SIAM J. CONTROL and OPTIMIZATION
Vol. 31, No. 5, pp. 1245–1266, September 1993

010

LINEAR SYSTEMS WITH SIGN-OBSERVATIONS*

RENÉE KOPLON† AND EDUARDO D. SONTAG‡

Abstract. This paper deals with systems that are obtained from linear time-invariant continuous- or discrete-time devices followed by a function that just provides the sign of each output. Such systems appear naturally in the study of quantized observations as well as in signal processing and neural network theory. Results are given on observability, minimal realizations, and other system-theoretic concepts. Certain major differences exist with the linear case, and other results generalize in a surprisingly straightforward manner.

Key words. observability, minimal realization, neural networks, quantization effects

AMS subject classifications. 93B07, 93B10, 93B15

1. Introduction.

A central issue in current control theory and signal processing concerns the interface between, on the one hand, the continuous, physical, world and, on the other hand, discrete devices such as digital computers, capable of symbolic processing. Classical control techniques, especially for linear systems, have proved spectacularly successful in automatically regulating relatively simple systems. However, for large-scale problems, controllers resulting from the application of the well-developed theory are used as building blocks of more complex systems. The integration of these systems is often accomplished by means of ad hoc techniques that combine pattern recognition devices, various types of switching controllers, and humans—or, more recently, expert systems—in supervisory capabilities.

Recently, there has been renewed interest in the formulation of mathematical models in which this interface between the continuous and the symbolic is naturally accomplished and system-theoretic questions can be formulated and resolved for the resulting models. Successful approaches will eventually allow the interplay of modern control theory with automata theory and other techniques from computer science. This interest has motivated much research into areas such as discrete-event systems, supervisory control, and, more generally, "intelligent control systems."

One possible first step in a systematic attack of this problem is the study of partial (discrete) measurements on the state of a continuous dynamical system. When no controls are present, this is closely related to classical work on symbolic dynamics, and in fact has been pursued in the control theory literature, where Ramadge studied in [9] the dynamical behavior of observation sequences corresponding to such systems.

If inputs are available, one of the first questions that we may address in this context is that of the nature of the information that can be deduced by a symbolic "supervisor" from data transmitted by such a "lower level" continuous device, using appropriate controls to obtain more information about the system. Here the work of Delchamps, especially in [3]–[5], is especially relevant. His work dealt with what we may call "single-experiment observability" of *constrained-output systems*, systems for which the dynamics are linear but the outputs reflect various limitations of measuring

* Received by the editors June 17, 1991; accepted for publication (in revised form) September 8, 1992. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-91-0343.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (koplon@hilbert.rutgers.edu).

‡ Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (sontag@hilbert.rutgers.edu).

devices. These are systems, in discrete or continuous time, whose equations can be expressed as

$$\begin{aligned} x(t+1)\ [\text{or}\ \dot{x}(t)\,] &= Ax(t)+Bu(t), \\ y(t) &= \sigma(Cx)\,, \end{aligned} \tag{1}$$

for some $n \times n$ real matrix A, $n \times m$ matrix B, and $p \times n$ matrix C, and where σ is a memory-free map: $\mathbb{R}^p \to \mathbb{R}^p$—in the case of Delchamp's work, a quantizer. (The simplest example of a constrained-output system occurs if σ is the identity. Then we are dealing with the class of all finite-dimensional linear systems. See [12] for precise definitions of "system" and related terms.) Models of the form (1) with quantizer σ arise also in a variety of other areas besides control. For instance, in signal processing, when modeling linear channels transmitting digital data from a quantized source, the channel equalization problem becomes one of systems inversion for such systems; see [2] and also the related paper [8].

In contrast to Delchamp's work, in this paper we look at the more standard notion of multiple-experiment observability, which is different for nonlinear systems from the single-experiment concept (for purely linear systems, both concepts do coincide, of course). We will be especially interested in the case in which σ simply takes the sign of each coordinate, that is, *sign-linear systems*, those for which

$$\sigma(x) = \text{sign}\,(x)$$

(applied to each coordinate independently), where

$$\text{sign}\,(x) = \begin{cases} 1 & \text{if} \quad x > 0, \\ 0 & \text{if} \quad x = 0, \\ -1 & \text{if} \quad x < 0\,. \end{cases}$$

Sign-linear systems correspond to the 1-bit quantization case of Delchamps' model and are also motivated by pattern recognition applications (see below), but many technical results will be given in the paper in somewhat more generality.

Among the most popular techniques in pattern classification are those based on the use of *perceptrons* or linear discriminants (see, e.g., [6], [13]). Mathematically, these are simply functions of the type

$$\mathbb{R}^n \mapsto \mathbb{R}^p, \qquad v \mapsto y = \text{sign}\,(Cv),$$

typically with large n and small p; again, the sign is understood as being taken in each coordinate separately. Perceptrons are used to classify input patterns $v = (v_1, \ldots, v_n)$ into classes, and they form the basis of many statistical techniques. In many practical situations arising in speech processing or learning finite automata and languages (see, e.g., [7]), the vector v really represents a finite window

$$u(t-1), \ldots, u(t-s) \tag{2}$$

of a sequence of m-dimensional inputs $u(1), u(2), \ldots$, where the components of (2) have been listed as v (and $sm = n$). In that case, the perceptron can be understood as a sign-linear system of dimension n, with a shift-register used to store the previous inputs (2). Borrowing from the signal processing terminology, perceptrons are "finite impulse response" sign-linear systems. As such, they are not suited to modeling time dependencies and recurrences in the data. More general sign-linear systems are called

for, and this motivated the introduction of such systems in [1], using the name "infinite impulse response" again by analogy to the classical linear case. In that paper, the authors studied practical problems of systems identification but did not address the more system-theoretic types of questions with which this paper deals.

As a final reason for studying sign-linear systems, we point out that such systems provide a natural class of nonsmooth nonlinear systems, a class that combines logical and switching devices together with more classical continuous variables. When the nonlinearities appear in the feedback loops, the problems become far more difficult; in that context, see, for instance, [11] for results about the computational power of systems of the type $x(t+1) = \sigma(Ax(t) + Bu(t))$.

1.1. Summary of paper. As mentioned earlier, the focus of this paper is the class of *sign-linear systems*, that is, those of the type (1) with $\sigma(x) = \text{sign}\,(x)$ (the sign is understood as being taken in each coordinate, so that the output value space could be taken simply as $\{-1, 0, 1\}^p$; careful definitions are given later). Also of interest are the associated *sign-linear input/output* (*i/o*) *maps* of the form

$$y(t) = \text{sign}\,(\mathcal{A}_1 u(t) + \cdots + \mathcal{A}_t u(1))$$

or the analogous continuous-time maps (convolution followed by sign).

For a system such as (1), we call any triple (A, B, C) such that the equations of the system can be expressed in terms of that triple, *a triple associated to* the system Σ. Note that in some cases there may be many triples associated to a single system:

Example 1.1. Let Σ be a sign-linear system with state-space $\mathbb{R}^2$ defined by the equations

$$x(t+1) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u(t),$$
$$y(t) = \text{sign}\,(x_1 + 2x_2).$$

Then the following triples are both associated to Σ:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 2 \end{pmatrix},$$
$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad C = \begin{pmatrix} 2 & 4 \end{pmatrix}.$$

The results that we describe parallel those known for standard linear systems, but with a few, perhaps unexpected, differences. We may summarize the main conclusions on sign-linear systems as follows:

(a) Although stronger than just observability of the pair (A, C), observability of sign-linear systems can be characterized in an elegant manner. The characterization is different in the continuous- and discrete-time cases, in contrast to what happens for linear systems. Moreover, in another characteristic that is typical of *non*linear systems, the degree of controllability of the system does affect observability.

(b) Minimal-dimensional realizations of sign-linear i/o maps by sign-linear systems are unique up to a change of variables in the state space and a positive rescaling of outputs. This is basically as in the linear case, except for the obvious need to rescale. Moreover, finite-dimensional realizability can be characterized in the usual manner using Hankel matrices.

(c) If a realization of a given sign-linear i/o map is controllable and observable, in the usual sense of control theory, then it is minimal. Conversely, a minimal realization is necessarily final-state observable (that is, there is a control allowing for determination of the state at the end of the interval of application) but, in the discrete-time case, minimal realizations may not be observable. (In continuous-time, final-state and plain —initial-state— observability coincide.)

(d) Because of the possible lack of observability of minimal realizations, for some discrete-time sign-linear i/o maps it is the case that the abstract "canonical" realization, known to exist from automata-theoretic arguments, is not given by a sign-linear system. We discuss the canonical systems that result when minimal realizations are not observable, obtaining a description in terms of cascades of finite automata and linear systems.

The paper ends in § 6, where we show how some of the continuous-time observability results can be seen as consequences of the corresponding discrete-time results by sampling at appropriate frequencies.

