[12] A. A. Jagers and E. A. Van Doorn, "On the continued Erlang loss function," *Oper. Res. Lett.*, vol. 5, pp. 43–47, 1986.
[13] F. P. Kelly, "Routing in circuit-switched networks: Optimization, shadow prices and decentralization," *Adv. Appl. Prob.*, vol. 20, pp. 112–144, 1988.
[14] L. Kleinrock, *Queueing Systems Volume II: Computer Applications*. New York: Wiley, 1976.
[15] P. Köchel, "Finite queueing systems—Structural investigations and optimal design," *Int. J. Prod. Econ.*, vol. 88, pp. 157–171, 2004.
[16] K. R. Krishnan, "The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1314–1316, Sep. 1990.
[17] J. M. Smith and F. R. B. Cruz, "The buffer allocation problem for general finite buffer queueing networks," *IIE Trans.*, vol. 37, pp. 343–365, 2005.
[18] E. Messerli, "Proof of a convexity property of the Erlang B formula," *Bell Syst. Tech. J.*, vol. 51, pp. 951–953, 1972.
[19] A. Pacheco, "Second-order properties of the loss probability in M/M/s/s+c systems," *Queueing Syst.*, vol. 15, pp. 289–308, 1994.
[20] J. W. Roberts, "A survey on statistical bandwidth sharing," *Computer Networks*, vol. 45, pp. 319–332, 2004.
[21] D. Sonderman, "Comparing multi-server queues with finite waiting rooms, I: Same number of servers," *Adv. Appl. Prob.*, vol. 11, pp. 439–447, 1979.
[22] W. Whitt, "Counterexamples for comparisons of queues with finite waiting rooms," *Queueing Syst.*, vol. 10, pp. 271–278, 1992.
[23] D. D. Yao and J. G. Shanthikumar, "The optimal input rates to a system of manufacturing cells," *Inform. Syst. Oper. Res.*, vol. 25, pp. 57–65, 1987.
[24] S. Ziya, H. Ayhan, R. D. Foley, and E. Pekoz, "A monotonicity result for a G/GI/c queue with balking or reneging," *J. Appl. Prob.*, vol. 43, pp. 1201–1205, 2006.

Necessary and Sufficient Conditions for State Equivalence to a Nonlinear Discrete-Time Observer Canonical Form

Hong-Gi Lee, *Member, IEEE*, Ari Arapostathis, *Fellow, IEEE*, and Steven I. Marcus, *Fellow, IEEE*

***Abstract*—In this technical note, we obtain necessary and sufficient conditions for a multi-input, multi-output, discrete-time nonlinear system to be state equivalent to a nonlinear observer form, and for an uncontrolled multi-output system to be state equivalent to a linear observer form. We adopt a geometric approach, and the proofs are constructive with respect to the required coordinate change.**

***Index Terms*—Nonlinear discrete-time control systems, nonlinear observer form, state equivalence.**

I. Introduction

The problem of observer design is prominent in control theory. Unlike linear systems, observer design for nonlinear systems is rather difficult. Observers for continuous time nonlinear systems were first investigated by Krener and Isidori [1] for time-invariant systems and Bestle and Zeitz [2] for time-varying systems, independently. The results were extended in [3]–[7]. For discrete-time systems, the problem has been investigated by several authors, see for example [8]–[22]. Lin and Byrnes [13] have obtained necessary and sufficient conditions for autonomous systems, but their approach does not seem to extend to systems with inputs. Califano *et al.* [8], Chung and Grizzle [9], and Lee and Nam [12], [15] have considered the problem under the restriction that the drift term is locally invertible. The work in [14] opened the path for direct nonlinear observer design without relying on the structure of linear observers. This was followed by the work in [10], [11], [16], [18]–[22].

In this technical note we revisit the problem of equivalence through a state transformation to a nonlinear observer canonical form [see (2)] of a discrete time system. We adopt a geometric approach and characterize equivalence through an auxiliary derived system [see (7)] whose dynamics are linked to those of the original system. Necessary and sufficient conditions for equivalence are given, and the proofs are constructive with respect to the required coordinate change. Concerning the autonomous system in (1b), a similar characterization is obtained in [13]. The method adopted allows us to characterize state equivalence of the multi-input, multi-output controlled system (1a) to the nonlinear observer form in (2). As far as we know such a characterization is lacking in the existing literature. Even some of the most recent papers that allow both state and output transformations (see for example

Manuscript received May 24, 2007; revised May 01, 2008 and June 11, 2008. Current version published December 10, 2008. This work was supported in part by the Chung-Ang University Research Fund, in part by the Office of Naval Research through the Electric Ship Research and Development Consortium, in part by the National Science Foundation under Grant ECS-0424169, and in part by the Air Force Office of Scientific Research under Grant F496200110161. Recommended by Associate Editor M. Xiao.

H.-G. Lee is with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 156-756, Korea.

A. Arapostathis is with the Department of Electrical and Computer Engineering at The University of Texas at Austin, Austin, TX 78712 USA (e-mail: ari@ece.utexas.edu).

S. I. Marcus is with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA.

Digital Object Identifier 10.1109/TAC.2008.2008321

[8]), restrict their attention to single-input, single-output, drift invertible systems. The direct method adopted utilizes the system impulse response (see Section II), and resembles the method used in [23].

Consider a discrete-time, time-invariant, controlled system of the form

$$\begin{aligned} x(t+1) &= f\left(x(t), u(t)\right), \quad f(0,0)=0 \\ y(t) &= h\left(x(t)\right), \quad h(0)=0 \end{aligned} \tag{1a}$$

or the autonomous system

$$\begin{aligned} x(t+1) &= f\left(x(t)\right), \quad f(0)=0 \\ y(t) &= h\left(x(t)\right), \quad h(0)=0 \end{aligned} \tag{1b}$$

with state $x \in \Sigma \simeq \mathbb{R}^n$, input $u \in \mathcal{U} \simeq \mathbb{R}^m$, and output $y \in \mathcal{Y} \simeq \mathbb{R}^p$.

Definition 1: System (1a), or (1b), is said to be state equivalent to a *nonlinear observer form*, if there exists a smooth diffeomorphism $T : V_0 \to \Sigma$, defined on some neighborhood of the origin $V_0 \subset \Sigma$, which transforms (1a), in the variable $z = T(x)$, to

$$\begin{aligned} z(t+1) &= Az(t) + \gamma\left(y(t), u(t)\right) \\ y(t) &= Cz(t) \end{aligned} \tag{2}$$

where $\gamma : \mathcal{Y} \times \mathcal{U} \to \Sigma$, or $\gamma : \mathcal{Y} \to \Sigma$, is a smooth function and (A, C) is an observable pair. In the case of (1b), if $\gamma \equiv 0$, then we say that it is state equivalent to a *linear observer form*.

