequence $\{u_t\}_{t=0}^{\infty} \in \mathcal{U}^{\mathbb{N}}$

$$x_0 \in \mathcal{S} \Rightarrow x_t \in \mathcal{S}, \text{ for all } t \in \mathbb{N}. \tag{222}$$

We are now ready to state the first problem of interest:

Problem 4. Given system (221), under what conditions on $A, B, \mathcal{U}$ does there exist a finite uniform bisimulation $\sim$ on some invariant set $\mathcal{S}$ of system (221)?

When Problem 4 has an affirmative answer, another set of problems naturally follows:

Problem 5. Given a system (221) that admits a finite uniform bisimulation on some invariant set $\mathcal{S}$, under what conditions on $A, B, \mathcal{U}$ can *an arbitrarily large number of* equivalence classes be generated by a finite uniform bisimulation?

Note that we seek (and propose) both analytical and constructive, algorithmic solutions to the above problems.

7.2.2 Comparison with Existing Work on Finite Bisimulations

Before presenting our main results, we briefly discuss the similarities and differences between the current problem of interest and some of the previous developments on finite bisimulations:

- Our definition of finite uniform bisimulation is stronger than that of finite bisimulation used in some of the literature, of which we pick [34] as a representative paper. In particular in that setting, the definition requires that if two states are bisimilar ($x \sim y$) and x transitions to x' under input u, then there exists an input u' such that y transition to y' under u' and $y' \sim y$. Note that u and u' need not be the same, and thus a finite bisimulation as in [34] is not necessarily a finite uniform bisimulation. We will use Example 6 in Section 7.5 to illustrate this difference.

- Our definition of finite uniform bisimulation is in accordance with the definitions of finite bisimulation introduced in [3, 35]. However, the sufficient conditions for existence of finite bisimulations derived in [3] concern linear vector fields, and as such correspond to special cases of (221) where B is the zero matrix, whereas the present contribution addresses the more general case where B is nonzero. Likewise, the dynamics of the system of interest in [35] are different, as the authors study systems of

the form $x_{t+1} = A_{\sigma(t)}x_t$, where $\sigma(t)$ is the switching signal and is considered to be the input.

- Finally, the finite input alphabet setup is unique in the literature, in contrast to typically studied setups where the input signal takes arbitrary instantaneous values in Euclidean space, or else the input signal is of certain form such as polynomial, exponential or sinusoidal as in [36].

7.3 Conditions for the Existence of Finite Uniform Bisimulations

7.3.1 Sufficient Conditions

We begin by defining a set that will be useful for formulating a sufficient condition for the existence of finite uniform bisimulations.

Definition 10. Given system (2), define set $\mathcal{A}_s$ as

$$\mathcal{A}_s = \{\alpha \in \mathbb{R}^n | \alpha = \sum_{\tau=0}^{t} A^{t-\tau} B u_\tau, u_{(\cdot)} \in \mathcal{U}, t \in \mathbb{N}\}. \tag{223}$$

Essentially, $\mathcal{A}_s$ is the collection of forced responses of system (2) in the state-space. Now, we are ready to propose a sufficient condition for the existence of finite uniform bisimulations on some invariant subset of the state space.

Theorem 13. Given system (221) with $0 \in \mathcal{U}$, assume that A has all eigenvalues within the unit disc. If $cl(\mathcal{A}_s)$ is not connected, then there exists a finite uniform bisimulation on a subset of $\mathbb{R}^n$ that is an invariant set of system (221).

To show this result, we first introduce several Lemmas which will be instrumental in this derivation of Theorem 13.

Lemma 6. Given system (221), if matrix A has all eigenvalues within the unit disc, then $cl(\mathcal{A}_s)$ is compact.

Proof. If $A \in \mathbb{R}^{n\times n}$ has all eigenvalues within the unit disc, then $\sum_{\tau=0}^{\infty} \|A^\tau\|_1$ converges (pp. 298, [24]). Since $\mathcal{U}$ is finite, $\max\{\|Bu\|_1 : u \in \mathcal{U}\}$ is also finite. Combining these

two facts, and applying triangle inequality, we conclude that $\mathcal{A}_s$ is bounded and therefore $cl(\mathcal{A}_s)$ is bounded. Since $cl(\mathcal{A}_s)$ is closed and bounded in $\mathbb{R}^n$, $cl(\mathcal{A}_s)$ is compact. $\square$

Next, we study the structure of set $\mathcal{A}_s$ as defined in (223). By the definition of $\mathcal{A}_s$ and $0 \in \mathcal{U}$, and recall (234), we have

$$\bigcup_{j=1}^{q} \mathcal{S}_j^1 = cl(\mathcal{A}_s). \tag{224}$$

Generally, for any $k \in \mathbb{Z}_+$, let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q^k}\}$ be an enumeration of the set $\mathcal{U}^k$, where $\mathbf{u}_j = (u_j^1, \ldots, u_j^k)$, $u_j^1, \ldots, u_j^k \in \mathcal{U}$, we define sets $\{\mathcal{S}_j^k\}_{j=1}^{q^k}$ as follows

$$\mathcal{S}_j^k = Bu_j^1 + ABu_j^2 + \cdots + A^{k-1}Bu_j^k + cl(A^k\mathcal{A}_s), \quad j = 1, 2, \ldots, q^k. \tag{225}$$

We also have

$$\bigcup_{j=1}^{q^k} \mathcal{S}_j^k = cl(\mathcal{A}_s). \tag{226}$$

Now we introduce the following Lemma.

Lemma 7. Given system (221), assume that A has all eigenvalues within the unit disc. If open sets $\mathcal{W}$ and $\mathcal{V}$ is a disconnection of $cl(\mathcal{A}_s)$, then there exists $k^* \in \mathbb{Z}_+$ such that for all $j \in \{1, \ldots, q^{k^*}\}$,

$$\mathcal{S}_j^{k^*} \cap \mathcal{W} \neq \varnothing \quad \Rightarrow \quad \mathcal{S}_j^{k^*} \subset \mathcal{W} \tag{227}$$

Proof. We show this Lemma by contradiction. We first assume that for all $k \in \mathbb{Z}_+$, there is $j(k) \in \{1, \ldots, q^k\}$ such that $\mathcal{S}_{j(k)}^k \cap \mathcal{W} \neq \varnothing$ and $\mathcal{S}_{j(k)}^k \cap \mathcal{V} \neq \varnothing$. For each k, choose $w_k \in \mathcal{S}_{j(k)}^k \cap \mathcal{W}$ and $v_k \in \mathcal{S}_{j(k)}^k \cap \mathcal{V}$. Then we have constructed two sequences $\{w_k\}_{k=1}^{\infty}$ and $\{v_k\}_{k=1}^{\infty}$.

Since $\{w_k\}_{k=1}^{\infty} \subset cl(\mathcal{A}_s)$, $\{v_k\}_{k=1}^{\infty} \subset cl(\mathcal{A}_s)$ and $cl(\mathcal{A}_s)$ is compact (by Lemma 6), there exists a subsequence $\{w_{k_l}\}_{l=1}^{\infty}$ that converges to a point in $cl(\mathcal{A}_s)$. Similarly, there also exists a subsequence of $\{v_{k_l}\}_{l=1}^{\infty}$ that converges to a point in $cl(\mathcal{A}_s)$. By relabeling, we

have found two sequences $\{w_{k_p}\}_{p=1}^{\infty}$ and $\{v_{k_p}\}_{p=1}^{\infty}$ such that

$$\lim_{p\to\infty} w_{k_p} = w, \quad \text{and} \quad \lim_{p\to\infty} v_{k_p} = v \tag{228}$$

where $w, v \in cl(\mathcal{A}_s)$.

By the construction of $\mathcal{S}_j^k$ (225), we see that for any j, $diam(\mathcal{S}_j^k) \leq \|A^k\|_1 diam(\mathcal{A}_s)$. Since A has all eigenvalues within the unit disc, $\lim_{k\to\infty} A^k = 0_{n\times n}$ (pp.298, [24]). By boundedness of set $\mathcal{A}_s$, $diam(\mathcal{A}_s)$ is finite. Therefore $diam(\mathcal{S}_j^k)$ goes to 0 as k tends to infinity. Note that $w_{k_p} \in \mathcal{S}_{j(k_p)}^{k_p}$ and $v_{k_p} \in \mathcal{S}_{j(k_p)}^{k_p}$, and $k_p \geq p$, therefore $\lim_{p\to\infty} \|w_{k_p} - v_{k_p}\|_1 = 0$. Combine with (228), we have $\lim_{p\to\infty} w_{k_p} = \lim_{p\to\infty} v_{k_p} = w$, where $w \in cl(\mathcal{A}_s)$. Without loss of generality, let $w \in \mathcal{W}$. Since $\mathcal{W}$ is open, there exist $\epsilon > 0$ such that the open ball $B_\epsilon(w) \subset \mathcal{W}$. Since $\mathcal{W} \cap \mathcal{V} = \varnothing$, $\{v_{k_p}\}_{p=1}^{\infty} \cap B_\epsilon(w) = \varnothing$. Therefore $\|v_{k_p} - w\|_1 \geq \epsilon$ for all p. This is a contradiction with $\lim_{p\to\infty} v_{k_p} = w$. Therefore (227) holds. □

Next, we introduce another Lemma which is based on Lemma 7.

