State-Estimators for Chemical Reaction Networks of Feinberg-Horn-Jackson Zero Deficiency Type*

Madalena Chaves†
Eduardo D. Sontag
Department of Mathematics
Rutgers University
New Brunswick, NJ 08903
E-mail: {madalena,sontag}@math.rutgers.edu

Abstract

This paper provides a necessary and sufficient condition for detectability for chemical reaction networks of the Feinberg-Horn-Jackson zero deficiency type. Under this condition, an explicit construction of globally convergent observers is obtained based on ISS techniques. The observers are easy to implement, and several robustness aspects are tested in numerical simulations.

Keywords: observers, chemical reaction systems, detectability

1 Introduction

One of the most interesting questions in control theory is that of constructing observers. Observers compute estimates of the internal states of a dynamical system, using data provided by measurement probes or partial state information. For linear systems, Luenberger observers (also known as "deterministic Kalman filters" since they amount to Kalman filters designed without regard to the statistics of measurement noise) provide a general solution, but, for nonlinear systems, establishing generally applicable conditions for existence and convergence of observers is an open and active area of research.

The question of constructing observers has long been of interest in chemical engineering, and in particular for bioreactors; see for instance the papers [2, 5, 14] and the book [3]. Observers are of great potential interest, in particular, for biological experimentation and biomedical applications, where the online monitoring of proteins involved in signaling pathways (using for instance fluorescent labelings of molecules) will lead to a better understanding of cellular dynamical processes.

The main objective of this paper is to construct (when and if possible) state observers for the class of systems which describe chemical reaction networks of the type introduced by Feinberg, Horn, and Jackson in [6, 7, 8, 9, 10, 11]. As outputs, we take a subset of state variables or, more generally, monomials in state variables (which could, in practice, be associated to measured reaction rates). We provide here a complete solution to the observer problem for this class. The results given in [19] provide a convenient formalism as well as a set of technical tools, in the

*Work was supported in part by US Air Force Grant F49620-98-1-0242
†Supported by Fundação para a Ciência e a Tecnologia, Portugal, under the grant PraxisXXI/BD/11322/97

form of Lyapunov estimates, which are central to our results, and we will repeatedly refer to that paper for basic concepts and results.

As a first step in this paper, we prove a necessary and sufficient theoretical condition for detectability. As a second step, we proceed to explicitly construct a full-state observer that is guaranteed to converge globally, under the hypothesis that the system is detectable.

We also provide simulations which test the behavior of our observer in the presence of observation noise and even of unknown inputs acting on the system; the observers turn out to be surprisingly robust to such effects, but we leave for future work the formulation of theoretical results which quantify this robustness. The advantadges of our observer over the standard constructions of Luenberger and the extended Kalman filter are also illustrated by simulations. Most of the examples worked out in this paper, as well as the simulations testing the robustness of our observer, all concern the kinetic proofreading model proposed by McKeithan in [13] for T-cell receptor signal transduction (which motivated [19]).

The organization of this paper is as follows: this section introduces the problem as well as some notations and definitions to be adopted in this work. Section 2 introduces the observer, and the main results are stated and proved. Some examples that illustrate the result are presented, applying the construction to the case of the McKeithan network. Section 3 shows how to adapt the proof of the main result to a more general model. Section 4 provides simulations, testing in particular the effect of noise and of unknown inputs acting on states, or state drift. A comparison between the Luenberger observer, an extended Kalman filter and our observer is also given here. An appendix collects various technical results.