Remark 1: If (1a) is state equivalent to a nonlinear observer form, then choosing $L \in \mathbb{R}^{n\times p}$ such that $(A - LC)$ is Hurwitz, we can design a state estimator

$$\begin{aligned} \hat{z}(t+1) &= (A-LC)\hat{z}(t) + \gamma\left(y(t), u(t)\right) + Ly(t) \\ \hat{x}(t) &= T^{-1}\left(\hat{z}(t)\right) \end{aligned}$$

which results in an asymptotically vanishing estimation error, i.e., $\lim_{t\to\infty} \|x(t) - \hat{x}(t)\| = 0$.

In this technical note, we obtain necessary and sufficient conditions for system (1b) to be state equivalent to a linear observer form, and for (1a), (1b) to be state equivalent to a nonlinear observer form. We also demonstrate the computations involved with several examples.

II. Notation and Definitions

In this section, we introduce some basic definitions. We refer the reader to [24]–[26] for basic results in nonlinear systems and differential geometry.

For a function $G : \Sigma \times \mathbb{R}^\ell \to \Sigma$, we define the "impulse response" $\hat{G}^i$ by

$$\begin{aligned} &\hat{G}^0(x,v) \triangleq x, \quad \hat{G}^1(x,v) \triangleq G(x,v), \\ &\hat{G}^i(x,v) \triangleq G\left(G^{i-1}(x,v), 0\right), \quad i \geq 2. \end{aligned}$$

If $G : \Sigma \to \Sigma$, then $\hat{G}^i$ is the i-fold composition of G, also denoted by G^i.

For a function $F : \Sigma \times \mathcal{U} \times \mathbb{R}^p \to \Sigma$ with $F(0,0,0) = 0$, we define $\Psi_k \equiv \Psi_k[F] : \mathbb{R}^{p\times k} \to \Sigma$, $k \in \mathbb{N}$, by $\Psi_1(w) \triangleq F(0,0,w)$, and

$$\Psi_{k+1}(w^1, w^2, \ldots, w^{k+1}) \triangleq F\left(\Psi_k(w^2, w^3, \ldots, w^{k+1}), 0, w^1\right)$$

with $w^i = (w_1^i, \ldots, w_p^i) \in \mathbb{R}^p$. We also adopt the convention $\Psi_0 \equiv 0$. If $F : \Sigma \times \mathbb{R}^p \to \Sigma$, the analogous definition applies. Observe that

$$\Psi_n(w^1, w^2, \ldots, w^n) = \Psi_{n+1}(w^1, w^2, \ldots, w^n, 0).$$

The symbol "$\circ$" denotes composition of functions. Let ν_i be the least nonnegative integer such that $d_x(h_i \circ \hat{f}^{\nu_i})(0,0)$ is linearly dependent on the vectors in the collection

$$\Big\{ d_x(h_j \circ \hat{f}^k)(0,0), d_x(h_\ell \circ \hat{f}^{\nu_i})(0,0), \\ 1 \leq j \leq p, 0 \leq k < \nu_i, 1 \leq \ell < i \Big\}.$$

The integers $\{\nu_1, \ldots, \nu_p\}$ are the *observability indices* of (1b) [or (1a)]. Let $\bar{\nu} \triangleq \max_i\{\nu_i\}$, and define the subspace $\Delta \subset \mathbb{R}^{p\times\bar{\nu}}$, by

$$\Delta \triangleq \left\{ w_j^i, 1 \leq i \leq \bar{\nu}, 1 \leq j \leq p | w_j^i = 0, \text{ if } i > \nu_j \right\}.$$

An element $w \in \Delta$ is also viewed as an element of $\mathbb{R}^\kappa$ in the ordered coordinates $(w_1^1, \ldots, w_1^{\nu_1}, \ldots, w_p^1, \ldots, w_p^{\nu_p})$, with $\kappa \triangleq \nu_1 + \cdots + \nu_p$. Thus if $M \in \mathbb{R}^{\ell\times\kappa}$, then $Mw \in \mathbb{R}^\ell$ is well defined.

Let $\bar{\Psi}$ denote the restriction of $\Psi_{\bar{\nu}}$ on Δ and define

$$\mathcal{F} \equiv \mathcal{F}[F] : \mathbb{R}^p \times \Delta \times \mathcal{U} \to \Sigma$$

by

$$\mathcal{F}(w^0, w^1, \ldots, w^{\bar{\nu}}, u) \triangleq F\left(\bar{\Psi}(w^1, \ldots, w^{\bar{\nu}}), u, w^0\right). \tag{3}$$

If $F : \Sigma \times \mathbb{R}^p \to \Sigma$, then $\mathcal{F}[F] : \mathbb{R}^p \times \Delta \to \Sigma$ is similarly defined.

III. Main Results

In this section, we obtain necessary and sufficient conditions for systems (1a), (1b) to be state equivalent to a nonlinear (or linear) observer form. Without loss of generality we assume throughout that $\nu_i \neq 0$, for $i = 1, \ldots, p$.

First, suppose system (1b) satisfies $\sum_{j=1}^p \nu_j = n$, or equivalently that

$$\operatorname{rank}\left\{ d(h_j \circ \hat{f}^k)(0), 1 \leq j \leq p, 0 \leq k < \nu_i \right\} = n.$$

Then, by the inverse function theorem, there exists a unique smooth function $F : V_0 \to \Sigma$, defined on some neighborhood of the origin $V_0 \subset \Sigma \times \mathbb{R}^p$ which satisfies, for $1 \leq j \leq p$

$$h_j \circ \hat{F}^i(x,w) = \begin{cases} h_j \circ f^i(x), & 1 \leq i < \nu_j \\ h_j \circ f^{\nu_j}(x) + w_j, & i = \nu_j. \end{cases} \tag{4}$$

Note that

$$h_j \circ f^{i-1} \circ F(x,w) = \begin{cases} h_j \circ f^i(x), & 1 \leq i < \nu_j \\ h_j \circ f^{\nu_j}(x) + w_j, & i = \nu_j. \end{cases} \tag{5}$$

Iterating (5), we obtain

$$h_j \circ f^{i-1} \circ \Psi_{\bar{\nu}}(w^1, \ldots, w^{\bar{\nu}}) = w_j^{\nu_j - i+1} \\ + h_j \circ f^{\nu_j} \circ \Psi_{\bar{\nu}-\nu_j+i-1}(w^{\nu_j-i+2}, \ldots, w^{\bar{\nu}}). \tag{6}$$

We define the *derived system* of (1b) by

$$x(t+1) = F\left(x(t), w(t)\right), \quad y(t) = h\left(x(t)\right). \tag{7}$$

The following theorem characterizes state equivalence to a linear observer form for system (1b). Note that the result is local in nature and the proposed observer is only guaranteed to work in some open neighborhood of the equilibrium point.

