Nonlinear Discrete-Time Observer Design with Linearizable Error Dynamics

MingQing Xiao, Nikolaos Kazantzis, Costas Kravaris, and Arthur J. Krener

***Abstract*—A necessary and sufficient condition for the existence of a discrete-time nonlinear observer with linearizable error dynamics is provided. The result can be applied to any real analytic nonlinear system whose linear part is observable. The necessary and sufficient condition is the solvability of a nonlinear functional equation. Furthermore, the well-known Siegel's theorem on the linearizability of a mapping is naturally reproduced in a corollary. The proposed observer design method is constructive and can be applied approximately to any sufficiently smooth, linearly observable system yielding a local observer with approximately linear error dynamics.**

***Index Terms*—Discrete-time nonlinear system, linearizable error dynamics, nonlinear observer, output injection.**

I. Introduction

We consider the problem of estimating the current state $x(k)$ of a nonlinear discrete-time dynamical system, described by a system of first-order difference equations

$$\begin{aligned} x(k+1) &= f(x(k)) = Fx(k) + O(x(k))^2 \\ y(k) &= h(x(k)) = Hx(k) + O(x(k))^2 \end{aligned} \tag{1.1}$$

from the past observations $y(s)$, $s \leq k$, where the discrete-time index $k \in \{0, 1, 2, \ldots\}$. The vector functions $f : \mathbb{R}^n \to \mathbb{R}^n$, and $h : \mathbb{R}^n \to \mathbb{R}^p$ are assumed to be sufficiently smooth with $f(0) = 0$, $h(0) = 0$ and $p \leq n$. Later, when convergence of a certain formal power series is established, we shall require f, h to be real analytic vector functions.

For the original system (1.1), an observer is a dynamical system driven by the observations $y(k)$

$$\hat{x}(k+1) = \hat{f}(\hat{x}(k), y(k)) \tag{1.2}$$

such that the estimation error $\tilde{x}(k) = x(k) - \hat{x}(k)$ goes to zero as $k \to \infty$. A local observer is one whose estimation error converges to zero for "small" $x(0)$ and $\tilde{x}(0)$.

One technique of constructing an observer for continuous-time systems is to find a nonlinear change of state and output coordinates which transforms the original system (1.1) into a system with linear dynamics driven by nonlinear output injection and with a linear output map (e.g., see [8]). The design of an observer for such a system is relatively easy since the error dynamics can be made linear in the transformed coordinates, thus allowing the subsequent employment of standard linear observer design methods. Recently, Kazantzis and Kravaris [6] proposed a new approach to the continuous-time nonlinear observer design problem, where the error dynamics can be made linear in the transformed coordinates as well. In particular, they seek a nonlinear change of state coordinates which transforms (1.1) into a system with linear dynamics driven by nonlinear output injection but with an arbitrary output map. This design objective is much easier to accomplish and the authors were able to do so for all linearly observable, real analytic systems whose spectrum of the linear part lies wholly in the left half or wholly in the right half of the complex plane (that is, the eigenvalues lie in the Poincaré domain [1]). Krener and Xiao [10]–[12] have extended this approach to linearly observable, real analytic systems where the spectrum of F is arbitrary (specifically in the Siegel domain [1]). They also showed that the sufficient condition for linearizable error dynamics is also necessary [12], and they further extended the method to some systems which are not linearly observable [13]. When the system is only C^r, this approach yields a local observer with approximately linear error dynamics.

The discrete-time nonlinear observer design problem poses a great challenge as well, particularly since digital technology is increasingly used at the implementation stage of model-based control laws or system condition monitoring schemes. Within the area of geometric control theory and in the direction of developing a systematic discrete-time nonlinear observer design methodology, the work in [2], [14], and [16] provided for the first-time a discrete-time analogue of the results obtained in [8]. From a practical point of view, however, it should be pointed out that all the above approaches rely on a set of rather restrictive conditions. Other interesting approaches are reported [5], [15] where some of the restrictive assumptions in [16] are relaxed using past values of the output variable and also in [4], where a discrete-time nonlinear observer design method is proposed that can be viewed as the discrete-time version of the observer presented in [3]. However, restrictive global Lipschitz continuity conditions inevitably arise in the discrete-time version as well. Another notable contribution to the discrete-time nonlinear observer design problem is technically based on Newton's algorithm for the simultaneous solution of a system of nonlinear equations[17]. Under proper interpretation, this discrete-time design method yields an asymptotic observer for a broad class of systems. However, this approach requires the iterative solution of a set of nonlinear equations for each time interval, and the convergence conditions derived can not be easily checked in practice. Recently, Kazantzis and Kravaris [7] extended the approach adopted in [6] to linearly observable, real analytic nonlinear discrete-time systems where the spectrum of the linear part lies wholly inside or wholly outside the unit disc (Poincaré domain [1]). In this note, a new approach to the nonlinear discrete-time observer design problem is introduced where the aforementioned requirement on the spectrum of the system's linear part is relaxed.

