



HAL
open science

Luenberger observers for discrete-time nonlinear systems

Lucas Brivadis, Vincent Andrieu, Ulysse Serres

► **To cite this version:**

Lucas Brivadis, Vincent Andrieu, Ulysse Serres. Luenberger observers for discrete-time nonlinear systems. 2019 IEEE 58th Conference on Decision and Control (CDC), Dec 2019, Nice, France. pp.3435-3440, 10.1109/CDC40024.2019.9029220 . hal-02467958

HAL Id: hal-02467958

<https://hal.science/hal-02467958v1>

Submitted on 5 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Luenberger observers for discrete-time nonlinear systems

Lucas Brivadis, Vincent Andrieu and Ulysse Serres

The authors are with Univ. Lyon, Université Claude Bernard Lyon 1, CNRS, LAGEPP UMR 5007, 43 bd du 11 novembre 1918, F-69100 Villeurbanne, France (e-mail: lucas.brivadis@gmail.com, vincent.andrieu@gmail.com, ulysse.serres@gmail.com)

February 5, 2020

Abstract

In this paper, we consider the problem of designing an asymptotic observer for a nonlinear dynamical system in discrete-time following Luenberger's original idea. This approach is a two-step design procedure. In a first step, the problem is to estimate a function of the state. The state estimation is obtained by inverting this mapping. Similarly to the continuous-time context, we show that the first step is always possible provided a linear and stable discrete-time system fed by the output is introduced. Based on a weak observability assumption, it is shown that picking the dimension of the stable auxiliary system sufficiently large, the estimated function of the state is invertible. This approach is illustrated on linear systems with polynomial output. The link with the Luenberger observer obtained in the continuous-time case is also investigated.

1 Introduction

1.1 Context

The design of observers for nonlinear discrete-time systems remains a challenging and open problem despite a burgeoning literature. Since no universal method exists, several approaches have been developed. Most of them have first been developed for continuous-time systems, and then extended to the discrete case. Some of them, such as the well-known *extended Kalman filter* ([4, 11]), provide only a local convergence of the observer, and are based on a linearization of the system. Others (as [5] or [7]) consist in applying an invertible change of coordinates that transforms the original system in an other form for which it is much more easier to design an observer. Still others deal with Lipschitz nonlinear systems ([13, 12], among others), that occur frequently in practice, and are based on linear matrix inequalities that provide Lyapunov functions for the error system.

A completely different idea is to try to reproduce the Luenberger's initial methodology originally developed for linear continuous-time system in [9], which differs from what is now usually called *Luenberger observer*. This path has been mapped in the case of discrete-time systems by N. Kazantzis and C. Kravaris in [8]. It consists to estimate first a function of the state, thanks to a linear stable system fed by the output, and then to inverse this mapping. However, strong assumptions such as analyticity of the system and observability of the linearized system are required, and the invertibility of the function is obtained only locally.

In the following, we relax those assumptions following the strategy developed in the continuous case in [2] and later in [1] and [3]. We require the system to be time reversible, and

replace the observability hypothesis of the linearized system by a backward distinguishability hypothesis on the nonlinear system itself. In so doing, we obtain the existence and the injectivity (not only locally) of such a function of the state.

This paper is organized as follows. In the next part of the introduction (Section 1.2), we state our problem in a more precise way and introduce some notations and definitions. We also prove a first result that guarantees the existence of an observer as soon as there exists a continuous uniformly injective map satisfying some functional equation. Our main results can be found in Section 2. We state sufficient conditions for the existence, injectivity and also unicity of such a map. We provide in Section 3 some examples and applications of those results. We examine linear systems with polynomial output and also discrete-time systems that approximate continuous-time systems.

Throughout the paper, we denote by $|\cdot|$ the usual Euclidean norm and by $\|\cdot\|$ the induced matrix norm.

1.2 Problem statement

We consider the discrete-time system

$$x_{k+1} = f(x_k), \quad y_k = h(x_k), \quad (1)$$

with state $x \in \mathbb{R}^n$, output $y \in \mathbb{R}^p$ and suitable functions f and h . In this paper, we deal with the problem of existence of an observer for system (1). We denote $X_k(x_0) = f^k(x_0)$ the value at time k of the unique solution of system (1) initialized at $x_0 \in \mathbb{R}^n$, and $Y_k(x_0) = h(X_k(x_0))$ the corresponding output. Let $\mathcal{X}_0 \subset \mathcal{X} \subset \mathbb{R}^n$ such that for all initial condition $x_0 \in \mathcal{X}_0$ and all $k \in \mathbb{N} \cup \{0\}$, $X_k(x_0) \in \mathcal{X}$.

Definition 1. *Let m be a positive integer, $\varphi : \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}^n$ and $\psi : \mathbb{R}^m \rightarrow \mathbb{R}^n$. The discrete-time dynamical system given by*

$$\xi_{k+1} = \varphi(\xi_k, y_k), \quad \hat{x}_k = \psi(\xi_k), \quad (2)$$

is called an observer for (1) if and only if, for all $(x_0, \xi_0) \in \mathcal{X}_0 \times \mathbb{R}^m$, the solution of the coupled system (1)-(2), denoted by $(X_k(x_0), \hat{X}_k(x_0, \xi_0))_{k \geq 0}$, satisfies

$$\lim_{k \rightarrow +\infty} |X_k(x_0) - \hat{X}_k(x_0, \xi_0)| = 0. \quad (3)$$

Note that, even if \hat{X}_k seems to depend directly of x_0 , it is actually not the case. As (2) says, \hat{X}_k depends only of the measurements $Y_0(x_0), Y_1(x_0), \dots, Y_{k-1}(x_0)$ through the dynamic of $(\xi_k)_{k \geq 0}$.