Some of the results to be given can be stated in more generality, in terms of constrained-output systems as in (1), where σ is a fixed nonlinearity satisfying some (or all) of the following axioms:

1. $\operatorname{sign}(\sigma(x)) = \operatorname{sign}(x)$.
2. Finite precision sensor: $\sigma(x)$ =constant for $x \in (0, \varepsilon]$ and $x \in [-\varepsilon, 0)$, for some $\varepsilon > 0$.
3. Sensor saturation: $\sigma(x)$ =constant for $x > K > 0$ and $x < -K < 0$, for some $K > 0$.
4. $\sigma(x)$ is not constant on $(0, \infty)$ or $(-\infty, 0)$.

(Again, for a vector $z \in \mathbb{R}^p$, the notation $\sigma(z)$ denotes the vector $(\sigma(z_1), \ldots, \sigma(z_p)) \in \mathbb{R}^p$.) The main systems of interest in this paper, sign-linear ones, are those for which $\sigma(x) = \operatorname{sign}(x)$, which satisfies axioms 1,2,3. Some other examples of constrained-output systems are as follows:

- Output-saturated systems (satisfying 1,3,4) are those with $\sigma(x) = s(x)$, where

$$s(x) = \begin{cases} 1 & \text{if} \quad x > 1, \\ x & \text{if} \quad |x| \leq 1, \\ -1 & \text{if} \quad x < -1. \end{cases}$$

  The output space for output-saturated systems is $[-1, 1]^p$.
- Quantized systems (satisfying 1,2,4) are defined by $\sigma(x) = \lfloor x \rfloor$, with output space $\mathbb{Z}^p$.
- Saturated-quantized systems (satisfying 1–4) are systems for which

$$\sigma(x) = \begin{cases} K \operatorname{sign}(x) & \text{if} \quad |x| > K, \\ \lfloor x \rfloor & \text{if} \quad |x| < K \end{cases}$$

  for some fixed $K > 0$. Saturated-quantized systems have output space $\{n \in \mathbb{Z} : |n| \leq K\}^p$.

More details about such functions, and results specific for some of these classes, are described in [10].

2. Observability. Our notion of observability is the usual concept of multiple-experiment observability. Let us recall the main ideas. For formal definitions, please refer to [12, §5.1].

DEFINITION 2.1. A system Σ is *observable* if for any two initial states, there is some control that produces different outputs for each of the two initial states. This is not in general equivalent to single-experiment observability in which there exists one control function (or sequence of controls) that distinguishes any pair of states. The control we use to distinguish two states may depend on the two given states. This concept of observability really tells us only that we may *distinguish* between any two initial states, not that we may *determine* the initial state using one special control. For linear systems, multiple- and single-experiment observability are equivalent. If a linear system is observable, then the zero control will distinguish any pair of states.

DEFINITION 2.2. A system Σ is *final-state observable* if for any two initial states, there is some control and some time T so that either the output before time T is different for each of the two states, or the states at time T are the same. For continuous-time systems, final-state observability is equivalent to observability ([12, Prop. 5.1.9]).

We now state a few general necessary conditions for observability of constrained-output systems. Later, we will provide necessary and sufficient conditions for the class of sign-linear systems. The following result is obvious.

LEMMA 2.3. *If Σ is an observable constrained-output system, then (A, C) is an observable pair.*

Conversely, if σ is one-to-one, then observability of the pair (A, C) implies oberv-ability of Σ, but in general the implication does not hold. The following lemma gives an additional necessary condition when σ is not one-to-one.

LEMMA 2.4. *If Σ is an observable discrete-time constrained-output system with a single output channel ($p = 1$), and σ is not one-to-one, then* $\det A \neq 0$.

Proof. Suppose $\det A = 0$. Then there exists a nonzero $x \in \ker A$. The output sequence for the initial state x is $\{\sigma(Cx), \ldots\}$ where the part not shown is independent of x. Since Σ is observable, (A, C) is an observable pair, so $Cx \neq 0$. Let $\sigma(\mu) = \sigma(\nu)$, $\mu \neq \nu$. Then, we may choose $\alpha_1 \neq \alpha_2$ so that $\alpha_1 Cx = \mu$ and $\alpha_2 Cx = \nu$. Then $\alpha_1 x \neq \alpha_2 x$ are indistinguishable, contradicting observability. □

For $p > 1$, this lemma is not necessarily true. Consider the following counterexample. We will use the notation $x^+(t)$ to mean $x(t+1)$, and we drop the argument t from now on.

Example 2.5. Let Σ be the system with equations $x^+ = 0$, $y_1 = \sigma(x)$, $y_2 = \sigma(2x)$; and

$$\sigma(x) = \begin{cases} x & x \notin [1, 2] \\ 1 & x \in [1, 2] \end{cases}.$$

The nonlinearity σ is not one-to-one, but the map $x \mapsto (\sigma(x), \sigma(2x))$ is one-to-one, so the system is observable. However, A is not invertible.

If the measurement limiter σ is some form of saturation or σ has finite precision near 0, observability does imply that A is invertible, even in the multiple output case, since the following lemma will apply.

LEMMA 2.6. *If Σ is an observable discrete-time constrained-output system and σ either models sensor saturation*

$$\sigma(x) = \textit{constant for } x > K > 0 \textit{ and } x < -K < 0$$

for some $K > 0$, or has finite precision

$$\sigma(x) = \textit{constant for } x \in (0, \varepsilon] \textit{ and } x \in [-\varepsilon, 0)$$

for some $\varepsilon > 0$, *then* $\det A \neq 0$.

Proof. If $\det A = 0$, then there exists a nonzero $x \in \ker A$. In the saturated case, choose λ so that for all i satisfying $C_i x \neq 0$, then $|\lambda C_i x| > K$. Then λx and $2\lambda x$ are indistinguishable. In the finite precision case, choose λ so that $|\lambda C_i x| \leq \varepsilon$ for all $i = 1, \ldots, p$. Then λx and $\frac{1}{2}\lambda x$ are indistinguishable. □

Before stating the next lemma we introduce the following definition.

DEFINITION 2.7. Let Σ be a constrained-output system and let (A, B, C) be any triple associated to Σ. Then the sequence of $p \times m$ matrices

$$\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \ldots\},$$

where

$$\mathcal{A}_i = CA^{i-1}B, \quad i = 1, 2, 3, \ldots$$

is called a *Markov parameter sequence* associated to Σ. Since in general C is not uniquely defined from the system equations, there may be more than one Markov sequence associated to a given system; this issue is discussed later.

LEMMA 2.8. *Assume that* Σ *is a single-output observable discrete-time constrained-output system defined by the triple* (A, B, C), *and* σ *has finite precision*

$$\sigma(x) = \textit{constant for } x \in (0, \varepsilon] \textit{ and } x \in [-\varepsilon, 0).$$

If A *has an eigenvalue* λ *satisfying* $|\lambda| \leq 1$, *then* $\mathcal{A} \not\equiv 0$, *for any Markov sequence associated to* Σ.

Proof. Let $\mathcal{A}$ be any Markov sequence associated to Σ. Let v be a nonzero eigenvector for A corresponding to λ and let $\gamma = \|v\|$ (where $\|\cdot\|$ denotes Euclidean norm). Then $A^k v = \lambda^k v$ for all k, so $\|A^k v\| \leq \gamma$ for all k. Write $v = v_1 + iv_2$, where v_1, v_2 are real vectors. Then $\|A^k v_1\| \leq \|A^k v\| \leq \gamma$ for all k. Note that $\|C\| \neq 0$ since (A, C) is an observable pair. Then

$$x := \frac{v_1 \varepsilon}{\gamma \|C\|}$$

satisfies

$$|CA^k x| \leq \|C\| \|A^k x\| \leq \frac{\|C\| \gamma \varepsilon}{\gamma \|C\|} = \varepsilon$$

for all k. If $\mathcal{A} \equiv 0$, then $x, \frac{1}{2}x$ are indistinguishable, contradicting observability. □

3. Sign-linear systems. Now we concentrate on the observability of sign-linear systems. Sign-linear input/output maps and their realizations will be discussed in §§ 4 and 5.

DEFINITION 3.1. A *sign-linear system* Σ is a system with state, input, and output-value spaces $\mathbb{R}^n$, $\mathbb{R}^m$, and $\{-1, 0, 1\}^p$, respectively, for which there exist matrices $A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{p \times n}$, so that the equations of Σ take the form

$$\begin{aligned} x^+ (\text{or } \dot{x}) &= Ax + Bu, \\ y &= \operatorname{sign}(Cx) \end{aligned}$$

in discrete- (or continuous-) time. If (A, B, C) is a triple like this, we denote $\Sigma = (A, B, C)_s$. Whether we are dealing with discrete- or continuous-time will be clear

from the context. The integer n is the *dimension* of the system. It is convenient to include the degenerate case $n = 0$, corresponding to the system with zero-dimensional state space.

Note that $(A, B, C)_s = (\hat{A}, \hat{B}, \hat{C})_s$ if and only if $A = \hat{A}$, $B = \hat{B}$, and $C = \Lambda\hat{C}$, where Λ is a *scaling matrix* in the following sense.