1.1 The Problem

The main objective of this paper is to construct (when and if possible) observers for a certain type of systems that provide a mathematical model for a class of chemical reactions. Several stability and other control-theoretic results applicable to such systems were discussed and surveyed (but without outputs) in the paper [19] (see also [4]). The systems that we study all have the generic form

$$\dot{x} = f(x), \quad y = h(x), \tag{1}$$

with the requirements on f and h specified next. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ is of the mass-action kinetics form

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_1^{b_{1j}} x_2^{b_{2j}} \dots x_n^{b_{nj}} (b_i - b_j), \tag{2}$$

where $m \leq n$ and each b_j is a column vector in $\mathbb{R}^n$ and has entries $b_{1j}, b_{2j}, \dots, b_{nj}$, which are nonnegative integers. It is assumed that $B := [b_1, b_2, \dots, b_m]$ has rank m and that none of its rows vanishes.

The constants a_{ij} are all nonnegative, and the matrix $A = [a_{ij}]$ is assumed to be irreducible. In the language of Feinberg et. al., irreducibility amounts to a restriction to "single linkage class" systems. This restriction can be removed, as explained in [19], provided that the space $\mathcal{D}$ introduced below is defined in a slightly different way to account for the number of connected components in the incidence graph of A. In order to simplify the presentation, the main result is stated for irreducible systems, and a sketch of how to treat the "multiple linkage classes" case is given in Section 3.

Although (2) is defined on all of $\mathbb{R}^n$, we will be interested only on those trajectories which evolve in the positive orthant $\mathbb{R}^n_{>0} = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_i > 0 \text{ for all } i\}$. It is easy to verify (cf. [19] and below) that $\mathbb{R}^n_{>0}$ is a forward invariant set for (1) when f has the form (2).

Define

$$\mathcal{D} := \text{span}\{b_i - b_j : i, j = 1, \ldots, m\}$$

and let the canonical basis of $\mathbb{R}^n$ be $\{e_i : i = 1, \ldots, n\}$.

The function f does not depend explicitly on time t; however, throughout the text, several variations of the system will be considered, obtained by adding input terms to f. The inputs will be assumed to be measurable and bounded functions $u : [0, +\infty) \to \mathbb{R}^p$, and the resulting right hand side will be denoted by $f^*(x, u)$.

In order to decide what kind of output functions $h : \mathbb{R}^n \to \mathbb{R}^p$ are natural to consider, we should think of the quantities that may be measured when performing a chemical experiment. Some possibilities are, for example, concentrations of some of the substances, or certain reaction rates (through markers, fluorescence, or energy released). This leads us to consider outputs whose coordinates are monomials. This kind of output includes both the case when the concentrations of some of the substances are measured (x_1, x_2, etc.) and the case when some of the reaction rates are measured (proportional to a monomial such as $x_1 x_2^3$).

Thus, we consider in this paper output maps $h : \mathbb{R}^n \to \mathbb{R}^p$ (typically, $p \leq n$), of the form:

$$h(x) = \begin{pmatrix} x_1^{c_{11}} x_2^{c_{12}} \cdots x_n^{c_{1n}} \\ \vdots \\ x_1^{c_{p1}} x_2^{c_{p2}} \ldots x_n^{c_{pn}} \end{pmatrix}, \tag{3}$$

where $C = (c_{ij})$ is a matrix whose entries are either 0 or real numbers ≥ 1. (In view of the preceding discussion, the most natural choice would be to take the entries of C to be nonnegative integers, but we allow more arbitrary exponents since the results do not require integers. The restriction $c_{ij} \geq 1$ is imposed in order to insure that $h(x)$ is locally Lipschitz, which is needed in order to guarantee uniqueness of solutions in the observer equations to be presented later. Although we are ultimately interested in behavior for positive x_i's, the ouputs make sense on $\mathbb{R}^n$, provided that we interpret exponents x_i^c as $|x_i|^c$ for negative x_i's.)

Let us introduce the following vector functions:

$$\rho_n(x) = (\ln x_1, \ldots, \ln x_n)' \quad \text{and} \quad \text{Exp}_n(v) = (e^{v_1}, \ldots, e^{v_n})'$$

defined on $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n$, respectively. (From now on, we will drop the subscript n of ρ_n and Exp_n, since its value is usually clear from the context.) Then, for $x \in \mathbb{R}^n_{>0}$,

$$\rho(h(x)) = C\rho(x) \quad \text{and} \quad h(x) = \text{Exp}(C\rho(x)),$$

as long as all state variables (concentrations, when dealing with chemical models) x_i are positive.