We follow the Luenberger-like methodology in order to design an observer for system (1). Let m be a positive integer. First, we try to transform (1) into

$$\xi_{k+1} = A\xi_k + By_k. \quad (4)$$

with $A \in \mathbb{R}^{m \times m}$ a matrix with spectral radius $\rho(A) < 1$ and $B \in \mathbb{R}^{m \times p}$. In order to do this, we look for a continuous map $T : \mathcal{X} \rightarrow \mathbb{R}^m$ such that, for any $x_0 \in \mathcal{X}_0$ and any $k \in \mathbb{N} \cup \{0\}$,

$$T(X_{k+1}(x_0)) = AT(X_k(x_0)) + BY_k(x_0). \quad (5)$$

Let $\Xi_k(x_0, \xi_0)$ denote the value at time k of the unique solution of system (4) with initial condition $\xi_0 \in \mathbb{R}^m$ and measurements $y_k = Y_k(x_0)$. Note that, for any $(x_0, \xi_0) \in \mathcal{X}_0 \times \mathbb{R}^m$,

$$\Xi_{k+1}(x_0, \xi_0) - T(X_{k+1}(x_0)) = A(\Xi_k(x_0, \xi_0) - T(X_k(x_0))) \quad (6)$$

and since $\rho(A) < 1$, $\Xi_k(x_0, \xi_0) - T(X_k(x_0))$ converges geometrically towards zero. Hence, implementing system (4), one can deduce an approximation of $T(x_k)$ as k goes to infinity. Then, if T is injective, one can estimate the state of system (1). More precisely, we have the following theorem.

Theorem 1. *Let m be a positive integer, $A \in \mathbb{R}^{m \times m}$ such that $\rho(A) < 1$ and $B \in \mathbb{R}^{m \times p}$. Let $T : \mathcal{X} \rightarrow \mathbb{R}^m$ be a continuous map. Assume the following:*

1. *For all $x \in \mathcal{X}$, T satisfies*

$$T(f(x)) = AT(x) + Bh(x). \quad (7)$$

2. *T is uniformly injective, that is, there exists α a class \mathcal{K}^∞ function such that for all $(x_1, x_2) \in \mathcal{X}^2$,*

$$|x_1 - x_2| \leq \alpha(|T(x_1) - T(x_2)|). \quad (8)$$

Then there exists a map $T^ : \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that $(\hat{X}_k)_{k \geq 0}$ defined by $\hat{X}_k(x_0, \xi_0) = T^*(\Xi_k(x_0, \xi_0))$ for all $(x_0, \xi_0) \in \mathcal{X}_0 \times \mathbb{R}^m$ is the solution of an observer for (1).*

Proof. Clearly, (7) implies that (5) is satisfied for all $x_0 \in \mathcal{X}_0$ and all $k \in \mathbb{N} \cup \{0\}$. Let $(x_0, \xi_0) \in \mathcal{X}_0 \times \mathbb{R}^m$. Since $\rho(A) < 1$, it follows from (6) that

$$\lim_{k \rightarrow +\infty} \Xi_k(x_0, \xi_0) - T(X_k(\xi_0)) = 0. \quad (9)$$

From the uniform injectivity of T , there exists a pseudo-inverse $T^{-1} : T(\mathcal{X}) \rightarrow \mathbb{R}^n$ such that for all x in \mathcal{X} $T^{-1}(T(x)) = x$ and for all $(\xi_1, \xi_2) \in T(\mathcal{X})^2$,

$$|T^{-1}(\xi_1) - T^{-1}(\xi_2)| \leq \alpha(|\xi_1 - \xi_2|). \quad (10)$$

According to [10, Theorem 2], there exists a function $T^* : \mathbb{R}^m \rightarrow \mathbb{R}^n$, that is an extension to \mathbb{R}^m of T^{-1} , satisfying (10) for all $(\xi_1, \xi_2) \in (\mathbb{R}^m)^2$. Hence,

$$|T^*(\xi) - x| \leq \alpha(|\xi - T(x)|) \quad \forall \xi \in \mathbb{R}^m, \forall x \in \mathcal{X}. \quad (11)$$

Thus $|T^*(\Xi_k(x_0, \xi_0)) - X_k(\xi_0)| \rightarrow 0$ as k goes to infinity. Setting $\varphi : (\xi, y) \in \mathbb{R}^n \times \mathbb{R}^p \mapsto A\xi + By$ and $\psi = T^*$, it follows from the Definition 1 that $(\hat{X}_k)_{k \geq 0}$ defined by $\hat{X}_k(x_0, \xi_0) = T^*(\Xi_k(x_0, \xi_0))$ is the solution of an observer for (1). \square

Then it is sufficient to prove the existence of a uniformly injective continuous map $T : \mathcal{X} \mapsto \mathbb{R}^m$ satisfying (7) for some positive integer m in order to design an observer for (1). In the next section, we state sufficient conditions for the existence, injectivity, and also unicity of a continuous map T solution of (7).

