Nonlinear Discrete-Time Observer Design with Linearizable Error Dynamics

MingQing Xiao
Department of Mathematics
Southern Illinois University
Carbondale, IL 62901/USA
mxiao@math.siu.edu

Nikolaos Kazantzis
Department of Chemical Engineering
Worcester Polytechnic Institute
Worcester, MA 01609 /USA
nikolas@wpi.edu

Costas Kravaris
Department of Chemical Engineering
University of Patras
GR-26500/Greece
kravaris@chemeng.upatras.gr

Arthur J. Krener
Department of Mathematics
University of California
Davis, CA 95616/USA
ajkrener@ucdavis.edu

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 derived is associated with the solvability of a nonlinear functional equation. A corollary is Siegel's theorem on the linearizability of a mapping. The method of observer design suggested by this theorem is constructive and can be applied approximately to any sufficiently smooth, linearly observable system yielding a local observer with approximately linear error dynamics.

Keywords: Discrete Time Nonlinear system, Nonlinear observer, Linearizable Error Dynamics, Output Injection, Functional Equations.

1 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

$$x(k+1) = f(x(k)) = Fx(k) + O(x(k))^2 \tag{1.1}$$

$$y(k) = h(x(k)) = Hx(k) + O(x(k))^2, \tag{1.2}$$

from the past observations $y(s), s \leq k$, where $k \in \{0, 1, 2, \dots\}$. The vector fields $f : I\!R^n \to I\!R^n$, and $h : I\!R^n \to I\!R^p$ are assumed to be sufficiently smooth with $f(0) = 0,\ h(0) = 0$ and $p \leq n$. Later, when showing convergence of a certain formal power series, we shall require that f, h are real analytic functions.

An observer is a dynamic system driven by the observations

$$\hat{x}(k+1) = \hat{f}(\hat{x}(k), y(k)) \tag{1.3}$$

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 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 system (1.2) into a system with linear dynamics driven by nonlinear output injection and with linear output map [10], [8], [2]. The design of an observer for such a system is relatively easy since the error dynamics can be made linear in the transformed coordinates. This approach was extended to discrete time systems in [16] , [17], [3]. For more on discrete time observer design we refer the reader to [5] and its references.

Recently, Kazantzis and Kravaris [6] proposed a new approach to finding a continuous-time observer whose error dynamics is also linear in the transformed coordinates. They seek a nonlinear change of state coordinates which transforms the system (1.2) into a system with linear dynamics driven by nonlinear output injection and with an arbitrary output map. This is much easier to accomplish and they were able to do so for all linearly observable, real analytic systems whose spectrum of the linear part lies wholly in the left half plane or wholly in the right half plane.

Krener and Xiao [12], [13] have extended this approach to linearly observable, real analytic systems where the spectrum of the linear part is arbitrary. They also showed that the sufficient condition for linearizable error dynamics is necessary [14] and extended the method to some systems which are not linearly observable [15]. When the system is only C^r, this approach yields a local observer with approximately linear error dynamics.

Kazantzis and Kravaris [7] have extended this approach to linearly observable, real analytic nonlinear discrete-time systems where the spectrum of the linear part lies wholly inside the unit disc or wholly outside the unit disc. The current paper drops the requirement on the spectrum of the linear part.

The basic approach is as follows. Suppose that there exists a change of state coordinates $z = \theta(x)$, an output injection term $\beta(y)$ and an $n \times n$ matrix A with eigenvalues strictly inside the unit disc such that the dynamics of (1.2) in the z-coordinates is linear and driven by the above nonlinear output injection term

$$z(k+1) = Az(k) + \beta(y(k)). \tag{1.4}$$

Then, one can construct an observer given by

$$\begin{aligned} \hat{z}(k+1) &= A\hat{z}(k) + \beta(y(k)) \\ \hat{x}(k) &= \theta^{-1}(\hat{z}(k)) \end{aligned} \tag{1.5}$$

whose error $\tilde{z} = z - \hat{z}$ dynamics is linear in the transformed coordinates

$$\tilde{z}(k+1) = A\tilde{z}(k). \tag{1.6}$$

Since the eigenvalues of A are strictly inside the unit disc we are assured that $\tilde{z}(k) \to 0$ as $k \to \infty$. Notice, that in the original coordinates the observer is given by

$$\hat{x}(k+1) = \theta^{-1}\left[\theta(f(\hat{x}(k))) + \beta(y) - \beta(h(\hat{x}(k)))\right]. \tag{1.7}$$