The basic steps of the proposed approach are as follows. Suppose that there exists a change of state coordinates $z = \theta(x)$, an output injection map $\beta(y)$ and an $n \times n$ matrix A with eigenvalues strictly inside the unit disc such that the dynamics of (1.1) in the z coordinates is linear and driven by the aforementioned nonlinear output injection

$$z(k+1) = Az(k) + \beta(y(k)). \tag{1.3}$$

Then one can readily design an observer for (1.3) whose dynamic equations are given by

$$\begin{aligned} \hat{z}(k+1) &= A\hat{z}(k) + \beta(y(k)) \\ \hat{x}(k) &= \theta^{-1}(\hat{z}(k)). \end{aligned} \tag{1.4}$$

Notice that under the above construction the error $\tilde{z} = z - \hat{z}$ dynamics becomes linear in the transformed coordinates, and the estimate of the state vector in the original coordinates $\hat{x}(k)$ can be recovered through

Manuscript received January 11, 2002; revised November 25, 2002. Recommended by Associate Editor J. Huang. This work was supported in part by the National Science Foundation under Grant CTS-0131809 and Grant 9907998, and in part by the Air Force Office of Scientific Research under Grant F49620-95-1-0409.

M. Xiao is with the Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 USA (mxiao@math.siu.edu).

N. Kazantzis is with the Department of Chemical Engineering, Worcester Polytechnic Institute, Worcester, MA 01609 USA (nikolas@wpi.edu).

C. Kravaris is with the Department of Chemical Engineering, University of Patras, Patras GR-26500, Greece (kravaris@chemeng.upatras.gr).

A. J. Krener is with the Department of Mathematics University of California, Davis, CA 95616 USA (ajkrener@ucdavis.edu).

Digital Object Identifier 10.1109/TAC.2003.809793

0018-9286/03$17.00 © 2003 IEEE

the inverse transformation map θ^{-1}. The error dynamics satisfies the linear equation

$$\tilde{z}(k+1) = A\tilde{z}(k). \tag{1.5}$$

Since the eigenvalues of A are strictly inside the unit disc we are assured that the error dynamics $\tilde{z}(k) \rightarrow 0$ it exponentially as $k \rightarrow \infty$. Furthermore, notice that in the original coordinates the observer is given by

$$\hat{x}(k+1) = \theta^{-1}\left[\theta(f(\hat{x}(k))) + \beta(y(k)) - \beta(h(\hat{x}(k)))\right]. \tag{1.6}$$

If one can only find a change of state coordinates $z = \theta(x)$, an output injection map $\beta(y)$, and an $n \times n$ Hurwitz matrix A such that the dynamics of (1.1) becomes almost linear driven by nonlinear output injection

$$z(k+1) = Az(k) + \beta\left(y(k)\right) + g\left(z(k)\right) \tag{1.7}$$

where $g(z) = O(z)^{d+1}$, then one can construct a local observer

$$\begin{aligned} \hat{z}(k+1) &= A\hat{z}(k) + \beta\left(y(k)\right) + g\left(\hat{z}(k)\right) \\ \hat{x}(k) &= \theta^{-1}\left(\hat{z}(k)\right) \end{aligned} \tag{1.8}$$

with approximately linear error dynamics

$$\begin{aligned} \tilde{z}(k+1) &= A\tilde{z}\left(k\right) + g(z(k)) - g\left(\hat{z}(k)\right) \Rightarrow \\ \tilde{z}(k+1) &= A\tilde{z}(k) + O\left(z(k), \tilde{z}(k)\right)^{d} O\left(\tilde{z}(k)\right). \end{aligned} \tag{1.9}$$

In this case, the observer in the original coordinates takes the same form (1.6) as before.

To transform the original system (1.1) to (1.3), $\theta(x)$, $\beta(y)$ and A must satisfy the following functional equation:

$$\theta\left(f(x)\right) = A\theta(x) + \beta\left(h(x)\right). \tag{1.10}$$

When the linear part of (1.1) is observable, we shall solve (1.10) by constructing a formal power series for the unknown transformation map $\theta(x)$ up to the degree of smoothness of the system for any suitable $\beta(y)$ and A. We shall show that this series converges if the system and the output injection are real analytic and A is chosen to satisfy a condition that holds almost everywhere. A truncation of the power series of degree d leads to a solution to the following functional equation:

$$\theta\left(f(x)\right) = A\theta(x) + \beta\left(h(x)\right) + g(z). \tag{1.11}$$

where $g(z) = O(z)^{d+1}$ and, hence, $z = \theta(x)$ transforms the system to (1.7). It should be pointed out, that the converse statement holds true as well. In particular, if there exists an observer with linear error dynamics in the transformed coordinates then it is because (1.10) is solvable. Finally, the analysis adopted in the present note is similar in spirit to that of [10], but it has some subtle differences.