Remark 1. *Note that if \mathcal{X} is a compact subset of \mathbb{R}^n , then every continuous injective map $T : \mathcal{X} \rightarrow \mathbb{R}^m$ is also uniformly injective in the sense of (8). In the following, we are interested in the injectivity of T . If uniform injectivity is required (for example to apply Theorem 1), then one must either assume \mathcal{X} compact or prove the uniform injectivity by other means.*

2 Results and comments

2.1 Existence of the transformation

First, we are interested in the existence of a map T satisfying (7). In [2], V. Andrieu and L. Praly have proved the existence of a so-called Kazantzis–Kravaris/Luenberger observer for continuous-time systems of the form

$$\dot{x} = f(x), \quad y = h(x). \quad (12)$$

We follow the same methodology and adapt it in the discrete case. We need to make some assumptions on the system.

Assumption 1. f is invertible and f^{-1} and h are continuous.

Assumption 2. There exist four non-negative constants C_1, C_2, C'_1 and C'_2 such that, for all $x \in \mathbb{R}^n$,

$$|x| \leq C_1 + C_2|f(x)|, \quad |h(x)| \leq C'_1 + C'_2|x|. \quad (13)$$

Remark 2. Note that Assumptions 1 and 2 are satisfied in particular if f is invertible and both f^{-1} and h are globally Lipschitz. We will use this remark in the next section about the injectivity of T .

For all non-negative integer i , we denote \circ the composition operator and

$$f^i = \underbrace{f \circ f \circ \dots \circ f}_{i \text{ times}}, \quad f^{-i} = (f^{-1})^i.$$

Theorem 2. Let m be a positive integer, $A \in \mathbb{R}^{m \times m}$ a normal matrix such that $\rho(A) < \min\{1, 1/C_2\}$ and $B \in \mathbb{R}^{m \times p}$. Assume that Assumptions 1 and 2 are satisfied. For all $x \in \mathcal{X}$, set

$$T(x) = \sum_{i=0}^{+\infty} A^i B h(f^{-(i+1)}(x)). \quad (14)$$

Then $T : \mathcal{X} \rightarrow \mathbb{R}^m$ is well defined, continuous, and satisfies (7).

Proof. For all $x \in \mathcal{X}$ and all non-negative integer i , let $a_i(x) = A^i B h(f^{-(i+1)}(x))$. According to Assumption 1, each a_i is continuous on \mathcal{X} . Note that, since A is normal, $\rho(A) = \|A\|$. Then, according to Assumption 2, we have for all $x \in \mathcal{X}$

$$|a_i(x)| \leq \rho(A)^i \|B\| \left(C'_1 + C'_2 \left(C_2^{i+1} |x| + C_1 \sum_{j=0}^i C_2^j \right) \right). \quad (15)$$

Since $\rho(A) < 1$ and $\rho(A)C_2 < 1$ the Lebesgue dominated convergence theorem applied on any compact set implies that (14) defines a continuous function. Moreover, for any $x \in \mathcal{X}$,

$$\begin{aligned} T(f(x)) &= \sum_{i=0}^{+\infty} A^i B h(f^{-(i+1)}(f(x))) \\ &= A \sum_{i=0}^{+\infty} A^{i-1} B h(f^{-i}(x)) \\ &= A \sum_{i=0}^{+\infty} A^i B h(f^{-(i+1)}(x)) + B h(x) \\ &= AT(x) + B h(x), \end{aligned}$$

which shows that T satisfies (7). □

2.2 Injectivity with backward distinguishability

In order to obtain that T defined by (14) is injective, we introduce the following *backward distinguishability* assumption on the system.

Assumption 3. *For all $(x_1, x_2) \in \mathcal{X}^2$, if $x_1 \neq x_2$, then there exists a positive integer i such that $h(f^{-i}(x_1)) \neq h(f^{-i}(x_2))$.*

We also need stronger hypothesis on the system than in the previous section.

Assumption 4. *f is invertible and f^{-1} and h are of class C^1 and globally Lipschitz.*

According to the Remark 2, if Assumption 4 holds, then Assumptions 1 and 2 are satisfied. We denote by I_k the identity $k \times k$ matrix, by \otimes the Kronecker product and by A^* the conjugate transpose matrix of A .

Theorem 3. *Let Assumptions 3 and 4 hold. Let $m = (n + 1)p$ and $B = (1, \dots, 1)^* \otimes I_p \in \mathbb{C}^{m \times p}$. Let $C_2 = \sup\{|(f^{-1})'(x)|, x \in \mathcal{X}\}$ and \mathcal{D} be the open disc of \mathbb{C} of radius $\min\{1, 1/C_2\}$. Then there exists a subset $\mathcal{R} \subset \mathcal{D}^{n+1}$ of zero Lebesgue measure in \mathbb{C}^{n+1} such that, for any $(\lambda_1, \dots, \lambda_{n+1}) \in \mathcal{D}^{n+1} \setminus \mathcal{R}$, the matrix $A = \text{diag}(\lambda_1, \dots, \lambda_{n+1}) \otimes I_p \in \mathbb{C}^{m \times m}$ is such that the map $T : \mathcal{X} \rightarrow \mathbb{C}^m$ defined by (14) is well-defined, of class C^1 and one-to-one.*

