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

---

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

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
Consider a discrete-time, time-invariant, controlled system of the form
\[ x(t + 1) = f(x(t), u(t)), \quad f(0, 0) = 0 \]
\[ y(t) = h(x(t)), \quad h(0) = 0 \] (1a)
or the autonomous system
\[ x(t + 1) = f(x(t)), \quad f(0) = 0 \]
\[ y(t) = h(x(t)), \quad h(0) = 0 \] (1b)
with state \( x \in \Sigma \cong \mathbb{R}^n \), input \( u \in \mathcal{U} \cong \mathbb{R}^m \), and output \( y \in \mathcal{Y} \cong \mathbb{R}^p \).

**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
\[ z(t + 1) = A(z(t)) + \gamma (y(t), u(t)) \] (2)
\[ y(t) = C(z(t)) \]
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 \( L \in \mathbb{R}^{p \times p} \) such that \((A - LC)\) is Hurwitz, we can design a state estimator
\[ \hat{z}(t + 1) = (A - LC)\hat{z}(t) + \gamma (y(t), u(t)) + Ly(t) \]
\[ \dot{x}(t) = T^{-1}(\hat{z}(t)) \]
which results in an asymptotically vanishing estimation error, i.e., \( \lim_{t \to \infty} \|x(t) - \hat{x}(t)\| \to 0 \).

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 \( G : \Sigma \times \mathbb{R}^r \to \Sigma \), we define the "impulse response" \( \hat{G}^i \) by
\[ \hat{G}^0(x, v) \triangleq x, \quad \hat{G}^i(x, v) \triangleq G\left(G^{-1}(x), v\right), \quad i \geq 2. \]

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

For a function \( F : \Sigma \times \mathcal{U} \times \mathbb{R}^r \to \Sigma \) with \( F(0, 0, 0) = 0 \), we define \( \Psi_k \equiv \Psi_k[F] : \mathbb{R}^{n \times b} \to \Sigma \), for \( k \in \mathbb{N} \), by \( \Psi_k(w) \triangleq F(0, 0, w) \), and
\[ \Psi_{k+1}(w^1, w^2, \ldots, w^{b+1}) \triangleq F\left(\Psi_k(w^2, w^3, \ldots, w^{b+1}), 0, w^1\right) \]
with \( w^t = (w^t_1, \ldots, w^t_m) \in \mathbb{R}^m \). We also adopt the convention \( \Psi_0 \equiv 0 \).

If \( F : \Sigma \times \mathbb{R}^r \to \Sigma \), 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 \( \nu \) be the least nonnegative integer such that \( d_{\nu}(h, \circ \hat{f}^k)(0, 0) \) is linearly dependent on the vectors in the collection
\[ \{d_{\nu}(h, \circ \hat{f}^k)(0, 0), d_{\nu}(h, \circ \hat{f}^{ni})(0, 0)\}, \quad 1 \leq j \leq p, 0 \leq k < \nu; 1 \leq t < i \} \]
The integers \( \{\nu_1, \ldots, \nu_p\} \) are the observability indices of (1b) [or (1a)]. Let \( \bar{p} \triangleq \max\{\nu_i\} \), and define the subspace \( \Delta \subset \mathbb{R}^{r \times \bar{p}} \), by
\[ \Delta \triangleq \left\{w^i, 1 \leq i \leq \bar{p}; 1 \leq j \leq p; 0 \leq k < \nu_j; 1 \leq t < i \right\}. \]

An element \( w \in \Delta \) is also viewed as an element of \( \mathbb{R}^{r \times \bar{p}} \) in the ordered coordinates \((w^1_1, \ldots, w^1_ar{p}, \ldots, w^n_1, \ldots, w^n_ar{p})\), with \( \kappa \triangleq \nu_1 + \cdots + \nu_p \). Thus if \( \bar{M} \in \mathbb{R}^{r \times \kappa} \), then \( \bar{M}w \in \bar{M} \mathbb{R}^{r \times \bar{p}} \) is well defined.

