SIAM J. CONTROL OPTIM.
Vol. 41, No. 3, pp. 932–953

NONLINEAR OBSERVER DESIGN IN THE SIEGEL DOMAIN*

ARTHUR J. KRENER† AND MINGQING XIAO‡

Abstract. We extend the method of Kazantzis and Kravaris [*Systems Control Lett.*, 34 (1998), pp. 241–247] for the design of an observer to a larger class of nonlinear systems. The extended method is applicable to any real analytic observable nonlinear system. It is based on the solution of a first-order, singular, nonlinear PDE. This solution yields a change of state coordinates which linearizes the error dynamics. Under very general conditions, the existence and uniqueness of the solution is proved. Lyapunov's auxiliary theorem and Siegel's theorem are obtained as corollaries. The technique is constructive and yields a method for constructing approximate solutions.

Key words. nonlinear systems, nonlinear observers, linearizable error dynamics, output injection, Siegel domain, Lyapunov's auxiliary theorem, Siegel's theorem

AMS subject classifications. 93, 35, 32

PII. S0363012900375330

1. Introduction.

We consider the problem of estimating the current state $x(t)$ of a nonlinear dynamical system, described by a system of first-order differential equations,

$$\begin{aligned} \dot{x} &= f(x), \\ y &= h(x), \end{aligned} \tag{1.1}$$

from the past observations $y(s), s \le t$. The vector fields $f : \mathbf{R}^n \to \mathbf{R}^n$ and $h : \mathbf{R}^n \to \mathbf{R}^p$ are assumed to be real analytic functions with $f(0) = 0,\ h(0) = 0$. One technique of constructing an observer is to find a nonlinear change of state and output coordinates which transforms the system (1.1) into a system with linear output map and linear dynamics driven by nonlinear output injection. The design of an observer for such systems is relatively easy [8], [6], [2], and the error dynamics is linear in the transformed coordinates. Recently Kazantzis and Kravaris proposed a simpler method [5]. One seeks a change of state coordinates $z = \theta(x)$ such that the dynamics of (1.1) is linear driven by nonlinear output injection

$$\dot{z} = Az - \beta(y), \tag{1.2}$$

where A is an $n \times n$ matrix and $\beta : \mathbf{R}^p \to \mathbf{R}^n$ is a real analytic vector field. One does not have to linearize the output map.

Such a θ must satisfy the following first-order PDE:

$$\frac{\partial \theta}{\partial x}(x) f(x) = A\theta(x) - \beta(h(x)). \tag{1.3}$$

Using a particular form of the Lyapunov auxiliary theorem [10], Kazantzis and Kravaris showed that (1.3) has a unique solution under certain assumptions.

*Received by the editors July 14, 2000; accepted for publication (in revised form) February 22, 2002; published electronically September 19, 2002. A preliminary version of this paper appeared in the Proceedings of the 2001 IEEE Conference on Decision and Control.
http://www.siam.org/journals/sicon/41-3/37533.html
†Department of Mathematics, University of California, Davis, CA 95616-8633 (ajkrener@ucdavis.edu). Research for this author was supported in part by NSF 9970998.
‡Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (mxiao@math.siu.edu).

THEOREM [10]. *Assume that* $f : \mathbf{R}^n \to \mathbf{R}^n, h : \mathbf{R}^n \to \mathbf{R}^p$, *and* $\beta : \mathbf{R}^p \to \mathbf{R}^n$ *are 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 the eigenvalues of* F *be* $(\lambda_1, \ldots, \lambda_n)$ *and the eigenvalues of* A *be* $(\mu_1, \ldots, \mu_n)$. *If*

1. 0 *does not lie in the convex hull of* $(\lambda_1, \ldots, \lambda_n)$,
2. *there do not exist nonnegative integers* $m_1, m_2, \ldots, m_n$ *not all zero such that* $\sum_{i=1}^n m_i \lambda_i = \mu_j$,

then the first-order PDE (1.3), *with initial condition* $\theta(0) = 0$, *admits a unique analytic solution* θ *in a neighborhood of* $x = 0$.

Based on the above theorem, Kazantzis and Kravaris proposed a nonlinear observer design method [5], where the state observer is constructed using the coordinate transformation $z = \theta(x)$ and the output injection $\beta(y)$.

KAZANTZIS AND KRAVARIS THEOREM [5]. *Assume that* f, h, θ, β *are as in the above theorem and additionally that*

3. θ *is a local diffeomorphism,*
4. A *is Hurwitz.*

Then the local state observer for (1.1) *given by*

$$\dot{\hat{x}} = f(\hat{x}) - \left[\frac{\partial \theta}{\partial \hat{x}}(\hat{x})\right]^{-1} (\beta(y) - \beta(h(\hat{x}))) \tag{1.4}$$

has locally asymptotically stable error dynamics. In z *coordinates, the system is given by* (1.2), *the observer is*

$$\dot{\hat{z}} = A\hat{z} - \beta(y), \tag{1.5}$$

and the error $\tilde{z} = z - \hat{z}$ *dynamics is*

$$\dot{\tilde{z}} = A\tilde{z}. \tag{1.6}$$

One can show that if the conditions of this theorem hold, then (H, F) is an observable pair and (A, B) is a controllable pair. On the other hand, if (H, F) is an observable pair, then one can choose an invertible T and B so that $A = (TF + BH)T^{-1}$ satisfies 2, 3, and if the solution of (1.3) exists for some β such that $\beta(0) = 0$, $\frac{\partial \beta}{\partial x}(0) = B$, then θ is a local diffeomorphism. The size of the neighborhood of 0 on which θ is a diffeomorphism varies with the higher derivatives of β, hence the advantage of allowing them to be different from zero.

The approach of Kazantzis and Kravaris has an advantage over that of Krener and Respondek [8] and similar attempts to transform the dynamics and output map into observer form. The former uses the Lyapunov auxiliary theorem, which depends on a nonresonance condition, assumption 2 above, while the latter depends on integrability conditions. The nonresonance condition is generically satisfied while the integrability conditions are generically not satisfied. However, assumption 1 of Kazantzis and Kravaris is quite restrictive, as it requires the system to be locally asymptotically stable to the origin in either forward or reverse time. If the system is stable in forward time, then an observer is not needed, as we know where it is going. If the system is stable in reverse time, then it is unstable in forward time, so what good is a local observer?

Assumption 1 requires that the eigenvalues of the linear part of $f(x)$ at the origin lie in the *Poincaré domain*, whose definition follows.

DEFINITION 1. *An n-tuple $\lambda = (\lambda_1, \ldots, \lambda_n)$ of complex numbers belongs to the Poincaré domain if the convex hull of $(\lambda_1, \ldots, \lambda_n)$ does not contain zero. An n-tuple of complex numbers belongs to the Siegel domain if zero lies in the convex hull of $(\lambda_1, \ldots, \lambda_n)$.*

Clearly, requiring the spectrum of F to be in the Poincaré domain rules out many interesting problems, including critical ones where there are eigenvalues on the imaginary axis [9]. In this paper we extend the observer design method of Kazantzis and Kravaris to the Siegel domain [1]. We start with a definition.