No Boundary Equilibra Assumption

For the rest of this paper, we will make the following assumption: *the system (1) has no boundary equilibrium in any positive stoichiometric class.* That is, if $x = (x_1, \ldots, x_n)$ is any vector with nonnegative components x_i, and some component x_i of x vanishes, and if $x - \bar{x} \in \text{span}\{b_i - b_j, \ \ i, j = 1, \ldots, m\}$ for some $\bar{x} \in \mathbb{R}^n_{>0}$, then $f(x) \neq 0$. This assumption amounts to saying that no reaction consistent with positive concentrations can be in equilibrium if one of the participating substances is at zero concentration. It is an assumption that is often satisfied in chemical reaction models, and is in particular satisfied in the main example (kinetic proofreading) to be discussed. It is possible to weaken this boundary assumption and still obtain significant (though more restricted) results (using the techniques developed in [19]) but we prefer not to do so in order to streamline the presentation.

Under the above assumption, the following result is a simple consequence of the theory developed by Feinberg et. al., see [19]. A (positive) class is any intersection of $\mathbb{R}^n_{>0}$ with an affine manifold of the form $\xi + \mathcal{D}$, where $\xi \in \mathbb{R}^n_{>0}$ and $\mathcal{D} = \mathrm{span}\{b_i - b_j : i, j = 1, \ldots, m\}$. We use the notation $|x|$ for Euclidean norms.

Theorem 1 For each positive class $\mathcal{C}$ there exists a (unique) state $\bar{x} = \bar{x}_{\mathcal{C}} \in \mathbb{R}^n_{>0}$ which is a globally asymptotically stable point relative to $\mathcal{C}$, i.e., for each $x_0 \in \mathcal{C}$, the solution of $\dot{x} = f(x),\ \ x(0) = x_0$ is defined for all $t \geq 0$, and $x(t) \to \bar{x}$ as $t \to \infty$, and for all $\varepsilon > 0$ there exists $\delta > 0$ such that, if $|\bar{x} - x_0| < \delta$, then $|\bar{x} - x(t)| < \varepsilon$ for all $t > 0$.

1.2 Detectability

Definition 1.1 The system (1) is *detectable* if, for every two trajectories x and z such that $x(t)$ evolves in $\mathbb{R}^n_{>0}$ and $z(t)$ evolves in $\mathbb{R}^n_{\geq 0}$ and both are defined for all $t \geq 0$,

$$h(x(t)) \equiv h(z(t)) \ \Rightarrow\ |x(t) - z(t)| \to 0 \text{ as } t \to \infty.$$

In particular, the system is *detectable on* $\mathbb{R}^n_{>0}$ if this implication is satisfied for every two trajectories $x(t)$ and $z(t)$ that evolve in $\mathbb{R}^n_{>0}$ and are defined for all $t \geq 0$.

Remark 1.2 Although we use it in this paper, this is not the most natural definition of detectability, because it is not "well posed" enough. In principle, one would want the definition of detectability to also include the property: "$h(x(t)) \approx h(z(t))$ for all t implies $|x(t) - z(t)|$ is asymptotically near zero as $t \to \infty$", which can be formulated as an "incremental output to state stability" (or more generally, "incremental input/output to state stability", if there are inputs) property. Such a more general concept, in the style of [12] and [21], can also be studied.

Definition 1.3 By a (full-state) *observer* for (1) we mean a system $\dot{z} = g(z, h(x))$ evolving on a state space $\mathcal{X}$ which is an open subset of $\mathbb{R}^n$ containing $\mathbb{R}^n_{>0}$, such that, for each $x(0) \in \mathbb{R}^n_{>0}$, $z(0) \in \mathbb{R}^n_{\geq 0}$, the composite system has solutions defined for all $t > 0$, and $|z(t) - x(t)| \to 0$ as $t \to +\infty$.