If one can only find a change of state coordinates $z = \theta(x)$, an output injection $\beta(y)$ and an $n \times n$ Hurwitz matrix A such that the dynamics of (1.2) becomes almost linear driven by the above nonlinear output injection term

$$z(k+1) = Az(k) + \beta(y(k)) + g(z(k)) \tag{1.8}$$

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

$$\begin{aligned} \hat{z}(k+1) &= A\hat{z}(k) + \beta(y(k)) + g(\hat{z}(k)) \\ \hat{x}(k) &= \theta^{-1}(\hat{z}(k)) \end{aligned} \tag{1.9}$$

with approximately linear error dynamics

$$\begin{aligned} \tilde{z}(k+1) &= A\tilde{z}(k) + g(z(k)) - g(\hat{z}(k)) \\ \tilde{z}(k+1) &= A\tilde{z}(k) + O(z(k), \hat{z}(k))^d O(\tilde{z}(k)). \end{aligned} \tag{1.10}$$

It should be pointed out, that the observer in the original coordinates takes the same form (1.7) as before.

To transform the original system to (1.4), $\theta(x)$, $\beta(y)$ and A must satisfy the following functional equation:

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

When the linear part of the system (1.2) is observable, we shall solve (1.11) by constructing a formal power series for θ 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. Truncating the power series at degree d yields a solution to

$$\theta(f(x)) = A\theta(x) + \beta(h(x)) + g(z). \tag{1.12}$$

where $g(z) = O(z)^{d+1}$ and hence $z = \theta(x)$ transforms the original system to (1.8). The analysis is similar to that of [12] but it has some subtle differences.

Moreover the converse holds. If there exists an observer with linear error dynamics in the transformed coordinates, then it is because (1.11) is solvable.

The paper is organized as follows. Section 2 states the main results of this paper and defines some needed concepts. The formal solution of (1.11) is given in Section 3. Furthermore, it is shown that the formal solution exists to a degree of differentiability that is equal to the degree of differentiability of the original system minus one. Section 4 provides an outline of the proof of convergence of the formal solution for real analytic systems. We also show that (1.11) has a unique solution for any choice of $\beta(y)$ as long as the eigenvalues are outside a set of zero measure in $\mathbb{C}$. The necessity is proved in Section 4. Two examples are provided in the last section to illustrate the main results of the paper.

2 Statement of Main Results

Before we state the main results we must introduce some concepts.

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

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

Definition 2.2. *Suppose that the eigenvalues of F are $\lambda = (\lambda_1, \ldots, \lambda_n)$. Given constants $C > 0,\ \nu > 0$, we say a complex number μ is of multiplicative type (C, ν) with respect to the spectrum of F if for any vector $m = (m_1, m_2, \ldots, m_n)$ of nonnegative integers, $|m| = \sum_k m_k$, we have*

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

Clearly if μ is of multiplicative type (C, ν) with respect to $\sigma(F)$ 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, ν) for some $C > 0$ is dense in the complex plane when ν is large enough (see Theorem 4.1). Notice, that occasionally we delete the modifiers "multiplicative" and "multiplicatively".

The following Theorems capture the present study's main results.

Theorem 2.1. *Assume that $f : I\!R^n \to I\!R^n, h : I\!R^n \to I\!R^p$ and $\beta : I\!R^p \to I\!R^n$ are C^{d+1} vector fields with $f(0) = 0,\ h(0) = 0,\ \beta(0) = 0$ and $F = \frac{\partial f}{\partial x}(0),\ H = \frac{\partial h}{\partial x}(0),\ B = \frac{\partial \beta}{\partial x}(0)$. 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 F of degree $1 \leq k \leq d$. Then there exists a unique degree d polynomial solution $z = \theta(x)$ to the functional equation (1.12) defined locally around $x = 0$ with $\frac{\partial \theta}{\partial x}(0) = T$, 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 (1.2) given by (1.7) has locally geometrically stable error dynamics which is approximately linear in the transformed coordinates (1.10).*

Theorem 2.2. *Assume that $f : I\!R^n \to I\!R^n, h : I\!R^n \to I\!R^p$ and $\beta : I\!R^p \to I\!R^n$ are real analytic vector fields with $f(0) = 0,\ h(0) = 0,\ \beta(0) = 0$ and $F = \frac{\partial f}{\partial x}(0),\ H = \frac{\partial h}{\partial x}(0),\ B = \frac{\partial \beta}{\partial x}(0)$. Let $A = T(F - T^{-1}BH)T^{-1}$ for some invertible matrix T. Suppose that there exists a $C > 0, \nu > 0$ such that all the eigenvalues of A are of multiplicative type (C, ν) with respect to the spectrum $\sigma(F)$ of F. Then, there exists a unique analytic solution $z = \theta(x)$ to the functional equation (1.11) defined locally around $x = 0$ with $\frac{\partial \theta}{\partial x}(0) = T$, 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 (1.2) given by (1.7) has locally geometrically stable error dynamics which is linear in the transformed coordinates (1.6).*