DEFINITION 2. *Given an $n \times n$ matrix F with spectrum $\sigma(F) = \lambda = (\lambda_1, \ldots, \lambda_n)$ and constants $C > 0$, $\nu > 0$, we say a complex number μ is of type (C, ν) with respect to $\sigma(F)$ if for any vector $m = (m_1, m_2, \ldots, m_n)$ of nonnegative integers, $|m| = \sum m_i > 0$, we have*

$$|\mu - m \cdot \lambda| \geq \frac{C}{|m|^\nu}. \tag{1.7}$$

Now we are ready to state the main result of this paper.

MAIN THEOREM. *Assume that $f : \mathbf{R}^n \to \mathbf{R}^n$, $h : \mathbf{R}^n \to \mathbf{R}^p$, and $\beta : \mathbf{R}^p \to \mathbf{R}^n$ are analytic vector fields with $f(0) = 0$, $h(0) = 0$, $\beta(0) = 0$ and $F = \frac{\partial f}{\partial x}(0)$, $H = \frac{\partial h}{\partial x}(0)$, $B = \frac{\partial \beta}{\partial y}(0)$. Suppose there exists*

1. *an invertible $n \times n$ matrix T so that $TF = AT - BH$;*
2. *a $C > 0$, $\nu > 0$ such that all the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$.*

Then there exists a unique analytic solution $z = \theta(x)$ to the PDE (1.3) locally around $x = 0$ with $\frac{\partial \theta}{\partial x}(0) = T$, so θ is a local diffeomorphism.

Notes. We have stated this theorem for real analytic functions because we are applying it to a real analytic system. However, it is true for complex analytic functions, as can be seen from the proof. Assumption 2 implies that the eigenvalues of A are distinct from those of F. We shall show the following. Assumptions 1 and 2 imply that (H, F) is an observable pair. On the other hand, if (H, F) is an observable pair, then one can let $T = I$ and set the spectrum of A arbitrarily by choice of B. Almost all complex numbers are of type (C, ν) with respect to $\sigma(F)$, so assumption 2 is hardly a restriction on A when (H, F) is an observable pair. If A is chosen to be Hurwitz, then the state estimator is given by (1.4) and the error dynamics is locally asymptotically stable as before. We defer the proof of the main theorem to the next section.

CONVERSE TO THE MAIN THEOREM. *Consider the class of nonlinear systems described by the following equation:*

$$\begin{aligned} \dot{z} &= g(z), \\ y &= h(z), \end{aligned} \tag{1.8}$$

where $z \in \mathbf{R}^n$, $y \in \mathbf{R}^p$, and g, h are continuous vector fields on $\mathbf{R}^n$, $\mathbf{R}^p$, respectively, with $g(0) = 0$ and $h(0) = 0$. If there exists a nonlinear observer

$$\dot{\hat{z}} = \hat{g}(\hat{z}, y) \tag{1.9}$$

such that the error $\tilde{z} = z - \hat{z}$ dynamics is linear,

$$\dot{\tilde{z}} = A\tilde{z}, \tag{1.10}$$

where A is an $n \times n$ matrix, then there exists a continuous vector field $\beta : \mathbf{R}^p \to \mathbf{R}^n$ such that

$$g(z) = Az - \beta(h(z)), \tag{1.11}$$

$$\hat{g}(\hat{z}, y) = A\hat{z} - \beta(y). \tag{1.12}$$

Proof. The error dynamics is

$$\dot{\tilde{z}} = A\tilde{z} = g(z) - \hat{g}(\hat{z}, y).$$

Assume $z = 0$. Then

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

Assume $\tilde{z} = 0$. Then

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

Define

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

Then

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

So

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

But the left side does not depend on $\hat{z}$, so neither does the right, and thus

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

Therefore

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

□

Note. This converse shows that if a system (1.1) admits an observer with linear error dynamics after a smooth change of coordinates, it is because the PDE (1.3) is solvable for some smooth θ and continuous β.

The rest of the paper is organized as follows. Section 2.1 discusses the relationship between the linear part of the nonlinear system (1.1) and the terms of degree 1 of the solution (1.3). A unique formal solution of (1.3) is given in section 2.2 and this is shown to be convergent in section 2.3. We also show in section 2.1 that (1.3) has a unique solution for any choice of the eigenvalues of A except for a set of zero measure in $\mathbf{C}^n$. Several examples are treated in section 3. Section 4 applies the main result to the case when the system has inputs.

2. Solution of the PDE.

2.1. Terms of degree 1. If we focus on the terms of degree 1 in (1.3), we obtain the equation

$$TF = AT - BH. \tag{2.1}$$

We view this as a linear equation for T in terms of given F, H, A, B.

LEMMA 1. *Equation* (2.1) *admits a unique solution* T *if and only if the eigenvalues of* F *and* A *are distinct, that is,* $\sigma(F) \cap \sigma(A) = \emptyset$.

Proof. We give the proof when F admits a basis of right eigenvectors, $\{\mathbf{v}^j, j = 1, \ldots, n\}$, and A admits a basis of left eigenvectors, $\{\mathbf{w}_i, i = 1, \ldots, n\}$. The general case is similar using bases of generalized eigenvectors. Define an operator $\mathcal{F} : T \longmapsto TF - AT$ on the space of $n \times n$ matrices $\{T\}$. Let λ_i be the eigenvalue of F corresponding to the right eigenvector $\mathbf{w}_i$, and let μ_j be the eigenvalue of F corresponding to left eigenvector $\mathbf{v}^j$. Now $\{\mathbf{v}^j\mathbf{w}_i, i, j = 1, \ldots, n\}$ is a basis for $\{T\}$ and

$$\begin{aligned}\mathcal{F}(\mathbf{v}^j\mathbf{w}_i) &= (\mathbf{v}^j\mathbf{w}_i)F - A(\mathbf{v}^j\mathbf{w}_i)\\ &= (\lambda_i - \mu_j)\mathbf{v}^j\mathbf{w}_i.\end{aligned}$$

Thus $\mathcal{F}$ is invertible if and only if $\lambda_i - \mu_j \neq 0$ for all possible i and j. Therefore $\mathcal{F}T = -BH$ admits a unique solution if and only if $\sigma(F) \cap \sigma(A) = \emptyset$. □

LEMMA 2. *Suppose* $\sigma(F) \cap \sigma(A) = \emptyset$. *If* T *is invertible, then* (H, F) *is observable and* (A, B) *is controllable.*

Proof. Suppose (H, F) is not observable. Then there exist $\lambda_i \in \sigma(F)$ and a vector $x \in \mathbf{R}^{n\times 1}$, $x \neq 0$, such that $Hx = 0$ and $Fx = \lambda_i x$. Multiply (2.1) by x to obtain

$$\lambda_i Tx = TFx = ATx + BHx = ATx.$$

Since $Tx \neq 0$, this implies that $\lambda_i \in \sigma(A)$, a contradiction.