This is a weak definition, on that only "attraction" and not stability is required; however, in our proofs we achieve a stability property as well, as will follow from "KL" estimates. Detectability on $\mathbb{R}^n_{>0}$ is, obviously, necessary for the existence of an observer, and detectability is also necessary if the observer is consistent in the sense that $g(z, h(z)) = f(z)$ for all z, as our observer will be.

For system (1), we denote by E the set of all equilibria, i.e., all $\bar{x}$ such that $f(\bar{x}) = 0$. Let E_+ denote the subset of E of strictly positive equilibria, and E_0 denote the set of all boundary equilibria. Then $E = E_+ \cup E_0$ as a disjoint union. In [19], the following is proved:

Lemma 1.4 (Corollaries 7.5 and 7.7 in [19]) Consider the system $\dot{x} = f(x)$. Pick any trajectory x evolving in $\mathbb{R}^n_{\geq 0}$. Then, either $x(t) \equiv \xi \in E_0$ or $x(t) \in \mathbb{R}^n_{>0}$ for all $t > 0$.

Lemma 1.5 For system (1), detectability is equivalent to:

$$[h(\bar{x}) = h(\bar{z}) \ \ \&\ \ \bar{x} \in E_+,\, \bar{z} \in E] \ \ \Rightarrow\ \ \bar{x} = \bar{z}\,. \tag{4}$$

Proof. [necessity] Suppose that the system is detectable, and pick $\bar{x} \neq \bar{z}$ distinct elements of E_+ and E, respectively, so that $h(\bar{x}) = h(\bar{z})$. Then $x(t) \equiv \bar{x}$ and $z(t) \equiv \bar{z}$ are two trajectories evolving in $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n_{\geq 0}$, respectively, and $h(x(t)) \equiv h(z(t))$ and distinct limits, a contradiction.

[sufficiency] Suppose that (4) holds and pick any two trajectories evolving in $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n_{\geq 0}$, respectively, and $h(x(t)) \equiv h(z(t))$. Since h is continuous, this implies $h(\bar{x}) = h(\bar{z})$ for the limits of x and z, which exist and belong to E_+ and E, respectively (to see this: by Lemma 1.4, either $z(t) \equiv z(0) \in E_0$, or $z(t) \in \mathbb{R}^n_{>0}$ for all $t > 0$. Under the assumption that each positive class contains no boundary equilibria, Theorem 1 says that $z(t) \to \bar{z} \in E$. Similarly, $x(t) \to \bar{x} \in E_+$). Then (4) says $\bar{x} = \bar{z}$, as we wanted to prove. ∎

Remark 1.6 A more symmetric detectability condition would be "$[h(\bar{x}) = h(\bar{z})\ \&\ \bar{x}, \bar{z} \in E] \Rightarrow \bar{x} = \bar{z}$", but this turns out to be quite strong. It is not reasonable to expect $h(\cdot)$ to distinguish between any two boundary equilibria. For instance, in Example 2.1 below, $h(x) \equiv 0$ on E_0.

Since the function Exp is one to one, for positive vectors $x, z \in \mathbb{R}^n_{>0}$ we have

$$h(x) = h(z) \ \Leftrightarrow \ C\rho(x) = C\rho(z),$$

so that the condition $h(\bar{x}) = h(\bar{z})$ and $\bar{x}, \bar{z} \in E_+$, becomes just $\rho(\bar{x}) - \rho(\bar{z}) \in \ker C$. Recall also this fact from [19]: if $\bar{x} \in E_+$, then, for any $\bar{z} \in \mathbb{R}^n_{>0}$,

$$\rho(\bar{x}) - \rho(\bar{z}) \in \mathcal{D}^{\perp} \quad \Longleftrightarrow \quad \bar{z} \in E_+ . \tag{5}$$

Theorem 2 The following statements are equivalent:

(a) The system (1) with f as in (2) and h as in (3) is detectable on $\mathbb{R}^n_{>0}$;

(b) $\forall x, z \in \mathbb{R}^n_{>0}$, if $\rho(x) - \rho(z) \in \ker C$ and $x, z \in E_+$, then $x = z$;

(c) $\mathcal{D}^{\perp} \bigcap \ker C \ = \ \{0\}$;

(d) $\mathcal{D} + \operatorname{im} C' \ = \ \mathbb{R}^n$.