DEFINITION 3.2. A $p \times p$ *scaling matrix* is a matrix of the type

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix},$$

where $\lambda_i > 0$, $i = 1, \ldots, p$. Any triple for which $\Sigma = (A, B, C)_s$ will be said to be *associated* to Σ. Observe that the properties of (A, B) being a controllable pair, (A, C) being an observable pair, and (A, B, C) being canonical (controllable and observable), in the usual linear systems sense, are independent of which of the associated triples is considered.

The following trivial observation will be used often.

Remark 3.3. If $\mathcal{H}$ is a real pre-Hilbert space (that is, a space with a nondegenerate inner product), and if $c \in \mathcal{H}, c$ nonzero, $a, b \in \mathbb{R}, a \neq b$, then there is a $u \in \mathcal{H}$ so that

$$\operatorname{sign}(a + \langle c, u\rangle) \neq \operatorname{sign}(b + \langle c, u\rangle).$$

Indeed, without loss of generality, we may assume that $a > b$. Let

$$\alpha := -\frac{a + b}{2\|c\|^2}.$$

Then $u := \alpha c$ satisfies $a + \langle c, u\rangle > 0$ and $b + \langle c, u\rangle < 0$.

Since controllability of a system does not depend on outputs, a sign-linear system is controllable if and only if (A, B) is a controllable pair in the usual sense ([12]). Observability requires a bit more than in the linear case, as illustrated by the system

$$x(t + 1) = u(t), \quad y(t) = \operatorname{sign}(x(t)).$$

Observability of the pair (A, C) is not sufficient to guarantee observability of the corresponding sign-linear system.

We will say that a triple (A, B, C) *has property* $\mathcal{P}$ if not only is (A, C) observable, but we can choose a subset of outputs which allow observability of the pair (A, C) and for each of which the corresponding row of the Markov sequence $\mathcal{A}$ is nonzero. In the case $p = 1$, this just means that (A, C) is an observable pair and $\mathcal{A} \not\equiv 0$, or equivalently, that (A, C) is an observable pair and $B \neq 0$. For $p > 1$, $\mathcal{A}$ has p rows and we only require that enough of those rows are nonzero. More precisely, let

$$I(\mathcal{A}) = \{i_1, \ldots, i_k\}$$

be the indices of the nonzero rows of $\mathcal{A}$; then property $\mathcal{P}$ is the condition that

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ q = 0, \ldots, n-1}} \ker(C_j A^q) = \{0\}, \tag{3}$$

where C_j denotes the jth row of C. Note that if C and $\hat{C}$ differ only by multiplication by a scaling matrix, property $\mathcal{P}$ holds for (A, B, C) if and only if it holds for $(A, B, \hat{C})$.

Thus there is no ambiguity in the following statements. For discrete- and continuous-time the following theorems state necessary and sufficient conditions for observability.

THEOREM 3.4. *Let $\Sigma = (A, B, C)_s$ be a sign-linear discrete-time system of dimension $n > 0$. Then, Σ is observable if and only if the following conditions hold:*

1. $\det A \neq 0$,
2. (A, B, C) *has property* $\mathcal{P}$.

Proof. Necessity. Suppose Σ is observable. We know $\det A \neq 0$ from Lemma 2.6. Now assume that property $\mathcal{P}$ would not hold, and pick $x \neq 0$ in the intersection in (3). The output sequence for any given control sequence $\{u_1, u_2, \ldots\}$ is $\{y(0), y(1), \ldots\}$ where

$$y(k) = \operatorname{sign}\left(CA^k x + \sum_{l=1}^{k} \mathcal{A}_l u_{k-l+1}\right).$$

For the chosen x in that intersection, each row of each term in the output sequence has the form

$$y(k)_j = \begin{cases} \operatorname{sign}(C_j A^k x + 0) & \text{if } j \notin I(\mathcal{A}), \\ \operatorname{sign}(0 + *) & \text{if } j \in I(\mathcal{A}), \end{cases}$$

where $*$ denotes a (possibly nonzero) function of the inputs and the Markov parameters. Then, x and λx for any $\lambda > 0$, $\lambda \neq 1$, cannot be distinguished, so observability is contradicted.

Sufficiency. Now suppose $\det A \neq 0$ and (A, B, C) has property $\mathcal{P}$. We must show that Σ is observable. Pick an integer $l > 0$ so that the ith row of

$$(\mathcal{A}_1 \, \mathcal{A}_2 \cdots \mathcal{A}_l)$$

is nonzero for every $i \in I(\mathcal{A})$. Note that since A is invertible,

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ q=0,\ldots,n-1}} \ker(C_j A^{q+l}) = \{0\}, \tag{4}$$

which follows from (3).

Now look at the following n terms in the output sequence for initial state x:

$$\begin{aligned} &\operatorname{sign}\left(CA^l x + \mathcal{A}_l u_1 + \cdots + \mathcal{A}_1 u_l\right), \\ &\operatorname{sign}\left(CA^{l+1} x + \mathcal{A}_{l+1} u_1 + \cdots + \mathcal{A}_1 u_{l+1}\right), \ldots, \\ &\operatorname{sign}\left(CA^{l+n-1} x + \mathcal{A}_{l+n-1} u_1 + \cdots + \mathcal{A}_1 u_{l+n-1}\right). \end{aligned}$$

Given $x \neq z$ we must show that x, z are distinguishable. If we can choose a sequence $u_1, u_2, \ldots u_{l+n-1}$ so that some row of some term above is different for the initial states x and z, then x, z are distinguishable. As $x - z \neq 0$, we may pick some $j \in I(\mathcal{A})$ and some $q = 0, \ldots, n-1$ so that

$$C_j A^{q+l} x \neq C_j A^{q+l} z.$$

Since $j \in I(\mathcal{A})$, the jth row of $(\mathcal{A}_1 \cdots \mathcal{A}_l)$ is nonzero by our choice of l. Denote $k := q + l$ so that the jth row of $(\mathcal{A}_1 \cdots \mathcal{A}_k)$ is also nonzero. Let $\mathcal{A}_i^j$ be the jth row of

$\mathcal{A}_i$. Then we may apply Remark 3.3 (with $\mathcal{H} = \mathbb{R}^k$ and the standard inner product) and obtain $u_1, u_2, \ldots, u_k$ so that

$$\begin{aligned}&\operatorname{sign}(C_j A^k x + \mathcal{A}_k^j u_1 + \cdots + \mathcal{A}_1^j u_k)\\ &\neq \operatorname{sign}(C_j A^k z + \mathcal{A}_k^j u_1 + \cdots + \mathcal{A}_1^j u_k).\end{aligned}$$

Thus, x and z are distinguishable. This completes the proof. □

We will say that a triple (A, B, C) is *discrete-time sign-linear observable* if the triple satisfies the observability conditions in Theorem 3.4.

For continuous-time sign-linear systems, the conditions for observability are slightly weaker, as invertibility of the matrix A is not needed.

THEOREM 3.5. *Let $\Sigma = (A, B, C)_s$ be a sign-linear continuous-time system of dimension $n > 0$. Then Σ is observable if and only if (A, B, C) has property $\mathcal{P}$.*

Proof. The proof is exactly the same as in the discrete-time case. Indeed, if (3) is not satisfied and $x \neq 0$ is in the intersection of the kernels, consider the output

$$y(t) = \operatorname{sign}\left(Ce^{At}x + \int_0^t \sum_{k=1}^{\infty} \frac{\mathcal{A}_k (t-s)^{k-1}}{(k-1)!} u(s)ds\right).$$

Each row has the form

$$y(t)_j = \begin{cases} \operatorname{sign}(C_j e^{At}x + 0) & \text{if } j \notin I(\mathcal{A}),\\ \operatorname{sign}(0 + *) & \text{if } j \in I(\mathcal{A}),\end{cases}$$

where $*$ denotes a (possibly nonzero) function of the inputs and the Markov parameters. Then x and λx for any $\lambda > 0$, $\lambda \neq 1$, are indistinguishable, contradicting observability.

Now suppose (A, B, C) satisfies property $\mathcal{P}$. We must show that Σ is observable. Look at the output function for initial state x:

$$y(t) = \operatorname{sign}\left(Ce^{At}x + \int_0^t K(t-s)u(s)ds\right),$$

where

$$K(t-s) := \sum_{k=1}^{\infty} \frac{\mathcal{A}_k (t-s)^{k-1}}{(k-1)!}.$$

Given $x \neq z$ we must show that x, z are distinguishable. If we can choose a t and a control function $u(\cdot)$ of length t so that some row of $y(t)$ is different for the initial states x and z, then x, z are distinguishable. As $x - z \neq 0$, we may pick some $j \in I(\mathcal{A})$ and some $t \geq 0$ so that

$$C_j e^{At}x \neq C_j e^{At}z,$$

by property $\mathcal{P}$. Since $C_j e^{At}x$ is an analytic function of t, this is true in a neighborhood of $t = 0$ so we may, in fact, fix a $t > 0$ so that the inequality holds.