Similarly, suppose (A, B) is not controllable. Then there is $\mu_j \in \sigma(A)$ and a vector $\xi \in \mathbf{R}^{1\times n}$ such that $\xi A = \mu_j \xi$ and $\xi B = 0$. Multiply (2.1) by ξ to obtain

$$\xi TF = \xi AT + \xi BH = \mu_j \xi T.$$

Since $\xi T \neq 0$, this implies that $\mu_j \in \sigma(F)$, a contradiction. □

LEMMA 3. *If* T *is an invertible solution to* (2.1)*, then* A *is conjugate to* F *modified by output injection.*

Proof. Since T satisfies equation

$$TF + BH = AT$$

and T is invertible, we thus have

$$T(F + T^{-1}BH)T^{-1} = A. \qquad \square$$

LEMMA 4. *If* $\sigma(F) \cap \sigma(A) = \emptyset$ *and* A *is conjugate to* F *modified by output injection, then there exists* B *such that the unique solution to* (2.1) *is invertible.*

Proof. Since A is conjugate to F modified by output injection, there exist an $n \times n$ invertible matrix S and an $n \times p$ matrix G such that

$$S(F + GH)S^{-1} = A.$$

Hence we have $SF = AS - SGH$. Let $B = SG$. Then $SF = AS - BH$, so $T = S$ according to Lemma 1. Therefore T is invertible. □

Loosely speaking, a complex number μ is of type (C, ν) with respect to $\sigma(F) = \lambda$ if $|\mu - m \cdot \lambda|$ is never zero and does not approach zero too fast as $|m| \to \infty$. If ν is large enough, then the set of μ's which are of type (C, ν) for some $C > 0$ is dense in the complex plane.

LEMMA 5. *If $C > 0$ and $\nu > \frac{n}{2}$, then*

$$\text{meas } \{\mu : \mu \text{ is not of type } (C, \nu)\} \leq k(n, \nu)C^2, \tag{2.2}$$

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

If $\nu > \frac{n}{2}$, then the set of points which are not of type (C, ν) for any $C > 0$ is a set of zero measure.

Proof. Clearly, the set $\{\mu : \mu \text{ is not of type } (C, \nu)\}$ is

$$\bigcup_{|m| \geq 1} \text{Ball}\left(m \cdot \lambda, \frac{C}{|m|^\nu}\right),$$

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

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

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

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

so (2.2) follows. □

2.2. The formal solution of the PDE. Assume the hypothesis of the main theorem holds. We show that there is a unique solution to the PDE (1.3) within the class of formal power series. It is convenient to assume that F and A are diagonal; the proof in the general case is similar but much messier. We expand the terms in power series

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

where $f^{[d]}$, $\beta^{[d]}$, and $\theta^{[d]}$ are homogeneous polynomial vector fields of degree d in x. The knowns are f, h, β, T and the unknowns are the higher degree terms $\theta^{[2]}, \theta^{[3]}, \ldots$. The linear terms satisfy (2.1) by the above assumption.

The degree d part of (1.3) is

$$\frac{\partial \theta^{[d]}}{\partial x}(x)Fx - A\theta^{[d]}(x) = -\tilde{\beta}^{[d]}(x), \tag{2.3}$$

where

$$\tilde{\beta}^{[d]}(x) = \beta^{[d]}(x) + Tf^{[d]}(x) + \sum_{j=2}^{d-1} \frac{\partial \theta^{[j]}}{\partial x}(x) f^{[d+1-j]}(x). \tag{2.4}$$

Let $\mathbf{e}^k$ denote the kth unit vector in z space and let $x^m = x_1^{m_1} \cdots x_n^{m_n}$. Then the above terms can be expanded as

$$\tilde{\beta}^{[d]}(x) = \sum_{k=1}^{n} \sum_{|m|=d} \tilde{\beta}_{k,m} \mathbf{e}^k x^m,$$

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

and we obtain the equations

$$(\mu_k - m \cdot \lambda)\theta_{k,m} = \tilde{\beta}_{k,m}. \tag{2.5}$$

These equations have unique solutions because $m \cdot \lambda - \mu_k \neq 0$. □

The formal approach yields a method for constructing an observer with approximately linear error dynamics. Start by choosing a T, A, B satisfying the linear equation (2.1). Then successively solve (2.3) up to some degree d. At each step $\beta^{[j]}$ can be chosen to make $\theta^{[j]}$ smaller and thereby try to keep $\theta(x)$ close to its globally invertible linear part Tx. The approximate solution

$$\theta(x) = Tx + \theta^{[2]}(x) + \theta^{[3]}(x) + \cdots + \theta^{[d]}(x),$$
$$\beta(y) = By + \beta^{[2]}(y) + \beta^{[3]}(y) + \cdots + \beta^{[d]}(y)$$

transforms the system (1.1) into

$$\dot{z} = Az - \beta(y) + O(x)^{d+1},$$

so the observer (1.4) has approximately linearizable error dynamics. The error is $O(x, \hat{x})^{d+1}$. When implementing the method, the matrices F, A need not be diagonal, but this makes solving (2.3) very easy.

2.3. Convergence of the formal solution. Let $|x| = \max\{|x_1|, \ldots, |x_n|\}$. We write

$$f(x) = Fx + \bar{f}(x),$$
$$\beta(y) = BHx + \bar{\beta}(x),$$

where $AT - TF = BH$. We first show that the sequence of PDEs

$$A\theta_2(x) - \frac{\partial}{\partial x}\theta_2(x)Fx = T\bar{f}(x) + \bar{\beta}(x),$$
$$A\theta_k(x) - \frac{\partial}{\partial x}\theta_k(x)Fx = \frac{\partial}{\partial x}\theta_{k-1}(x)\bar{f}(x)$$

admits a sequence of analytical solutions $\theta_2(x), \theta_3(x), \ldots$ in some neighborhood of the origin. Then we show that the sum

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

converges to an analytic function which solves (1.3).

We define a positive real function $b_k : [0,1) \to [0,\infty)$ to be

$$b_k(q) := \max_{d \in \mathbf{Z}_+, d \geq k} \left[C^{-1} d^{\nu} q^{\frac{d}{2}} \right],$$

where $C > 0$ and $\nu > 0$ are given. We start with an important theorem.

THEOREM 1. *Let $P(x)$ be a real analytic function in $|x| < r$ with $P(0) = 0$ and $\frac{\partial P}{\partial x}(0) = 0$. Suppose all of the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$. Then the first-order PDE*

$$A\theta(x) - \frac{\partial \theta(x)}{\partial x} Fx = P(x) \tag{2.6}$$

admits a unique analytic solution $\theta(x)$ in $|x| < r$ with $\theta(0) = 0$.

Proof. The analyticity of $P(x)$ implies that $P(x)$ can be expanded into a Taylor series

$$P(x) = P^{[k]}(x) + P^{[k+1]}(x) + \cdots \qquad \text{for} \quad |x| < r$$

with

$$P^{[d]}(x) = \sum_{j=k}^{n} \sum_{|m|=d} p_{j,m} \mathbf{e}^j x^m,$$

where $k \geq 2$ is the lowest degree of $P(x)$. We assume a series solution

$$\theta(x) = \theta^{[k]}(x) + \theta^{[k+1]}(x) + \cdots + \theta^{[d]}(x) + \cdots \tag{2.7}$$

with

$$\theta^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} \theta_{j,m} \mathbf{e}^j x^m.$$

If we plug (2.7) into (2.6), then we have

$$\theta_{j,m} = \frac{p_{j,m}}{\mu_j - m \cdot \lambda} \qquad \text{for } |m| \geq k, \quad 1 \leq j \leq n.$$

Since the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$, we have

$$|\theta_{j,m}| = \left| \frac{p_{j,m}}{\mu_j - m \cdot \lambda} \right| \leq \frac{|m|^{\nu} |p_{j,m}|}{C}.$$

We shall show that (2.7) converges on the closed polydisk $|x| \leq qr$ for any $0 < q < 1$. Hence (2.7) converges on $|x| < r$.