II. Main Results

Definition 2.1: Suppose that the eigenvalues of F are $\lambda = (\lambda_1, \ldots, \lambda_n)$. A complex number μ is multiplicatively resonant with the spectrum $\sigma(F)$ of F of degree $d > 0$ if there exist nonnegative integers $(m_1, m_2, \ldots, m_n)$ with $|m| = \sum_k m_k = d$ such that

$$\lambda^m = \lambda_1^{m_1}\lambda_2^{m_2}\ldots\lambda_n^{m_n} = \mu.$$

Definition 2.2: A collection of eigenvalues belongs to the Poincaré domain if the moduli of the eigenvalues are all smaller or all greater than one. The complement of the Poincaré domain is the Siegel domain.

According to the aforementioned definition a system whose eigenvalues of the linear part belong to the Poincaré domain is either asymptotically stable (if $|\lambda| < 1$) or complete unstable (if $|\lambda| > 1$).

Definition 2.3: Suppose that the eigenvalues of F are $\lambda = (\lambda_1, \ldots, \lambda_n)$. Given constants $C > 0$, $v > 0$, we say a complex number μ is of multiplicative type (C, v) with respect to the spectrum $\sigma(F)$ of F if for any vector $m = (m_1, m_2, \ldots, m_n)$ of nonnegative integers, $|m| = \sum_k m_k$, we have

$$|\mu - \lambda^m| \geq \frac{C}{|m|^v}. \tag{2.1}$$

Clearly, if μ is of multiplicative type (C, v) with respect to the spectrum $\sigma(F)$ of F, then it is not multiplicatively resonant with the spectrum of F of any degree. The set of μ's which are of multiplicative type (C, v) for some $C > 0$ is dense in the complex plane when v is large enough (see Lemma 2.1). Indeed, we shall now prove that almost all $\mu \in \mathbb{C}$ are of multiplicative type (C, v) with respect to the spectrum of F. We state the following lemma without proof, which is similar to the one given in [10].

Lemma 2.1: For fixed $v > (n/2)$ and any $C > 0$, define $M(C)$ as the set of all μ which are not of multiplicative type (C, v) with respect to the spectrum $\sigma(F)$ of F for degree $d \geq 1$. Then

$$\text{measure } M(C) \leq k(n, v)C^2$$

where $k(n, v)$ is a constant which depends only on n and v. Therefore

$$\text{measure} \bigcap_{C>0} M(C) = 0$$

Theorem 2.1: Assume that $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$, and $\beta : \mathbb{R}^p \rightarrow \mathbb{R}^n$ are C^{d+1} vector functions with $f(0) = 0$, $h(0) = 0$, $\beta(0) = 0$, and $F = (\partial f)/(\partial x)(0)$, $H = (\partial h)/(\partial x)(0)$, $B = (\partial\beta)/(\partial x)(0)$. Let $A = T(F - T^{-1}BH)T^{-1}$ for some invertible matrix T. Suppose that none of the eigenvalues of A are multiplicatively resonant with the spectrum $\sigma(F)$ of degree $1 \leq k \leq d$. Then, there exists a unique degree d polynomial solution $z = \theta(x)$ to the functional equation (1.11) locally around $x = 0$ with $(\partial\theta)/(\partial x)(0) = T$, so that θ is a local diffeomorphism. If all eigenvalues of A are chosen to be inside the unit disk, then the local state observer for (1.1) given by (1.6) has locally geometrically stable error dynamics (1.9) which is approximately linear in the transformed coordinates.

Proof: We prove Theorem 2.1 by induction on $d \geq 2$. The terms of degree 2 of (1.10) are

$$A\theta^{[2]}(x) = \theta^{[2]}(Fx) = Tf^{[2]}(x) - \beta^{[2]}(x). \tag{2.2}$$

The left-hand side is a linear function of $\theta^{[2]}$ and the right-hand side is a known quantity. In fact, the map

$$\theta^{[2]}(x) \rightarrow A\theta^{[2]}(x) - \theta^{[2]}(Fx) \tag{2.3}$$

is a linear operator on the space of quadratic vector functions. Let $\mathbf{e}^k$ be the k^{th} unit vector in z space. If $m = (m_1, \ldots, m_n)$ then $x^m = x_1^{m_1}\ldots x_m^{m_n}$. The linear operator (2.3) maps $\mathbf{e}^k x_i x_j$ to $(\mu_k - \lambda_i\lambda_j)\mathbf{e}^k x_i x_j$. Hence, if there is no multiplicative resonance of degree 2, the operator (2.3) is invertible. If