Proof. Let $(\lambda_1, \dots, \lambda_{n+1}) \in \mathcal{D}^{n+1}$ and $A = \text{diag}(\lambda_1, \dots, \lambda_{n+1}) \otimes I_p \in \mathbb{C}^{m \times m}$. Let $T : \mathcal{X} \rightarrow \mathbb{C}^m$ be defined as in (14). For all $\lambda \in \mathcal{D}$, let

$$T_\lambda(x) = \sum_{i=0}^{+\infty} \lambda^i h(f^{-(i+1)}(x)), \quad \forall x \in \mathcal{X}. \quad (16)$$

Let $a_i(x) = \lambda^i h(f^{-(i+1)}(x))$ for all $x \in \mathcal{X}$. Then each a_i is of class C^1 on \mathcal{X} by Assumption 4, and we have the following domination:

$$|a'_i(x)| \leq \lambda^i C'_2 C_2^{i+1}$$

with $C'_2 = \sup\{|h'(x)|, x \in \mathcal{X}\}$. Moreover, $\lambda C_2 < 1$. So the Lebesgue dominated convergence theorem implies that for each $\lambda \in \mathcal{D}$, $T_\lambda : \mathcal{X} \rightarrow \mathbb{C}^p$ is well-defined and of class C^1 . Considering the structure of A and B , remark that up to a permutation of coordinates we have

$$T(x) = (T_{\lambda_1}(x), \dots, T_{\lambda_{n+1}}(x))^*$$

It is sufficient to prove that $T : \mathcal{X} \rightarrow \mathbb{C}^p$ is one-to-one for almost all $(\lambda_1, \dots, \lambda_{n+1}) \in \mathcal{D}^{n+1}$.

In order to do this, we need the following lemma, established by L. Praly and V. Andrieu in [2, Lemma 1], which is a modified version of [6, Lemma 3.2] due to J.-M. Coron.

Lemma 1. *Let \mathcal{D} and Γ be open subsets of \mathbb{C} and \mathbb{R}^{2n} , respectively. Let $g : \Gamma \times \mathcal{D} \rightarrow \mathbb{C}^p$ be a function which is holomorphic in λ for each $\underline{x} \in \Gamma$ and C^1 in \underline{x} for each $\lambda \in \mathcal{D}$. If for each $\underline{x} \in \Gamma$, the function $\lambda \in \mathcal{D} \mapsto g(\underline{x}, \lambda)$ is not constantly zero, then the set*

$$\mathcal{R} = \bigcup_{\underline{x} \in \Gamma} \left\{ (\lambda_1, \dots, \lambda_{n+1}) \in \mathcal{D}^{n+1} \mid \forall i \in \{1, \dots, n+1\}, g(\underline{x}, \lambda_i) = 0 \right\} \quad (17)$$

has zero Lebesgue measure in \mathbb{C}^{n+1} .

We apply this lemma to $\Gamma = \{(x_1, x_2) \in \mathcal{X}^2 \mid x_1 \neq x_2\}$ and $g = \Delta T$ defined as follows:

$$\Delta T : (x_1, x_2, \lambda) \in \mathcal{X}^2 \times \mathcal{D} \mapsto T_\lambda(x_1) - T_\lambda(x_2) \quad (18)$$

Clearly, $\Delta T(x_1, x_2, \cdot)$ is holomorphic on \mathcal{D} for each $(x_1, x_2) \in \mathcal{X}^2$ and $\Delta T(\cdot, \lambda)$ is of class C^1 on \mathcal{X}^2 for each $\lambda \in \mathcal{D}$. Fix $(x_1, x_2) \in \Gamma$. Now, we prove that $\Delta T(x_1, x_2, \cdot)$ is not identically zero on \mathcal{D} . Assume the contrary. By unicity of the power series expansion, we get that for all positive integer i ,

$$h(f^{-i}(x_1)) = h(f^{-i}(x_2)) \quad (19)$$

According to the *backward distinguishability* Assumption 3, it implies that $x_1 = x_2$ which is contradictory with the fact that $(x_1, x_2) \in \Gamma$. Hence, $\Delta T(x_1, x_2, \cdot)$ is not identically zero on \mathcal{D} .

Since \mathcal{D} is a convex subset of \mathbb{C} and $\Delta T(x_1, x_2, \cdot)$ is holomorphic, its zero are isolated and with finite multiplicity. Hence the hypotheses of Lemma 1 are satisfied. Thus, $\mathcal{R} \subset \mathcal{D}^{n+1}$ has zero Lebesgue measure and for all $(\lambda_1, \dots, \lambda_{n+1}) \in \mathcal{D}^{n+1} \setminus \mathcal{R}$, T is injective by definition of ΔT . \square

Remark 3. *The function T and the matrices A and B defined Theorem 3 take complex values while previous Theorems 1 and 2 remain in the real frame. However, one can choose two different ways to bridge this gap.*