Consider a new series

$$\hat{P}(x) = \hat{P}^{[k]}(x) + \hat{P}^{[k+1]}(x) + \cdots \tag{2.8}$$

with

$$\hat{P}^{[d]}(x) = \sum_{j=k}^{n} \sum_{|m|=d} \frac{|m|^{\nu} |p_{j,m}|}{C} \mathbf{e}^j x^m, \qquad d \geq k.$$

We next claim that (2.8) converges in $|x| \le qr$. Let $\xi := (qr, qr, \dots, qr)$. Then

$$\begin{aligned}|\hat{P}^{[d]}(x)| &\le \max_{1\le j\le n} \sum_{|m|=d} \frac{|m|^\nu |p_{j,m}|}{C} |x|^m \le \max_{1\le j\le n} \sum_{|m|=d} \frac{|m|^\nu |p_{j,m}|}{C} |\xi|^m \\ &\le \max_{1\le j\le n} \sum_{|m|=d} |m|^\nu C^{-1} q^{\frac{|m|}{2}} |p_{j,m}| (\sqrt{q}r)^{|m|} \\ &\le b_k(q) \max_{1\le j\le n} \sum_{|m|=d} |p_{j,m}| (\sqrt{q}r)^{|m|}.\end{aligned}$$

Notice that $P(x)$ is an analytic function for $|x| < r$, so its Taylor series converges there absolutely, which yields

$$|\hat{P}(x)| \le b_k(q) \max_{1\le j\le n} \sum_{d=k}^{\infty} \left(\sum_{|m|=d} |p_{j,m}| (\sqrt{q}r)^{|m|} \right) < +\infty.$$

Thus (2.7) defines an analytic function $\theta(x)$ for $|x| < r$, which solves (2.6). ∎

From Theorem 1, we immediately have the following corollary.

COROLLARY 1. *Suppose all of the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$. The PDEs*

$$A\theta_2(x) - \frac{\partial \theta_2}{\partial x}(x) Fx = T\bar{f}(x) + \bar{\beta}(x), \qquad \theta_2(0) = 0, \tag{2.9}$$

$$A\theta_k(x) - \frac{\partial \theta_k}{\partial x}(x) Fx = \frac{\partial \theta_{k-1}}{\partial x}(x) \bar{f}(x), \qquad \theta_k(0) = 0, \quad k \ge 3, \tag{2.10}$$

admit analytic solutions in $|x| < r$.

The next step is to prove that

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

converges near the origin and solves the PDE (1.3).

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:

$$\bar{f}(x) = f^{[2]}(x) + f^{[3]}(x) + \cdots, \qquad |x| \le r,$$

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 $j = 1, 2, \dots, n$. We define

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

and

$$\|P(x)\| := \max_{1\le i\le n} \sum_m |p_{i,m} x^m|$$

if $P(x)$ is analytic in $|x| < r$ with

$$P(x) = \left(\sum_m p_{1,m} x^m, \sum_m p_{2,m} x^m, \ldots, \sum_m p_{n,m} x^m \right).$$

LEMMA 6. *There exists $0 < r_1 < r$ such that if $P(x)$ is analytic in $|x| < r_1$, where $\|P(x)\| \le N$, then*

$$\left\| \frac{\partial P}{\partial x}(x) \bar{f}(x) \right\| \le N \qquad \text{in} \quad |x| < r_1.$$

Proof. First it is easy to see that for any $r_1 < r$ we have

$$|\bar{f}(x)| \le r_1^2 \bar{M}_f \qquad \text{for} \quad |x| \le r_1,$$

since for $j = 1, 2, \ldots, n$

$$\begin{aligned} &\sum_{|m|=2} |f_{j,m}| r_1^2 + \sum_{|m|=3} |f_{j,m}| r_1^3 + \cdots \\ &\quad = r_1^2 \left(\sum_{|m|=2} |f_{j,m}| + \sum_{|m|=3} |f_{j,m}| r_1^1 + \cdots \right) \\ &\quad \le r_1^2 \frac{M_j}{r^2} \le r_1^2 \bar{M}_f. \end{aligned} \tag{2.11}$$

Next let

$$P(x) = (P_1(x), P_2(x), \ldots, P_n(x)),$$

with $P_i(x) = \sum_m p_{i,m} x^m$ and

$$N(r) := \max_{|x| \le r} \|P(x)\|.$$

The analyticity of $P(x)$ implies that

$$\frac{\partial P_i}{\partial x_j}(x) = \sum_m \frac{\partial}{\partial x_j} (p_{i,m} x^m) = \sum_m p_{i,m} m_j x_1^{m_1} \cdots x_j^{m_j - 1} \cdots x_n^{m_n}, \qquad |x| < r_1,$$

and for any given $\varepsilon > 0$ there exists $K > 0$ such that when $|m| > K$

$$\sum_{m, |m| \ge K} \left| p_{i,m} m_j x_1^{m_1} \cdots x_j^{m_j - 1} \cdots x_n^{m_n} \right| < \varepsilon$$

for $|x| < r_1$. Thus

$$\sum_{m, |m| \le K} |p_{i,m} m_j x_1^{m_1} \cdots x_j^{m_j - 1} \cdots x_n^{m_n}| \|\bar{f}_j(x)\| \le \sum_{m, |m| \le K} |p_{i,m}| m_j {r_1}^{|m|} r_1 \bar{M}_f.$$

Let r_1 be small enough such that

$$\sum_{m, |m| \le K} |p_{i,m}| m_j {r_1}^{|m|} r_1 \bar{M}_f \le \frac{N(r_1)}{n}.$$

Then for $|x| < r_1$

$$\sum_m \left| \frac{\partial}{\partial x_j} \left(p_{i,m} x^m \right) \right| \left\| \bar{f}_j(x) \right\| < \frac{N(r_1)}{n} + \varepsilon r_1^2 \bar{M}_f.$$

Thus we have

$$\left\| \frac{\partial P_i}{\partial x_j}(x) \bar{f}_j(x) \right\| \leq \sum_m \left| \left(\frac{\partial}{\partial x_j} \left(p_{i,m} x^m \right) \right) \right| \left\| \bar{f}_j(x) \right\| \leq \frac{N(r_1)}{n}.$$

Therefore

$$\left\| \frac{\partial P}{\partial x}(x) \bar{f}(x) \right\| \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} \left\| \frac{\partial P_i}{\partial x_j}(x) \bar{f}_j(x) \right\| \leq N(r_1). \qquad \square$$

In the definition of type (C, ν), without lose of generality we can assume that ν is a positive integer since if ν is not, we can replace it by a larger integer.

LEMMA 7. *Let $r_2 := r_1/n$, where r_1 is given in Lemma 6. Let $\theta_k(x)$ be the solution of*

$$A\theta_k(x) - \frac{\partial \theta_k}{\partial x}(x) Fx = \frac{\partial \theta_{k-1}}{\partial x}(x) \bar{f}(x).$$

Then if $\|\theta_{k-1}(x)\| \leq N$ for $|x| < r_2$, we have

$$\|\theta_k(x)\| \leq \frac{NP(|x_1| + |x_2| + \cdots + |x_n|)}{C\left(r_1 - (|x_1| + |x_2| + \cdots + |x_n|)\right)^{\nu+1}}$$

for $|x| < r_2$, where P is a polynomial of degree ν with coefficients depending only on r_1.