Theorem 1: Let $\mathcal{F}$ be the map defined in (3), relative to F in (7). System (1b) is state equivalent to a linear observer form if and only if, in V_0, a neighborhood of the origin:

i) $\sum_{j=1}^{p} \nu_j = n$.

ii) $d(h_j \circ f^{\nu_j}) \in \mathrm{span}\{d(h_i \circ f^{\ell-1}), 1 \leq i \leq p, 1 \leq \ell \leq \nu_j\}$, for $1 \leq j \leq p$.

iii) $\mathcal{F}_*(\partial/\partial w_j^i)$, $1 \leq j \leq p, 0 \leq i \leq \nu_j$ are well-defined vector fields.

Furthermore, $\xi = T(x) = \bar{\Psi}^{-1}(x)$ is a linearizing coordinate change.

Proof: (Necessity) Suppose that there exists $z = T(x)$ such that

$$z(t+1) = Az(t), \quad y(t) = Cz(t).$$

We can assume without loss of generality that A and C are in observer canonical form, i.e.,

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix} \tag{8a}$$

and $C = \mathrm{blockdiag}\{C_1, \ldots, C_p\}$, with $A_{ji} \in \mathbb{R}^{\nu_j \times \nu_i}$ given by

$$A_{ii} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha_{i1}^i \\ 1 & 0 & \cdots & 0 & \alpha_{i2}^i \\ 0 & 1 & \cdots & 0 & \alpha_{i3}^i \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha_{i\nu_i}^i \end{bmatrix} \tag{8b}$$

and

$$A_{ji} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha_{j1}^i \\ 0 & 0 & \cdots & 0 & \alpha_{j2}^i \\ 0 & 0 & \cdots & 0 & \alpha_{j3}^i \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \alpha_{j\nu_j}^i \end{bmatrix}, \quad i \neq j \tag{8c}$$

and $C_j = [0 \cdots 0\ 1] \in \mathbb{R}^{\nu_j}$. Let $\tilde{f}(z) = Az$, $\tilde{h}(z) = Cz$. Then, since $f = T^{-1} \circ \tilde{f} \circ T$ and $h = \tilde{h} \circ T$

$$h_j \circ f^i = \tilde{h}_j \circ \tilde{f}^i \circ T = C_{j,\cdot} A^i T, \quad 0 \leq i \leq \nu_j \tag{9}$$

where $C_{j,\cdot}$ denotes the jth row of C. Let

$$\mathcal{V}_j \triangleq \begin{bmatrix} dh_j(0) \\ d(h_j \circ f)(0) \\ \vdots \\ d(h \circ f^{\nu_j - 1})(0) \end{bmatrix}, \ \tilde{\mathcal{V}}_j \triangleq \begin{bmatrix} C_{j,\cdot} \\ C_{j,\cdot} A \\ \vdots \\ C_{j,\cdot} A^{\nu_j - 1} \end{bmatrix}; \ \mathcal{V} \triangleq \begin{bmatrix} \mathcal{V}_1 \\ \vdots \\ \mathcal{V}_p \end{bmatrix}$$

and similarly for $\tilde{\mathcal{V}}$. By (9), $\mathcal{V} = \tilde{\mathcal{V}} DT(0)$, and since, by i), $\mathcal{V}$ is nonsingular, T is a local diffeomorphism. Also, $d(h_j \circ f^{\nu_j}) = C_{j,\cdot} A^{\nu_j} DT$, and ii) follows since:

$$C_{j,\cdot} A^{\nu_j} \in \mathrm{span}\{C_{i,\cdot} A^{\ell - 1}, 1 \leq i \leq p, 1 \leq \ell \leq \nu_j\}. \tag{10}$$

By (5) and (9)

$$C_{j,\cdot} A^i T \circ F(x,w) = \begin{cases} C_{j,\cdot} A^{i+1} T(x), & 0 \leq i < \nu_j - 1 \\ C_{j,\cdot} A^{\nu_j} T(x) + w_j, & i = \nu_j - 1 \end{cases}$$

which, written in matrix form, yields $\tilde{\mathcal{V}} T \circ F(x,w) = \tilde{\mathcal{V}} A T(x) + C^{\mathsf{T}} w$, or

$$\tilde{F}(z,w) \triangleq T \circ F\left(T^{-1}(z), w\right) = Az + Bw \tag{11}$$

with

$$B \triangleq \tilde{\mathcal{V}}^{-1} C^{\mathsf{T}} = \mathrm{blockdiag}\left([1\ 0\ \cdots\ 0]^{\mathsf{T}} \in \mathbb{R}^{\nu_i}, 1 \leq i \leq p\right). \tag{12}$$

Since $T(0) = 0$, using the relation $F(x,w) = T^{-1} \circ \tilde{F}(T(x), w)$ and (11), we obtain

$$\mathcal{F}(w^0, w^1, \ldots, w^{\bar{\nu}}) = T^{-1}(A^{\bar{\nu}} B w^{\bar{\nu}} + \cdots + ABw^1 + Bw^0).$$

Thus

$$\mathcal{F}_* \frac{\partial}{\partial w_j^i} = T_*^{-1}(A^i B_{\cdot,j}), \quad 0 \leq i \leq \nu_j, \quad 1 \leq j \leq p$$

and condition iii) holds.

(Sufficiency) Suppose that conditions i)–iii) are satisfied. By (6)

$$d(h_j \circ f^{i-1})(0) \frac{\partial \Psi_{\bar{\nu}}}{\partial w_k^\ell} = \begin{cases} 1, & k = j,\ \ell = \nu_j - i + 1 \\ 0, & k \neq j,\ \ell = \nu_j - i + 1 \\ 0, & \ell < \nu_j - i + 1. \end{cases} \tag{13}$$

It is straightforward to show using (13) that $\mathcal{V} D\bar{\Psi}(0)$ is nonsingular. Thus, $\bar{\Psi}$ is a local diffeomorphism.