Lemma 8. Given system (221), assume that A has all eigenvalues within the unit disc. If open sets $\mathcal{W}$ and $\mathcal{V}$ is a disconnection of $cl(\mathcal{A}_s)$, then there exist open sets $\mathcal{W}'$ and $\mathcal{V}'$ in $\mathbb{R}^n$ such that the pair $\mathcal{W}'$ and $\mathcal{V}'$ is also a disconnection of $cl(\mathcal{A}_s)$, and for all $j \in \{1, \ldots, q\}$

$$\mathcal{S}_j^1 \cap \mathcal{W}' \neq \varnothing \quad \Rightarrow \quad \mathcal{S}_j^1 \subset \mathcal{W}' \tag{229}$$

Proof. By Lemma 7 , (227) holds, and we only need to consider the case when $k^* \geq 2$.

Define a function $f : \mathbb{R}^n \to \mathbb{R}^n$ as :$f(x) = Ax$. Clearly f is continuous. For any set $\mathcal{S}$, use $f^{-1}(\mathcal{S})$ to denote the set $f^{-1}(\mathcal{S}) = \{x \in \mathbb{R}^n | f(x) \in \mathcal{S}\}$.

For $\{\mathcal{S}_j^{k^*}\}_{j=1}^{q^{k^*}}$ as constructed in (225), let u be an element of $\mathcal{U}$, then define an index set $\mathcal{J}$ as

$$\mathcal{J} = \{j \in \{1, \ldots, q^{k^*}\} : u_j^1 = u\},$$

then $|\mathcal{J}| = q^{k^*-1}$. Define sets

$$\tilde{\mathcal{S}}_j^{k^*} = -Bu + \mathcal{S}_j^{k^*}, j \in \mathcal{J}. \tag{230}$$

For any $j \in \mathcal{J}$, by (227), either $\tilde{\mathcal{S}}_j^{k^*} \subset -Bu + \mathcal{W}$ or $\tilde{\mathcal{S}}_j^{k^*} \subset -Bu + \mathcal{V}$. Write $\mathcal{W}' = f^{-1}(-Bu + \mathcal{W})$ and $\mathcal{V}' = f^{-1}(-Bu + \mathcal{V})$, then either $f^{-1}(\tilde{\mathcal{S}}_j^{k^*}) \subset \mathcal{W}'$ or $f^{-1}(\tilde{\mathcal{S}}_j^{k^*}) \subset \mathcal{V}'$.

For each $j \in \mathcal{J}$, by (225), (230), and the compactness of $cl(\mathcal{A}_s)$, we have

$$\tilde{\mathcal{S}}_j^{k^*} = A(Bu_j^2 + \cdots + A^{k^*-2}Bu_j^{k^*} + cl(A^{k^*-1}\mathcal{A}_s))$$

for some $(u_j^2, \ldots, u_j^{k^*}) \in \mathcal{U}^{k^*-1}$. Consequently, we can determine one and only one $j' \in \{1, \ldots, q^{k^*-1}\}$ such that

$$\mathcal{S}_{j'}^{k^*-1} \subset f^{-1}(\tilde{\mathcal{S}}_j^{k^*}). \tag{231}$$

We also observe that

$$\bigcup_{j\in\mathcal{J}} (u_j^2, u_j^3, \ldots, u_j^{k^*}) = \mathcal{U}^{k^*-1}. \tag{232}$$

Recall (226), (231), we see that

$$cl(\mathcal{A}_s) = \bigcup_{j'=1}^{q^{k^*-1}} \mathcal{S}_{j'}^{k^*-1} \subset \bigcup_{j\in\mathcal{J}} f^{-1}(\tilde{\mathcal{S}}_j^{k^*}) \subset \mathcal{W}' \cup \mathcal{V}'. \tag{233}$$

It is clear that $\mathcal{W}'$ and $\mathcal{V}'$ are disjoint open sets. Therefore (227) holds for $k^* - 1$ and $\mathcal{W}'$, $\mathcal{V}'$. Repeat this argument $k^* - 1$ times, we conclude that (229) holds. □

Finally, we provide the proof of Theorem 13.

Proof. (of Theorem 13) Since $cl(\mathcal{A}_s)$ is not connected, let $\mathcal{W}$ and $\mathcal{V}$ be a disconnection of $cl(\mathcal{A}_s)$. Then by Lemma 8, (229) holds. We propose an equivalence relation on $\mathcal{A}_s$. Since $\mathcal{A}_s$ is an invariant set of system (221), the proof is complete if we can show that this equivalence relation satisfies (217) and (218).

Given open sets $\mathcal{W}'$ and $\mathcal{V}'$ that satisfy (229), let $\mathcal{X}_1 = \mathcal{A}_s \cap \mathcal{W}'$ and $\mathcal{X}_2 = \mathcal{A}_s \cap \mathcal{V}'$.

Define an equivalence relation $\sim$ as

$$x \sim x' \Leftrightarrow x \in \mathcal{X}_i \text{ and } x' \in \mathcal{X}_i \text{ for some } i \in \{1, 2\}.$$

For any $x, x' \in \mathcal{A}_s$, any $u_j \in \mathcal{U}$, if $x \sim x'$, then $Ax + Bu_j \in \mathcal{S}_j^1$ and $Ax' + Bu_j \in \mathcal{S}_j^1$. By (229), we see that $Ax + Bu_j \sim Ax' + Bu_j$. Therefore (217) is satisfied. Since $1 < 2 < \infty$, (218) is also satisfied. This completes the proof. □

Next, we continue to study Problem 5. It turns out that additional assumptions are needed to guarantee the existence of arbitrarily many equivalence classes, as we shall see in Section 7.5 Example 7. In order to describe such conditions, we first define a relevant collection of subsets of the state space $\mathbb{R}^n$: Given system (221), let $\mathcal{U} = \{u_1, u_2, \ldots, u_q\}$ for $q \in \mathbb{Z}_+$ and define sets $\{\mathcal{S}_j^1\}_{j=1}^q$ as follows

$$\mathcal{S}_j^1 = Bu_j + cl(A\mathcal{A}_s), \quad j = 1, 2, \ldots, q. \tag{234}$$

We can now propose a sufficient condition for the existence of an arbitrarily large number of equivalence classes.

Theorem 14. Given system (221) with $0 \in \mathcal{U}$ and $|\mathcal{U}| > 1$, assume that A has all eigenvalues within the unit disc. If A is invertible, and $\{S_j^1\}_{j=1}^q$ (234) are disjoint, then for any $z \in \mathbb{Z}_+$ there is a finite uniform bisimulation $\sim$ of system (221) such that the number of equivalence classes associated with $\sim$ is greater than z.

The derivation of this result is given in Section 7.4.

7.3.2 Necessary Conditions

Next, we investigate necessary conditions for the existence of finite uniform bisimulations. We quickly realize that system (221) may admit "pathological" finite uniform bisimulations: If $A, B, \mathcal{U}$ have entries in $\mathbb{Q}$, then the partition $\mathbb{Q}^n$ and $\mathbb{R}^n \setminus \mathbb{Q}^n$ affords a finite uniform bisimulation of system (221). This motivates us to study regular finite uni-

form bisimulations. We propose a necessary condition for the existence of regular finite uniform bisimulations.

Theorem 15. Given system (221) with $0 \in \mathcal{U}$. If $\sim$ is a regular finite uniform bisimulation on an invariant set $\mathcal{S}$ of system (221), $0 \in int([0])$, and $[0]$ is bounded, then $\rho(A) \leq 1$.

Remark. Theorem 15 states that under certain assumptions, there do not exist regular finite uniform bisimulations for Schur unstable systems (221). This justifies why we study Schur stable systems in Theorem 13.

Proof. (of Theorem 15) We will prove by contradiction. Assume $\rho(A) > 1$, let $Av = \lambda v$ with $|\lambda| > 1$, $\|v\|_1 = 1$, $\lambda \in \mathbb{C}$, $v \in \mathbb{C}^n$. And for any $w \in \mathbb{C}^n$, we use $Re(w)$ to denote the real part of w. Define a set $\mathcal{O}$ as

$$\mathcal{O} = \{\alpha \in \mathbb{R}_+ | Re(\gamma v) \in [0], \text{ for all } |\gamma| \leq \alpha, \gamma \in \mathbb{C}\}. \tag{235}$$

We show that $\mathcal{O}$ is non-empty and bounded in the following. Write $v = [v_1\ v_2\ \dots\ v_n]^T$, where $v_1, \dots, v_n \in \mathbb{C}$ and $|v_1| + \dots + |v_n| = 1$. For any $\gamma \in \mathbb{C}$, we have $|Re(\gamma v_i)| \leq |\gamma||v_i|$, therefore

$$\|Re(\gamma v)\|_1 = \sum_{i=1}^{n} |Re(\gamma v_i)| \leq |\gamma| \sum_{i=1}^{n} |v_i| = |\gamma|.$$

Since $B_r(0) \subset [0]$ for some $r > 0$ by assumption, for all γ with $|\gamma| \leq r/2$, $Re(\gamma v) \in B_r(0)$. Therefore $r/2 \in \mathcal{O}$, and $\mathcal{O}$ is nonempty.