Proof. We first let $g(x) := \frac{\partial \theta_{k-1}}{\partial x}(x) \bar{f}(x)$ and

$$\phi(x) := \frac{N r_1}{r_1 - (x_1 + \cdots + x_n)}.$$

Clearly for $|x| < r_2$,

$$\begin{aligned} \phi(x) &= \frac{N}{1 - (x_1 + \cdots + x_n)/r_1} = N \sum_{d=0}^{\infty} \left(\frac{x_1 + \cdots + x_n}{r_1} \right)^d \\ &= N \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|!}{m!} x^m \end{aligned}$$

and

$$D^m \phi(0) = N |m|! r_1^{-|m|}.$$

By the previous lemma, $|g(x)| \leq N$ for $|x| < r_1$, so the Cauchy estimate yields

$$|D^m g(0)| \leq N |m|! r_1^{-|m|},$$

where D^m is a partial differential operator of order m defined to be

$$D^m = \frac{\partial^m}{\partial x_1^{m_1} \cdots \partial x_n^{m_n}}.$$

Let

$$g(x) = g^{[k]}(x) + g^{[k+1]}(x) + \cdots + g^{[d]}(x) + \cdots$$

with $g^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} g_{j,m} \mathbf{e}^j x^m$, where

$$|g_{j,m}| = \left| \frac{1}{m!} D^m g(0) \right| \le N \frac{|m|!}{m!} r_1^{-|m|}$$

and

$$\theta_k(x) = \theta_k^{[k]}(x) + \theta_k^{[k+1]}(x) + \cdots + \theta_k^{[d]}(x) + \cdots$$

with $\theta_k^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} \theta_{j,m} \mathbf{e}^j x^m$. Then (2.12) implies that

$$\theta_{j,m} = \frac{g_{j,m}}{\mu_j - \lambda \cdot m}.$$

Since the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$, it follows that

$$|\theta_{j,m}| = \left| \frac{g_{j,m}}{\mu_j - \lambda \cdot m} \right| \le \frac{|m|^\nu}{C} |g_{j,m}| \le \frac{|m|^\nu |m|!}{Cm!} r_1^{-|m|}.$$

Next we claim that

$$N \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|^\nu |m|!}{m! C} x^m = \frac{NP(x_1 + x_2 + \cdots + x_n)}{C(r_1 - x_1 - x_2 - \cdots - x_n)^{\nu+1}}.$$

For convenience, we denote $\hat{x} = x_1 + \cdots + x_n$. Notice that for $|x| < r_2$,

$$\frac{r_1}{r_1 - (x_1 + \cdots + x_n)} = \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|!}{m!} x^m.$$

We differentiate above both sides with respect to $\hat{x}$ and then multiply both sides by $\hat{x}$,

$$\frac{r_1 \hat{x}}{(r_1 - \hat{x})^2} = \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m||m|!}{m!} x^m.$$

We repeat this procedure ν times and obtain

$$\frac{P(\hat{x})}{(r_1 - \hat{x})^{\nu+1}} = \sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|^\nu |m|!}{m!} x^m,$$

where $P(\hat{x})$ is a polynomial of degree ν with coefficients depending only on r_1. Hence

$$\sum_{d=0}^{\infty} \frac{1}{r_1^d} \sum_{|m|=d} \frac{|m|^\nu |m|!}{m!} |x^m| = \frac{P(|x_1| + \cdots + |x_n|)}{(r_1 - (|x_1| + \cdots + |x_n|))^{\nu+1}},$$

which yields the conclusion. □

Let $r_3 := r_2/2$ and

$$\hat{N} := \max_{|x| \le r_3} \frac{P(|x_1| + \cdots + |x_n|)}{C(r_1 - (|x_1| + \cdots + |x_n|))^{\nu+1}}$$

and

$$M := \max_{|x|\le r} \sum_{d=2}^{\infty} \left(|\beta^{[d]}(x)| + |Tf^{[d]}(x)| \right).$$

THEOREM 2. *Let $\theta_k(x)$ be the solution of*

$$A\theta_k(x) - \frac{\partial \theta_k}{\partial x}(x)Fx = \frac{\partial \theta_{k-1}}{\partial x}(x)\bar{f}(x), \qquad \theta_k(0) = 0.$$

Then for any $|x| \le qr_3$ with $0 < q < 1$ we have

$$\|\theta_k(x)\| \le b_k(q)\hat{N}^{k-2}M.$$

Proof. According to the previous lemma, we know that

$$\|\theta_2(x)\| \le M\hat{N} \quad \text{for} \quad |x| \le r_3.$$

Applying the lemma in a recursive way yields

$$\|\theta_k(x)\| \le M\hat{N}^{k-1} \quad \text{for} \quad |x| \le r_3 \quad \text{and} \quad k = 3, 4, \ldots.$$

Let $g(x) = \frac{\partial \theta_{k-1}}{\partial x}(x)\bar{f}(x)$. Then $g(x)$ can be expanded into a Taylor series in $|x| \le r_3$:

$$g(x) = g^{[k]}(x) + g^{[k+1]}(x) + \cdots$$

with $g^{[d]}(x) = \sum_{j=1}^{n} \sum_{|m|=d} g_{j,m}\mathbf{e}^j x^m$. Similar to the proof given in Theorem 1,

$$\|\theta_k(x)\| \le b_k(q) \sum_{d=k}^{\infty} \left(\sum_{|m|=d} |g_{j,m}| (\sqrt{q}r_3)^{|m|} \right) \le b_k(q)\hat{N}^{k-2}M,$$

and the proof is complete. ∎

COROLLARY 2. *When q is small enough, the series*

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

converges in $|x| \le qr_3$, where $\theta_d(x)$ for $d = 2, 3, \ldots$ is the solution of (2.10).

Proof. Let $q \le \frac{1}{2^{\nu+1}\hat{N}}$. It is sufficient to show that

$$\theta_k(x) + \theta_{k+1}(x) + \cdots$$

converges for some fixed k in $|x| \le qr_3$. According to the definition of $b_k(q)$, we know that when $k \ge 2\nu / \ln \frac{1}{q}$, the following holds:

$$b_k(q) > b_{k+1}(q) > \cdots > b_d(q) > \cdots \to 0 \qquad \text{as} \quad d \to \infty.$$

Choose $k \ge 2\nu / \ln \frac{1}{q}$ and notice that

$$b_k(q) = k^\nu q^k, \quad b_{k+1} = (k+1)^\nu q^{k+1}, \ldots, \quad b_d(q) = d^\nu q^d, \ldots.$$

According to Theorem 2, we have

$$\|\theta_k(x)\| + \|\theta_{k+1}(x)\| + \cdots \le b_k(q)\hat{N}^{k-2}M + b_{k+1}(q)\hat{N}^{k-1}M + \cdots.$$

Since

$$\frac{b_{d+1}(q)\hat{N}^{d-1}M}{b_d(q)\hat{N}^{d-2}M} = \left(1+\frac{1}{d}\right)^{\nu} q\hat{N} < 2^{\nu} q\hat{N} \le \frac{1}{2}, \qquad d \ge k,$$

we thus complete the proof. □

From Corollary 2, we know that series

$$\theta_2(x) + \theta_3(x) + \cdots + \theta_d(x) + \cdots \tag{2.12}$$

defines an analytic function in $|x| \le qr_3$. Now we are ready to prove the main result of this paper.