Next we show that $\tilde{h}(\xi) = h \circ \bar{\Psi}(\xi) = C\xi$. By (6), for $\xi = (\xi^1, \ldots, \xi^{\bar{\nu}}) \in \Delta$

$$h_j \circ \Psi_{\bar{\nu}}(\xi^1, \ldots, \xi^{\bar{\nu}}) = h_j \circ f^{\nu_j} \circ \Psi_{\bar{\nu} - \nu_j}(\xi^{\nu_j + 1}, \ldots, \xi^{\bar{\nu}}) + \xi_j^{\nu_j}. \tag{14}$$

We claim that $h_j \circ f^{\nu_j} \circ \Psi_{\bar{\nu} - \nu_j}(\xi^{\nu_j + 1}, \ldots, \xi^{\bar{\nu}}) = 0$. Then the result follows from (14) and this claim. To prove the claim note that assumption ii) implies that, for $1 \leq j \leq p$ and $i \geq \nu_j$

$$d(h_j \circ f^{\nu_j}) \in \mathrm{span}\left\{d(h_k \circ f^{\ell - 1}),\ 1 \leq k \leq p,\ 1 \leq \ell < \nu_j\right\}. \tag{15}$$

For $k = 1, \ldots, \bar{\nu}$, define

$$\Gamma_k \triangleq \Big\{d(h_j \circ f^{k-i} \circ \Psi_i)(\xi^{\bar{\nu} - i + 1}, \ldots, \xi^{\bar{\nu}}), \quad 1 \leq j \leq p,\ 1 \leq i \leq k\Big\}.$$

If $k < \bar{\nu}$, then by (5)

$$\begin{aligned} &h_j \circ f^{k-i} \circ \Psi_i(\xi^{\bar{\nu} - i + 1}, \ldots, \xi^{\bar{\nu}}) \\ &= \begin{cases} 0, & k < \nu_j \\ h_j \circ f^{\nu_j} \circ \Psi_{k - \nu_j}(\xi^{\bar{\nu} + \nu_j - k + 1}, \ldots, \xi^{\bar{\nu}}) + \xi_j^{\bar{\nu} + \nu_j - k}, & k \geq \nu_j. \end{cases} \end{aligned}$$

Therefore, since $\xi_j^\ell = 0$, for $\ell > \nu_j$, when $\xi \in \Delta$, it follows by (15) that for $k < \bar{\nu}$

$$\Gamma_k \subseteq \mathrm{span}\{\Gamma_i,\ 0 \leq i < k\}, \quad \text{on } \Delta. \tag{16}$$

Since $\Gamma_0 \equiv 0$, the claim follows by iterating (16).

It remains to show that $\tilde{f}(\xi) = \bar{\Psi}^{-1} \circ f \circ \bar{\Psi}(\xi) = A\xi$. Let

$$Y_j^i = \mathcal{F}_* \frac{\partial}{\partial w_j^{i-1}}, \quad 1 \leq j \leq p, \quad 1 \leq i \leq \nu_j + 1.$$

Then

$$Y_j^i = \mathcal{F}_* \frac{\partial}{\partial w_j^{i-1}} = \bar{\Psi}_* \frac{\partial}{\partial \xi_j^i}, \quad 1 \leq j \leq p, \quad 1 \leq i \leq \nu_j$$

which implies that $\{Y_j^i | 1 \leq j \leq p, 1 \leq i \leq \nu_j\}$ is a set of linearly independent vector fields. Since, for $1 \leq j, k \leq p$

$$\left[Y_j^i, Y_k^\ell\right] = \mathcal{F}_* \left[\frac{\partial}{\partial w_j^{i-1}}, \frac{\partial}{\partial w_k^{\ell-1}}\right] = 0$$

for all $i = 1, \ldots, \nu_j + 1$ and $\ell = 1, \ldots, \nu_k + 1$, it follows that $\{Y_j^i | 1 \leq j \leq p, 1 \leq i \leq \nu_j + 1\}$ is a set of $n+p$ commuting vector fields. Thus

$$Y_j^{\nu_j+1} = \sum_{k=1}^{p} \sum_{i=1}^{\nu_k} \alpha_{ji}^k Y_k^i$$

for some $\alpha_{ji}^k \in \mathbb{R}$, $1 \leq j, k \leq p$, $1 \leq i \leq \nu_k$. Since

$$\tilde{F}(\xi, w) = \bar{\Psi}^{-1} \circ F\left(\bar{\Psi}(\xi), w\right) = \bar{\Psi}^{-1} \circ \mathcal{F}(w, \xi), \quad \xi \in \Delta,$$

we obtain

$$\tilde{F}_* \frac{\partial}{\partial w_j} = (\bar{\Psi}^{-1} \circ \mathcal{F})_* \frac{\partial}{\partial w_j} = (\bar{\Psi}^{-1})_* Y_j^1 = \frac{\partial}{\partial \xi_j^1}. \quad (17)$$

Similarly, for $1 \leq i < \nu_j$

$$\tilde{F}_* \frac{\partial}{\partial \xi_j^i} = (\bar{\Psi}^{-1} \circ \mathcal{F})_* \frac{\partial}{\partial \xi_j^i} = (\bar{\Psi}^{-1})_* Y_j^{i+1} = \frac{\partial}{\partial \xi_j^{i+1}} \quad (18)$$

and

$$\begin{aligned}\tilde{F}_* \frac{\partial}{\partial \xi_j^{\nu_j}} &= (\bar{\Psi}^{-1} \circ \mathcal{F})_* \frac{\partial}{\partial \xi_j^{\nu_j}} = (\bar{\Psi}^{-1})_* Y_j^{\nu_j+1} \\ &= (\bar{\Psi}^{-1})_* \left(\sum_{k=1}^{p} \sum_{i=1}^{\nu_k} \alpha_{ji}^k Y_k^i\right) \\ &= \sum_{k=1}^{p} \sum_{i=1}^{\nu_k} \alpha_{ji}^k \frac{\partial}{\partial \xi_k^i}. \end{aligned} \quad (19)$$

By (17)–(19), it follows that $\tilde{F}(\xi, w) = A\xi + Bw$. By (5)

$$\tilde{h}_j \circ \tilde{f}^{i-1} \circ \tilde{F}(\xi, 0) = \tilde{h}_j \circ \tilde{f}^i(\xi), \quad 1 \leq j \leq p, \quad 1 \leq i \leq \nu_j. \quad (20)$$

Hence, (20) and assumption i) imply that for some local diffeomorphism $G : \Sigma \to \Sigma$ whose components are the functions $\{\tilde{h}_j \circ \tilde{f}^{i-1}\}$, we have $G(A\xi) = G \circ \tilde{F}(\xi, 0) = G \circ \tilde{f}(\xi)$. This shows that $\tilde{f}(\xi) = A\xi$. ∎

Next, we consider the observer problem for system (1a). Assuming that $\sum_{j=1}^{p} \nu_j = n$, we define the derived system of (1a) by

$$x(t+1) = F\left(x(t), u(t), w(t)\right), \quad y(t) = h\left(x(t)\right).$$

The function $F : \Sigma \times \mathcal{U} \times \mathbb{R}^p \to \Sigma$ is defined as

$$h_j \circ \hat{F}^i(x, u, w) = \begin{cases} h_j \circ \hat{f}^i(x, u), & 1 \leq i < \nu_j \\ h_j \circ \hat{f}^{\nu_j}(x, u) + w_j, & i = \nu_j. \end{cases} \quad (21)$$

Existence of F is guaranteed by the inverse function theorem. It holds that

$$h_j \circ \hat{f}^{i-1}\left(F(x, u, w), 0\right) = \begin{cases} h_j \circ \hat{f}^i(x, u), & 1 \leq i < \nu_j \\ h_j \circ \hat{f}^{\nu_j}(x, u) + w_j, & i = \nu_j. \end{cases} \quad (22)$$

Theorem 2: Let $\mathcal{F} = \mathcal{F}[F]$, with F the function defined in (21). System (1a) is state equivalent to a nonlinear observer form if and only if, in V_0, a neighborhood of the origin:

i) $\sum_{j=1}^{p} \nu_j = n$.

ii) $d_x(h_j \circ \hat{f}^{\nu_j}) \in \text{span}\{d_x(h_i \circ \hat{f}^{\ell-1}), 1 \leq i \leq p, 1 \leq \ell \leq \nu_j\}$, for $1 \leq j \leq p$.

iii) $\mathcal{F}_*(\partial/\partial w_j^i)$, $1 \leq j \leq p$, $0 \leq i < \nu_j$, are well-defined vector fields.