Proof. [(a) $\Leftrightarrow$ (b)] That condition (4) is equivalent to (b) follows immediately from the discussion above.

[(b) $\Rightarrow$ (c)] Pick any $y \in \mathcal{D}^{\perp} \bigcap \ker C$; we need to show that $y = 0$. Let $\bar{x}$ be any point of E_+ and put $\tilde{y} = \rho(\bar{x}) - y$, $\tilde{y} \in \mathbb{R}^n$. Then find $z = \operatorname{Exp}(\tilde{y}) \in \mathbb{R}^n_{>0}$ so that $\tilde{y} = \rho(z)$. Thus $y = \rho(\bar{x}) - \rho(z)$ with $\bar{x} \in E_+$ and $z \in \mathbb{R}^n_{>0}$. By definition of y, $\rho(\bar{x}) - \rho(z)$ is contained both in $\mathcal{D}^{\perp}$ and in $\ker C$. Condition (5) now implies that $z \in E_+$. By assumption (b), we now conclude that $\bar{x} = z$, or equivalently, $y = 0$ as wanted.

[(c) $\Rightarrow$ (b)] Let $x, z \in \mathbb{R}^n_{>0}$ satisfy both $\rho(x) - \rho(z) \in \ker C$ and $x, z \in E_+$. Then, from (5), if follows that $\rho(x) - \rho(z) \in \mathcal{D}^{\perp}$. Therefore, $\rho(x) - \rho(z) \in \mathcal{D}^{\perp} \bigcap \ker C$. By assumption (c) $\rho(x) - \rho(z) = 0$, and therefore, since $\rho(\cdot)$ is a bijective function on $\mathbb{R}^n_{>0}$, we have $x = z$.

[(c) $\Leftrightarrow$ (d)] This equivalence follows by duality. ∎

Corollary 1.7 The system (1) is detectable if and only if it is detectable on $\mathbb{R}^n_{>0}$ and $h(\bar{x}) \neq h(\bar{z})$ whenever $\bar{x} \in E_+$ and $\bar{z} \in E_0$.

Remark 1.8 A useful sufficient condition, on the matrix C, for system (1) to be detectable is now given. This condition is straightforward from the results above and depends only on the stoichiometric space $\mathcal{D}$ (more precisely, on the matrix B).

Since $h_i(\bar{x}) > 0$ for all i and all $\bar{x} \in E_+$, the condition

$$(\forall\ \bar{z} \in E_0)\ (\exists\ i \in \{1, \ldots, p\})\ h_i(\bar{z}) = 0, \tag{6}$$

is certainly sufficient for h to distinguish between interior and boundary equilibrium points.

One can show that every boundary equilibrium $\bar{z}$ satisfies $\bar{z}_1^{b_{1j}} \bar{z}_2^{b_{2j}} \dots \bar{z}_n^{b_{nj}} = 0$ for all $j = 1, \dots, m$, see Lemma 6.2 in [19]. An easy way to satisfy (6) is to ask that one of the columns of C' is a multiple of one of the columns of B — in other words, as discussed in the introduction, *choose one of the measured quantities to be one of the reaction rates.* (In fact, in this case, $h_i(\bar{z}) = 0$ for all $z \in E_0$, where i is the column in question.) Since dim $\mathcal{D} = m - 1$ and rank $B = m$, it is very easy to construct C so that both condition (6) and $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$ are satisfied, and thus system (1) detectable.