Next, we show that $\mathcal{O}$ is bounded. Since $[0]$ is bounded by assumption, let $[0] \subset B_\sigma(0)$ for some $\sigma > 0$. Since $v = [v_1\ v_2\ \dots\ v_n]^T \neq 0_{n\times 1}$, let $|v_k| > 0$ for some $1 \leq k \leq n$. Write v_k as $v_k = |v_k|e^{i\phi}$ for some $\phi \in [0, 2\pi)$. Assume $\mathcal{O}$ is unbounded, then there exist $\alpha \in \mathcal{O}$ with $|\alpha| > 2\sigma/|v_k|$. Let $\gamma = (2\sigma/|v_k|)e^{i(-\phi)}$, then $|\gamma| < \alpha$. By the definition of $\mathcal{O}$ (235), we have $Re(\gamma v) \in [0]$. Observe that

$$\|Re(\gamma v)\|_1 \geq |Re(\gamma v_k)| = |Re(\frac{2\sigma}{|v_k|} e^{i(-\phi)} |v_k| e^{i\phi})| = |Re(2\sigma)| = 2\sigma.$$

Therefore $Re(\gamma v) \notin B_\sigma(0)$, and consequently $Re(\gamma v) \notin [0]$, which draws a contradiction.

Therefore $\mathcal{O}$ is bounded.

Next, we define $\beta = \sup \mathcal{O}$. Since $\mathcal{O}$ is non-empty and bounded, we have $0 < \beta < \infty$. Then for any $\epsilon > 0$, there is $0 \leq \delta < \epsilon$ such that $Re(\kappa v) \notin [0]$ for some $\kappa \in \mathbb{C}$ and $|\kappa| = \beta + \delta$. Choose $\epsilon = (\frac{|\lambda|-1}{2})\beta$, and let $\kappa' = \frac{\kappa}{\lambda}$, then

$$|\kappa'| = \frac{|\kappa|}{|\lambda|} = \frac{\beta + \delta}{|\lambda|} < \frac{\beta + \epsilon}{|\lambda|} < \frac{\beta + (|\lambda| - 1)\beta}{|\lambda|} = \beta.$$

Therefore $|\kappa'| < \beta$. Since $\beta = \sup \mathcal{O}$, there exists $\alpha \in \mathcal{O}$ such that $\alpha > |\kappa'|$. By (235), we see that $Re(\kappa' v) \in [0]$, or equivalently $Re(\kappa' v) \sim 0$. Since $\sim$ is a finite uniform bisimulation, by (217) and letting the input u be zero, we have $ARe(\kappa' v) \sim 0$. We observe that

$$ARe(\kappa' v) = Re(A\kappa' v) = Re(\kappa'(Av)) = Re(\kappa' \lambda v) = Re(\kappa v),$$

therefore $Re(\kappa v) \sim 0$,which draws a contradiction. We conclude that the assumption $\rho(A) > 1$ is false, and therefore $\rho(A) \leq 1$.

□

We point out that the condition "$[0]$ is bounded" in Theorem 15 cannot be dropped (see Example 8 in Section 7.5). However, the condition "$[0]$ is bounded" in Theorem 15 can be dropped for scalar systems, where we restrict our attention to instances of (221) described by

$$x_{t+1} = ax_t + bu_t \tag{236}$$

where $x_t \in \mathbb{R}$, $u_t \in \mathcal{U}$, and $a, b \in \mathbb{R}$. $\mathcal{U}$ is a finite subset of $\mathbb{R}$.

Corollary 1. Given system (236) with $0 \in \mathcal{U}$. If $\sim$ is a regular finite uniform bisimulation on an invariant set $\mathcal{S}$ of system (236) and $0 \in int([0])$, then $|a| \leq 1$.

Proof. We will prove by contradiction. Assume $|a| > 1$, and use $[0]$ to denote the equivalence class $[0] = \{x \in \mathcal{S} | x \sim 0\}$. By the assumption $B_r(0) \subset [0]$ for some $r > 0$, define β as

$$\beta = \sup\{x \in \mathcal{S} | [0, x] \subset [0]\}, \tag{237}$$

where $[0, x]$ is the closed interval between 0 and x. Since $int([0])$ is nonempty, there is ϵ such that $[0, \epsilon) \subset [0]$, therefore the supremum is well defined, and $\beta > 0$.

First, we consider the case $\beta < \infty$. Clearly $[0, \beta) \subset [0]$. By the definition of β, we have that for any $\epsilon > 0$, there is $0 \leq \delta < \epsilon$ such that

$$\beta + \delta \notin [0]. \tag{238}$$

Let $\epsilon = (a^2 - 1)\beta > 0$, and let δ denote the nonnegative number that satisfy (238). We observe that

$$z = \frac{\beta + \delta}{a^2} < \beta,$$

therefore $z \sim 0$. Since $\sim$ is a finite uniform bisimulation, when the input is 0 we have $az \sim 0$, and $a^2 z \sim 0$. This draws a contradiction with (238).

For the case $\beta = \infty$, let $\beta' = \inf\{x \in \mathcal{S} | [x, 0] \subset [0]\}$, then $\beta' > -\infty$, otherwise for any $x \in \mathbb{R}$, $x \in [0]$, which implies $\mathbb{R} = [0]$ and there is only one equivalence class. Next, for any $\epsilon > 0$, there is $0 \leq \delta < \epsilon$ such that $\beta' - \delta \notin [0]$. Choose $\epsilon = (1 - a^2)\beta'$ and $z = (\beta' - \delta)/a^2$, then the preceding argument follows. □

7.4 Constructive Algorithms

In this section, we present algorithms for computing finite uniform bisimulations when the sufficient conditions are satisfied. First, when the conditions in Theorem 13 are satisfied, we propose an algorithm, and we show that the proposed Algorithm 1 is guaranteed to generate a finite uniform bisimulation when the sufficient condition is satisfied.

We begin by introducing the notation of binary partitions of the finite input set $\mathcal{U}$ with $|\mathcal{U}| > 1$: A pair $(\mathcal{U}_1, \mathcal{U}_2)$ is a binary partition of $\mathcal{U}$ if $\mathcal{U}_1, \mathcal{U}_2$ are nonempty, disjoint subsets of $\mathcal{U}$, and $\mathcal{U}_1 \cup \mathcal{U}_2 = \mathcal{U}$. The order of $\mathcal{U}_1, \mathcal{U}_2$ is not relevant: $(\mathcal{U}_1, \mathcal{U}_2)$ is the same as $(\mathcal{U}_2, \mathcal{U}_1)$. Since $\mathcal{U}$ is a finite set, there are finitely many distinct binary partitions of $\mathcal{U}$. We use $\{(\mathcal{U}_1^{(i)}, \mathcal{U}_2^{(i)}) : i = 1, \ldots, r\}$ to denote the collection of all binary partitions of $\mathcal{U}$. Here $r = (C_q^1 + C_q^2 + \cdots + C_q^{q-1})/2$, where $q = |\mathcal{U}|$, and $C_q^j = \frac{q!}{j!(q-j)!}$ represents the quantity "q

choose j". Now we are ready to present the following algorithm to compute finite uniform bisimulations of system (221).

Algorithm 1 Computing a Finite Uniform Bisimulation

Input: Matrix A, B, set $\mathcal{U}$
Compute: $h = \max\{\|Bu\|_1 : u \in \mathcal{U}\}$
Choose: ϵ such that $0 < \epsilon < 1 - \rho(A)$.
Compute: Matrix T, invertible, such that $\|T^{-1}AT\|_1 \leq \rho(A) + \epsilon$.
Compute: All binary partitions of $\mathcal{U}$: $(\mathcal{U}_1^{(i)}, \mathcal{U}_2^{(i)}), i = 1, \ldots, r$.
Compute: $\kappa = \frac{2\|T\|_1\|T^{-1}\|_1}{1-\rho(A)-\epsilon}$
$k \leftarrow 1$.
loop
    Compute: $l_k = h\|A^k\|_1$
    $i \leftarrow 1$.
    while $i \leq r$ **do**
        Compute: $\mathcal{C}_1^{(i)} = \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 \in \mathcal{U}_1^{(i)}, u_2, \ldots, u_k \in \mathcal{U}\}$
        $\mathcal{C}_2^{(i)} = \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 \in \mathcal{U}_2^{(i)}, u_2, \ldots, u_k \in \mathcal{U}\}$
        Compute: $d_k^{(i)} = \min\{\|\alpha - \beta\|_1 : \alpha \in \mathcal{C}_1^{(i)}, \beta \in \mathcal{C}_2^{(i)}\}$
        if $d_k^{(i)} \geq \kappa l_k$ **then**
            $\tilde{i} \leftarrow i, \tilde{k} \leftarrow k$.
            Exit the loop
        end if
        $i \leftarrow i + 1$.
    end while
    $k \leftarrow k + 1$.
end loop
Compute: $\mathcal{S} = \{x \in \mathbb{R}^n : \|T^{-1}x\|_1 < \frac{d_{\tilde{k}}^{(\tilde{i})}}{2\|T\|_1}\}$
Compute: $\mathcal{X}_1 = \mathcal{C}_1^{(\tilde{i})} + \mathcal{S}, \mathcal{X}_2 = \mathcal{C}_2^{(\tilde{i})} + \mathcal{S}$
Return: $\mathcal{X}_1, \mathcal{X}_2$

Remark. In the preceding algorithm, one approach to compute matrix T involves Schur's triangularization of matrix A (pp. 79, [24]). We refer interested readers to [24] on the specifics of computing matrix T such that $\|T^{-1}AT\|_1 \leq \rho(A) + \epsilon$ is satisfied.