Proof of the main theorem. We first define two functions in $|x| \le qr_3$:

$$\theta(x) := Tx + \theta_2(x) + \theta_3(x) + \cdots + \theta_d(x) + \cdots$$

and

$$\theta^L(x) := Tx + \theta_2(x) + \theta_3(x) + \cdots + \theta_L(x),$$

where $\theta_2(x), \theta_3(x), \ldots$ are the solutions of (2.9), (2.10). We next show that $\theta(x)$ solves (1.3). Now

$$\begin{aligned} A\theta^L(x) - \frac{\partial \theta^L(x)}{\partial x} f(x) - \beta(h(x)) \\ = A\theta^L(x) - \frac{\partial \theta^L(x)}{\partial x}(Fx + \bar{f}(x)) - (BHx + \bar{\beta}(x)) \\ = \frac{\partial \theta_L(x)}{\partial x}\bar{f}(x). \end{aligned}$$

If $|x| \le qr_3$, then $\|\theta_L(x)\| \le b_L(q)\hat{N}^{L-2}M$ and

$$\left\| \frac{\partial \theta_L(x)}{\partial x}\bar{f}(x) \right\| \le b_L(q)\hat{N}^{L-2}M \to 0 \qquad \text{as} \quad L \to \infty$$

since series

$$b_k(q)\hat{N}^{k-2}M + b_{k+1}(q)\hat{N}^{k-1}M + \cdots$$

converges. Therefore $\theta(x)$ is an analytic solution of (1.3). Uniqueness follows from the uniqueness of the formal power series. □

A slight modification of the proof of the main theorem yields the following.

COROLLARY 3 (Lyapunov's auxiliary theorem). *Assume that $f : \mathbf{R}^n \to \mathbf{R}^n$ and $\gamma : \mathbf{R}^n \to \mathbf{R}^n$ are analytic vector fields with $f(0) = 0$, $\frac{\partial f}{\partial x}(0) = F$, and $\gamma(0) = 0$. Suppose that the eigenvalues $\lambda_1, \ldots, \lambda_n$ of F lie wholly in the open left half plane or lie wholly in the open right half plane. Let A be an $n \times n$ matrix with eigenvalues $\mu_1, \ldots, \mu_n$ such that there do not exist nonnegative integers $m_1, m_2, \ldots, m_n$ not all zero such that $\sum_{i=1}^n m_i\lambda_i = \mu_j$. Then there is a unique analytic solution in some neighborhood of the origin of the first-order PDE:*

$$\frac{\partial \theta}{\partial x}(x) f(x) - A\theta(x) + \gamma(x) = 0$$

with initial condition $\theta(0) = 0$.

Proof. Let $h(x) = x$ and $\beta(h(x) = \gamma(x)$. The main theorem cannot be applied directly because the Lyapunov auxiliary theorem does not require $\theta(x)$ to be a local diffeomorphism. But the proof stills holds provided we can show that the spectrum of A is of class (C, ν) with respect to the spectrum of F. Suppose the spectrum of F lies wholly in the open right half plane. Then there is a constant $c > 0$ such that $c \leq Re\ \lambda_i$, $i = 1, \ldots, n$. Suppose $M \geq Re\ \mu_j$, $j = 1, \ldots, n$. Then

$$|m \cdot \lambda - \mu_j| \geq 1$$

whenever $|m| \geq \frac{M+1}{c}$. Let $\nu = 1$ and choose $0 < C \leq 1$ satisfying

$$|m \cdot \lambda - \mu_j| \geq C$$

whenever $|m| < \frac{M+1}{c}$. This is possible because the left side is never zero. We have shown that the spectrum of A is of class (C, ν) with respect to the spectrum of F. □

COROLLARY 4 (Siegel's theorem). *Assume that $f : \mathbf{R}^n \to \mathbf{R}^n$ is an analytic vector field with $f(0) = 0$, $\frac{\partial f}{\partial x}(0) = F$. Suppose, for some $C > 0, \nu > 0$, the eigenvalues of F are of type (C, ν) with respect to $\sigma(F)$. Then there is an analytic solution in some neighborhood of the origin of the first-order PDE:*

$$\frac{\partial \theta}{\partial x}(x) f(x) = F\theta(x)$$

with initial condition $\theta(0) = 0$. Moreover $z = \theta(x)$ is a local analytic diffeomorphism around $x = 0$ which transforms the differential equation

$$\dot{x} = f(x)$$

into its linear part

$$\dot{z} = Fz.$$

Proof. Apply the main theorem with $\beta = 0$, $A = F$, and $T = I$. □

Note. Lyapunov's auxiliary theorem and Siegel's theorem are usually stated for complex analytic vector fields. We have stated them for real analytic vector fields since we stated our main theorem that way. But the proof of the main theorem holds for complex vector fields too.

3. Examples. As discussed in the introduction, there are distinct advantages to considering *nonlinear output injection* $\beta(y)$. It is desirable that θ be a diffeomorphism over as large a range as possible, for this is the domain of convergence of the observer. Nonlinear output injection can make θ a global diffeomorphism.

To illustrate this, we consider a Duffing oscillator

$$\begin{aligned} \ddot{x} &= x - x^3, \\ y &= x, \end{aligned}$$

which is equivalent to the planar system

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -x_1^3 \end{bmatrix}, \\ y &= x_1. \end{aligned}$$

This system is trivially transformed into a linear system with output injection (1.2)

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} - \begin{bmatrix} -2y \\ -3y + y^3 \end{bmatrix}$$

by

$$\begin{aligned} \theta(x) &= x, \\ \beta(y) &= \begin{bmatrix} -2y \\ -3y + y^3 \end{bmatrix}. \end{aligned}$$

Notice that β is nonlinear and θ is trivially a global diffeomorphism. The observer (1.4) is

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} - \begin{bmatrix} -2y \\ -3y + y^3 \end{bmatrix},$$

and the error dynamics

$$\begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}$$

is linear and exponentially stable with poles at $-1 \pm i$.

The example is trivial but illustrates two important facts. The first is the advantage of allowing nonlinear β. We could take it to be linear,

$$\beta(y) = \begin{bmatrix} -2 \\ -3 \end{bmatrix} y,$$

and still solve the PDE (1.3) for θ. But the solution might be hard to find, it could have an infinite power series expansion, and it might not be a global diffeomorphism.

The second point is that the Duffing oscillator is truly nonlinear; it has three equilibria and two homoclinic orbits, and the rest of the trajectories are limit cycles. Yet it is possible to build a globally convergent error with linear error dynamics.

Next we consider a Van der Pol oscillator,

$$\begin{aligned} \ddot{x} &= -(x^2 - 1)\dot{x} - x, \\ y &= x, \end{aligned}$$

which is equivalent to the planar system

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ x_1^2 x_2 \end{bmatrix}, \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

Now we have

$$\begin{aligned} f(x) &= \begin{bmatrix} x_2 \\ -x_1 + x_2 - x_1^2 x_2 \end{bmatrix}, \qquad h(x) = x_1, \\ F &= \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \qquad H = \begin{bmatrix} 1 & 0 \end{bmatrix}. \end{aligned}$$

We look for a nonlinear coordinate transformation $z = \theta(x)$ such that in the new coordinates z, the system can be described in the form

$$\dot{z} = Az - \beta(y).$$

Let us choose A and β to be

$$A = \begin{bmatrix} b_1 & 1 \\ b_2 - 1 & 1 \end{bmatrix}, \qquad \beta(y) = \begin{bmatrix} b_1 y + \frac{y^3}{3} \\ b_2 y + \frac{y^3}{3} \end{bmatrix},$$

where b_1, b_2 are constants such that $1 + b_1 < 0$, $b_1 - b_2 + 1 > 0$. Clearly, A is stable since $\operatorname{trace}(A) = 1 + b_1 < 0$ and $\det(A) = b_1 - b_2 + 1 > 0$. Moreover $A = F + BH$ with $B = [b_1, b_2]'$. The solution of (1.3) in this case is given by

$$\theta(x) = \begin{bmatrix} x_1 \\ x_2 + \frac{x_1^3}{3} \end{bmatrix}.$$

Note that θ is polynomial and *globally invertible* on $\mathbf{R}^2$. This is because we chose a nonlinear β. The resulting observer is again globally convergent with exponentially stable linear error dynamics in $\tilde{z}$ coordinates despite the nonlinearities of the Van der Pol oscillator. See Figure 1.