$$\theta^{[2]}(x) = \sum_{1\leq k\leq n}\;\sum_{1\leq i\leq j\leq n} \theta_k^{ij}\mathbf{e}^k x_i x_j \tag{2.4}$$

and

$$Tf^{[2]}(x) - \beta^{[2]}(x) = \sum_{1\leq k\leq n}\;\sum_{1\leq i\leq j\leq n} \gamma_k^{ij}\mathbf{e}^k x_i x_j \tag{2.5}$$

then

$$\theta_k^{ij} = \frac{\gamma_k^{ij}}{\mu_k - \lambda_i\lambda_j}. \tag{2.6}$$

Assume that the unique solution

$$\theta(x) = \theta^{[1]}(x) + \theta^{[2]}(x) + \theta^{[3]}(x) + \ldots\theta^{[d-1]}(x)$$

to (1.11) through terms up to $d-1$ has been found where $\theta^{[1]}(x) = Tx$. The next term $\theta^{[d]}(x)$ must satisfy

$$A\theta^{[d]}(x) - \theta^{[d]}(Fx) = \left[\sum_{i=1}^{d-1} \theta^{[i]}(f(x))\right]^{[d]} - \beta^{[d]}(x). \qquad (2.7)$$

The notation $[\ldots]^{[d]}$ means the degree part of the bracketed expression. Again, the left hand side is linear in the unknown $\theta^{[d]}(x)$, and the right-hand side involves only known or previously computed terms. The map

$$\theta^{[d]}(x) \to A\theta^{[d]}(x) - \theta^{[d]}(Fx) \qquad (2.8)$$

is a linear operator on the space of degree d vector functions. It maps $\mathbf{e}^k x^m$ to $(\mu_k - \lambda^m)\mathbf{e}^k x^m$ where $|m| = d$. Hence, if here is no multiplicative resonance of degree d then (2.7) as a unique solution. If

$$\theta^{[d]}(x) = \sum_{1\le k\le n} \sum_{|m|=d} \theta_{k,m}\mathbf{e}^k x^m \qquad (2.9)$$

and

$$\left[\sum_{i=1}^{d-1} \theta^{[i]}(f(x))\right]^{[d]} - \beta^{[d]}(x) = \sum_{1\le k\le n} \sum_{|m|=d} \gamma_{k,m}\mathbf{e}^k x^m \qquad (2.10)$$

then

$$\theta_{k,m} = \frac{\gamma_{k,m}}{\mu_k - \lambda^m}. \qquad (2.11)$$

The rest of Theorem 2.1 follows rather easily from the discussion in the introduction. □

When f, h, and β are real analytic we may apply Theorem 2.1 and conclude that there is an infinite series

$$\theta(x) = \sum_{d=1}^{\infty} \theta^{[d]}(x)$$

which satisfies (1.11) for all d. When θ is constructed, one would expect that a judicious choice may enhance the convergence properties of the series (speed, radius of convergence, etc.). Therefore, the output injection β in general should be chosen such that the first several terms in the series carry more "weight" than higher order terms. Due to space limitations, we only give an outline of the proof of Theorem 2.2.

Theorem 2.2: Assume that $f : \mathbb{R}^n \to \mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}^p$ and $\beta : \mathbb{R}^p \to \mathbb{R}^n$ are real analytic vector functions with $f(0) = 0$, $h(0) = 0$, $\beta(0) = 0$, and $F = (\partial f)/(\partial x)(0)$, $H = (\partial h)/(\partial x)(0)$, $B = (\partial\beta)/(\partial x)(0)$. Let $A = T(F - T^{-1}BH)T^{-1}$ for some invertible matrix T. Suppose there exists a $C > 0$, $v > 0$ such that all the eigenvalues of A are of multiplicative type (C, v) with respect to the spectrum $\sigma(F)$ of F. Then there exists a unique analytic solution $z = \theta(x)$ to the functional equation (1.10) locally around $x = 0$ with $(\partial\theta)/(\partial x)(0) = T$, so that θ is a local diffeomorphism. If all eigenvalues of A are chosen to be inside the unit disk, then the local state observer for (1.1) given by (1.6) has locally geometrically stable error dynamics (1.5) which is linear in the transformed coordinates.