Furthermore, $\xi = T(x) = \bar{\Psi}^{-1}(x)$ is a linearizing coordinate change.

Proof: (Necessity); Suppose that there exists $z = T(x)$ such that (2) holds. Thus

$$\begin{aligned}\tilde{f}(z) &= T \circ f\left(T^{-1}(z), u\right) = A_0 z + \gamma(Cz, u) \\ \tilde{h}(z) &= h \circ T^{-1}(z) = Cz.\end{aligned}$$

We assume, without loss of generality, that A_0 and C take the form (8), with $\alpha_{jk}^i \equiv 0$. We have

$$h \circ \hat{f}^i(x, u) = \tilde{h} \circ \hat{\tilde{f}}^i\left(T(x), u\right), \quad i \geq 0. \quad (23)$$

Expand $\hat{\tilde{f}}^i$ as

$$\begin{aligned}\hat{\tilde{f}}^i(z, u) &= A_0^i z + A_0^{i-1}\gamma(Cz, u) \\ &\quad + A_0^{i-2}\gamma\left(C\hat{\tilde{f}}^1(z, u), 0\right) + \cdots \\ &= A_0^i z + \sum_{j=1}^{i} A_0^{i-j}\gamma\left(C\hat{\tilde{f}}^{j-1}(z, u), 0\right). \end{aligned} \quad (24)$$

Then, using (24),

$$\begin{bmatrix} d\tilde{h} \\ d(\tilde{h} \circ \hat{\tilde{f}}) \\ \vdots \\ d(\tilde{h} \circ \hat{\tilde{f}}^{n-1}) \end{bmatrix}(0, 0) = \begin{bmatrix} I & 0 & \cdots & 0 \\ CL & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CM_0^{n-2}L & CM_0^{n-3}L & \cdots & I \end{bmatrix} \times \begin{bmatrix} C \\ CA_0 \\ \vdots \\ CA_0^{n-1} \end{bmatrix} \quad (25)$$

where $L \triangleq D_1\gamma(0, 0)$, and $M_0 \triangleq A_0 + LC$. Necessity of i) then follows from (25). By (23) and (24)

$$\begin{aligned}h_j \circ \hat{f}^{\nu_j}(x, u) &= C_{j,\cdot} A_0^{\nu_j} T(x) \\ &\quad + \sum_{k=1}^{\nu_j} C_{j,\cdot} A_0^{\nu_j - k}\gamma\left(C\hat{\tilde{f}}^{k-1}\left(T(x), u\right), 0\right). \end{aligned} \quad (26)$$

Condition ii) then follows from (10) and (26).

By (22) and (23)

$$\begin{aligned}&\tilde{h}_j \circ \hat{\tilde{f}}^i\left(T \circ F(x, u, w), 0\right) \\ &\quad = \begin{cases} \tilde{h}_j \circ \hat{\tilde{f}}^{i+1}\left(T(x), u\right), & 0 \leq i < \nu_j - 1 \\ \tilde{h}_j \circ \hat{\tilde{f}}^{\nu_j}\left(T(x), u\right) + w_j, & i = \nu_j - 1. \end{cases} \end{aligned} \quad (27)$$

Using (24) we can verify that

$$T \circ F(x, u, w) = A_0 T(x) + \gamma\left(CT(x), u\right) + Bw \quad (28)$$

is a solution to (27), with B as in (12). The uniqueness of the solution to (27), as mentioned earlier, implies that (28) is the only solution. Thus $\tilde{F}(z, u, w) = A_0 z + \gamma(Cz, u) + Bw$, and

$$\begin{aligned}\bar{\Psi}(\xi^1, \ldots, \xi^{\bar{\nu}}) &= T^{-1}(A^{\bar{\nu}-1}B\xi^{\bar{\nu}} + \cdots + B\xi^1) \\ \mathcal{F}(w^0, \cdots, w^{\bar{\nu}}, u) &= T^{-1}\left(\gamma\left(w_1^{\nu_1}, \ldots, w_p^{\nu_p}, u\right) + A_0^{\bar{\nu}}Bw^{\bar{\nu}} \right. \\ &\quad \left. + \cdots + A_0 Bw^1 + Bw^0\right).\end{aligned}$$

Thus

$$\mathcal{F}_*\frac{\partial}{\partial w_j^i} = T_*^{-1}(A^iB_{\cdot,j}), \quad 1 \le j \le p, \quad 0 \le i < \nu_j$$

and condition iii) holds.

(Sufficiency); Suppose that conditions i)–iii) are satisfied. That $\bar{\Psi}$ is a local diffeomorphism and $\tilde{h}(\xi) = h(\bar{\Psi}(\xi)) = C\xi$, follow using the method in the proof of Theorem 1. It remains to show that $\tilde{f}(\xi, u) = A_0\xi + \gamma(C\xi, u)$. Let

$$Y_j^i = \mathcal{F}_*\frac{\partial}{\partial w_j^{i-1}}, \quad 1 \le j \le p, \quad 1 \le i \le \nu_j.$$

Then

$$Y_j^i = \mathcal{F}_*\frac{\partial}{\partial w_j^{i-1}} = \bar{\Psi}_*\frac{\partial}{\partial \xi_j^i}, \quad 1 \le j \le p, \quad 1 \le i \le \nu_j.$$

Since $\tilde{F}(\xi, u, w) = \bar{\Psi}^{-1} \circ F(\bar{\Psi}(\xi), u, w) = \bar{\Psi}^{-1} \circ \mathcal{F}(w, \xi, u)$, for $w \in \mathbb{R}^p$ and $\xi \in \Delta$, it follows that:

$$\tilde{F}_*\frac{\partial}{\partial w} = (\bar{\Psi}^{-1} \circ \mathcal{F})_*\frac{\partial}{\partial w_j} = (\bar{\Psi}^{-1})_*Y_j^1 = \frac{\partial}{\partial \xi_j^1}. \tag{29}$$