Let \( \tilde{\Psi} \) denote the restriction of \( \Psi_k \) on \( \Delta \) and define
\[ \mathcal{F} \equiv \mathcal{F}[F] : \mathbb{R}^r \times \Delta \times \mathcal{U} \to \Sigma \]
by
\[ \mathcal{F}(w^0, w^1, \ldots, w^p, u) \triangleq F\left(\tilde{\Psi}(w^1, \ldots, w^p), u, w^0\right). \]

If \( F : \Sigma \times \mathbb{R}^r \to \Sigma \), then \( \mathcal{F}[F] : \mathbb{R}^r \times \Delta \to \Sigma \) 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 \( \sum_{i=1}^p \nu_i = n \), or equivalently that
\[ \text{rank} \left\{d(h, \circ \hat{f}^k)(0), 1 \leq j \leq p, 0 \leq k < \nu_j\right\} = n. \]
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 \( 1 \leq j \leq p \)
\[ h_j \circ \hat{f}^i(x, v) = \begin{cases} h_j \circ f^{i}(x), & 1 \leq i < \nu_j \\ h_j \circ f^{\nu_j}(x) + w_j, & i = \nu_j. \end{cases} \]

Note that
\[ h_j \circ f^{i-1} \circ F(x, w) = \begin{cases} h_j \circ f^{i}(x), & 1 \leq i < \nu_j \\ h_j \circ f^{\nu_j}(x) + w_j, & i = \nu_j. \end{cases} \]

Iterating (5), we obtain
\[ h_j \circ f^{i-1} \circ \Psi(\mathbf{w}^1, \ldots, \mathbf{w}^p) = f_w^{j-\nu_j+1} + h_j \circ f^{\nu_j} \circ \Psi_{j-\nu_j+2}(\mathbf{w}^{j-\nu_j+2}, \ldots, \mathbf{w}^p). \]

We define the derived system of (1b) by
\[ x(t + 1) = F(x(t), u(t)), \quad y(t) = h(x(t)). \]

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

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

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

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

**Proof:** (Necessity) Suppose that there exists $z = T(x)$ such that $z(t+1) = A z(t), \ y(t) = C z(t)$.

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

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

and $C \equiv \text{blockdiag} \{ C_1, \ldots, C_p \}$, with $A_{ji} \in \mathbb{R}^{r \times r}$, given by

$$ A_{ii} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha_{i1j}^{j} \\ 0 & 1 & \cdots & 0 & \alpha_{i2j}^{j} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha_{innj} \\ \end{bmatrix} \quad (8b) $$

and

$$ A_{ij} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \alpha_{i1j}^{j} \\ 0 & 0 & \cdots & 0 & \alpha_{i2j}^{j} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \alpha_{innj} \\ \end{bmatrix} \quad i \neq j \quad (8c) $$

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

$$ h_j \circ f^i = \tilde{h}_j \circ \tilde{f}^i \circ T = C_j A^i T, \ 0 \leq i \leq \nu_j \quad (9) $$

where $C_j$ denotes the $j$th row of $C$. Let

$$ \mathcal{V}_j \triangleq \begin{bmatrix} dh_j(0) \\ d(h_j \circ f(0)) \\ \vdots \\ d(h_k \circ f^{k-1}(0)) \end{bmatrix}, \ \mathcal{V}_j \triangleq \begin{bmatrix} C_j \\ C_j A^{i} \\ \vdots \\ C_j A^{i-1} \end{bmatrix}, \ \mathcal{V} \triangleq \begin{bmatrix} \mathcal{V}_1 \\ \vdots \\ \mathcal{V}_p \end{bmatrix} $$

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

$$ C_j A^{i} \in \text{span} \{ C_i, A^{i-1}, 1 \leq i \leq p, 1 \leq \ell \leq \nu_j \}. \quad (10) $$

By (5) and (9)

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

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

$$ \hat{F}(z, w) = \hat{T} \circ F \left( T^{-1}(z), w \right) = A z + B w \quad (11) $$

with $B = \hat{T}^{-1} C^T = \text{blockdiag} \left( [1 \ 0 \ \cdots \ 0]^T \in \mathbb{R}^{r \times 1}, 1 \leq i \leq p \right)$. (12)

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

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

Thus

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

and condition iii) holds.