Next note that since $j \in I(\mathcal{A})$, $\mathcal{A}^j \not\equiv 0$ so also $K_j(\cdot) \not\equiv 0$. Now apply Remark 3.3 with $a = C_j e^{At}x, b = C_j e^{At}z, \mathcal{H} = \mathcal{L}^\infty[0, t]$ with the $\mathcal{L}^2$ inner product

$$\langle v(\cdot), u(\cdot)\rangle := \int_0^t v(s)u(s)ds,$$

and $c = K_j(t-s) \in \mathcal{H}$. Thus we may choose a measurable essentially bounded $u(\cdot)$ so that

$$\text{sign}\left(Ce^{At}x + \int_0^t K(t-s)u(s)ds\right) \neq \text{sign}\left(Ce^{At}z + \int_0^t K(t-s)u(s)ds\right).$$

This $u(\cdot)$ distinguishes x, z and the proof is complete. □

4. Sign-linear realizations. We now focus on questions of realizability for the class of sign-linear systems. As we mentioned earlier, a sign-linear system does not have a *unique* associated Markov sequence. However, for sign-linear systems, we have the following obvious fact.

Remark 4.1. A Markov parameter sequence associated to $\Sigma = (A, B, C)_s$ is any sequence of $p \times m$ matrices $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \ldots\}$ so that

$$\mathcal{A}_i = \Lambda C A^{i-1} B, \quad i = 1, 2, 3, \ldots \tag{5}$$

for some scaling matrix Λ.

For the degenerate system, its (only) associated sequence is $\mathcal{A} \equiv 0$. If $\mathcal{A}$ is associated to Σ, we also say that Σ *realizes* $\mathcal{A}$. If (A, B, C) is a triple of matrices and $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots\}$ is a Markov sequence so that $\mathcal{A}_i = CA^{i-1}B$ holds, we will say that (A, B, C) is a *linear representation* of $\mathcal{A}$. The standard terminology is "realization" (as a linear system), but this can lead to confusion here, since we are interested in sign-linear realizations. Note that the above definitions imply that for any given triple (A, B, C), and any sequence of $p \times m$ matrices $\mathcal{A}$, the sign-linear system $\Sigma = (A, B, C)_s$ realizes $\mathcal{A}$ if and only if $(A, B, \Lambda C)$ is a linear representation of $\mathcal{A}$ for some scaling matrix Λ. In other words, there must exist a triple associated to Σ that represents $\mathcal{A}$.

The matrix

$$H_{s,t} = \begin{pmatrix} \mathcal{A}_1 & \cdots & \mathcal{A}_t \\ \vdots & \ddots & \vdots \\ \mathcal{A}_s & \cdots & \mathcal{A}_{s+t-1} \end{pmatrix}$$

is called the $s \times t$ *Hankel matrix* for the Markov sequence $\mathcal{A}$. The (*Hankel*) *rank* of a sequence $\mathcal{A}$ is defined to be

$$\sup_{s,t} \text{rank}\, H_{s,t}.$$

An i/o map (for a precise definition, see [12, Rem. 2.2.2], is a function of controls u defined on some time interval $[\sigma, \tau)$, which gives the entire output function for the time interval $[\sigma, \tau]$.

DEFINITION 4.2. A $(p \times m)$ *discrete-time sign-linear i/o map* α is a discrete-time i/o map for which there exists some sequence of $(p \times m)$ matrices $\mathcal{A}_1, \mathcal{A}_2, \ldots$, so that

$$\alpha(u)(j) = \text{sign}\,(\mathcal{A}_j u_1 + \cdots + \mathcal{A}_1 u_j) \tag{6}$$

for each input sequence $\{u_1, u_2, u_3, \ldots\}$. A *continuous-time sign-linear i/o map* is a continuous-time i/o map α for which there exists an analytic kernel $K(t)$ with expansion

$$K(t) = \sum_{i=1}^{\infty} \mathcal{A}_i \frac{t^{i-1}}{(i-1)!} \tag{7}$$

so that

$$\alpha(u)(t) = \text{sign}\left(\int_0^t K(t-s)u(s)ds\right) \tag{8}$$

for every measurable essentially bounded control function $u(\cdot)$. In either case, any sequence of matrices $\mathcal{A}_1, \mathcal{A}_2, \ldots$ as above is called a *Markov sequence* of the map α. We will study realizations of these i/o maps by sign-linear systems. It will be helpful to have a simple example in mind as we go through the definitions and results.

Example 4.3. Let $\mathcal{A} = \{1, -1, 1, -1, \ldots\}$. Then α is a 1×1 discrete-time i/o map, where

$$\alpha(u)(j) = \text{sign}\left((-1)^{j-1}u_1 + \cdots + u_{j-2} - u_{j-1} + u_j\right).$$

For the control $u = \{1, 0, 0, \ldots\}$, $\alpha(u)(j) = (-1)^{j-1}$ for all j. For the control $u = \{1, 2, 3, 0, 0, 0, \ldots\}$, the values of the i/o function are

$$\begin{aligned}
\alpha(u)(1) &= \text{sign}\,(1) = 1,\\
\alpha(u)(2) &= \text{sign}\,(-1+2) = 1,\\
\alpha(u)(3) &= \text{sign}\,(1-2+3) = 1,\\
\alpha(u)(4) &= \text{sign}\,(-1+2-3) = -1,\\
\alpha(u)(j) &= (-1)^{j-1} \text{ for } j > 4.
\end{aligned}$$

DEFINITION 4.4. Two triples (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$ are *sign-similar* if there is a $T \in Gl(n)$ and a scaling matrix Λ such that

$$\begin{aligned}
T^{-1}AT &= \hat{A},\\
T^{-1}B &= \hat{B},\\
CT &= \Lambda\hat{C}.
\end{aligned}$$

If $\Sigma = (A, B, C)_s$ and $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C})_s$ are sign-linear systems, they are called *sign-similar* if the corresponding triples are. Note that sign-similarity is an equivalence relation, and that the ambiguity in defining a triple associated to Σ causes no difficulties in the above definition.

DEFINITION 4.5. Two Markov sequences

$$\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots\}, \quad \text{and} \quad \hat{\mathcal{A}} = \{\hat{\mathcal{A}}_1, \hat{\mathcal{A}}_2, \ldots\}$$

are *sign-equivalent* if there exists a scaling matrix Λ so that

$$\mathcal{A}_j = \Lambda\hat{\mathcal{A}}_j, \quad j = 1, 2, 3, \ldots.$$

Note that if $K(\cdot)$ and $\hat{K}(\cdot)$ are as in (7) for the sequences $\mathcal{A}$ and $\hat{\mathcal{A}}$, sign-equivalence of $\mathcal{A}$ and $\hat{\mathcal{A}}$ is the same as asking that $K(t) = \Lambda\hat{K}(t)$ for all t, where Λ is a scaling matrix. The Markov sequence $\mathcal{A}$ in Example 4.3 is sign-equivalent to $\hat{\mathcal{A}} = \{a, -a, a, -a, \ldots\}$ for any $a > 0$.

4.1. Basic facts about realizations. The next lemma says that the Markov sequence $\mathcal{A}$ is uniquely determined by a sign-linear i/o map α up to multiplication by a scaling matrix. That is, a sign-linear i/o map is defined by many Markov sequences, but these sequences are related by scaling.

Observe that the impulse response of a sign-linear i/o map (e.g., for discrete-time systems, the response to the input $u = \{1, 0, 0, 0, \ldots\}$) is not enough to uniquely characterize the i/o map. For a discrete-time sign-linear i/o map α, the impulse response is just the sequence of signs of the Markov parameters: $\{\text{sign}\,(\mathcal{A}_1), \text{sign}\,(\mathcal{A}_2), \ldots\}$. Such a sign sequence represents infinitely many different families of sign-equivalent Markov sequences, as illustrated by the following example.

Example 4.6. Let α_1 be the discrete-time sign-linear i/o map defined by $\mathcal{A} = \{1, 3, 1, 3, \ldots\}$ and $\hat{\alpha}$ the map defined by $\hat{\mathcal{A}} = \{3, 1, 3, 1, \ldots\}$. Then the impulse response for both i/o maps is $\{+1, +1, +1, \ldots\}$; however the two maps are not the same as shown by considering the output corresponding to the input $u = \{1, -1, 1, -1, \ldots\}$:

$$\begin{array}{rclrcl} \alpha(u)(1) & = & +1; & \hat{\alpha}(u)(1) & = & +1, \\ \alpha(u)(2) & = & +1; & \hat{\alpha}(u)(2) & = & -1, \\ \alpha(u)(3) & = & -1; & \hat{\alpha}(u)(3) & = & +1, \\ \alpha(u)(4) & = & +1; & \hat{\alpha}(u)(4) & = & -1. \end{array}$$

LEMMA 4.7. *$\mathcal{A}$ and $\hat{\mathcal{A}}$ define the same i/o map if and only if they are sign-equivalent.*

Proof. If $\mathcal{A}$ and $\hat{\mathcal{A}}$ are sign-equivalent, then $\mathcal{A}_i = \Lambda \hat{\mathcal{A}}_i$ for all i, where Λ is a scaling matrix. It is then clear from (6)–(8) that the corresponding i/o maps coincide. To prove the converse, we can assume, without loss of generality, looking at each component of the output and each row of $\mathcal{A}$, that $p = 1$. We first prove the following easy observation.