Theorem 2.3. *Conversely, assume that $f : I\!R^n \to I\!R^n, h : I\!R^n \to I\!R^p$ are C^{d+1} vector fields with $f(0) = 0,\ h(0) = 0$. If there exists an observer (1.3) and a C^{d+1} change of coordinates $z = \theta(x)$ such that the error dynamics in transformed coordinates is linear (1.6), then θ satisfies (1.11) for some A and a C^{d+1} output injection term $\beta(y)$. If the system and the change of coordinates are real analytic, then $\beta(y)$ is real analytic as well.*

Remarks:

We have stated the above Theorems assuming (1.2) is real but all the results hold for complex systems too. If θ is a locally diffeomorphic solution to (1.11) for some A and β and the spectra of F and A are disjoint, then it is not hard to show that (H, F) is an observable pair and (A, B) is a controllable pair [12]. On the other hand, if (H, F) is an observable pair then one can set the spectrum of $A = T(F - T^{-1}BH)T^{-1}$ arbitrarily by an appropriate choice of B.

It is with no loss of generality that we can fix $\dfrac{\partial\theta}{\partial x}(0)$ to be an arbitrary invertible matrix T. If $\bar{\theta}$ is a local diffeomorphism and $\dfrac{\partial\bar{\theta}}{\partial x}(0) = \bar{T}$, then $\bar{T}$ is invertible, so we can define $\theta = T\bar{T}^{-1}\bar{\theta}$. Then $\dfrac{\partial\theta}{\partial x}(0) = T$. In the proofs it will be convenient to consider a particular invertible matrix T. Of course changing T changes A because $A = T(F - T^{-1}BH)T^{-1}$.

According to Theorem 4.1 of this paper, almost all complex numbers are of type (C, ν) with respect to $\sigma(F)$ for degree 1 and higher, so this assumption on the spectrum of A and the corresponding one of no resonances are hardly restrictions on the choice of A when (H, F) is an observable pair.

It should be also emphasized, that if we let $\beta = 0$, and $A = F$, then Theorem 2.2 reduces to the well-known *Siegel mapping theorem* (Page 214, [1]) which we now state in accordance to the present context of analysis.

Siegel Mapping Theorem *Assume that $f : I\!R^n \to I\!R^n$ is a real analytic vector field with $f(0) = 0$ and $F = \dfrac{\partial f}{\partial x}(0)$. Assume that the eigenvalues of F are of multiplicative type (C, ν) for $|m| > 1$ with respect to each other, i.e. the spectrum $\sigma(F)$ of F. Then, there is a real analytic diffeomorphism satisfying functional equation (1.11) with $A = F$ and $\beta = 0$.*

Notation:

In this paper we consider vectors $x = (x_1, x_2, \ldots, x_n)$ with n real components x_k or equivalently points x in $I\!R^n$. Let $|x| = \max\{|x_1|, \ldots, |x_n|\}$. The vectors $m = (m_1, m_2, \ldots, m_n)$ whose components are non-negative integers are called multi-indices and form a subset $\mathbf{Z}^n_{\geq 0}$ of $I\!R^n$. We define for $m \in \mathbf{Z}^n_{\geq 0}$ the scalars

$$|m| = m_1 + m_2 + \cdots + m_n; \qquad m! = m_1!m_2!\cdots m_n!,$$

and for $x \in I\!R^n$, $m \in \mathbf{Z}^n_{\geq 0}$ the monomial

$$x^m = x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}.$$

In $\mathbf{Z}^n_{\geq 0}$ we define a partial ordering by writing

$$m \geq \alpha \qquad \text{whenever} \quad m_i \geq \alpha_i \qquad \text{for } i = 1, 2, \ldots, n.$$

The general d-th degree polynomial in x_1, x_2, ... , x_n is of the form

$$P(x) = \sum_{|m| \leq d} p_m x^m$$

with $m \in \mathbf{Z}^n_{\geq 0}$, $x \in I\!R^n$, $p_m \in I\!R$.

Let $f(x)$ be a function whose domain is an open set Ω in $I\!R^n$ and whose range lies in $I\!R$. For $y \in \Omega$ we call f real analytic at y if there exist $f_m \in I\!R$ and a neighborhood N of y such that

$$f(x) = \sum_m f_m (x - y)^m \tag{2.2}$$

for all x in N, where the sum is taken over all multi-indices m. The convergence of (2.2) means *absolute convergence* for each $x \in N$. We say f is real analytic in Ω if f is real analytic at each $y \in \Omega$. A vector $f(x) = (f_1(x), \ldots, f_n(x))$ defined in Ω is called real analytic if each of its components is real analytic. Convergence then means *component-wise* absolute convergence.