Outline of the Proof: Let

$$\begin{aligned} f(x) &= Fx + \bar{f}(x) \\ \beta(y) &= BHx + \bar{\beta}(x). \end{aligned}$$

As before, we assume that x coordinates can be chosen so that F becomes diagonal. Furthermore, we choose B to appropriately set the spectrum of A. The rest of the output injection $\bar{\beta}(x)$ is an arbitrary real analytic function. We also assume that T can be chosen so that A is diagonal. We now construct a sequence of functions $\{\theta_k(x)\}_{1=2}^{\infty}$ which satisfies the following system of homological equations:

$$\begin{aligned} \theta_1(x) &= Tx \\ A\theta_2(x) - \theta_2(Fx) &= T\bar{f}(x) - \bar{\beta}(x) \\ A\theta_k(x) - \theta_k(Fx) &= \theta_{k-1}(f(x)) - \theta_{k-1}(Fx), \qquad k \ge 3. \end{aligned} \qquad (2.12)$$

Clearly $\theta_2(x)$ starts with terms of degree two and it is easy to show by induction that $\theta_k(x)$ starts with terms of degree k. We define a family of positive real functions $b_k : [0,1) \to [0,\infty)$ to be

$$b_k(q) := \max_{d\in \mathrm{Z}_{\ge 0}, d\ge k} \left[C^{-1}d^v q^{\frac{d}{2}}\right], \qquad k \in \{1,2,3,\ldots,\} = \mathbf{Z}_+$$

where $C > 0$ and $v > 0$ are given. Suppose that $\Phi(x) = (\phi_1(x),\ldots,\phi_n(x))$ is analytic in $|x| < r$ and

$$\phi_i(x) = \sum_m \phi_{i,m}x^m, \qquad i = 1,2,\ldots,n \qquad (2.13)$$

We denote

$$\|\Phi(x)\| := \max_{1\le i\le n} \sum_m |\phi_{i,m}x^m|.$$

One can show that if all of the eigenvalues of A are of type (C, v) with respect to the spectrum $\sigma(F)$ of F, then there exists a sequence of analytic functions $\{\theta_k(x)\}_{k=2}^{\infty}$ defined in $\mathbb{R}^n$ which solve the system of homological equations (2.12). The next step is to prove that

$$Tx + \theta_2(x) + \theta_3(x) + \cdots + \theta_k(x) + \cdots$$

converges near the origin. Recall that $f(x) = Fx + \bar{f}(x)$. Since $\bar{f}(x) = O(|x|^2)$ is an analytic function in the polydisk $\{|x| \le r\}$, it can be expanded into a Taylor series

$$\bar{f}(x) = f^{[2]}(x) + f^{[3]}(x) + \ldots \quad |x| \le r$$

where $f^{[d]}(x) = \sum_{j=1}^{n}\sum_{|m|=d} f_{j,m}\mathbf{e}^j x^m$. Thus, the following series converges:

$$\sum_{|m|=2} |f_{j,m}|r^2 + \sum_{|m|=3} |f_{j,m}|r^3 + \cdots := M_j$$

for $j = 1,2,\ldots,n$. We now define

$$\bar{M}_f := \max\left\{\frac{M_1}{r^2},\ldots,\frac{M_n}{r^2}\right\}.$$

A direct estimation leads to the fact that there exist $r_1 \le r$ and a continuous function $H(x) > 0$ which is defined in $|x| < r_1$ such that

$$\|\theta_k(x)\| \le \bar{M}_f H(x).$$

Then, we let $r_2 := r_1/2$ and

$$\hat{N} := \max_{|x|\le r_2} H(x).$$

and

$$M := \max_{|x|\le r_2} \|\theta_2(x)\|. \qquad (2.14)$$

It can be shown that

$$\|\theta_k(x)\| \le b_k(q)\hat{N}^{k-3}M, \text{ for } k = 3,4,5, \text{ and } 0 < q < 1\ldots. \qquad (2.15)$$

Hence, there exists a r_3 with $r_2 \leq r_3$ such that (1.10) converges in $|x| < r_3$. □

Corollary 2.1 (Siegel's Mapping Theorem): Assume that $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a real analytic vector function with $f(0) = 0$ and $F = (\partial f)/(\partial x)(0)$. Assume that the eigenvalues of F are of multiplicative type (C, v) for $|m| > 1$ with respect to the spectrum of F. Then, there is a real analytic diffeomorphism satisfying (1.10) with $A = F$ and $\beta = 0$.

Proof: Apply Theorem 2.2 with $\beta = 0$ and $A = F$. □

Theorem 2.3: Conversely, assume that $f : \mathbb{R}^n \rightarrow \mathbb{R}^n, h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ are C^{d+1} vector functions with $f(0) = 0, h(0) = 0$. If there exists an observer (1.2) and a C^{d+1} change of coordinates $z = \theta(x)$ such that the error dynamics (1.5) in the transformed coordinates is linear, then θ satisfies (1.10) for some A and C^{d+1} output injection $\beta(y)$. If the system and the change of coordinates are real analytic then $\beta(y)$ is real analytic as well.