Remark. Here we explain why Algorithm 1 returns two equivalence classes. We first point out that if the conditions in Theorem 13 are satisfied, the number of equivalence classes generated by a finite uniform bisimulation could be greater than two, which is the case in Example 9 in Section 7.5. However, for certain systems (see Example 7 in Section 7.5),

two, and only two equivalence classes can be generated based on the analytical result stated in Theorem 13. Therefore Algorithm 1 returns two equivalence classes, since it is capable of computing finite uniform bisimulations for *any* system that satisfies the conditions in Theorem 13. As we shall see next, we propose another algorithm in case more equivalence classes are desired.

We claim that Algorithm 1 is guaranteed to generate a finite uniform bisimulation when the sufficient condition is satisfied.

Theorem 16. Given system (221), and let the hypotheses in Theorem 13 hold, then Algorithm 1 terminates, and returns $\mathcal{X}_1, \mathcal{X}_2$ such that $\mathcal{X}_1, \mathcal{X}_2$ afford a finite uniform bisimulation on an invariant set $\mathcal{S}$, namely $\mathcal{S} = \mathcal{X}_1 \cup \mathcal{X}_2$, of system (221).

Proof. To derive Theorem 16, we first show that Algorithm 1 terminates, and then show that the equivalence classes $\mathcal{X}_1, \mathcal{X}_2$ returned by Algorithm 1 afford a finite uniform bisimulation on $\mathcal{X}_1 \cup \mathcal{X}_2$.

Given system (221), since matrix A has all eigenvalues within the unit disc, and $cl(\mathcal{A}_s)$ is not connected, by Lemma 8, there is a disconnection of $cl(\mathcal{A}_s)$, $\mathcal{W}$ and $\mathcal{V}$, such that for all $j \in \{1, \ldots, q\}$

$$\mathcal{S}_j^1 \cap \mathcal{W} \neq \varnothing \quad \Rightarrow \quad \mathcal{S}_j^1 \subset \mathcal{W} \tag{239}$$

where $q = |\mathcal{U}|$. Let $\mathcal{U}_1^* = \{u_j \in \mathcal{U} | \mathcal{S}_j^1 \cap \mathcal{W} \neq \varnothing\}$, and $\mathcal{U}_2^* = \mathcal{U} \setminus \mathcal{U}_1^*$. Recall (224), we see that $\mathcal{U}_1^*$ is nonempty, otherwise $cl(\mathcal{A}_s) \cap \mathcal{W} = \varnothing$, which contradicts with $\mathcal{W}$ and $\mathcal{V}$ being a disconnection of $cl(\mathcal{A}_s)$. $\mathcal{U}_2^*$ is also nonempty, otherwise $cl(\mathcal{A}_s) \subset \mathcal{W}$, then $cl(\mathcal{A}_s) \cap \mathcal{V} = \varnothing$, which draws a contradiction. We also observe that $|\mathcal{U}| > 1$, otherwise $\mathcal{U} = 0$ by assumption, and $cl(\mathcal{A}_s) = 0$ is connected. Therefore the binary partitions of $\mathcal{U}$ are well-defined. Since $\mathcal{U}_1^*$ and $\mathcal{U}_2^*$ are nonempty, disjoint subsets of $\mathcal{U}$, and $\mathcal{U}_1^* \cup \mathcal{U}_2^* = \mathcal{U}$, there is a binary partition of $\mathcal{U}$, $(\mathcal{U}_1^{(i^*)}, \mathcal{U}_2^{(i^*)})$, such that

$$(\mathcal{U}_1^{(i^*)}, \mathcal{U}_2^{(i^*)}) = (\mathcal{U}_1^*, \mathcal{U}_2^*) \tag{240}$$

where i^* is an integer between 1 and r.

Since for any $k \in \mathbb{Z}_+$,

$$d_k^{(i)} = \min\{\|\alpha - \beta\|_1 : \alpha \in \mathcal{C}_1^{(i)}, \beta \in \mathcal{C}_2^{(i)}\}, \tag{241}$$

we claim that $d_k^{(i^*)}$ (241) is uniformly bounded away from zero, that is: There exists $d > 0$ such that

$$d_k^{(i^*)} \geq d, \quad \text{for all} \quad k \in \mathbb{Z}_+. \tag{242}$$

To see this claim, we define two sets $\mathcal{G}_1, \mathcal{G}_2$ by

$$\mathcal{G}_1 = \bigcup_{j \in \mathcal{U}_1^*} \mathcal{S}_j^1, \quad \mathcal{G}_2 = \bigcup_{j \in \mathcal{U}_2^*} \mathcal{S}_j^1. \tag{243}$$

By the definition of $\mathcal{U}_1^*$, we see that $\mathcal{G}_1 \subset \mathcal{W}$. Recall (224) and that $\mathcal{W}$ and $\mathcal{V}$ is a disconnection of $cl(\mathcal{A}_s)$, we see that $\mathcal{G}_2 \subset \mathcal{V}$. Because $\mathcal{V}$ and $\mathcal{W}$ are disjoint, $\mathcal{G}_1$ and $\mathcal{G}_2$ are also disjoint. Since $\mathcal{G}_1$ is a finite union of closed sets, $\mathcal{G}_1$ is closed. By Lemma 6, $cl(\mathcal{A}_s)$ is bounded, and therefore $\mathcal{G}_1$ is bounded. We see that $\mathcal{G}_1$ is closed, bounded, and therefore compact. Similarly, $\mathcal{G}_2$ is also compact. By an observation in analysis: The distance between two disjoint compact sets is positive (pp. 18, [31]), we have

$$d = \inf\{\|\alpha - \beta\|_1 : \alpha \in \mathcal{G}_1, \beta \in \mathcal{G}_2\} > 0. \tag{244}$$

Since

$$\begin{aligned} \mathcal{C}_1^{(i)} &= \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 \in \mathcal{U}_1^{(i)}, u_2, \ldots, u_k \in \mathcal{U}\} \\ \mathcal{C}_2^{(i)} &= \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 \in \mathcal{U}_2^{(i)}, u_2, \ldots, u_k \in \mathcal{U}\} \end{aligned} \tag{245}$$

and recall (234), (240), and (243), we observe that: For all $k \in \mathbb{Z}_+$,

$$\mathcal{C}_1^{(i^*)} \subset \mathcal{G}_1, \quad \mathcal{C}_2^{(i^*)} \subset \mathcal{G}_2. \tag{246}$$

Recall (241), we have $d_k^{(i^*)} \geq d > 0$ for all $k \in \mathbb{Z}_+$.

Since matrix A is Schur stable, we see that $l_k = h\|A^k\|_1 \to 0$ as $k \to \infty$. Consequently, there exists $k^* \in \mathbb{Z}_+$ such that

$$d_{k^*}^{(i^*)} \geq \kappa l_{k^*} = \frac{2\|T\|_1\|T^{-1}\|_1}{1-\rho(A)-\epsilon} l_{k^*}.$$

Now we see that the loop in Algorithm 1 terminates, and returns two sets $\mathcal{X}_1, \mathcal{X}_2$:

$$\begin{aligned} \mathcal{X}_1 &= \mathcal{C}_1^{(\tilde{i})} + \mathcal{S}, \\ \mathcal{X}_2 &= \mathcal{C}_2^{(\tilde{i})} + \mathcal{S}. \end{aligned} \tag{247}$$

For the second part of this derivation, we show that $\mathcal{X}_1 \cup \mathcal{X}_2$ is an invariant set of system (221), and that $\mathcal{X}_1, \mathcal{X}_2$ afford a finite uniform bisimulation on $\mathcal{X}_1 \cup \mathcal{X}_2$.

For any $x \in \mathcal{X}_1 \cup \mathcal{X}_2$, by (245) and (247), there exist $(u_1, \ldots, u_{\tilde{k}}) \in \mathcal{U}^{\tilde{k}}$ and $s \in \mathcal{S}$ such that

$$x = Bu_1 + ABu_2 + \cdots + A^{\tilde{k}-1}Bu_{\tilde{k}} + s. \tag{248}$$

Then for any $u \in \mathcal{U}$,

$$Ax + Bu = (Bu + ABu_1 + \cdots + A^{\tilde{k}-1}Bu_{\tilde{k}-1}) + (A^{\tilde{k}}Bu_{\tilde{k}} + As). \tag{249}$$

Recall $\|T^{-1}AT\|_1 \leq \rho(A) + \epsilon$, $\kappa = \frac{2\|T\|_1\|T^{-1}\|_1}{1-\rho(A)-\epsilon}$, and $h = \max\{\|Bu\|_1 : u \in \mathcal{U}\}$, we

observe that

$$\begin{aligned}
&\|T^{-1}(A^{\tilde{k}}Bu_{\tilde{k}} + As)\|_1 \\
&\leq \|T^{-1}A^{\tilde{k}}Bu_{\tilde{k}}\|_1 + \|T^{-1}As\|_1 \\
&\leq \|T^{-1}\|_1\|A^{\tilde{k}}\|_1\|Bu_{\tilde{k}}\|_1 + \|(T^{-1}AT)T^{-1}s\|_1 \\
&\leq \|T^{-1}\|_1 l_{\tilde{k}} + \|(T^{-1}AT)\|_1\|T^{-1}s\|_1 \\
&< \|T^{-1}\|_1\frac{1-\rho(A)-\epsilon}{2\|T\|_1\|T^{-1}\|_1}d_{\tilde{k}}^{(\tilde{i})} + (\rho(A)+\epsilon)\frac{d_{\tilde{k}}^{(\tilde{i})}}{2\|T\|_1} \\
&= \frac{d_{\tilde{k}}^{(\tilde{i})}}{2\|T\|_1}.
\end{aligned}$$

Therefore $(A^{\tilde{k}}Bu_{\tilde{k}} + As) \in \mathcal{S}$. By (245), we see that $(Bu + ABu_1 + \cdots + A^{\tilde{k}-1}Bu_{\tilde{k}-1}) \in \mathcal{C}_1^{\tilde{i}} \cup \mathcal{C}_2^{\tilde{i}}$, therefore, we have

$$Ax + Bu \in \mathcal{X}_1 \cup \mathcal{X}_2. \tag{250}$$

We conclude that $\mathcal{X}_1 \cup \mathcal{X}_2$ is an invariant set of system (221).