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

$$ d(h_j \circ f^{i-1}) (0) \frac{\partial \Psi_j}{\partial w_k} = \begin{cases} 1, & \ell = \nu_j - i + 1 \\ 0, & \ell = \nu_j - i + 1 \end{cases} \quad (13) $$

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

Next we show that $\xi(z) = h \circ \tilde{\Psi}(\xi) = C \xi$. By (6), for $\xi = (\xi_1, \ldots, \xi_p) \in \Delta$

$$ h_j \circ \Psi_j \xi_j = h_j \circ \tilde{f}^j \circ \Psi_j \tilde{\psi}_{j-\nu_j}(\xi_\nu_j^{j-1}, \ldots, \xi^j) + \xi_j^j. \quad (14) $$

We claim that $h_j \circ f^{j} \circ \Psi_j \tilde{\psi}_{j-\nu_j}(\xi_\nu_j^{j-1}, \ldots, \xi^j) = 0$. 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 $i \geq \nu_j$

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

For $k = 1, \ldots, \nu_j$, define

$$ \Gamma_k \triangleq \left\{ d(h_j \circ f^{k-1}) \circ \Psi_j (\xi^{\nu_j+1}, \ldots, \xi^j) \right\}, \ 1 \leq j \leq p, 1 \leq i \leq k \}.$$ 

If $k < \nu_j$, then by (5)

$$ h_j \circ f^{k-1} \circ \Psi_j (\xi^{\nu_j+1}, \ldots, \xi^j) = 0, \quad \begin{cases} \ell = \nu_j - k, & \nu_j < \nu_j \\ \ell = \nu_j, & \ell \geq \nu_j \end{cases} \quad (16) $$

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

$$ \Gamma_k \subseteq \text{span} \left\{ \Gamma_i, 1 \leq i \leq k \right\} \quad (16) $$

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

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

$$ Y_j^i = \mathcal{F}, \frac{\partial}{\partial w_j^i}, \ 1 \leq j \leq p, \ 1 \leq i \leq \nu_j $$