Proof: Suppose that the change of coordinates $z = \theta(x)$ transforms the original system (1.1) to the following one:

$$\begin{aligned} z(k+1) &= g\left(z(k)\right) \\ y(k) &= h\left(\theta^{-1}(z(k))\right). \end{aligned} \tag{2.16}$$

Suppose there exists a nonlinear observer

$$\hat{z}(k+1) = \hat{g}\left(\hat{z}(k), y(k)\right) \tag{2.17}$$

such that the error dynamics is linear and in the form of (1.5). Recall that $\tilde{z} = z - \hat{z}$. For $z = 0$, (1.5) yields $A\hat{z} = \hat{g}(\hat{z}, 0)$ since $z = 0$ implies $y = 0$. On the other hand, for $\tilde{z} = 0$ (1.5) yields $g(z) = \hat{g}(z, h(z))$. We now define

$$\beta(\hat{z}, y) := \hat{g}(\hat{z}, y) - \hat{g}(\hat{z}, 0)$$

then

$$\begin{aligned} A\tilde{z} &= g(z) - \hat{g}\left(\hat{z}, h(z)\right) - \hat{g}\left(z, h(z)\right) - \hat{g}\left(\hat{z}, h(z)\right) \\ &= \hat{g}(z, 0) + \beta\left(z, h(z)\right) - \hat{g}(\hat{z}, 0) - \beta\left(\hat{z}, h(z)\right) \\ &= Az + \beta\left(z, h(z)\right) - A\hat{z} - \beta\left(\hat{z}, h(z)\right). \end{aligned}$$

and thus $\beta(z, h(z)) = \beta(\hat{z}, h(z))$. Notice, that the left hand side of the previous equality does not depend on $\hat{z}$ and, therefore, the right-hand side does not as well. Hence, $\beta(z, y) = \beta(y)$. Therefore, we have

$$\begin{aligned} \hat{g}(\hat{z}, 0) &= A\hat{z} + \beta(y) \\ g(z) &= Az + \beta\left(h(z)\right) \end{aligned}$$

which indicates that $z = \theta(x)$ must satisfy the functional equation (1.10). Clearly, $\beta(y)$ is a C^{d+1} function. □

III. Illustrative Example

Example 1: Consider the following nonlinear discrete-time dynamic system:

$$\begin{aligned} y(k+1) &= \frac{0.5\left(\frac{y(k)}{1+y(k)}\right) - 0.9w(k)}{1 - 0.5\left(\frac{y(k)}{1+y(k)}\right) + 0.9w(k)} \\ w(k+1) &= w(k) \end{aligned} \tag{3.18}$$

with y being the measured state variable and w an unknown constant parameter or disturbance term that needs to be estimated. The Jacobian matrix F of (3.18) evaluated at the origin is

$$F = \begin{pmatrix} 0.5 & -0.9 \\ 0 & 1 \end{pmatrix}. \tag{3.19}$$

The eigenvalues of F are: $\lambda_1 = 0.5$ and $\lambda_2 = 1$. Notice that the nonlinear discrete-time observer design method developed by Kazantzis and Kravaris [7] can not be applied to this case, since one of the eigenvalues of F lies on the unit circle. Notice also that the system's linear part is observable.

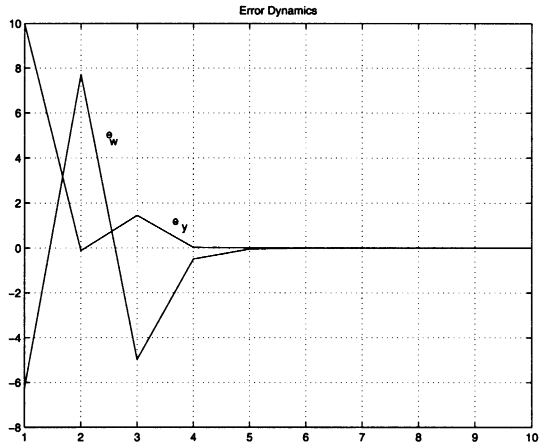

Fig. 1. Error dynamics for $\boldsymbol{k = 10}$.

Consider now the following matrix A and output injection:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 0.1 \end{pmatrix} \quad \beta(y) = \begin{pmatrix} 0.5\left(\frac{y(k)}{1+y(k)}\right) \\ \frac{y(k)}{1+y(k)} \end{pmatrix}. \tag{3.20}$$

When expanding β, one can find that functional equation (1.10) admits a unique analytic solution

$$\begin{aligned} \theta_1(y, w) &= \frac{y}{1+y} + 0.9w \\ \theta_2(y, w) &= \frac{5}{2}\left(\frac{y}{1+y} + w\right). \end{aligned} \tag{3.21}$$

The proposed discrete-time observer is then given by the following dynamic equations:

$$\begin{aligned} \hat{z}_1(k+1) &= 0.5\left(\frac{y(k)}{1+y(k)}\right) \\ \hat{z}_2(k+1) &= 0.1\hat{z}_2(k) + \frac{y(k)}{1+y(k)} \\ \hat{y}(k) &= \frac{10\hat{z}_1(k) - 3.6\hat{z}_2(k)}{1 - 10\hat{z}_1(k) + 3.6\hat{z}_2(k)} \\ \hat{w}(k) &= 4\hat{z}_2 - 10\hat{z}_1(k). \end{aligned} \tag{3.22}$$

Let $e_y(k) := y(k) - \hat{y}(k)$ and $e_w := w(k) - \hat{w}(k)$. Then, the simulation of error dynamics is shown in Fig. 1, where $y(1) = 10$, $w(1) = -2\pi$ and $\hat{z}_1(1) = \hat{z}_2(1) = 1$.