- *State Theorems 1 and 2 in the complex frame. The proofs remain identical. One should simply change the domains and codomains of f and h .*
- *Instead of considering $A = \text{diag}(\lambda_1, \dots, \lambda_{n+1}) \otimes I_p \in \mathbb{C}^{m \times m}$ and $B = (1, \dots, 1)^* \otimes I_p \in \mathbb{C}^{m \times p}$, one should either consider $\tilde{A} = \text{diag}(\Lambda_1, \dots, \Lambda_{n+1}) \otimes I_p \in \mathbb{R}^{2m \times 2m}$ and $\tilde{B} = (\mathbb{I}, \dots, \mathbb{I})^* \otimes I_p \in \mathbb{R}^{2m \times p}$, where*

$$\Lambda_i = \begin{pmatrix} \Re(\lambda_i) & -\Im(\lambda_i) \\ \Im(\lambda_i) & \Re(\lambda_i) \end{pmatrix}, \quad \mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then for all real sequence of measurements $(y_k)_{k \geq 0}$ the solutions of $\tilde{\xi}_{k+1} = \tilde{A}\tilde{\xi}_k + \tilde{B}y_k$ contain the real and imaginary parts of the solutions of $\xi_{k+1} = A\xi_k + By_k$.

2.3 Unicity

One can also wonder in which cases does the unicity of T satisfying (7) holds. More than a theoretical question, this fact may be useful in practice in order to obtain the injectivity of T . Most of the time, the function T given by (14) is difficult to compute. Since the matrix A has spectral radius strictly inferior to 1, an approximation of T is given by

$$T_N(x) = \sum_{i=0}^N A^i B h(f^{-(i+1)}(x)), \quad \forall x \in \mathcal{X}. \quad (20)$$

for all $N \geq 0$. Then $|T(x) - T_N(x)| \rightarrow 0$ as $N \rightarrow +\infty$. However, if f and h have more properties (for example if f is linear and h is polynomial, see Section 3.1), there may exist another solution \tilde{T} of (7) much more easier to compute than T . Then, the question of the injectivity of that new \tilde{T} remains open *a priori*. But if (7) has a unique solution for A and B complex matrices chosen as in Theorem 3, then $T = \tilde{T}$ and hence \tilde{T} is injective. Now, we state our unicity theorem.

Theorem 4. *Let m be a positive integer, $A \in \mathbb{R}^{m \times m}$ such that $\rho(A) < 1$ and $B \in \mathbb{R}^{m \times p}$. Let Assumption 1 hold and make the following backward stability hypothesis on \mathcal{X} :*

$$\forall x \in \mathcal{X}, \forall i \geq 1, \quad f^{-i}(x) \in \mathcal{X}. \quad (21)$$

Assume also that \mathcal{X} is compact. Then there exists one and only one continuous function $T : \mathcal{X} \rightarrow \mathbb{R}^m$ that satisfy (7) for all $x \in \mathcal{X}$.

Proof. First, we prove that the continuous solution of (7) is unique. Let $T_1, T_2 : \mathcal{X} \rightarrow \mathbb{R}^m$ be two continuous solutions of (7). Let $x \in \mathcal{X}$. Then for all $i \in \mathbb{N} \cup \{0\}$,

$$\begin{aligned} T_1(x) - T_2(x) &= (T_1 - T_2)(f^i(f^{-i}(x))) \\ &= A^i(T_1 - T_2)(f^{-i}(x)). \end{aligned} \quad (\text{from (7)})$$

Since \mathcal{X} is compact, satisfy (21) and T_1 and T_2 are continuous, there exists a constant $K > 0$ such that $|(T_1 - T_2)(f^{-i}(x))| \leq K$ for all $i \in \mathbb{N} \cup \{0\}$. Since moreover $\rho(A) < 1$, $A^i(T_1 - T_2)(f^{-i}(x)) \rightarrow 0$ as $i \rightarrow +\infty$. Thus $T_1(x) - T_2(x) = 0$.

The existence of a continuous T satisfying (7) follows from the Theorem 2 and from the fact that Assumption 2 can be replaced in its proof by the fact that \mathcal{X} is compact and backward stable¹. Indeed, the series (14) still defines a continuous function since the domination

$$|a_i(x)| \leq \rho(A)^i \|B\| \sup_{\tilde{x} \in \mathcal{X}} h(\tilde{x}) \quad (22)$$

holds for all $x \in \mathcal{X}$ and can replace (15). Then one may apply the Lebesgue dominated convergence on \mathcal{X} . \square

To conclude this section, recall that we have now at our disposal three theorems that ensures under different conditions on (1) the existence, unicity and injectivity of a continuous map T satisfying (7). In the next section, we illustrate on examples how to use those tools. In particular, we study systems with linear dynamics and polynomial output, and emphasize the link between the Luenberger observers developed in [2] for continuous-time systems and the discrete-time observers developed in this paper for their first-order approximations.

3 Examples

3.1 Linear dynamics with polynomial output

We consider first the system with linear dynamic and polynomial output of degree d

$$x_{k+1} = Fx_k, \quad y_k = HP_d(x) \quad (23)$$

with $P_d : \mathbb{R}^n \rightarrow \mathbb{R}^{k_d}$ a vector containing the k_d possible monomials with degree less or equal than d , $F \in \mathbb{R}^{n \times n}$ and $H \in \mathbb{R}^{p \times k_d}$. Then we have the following proposition.