Similarly, for $1 \le i < \nu_j$

$$\tilde{F}_*\frac{\partial}{\partial \xi_j^i} = (\bar{\Psi}^{-1} \circ \mathcal{F})_*\frac{\partial}{\partial \xi_j^i} = (\bar{\Psi}^{-1})_*Y_j^{i+1} = \frac{\partial}{\partial \xi_j^{i+1}}. \tag{30}$$

By (29) and (30), $\tilde{F}(\xi, u, w) = A\xi + \gamma(C\xi, u) + Bw$, for some smooth γ. By (22)

$$\tilde{h}_j \circ \hat{\tilde{f}}^{i-1}\left(\tilde{F}(\xi, u, w), 0\right) = \begin{cases} \tilde{h}_j \circ \hat{\tilde{f}}^{i}(\xi, u), & 1 \le i < \nu_j \\ \tilde{h}_j \circ \hat{\tilde{f}}^{\nu_j}(\xi, u) + w_j, & i = \nu_j \end{cases}$$

and arguing as in the proof of Theorem 1, we obtain $\tilde{F}(\xi, u, 0) = \tilde{f}(\xi, u) = A_0\xi + \gamma(C\xi, u)$. ■

Remark 2: For a single-output system, hypotheses ii) of Theorem 1 and Theorem 2 are not needed, since in this case ii) is always satisfied.

Remark 3: Hypotheses iii) of Theorems 1 and 2 can be replaced by iii′) and iii″), respectively, [23], [27]:

iii′) $[\partial/\partial w_j^i, \ker\mathcal{F}_*] \subset \ker\mathcal{F}_*, 1 \le j \le p, 0 \le i \le \nu_j$.

iii″) $[\partial/\partial w_j^i, \ker\mathcal{F}_*] \subset \ker\mathcal{F}_*, 1 \le j \le p, 0 \le i < \nu_j$.

The following corollary applies to system (1b).

Corollary 1: The autonomous system (1b) is state equivalent to a nonlinear observer form if and only if, in V_0, a neighborhood of the origin:

i) $\sum_{j=1}^p \nu_j = n$.

ii) $d(h_j \circ f^{\nu_j}) \in \operatorname{span}\{d(h_i \circ f^{\ell-1}), 1 \le i \le p, 1 \le \ell \le \nu_j\}$, for $1 \le j \le p$.

iii) $\mathcal{F}_*(\partial/\partial w_j^i)$, $1 \le j \le p$, $0 \le i < \nu_j$, are well-defined vector fields.

Furthermore, $\xi = T(x) = \bar{\Psi}^{-1}(x)$ is a linearizing coordinate transformation.

IV. Examples

In this section, we employ four examples to demonstrate the computations involved. Since the systems used here are not drift invertible, the conditions in [8], [9], [12], [15] cannot be applied.

Example 1: Consider a system as in (1a) with

$$f(x, u) = \begin{pmatrix} x_2 - x_3^2 + x_1^2 \\ x_3 + x_1u + u^2 \\ u \end{pmatrix}, \quad h(x) = x_1.$$

A straightforward calculation yields

$$\begin{aligned} h \circ f(x, u) &= x_2 - x_3^2 + x_1^2, \\ h \circ \hat{f}^2(x, u) &= x_3 + x_1u + \left(x_2 - x_3^2 + x_1^2\right)^2, \\ h \circ \hat{f}^3(x, u) &= u + \left(x_3 + x_1u + \left(x_2 - x_3^2 + x_1^2\right)^2\right)^2. \end{aligned}$$

Thus, condition i) of Theorem 2 is satisfied. Condition ii) of Theorem 2 is trivially satisfied for a single-output system. With $F = (F_1 F_2 F_3)^{\mathsf{T}}$, (4) yields

$$\begin{aligned} h \circ F &= F_1 \\ &= x_2 - x_3^2 + x_1^2 \\ h \circ \hat{F}^2 &= F_2 - F_3^2 + F_1^2 \\ &= x_3 + x_1u + \left(x_2 - x_3^2 + x_1^2\right)^2 \\ h \circ \hat{F}^3 &= F_3 + \left(F_2 - F_3^2 + F_1^2\right)^2 \\ &= u + \left(x_3 + x_1u + \left(x_2 - x_3^2 + x_1^2\right)^2\right)^2 + w. \end{aligned} \tag{31}$$

Solving (31), we obtain

$$F(x, u, w) = \begin{pmatrix} x_2 - x_3^2 + x_1^2 \\ x_3 + x_1u + (u + w)^2 \\ u + w \end{pmatrix}.$$

Thus

$$\begin{aligned} \bar{\Psi}(w^1, w^2, w^3) &= \begin{pmatrix} w^3 \\ w^2 + (w^1)^2 \\ w^1 \end{pmatrix} \\ \mathcal{F}(w^0, w^1, w^2, w^3, u) &= \begin{pmatrix} w^2 + (w^3)^2 \\ w^1 + w^3u + (u + w^0)^2 \\ u + w^0 \end{pmatrix} \end{aligned}$$

and

$$\ker\mathcal{F}_* = \operatorname{span}\left\{-u\frac{\partial}{\partial w^1} - 2w^3\frac{\partial}{\partial w^2} + \frac{\partial}{\partial w^3}, -\frac{\partial}{\partial w^0} + \frac{\partial}{\partial u}\right\}.$$

It follows that $[\partial/\partial w^i, \ker\mathcal{F}_*] \subset \ker\mathcal{F}_*$, $0 \le i \le 2$, which implies that condition iii) of Theorem 2 holds. Hence, using the transformation $\bar{\Psi}^{-1}(x) = \begin{pmatrix} x_3 \\ x_2 - x_3^2 \\ x_1 \end{pmatrix}$, we obtain a nonlinear observer in canonical form as in Theorem 2, with $\gamma(y, u) = \begin{pmatrix} u \\ yu \\ y^2 \end{pmatrix}$.

Next are two examples of systems which are not state equivalent to an observer form.