Remark 4.8. If $\mathcal{V}$ is a real pre-Hilbert space and if $v, w \in \mathcal{V}$ are nonzero and such that

$$\text{sign}\,\langle v, u\rangle = \text{sign}\,\langle w, u\rangle$$

for all $u \in \mathcal{V}$, then there exists $\lambda > 0$ so that $v = \lambda w$.

Proof. Suppose first that v and w are linearly independent, and consider the plane they span. Let $u \neq 0$ be in this plane and perpendicular to $v + w$. As $\langle v + w, u\rangle = 0$, necessarily $\langle v, u\rangle \neq 0$ and $\langle w, u\rangle \neq 0$, since if either of these is zero, then the other one is too, and that would contradict linear independence. Then

$$\langle v, u\rangle + \langle w, u\rangle = \langle v + w, u\rangle = 0.$$

So $\langle v, u\rangle = -\langle w, u\rangle \neq 0$, contradicting the assumption. Thus, either $v = \lambda w$ with $\lambda > 0$, or $v = -\mu w$ with $\mu > 0$. If $v = -\mu w$, then

$$\langle v, u\rangle = \langle -\mu w, u\rangle = -\mu \langle w, u\rangle \neq 0$$

and so sign $\langle v, u\rangle \neq$ sign $\langle w, u\rangle$, again a contradiction. The only remaining possibility is that there exists a $\lambda > 0$ so that $v = \lambda w$. □

Now we can continue the proof of Lemma 4.7. For discrete-time, we must show that if

$$\text{sign}\left(\sum_{i=1}^{l} \mathcal{A}_i u_{l-i+1}\right) = \text{sign}\left(\sum_{i=1}^{l} \hat{\mathcal{A}}_i u_{l-i+1}\right)$$

for all $l \geq 1$ and for all $u_1, u_2, \ldots, u_l$, then $\mathcal{A} = \lambda\hat{\mathcal{A}}$ for some $\lambda > 0$. First choose an l so that $(\mathcal{A}_1, \ldots, \mathcal{A}_l) \neq 0$. Note that $\mathbb{R}^{lm} = (\mathbb{R}^m)^l$ forms a pre-Hilbert space with the standard inner product

$$\left\langle \begin{pmatrix} v_1 \\ \vdots \\ v_l \end{pmatrix}, \begin{pmatrix} u_1 \\ \vdots \\ u_l \end{pmatrix} \right\rangle := \sum_1^l v_i' u_i.$$

Applying Remark 4.8 with $v = (\mathcal{A}_1, \ldots, \mathcal{A}_l)'$, $w = (\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_l)'$ and $u = (u_l', \ldots, u_1')'$, we see that there exists a $\lambda > 0$ so that

$$(\mathcal{A}_1, \ldots, \mathcal{A}_l) = \lambda(\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_l).$$

Now pick any $q > l$. Applying the same argument to $(\mathbb{R}^m)^q$, we obtain a $\lambda_q > 0$ so that

$$(\mathcal{A}_1, \ldots, \mathcal{A}_q) = \lambda_q(\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_q).$$

Since $(\mathcal{A}_1, \ldots, \mathcal{A}_l)$ is a subvector of $(\mathcal{A}_1, \ldots, \mathcal{A}_q)$, and similarly $(\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_l)$ a subvector of $(\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_q)$, this implies that $\lambda = \lambda_q$. Thus $\mathcal{A}_q = \lambda\hat{\mathcal{A}}_q$ for all $q \geq 1$.

For continuous-time, we need to show that if

$$\operatorname{sign}\left(\int_0^t K(t-s)u(s)ds\right) = \operatorname{sign}\left(\int_0^t \hat{K}(t-s)u(s)ds\right)$$

for all $t \in [0, \infty)$ and for all $u(\cdot)$, measurable and essentially bounded on $[0, t]$, then $K(t) = \lambda\hat{K}(t)$ for some $\lambda > 0$ and for all $t \geq 0$. Note that $\mathcal{L}^\infty[0, t]$ forms a pre-Hilbert space with the $\mathcal{L}^2$ inner product

$$\langle v(\cdot), u(\cdot)\rangle := \int_0^t v(s)u(s)ds.$$

Applying Remark 4.8 with $v(s) = K(t-s)$ and $w(s) = \hat{K}(t-s)$, we see that there exists a $\lambda_t > 0$ so that $K(\cdot)|_{[0,t]} = \lambda_t\hat{K}(\cdot)|_{[0,t]}$. Using an argument similar to the one used in the discrete-time case, we can conclude that there exists a $\lambda > 0$ so that $K(t) = \lambda\hat{K}(t)$ for all $t \geq 0$. □

COROLLARY 4.9. *Let α be a sign-linear i/o map, with Markov sequence $\mathcal{A}$, and let $\Sigma = (A, B, C)_s$ be any sign-linear realization of α, with Markov sequence $\hat{\mathcal{A}}$. Then $\mathcal{A}$ and $\hat{\mathcal{A}}$ are sign-equivalent.*

Proof. Just note that $\mathcal{A}$ and $\hat{\mathcal{A}}$ define the same i/o map, namely α. Thus, the previous lemma applies. □

4.2. Minimality.

DEFINITION 4.10. A sign-linear system of dimension n is *minimal* if any other sign-linear system realizing the same i/o map has dimension $n_1 \geq n$. Recall that a triple (A, B, C) is *canonical* if and only if it is a minimal-dimensional linear representation of its Markov sequence ([12, Thm. 20]). The next lemma states that minimality of a sign-linear system is equivalent to minimality of the associated linear system.

LEMMA 4.11. *The sign-linear system $(A, B, C)_s$ is a minimal realization of α if and only if the triple (A, B, C) is canonical.*

Proof. Suppose the sign-linear system $\Sigma = (A, B, C)_s$ is a minimal realization of the i/o map α and $\mathcal{A}$ is a Markov sequence associated to α. If (A, B, C) is not canonical, then there exists another triple $(\hat{A}, \hat{B}, \hat{C})$ of smaller dimension that is a linear representation of the same Markov sequence $\hat{\mathcal{A}}$ as (A, B, C). Then $\Sigma = (A, B, C)_s$ also realizes $\hat{\mathcal{A}}$ so $\hat{\mathcal{A}}$ is sign-equivalent to $\mathcal{A}$ by Corollary 4.9. But then since $(\hat{A}, \hat{B}, \hat{C})_s$ realizes $\hat{\mathcal{A}}$, it also realizes $\mathcal{A}$ (a Markov sequence associated to a sign-linear system is only determined up to sign-equivalence). Thus, $(\hat{A}, \hat{B}, \hat{C})_s$ is a sign-linear system realizing α and of smaller dimension than $(A, B, C)_s$, contradicting minimality.

Conversely, suppose the triple (A, B, C) is canonical of dimension n, which implies, in particular, that it is minimal. If $(A, B, C)_s$ is not also minimal, then there exists a sign-linear system $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C})_s$ of dimension $n_1 < n$ that realizes the same i/o map α as $(A, B, C)_s$. Let $\mathcal{A}$ be the Markov sequence represented by (A, B, C) and $\hat{\mathcal{A}}$ the sequence represented by $(\hat{A}, \hat{B}, \hat{C})$. Then $(A, B, C)_s$ realizes $\mathcal{A}$ and $(\hat{A}, \hat{B}, \hat{C})_s$ realizes $\hat{\mathcal{A}}$. Since the two sign-linear systems realize the same i/o map α, $\mathcal{A}$ and $\hat{\mathcal{A}}$ are sign-equivalent (Corollary 4.9). Thus, there exists a scaling matrix Λ so that $\mathcal{A}_j = \Lambda \hat{\mathcal{A}}_j$ for all $j \geq 1$. So

$$CA^{i-1}B = \Lambda \hat{C} \hat{A}^{i-1} \hat{B}, \quad i = 1, 2, 3, \ldots.$$

But then $(\hat{A}, \hat{B}, \Lambda \hat{C})$ is a linear representation of $\mathcal{A}$ of dimension $n_1 < n$, contradicting the minimality of (A, B, C). □

Remark 4.12. If $\mathcal{A}$ and $\hat{\mathcal{A}}$ are two Markov sequences associated to the same i/o map α, then $\operatorname{rank} \mathcal{A} = \operatorname{rank} \hat{\mathcal{A}}$. Thus, we can define the *Hankel rank* of α as the rank of any of the associated Markov sequences. Indeed, by Lemma 4.7, we know that $\mathcal{A}$ and $\hat{\mathcal{A}}$ are sign-equivalent. Thus there is a scaling matrix Λ with $\mathcal{A}_j = \Lambda \hat{\mathcal{A}}_j$, $j = 1, 2, 3, \ldots$. We then have, for any $s, t \geq 1$,

$$H_{s,t} = \Lambda_s \hat{H}_{s,t},$$

where $\hat{H}_{s,t}$ is the $s \times t$ Hankel matrix for $\hat{\mathcal{A}}$ and $\Lambda_s = \operatorname{diag}(\Lambda, \ldots, \Lambda)$. Since this is true for every s, t,

$$\begin{aligned} \operatorname{rank}(\mathcal{A}) &= \sup_{s,t} \{\operatorname{rank}(H_{s,t})\} \\ &= \sup_{s,t} \{\operatorname{rank}(\hat{H}_{s,t})\} = \operatorname{rank}(\hat{\mathcal{A}}), \end{aligned}$$

as claimed.