Next, we show $\mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing$. We show by contradiction: Assume $z \in \mathcal{X}_1 \cap \mathcal{X}_2$, then by (247), there exist $c_1 \in \mathcal{C}_1^{\tilde{i}}$, $c_2 \in \mathcal{C}_2^{\tilde{i}}$, $s_1 \in \mathcal{S}$, and $s_2 \in \mathcal{S}$ such that $z = c_1 + s_1,$ and $z = c_2 + s_2$, and recall $\mathcal{S} = \{x \in \mathbb{R}^n : \|T^{-1}x\|_1 < d_{\tilde{k}}^{(\tilde{i})}/(2\|T\|_1)\}$, we have

$$\begin{aligned}
\|c_1 - c_2\|_1 &\leq \|c_1 - z\|_1 + \|z - c_2\|_1 \\
&= \|s_1\|_1 + \|s_2\|_1 \\
&= \|T(T^{-1}s_1)\|_1 + \|T(T^{-1}s_2)\|_1 \\
&\leq \|T\|_1(\|T^{-1}s_1\|_1 + \|T^{-1}s_2\|_1) \\
&< d_{\tilde{k}}^{(\tilde{i})}.
\end{aligned}$$

But by (241), we have $\|c_1 - c_2\|_1 \geq d_{\tilde{k}}^{(\tilde{i})}$, which draws a contradiction. Therefore $\mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing$.

Now we are ready to define an equivalence relation $\sim$ on $\mathcal{X}_1 \cup \mathcal{X}_2$ as:

$$x \sim x' \Leftrightarrow x \in \mathcal{X}_i \text{ and } x' \in \mathcal{X}_i \text{ for some } i \in \{1, 2\}.$$

We show that $\sim$ is a finite uniform bisimulation on $\mathcal{X}_1 \cup \mathcal{X}_2$. For any $x, x' \in \mathcal{X}_1 \cup \mathcal{X}_2$, and any $u \in \mathcal{U}$, if $x \sim x'$, we consider two cases: If $u \in \mathcal{U}_1^{\tilde{i}}$, recall (245), (247), and (249), we see that $Ax + Bu \in \mathcal{X}_1$ and $Ax' + Bu \in \mathcal{X}_1$, therefore $Ax + Bu \sim Ax' + Bu$. Similarly, if $u \in \mathcal{U}_2^{\tilde{i}}$, then $Ax + Bu \in \mathcal{X}_2$ and $Ax' + Bu \in \mathcal{X}_2$, therefore $Ax + Bu \sim Ax' + Bu$. Since $(\mathcal{U}_1^{\tilde{i}}, \mathcal{U}_2^{\tilde{i}})$ is a binary partition of $\mathcal{U}$, we see that (217) is satisfied.

Since $\{[x] | x \in \mathcal{X}_1 \cup \mathcal{X}_2\} = \{\mathcal{X}_1, \mathcal{X}_2\}$, we have $|\{[x] | x \in \mathcal{X}_1 \cup \mathcal{X}_2\}| = 2$, and (218) is satisfied. Therefore $\sim$ is a finite uniform bisimulation on $\mathcal{X}_1 \cup \mathcal{X}_2$. This completes the proof. $\square$

Next, we present a second algorithm, which is an extended version of Algorithm 1, to generate an arbitrarily large number of equivalence classes when the conditions in Theorem 14 are satisfied.

Algorithm 2 Computing a Finite Uniform Bisimulation with Many Equivalence Classes

Input: Matrix A, B, set $\mathcal{U} = \{u_{(1)}, u_{(2)}, \ldots, u_{(q)}\}$, integer z: Lower bound of the number of equivalence classes.

Compute: $h = \max\{\|Bu\|_1 : u \in \mathcal{U}\}$
Choose: ϵ such that $0 < \epsilon < 1 - \rho(A)$.
Compute: Matrix T, invertible, such that $\|T^{-1}AT\|_1 \leq \rho(A) + \epsilon$.
Compute: $\kappa = \frac{2\|T\|_1\|T^{-1}\|_1}{1-\rho(A)-\epsilon}$
$k \leftarrow 1$.
loop
  Compute: $l_k = h\|A^k\|_1$
  Compute:
$$\mathcal{C}_1^{(k)} = \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 = u_{(1)}, u_2, \ldots, u_k \in \mathcal{U}\}$$
$$\mathcal{C}_2^{(k)} = \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 = u_{(2)}, u_2, \ldots, u_k \in \mathcal{U}\}$$
$$\vdots$$
$$\mathcal{C}_q^{(k)} = \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 = u_{(q)}, u_2, \ldots, u_k \in \mathcal{U}\}$$
  Compute: $d_k = \min\{\|\alpha - \beta\|_1 : \alpha \in \mathcal{C}_v^{(i)}, \beta \in \mathcal{C}_w^{(i)}, w \neq v, 1 \leq w, v \leq q\}$
  if $d_k \geq \kappa l_k$ **then**
    $\tilde{k} \leftarrow k$.
    Exit the loop
  end if
  $k \leftarrow k + 1$.
end loop
Compute: $\mathcal{S} = \{x \in \mathbb{R}^n : \|T^{-1}x\|_1 < \frac{d_{\tilde{k}}}{2\|T\|_1}\}$
Compute: $\bar{\mathcal{X}}_1 = \mathcal{C}_1^{(\tilde{k})} + \mathcal{S}, \bar{\mathcal{X}}_2 = \mathcal{C}_2^{(\tilde{k})} + \mathcal{S}, \ldots, \bar{\mathcal{X}}_q = \mathcal{C}_q^{(\tilde{k})} + \mathcal{S}$
Choose: $\eta \in \mathbb{Z}_+$ such that $q^{\eta+1} > z$.
Compute: An enumeration $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{q^\eta}\}$ of the set $\mathcal{U}^\eta$, where $\mathbf{u}_j = (u_j^1, \ldots, u_j^\eta)$.
Compute: $\mathcal{X}_k = Bu_1 + ABu_2 + \cdots + A^{\eta-1}Bu_\eta + A^\eta\bar{\mathcal{X}}_i,\ 1 \leq k \leq q^{\eta+1}$, where $(u_1, \ldots, u_\eta) = \mathbf{u}_j$ for some $1 \leq j \leq q^\eta$, and $1 \leq i \leq q$.
Return: $\mathcal{X}_1, \ldots, \mathcal{X}_{q^{\eta+1}}$

Again, we claim that Algorithm 2 is guaranteed to generate a finite uniform bisimulation with many equivalence classes when the corresponding conditions are satisfied. In particular, we claim that the sets $\mathcal{X}_1, \ldots, \mathcal{X}_{q^{\eta+1}}$ returned by Algorithm 2 afford a finite uniform bisimulation on $\cup_{k=1}^{q^{\eta+1}} \mathcal{X}_k$ of system (221).

Corollary 2. Given system (221), and let the hypotheses in Theorem 14 hold, then for any $z \in \mathbb{Z}_+$, Algorithm 2 terminates, and returns a finite uniform bisimulation $\sim$ that has more than z equivalence classes.

Remark. These equivalence classes computed by Algorithm 2 can also be made arbitrarily

fine, that is to say, the diameter of each equivalence class can be made arbitrarily small (see the following derivation).

Proof. (of Theorem 14 and Corollary 2)

To show Theorem 14 and Corollary 2, it suffices to show that Algorithm 2 terminates, and that the sets $\mathcal{X}_1, \ldots, \mathcal{X}_{q^{\eta+1}}$:

$$\mathcal{X}_k = Bu_1 + ABu_2 + \cdots + A^{\eta-1}Bu_\eta + A^\eta \bar{\mathcal{X}}_i, \;\; 1 \leq k \leq q^{\eta+1}, \tag{251}$$

returned by Algorithm 2 afford a finite uniform bisimulation $\sim$ on $\cup_{k=1}^{q^{\eta+1}} \mathcal{X}_k$ of system (221). By Algorithm 2, the number of equivalence classes $q^{\eta+1}$ is guaranteed to be greater than z.