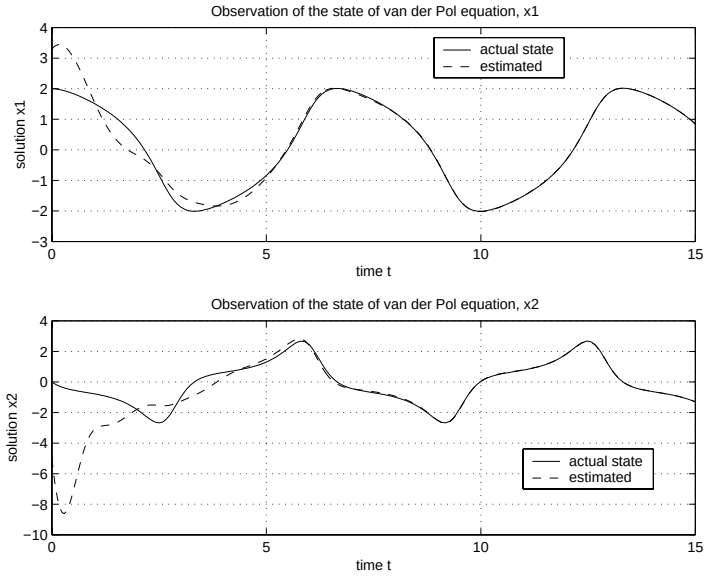

FIG. 1. *Observation of Van der Pol oscillator.*

Both these examples could be treated by the method of Krener and Respondek [8]. In particular, they showed that any observable two-dimensional system of the form

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= x_2, \\ \dot{x}_2 &= f_2(x) = a(x_1) + b(x_1)x_2 + c(x_1)x_2^2, \end{aligned}$$

where $a(x_1), b(x_1), c(x_1)$ are smooth functions, admits a local observer with linear error dynamics in transformed coordinates. But their method is not applicable to more general f_2. The above method is applicable to any observable system with arbitrary f_2. The conditions of Krener and Respondek become more restrictive as the dimension of the system is increased, while there are no additional conditions for the above method.

The next example cannot be treated by the method of Krener and Respondek:

$$\begin{aligned} y &= x_1, \\ \dot{x}_1 &= 2x_2, \\ \dot{x}_2 &= 2x_1 - 3x_1^2 - x_2(x_1^3 - x_1^2 + x_2^2). \end{aligned}$$

There is a saddle at $(0,0)$ and an unstable source at $(2/3,0)$. The stable and unstable manifolds of the saddle form a homoclinic orbit given by $x_1^3 - x_1^2 + x_2^2 = 0$ which wraps around the unstable source.

The system is linearly observable around $x = 0$ with

$$F = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

The spectrum of F is $\lambda = (2, -2)$. We choose a linear output injection based on a long time Kalman filter for the linear part of the system corrupted by standard white noises, and this leads to

$$A = \begin{bmatrix} -\sqrt{17} & 2 \\ -2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -\sqrt{17} \\ -4 \end{bmatrix}.$$

The spectrum of A is

$$\frac{-\sqrt{17} \pm 1}{2},$$

and clearly these are not resonant with the spectrum of F because they are not even integers.

First we compute θ for up to degree 3 with $\beta^{[2]} = 0$, $\beta^{[3]} = 0$:

$$\begin{aligned} \theta^{[1]}(x) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\ \theta^{[2]}(x) &= \begin{bmatrix} 1.2188 & -0.7731 & -0.2812 \\ 1.7394 & 0.2812 & -1.3529 \end{bmatrix} \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}, \\ \theta^{[3]}(x) &= \begin{bmatrix} -20.4026 & 19.1878 & 20.8159 & -20.1972 \\ -21.7136 & 20.8245 & 20.6972 & -20.8216 \end{bmatrix} \begin{bmatrix} x_1^3 \\ x_1^2 x_2 \\ x_1 x_2^2 \\ x_2^3 \end{bmatrix}. \end{aligned}$$

Figure 2 shows the system starting at $x_1 = 0.5, x_2 = 0$ and the observer starting at $\hat{x}_1 = 0, \hat{x}_2 = 0$. Clearly this observer does not converge; in particular, the observer seems to stall around $(0.3, 0.4)$. The problem appears to be caused by the large sizes of $\theta^{[2]}$ and $\theta^{[3]}$.

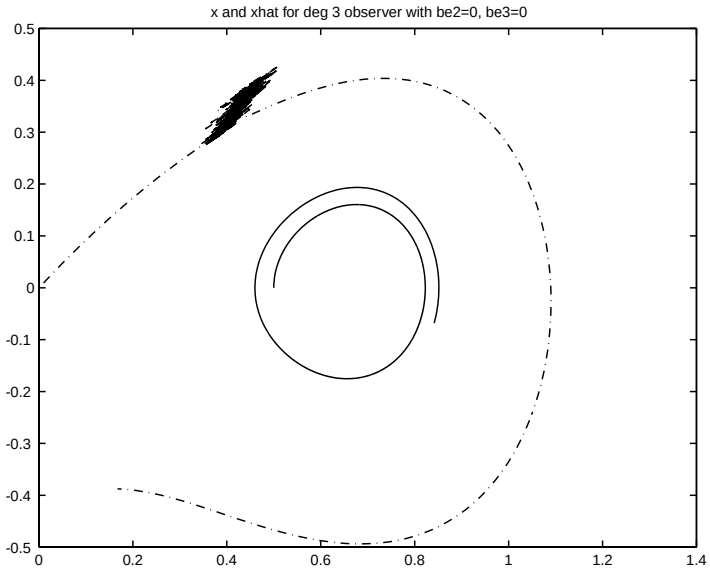

FIG. 2. *Solid line: state trajectory. Dashed line: observer trajectory.*

Next we choose $\beta^{[2]}$ to minimize the Euclidean norm of the coefficients of $\theta^{[2]}$, and then we choose $\beta^{[3]}$ to minimize the Euclidean norm of the coefficients of $\theta^{[3]}$. The result is

$$\theta^{[1]}(x) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

$$\theta^{[2]}(x) = \begin{bmatrix} 0.0000 & -0.0000 & 0.0000 \\ 0.0000 & -0.0000 & 0.0000 \end{bmatrix} \begin{bmatrix} x_1^2 \\ x_1x_2 \\ x_2^2 \end{bmatrix},$$

$$\theta^{[3]}(x) = \begin{bmatrix} 0.0330 & 0.0938 & -0.2219 & 0.1925 \\ -0.4030 & -0.1514 & 0.3075 & 0.1749 \end{bmatrix} \begin{bmatrix} x_1^3 \\ x_1^2x_2 \\ x_1x_2^2 \\ x_2^3 \end{bmatrix},$$

$$\beta(y) = \begin{bmatrix} -4.1231 & 0.0000 & -1.1296 \\ -4.0000 & 3.0000 & 0.2368 \end{bmatrix} \begin{bmatrix} y \\ y^2 \\ y^3 \end{bmatrix}.$$

Notice how much smaller $\theta^{[2]}$ and $\theta^{[3]}$ are. The resulting observer performs much better, as can be seen from Figure 3.