THEOREM 4.13. *Let α be a sign-linear i/o map. Then α is realizable by a sign-linear system if and only α has finite Hankel rank.*

Proof. If α has finite rank then any Markov sequence for α, $\mathcal{A}$, has finite rank. It then follows that there exists a linear representation for $\mathcal{A}$, (A, B, C). Then the corresponding sign-linear system $(A, B, C)_s$ realizes α.

Conversely, given a sign-linear i/o map α that is realizable by a sign-linear system $(A, B, C)_s$, we would like to show that α has finite rank. One Markov sequence for $(A, B, C)_s$ is the impulse response of the linear system (A, B, C). This impulse response $\mathcal{A}$ is a Markov sequence for α. From linear realization theory, we know that $\mathcal{A}$ has finite rank. Thus, by the remark above, α has finite rank. □

LEMMA 4.14. *If (A, B, C) is a canonical representation of a Markov sequence $\mathcal{A}$, then (A, B, C) satisfies property $\mathcal{P}$.*

Proof. Suppose (A, B, C) does not satisfy property $\mathcal{P}$. Then by observability there is some $i \notin I(\mathcal{A})$ (i.e., so that the ith row $\mathcal{A}^i$ of $\mathcal{A}$ is zero) so that

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ l=0,\ldots,n-1}} \ker (C_j A^l) \neq \{0\},$$

but

$$\bigcap_{\substack{j \in I(\mathcal{A}) \cup \{i\} \\ l=0,\ldots,n-1}} \ker (C_j A^l) \subsetneqq \bigcap_{\substack{j \in I(\mathcal{A}) \\ l=0,\ldots,n-1}} \ker (C_j A^l). \tag{9}$$

Since $\mathcal{A}^i \equiv 0$ then $C_i(A^j B) = 0$ for all j, so in particular,

$$C_i \begin{pmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{pmatrix} = 0.$$

The pair (A, B) is controllable so

$$\begin{pmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{pmatrix}$$

has full row rank. Thus $C_i = 0$, contradicting (9). So property $\mathcal{P}$ indeed holds. □

LEMMA 4.15. *If (A, B, C) is a triple satisfying property $\mathcal{P}$, then the sign-linear system $\Sigma = (A, B, C)_s$ is final-state observable.*

Proof. First suppose $\Sigma = (A, B, C)_s$ is a discrete-time sign-linear system. Perform a change of variables in the state space. Let

$$z = T^{-1}x = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix},$$

where $T \in Gl(n)$ is chosen so that

$$T^{-1}AT = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},$$

with A_1 of size $n_1 \times n_1$ nilpotent and A_2 of size $n_2 \times n_2$ nonsingular. (This can be done, for instance, by first putting A in real canonical form and then reordering the blocks so that the blocks corresponding to 0 eigenvalues come first.) Then

$$y = \text{sign } (CTz)$$

can be written as

$$y = \text{sign } \left[\begin{pmatrix} C_1 & C_2 \end{pmatrix} z \right],$$

and we can also write

$$T^{-1}B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

Since (A, C) is an observable pair, the n columns of

$$\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = \begin{pmatrix} C_1 & C_2 \\ C_1A_1 & C_2A_2 \\ \vdots & \vdots \\ C_1A_1^{n-1} & C_2A_2^{n-1} \end{pmatrix} T^{-1}$$

are linearly independent. As T^{-1} is an invertible matrix, both (A_1, C_1) and (A_2, C_2) must be observable pairs. Property $\mathcal{P}$ implies that the subset of outputs indexed by $I(\mathcal{A})$ allows observability of the pair (A, C). Then the outputs indexed by $I(\mathcal{A})$ also allow observability of the pair (A_2, C_2). We know that $\mathcal{A}^i \not\equiv 0$ for $i \in I(\mathcal{A})$. Since A_1 is nilpotent, after $n+1$ steps the output sequence looks like

$$\operatorname{sign}(C_2 A_2^{n+1} z_2 + \mathcal{A}_1 u_{n+1} + \cdots + \mathcal{A}_{n+1} u_1),$$
$$\operatorname{sign}(C_2 A_2^{n+2} z_2 + \mathcal{A}_1 u_{n+2} + \cdots + \mathcal{A}_{n+2} u_1), \ldots.$$

Now we have (A_2, C_2) is an observable pair, $\det A_2 \neq 0$, and

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ q=0,\ldots,n_2-1}} \ker((C_2)_j A_2^q) = \{0\},$$

where $\mathcal{A}^i \not\equiv 0$ for $i \in I(\mathcal{A})$. Now using Remark 3.3, we may always choose appropriate controls to distinguish any distinct z_2 and $\tilde{z}_2$. Also, z_1 goes to zero (in less than n time steps). So the system is final-state observable.

For a continuous-time sign-linear system, $\Sigma = (A, B, C)_s$, property $\mathcal{P}$ alone implies that the system is observable, by Lemma 3.5. Hence, Σ is also final-state observable. (Observability and final-state observability are equivalent in continuous-time.) □

THEOREM 4.16.

1. *If a sign-linear realization is controllable and observable then it is minimal.*
2. *If it is minimal then it is controllable and final-state observable.*
3. *Any two minimal sign-linear realizations are sign-similar.*

Proof. 1. If the sign-linear system $(A, B, C)_s$ is controllable and observable then in particular the triple (A, B, C) is canonical so the sign-linear system is minimal by Lemma 4.11.

2. If the system $\Sigma = (A, B, C)_s$ is minimal, then the triple (A, B, C) is canonical. If $\mathcal{A} \equiv 0$, then the minimal realization has dimension 0 and is trivially final-state observable. So now assume that we are dealing with dimension $n > 0$. We know that the triple (A, B, C) is canonical, so it satisfies property $\mathcal{P}$ (Lemma 4.14). Next, applying Lemma 4.15, we conclude that $\Sigma = (A, B, C)_s$ is final-state observable.

3. Given two minimal realizations $(A, B, C)_s$ and $(\hat{A}, \hat{B}, \hat{C})_s$, of a sign-linear map α, with Markov sequence $\mathcal{A}$, we must show that they are sign-similar. The corresponding triples (A, B, C) and $(\hat{A}, \hat{B}, \hat{C})$ represent Markov sequences $\mathcal{A}^1$ and $\mathcal{A}^2$, respectively, which are both sign-equivalent to $\mathcal{A}$ (Corollary 4.9). That is, we have scaling matrices Λ_1, Λ_2 satisfying

$$\mathcal{A} = \Lambda_1 \mathcal{A}^1, \qquad \mathcal{A} = \Lambda_2 \mathcal{A}^2.$$

Since the sign-linear realizations are minimal, the linear representations are canonical (Lemma 4.11). Since Λ_1 and Λ_2 have full rank, this implies that $(A, \Lambda_1 C)$ and $(\hat{A}, \Lambda_2 \hat{C})$ are also observable pairs. Thus, $(A, B, \Lambda_1 C)$ and $(\hat{A}, \hat{B}, \Lambda_2 \hat{C})$ are both canonical linear representations of the same Markov sequence $\mathcal{A}$. By [12], Thm. 20, they must be similar, i.e., there exists some $T \in Gl(n)$ so that $T^{-1} A T = \hat{A}$, $T^{-1} B = \hat{B}$, and $\Lambda_1 C T = \Lambda_2 \hat{C}$. Thus $(A, B, C)_s$ and $(\hat{A}, \hat{B}, \hat{C})_s$ are sign-similar, with T as above and scaling matrix $\Lambda = \Lambda_1^{-1} \Lambda_2$. □

The rank of $\mathcal{A} = \{1, -1, 1, -1, \ldots\}$ from Example 4.3, is 1, which is clearly finite. The triple $A = -1$, $B = 1$, $C = 1$ is a realization of the i/o map α, which is

controllable and observable; hence it is minimal. An example of a nonminimal sign-linear realization of the same α is

$$A = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix};\ B = \begin{pmatrix} 1 \\ 1 \end{pmatrix};\ C = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

In this case (A, B) is a controllable pair, but (A, C) is not an observable pair.