By assumption, $\{S_j^1\}_{j=1}^q$ (234) are disjoint. By Lemma 6, S_j^1 is also compact for all $j \in \{1, \ldots, q\}$. Since the distance between two disjoint compact sets is positive, we have

$$\min\{d(S_w^1, S_v^1) : w \neq v, 1 \leq w, v \leq q\} > 0.$$

Recall

$$\begin{aligned}
\mathcal{C}_1^{(k)} &= \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 = u_{(1)}, u_2, \ldots, u_k \in \mathcal{U}\}, \\
\mathcal{C}_2^{(k)} &= \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 = u_{(2)}, u_2, \ldots, u_k \in \mathcal{U}\}, \\
&\vdots \\
\mathcal{C}_q^{(k)} &= \{Bu_1 + ABu_2 + \cdots + A^{k-1}Bu_k : u_1 = u_{(q)}, u_2, \ldots, u_k \in \mathcal{U}\},
\end{aligned} \tag{252}$$

we observe that $\mathcal{C}_j^{(k)}$ (252) is a subset of S_j^1 for any $j \in \{1, \ldots, q\}$ and any $k \in \mathbb{Z}_+$, therefore $d_k = \min\{\|\alpha - \beta\|_1 : \alpha \in \mathcal{C}_v^{(i)}, \beta \in \mathcal{C}_w^{(i)}, w \neq v, 1 \leq w, v \leq q\}$ is uniformly bounded away from zero:

$$d_k \geq \min\{d(S_w^1, S_v^1) : w \neq v, 1 \leq w, v \leq q\} > 0, \; \forall \, k \in \mathbb{Z}_+. \tag{253}$$

Since l_k tends to zero as k tends to infinity, we see that Algorithm 2 terminates.

Recall

$$\begin{aligned}\bar{\mathcal{X}}_1 &= \mathcal{C}_1^{(\tilde{k})} + \mathcal{S},\\ \bar{\mathcal{X}}_2 &= \mathcal{C}_2^{(\tilde{k})} + \mathcal{S},\\ &\vdots\\ \bar{\mathcal{X}}_q &= \mathcal{C}_q^{(\tilde{k})} + \mathcal{S},\end{aligned} \tag{254}$$

we observe that $\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_q$ afford a finite uniform bisimulation on $\cup_{j=1}^{q} \bar{\mathcal{X}}_j$ of system (221) by the derivation of Theorem 16. We will use this observation to show that sets $\mathcal{X}_1, \ldots, \mathcal{X}_{q^{\eta+1}}$ (251) also afford a finite uniform bisimulation.

We first show that $\cup_{k=1}^{q^{\eta+1}} \mathcal{X}_k$ is an invariant set of system (221). For any $x \in \mathcal{X}_k$, by (251), we can write

$$x = Bu_1 + ABu_2 + \cdots + A^{\eta-1}Bu_\eta + A^\eta \bar{x}$$

for some $(u_1, \ldots, u_\eta) \in \mathcal{U}^\eta$ and some $\bar{x} \in \bar{\mathcal{X}}_i$ with $1 \leq i \leq q$. Then for any $u \in \mathcal{U}$,

$$Ax + Bu = Bu + ABu_1 + A^2Bu_2 + \cdots + A^{\eta-1}Bu_{\eta-1} + A^\eta(A\bar{x} + Bu_\eta).$$

Since $\cup_{j=1}^{q} \bar{\mathcal{X}}_j$ is an invariant set of system (221), we have $(A\bar{x} + Bu_\eta) \in \bar{\mathcal{X}}_j$ for some $1 \leq j \leq q$. Recall (251), we see that $(Ax + Bu) \in \mathcal{X}_k$ for some $1 \leq k \leq q^{\eta+1}$, and therefore $\cup_{k=1}^{q^{\eta+1}} \mathcal{X}_k$ is an invariant set of system (221).

Next, we use an inductive approach to show that the sets $\mathcal{X}_k$, $k = 1, \ldots, q^{\eta+1}$ (254) are disjoint. Write $\mathcal{U} = \{u_{(1)}, \ldots, u_{(q)}\}$, we observe that the q^2 sets $Bu_{(i)} + A\bar{\mathcal{X}}_j$, $i = 1, \ldots, q$, $j = 1, \ldots, q$ are disjoint. Indeed, consider any $Bu_{(i^1)} + A\bar{\mathcal{X}}_{j^1}$ and $Bu_{(i^2)} + A\bar{\mathcal{X}}_{j^2}$ with $(i^1, j^1) \neq (i^2, j^2)$. If $i^1 = i^2$, then $j^1 \neq j^2$. Since $\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_q$ are disjoint, we have $\bar{\mathcal{X}}_{j^1} \cap \bar{\mathcal{X}}_{j^2} = \varnothing$. Since A is invertible by assumption, we have $A\bar{\mathcal{X}}_{j^1} \cap A\bar{\mathcal{X}}_{j^2} = \varnothing$, and therefore $(Bu_{(i^1)} + A\bar{\mathcal{X}}_{j^1}) \cap (Bu_{(i^2)} + A\bar{\mathcal{X}}_{j^2}) = \varnothing$. If $i^1 \neq i^2$, from the second part of the derivation of Theorem 16 (equation (248) through (250)) and the construction of $\bar{\mathcal{X}}_j$ (252), (254), we see that $(Bu_{(i^1)} + A\bar{\mathcal{X}}_{j^1}) \subset \bar{\mathcal{X}}_{i^1}$ and $(Bu_{(i^2)} + A\bar{\mathcal{X}}_{j^2}) \subset \bar{\mathcal{X}}_{i^2}$. Since $\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_q$

are disjoint, we have $\bar{\mathcal{X}}_{i^1} \cap \bar{\mathcal{X}}_{i^2} = \varnothing$, and therefore $(Bu_{(i^1)} + A\bar{\mathcal{X}}_{j^1}) \cap (Bu_{(i^2)} + A\bar{\mathcal{X}}_{j^2}) = \varnothing$. We conclude that the sets $Bu_{(i)} + A\bar{\mathcal{X}}_j$, $i = 1, \ldots, q$, $j = 1, \ldots, q$ are disjoint, where $\mathcal{U} = \{u_{(1)}, \ldots, u_{(q)}\}$.

For the ease of exposition, we use $\mathcal{X}_j^1$, $j = 1, \ldots, q^2$ to denote the q^2 disjoint sets $Bu_{(i)} + A\bar{\mathcal{X}}_j$, $i = 1, \ldots, q$, $j = 1, \ldots, q$. We observe that the q^3 sets $Bu_{(i)} + A\mathcal{X}_j^1$, $i = 1, \ldots, q$, $j = 1, \ldots, q^2$ are also disjoint. Indeed, consider any $Bu_{(i^1)} + A\mathcal{X}_{j^1}^1$ and $Bu_{(i^2)} + A\mathcal{X}_{j^2}^1$ with $(i^1, j^1) \neq (i^2, j^2)$. If $i^1 = i^2$, then $j^1 \neq j^2$. Since $\mathcal{X}_j^1$, $j = 1, \ldots, q^2$ are disjoint, we have $\mathcal{X}_{j^1}^1 \cap \mathcal{X}_{j^2}^1 = \varnothing$. Since A is invertible by assumption, we have $A\mathcal{X}_{j^1}^1 \cap A\mathcal{X}_{j^2}^1 = \varnothing$, and therefore $(Bu_{(i^1)} + A\mathcal{X}_{j^1}^1) \cap (Bu_{(i^2)} + A\mathcal{X}_{j^2}^1) = \varnothing$. If $i^1 \neq i^2$, by the preceding paragraph, we see that $\mathcal{X}_{j^1}^1 \subset \bar{\mathcal{X}}_l$ for some $1 \leq l \leq q$, and therefore

$$(Bu_{(i^1)} + A\mathcal{X}_{j^1}^1) \subset (Bu_{(i^1)} + A\bar{\mathcal{X}}_l) \subset \bar{\mathcal{X}}_{i^1}.$$

Similarly, we see that $(Bu_{(i^2)} + A\mathcal{X}_{j^2}^1) \subset \bar{\mathcal{X}}_{i^2}$. Since $\bar{\mathcal{X}}_1, \ldots, \bar{\mathcal{X}}_q$ are disjoint, we have $\bar{\mathcal{X}}_{i^1} \cap \bar{\mathcal{X}}_{i^2} = \varnothing$, and therefore $(Bu_{(i^1)} + A\mathcal{X}_{j^1}^1) \cap (Bu_{(i^2)} + A\mathcal{X}_{j^2}^1) = \varnothing$. We conclude that the sets $Bu_{(i)} + A\mathcal{X}_j^1$, $i = 1, \ldots, q$, $j = 1, \ldots, q^2$ are disjoint.

Repeating this argument η times, we conclude that the $q^{\eta+1}$ sets $\mathcal{X}_k$, $k = 1, \ldots, q^{\eta+1}$ (254) are disjoint.