Then

$$ Y_j^i = \mathcal{F}, \frac{\partial}{\partial \xi_j^i}, \ 1 \leq j \leq p, \ 1 \leq i \leq \nu_j $$
which implies that \( \{Y_j^r | 1 \leq j \leq p, 1 \leq i \leq \nu_j \} \) is a set of linearly independent vector fields. Since, for \( 1 \leq j, k \leq p \)
\[
[Y_j^r, Y_k^r] = \mathcal{F}_r \left[ \frac{\partial}{\partial w_{j+k}^{r-1}}, \frac{\partial}{\partial w_k^{r-1}} \right] = 0
\]
for all \( i = 1, \ldots, \nu_j \) and \( j = 1, \ldots, \nu_k \), it follows that \( \{Y_j^r | 1 \leq j \leq p, 1 \leq i \leq \nu_j \} \) is a set of \( n \) commuting vector fields. Thus
\[
Y_j^{r+1} = \sum_{k=1}^n \sum_{i=1}^{\nu_k} \alpha_{j,k} Y_k^i
\]
for some \( \alpha_{j,k} \in \mathbb{R}, 1 \leq j, k \leq p, 1 \leq i \leq \nu_k \). Since
\[
\tilde{F}(\xi, w) = \tilde{\Psi}^{-1} \circ F(\tilde{\Psi}(\xi), w) = \tilde{\Psi}^{-1} \circ \mathcal{F}(v, \xi), \quad \xi \in \Delta,
\]
we obtain
\[
\tilde{F}, \frac{\partial}{\partial w_j} = (\tilde{\Psi}^{-1} \circ \mathcal{F}), \frac{\partial}{\partial w_j} = (\tilde{\Psi}^{-1}), Y_j^1 = \frac{\partial}{\partial \xi_j}.
\] (17)
Similarly, for \( 1 \leq i < \nu_j \)
\[
\tilde{F}, \frac{\partial}{\partial \xi_j^i} = (\tilde{\Psi}^{-1} \circ \mathcal{F}), \frac{\partial}{\partial \xi_j^i} = (\tilde{\Psi}^{-1}), Y_j^{i+1} = \frac{\partial}{\partial \xi_j^i}
\] (18)
and
\[
\tilde{F}, \frac{\partial}{\partial \xi_j^{r+1}} = (\tilde{\Psi}^{-1} \circ \mathcal{F}), \frac{\partial}{\partial \xi_j^{r+1}} = (\tilde{\Psi}^{-1}), Y_j^{r+1} = \frac{\partial}{\partial \xi_j^{r+1}}
\] (19)
By (17)–(19), it follows that \( \tilde{F}(\xi, w) = A \xi + B w \). By (5)
\[
\tilde{h}_j \circ \tilde{f}^{-1} \circ \tilde{F} (\xi, 0) = \tilde{h}_j \circ \tilde{f}^{i} (\xi), \quad 1 \leq j \leq p, \quad 1 \leq i \leq \nu_j.
\] (20)
Hence, (20) and assumption i) imply that for some local diffeomorphism \( G : \Sigma \rightarrow \Sigma \) whose components are the functions \( \{h_j \circ \tilde{f}^{-1} \}, \) 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 \( \nu_0 = n \), we derive the defined system of (1a) by
\[
x(t + 1) = F(x(t), u(t), w(t)), \quad y(t) = h(x(t)).
\]
The function \( F : \Sigma \times \mathbb{R} \times \mathbb{R}^m \rightarrow \Sigma \) is defined as
\[
h_j \circ \tilde{f}^{-1}(x, u, w) = \begin{cases} h_j \circ \tilde{f}(x, u), & 1 \leq i < \nu_j \\ h_j \circ \tilde{f}^r(x, u) + w_j, & i = \nu_j. \end{cases}
\] (21)
Existence of \( F \) is guaranteed by the inverse function theorem. It holds that
\[
h_j \circ \tilde{f}^{-1}(F(x, u, w), 0) = \begin{cases} h_j \circ \tilde{f}(x, u), & 1 \leq i < \nu_j \\ h_j \circ \tilde{f}^r(x, u) + w_j, & i = \nu_j. \end{cases}
\] (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) \( \sum_{i=1}^n \nu_i = n \).

ii) \( d_{\nu}(h_j \circ \tilde{f}^{-1}), 1 \leq i \leq p, 1 \leq \ell \leq \nu_j \), for \( 1 \leq j \leq p \).

iii) \( \mathcal{F}_r(\partial \nu \partial w_j^{r-1}), 1 \leq j \leq p, 0 \leq i < \nu_j \), are well-defined vector fields.