3 The Formal Solution of the Functional Equation

In this section we shall prove Theorem 2.1. We expand the terms of (1.11) in power series as follows

$$\begin{aligned} f(x) &= Fx + f^{[2]}(x) + f^{[3]}(x) + \cdots + f^{[d]}(x) + O(x)^{d+1}, \\ \beta(h(x)) &= BHx + \beta^{[2]}(x) + \beta^{[3]}(x) + \cdots + \beta^{[d]}(x) + O(x)^{d+1}, \\ \theta(x) &= Tx + \theta^{[2]}(x) + \theta^{[3]}(x) + \cdots + \theta^{[d]}(x), \end{aligned}$$

where $f^{[d]}$, $\beta^{[d]}$, and $\theta^{[d]}$ are homogeneous polynomial vector fields of degree d in x. It is assumed, that $\beta(y), T$ have been chosen such that $A = T(F - T^{-1}BH)T^{-1}$. Therefore, the known quantities are $f,\ h,\ \beta,\ A$ and the unknown ones are the higher degree terms $\theta^{[2]}, \theta^{[3]}, \ldots$. Since $\{F,\ H\}$ is a linearly observable pair we can choose B to fix the spectrum of A. We assume that we can choose x coordinates so that F is diagonal. When the Jordan form of F is not diagonal the proofs below need to be modified. In this case, the proofs become algebraically involved, and thus, they are omitted for brevity. We assume that we can choose z coordinates so that A is also diagonal by an appropriate choice of T. We can always choose A to have distinct eigenvalues and hence its Jordan form will be diagonal. Complex eigenvalues introduce no difficulties. In fact all the results hold for complex $x, z, f, h, A, \beta, \theta, \ldots$ as well.

Proof of Theorem 2.1 The proof is based on induction for $d \geq 2$. The terms of degree 2 of (1.11) are

$$A\theta^{[2]}(x) - \theta^{[2]}(Fx) = Tf^{[2]}(x) - \beta^{[2]}(x). \tag{3.1}$$

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

$$\theta^{[2]}(x) \to A\theta^{[2]}(x) - \theta^{[2]}(Fx) \tag{3.2}$$

is a linear operator on the space of quadratic vector fields. 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 (3.2) 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 (3.2) is invertible. If

$$\theta^{[2]}(x) = \sum_{1 \leq k \leq n} \sum_{1 \leq i \leq j \leq n} \theta_k^{ij} \mathbf{e}^k x_i x_j \tag{3.3}$$

and

$$Tf^{[2]}(x) - \beta^{[2]}(x) = \sum_{1 \leq k \leq n} \sum_{1 \leq i \leq j \leq n} \gamma_k^{ij} \mathbf{e}^k x_i x_j \tag{3.4}$$

then

$$\theta_k^{ij} = \frac{\gamma_k^{ij}}{\mu_k - \lambda_i\lambda_j}. \tag{3.5}$$

Assume that the unique solution

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

to (1.12) through terms up to $d-1$ has been found where $\theta^{[1]}(x) = Tx$. The next term $\theta^{[d]}(x)$ must satisfy

$$A\theta^{[d]}(x) - \theta^{[d]}(Fx) = \left[\sum_{i=1}^{d-1} \theta^{[i]}(f(x))\right]^{[d]} - \beta^{[d]}(x). \tag{3.6}$$

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

$$\theta^{[d]}(x) \to A\theta^{[d]}(x) - \theta^{[d]}(Fx) \tag{3.7}$$

is a linear operator on the space of degree d vector fields. In particular, it maps $\mathbf{e}^k x^m$ to $(\mu_k - \lambda^m)\mathbf{e}^k x^m$ where $|m| = d$. Hence if there is no multiplicative resonance of degree d, then (3.6) has a unique solution. If

$$\theta^{[d]}(x) \quad = \quad \sum_{1\le k\le n} \sum_{|m|=d} \theta_{k,m}\mathbf{e}^k x^m \tag{3.8}$$

and

$$\left[\sum_{i=1}^{d-1} \theta^{[i]}(f(x))\right]^{[d]} - \beta^{[d]}(x) = \sum_{1\le k\le n} \sum_{|m|=d} \gamma_k^m \mathbf{e}^k x^m \tag{3.9}$$

then

$$\theta_k^m \quad = \quad \frac{\gamma_k^m}{\mu_k - \lambda^m}. \tag{3.10}$$

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

4 Convergence of the Formal Solution

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

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

which satisfies (1.12) for all d. We call this a formal solution to (1.11). To prove Theorem 2.2 we must show that this series converges and that it satisfies (1.11).

The equation (3.6) defines $\theta^{[d]}(x)$ in terms of f, $\beta^{[d]}(x)$ and $\theta^{[1]}(x), \dots, \theta^{[d-1]}(x)$. But it is difficult to use it to estimate the size of $\theta^{[d]}(x)$ because of the terms in the sum on the right hand-side. Therefore, we sum the series in a different fashion.