References

[1] V. I. Arnol'd, *Geometrical Methods in the Theory of Ordinary Differential Equations*. Berlin, Germany: Springer-Verlag, 1988.

[2] S. T. Chung and J. W. Grizzle, "Sampled-data observer error linearization," *Automatica*, vol. 26, p. 997, 1990.

[3] G. Ciccarela, M. Dalla Mora, and A. Germani, "A Luenberger-like observer for nonlinear systems," *Int. J. Control*, vol. 57, p. 537, 1993.

[4] ——, "Observers for discrete-time nonlinear systems," *Syst. Control Lett.*, vol. 20, p. 373, 1993.

[5] H. J. C. Huijberts, T. Lilge, and H. Nijmeijer, "Nonlinear discrete-time synchronization via extended observers," *Int. J. Bifur. Chaos.*, vol. 11, pp. 1997–2006, 2001.

[6] N. Kazantzis and C. Kravaris, "Nonlinear observer design using Lyapunov's auxiliary theorem," *Syst. Control Lett.*, vol. 34, pp. 241–247, 1998.

[7] ——, "Discrete-time nonlinear observer design using functional equations," *Syst. Control Lett.*, vol. 42, pp. 81–94, 2001.

[8] A. J. Krener and A. Isidori, "Linearization by output injection and nonlinear observers," *Syst. Control Lett.*, no. 3, pp. 47–52, 1983.

[9] A. J. Krener, "Nonlinear stabilizability and detectability," in *In Systems and Networks: Mathematical Theory and Applications*, U. Helmke, R. Mennicken, and J. Saurer, Eds. Berlin, Germany: Akademie-Verlag, 1994, pp. 231–250.

[10] A. J. Krener and M. Xiao, "Nonlinear observer design in the Siegel domain," *SIAM J. Control Optim.*, vol. 41, no. 2, pp. 932–953, 2002.

[11] ——, "Nonlinear observer design in the Siegel domain through coordinate transformation," in *Proc. 5th Int. Fed. Automatic Control*, St. Petersburg, Russia, 2001, pp. 557–562.

[12] ——, "Necessary and sufficient condition for nonlinear observer with linearizable error dynamics," presented at the *40th IEEE Conf. Decision Control*, Orlando, FL, 2001.

[13] ——, "Observers for linearly unobservable nonlinear systems," *Syst. Control Lett.*, vol. 46, pp. 281–288, 2002.

[14] W. Lee and K. Nam, "Observer design for autonomous discrete-time nonlinear systems," *Syst. Control Lett.*, vol. 17, p. 49, 1991.

[15] T. Lilge, "On observer design for nonlinear discrete-time systems," *Eur. J. Control*, vol. 4, pp. 306–319, 1998.

[16] W. Lin and C. I. Byrnes, "Remarks on linearization of discrete-time autonomous systems and nonlinear observer design," *Syst. Control Lett.*, vol. 25, p. 31, 1995.

[17] P. E. Moraal and J. E. Grizzle, "Observer design for nonlinear systems with discrete-time measurements," *IEEE Trans. Automat. Contr.*, vol. 40, p. 395, Mar. 1995.

Stability of Linear Discrete Dynamics Employing State Saturation Arithmetic

Tatsushi Ooba

***Abstract*—This note is concerned with the stability of discrete-time dynamical systems employing saturation arithmetic in the state–space. A matrix measure is introduced so that it can administer the proximity evaluation of a matrix to the set of diagonal matrices, and the measure is utilized for making an additional condition to the Lyapunov–Stein matrix inequality. The solvability of the modified matrix inequality ensures not only the stability but also the absence of overflow oscillation under the state saturation arithmetic, and this approach has the advantage of being free from auxiliary parameters. As an application, the obtained result is applied to the stability analysis of two-dimensional dynamics. Numerical examples are given to illustrate the results.**

***Index Terms*—Free overflow oscillation, matrix inequalities, stability of discrete-time dynamical systems, state saturation nonlinearity.**