4.3. Counterexamples. Note that the converses of parts 1 and 2 of Theorem 4.16 are not true for discrete-time systems. If a sign-linear system is minimal, it is not necessarily observable. For example, the system with $x^+ = u$ and $y = \text{sign}\,(x)$ is minimal, but $A = 0$ so it is not observable. Also, a system may be final-state observable, and yet not be minimal. For example,

$$\begin{aligned} x^+ &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u, \\ y &= \text{sign}\,\left[\begin{pmatrix} 0 & 1 \end{pmatrix} x\right] \end{aligned}$$

is final-state observable. After k steps (for any $k \geq 1$) any state (x_1, x_2) ends up at (u_k, x_2) and x_2 can be identified. However, (A, C) is not an observable pair. If this system would be minimal, then the corresponding triple would be canonical by Lemma 4.11. But then (A, C) would have to be an observable pair. The minimal system for this i/o map is one of dimension 1, namely, $x^+ = x + u,\ y = \text{sign}\,(x)$.

5. Canonical realizations of sign-linear i/o maps. We noted that for sign-linear systems (unlike for linear systems) it is not true that a system is minimal if and only if it is canonical. The problem is that a minimal *discrete-time* sign-linear system may have $\det A = 0$, in which case it is not observable (Theorem 3.4). We may then ask—what is the canonical realization of a minimal sign-linear system which is guaranteed to exist by abstract realization theory ([12, §5.8])? The answer, for $p = 1$, is that for any α realizable by a sign-linear system, there exists a canonical (reachable and observable) system $\tilde{\Sigma}$ that realizes α, where $\tilde{\Sigma}$ is in the form of a cascade of a sign-linear system and shift registers. (In the general case, $p > 1$, the result has to be modified: we can only conclude that there is a system of this cascade form in which the minimal system may be embedded.)

We know there exists some canonical realization. We need only to show that there is a canonical realization of the form described above. Next we sketch the construction for the single-output case ($p = 1$). First find a minimal sign-linear realization Σ of α. Then we know (A, B, C) is a canonical triple and satisfies property $\mathcal{P}$ (Lemmas 4.11 and 4.14). Perform a change of variables in the state space so that A has the form

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

with A_1 an $n_1 \times n_1$ invertible matrix and A_2 an $n_2 \times n_2$ nilpotent matrix. (Note that if Σ is already observable, then A is invertible and there is no A_2. This Σ is already in the canonical form we are looking for.) Now the system equations have the form

$$\begin{aligned} x_1^+ &= A_1 x_1 + B_1 u, \\ x_2^+ &= A_2 x_2 + B_2 u, \\ y &= \text{sign}\,(C_1 x_1 + C_2 x_2). \end{aligned}$$

From now on, assume Σ has the form described above. Let κ be the relative degree of the system and $l := \min\{\kappa, n_2\}$. Let $\tilde{\Sigma}$ be the discrete-time system with state space $\mathbb{R}^{n-l} \times \{-1, 0, 1\}^l$, and system equations

$$
\begin{aligned}
\xi^+ &= F\xi + Gu, \\
\zeta_1^+ &= \operatorname{sign}(\xi_{n-l} + \mathcal{A}_l u), \\
\zeta_2^+ &= \zeta_1, \\
&\vdots \\
\zeta_l^+ &= \zeta_{l-1}, \\
\eta &= \zeta_l,
\end{aligned} \tag{10}
$$

where $(F, G) = (A_1, B_1)$ when $l = n_2$, and when $l < n_2$,

$$
F = \begin{pmatrix} A_1 & 0 & 0 \\ C_1 A_1^{n_2} & 0 & 0 \\ 0 & I & 0 \end{pmatrix}, \quad G = \begin{pmatrix} B_1 \\ \mathcal{A}_{n_2} \\ \mathcal{A}_{n_2-1} \\ \vdots \\ \mathcal{A}_{l+1} \end{pmatrix},
$$

and I is the identity matrix of size $n_2 - l - 1$. (When $l = n_2 - 1$, there is no "I" part.) This can be seen as a cascade of a sign-linear system and shift registers.

LEMMA 5.1. *The system $\tilde{\Sigma}$ is the observable reduction of Σ.*

Proof. First we show that two states x and z are indistinguishable for Σ if and only if

$$
x_1 = z_1 \tag{11}
$$

$$
\begin{cases} CA^{n_2-1}x &= CA^{n_2-1}z \\ &\vdots \\ CA^{l+1}x &= CA^{l+1}z \\ CA^l x &= CA^l z \end{cases} \tag{12}
$$

$$
\begin{cases} \operatorname{sign}(CA^{l-1}x) &= \operatorname{sign}(CA^{l-1}z) \\ &\vdots \\ \operatorname{sign}(Cx) &= \operatorname{sign}(Cz). \end{cases} \tag{13}
$$

In the case $l = n_2$, we have only (11) and (13). Suppose all equalities hold. Since $l <$ relative degree, the first l output terms for Σ are independent of the control. Then the last l equalities imply that the first l output terms coincide for x and z, for any input. Equations (12) imply that actually the first n_2 output terms coincide for x and z.

The remaining outputs only involve the first n_1 components of the state because of the nilpotency of A_2. So if $x_1 = z_1$, then we see that all the remaining output terms are equal for initial states x and z. Thus, x and z are indistinguishable.

On the other hand, if x, z are indistinguishable, then using any control sequence, the outputs for the two initial states are always equal. In particular, the first l output terms are independent of the control so we obtain the last l equalities directly. For equalities (12) (in the case $l < n_2$) look at the next $n_2 - l$ output terms. If

$$
CA^k x \neq CA^k z, \text{ for some } l \leq k \leq n_2 - 1,
$$

then property $\mathcal{P}$ and Remark 3.3 would imply that there is some control that would cause the kth output to be different for x, z, contradicting indistinguishability. Thus, those equalities hold too. Finally, for (11), we may focus on the output terms $y(k)$ for $k \geq n_2$. Indistinguishability implies that in particular, for the 0 control, all output terms are equal. Then $CA^k x_1 = CA^k z_1$ for all $k \geq n_2$. But (A_1, C_1) is an observable pair and $\det A_1 \neq 0$ so this implies $x_1 = z_1$.

Now consider the mapping $\phi : \mathbb{R}^n \to \mathbb{R}^{n-l} \times \{-1, 0, 1\}^l$ given by $\phi(x) = (\xi, \zeta)$, where

$$\xi = \begin{pmatrix} x_1 \\ CA^{n_2-1}x \\ \vdots \\ CA^l x \end{pmatrix} \in \mathbb{R}^{n-l},$$

$$\zeta = \begin{pmatrix} \operatorname{sign}(CA^{l-1}x) \\ \vdots \\ \operatorname{sign}(CAx) \\ \operatorname{sign}(Cx) \end{pmatrix} \in \{-1, 0, 1\}^l$$

and $\xi = x_1$ if $l = n_2$. We just proved that x and z are indistinguishable if and only if $\phi(x) = \phi(z)$. To show that the map is onto, we must show that for any $(\xi, \zeta) \in \mathbb{R}^{n-l} \times \{-1, 0, 1\}^l$, there is some $x \in \mathbb{R}^n$ so that $\phi(x) = (\xi, \zeta)$. Since (A, C) is an observable pair, (A_2, C_2) is also an observable pair. Thus, we may let x_1 be the first n_1 components of ξ and x_2 the solution to

$$\begin{pmatrix} C_2 A_2^{n_2-1} \\ \vdots \\ C_2 A_2 \\ C_2 \end{pmatrix} x_2 = \begin{pmatrix} \xi_{n_1+1} - C_1 A_1^{n_2-1} x_1 \\ \vdots \\ \xi_{n-l} - C_1 A_1^l x_1 \\ \zeta_1 - C_1 A_1^{l-1} x_1 \\ \vdots \\ \zeta_l - C_1 x_1 \end{pmatrix}.$$

Then clearly, $\phi(x) = (\xi, \zeta)$. Furthermore, it is easy to verify that ϕ commutes with the dynamics of Σ so it is a system morphism in the sense of [12], §5.8. □

LEMMA 5.2. *The system $\tilde{\Sigma}$ is reachable and observable and realizes the same i/o behavior as Σ.*

Proof. Since $\tilde{\Sigma}$ is the observable reduction of Σ, and Σ is reachable, [12, Lemma 5.8.3] implies that $\tilde{\Sigma}$ is both reachable and observable with the same input/output behavior as Σ. □

Example 5.3. Let Σ be the system with state space $\mathbb{R}^2$ and

$$\begin{aligned} x_1^+ &= u, \\ x_2^+ &= x_2 + u, \\ y &= \operatorname{sign}(x_1 + x_2). \end{aligned}$$

Then Σ is minimal. But this sign-linear system is not observable, since $\det A = 0$. Perform a change of variables in the state space so that the A matrix is in the form discussed above. In the new coordinates, (z_1, z_2), the equations take the form

$$z_1^+ = z_1 - u,$$

$$z_2^+ = u,$$
$$y = \text{sign}\,(-z_1 + z_2).$$

The Markov sequence is $\mathcal{A} = \{2, 1, 1, 1, \ldots\}$, $n_2 = 1$, relative degree $= 1$, so $l = 1$. The state space for $\tilde{\Sigma}$ is $\mathbb{R} \times \{-1, 0, 1\}$ and the equations for $\tilde{\Sigma}$ are

$$\xi^+ = \xi - u,$$
$$\zeta^+ = \text{sign}\,(\xi + 2u),$$
$$\eta = \zeta.$$

This system is reachable and observable.