4. Nonlinear observer design with inputs. We now consider a nonlinear system with inputs:

$$\dot{x} = f(x,u), \tag{4.1}$$

$$y = h(x,u), \tag{4.2}$$

where $f : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ and $h : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^p$ are continuous. We assume here that

$$f(x,u) = f_0(x) + f_1(x,u), \qquad h(x,u) = h_0(x) + h_1(x,u)$$

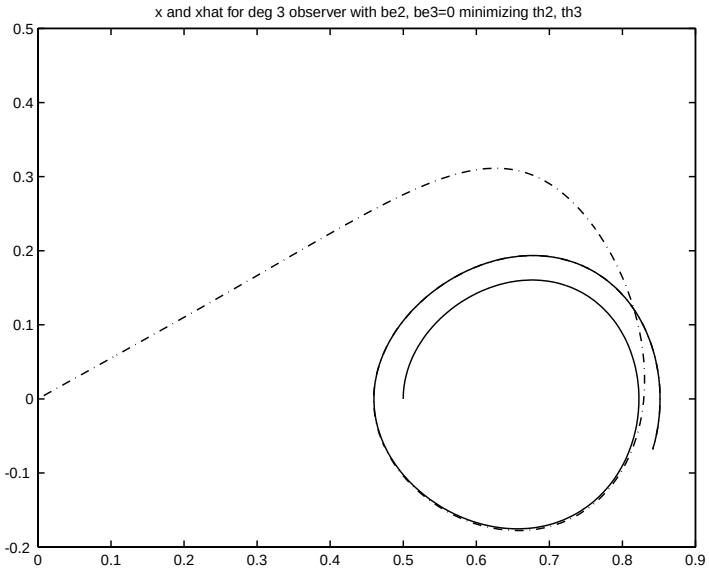

FIG. 3. *Solid line: state trajectory. Dashed line: observer trajectory.*

with $f_1(x,0) \equiv 0, h_1(x,0) \equiv 0$, and $f_0 : \mathbf{R}^n \to \mathbf{R}^n$ and $h_0 : \mathbf{R}^n \to \mathbf{R}^p$ are real analytic functions with $f_0(0) = 0, h_0(0) = 0$. Let $\beta : \mathbf{R}^p \to \mathbf{R}^n$ be a real analytic function and $F = \frac{\partial f_0}{\partial x}(0)$, $H = \frac{\partial h_0}{\partial x}(0)$, and $B = \frac{\partial \beta}{\partial x}(0)$. We further assume that

1. for a given $n \times n$ matrix A, there exists an invertible $n \times n$ matrix T so that $TFT^{-1} = A - BH$;
2. there exists a $C > 0, \nu > 0$ such that all the eigenvalues of A are of type (C, ν) with respect to $\sigma(F)$.

Then according to the main result of this paper, we know that the first-order PDE

$$\frac{\partial \phi}{\partial x}(x) f_0(x) = A\phi(x) - \beta(h_0(x)) \tag{4.3}$$

has a unique analytic solution $z = \phi$, which is a diffeomorphism in some neighborhood U of the origin with $\frac{\partial \phi}{\partial x}(0) = T$.

Now we let the estimate of the true state obey the equation

$$\dot{\hat{x}} = f(\hat{x}, u) - \left[\frac{\partial \phi}{\partial \hat{x}}\right]^{-1} (\beta(y) - \beta(h(\hat{x}, u))). \tag{4.4}$$

Let e denote

$$e = \phi(\hat{x}) - \phi(x).$$

Then e satisfies the differential equation

$$\begin{aligned} \dot{e} &= \frac{\partial \phi}{\partial \hat{x}} f(\hat{x}, u) - (\beta(y) - \beta(h(\hat{x}, u))) - \frac{\partial \phi}{\partial x} f(x, u) \\ &= \frac{\partial \phi}{\partial \hat{x}} (f_0(\hat{x}) + f_1)(\hat{x}, u) - (\beta(y) - \beta(h(\hat{x}, u))) - \frac{\partial \phi}{\partial x} (f_0(x) + f_1(x, u)). \end{aligned}$$

Since

$$\frac{\partial \phi}{\partial \hat{x}} f_0(\hat{x}) = A\phi(\hat{x}) - \beta(h_0(\hat{x})),$$
$$\frac{\partial \phi}{\partial x} f_0(x) = A\phi(x) - \beta(h_0(x)),$$

this yields

$$\dot{e} = Ae + N(\hat{x}, u) - N(x, u), \tag{4.5}$$

where the nonlinear function N is defined to be

$$N(x, u) := \frac{\partial \phi}{\partial x}(x) f_1(x, u) + \beta(h(x, u)) - \beta(h_0(x)). \tag{4.6}$$

We further assume that $f_1(\cdot, u)$ is locally Lipschitz about the origin; then there exists a positive constant $L(u)$ such that

$$\|N(x_1, u) - N(x_2, u)\| \le L(u)\|x_1 - x_2\|$$

for all x_1, x_2 in some open neighborhood U containing the origin. If we choose A to be Hurwitz, then for any given positive-definite $Q \in \mathbf{R}^{n\times n}$ there exists a unique positive-definite $P \in \mathbf{R}^{n\times n}$ such that

$$A^T P + PA = -2Q.$$

Now we consider the Lyapunov function

$$V(e) = e^T P e.$$

The derivative of $V(e)$ evaluated along the solution of the error dynamics is given by

$$\dot{V}(e) = \dot{e}^T P e + e^T P \dot{e} = -2e^T Q e + 2e^T P[N(x + e, u) - N(x, u)].$$

Therefore we have

$$\begin{aligned} \dot{V}(e) &\le -2e^T Q e + 2L(u)\|Pe\|\|e\| \\ &\le (-2\lambda_{\min}(Q) + 2L(u)\lambda_{\max}(P))\|e\|, \end{aligned}$$

where $\lambda_{\min}(Q)$ is the minimum eigenvalue of Q and $\lambda_{\max}(P)$ is the maximum eigenvalue of P. Hence if

$$\lambda_{\min}(Q)/\lambda_{\max}(P) > L(u),$$

then $e = 0$ is local asymptotically stable.

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*, Internat. J. Control, 38 (1983), pp. 419–431.

[3] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

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

[5] N. Kazantzis and C. Kravaris, *Nonlinear observer design using Lyapunov's auxiliary theorem*, Systems Control Lett., 34 (1998), pp. 241–247.
[6] A. J. Krener and A. Isidori, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.
[7] A. J. Krener, *Approximate linearization by state feedback and coordinate change*, Systems Control Lett., 5 (1984), pp. 181–185.
[8] A. J. Krener and W. Respondek, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.
[9] 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, 1994, pp. 231–250.
[10] A. M. Liapunov, *Stability of Motion*, Academic Press, New York, London, 1966.