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 is diagonal. We choose B to set the spectrum of A. The rest of the output injection $\bar{\beta}(x)$ is an arbitrary real analytic function. We assume that T can be chosen so that A is diagonal. We construct a sequence of functions $\{\theta_k(x)\}_{1=2}^{\infty}$ which satisfy the homological equations

$$\begin{array}{l} \theta_1(x) = Tx \\ A\theta_2(x) - \theta_2(Fx) = T\bar{f}(x) - \bar{\beta}(x) \\ A\theta_k(x) - \theta_k(Fx) = \theta_{k-1}(f(x)) - \theta_{k-1}(Fx). \end{array} \tag{4.1}$$

for $k \geq 3$. 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 have shown that the sum

$$\theta_1(x) + \theta_2(x) + \theta_3(x) + \dots \tag{4.2}$$

converges to an analytic function which solves (1.11) in some polydisc around the origin. Due to space limitation, we omit the proof of the convergence of (4.2). The detail can be found in the full version of the paper [18]. As a final, yet important remark, and before closing this section we shall prove that almost all $\mu \in \mathbb{C}$ are of multiplicative type (C, ν) with respect to the spectrum of F.

Theorem 4.1. *For fixed $\nu > \frac{n}{2}$ and any $C > 0$, define $M(C)$ as the set of all μ which are not of multiplicative type (C, ν) with respect to the spectrum of F for degree $d \geq 1$. Then*

$$\textit{measure } M(C) \leq k(n, \nu) C^2,$$

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

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

Proof. Clearly,

$$M(C) = \bigcup_{|m| \geq 1} \mathrm{Ball}\big(\lambda^m, \frac{C}{|m|^\nu}\big)$$

where $\mathrm{Ball}(p, r)$ stands for an open ball in $\mathbb{C}$ centered at $p \in \mathbb{C}$ with radius r. The measure of the $\mathrm{Ball}\big(\lambda^m, \frac{C}{|m|^\nu}\big)$ is $\dfrac{\pi C^2}{|m|^{2\nu}}$. There are no more than $(d+1)^{n-1}$ choices of $m = (m_1, m_2, \dots, m_n)$ such that $|m| = d$. To see this note that each of $m_1, \dots, m_{n-1}$ must lie between 0 and d and then $m_n = d - m_1 - \dots - m_{n-1}$. Since $(d+1) \leq 2d$, we have

$$\mathrm{meas} \bigcup_{|m|=d} \mathrm{Ball}\big(\lambda^m, \frac{C}{|m|^\nu}\big) \leq \pi C^2 (2d)^{n-1-2\nu}.$$

Therefore, if $n - 1 - 2\nu < -1$ then

$$\mathrm{meas} \bigcup_{|m|>0} \mathrm{Ball}\big(\lambda^m, \frac{C}{|m|^\nu}\big) \leq \pi C^2 \Big(\sum_{d=1}^{\infty} (2d)^{n-1-2\nu}\Big),$$

and the result follows. □

5 Necessity

In this section we prove the converse statement of the error dynamics linarization problem of interest.

Proof of Theorem 2.3 Suppose that the change of coordinates $z = \theta(x)$ transforms the original system (1.2) to

$$\begin{aligned} z(k+1) &= g(z(k)) \\ y(k) &= h(z(k)) = h(\theta^{-1}(z(k))) \end{aligned} \tag{5.3}$$

where $0 = \theta(0)$, $g(0) = 0$ and $h(0) = 0$. Suppose there exists a nonlinear observer

$$\hat{z}(k+1) = \hat{g}(\hat{z}(k), y(k)) \tag{5.4}$$

such that the error dynamics is linear (1.6).

Assuming $z = 0$, (1.6) yields

$$A\hat{z} = \hat{g}(\hat{z}, 0). \tag{5.5}$$

Assume now that $\tilde{z} = 0$. Then (1.6) implies

$$g(z) = \hat{g}(z, h(z)). \tag{5.6}$$

Define

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

then

$$\begin{aligned} A\tilde{z} &= g(z) - \hat{g}(\hat{z}, h(z)) = \hat{g}(z, h(z)) - \hat{g}(\hat{z}, h(z)) \\ &= \hat{g}(z, 0) - \beta(z, h(z)) - \hat{g}(\hat{z}, 0) + \beta(\hat{z}, h(z)) \\ &= Az - \beta(z, h(z)) - A\hat{z} + \beta(\hat{z}, h(z)). \end{aligned} \tag{5.7}$$

Therefore

$$\beta(z, h(z)) = \beta(\hat{z}, h(z)). \tag{5.8}$$

Notice, that the left hand-side of this equation does not depend on $\hat{z}$, hence $\beta(z, y) = \beta(y)$. Furthermore

$$\begin{aligned} \hat{g}(\hat{z}, y) &= A\hat{z} - \beta(y) \\ g(z) &= Az - \beta(h(z)). \end{aligned} \tag{5.9}$$

so $\theta(x)$ must satisfy functional equation (1.11). Clearly $\beta(y)$ is C^{d+1}.