Example 2:

$$f(x, u) = \begin{pmatrix} x_2^3 + u - \left(x_1 + 2x_2^2 + x_1u^2\right)^2 \\ x_1 + 2x_2^2 + x_1u^2 \end{pmatrix}, \quad h(x) = x_2.$$

We obtain

$$\begin{aligned} h \circ f &= x_1 + 2x_2^2 + x_1u^2 \\ h \circ \hat{f}^2 &= x_2^3 + u + \left(x_1 + 2x_2^2 + x_1u^2\right)^2. \end{aligned}$$

Solving for F and $\mathcal{F}$, we obtain

$$F(x,u,w) = \begin{pmatrix} x_2^3 + u - \left(x_1+2x_2^2+x_1u^2\right)^2 + w \\ x_1+2x_2^2+x_1u^2 \end{pmatrix}$$

$$\mathcal{F}(w^0,w^1,w^2,u) = \begin{pmatrix} (w^2)^3+u-\left(w^1+(w^2)^2+\left(w^1-(w^2)^2\right)u^2\right)^2+w^0 \\ w^1+(w^2)^2+(w^1-(w^2)^2)\,u^2 \end{pmatrix}.$$

Thus,

$$\ker\mathcal{F}_* = \operatorname{span}\left\{3(w^2)^2\frac{\partial}{\partial w^0} - \frac{2(w^2)(1-u^2)}{1+u^2}\frac{\partial}{\partial w^1} + \frac{\partial}{\partial w^2}, \left(-1+4u\left(w^1-(w^2)^2\right) \times \left(w^1+(w^2)^2+\left(w^1-(w^2)^2\right)u^2\right)\right) \times \frac{\partial}{\partial w^0} + \frac{\partial}{\partial u}\right\}$$

and $[\partial/\partial w^1, \ker\mathcal{F}_*] \not\subset \ker\mathcal{F}_*$, which implies that condition iii) of Theorem 2 does not hold.

Example 3: Consider the system

$$f(x,u) = \begin{pmatrix} x_2+u^2 \\ x_2x_3^2+u \\ x_1 \end{pmatrix}, \quad h(x) = \begin{pmatrix} x_1 \\ x_1+x_3 \end{pmatrix}. \tag{32}$$

Thus

$$h_1 = x_1, \quad h_1\circ f = x_2+u^2, \quad h_1\circ \hat{f}^2 = x_2x_3^2+u$$
$$h_2 = x_1+x_3, \quad h_2\circ f = x_2+x_1+u^2$$

and we obtain $\nu_1 = 2, \nu_2 = 1$. Condition ii) of Theorem 2 fails, since $dh_2\circ f \notin \operatorname{span}\{dh_1, dh_2\}$. Hence, (32) is not state equivalent to a nonlinear observer form.

Example 4: Consider the system

$$f(x,u) = \begin{pmatrix} x_3 \\ \left(x_2-x_3^2\right)^2+x_3^2+\left(1+x_1^2\right)u \\ x_2-x_3^2 \end{pmatrix}, \quad h(x) = \begin{pmatrix} x_1 \\ x_1+x_3 \end{pmatrix}. \tag{33}$$

We have

$$h_1 = x_1, \quad h_1\circ f = x_3,$$
$$h_2 = x_1+x_3, \quad h_2\circ f = x_3+x_2-x_3^2$$
$$h_2\circ\hat{f}^2 = x_2+\left(1+x_1^2\right)u,$$

and $\nu_1 = 1, \nu_2 = 2$. Thus, conditions i) and ii) of Theorem 2 are satisfied. If we let $F = (F_1F_2F_3)^\top$, then by (4)

$$\begin{aligned} F_1 &= x_3+w_1, \\ F_1+F_3 &= x_3+x_2-x_3^2 \\ F_3+F_2-F_3^2 &= x_2+\left(1+x_1^2\right)u+w_2 \end{aligned} \tag{34}$$

and solving (34) we obtain

$$F(x,u,w) = \begin{pmatrix} x_3+w_1 \\ x_3^2+\left(1+x_1^2\right)u+w_1+w_2+\left(x_2-x_3^2-w_1\right)^2 \\ x_2-x_3^2-w_1 \end{pmatrix}.$$

Thus

$$\mathcal{F}\left(w^0,w^1,w^2,u\right) = \begin{pmatrix} w_2^2-w_1^1+w_1^0 \\ \left(1+\left(w_1^1\right)^2\right)u+w_1^0+w_2^0+\left(w_1^1+w_2^1-w_1^0\right)^2 \\ w_1^1+w_2^1-w_1^0 \end{pmatrix}$$

and

$$\ker\mathcal{F}_* = \operatorname{span}\left\{\frac{\partial}{\partial w_1^0} - \left(1+2w_1^1u\right)\frac{\partial}{\partial w_2^0} + \frac{\partial}{\partial w_1^1}, -\frac{\partial}{\partial w_1^0} + \frac{\partial}{\partial w_2^0} - \frac{\partial}{\partial w_2^1} + \frac{\partial}{\partial w_2^2}, -\left(1+\left(w_1^1\right)^2\right)\frac{\partial}{\partial w_2^0} + \frac{\partial}{\partial u}\right\}.$$

Therefore

$$\left[\frac{\partial}{\partial w_j^i}, \ker\mathcal{F}_*\right] \subset \ker\mathcal{F}_*, \qquad 1\le j\le 2, \quad 0\le i<\nu_j$$

which implies that condition iii) of Theorem 2 is satisfied. Hence, system (33) is state equivalent to a nonlinear observer form. We have

$$\bar{\Psi}(w^1,w^2) = \begin{pmatrix} w_1^1 \\ w_1^1+w_2^1+\left(w_2^2-w_1^1\right)^2 \\ w_2^2-w_1^1 \end{pmatrix},$$

$$\bar{\Psi}^{-1}(x) = \begin{pmatrix} x_1 \\ x_2-x_1-x_3^2 \\ x_3+x_1 \end{pmatrix}$$

and transforming to the z-coordinates, we obtain a nonlinear observer in canonical form with

$$\gamma(y,u) = \begin{pmatrix} y_2-y_1 \\ -y_2+y_1+(y_2-y_1)^2+(1+y_1^2)u \\ y_2 \end{pmatrix}.$$

REFERENCES

[1] A. J. Krener and A. Isidori, "Linearization by output injection and nonlinear observers," *Syst. Control Lett.*, vol. 3, no. 1, pp. 47–52, 1983.

[2] D. Bestle and M. Zeitz, "Canonical form observer design for non-linear time-variable systems," *Int. J. Control*, vol. 38, no. 2, pp. 419–431, 1983.

[3] A. J. Krener and W. Respondek, "Nonlinear observers with linearizable error dynamics," *SIAM J. Control Optim.*, vol. 23, no. 2, pp. 197–216, 1985.

[4] X. H. Xia and W.-B. Gao, "Nonlinear observer design by observer canonical forms," *Int. J. Control*, vol. 47, no. 4, pp. 1081–1100, 1988.

[5] X. H. Xia and W.-B. Gao, "Nonlinear observer design by observer error linearization," *SIAM J. Control Optim.*, vol. 27, no. 1, pp. 199–216, 1989.

[6] A. J. Krener and M. Xiao, "Nonlinear observer design in the Siegel domain," *SIAM J. Control Optim.*, vol. 41, no. 3, pp. 932–953, 2002.

[7] A. J. Krener and M. Xiao, "Erratum: "Nonlinear observer design in the Siegel domain"," *SIAM J. Control Optim.*, vol. 43, no. 1, pp. 377–378, 2004.