Next, we define an equivalence relation $\sim$ on $\cup_{k=1}^{q^{\eta+1}} \mathcal{X}_k$ as

$$x \sim y \iff x \in \mathcal{X}_k \text{ and } y \in \mathcal{X}_k \text{ for some } 1 \leq k \leq q^{\eta+1}.$$

We claim that $\sim$ is a finite uniform bisimulation. Indeed, for any $1 \leq k \leq q^{\eta+1}$, by (251), write $\mathcal{X}_k$ as $\mathcal{X}_k = Bu_1 + ABu_2 + \cdots + A^{\eta-1}Bu_\eta + A^\eta \bar{\mathcal{X}}_i$. Then for any $u \in \mathcal{U}$, $A\mathcal{X}_k + Bu = Bu + ABu_1 + A^2Bu_2 + \cdots + A^{\eta-1}Bu_{\eta-1} + A^\eta(A\bar{\mathcal{X}}_i + Bu_\eta)$. Since $(A\bar{\mathcal{X}}_i + Bu_\eta) \subset \bar{\mathcal{X}}_j$ for some $1 \leq j \leq q$, we have

$$(A\mathcal{X}_k + Bu) \subset (Bu + ABu_1 + A^2Bu_2 + \cdots + A^{\eta-1}Bu_{\eta-1} + A^\eta \bar{\mathcal{X}}_j) = \mathcal{X}_{k'}$$

for some $1 \leq k' \leq q^{\eta+1}$. Therefore (217) is satisfied. Since $q^{\eta+1}$ is finite, (218) is also satisfied. This completes the proof of Theorem 14 and Corollary 2.

Lastly, we comment on the fact that the diameter of the equivalence classes $\mathcal{X}_k$ can be made arbitrarily small. For any $1 \leq k \leq q^{\eta+1}$, we have

$$diam(\mathcal{X}_k) \leq \|A^\eta\|_1 diam(\mathcal{C}_i^{(\tilde{k})} + \mathcal{S}) \leq \|A^\eta\|_1 (diam(\mathcal{A}_s) + diam(\mathcal{S}))$$

Since A is Schur-stable, $diam(\mathcal{A}_s)$ is finite, and $\|A^\eta\|_1$ can be made arbitrarily small by choosing η large enough. $diam(\mathcal{S})$ is finite by construction, and we conclude that $diam(\mathcal{X}_k)$ can be made arbitrarily small by choosing η sufficiently large. □

7.5 Illustrative Examples

In this section, we present a set of illustrative examples: In Example 6, we illustrate the difference between the notion of finite uniform bisimulation and the notion of finite bisimulation stated in [34]; in Example 7, we show that additional assumptions, besides the conditions in Theorem 13, are needed to guarantee the existence of arbitrarily many equivalence classes; in Example 8, we show that the condition "$[0]$ is bounded" in Theorem 15 cannot be dropped; in Example 9, we illustrate the analytical result in Theorem 13, discuss how to construct a DFM approximation of the original system, and apply Algorithm 2 to construct many equivalence classes.

Example 6. (Example 2.14, [34]) Consider system (221) with parameters

$$A = \begin{bmatrix} 2 & 0 & -1 \\ -1 & -7 & 11 \\ 0 & 4 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

According to [34], a finite bisimulation with eight equivalence classes $\{q_1, \ldots, q_8\}$ is constructed. If we choose $x = [1\ -2\ -3]^T \in q_1$, $x' = [8, -18, -24]^T \in q_1$ and let input $u = [0\ 60]^T$, then $Ax + Bu = [125\ 40\ 34]^T \in q_2$, and $Ax' + Bu = [160\ -86\ -156]^T \in$

q_1. Therefore this finite bisimulation is not a “finite uniform bisimulation" as defined in Definition 6.

Example 7. Consider system (221) with parameters

$$A = \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{255}$$

and

$$\mathcal{U} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

We calculate, and plot $cl(\mathcal{A}_s)$:

$$cl(\mathcal{A}_s) = \{(x, y) \in \mathbb{R}^2 | x = 0, -2 \leq y \leq 2\} \cup \{(x, y) \in \mathbb{R}^2 | x = 1, -1 \leq y \leq 1\} \tag{256}$$

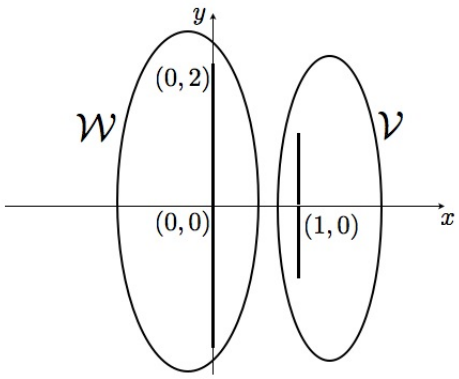


Figure 7: 2 and only 2 equivalence classes.

In the above figure, $\mathcal{W}$ and $\mathcal{V}$ represents a disconnection of $cl(\mathcal{A}_s)$. We see that both $cl(\mathcal{A}_s) \cap \mathcal{W}$ and $cl(\mathcal{A}_s) \cap \mathcal{V}$ are connected. Therefore, we cannot apply the analytical result in Theorem 13 to generate more than two equivalence classes, because such result relies on the disconnectedness of an invariant set.

Example 8. Given system (221) with parameters: $A = diag(\{2, 0.5\})$ (a diagonal matrix with diagonal entries 2 and 0.5), B is the identity matrix, and $\mathcal{U} = \{[0\ 0]^T\}$. Let $\mathcal{X}_1 =$

$\{(x,y) \in \mathbb{R}^2 : 1 < |y| < 2\}$, and $\mathcal{X}_2 = \{(x,y) \in \mathbb{R}^2 : |y| < 1\}$, then we see that $\mathcal{X}_1, \mathcal{X}_2$ afford a regular finite uniform bisimulation on $\mathcal{X}_1 \cup \mathcal{X}_2$, which is an invariant set, and $B_r(0) \subset [0]$ for $r = 0.5$, and $\rho(A) = 2 > 1$.

Example 9. Consider system (221) with parameters:

$$A = \begin{bmatrix} 0.25 & -0.15 \\ 0 & 0.1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{257}$$

and

$$\mathcal{U} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

Since A is diagonalizable, we have

$$A^n = \begin{bmatrix} (1/4)^n & (1/10)^n - (1/4)^n \\ 0 & (1/10)^n \end{bmatrix}, n = 0, 1, 2, \cdots$$

and we can show that $cl(\mathcal{A}_s)$ is a subset of:

$$\bigcup \{(\pm 1, \pm 1), (0,0)\} + \{(x,y) : x \in [-\frac{4}{9}, \frac{4}{9}],\ y \in [-\frac{1}{9}, \frac{1}{9}]\}$$

Therefore $cl(\mathcal{A}_s)$ is not connected.

By the derivation of Theorem 13, we find a finite uniform bisimulation $\sim$ on an invariant set of this system:

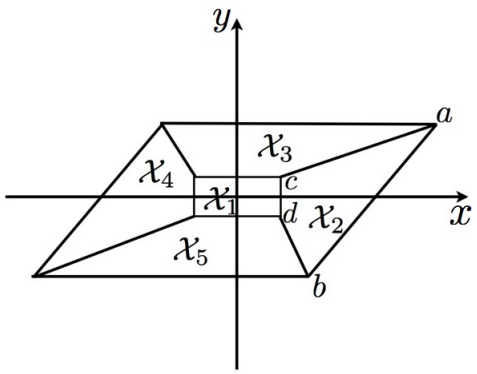


Figure 8: 2-d finite uniform bisimulation example.

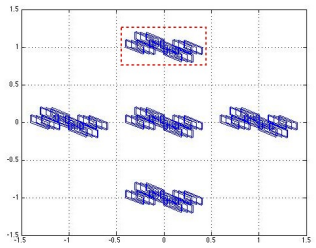
$\mathcal{X}_1, \ldots, \mathcal{X}_5$ shown in Figure 8 afford a finite uniform bisimulation $\sim$ on an invariant set $\mathcal{S} = \cup_{i=1}^{5} \mathcal{X}_i$ of system (221). The points a, b, c, d are given by:

$$a = (\frac{22}{9}, \frac{10}{9}), b = (\frac{10}{9}, -\frac{10}{9}), c = (\frac{4}{9}, \frac{1}{9}), d = (\frac{4}{9}, -\frac{1}{9})$$

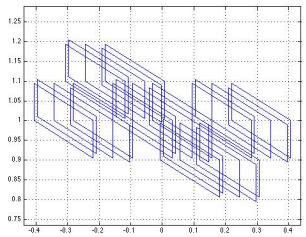
and Figure 8 is symmetric with respect to the origin. Particularly, the set $\mathcal{S}$ is the convex hull of points: $\{a, b, -a, -b\}$.

Given $\sim$, we can construct a DFM that is uniformly bisimilar to the original system. Particularly, we associate each equivalence class $\mathcal{X}_i$ to a discrete state q_i of the DFM, $i = 1, \ldots, 5$. The state transitions of the DFM can be determined based on (220): For instance, if the current state of the DFM is q_1, and the current input is $[0 \quad 1]^T$, then the next state of the DFM is q_3.

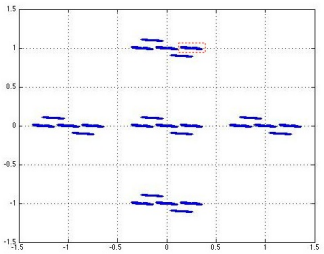
Since this example also satisfies the conditions in Theorem 14, we can also use Algorithm 2 to generate a finite uniform bisimulation with an arbitrarily large number of equivalence classes. In particular, we generate two finite uniform bisimulations with 5 equivalence classes, and 25 equivalence classes respectively.