6 Illustrative Examples

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

$$\begin{aligned} x_1(k+1) &= \exp(1.3\frac{x_2(k)}{1+x_2(k)})\sqrt{1+x_1(k)+x_2(k)} - 1 - 0.4x_2(k) - 0.5\ln(1+x_1(k)+x_2(k)) \\ x_2(k+1) &= 0.5\ln(1+x_1(k)+x_2(k)) + 0.4x_2(k) \\ y(k) &= x_2(k) \end{aligned} \tag{6.10}$$

with $(x_1, x_2) = (0, 0)$ being an equilibrium point of (6.10) and $1 + x_1 + x_2 > 0$ for all $x_1, x_2 \in R$. The Jacobian matrix F of (6.10) evaluated at the above equilibrium point is:

$$F = \begin{pmatrix} 0 & 0.9 \\ 0.5 & 0.9 \end{pmatrix}. \tag{6.11}$$

The eigenvalues of F are: $\lambda_1 = -0.358$ and $\lambda_2 = 1.258$, and therefore the above system is unstable ($\lambda_2 > 1$). Furthermore, notice that the nonlinear discrete-time observer design method developed by Kazantzis and Kravaris [7] can not be applied to this case since the eigenvalues of F do not belong to the Poincaré domain. Notice also that the system's linear part is observable.

Consider now the following matrix A:

$$A = \begin{pmatrix} 0.5 & 0.3 \\ 0.5 & 0.4 \end{pmatrix} \tag{6.12}$$

with eigenvalues: $\mu_1 = 0.8405$ and $\mu_2 = 0.0595$. Under the above choice for A, its eigenvalues are of multiplicative type (C, ν) with respect to the spectrum $\sigma(F)$ of F. Moreover, the following output injection map was chosen:

$$\beta(y) = \begin{pmatrix} 1.3(\frac{y}{1+y}) - 0.3y \\ 0 \end{pmatrix} \tag{6.13}$$

such that the pair (A, B) is controllable.

Under the above remarks and suitable selection of the pair (A, B), the functional equation (1.11) is uniquely solvable in the class of real analytic functions (Theorem (2.2)) and assumes the following form:

$$\begin{aligned} \theta_1\Big(\exp{(1.3\frac{x_2}{1+x_2})}\sqrt{1+x_1+x_2} - 1 - 0.4x_2 &- 0.5\ln{(1+x_1+x_2)}, 0.5\ln{(1+x_1+x_2)} + 0.4x_2\Big) \\ &= 0.5\theta_1 + 0.3\theta_2 + \beta_1(x_2) \\ \theta_2\Big(\exp{(1.3\frac{x_2}{1+x_2})}\sqrt{1+x_1+x_2} - 1 - 0.4x_2 &- 0.5\ln{(1+x_1+x_2)}, 0.5\ln{(1+x_1+x_2)} + 0.4x_2\Big) \\ &= 0.5\theta_1 + 0.4\theta_2 \\ \theta_1(0,0) &= 0 \\ \theta_2(0,0) &= 0 \end{aligned} \tag{6.14}$$

where the last two equations (initial conditions) merely reflect the fact that equilibrium properties should be preserved under the nonlinear coordinate transformation. Indeed, the above functional equation (6.14) admits a unique analytic solution in closed form:

$$\begin{aligned} \theta_1(x_1, x_2) &= \ln{(1+x_1+x_2)} \\ \theta_2(x_1, x_2) &= x_2 \end{aligned} \tag{6.15}$$

Notice that $\theta_1(0, 0) = 0$ and $\theta_2(0, 0) = 0$, and therefore the initial conditions are satisfied. Moreover, since:

$$J = \frac{\partial \theta}{\partial x}(0, 0) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{6.16}$$

and $\det[J] \neq 0$, the above solution $\theta(x)$ is indeed locally invertible around the equilibrium point $(x_1, x_2) = (0, 0)$ (diffeomorphism). The proposed discrete-time observer in z-coordinate is given by:

$$\begin{array}{rcl} \hat{z}_1(k+1) & = & 0.5\hat{z}_1(k) + 0.3\hat{z}_2(k) + 1.3(\dfrac{y(k)}{1+y(k)}) - 0.3y(k) \\ \hat{z}_2(k+1) & = & 0.5\hat{z}_1(k) + 0.4\hat{z}_2(k) \\ \hat{x}_1(k) & = & \exp(\hat{z}_1(k)) - \hat{z}_2(k) - 1 \\ \hat{x}_2(k) & = & \hat{z}_2(k). \end{array} \qquad (6.17)$$

Let $e_1(k) := x_1(k) - \hat{x}_1(k)$ and $e_2(k) := x_2(k) - \hat{x}_2(k)$. The error dynamics is shown in Fig.1, where $x_1(1) = 0.5, x_2(1) = 0.8$ and $\hat{x}_1(1) = \hat{x}_2(1) = 0$.