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

**Proof:** (Necessity): Suppose that there exists \( z = T(x) \) such that (2) holds. Thus
\[
\hat{f}(z) = T \circ f (T^{-1}(z), u) = A_0 z + \gamma (C z, u)
\]
\[
\hat{h}(z) = h \circ T^{-1}(z) = C z.
\]
We assume, without loss of generality, that \( A_0 \) and \( C \) take the form (8), with \( \alpha_{j,k} \equiv 0 \). We have
\[
h \circ \tilde{f}^i (x, u) = h \circ \tilde{f}^i (T(x), u), \quad i \geq 0.
\] (23)
Expand \( \tilde{f}^i \) as
\[
\tilde{f}^i (z, u) = A_0^i z + A_0^{i-1} \gamma (C z, u) + A_0^{i-2} \gamma \left(C \tilde{f}^i (z, u), 0\right) + \cdots
\]
\[
= A_0^i z + \sum_{i=1}^i A_0^{i-1} \gamma \left(C \tilde{f}^i (z, u), 0\right).
\] (24)
Then, using (24),
\[
\begin{bmatrix}
\frac{d\hat{h}}{dt} \\
\frac{d(h \circ \tilde{f}^i)}{dt} \\
\vdots
\end{bmatrix}
(0, 0) =
\begin{bmatrix}
I & 0 & \cdots & 0 \\
CL & I & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
CM_{n-2} & C M_{n-3} L & \cdots & I
\end{bmatrix}
\times
\begin{bmatrix}
C \\
C A_0 \\
\vdots \\
C A_0^{i-1}
\end{bmatrix}
\] (25)
where \( L \triangleq D_1 \gamma (0, 0) \), and \( M_i \triangleq A_0 + LC_i \). Necessity of i) then follows from (25). By (23) and (24)
\[
h_j \circ \tilde{f}^i (x, u, w, 0) = C_j \circ A_0^{i-1} T(x)
\]
\[
+ \sum_{k=1}^i C_j \circ A_0^{i-k} \gamma \left( C \tilde{f}^r (T(x), u), 0 \right).
\] (26)
Condition ii) then follows from (10) and (26).
By (22) and (23)
\[
\hat{h}_j \circ \tilde{f}^{i}(T \circ F(x, u, w), 0)
\]
\[
= \begin{cases} 
\hat{h}_j \circ \tilde{f}^{i-1}(T(x), u), & 0 \leq i < \nu_j - 1 \\
\hat{h}_j \circ \tilde{f}^{i}(T(x), u) + w_j, & i = \nu_j - 1. 
\end{cases}
\] (27)
Using (24) we can verify that
\[
T \circ F(x, u, w) = A_0 T(x) + \gamma (C T(x), u) + B w
\] (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
\[
\tilde{F}(x, u, w) = A_0 z + \gamma (C z, u) + B w, \quad \hat{X}
\]
\[
= T^{-1} (A_0^{i-1} B \xi^2 + \cdots + B \xi^1)
\]
\[
= \left( \gamma (w_1^{1}, \ldots, w_p^{1}, u) + A_0 B w^1 + \cdots + A_0 B w^{1} \right).
\]
Thus
\[ F_i \partial_{w_{ij}} = T_i^{-1}(A_i B_{i..j}), \quad 1 \leq j \leq p, \quad 0 \leq i < \nu_j \]
and condition iii) holds.

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

\[ Y_j = \hat{F}_i \partial_{w_{ij}} = \tilde{\Psi}_j \partial_{\tilde{\xi}_j}, \quad 1 \leq j \leq p, \quad 1 \leq i < \nu_j. \]

Then
\[ Y_j = \hat{F}_i \partial_{w_{ij}} = \tilde{\Psi}_j \partial_{\tilde{\xi}_j}, \quad 1 \leq j \leq p, \quad 1 \leq i < \nu_j. \]

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

\[ \hat{F}_i \partial_{w} = (\tilde{\Psi}^{-1} \circ F), \quad \hat{F}_i \partial_{w_{ij}} = (\tilde{\Psi}^{-1}), Y_j = \partial_{\tilde{\xi}_j}. \tag{29} \]

Similarly, for \( 1 \leq i < \nu_j \)

\[ \hat{F}_i \partial_{\tilde{\xi}_j} = (\tilde{\Psi}^{-1} \circ F), \quad \hat{F}_i \partial_{\tilde{\xi}_j} = (\tilde{\Psi}^{-1}), Y_j^{ij+1} = \partial_{\tilde{\xi}_j}. \tag{30} \]

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

\[ h \circ F(x, u, w) = \begin{pmatrix} x_2 - x_3^2 + x_1^2 \\ x_3 + x_1 u + u^2 \\ x_4 + x_2 u + x_1^2 u^2 \end{pmatrix}, \quad h(x) = x_1. \]

A straightforward calculation yields

\[ h \circ f(x, u) = x_2 - x_3^2 + x_1^2, \]
\[ h \circ f^2(x, u) = x_3 + x_1 u + (x_2 - x_3^2 + x_1^2)^2, \]
\[ h \circ f^3(x, u) = u + (x_3 + x_1 u + (x_2 - x_3^2 + x_1^2)^2)^2. \]

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

\[ h \circ F = F_1 = x_2 - x_3^2 + x_1^2, \]
\[ h \circ F^2 = F_2 - F_3^2 + F_3^2 = x_3 + x_1 u + (x_2 - x_3^2 + x_1^2)^2, \]
\[ h \circ F^3 = F_3 + (F_2 - F_3^2 + F_3^2)^2 = u + (x_3 + x_1 u + (x_2 - x_3^2 + x_1^2)^2)^2. \]