Proposition 1. *Let m be a positive integer and $B \in \mathbb{R}^{m \times p}$. There exists a subset \mathcal{S} of zero Lebesgue measure in $\mathbb{R}^{m \times m}$ such that for all $A \in \mathbb{R}^{m \times m} \setminus \mathcal{S}$, there exists a function $T : \mathbb{R}^n \mapsto \mathbb{R}^m$ of the form*

$$T(x) = MP_d(x), \quad \forall x \in \mathbb{R}^n \quad (24)$$

for some $M \in \mathbb{R}^{m \times k_d}$, that satisfies (7) for any $x \in \mathbb{R}^n$.

¹ Similarly, using the same trick, one can easily show that the hypothesis of *globally Lipschitz* in Assumption 4 can be replaced in the proof of Theorem 3 by the fact that \mathcal{X} is compact and backward stable.

Proof. First, note that since $P_d(Fx)$ is a vector containing polynomials of x with degree inferior to d , there exists a matrix $D \in \mathbb{R}^{k_d \times k_d}$ such that

$$P_d(Fx) = DP_d(x), \quad \forall x \in \mathbb{R}^n. \quad (25)$$

Since the set of eigenvalues of D is finite, the spectra of D and $-A$ are disjoint for almost all $A \in \mathbb{R}^{m \times m}$ *i.e.* there exists a subset $\mathcal{S} \subset \mathbb{R}^{m \times m}$ of zero Lebesgue measure such that the spectra of D and $-A$ are disjoint for all $A \in \mathbb{R}^{m \times m} \setminus \mathcal{S}$. For such matrices A the Sylvester equation

$$MD = AM + BH \quad (26)$$

has a unique solution $M \in \mathbb{R}^{m \times k_d}$. Set T as in (24). It remains to check that (7) is satisfied for $f = F$ and $h = HP_d$. For all $x \in \mathbb{R}^n$,

$$\begin{aligned} T(Fx) &= MP_d(Fx) && \text{(from (24))} \\ &= MDP_d(x) && \text{(from (25))} \\ &= AMP_d(x) + BHP_d(x) && \text{(from (26))} \\ &= AT(x) + BHP_d(x). \end{aligned}$$

□

Remark 4. Note that the result is still true if A and B are complex matrices. Then T takes complex values. The proof remains identical.

Remark 5. Choose a set $\mathcal{X}_0 \subset \mathbb{R}^n$ of initial condition and let \mathcal{X} be as usual such that $X_k(x_0) \in \mathcal{X}$ for all $x_0 \in \mathcal{X}_0$ and all $k \in \mathbb{N} \cup \{0\}$. Note that if F is invertible and if \mathcal{X} is compact and backward stable, then the assumptions of the Theorem 4 hold. Assume also that Assumptions 3 and 4 hold and apply Theorem 3 with $m = (n+1)p$. Then, for almost all $(\lambda_1, \dots, \lambda_{n+1}) \in \mathbb{C}^{n+1}$, and for complex matrices A and B as in Theorem 3, we have

$$T(x) = MP_d(x) = \sum_{i=0}^{+\infty} A^i B h(f^{-(i+1)}(x)) \quad (27)$$

for all $x \in \mathcal{X}$. In particular, T defined by (24) is injective.

3.2 Link with the continuous Luenberger observer

In this section, we are interested in the link between the continuous Luenberger observer developed in [2] for system (12) and the discrete observer developed in the previous sections for a discrete-time version of (12).

3.2.1 Continuous-time system

We consider the following example with linear dynamic and polynomial output:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1 \end{cases}, \quad y = x_1^2 - x_2^2 + x_1 + x_2. \quad (28)$$

It can be shown that this system is weakly differentially observable² of order 4 on \mathbb{R}^2 in the sense of [3, Definition 1]. Following [3], we seek $T_\lambda : \mathbb{R} \rightarrow \mathbb{R}^n$ such that

$$\frac{d}{dt} T_\lambda(x) = \lambda T_\lambda(x) + y \quad (29)$$

² First, the map $(x_1^2 - x_2^2, x_1 + x_2) \mapsto (y, \dot{y})$ is injective. Similarly, $(x_1 x_2, x_1 - x_2) \mapsto (\dot{y}, \ddot{y})$ is also injective. Combining those results, we get that $(x_1, x_2) \mapsto (y, \dot{y}, \ddot{y}, \ddot{\dot{y}})$ is injective.

for some $\lambda < 0$. Since (28) has linear dynamic and polynomial output of degree 2, one can look for T of the form

$$T_\lambda(x) = x^* \begin{pmatrix} a & c/2 \\ c/2 & b \end{pmatrix} x + (d \ e) x \quad (30)$$

for some $(a, b, c, d, e) \in \mathbb{R}^5$. Then (29) holds if and only if

$$\begin{aligned} -c &= \lambda a + 1, & c &= \lambda b - 1, & 2(a - b) &= \lambda c, \\ -e &= \lambda d + 1, & d &= \lambda e + 1. \end{aligned} \quad (31)$$

The only solution of this equation is

$$\begin{aligned} a &= -\frac{\lambda}{4 + \lambda^2}, & b &= \frac{\lambda}{4 + \lambda^2}, & c &= -\frac{4}{4 + \lambda^2}, \\ d &= \frac{1 - \lambda}{1 + \lambda^2}, & e &= -\frac{1 + \lambda}{1 + \lambda^2}. \end{aligned} \quad (32)$$

Since T_λ is stationary, one could believe that this function provide an observer that could be efficient even for a numerical approximation of (28). However, as we will see in the following, it is not the case: for a given discrete approximation of (28), it is better to design an observer based on the discrete-time system rather than to use the one given by T_λ .