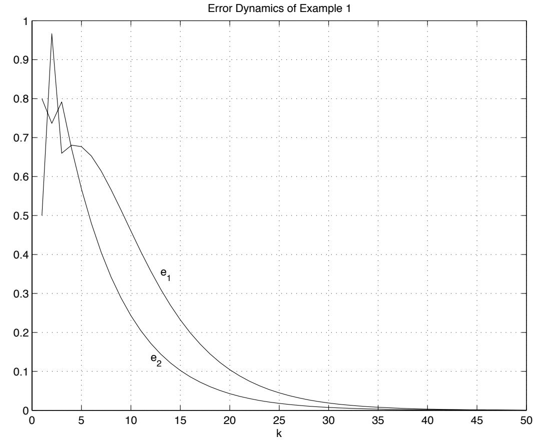

Figure 1: Error Dynamics for $k = 50$.

Example 2: Consider the following nonlinear discrete-time dynamic system:

$$\begin{array}{rcl} y(k+1) & = & \dfrac{0.5(\dfrac{y(k)}{1+y(k)}) - 0.9w(k)}{1 - 0.5(\dfrac{y(k)}{1+y(k)}) + 0.9w(k)} \\ w(k+1) & = & w(k) \end{array} \qquad (6.18)$$

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 (6.18) evaluated at the origin is:

$$F = \begin{pmatrix} 0.5 & -0.9 \\ 0 & 1 \end{pmatrix}. \qquad (6.19)$$

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.

Consider now the following matrix A:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 0.1 \end{pmatrix} \qquad (6.20)$$

with eigenvalues: $\mu_1 = 0$ and $\mu_2 = 0.1$. Furthermore, the following nonlinear output injection map was chosen:

$$\beta(y) = \left(\begin{array}{c} 0.5(\frac{y(k)}{1+y(k)}) \\ \frac{y(k)}{1+y(k)} \end{array} \right). \tag{6.21}$$

Notice that the pair $(A, B = \frac{\partial \beta}{\partial y}(0))$ is controllable.

Under the above remarks the functional equation (1.11) is uniquely solvable in the class of real analytic functions (Theorem (2.2)) and assumes the following form:

$$\begin{aligned} \theta_1\Big(\frac{0.5(\frac{y}{1+y}) - 0.9w}{1 - 0.5(\frac{y}{1+y}) + 0.9w}, w\Big) &= 0.5(\frac{y}{1+y}) \\ \theta_2\Big(\frac{0.5(\frac{y}{1+y}) - 0.9w}{1 - 0.5(\frac{y}{1+y}) + 0.9w}, w\Big) &= 0.1\theta_2 + \frac{y}{1+y} \\ \theta_1(0,0) &= 0 \\ \theta_2(0,0) &= 0 \end{aligned} \tag{6.22}$$

The above functional equation (6.22) admits a unique analytic solution:

$$\begin{aligned} \theta_1(y,w) &= \frac{y}{1+y} + 0.9w \\ \theta_2(y,w) &= \frac{5}{2}(\frac{y}{1+y} + w) \end{aligned} \tag{6.23}$$

Notice that $\theta_1(0,0) = 0$ and $\theta_2(0,0) = 0$, and therefore the initial conditions are satisfied. Moreover, since:

$$J = \frac{\partial \theta}{\partial x}(0,0) = \left(\begin{array}{cc} 1 & 0.9 \\ 2.5 & 2.5 \end{array} \right) \tag{6.24}$$

and $\det[J] \neq 0$, the above solution $\theta(x)$ is indeed locally invertible around the origin. The proposed discrete-time observer is then given by the following dynamic equations:

$$\begin{aligned} \hat{z}_1(k+1) &= 0.5(\frac{y(k)}{1+y(k)}) \\ \hat{z}_2(k+1) &= 0.1\hat{z}_2(k) + \frac{y(k)}{1+y(k)} \\ \hat{y}(k) &= \frac{10\hat{z}_1(k) - 3.6\hat{z}_2(k)}{1 - 10\hat{z}_1(k) + 3.6\hat{z}_2(k)} \\ \hat{w}(k) &= 4\hat{z}_2(k) - 10\hat{z}_1(k). \end{aligned} \tag{6.25}$$

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.2, where $y(1) = 10, w(1) = -2\pi$ and $\hat{z}_1(1) = \hat{z}_2(1) = 1$.