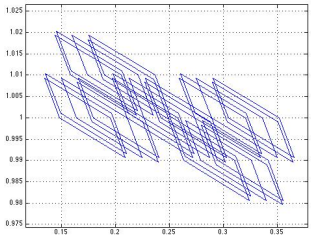
(a) 5 classes.



(b) Zoom in on 1 class.



(c) 25 classes.



(d) Zoom in on 1 class.

Figure 9: Finite uniform bisimulations with many equivalence classes.

In the above, Figure 9a shows the 5 equivalence classes generated by Algorithm 2, and Figure 9b shows one particular equivalence class (the boxed rectangular area in Figure 9a). Similarly Figure 9c shows the 25 equivalence classes, and Figure 9d shows one particular equivalence class. As shown in Figure 9b and Figure 9d, an equivalence class computed by Algorithm 2 is the union of all the polytopes (in this case parallelograms). This is in accordance with the construction of the equivalence classes.

7.6 Summary

In this section we propose notions of finite uniform bisimulation and regular finite uniform bisimulation. We then present a sufficient condition for the existence of finite uniform bisimulations: If the forced response of a Schur stable system is not connected, then the system admits a finite uniform bisimulation. In this case, we construct an algorithm to compute finite uniform bisimulations. Furthermore, we discuss the existence and construction of an arbitrarily large number of equivalence classes. We also present a necessary condition for the existence of regular finite uniform bisimulation. Future works include closing

the gap between necessary conditions and sufficient conditions, and extending the current result to systems with more general dynamics.

8 Conclusions and Future Work

8.1 Conclusions

In this dissertation, we motivate the need for and formulate a notion of observability of systems over finite alphabets in the sense of how well the output of the system can be estimated based on past input and output information. We characterize this proposed notion by deriving both necessary and sufficient conditions of observability in terms of system parameters. For system (2), such conditions involve both the dynamics of the underlying LTI system and the discontinuous points of the quantizer. Based on this notion of observability, we propose a control design problem which has the flavor of predictive control and reachability analysis for systems (2) with constraints on the system state. We also discuss a new construction of DFM observers, their connections to existing results on DFM approximations, and we study conditions under which an existing construct for finite state approximation can be simplified. Finally, we apply our results to address a relevantly open problem in the theory of bisimulation, bringing in a topological approach.

8.2 Directions for Future Work

An immediate direction is to continue charactering the proposed notions of observability with an eye on systems with more general dynamics. A more interesting direction would be further developing the observability analysis in order to address applications in control design and compare our approach with that of contemporary researchers.

Intended to be blank.

9 References

[1] D. F. Delchamps, "Stabilizing a linear system with quantized state feedback," *IEEE Transactions on Automatic Control*, vol. 35, no. 8, pp. 916–924, 1990.

[2] R. W. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1279–1289, 2000.

[3] G. Lafferriere, G. J. Pappas, and S. Sastry, "O-minimal hybrid systems," *Mathematics of Control, Signals and Systems*, vol. 13, no. 1, pp. 1–21, 2000.

[4] P. Tabuada, *Verification and control of hybrid systems: a symbolic approach*. Springer, 2009.

[5] A. Tanwani, H. Shim, and D. Liberzon, "Observability for switched linear systems: Characterization and observer design," *IEEE Transactions on Automatic Control*, vol. 58, no. 4, pp. 891–904, 2013.

[6] B. Yordanov, J. Tumová, I. Cerná, J. Barnat, and C. Belta, "Temporal logic control of discrete-time piecewise affine systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 6, pp. 1491–1504, 2012.

[7] D. C. Tarraf, A. Megretski, and M. A. Dahleh, "A framework for robust stability of systems over finite alphabets," *IEEE Transactions on Automatic Control*, vol. 53, no. 5, pp. 1133–1146, 2008.

[8] J. P. Hespanha, *Linear Systems Theory*. Princeton University Press, 2009.

[9] K. S. Narendra and A. M. Annaswamy, *Stable adaptive systems*. Courier Dover Publications, 2012.

[10] D. C. Tarraf, "A control-oriented notion of finite state approximation," *IEEE Transactions on Automatic Control*, vol. 57, no. 12, pp. 3197–3202, 2012.

[11] D. C. Tarraf, "Finite approximations of switched homogeneous systems for controller synthesis," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1140–1145, 2011.

[12] D. C. Tarraf, "An input-output construction of finite state ρ/μ approximations for control design," *IEEE Transactions on Automatic Control, Special Issue on Control of Cyber-Physical Systems*, vol. 59, no. 12, pp. 3164–3177, 2014.

[13] R. Hermann and A. J. Krener, "Nonlinear controllability and observability," *IEEE Transactions on Automatic Control*, vol. 22, no. 5, pp. 728–740, 1977.

[14] A. Balluchi, L. Benvenuti, M. D. Di Benedetto, and A. L. Sangiovanni-Vincentelli, "Observability for hybrid systems," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, (Maui, HI), pp. 1159–1164, 2003.

[15] G. Xie and L. Wang, "Necessary and sufficient conditions for controllability and observability of switched impulsive control systems," *IEEE Transactions on Automatic Control*, vol. 49, no. 6, pp. 960–966, 2004.

[16] D. F. Delchamps, "Extracting state information from a quantized output record," *Systems & Control Letters*, vol. 13, pp. 365–372, 1989.

[17] J. Sur and B. Paden, "Observers for linear systems with quantized outputs," in *Procedings of the American Control Conference*, (Albuquerque, NM), pp. 3012–3016, June 1997.

[18] J. Raisch, "Controllability and observability of simple hybrid control systems-FDLTI plants with symbolic measurements and quantized control inputs," in *International Conference on Control'94*, vol. 1, (Coventry, UK), pp. 595–600, 1994.

[19] D. Delvecchio, R. M. Murray, and E. Klavins, "Discrete state estimators for systems on a lattice," *Automatica*, vol. 42, no. 2, pp. 271–285, 2006.

[20] R. Ehlers and U. Topcu, "Estimator-based reactive synthesis under incomplete information," in *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, (Seattle, WA), pp. 249–258, 2015.

[21] O. Mickelin, N. Ozay, and R. M. Murray, "Synthesis of correct-by-construction control protocols for hybrid systems using partial state information," in *Proceedings of the American Control Conference, 2014*, (Portland, OR), pp. 2305–2311, 2014.

[22] W. K. Nicholson, *Introduction to Abstract Algebra*. Wiley, 2012.

[23] N. L. Carothers, *Real Analysis*. Cambridge University Press, 1999.

[24] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[25] D. Fan and D. C. Tarraf, "On finite memory observability of a class of systems over finite alphabets with linear dynamics," in *Proceedings of the 53rd IEEE Conference on Decision and Control*, (Los Angeles, CA), pp. 3884–3891, 2014.

[26] R. Bronson, *Matrix Methods: An Introduction*. Academic Press, 2014.

[27] E. G. Gilbert and K. T. Tan, "Linear systems with state and control constraints: The theory and application of maximal output admissible sets," *IEEE Transactions on Automatic Control*, vol. 36, no. 9, pp. 1008–1020, 1991.

[28] F. Blanchini and S. Miani, *Set-theoretic methods in control*. Springer Science & Business Media, 2007.

[29] D. Fan and D. C. Tarraf, "On initialization of finite state rho / mu approximations of systems over finite alphabets," in *Proceedings of the 54th IEEE International Conference on Decision and Control*, (Osaka, Japan), December 2015.

[30] F. Aalamifar and D. Tarraf, "An iterative algorithmic implementation of input-output finite state approximations," in *Proceedings of the 51st IEEE International Conference on Decision and Control*, (Maui, HI), pp. 6735–6741, December 2012.

[31] E. M. Stein and R. Shakarchi, *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

[32] D. Fan and D. C. Tarraf, "On existence of finite uniform bisimulations for linear systems with finite input alphabets," in *Proceedings of the 54th IEEE International Conference on Decision and Control*, (Osaka, Japan), December 2015.

[33] D. Fan and D. C. Tarraf, "Finite uniform bisimulations for linear systems with finite input alphabets," *arXiv preprint arXiv:1510.04209*, 2015. (Under review for journal publication, manuscript available at http://arxiv.org/abs/1510.04209.).

[34] P. Tabuada and G. J. Pappas, "Linear time logic control of discrete-time linear systems," *IEEE Transactions on Automatic Control*, vol. 51, no. 12, pp. 1862–1877, 2006.

[35] E. A. Gol, X. Ding, M. Lazar, and C. Belta, "Finite bisimulations for switched linear systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3122–3134, 2014.

[36] R. Alur, T. Henzinger, G. Lafferriere, and G. Pappas, "Discrete abstractions of hybrid systems," *Proceedings of the IEEE*, vol. 88, no. 7, pp. 971–984, 2000.

CURRICULUM VITAE

Donglei Fan was born in China in 1987. He received the B.E. degree from Tsinghua University, Beijing, in 2010, and the M.S. degree from California Institute of Technology, Pasadena, CA, in 2012. He started his doctoral studies in the ECE department of Johns Hopkins University, Baltimore, MD, in 2012. He is currently a PhD candidate. His research interests include observability and state estimation of hybrid systems (specifically quantized systems), finite state controller synthesis, and problems on system abstraction.