3.2.2 Associated first-order discrete-time system

For some discretization parameter $dt > 0$, the associated first-order approximation³ of (28) is

$$\begin{cases} x_1(k+1) = x_1(k) + dt x_2(k) \\ x_2(k+1) = x_2(k) - dt x_1(k) \\ y_k = x_1(k)^2 - x_2(k)^2 + x_1(k) + x_2(k) \end{cases} . \quad (33)$$

We seek a function $T_\lambda^d : \mathbb{R} \rightarrow \mathbb{R}^n$ satisfying a first-order approximation of (29) given by the Euler explicit method:

$$T_\lambda^d(x(k+1)) = (1 + \lambda dt) T_\lambda^d(x(k)) + dt y_k. \quad (34)$$

Since $\lambda < 0$, it is sufficient to choose $\lambda dt > -2$ to have $-1 < 1 + \lambda dt < 1$. Now, we seek T_λ^d of the form

$$T_\lambda^d(x) = x^* \begin{pmatrix} a' & c'/2 \\ c'/2 & b' \end{pmatrix} x + (d' \ e') x \quad (35)$$

for some $(a', b', c', d', e') \in \mathbb{R}^5$. Then (34) holds if and only if (d', e') satisfy the same equation that (d, e) in (31) and (a', b', c') satisfy

$$\begin{cases} -c' + b' dt = \lambda a' + 1, \\ c' + a' dt = \lambda b' - 1, \\ 2(a' - b') - c' dt = \lambda c'. \end{cases} \quad (36)$$

Remark that this equation is the same than (32) when $dt = 0$. This is coherent with the fact that (33) is a discretization of (28). Then, the only solution of (36) is such that $(d', e') = (d, e)$

³ Since (28) is weakly differentially observable, it can be shown that (33) is backward distinguishable as soon as dt is small enough.

for all $dt > 0$ and (a', b', c') converges to (a, b, c) as dt goes to 0:

$$\begin{cases} a' = -\frac{\lambda + dt}{4 + (\lambda + dt)^2}, \\ b' = \frac{\lambda + dt}{4 + (\lambda + dt)^2}, \\ c' = -\frac{4}{4 + (\lambda + dt)^2}. \end{cases} \quad (37)$$

For $dt > 0$, the discrete observer given by T_{λ}^d is therefore different from the continuous observer given by T_{λ} , even if their difference goes to 0 as dt goes to 0.

3.2.3 Comparison of the observers

Consider a numerical simulation of the continuous-time system (28) obtained by the Euler explicit first-order method, which corresponds to the discrete-time system (33). Then the map T_{λ}^d given by (34) is much more adapted to the design of a numerically efficient observer than the function T_{λ} given by (29) that has been designed for (28). More generally, in order to implement an observer for a continuous-time varying system, it is better to develop a discrete-time observer based on the numerical approximation of the system, rather than a continuous-time observer based on the original system itself.

In order to highlight numerically this fact, we simulate the system (28) thanks to (33) and compare the accuracy of two observers: one based on functions of the form T_{λ}^d , and another based on functions of the form T_{λ} . To obtain the observers, we fix $dt > 0$ and three arbitrary values $\lambda_i < 0$ satisfying $\lambda_i dt > -2$ and use the fact that

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ a_1 & c_1 & d_1 & e_1 \\ a_2 & c_2 & d_2 & e_2 \\ a_3 & c_3 & d_3 & e_3 \end{pmatrix} \begin{pmatrix} x_1^2 - x_2^2 \\ x_1 x_2 \\ x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} y \\ T_1(x) \\ T_2(x) \\ T_3(x) \end{pmatrix}. \quad (38)$$

where (a_i, c_i, d_i, e_i) is given by (32) (resp. (37)) with $\lambda = \lambda_i$ and $T_i = T_{\lambda_i}$ (resp. $T_i = T_{\lambda_i}^d$). Fix the following parameters and initial conditions:

$$dt = 0.01, \quad x(0) = (1, 0), \quad \lambda_i = -10 \times i, \quad \xi^i(0) = 0. \quad (39)$$

Then the 4×4 matrix defined in (38) is invertible. Hence one can reconstruct an approximation (\hat{x}_1, \hat{x}_2) of the state (x_1, x_2) from the measurement y and approximations of $T_i(x)$ given by the dynamic $\xi_{k+1}^i = (1 + \lambda_i dt)\xi_k^i + dt y_k$.

On Fig. 1, we plot on a semi-log scale the evolution of the absolute error $\varepsilon_k = |x_k - \hat{x}_k|$ between the state and its observer for $k \in \{0, \dots, 500\}$ (*i.e.* $t \in [0, 5]$) for the observer based on functions T_{λ_i} designed for the original continuous-time system. Similarly, we make on Fig. 2 the same plot but for the observer based on functions $T_{\lambda_i}^d$ designed for the discrete-time system. We clearly see that the observer based on $T_{\lambda_i}^d$ is much more efficient than the one based on T_{λ_i} . On one hand, using $T_{\lambda_i}^d$, the error goes to zero until it achieves 10^{-12} , which is close to the machine epsilon ($\approx 10^{-16}$). Moreover, the state observer seems to converge exponentially to the state, with a rate $r \approx -4.58$ (estimation based on a linear regression made on $[0.5, 3]$). On the other hand, with T_{λ_i} , the observer does not converge to the state: it keeps an absolute error oscillating around 10^{-2} . This phenomenon is due to the fact that the trajectory of (33) is not invariant for this observer: even if it is well initialized (*i.e.* $x(0) = \hat{x}(0)$), the observer will oscillate around the state.