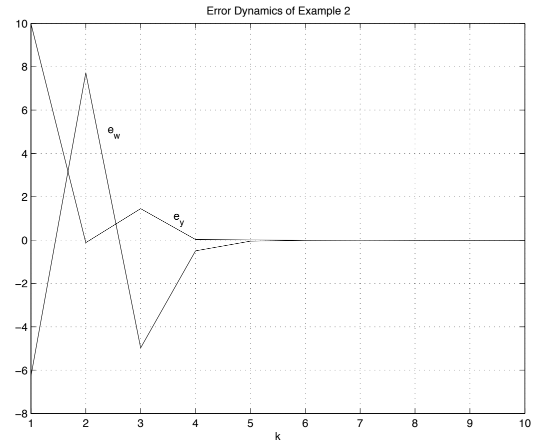

Figure 2: Error Dynamics for $k = 10$.

7 Final Remarks

We have given a necessary and sufficient condition for a smooth nonlinear discrete-time system to admit an observer with linear error dynamics in smoothly transformed coordinates. The condition is that the functional equation (1.11) admits a smooth solution for some A and a smooth output injection term $\beta(y)$.

We have also shown that, for real analytic systems with observable linear part, the functional equation (1.11) is locally solvable for any real analytic $\beta(y)$ that makes the eigenvalues of A of multiplicative type (C, ν) with respect to the spectrum of the linear part F of the dynamics. This allows one to construct a local observer with linear error dynamics in the transformed coordinates.

If the system is only smooth, then the functional equation (1.11) is made formally solvable up to the desirable degree of smoothness by choosing the eigenvalues of A to be nonresonant with the spectrum of the linear part of the dynamics. This allows one to construct a local observer with approximately linear error dynamics in the transformed coordinates.

References

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

[2] D. Bestle and M. Zeitz. Canonical form observer design for nonlinear time-variable systems. *Int. J. Control*, Vol. 38, No. 2, 1983, pp. 419-431.

[3] S.T. Chung and J.W. Grizzle. Sampled-data observer error linearization. *Autom.*, Vol. 26, 1990, pp. 997.

[4] P. Glendinning. Stability, Instability and Chaos: An Introduction to the Theory of Nonlinear Differential Equations. *Cambridge University Press*, 1994.

[5] H. J. C. Huijberts, T. Lilge and H. Nijmeijer. Nonlinear discrete-time synchronization via extended observers. *Int. J. Bifurcations and Chaos*, 7, 2001, pp. 1997-2006.

[6] N. Kazantzis and C. Kravaris. Nonlinear observer design using Lyapunov's auxiliary theorem. *Systems & Control Letters*, 34, 1998, pp. 241-247.

[7] N. Kazantzis and C. Kravaris. Discrete-time nonlinear observer design using functional equations. *Systems & Control Letters*, 42, 2001, pp. 81-94.

[8] A. J. Krener and A. Isidori. Linearization by output injection and nonlinear observers. *Systems & Control Letters*, 3, 1983, pp. 47-52.

[9] A. J. Krener. Approximate Linearization by state feedback and coordinate change. *Systems & Control Letters*, 5, 1984, pp. 181-185.

[10] A. J. Krener and W. Respondek. Nonlinear observers with linearizable error dynamics. *SIAM J. Control and Optimization*, Vol. 23, No. 2, 1985, pp. 197-216.

[11] A. J. Krener. Nonlinear stabilizability and detectability. *In Systems and Networks: Mathematical Theory and Applications*, U. Helmke, R. Mennicken and J. Saurer (eds.), Akademie Verlag, Berlin, pp. 231-250, 1994.

[12] A. J. Krener and M. Xiao. Nonlinear observer design in the Siegel domain. *SIAM J. Control and Optimization*, to appear, 2002.

[13] A. J. Krener and M. Xiao. Nonlinear observer design in the Siegel domain through coordinate transformation. *Proc. of the 5th International Federation of Automatic Control*, Saint-Petersburg, Russia, 2001, pp. 557-562.

[14] A. J. Krener and M. Xiao. Necessary and Sufficient Condition for Nonlinear Observer with linearizable Error Dynamics. *Proc. of the 40th IEEE Conference on Decision and Control*, Orlando, Florida, 2001.

[15] A. J. Krener and M. Xiao. Observers for Linearly Unobservable Nonlinear Systems. *Systems & Control Letters*, to appear, 2002.

[16] H. G. Lee and K. Nam. Observer design for autonomous discrete-time nonlinear systems. *Systems & Control Letters*, 17, 1991, pp. 49-58.

[17] W. Lin and C. I. Byrnes. Remarks on linearization of discrete-time autonomous systems and nonlinear observer design. *Systems & Control Letters*, 25, 1998, pp. 31-41.

[18] M. Xiao, N. Kazantzis, C. Kravaris, and A. J. Krener. Nonlinear Discrete-Time Observer Design with Linearizable Error Dynamics. *Preprint*, September, 2001.