I. Introduction

Saturation is one of the familiar nonlinear phenomena observed in the real world. The stability of dynamical systems employing saturation arithmetic is therefore thought to be an important object of system theoretic study. Consider the linear discrete-time dynamical equation

$$x(t) = Ax(t-1) \tag{1.1}$$

where $x(\cdot) \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. With any $x = (\xi_1, \ldots, \xi_n)^T$, we associate a rectangular region

$$\Theta\left((\xi_1, \ldots, \xi_n)^T\right) = \left\{(\theta_1\xi_1, \ldots, \theta_n\xi_n)^T;\ 0 \le \theta_i \le 1,\ i = 1, \ldots, n\right\}. \tag{1.2}$$

A map $V : \mathbb{R}^n \to \mathbb{R}^n$ is said to be *a saturation arithmetic map* if it satisfies $V(x) \in \Theta(x)$ for any $x \in \mathbb{R}^n$. For the sake of convenience, we denote by $\mathfrak{D}^{\mathfrak{n}}$ the set of all saturation arithmetic maps on $\mathbb{R}^n$. We will consider the following dynamics with saturation arithmetic:

$$x(t) = V_t\left(Ax(t-1)\right) \qquad V_t \in \mathfrak{D}^{\mathfrak{n}}. \tag{1.3}$$

Let us denote by $\|\cdot\|_2$ the Euclidian norm of the vector involved. It is readily seen from (1.2) that any map $V \in \mathfrak{D}^{\mathfrak{n}}$ brings about the relation $\|Vx\|_2 \le \|x\|_2$ for any $x \in \mathbb{R}^n$. In spite of the apparent contractive property, the occurrence of saturation arithmetic in the state–space is a distress to dynamical systems because it is capable of generating divergent state sequences in a nominal stable dynamics if the worst happens. Such wandering state behavior caused by saturation arithmetics is often called *overflow oscillation*.

Definition 1.1: The dynamics of (1.1) is said to be *free from overflow oscillation* if any state evolution $\{x(t)\}_{t=0,1,2,\ldots}$ through (1.3) converges to zero no matter what map V_t is taken from $\mathfrak{D}^{\mathfrak{n}}$ at each t.

It is widely known as the Mills–Mullis–Roberts criterion [1] that (1.1) is free from overflow oscillation if A is diagonally stable (A matrix A is said to be diagonally stable if there exists a positive diagonal matrix D such that $D - A^TDA > 0$). We refer to [2]–[6] and their references for details about the diagonal stability of matrices. Although the diagonal stability is a pithy notion, it is too stiff a condition for ensuring the absence of overflow oscillation in (1.3). A less conservative result is known as the Singh's criterion [7], on which we can check the absence of overflow oscillation by confirming the existence of a positive–definite matrix P and a positive–diagonal matrix C satisfying

$$\begin{pmatrix} P & -A^TC \\ -CA & 2C - P \end{pmatrix} > 0. \tag{1.4}$$

As can be seen from the proof of the previous result, the matrix P in (1.4) necessarily serves as the kernel of a quadratic Lyapunov function for (1.1). A rational way of interpreting the composite matrix inequality (1.4) is, therefore, to regard it as an additional condition attached to the solutions to the Lyapunov–Stein matrix inequality $P - A^TPA > 0$. It can be said within this context that the parameters in the matrix C are just auxiliaries which help search for a specific solution to the Lyapunov–Stein matrix inequality. In any case, the search for (P, C) is by no means an easy task. So one might wish that the search for C could be replaced with some decisive measuring. The motivation for writing this note is the author's wish to draw up a stability condition based firmly on the Lyapunov–Stein matrix inequality in which we need not conduct a search of auxiliary parameters. In Section II, we will propose a new condition for ensuring the absence of overflow oscillation in (1.3) which fits for the intended purpose.

Throughout this note, we use the following notation. The matrix inequality $P > 0$ (resp. $P \ge 0$) means that P is a positive–definite (respectively, positive–semidefinite) matrix. The notation x^T and A^T mean their transpose, $\|x\|_P$ means the quadratic norm of x weighted by $P > 0$, i.e., $\|x\|_P = (x^TPx)^{1/2}$. The real valued functions $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are applied to any matrix whose eigenvalues are all real numbers and these two functions represent the maximum eigenvalue and the minimum eigenvalue respectively. Notice that the product of a pair of symmetric matrices has its eigenvalues in all real numbers if one of the matrices is positive semidefinite.

II. Main Results

In this section, we present a matrix inequality whose solvability ensures the absence of overflow oscillation in (1.3). The result provides an alternative to the Singh's criterion. We make an obvious assumption that the spectral radius of A, $\rho(A)$, is less than unity. Let δ be such that

Manuscript received June 19, 2002; revised November 4, 2002. Recommended by Associate Editor Z. Lin.

The author is with the Department of Mechanical Engineering, Nagoya Institute of Showa-ku, Nagoya 466-8555, Japan (e-mail: ooba@cjn.or.jp).

Digital Object Identifier 10.1109/TAC.2003.809792