Solving (31), we obtain

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

Thus

\[ \tilde{\Psi}(w^0, w^1, w^2, w^3) = \begin{pmatrix} w_1 \\ w_2 + (w_1)^2 \\ w_1 \\ w_1 + w_2 u + (u + w_0)^2 \end{pmatrix}, \]
\[ F(w^0, w^1, w^2, w^3, u) = \begin{pmatrix} u \\ w_1 \\ w_1 + (w_0)^2 \end{pmatrix}. \]

and

\[ \ker F = \text{span} \left\{ u \partial_{w^1} - 2u^3 \partial_{w^2} + \partial_{w^3} - \partial_{w^1} + \partial_u \right\}. \]

It follows that \( \partial / \partial w^1 \subset \ker F \), which implies that condition iii) of Theorem 2 holds. Hence, using the transformation \( \tilde{\Psi}^{-1}(x) = \begin{pmatrix} x_3 \\ x_2 - x_3^2 \end{pmatrix} \), 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:**

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

We obtain

\[ h \circ f = x_1 + 2x_2^2 + x_1 u^2, \]
\[ h \circ f^2 = x_3 + x_1 u + (x_2 + x_1 u^2)^2. \]
Solving for $F$ and $\mathcal{F}$, we obtain

$$F(x, u, w) = \left( x_3^3 + u - (x_1 + 2x_2 + x_1 u^2)^2 + w \right) \frac{\partial}{\partial w} = \left( \left( 1 + (w_1)^2 \right) u + w_1 + w_2 + (w_1 + w_2 - w_1^2) \right)$$

$$\mathcal{F}(w_1, w_2, w, u) = \left( \begin{array}{c} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \frac{\partial}{\partial u} \end{array} \right) \left( \begin{array}{c} w_2 - w_1 + w_0 \\ w_1 + w_2 - w_1^2 \end{array} \right)$$

Thus,

$$\ker \mathcal{F} = \text{span} \left\{ 3(w_2)^2 \frac{\partial}{\partial w_2} - \frac{2(w_2)^2(1-w_2)}{1+w_2} \frac{\partial}{\partial w_1} + \frac{\partial}{\partial u}, \left(-1+4u \left( w_1 - (w_2)^2 \right) \right) \frac{\partial}{\partial w_1} + \frac{\partial}{\partial u} \right\}$$

and $[\partial / \partial w_1, \ker \mathcal{F}] \subset \ker \mathcal{F}$, which implies that condition iii) of Theorem 2 does not hold.

**Example 3:** Consider the system

$$f(x, u) = \left( x_2 + u \frac{\partial}{\partial w} \right), \quad h(x) = \left( x_1 + x_3 \right). \quad (32)$$

Thus

$$h_1 = x_1, \quad h_2 = x_1 + x_3, \quad h_2 = x_2 + x_1 + u^2$$

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

**Example 4:** Consider the system

$$f(x, u) = \left( x_2 - x_3^2 \right), \quad h(x) = \left( x_3 \right). \quad (33)$$

We have

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

and $\nu_1 = 1$, $\nu_2 = 2$. Thus, conditions i) and ii) of Theorem 2 are satisfied. If we let $F = (F_1, F_2, F_3)^T$, then by (4)

$$F_1 = x_3 + w_1, \quad F_1 = x_3 + x_2 + x_3 \frac{\partial}{\partial w_1} + x_2 + x_3 \frac{\partial}{\partial w_2} = x_2 + (1 + x_1^2) u + w_2$$

and solving (34) we obtain

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

Therefore

$$\Psi(w_1, w_2) = \left( \begin{array}{c} w_1 \\ w_2 - w_1 \end{array} \right), \quad \Psi^{-1}(x) = \left( \begin{array}{c} x_1 \\ x_2 - x_1 + x_3 \\ x_3 + x_1 \end{array} \right)$$

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

$$\gamma(y, u) = \left( \begin{array}{c} y_2 - y_1 \\ y_2 \end{array} \right).$$

**REFERENCES**


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