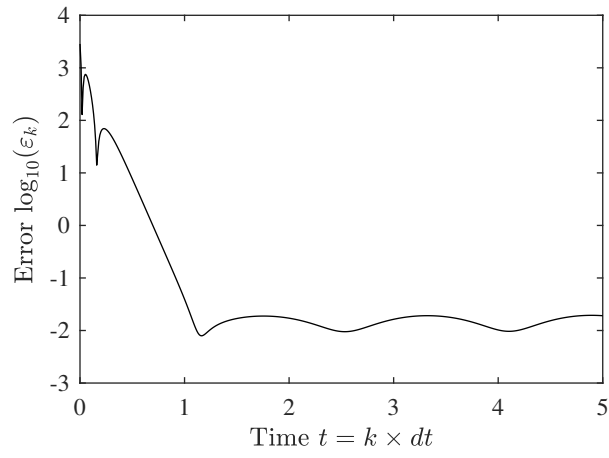


Figure 1: Evolution of the error between the state and the observer based on T_{λ_i} in semi-log scale

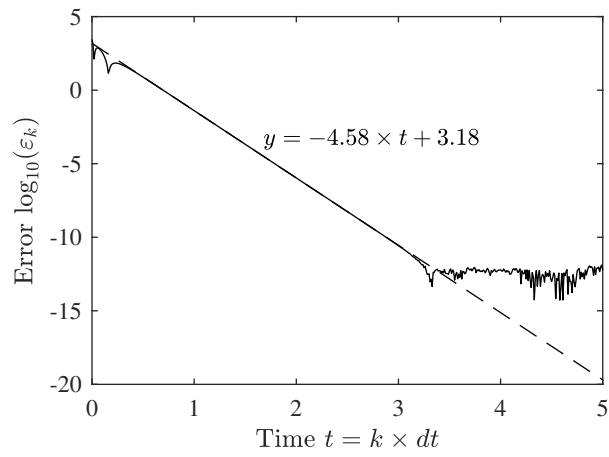


Figure 2: Evolution of the error between the state and the observer based on $T_{\lambda_i}^d$ in semi-log scale

4 Conclusion

We have shown how the initial Luenberger methodology can be applied to nonlinear discrete-time systems. It is based on the existence of a map satisfying some functional equation linked to the system, that transform the original system into a linear asymptotically stable one fed by the output. As soon as this map is uniformly injective, it allows us to estimate the state of the nonlinear system by simulating an autonomous system fed by the output and inverting this map. We stated sufficient conditions for the existence of such a map. In particular, we need the system to be reversible in time. Under a backward distinguishability hypothesis, we also proved that this map is injective.

References

- [1] V. Andrieu. Convergence speed of nonlinear luenberger observers. *SIAM Journal on Control and Optimization*, 52(5):2831–2856, 2014.
- [2] Vincent Andrieu and Laurent Praly. On the existence of a kazantzis–kravaris/luenberger observer. *SIAM J. Control and Optimization*, 45:432–456, 02 2006.
- [3] Pauline Bernard and Vincent Andrieu. Luenberger observers for nonautonomous nonlinear systems. *IEEE Transactions on Automatic Control*, PP:1–1, 09 2018.
- [4] M. Boutayeb and D. Aubry. A strong tracking extended kalman observer for nonlinear discrete-time systems. *IEEE Transactions on Automatic Control*, 44(8):1550–1556, Aug 1999.
- [5] C. Califano, S. Monaco, and D. Normand-Cyrot. On the observer design in discrete-time. *Systems & Control Letters*, 49(4):255 – 265, 2003.
- [6] Jean-Michel Coron. On the stabilization of controllable and observable systems by an output feedback law. *Mathematics of Control, Signals and Systems*, 7(3):187–216, 1994.
- [7] Henri Huijberts. On existence of extended observers for nonlinear discrete-time systems. *Lecture Notes in Control and Information Sciences*, 244, 03 1999.
- [8] Nikolaos Kazantzis and Costas Kravaris. Discrete-time nonlinear observer design using functional equations. *Systems & Control Letters*, 42(2):81 – 94, 2001.
- [9] D. G. Luenberger. Observing the state of a linear system. *IEEE Transactions on Military Electronics*, 8(2):74–80, April 1964.
- [10] E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 12 1934.
- [11] K Reif and R Unbehauen. The extended kalman filter as an exponential observer for nonlinear systems. *Signal Processing, IEEE Transactions on*, 47:2324 – 2328, 09 1999.
- [12] A. Zemouche and M. Boutayeb. Observer design for lipschitz nonlinear systems: The discrete-time case. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 53(8):777–781, Aug 2006.
- [13] A Zemouche and M Boutayeb. Observers design for discrete-time lipschitz nonlinear systems. state of the art and new results. *Proceedings of the IEEE Conference on Decision and Control*, pages 4780–4785, 12 2012.