6. Sampling. In this section we make some remarks about the time-sampling of sign-linear systems. This is the process of replacing a given continuous-time sign-linear system by the discrete-time one that results when only piecewise constant inputs (with a fixed sampling time) are used. The results in this section can be used to obtain the continuous-time results of Theorem 3.5 as a consequence of those of Theorem 3.4, and they clarify the differences between the two types of results, in particular, the fact that invertibility of the A matrix is not needed in the continuous-time case.

Remark 6.1. Suppose that (A, B, C) has property $\mathcal{P}$. Then the continuous-time sign-linear system $(A, B, C)_s$ is observable.

Proof. We will prove this by studying the associated sampled system. Using the notations and terminology in [12, §2.10], for each $\delta > 0$, the δ-sampled system corresponding to Σ is

$$\sum_\delta : \begin{cases} x^+ &= Fx + Gu, \\ y &= \text{sign}\,(Cx), \end{cases}$$

where $F = e^{\delta A}$, $G = A^{(\delta)}B$, and $A^{(\delta)} = \int_0^\delta e^{(\delta - s)A} ds$. We want to show that there is a $\delta > 0$ so that if (A, B, C) has property $\mathcal{P}$ then the δ-sampled system satisfies condition 2 of Theorem 3.4. If this is true then the sampled system would be observable (clearly $\det e^{\delta A} \neq 0$). Hence, Σ is observable using only piecewise constant controls that are constant on intervals of length δ, and the result is proved.

Apply Kalman's sampling theorem (see [12, Prop. 5.2.11]), to the pair $(A, \hat{C})$ obtained by dropping the rows of C not in $I(\mathcal{A})$. For any δ satisfying

$$\delta(\lambda - \mu) \neq 2\pi i k, \quad k = \pm 1, \pm 2, \pm 3, \ldots, \tag{14}$$

for every two eigenvalues λ, μ of A, we have that

$$\bigcap_{\substack{j \in I(\mathcal{A}) \\ q = 0, \ldots, n-1}} \ker\,(C_j (e^{\delta A})^q) = \{0\}.$$

What is left is to show that

$$I(\mathcal{A}) = I(\mathcal{A}_\delta),$$

where $\mathcal{A}_\delta$ is the Markov sequence of $(e^{\delta A}, A^{(\delta)}B, C)$. Note that $I(\mathcal{A}_\delta) \subseteq I(\mathcal{A})$ is always true for any δ, so the other inclusion is the interesting one. We will prove that if the kth row $\mathcal{A}^k$ of $\mathcal{A}$ is nonzero then the kth row $\mathcal{A}_\delta^k$ of $\mathcal{A}_\delta$ is nonzero for all δ satisfying (14). This will be done by showing the stronger result that $\mathcal{A}^k$ and $\mathcal{A}_\delta^k$ have the same Hankel rank.

Fix a $k \in I(\mathcal{A})$. By restricting our attention to the linear system described by (A, B, C_k), whose Markov sequence is $\mathcal{A}^k$ and sampled-Markov sequence is $\mathcal{A}^k_\delta$, we may, and will, assume without loss of generality that $\mathcal{A}$ is a sequence with $p = 1$ and C has only one row. Thus we need to show that if $\mathcal{A}$ is a Markov sequence with $p = 1$ represented by the triple (A, B, C) and if δ satisfies (14) then $\mathcal{A}_\delta$, the Markov sequence of $(e^{\delta A}, A^{(\delta)}B, C)$, has the same Hankel rank as $\mathcal{A}$.

So let $\delta, \mathcal{A}, \mathcal{A}_\delta$ and (A, B, C) be as described. Next define a sequence $\mathcal{A}^{(\delta)}$ as follows. If

$$K(t) = \sum_{i=1}^{\infty} \mathcal{A}_i \frac{t^{i-1}}{(i-1)!},$$

then the output function for $\Sigma = (A, B, C)$ is $y(t) = \int_0^t K(t-s)u(s)ds$. If we restrict to sampled controls of length δ, then

$$y(l\delta) = \sum_{j=0}^{l-1} \left[\int_{j\delta}^{(j+1)\delta} K(l\delta - s)ds \right] u_{j+1}.$$

Letting

$$\mathcal{A}_j^{(\delta)} = \int_{j\delta}^{(j+1)\delta} K(l\delta - s)ds, \ j = 0, 1, 2, \ldots$$

we get

$$y(l\delta) = \sum_{j=0}^{l-1} \mathcal{A}_j^{(\delta)} u_{j+1}.$$

Look at any linear representation of the Markov sequence $\mathcal{A}$. Take the δ-sampled system for that representation. The Markov sequence for the δ-sampled system is $\mathcal{A}^{(\delta)}$. In particular, applied to the given triple (A, B, C), this means that

$$\mathcal{A}^{(\delta)} = \mathcal{A}_\delta.$$

Take now a canonical representation (A^c, B^c, C^c) of $\mathcal{A}$ of dimension n^c. Its eigenvalues, i.e., the eigenvalues of the matrix A^c, are among the eigenvalues of the (possibly non-canonical) original triple (A, B, C). Thus, δ also satisfies $\delta(\lambda - \mu) \neq 2\pi ik$, $k = \pm 1, \pm 2, \pm 3, \ldots$, for any two eigenvalues λ and μ of A^c. Then controllability and observability of (A^c, B^c, C^c) are preserved by sampling at this δ; and thus the sampled triple $(e^{\delta A^c}, (A^c)^{(\delta)}B^c, C^c)$ is itself canonical and is a linear representation of $\mathcal{A}_\delta = \mathcal{A}^{(\delta)}$. The rank of a Markov sequence is equal to the dimension of a canonical linear representation of that sequence ([12, Cor. 5.5.7]). Therefore,

$$\operatorname{rank} \mathcal{A}^{(\delta)} = n^c = \operatorname{rank} \mathcal{A},$$

as desired. □

REFERENCES

[1] A.D. Back and A.C. Tsoi, *FIR and IIR synapses, a new neural network architecture for time-series modeling*, Neural Computation, 3 (1991), pp. 375–385.

[2] A.M. BAKSHO, S. DASGUPTA, J.S. GARNETT, AND C.R. JOHNSON, *On the similarity of conditions for an open-eye channel and for signed filtered error adaptive filter stability*, Proc. IEEE Conf. Decision and Control, Brighton, UK, Dec. 1991, IEEE Publications, 1991, pp. 1786–1787.

[3] D. F. DELCHAMPS, *Extracting State Information from a Quantized Output Record*, Systems Control Lett., 13 (1989), pp. 365-372.

[4] ———, *Controlling the Flow of Information in Feedback Systems with Measurement Quantization*, Proc. IEEE Conf. Decision and Control, Tampa, Dec. 1989, IEEE Publications, 1989, pp. 2355–2360.

[5] ———, *Stabilizing a Linear System With Quantized State Feedback*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 916–924.

[6] R.O. DUDA AND P.E. HART, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

[7] C.E. GILES, G.Z. SUN, H.H. CHEN, Y.C. LEE, AND D. CHEN, *Higher order networks recurrent and grammatical inference*, Advances in Neural Information Processing Systems 2, D.S. Touretzky, ed., Morgan Kaufmann, San Mateo, CA, 1990.

[8] G.W. PULFORD, R.A. KENNEDY, AND B.D.O. ANDERSON, *Neural network structure for emulating decision feedback equalizers*, Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Toronto, Canada, May 1991, pp. 1517–1520.

[9] P. RAMADGE, *On the Periodicity of Symbolic Observations of Piecewise Smooth Discrete-Time Systems*, IEEE Trans. Automat. Control, AC-35 (1990) pp. 807–813.

[10] R. SCHWARZSCHILD (KOPLON) AND E.D. SONTAG, *Linear systems with constrained observations, Part* I, Report SYCON-91-07, Rutgers Center for Systems and Control, Rutgers University, May 1991.

[11] H. SIEGELMANN AND E.D. SONTAG, *Turing computability with neural nets*, Appl. Math. Lett., 4 (1991) pp. 77–80.

[12] E.D. SONTAG, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.

[13] E.D. SONTAG AND H. SUSSMANN, *Backpropagation separates where perceptrons do*, Neural Networks, 4 (1991) pp. 243–249.