[8] C. Califano, S. Monaco, and D. Normand-Cyrot, "On the observer design in discrete-time," *Syst. Control Lett.*, vol. 49, no. 4, pp. 255–265, 2003.

[9] S.-T. Chung and J. W. Grizzle, "Sampled-data observer error linearization," *Automatica J. IFAC*, vol. 26, no. 6, pp. 997–1007, 1990.

[10] G. Ciccarella, M. Dalla Mora, and A. Germani, "A robust observer for discrete time nonlinear systems," *Syst. Control Lett.*, vol. 24, no. 4, pp. 291–300, 1995.

[11] N. Kazantzis and C. Kravaris, "Discrete-time nonlinear observer design using functional equations," *Syst. Control Lett.*, vol. 42, no. 2, pp. 81–94, 2001.

[12] W. Lee and K. Nam, "Observer design for autonomous discrete-time nonlinear systems," *Syst. Control Lett.*, vol. 17, no. 1, pp. 49–58, 1991.

[13] W. Lin and C. I. Byrnes, "Remarks on linearization of discrete-time autonomous systems and nonlinear observer design," *Syst. Control Lett.*, vol. 25, no. 1, pp. 31–40, 1995.
[14] P. E. Moraal and J. W. Grizzle, "Observer design for nonlinear systems with discrete-time measurements," *IEEE Trans. Automat. Control*, vol. 40, no. 3, pp. 395–404, 1995.
[15] K. Nam and W. Lee, "Observers for nonautonomous discrete-time nonlinear systems," *J. KIEE*, vol. 5, pp. 32–38, 1992.
[16] K. Shulan, Z. Huanshui, and Z. Zhaosheng, "Observers design for discrete-time nonlinear systems," *Int. J. Pure Appl. Math.*, vol. 28, no. 1, pp. 1–11, 2006.
[17] Y. Song and J. W. Grizzle, "The extended Kalman filter as a local asymptotic observer for discrete-time nonlinear systems," *J. Math. Syst. Estim. Control*, vol. 5, no. 1, pp. 59–78, 1995.
[18] V. Sundarapandian, "Observer design for discrete-time nonlinear systems," *Math. Comput. Modelling*, vol. 35, no. 1–2, pp. 37–44, 2002.
[19] V. Sundarapandian, "New results on general observers for discrete-time nonlinear systems," *Appl. Math. Lett.*, vol. 17, no. 12, pp. 1415–1420, 2004.
[20] V. Sundarapandian, "General observers for discrete-time nonlinear systems," *Math. Comput. Modelling*, vol. 39, no. 1, pp. 87–95, 2004.
[21] M. Xiao, N. Kazantzis, C. Kravaris, and A. J. Krener, "Nonlinear discrete-time observer design with linearizable error dynamics," *IEEE Trans. Automat. Control*, vol. 48, no. 4, pp. 622–626, 2003.
[22] M. Xiao, "A direct method for the construction of nonlinear discrete-time observer with linearizable error dynamics," *IEEE Trans. Automat. Control*, vol. 51, no. 1, pp. 128–135, Jan. 2006.
[23] H.-G. Lee, A. Arapostathis, and S. I. Marcus, "Linearization of discrete-time systems," *Int. J. Control*, vol. 45, no. 5, pp. 1803–1822, 1987.
[24] A. Isidori, *Nonlinear Control Systems*, 3rd ed. New York: Springer-Verlag, 1995.
[25] R. Marino and P. Tomei, *Nonlinear Control Design*. New York: Prentice-Hall, 1995.
[26] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*. New York: Springer-Verlag, 1990.
[27] A. J. van der Schaft, "Observability and controllability for smooth nonlinear systems," *SIAM J. Control Optim.*, vol. 20, no. 3, pp. 338–354, 1982.

Stability of the Extended Kalman Filter When the States are Constrained

Esin Koksal Babacan, Levent Ozbek, and Murat Efe, *Member, IEEE*

***Abstract*—In this note, stability of the projection-based constrained discrete time extended Kalman filter (EKF) when applied to deterministic nonlinear systems has been studied. It is proved that, like the unconstrained case, under certain assumptions, the EKF with state equality constraints is an exponential observer, i.e., it keeps the dynamics of its estimation error exponentially stable. Also, it has been shown that a simple modification to the general definition of the EKF with exponential weighting increases the filter's degree of stability and convergence speed with or without state constraints.**

***Index Terms*—Asymptotic stability, Kalman filtering, nonlinear systems, state equality constraints.**

I. Introduction

Since its invention, Kalman filter and its derivations have been extensively used to address both linear and nonlinear state estimation problems [1]. It is known to be an optimal estimator for linear dynamic systems subject to white process and measurement noise. Kalman filter has also been utilized to address the estimation problem for both nonlinear stochastic [2] and nonlinear deterministic systems [3]. The most common way of estimating the states in nonlinear deterministic systems is firstly to design a dynamic state observer that comprises the model of the system and secondly feed the outputs in an appropriate manner [3], [4]. In [5] the extended Kalman filter (EKF) was proposed as an observer for nonlinear deterministic systems and it was proved, through the use of second method of Lyapunov [6], that the EKF was an exponential observer. Furthermore, in the same study a slightly more general definition of the standard extended Kalman filter, that is the EKF with exponential data weighting [7], [8], was applied to nonlinear systems and it was proved that the resulting observer had a predetermined degree of stability, described as the time constant for the error decline, which also affected the convergence of the extended Kalman filter.

In the past, researchers used to be reluctant to utilize constrained Kalman filtering, partly because constraints can be difficult to model and partly because of the increased computational burden (e.g., due to the additional information the error covariance matrix can get tighter). As a result, equality constraints used to be often neglected in standard Kalman Filtering applications. However, the benefits of incorporating constraints can outweigh the computational cost associated with constraining the estimate. Also, with cheap computational power and practical formulations to incorporate constraints in the filter equations readily available, there is increased interest in using constrained Kalman filtering. Thus any study on statistical properties of the resulting filter will be of utmost importance for researchers using this filter in their applications.

In this note, therefore, we deal with state estimation in deterministic nonlinear systems where constraints are imposed on the states and investigate the stability and convergence of the constrained EKF. The particular constrained Kalman filter studied in this study is the

Manuscript received November 13, 2007; revised February 02, 2008. Current version published December 10, 2008. Recommended by Associate Editor V. Krishnamurthy.

E. K. Babacan and L. Ozbek are with the Statistics Department, Faculty of Science, Ankara University, Ankara 06100, Turkey (e-mail: esin.koksal@.science.ankara.edu.tr). ozbek@.science.ankara.edu.tr).

M. Efe is with the Electronics Engineering Department, Faculty of Engineering, Ankara University, Ankara 06100, Turkey (e-mail: efe@eng.ankara.edu.tr).

Digital Object Identifier 10.1109/TAC.2008.2008333