Remark 1.9 Note that dim$\mathcal{D} = m - 1$ and that detectability implies $m - 1 + \operatorname{rank} C' = n$. In the case the matrix C has full rank, then rank $C' = p$ and detectability implies $m - 1 + p = n$.

2 Constructing Observers

The main result is stated next and its proof is given in Section 2.2:

Theorem 3 Consider the system (1) and assume that it is detectable. Then the following system, with state space $\mathcal{X} = \mathbb{R}^n$, is an observer for the system (1):

$$\dot{z} = f(z) + C'(h(x) - h(z)). \tag{7}$$

Example 2.1 Consider the system with $n = 4$ and $m = 3$, determined by the vectors:

$$b_1 = (1, 1, 0, 0)', \quad b_2 = (0, 0, 1, 0)', \quad \text{and} \quad b_3 = (0, 0, 0, 1)'.$$

Then $\mathcal{D} = \operatorname{span}\{(1, 1, -1, 0), (1, 1, 0, -1)\}$, and for positive constants k, k_3, k_4 and β_3, the system (1) becomes the "McKeithan network" ([13]),

$$\begin{aligned}
\dot{x}_1 &= -kx_1x_2 + k_3x_3 + k_4x_4 \\
\dot{x}_2 &= -kx_1x_2 + k_3x_3 + k_4x_4 \\
\dot{x}_3 &= kx_1x_2 - (k_3 + \beta_3)x_3 \\
\dot{x}_4 &= \beta_3x_3 - k_4x_4.
\end{aligned} \tag{8}$$

The boundary equilibria of the system are given by $x_1x_2 = x_3 = x_4 = 0$, i.e., elements of E_0 have the form $(r, 0, 0, 0)'$ or $(0, r, 0, 0)'$ for $r \geq 0$. The positive classes are characterized by $x_1 + x_3 + x_4 = \alpha, x_2 + x_3 + x_4 = \beta$ for each pair of positive constants α, β. Suppose that the output is given by $h(x) = (x_1x_2^2, x_1x_4)'$. It is easy to check that h satisfies the detectability conditions (in fact, for this example, the condition $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$ is necessary and sufficient for detectability, since any such matrix C also satisfies (6)). Then, we can construct the following observer:

$$\begin{aligned}
\dot{z}_1 &= -kz_1z_2 + k_3z_3 + z_4z_4 \quad +(x_1x_2^2 - z_1z_2^2) + (x_1x_4 - z_1z_4) \\
\dot{z}_2 &= -kz_1z_2 + k_3z_3 + z_4z_4 \quad +2(x_1x_2^2 - z_1z_2^2) \\
\dot{z}_3 &= kz_1z_2 - (k_3 + \beta_3)z_3 \\
\dot{z}_4 &= \beta_3z_3 - k_4z_4 \quad +(x_1x_4 - z_1z_4).
\end{aligned}$$

The remainder of this section will be devoted to proving this theorem. The basic idea is to study the stability properties of system (1) when a certain input is added to the function f, specifically, the system with right-hand side: $f^*(z, u) := f(z) + C'(u - h(z))$. We will show that, for the system thus obtained, an "input to state stability" condition holds. The observer for (1) is obtained by letting the input be $u(t) = h(x(t))$. The analysis of this system with inputs is interesting in its own right, since it provides a means of studying the behavior of the model under bounded inputs.

2.1 An ISS Property

The definition of an input-to-state stable (ISS) system was introduced in [17]. Here, we adapt this notion to deal with constrained inputs and relative equilibria, as well as positive states (i.e., those states with all coordinates in the strictly positive half-line). We also use a notion of *semi-global ISS*, for dealing with systems that evolve in a compact set of their state-space.

From now on, whenever we mention an *input* $u(\cdot)$, we will mean a measurable essentially bounded function $u : [0, +\infty) \to \mathbb{R}^p$, possibly restricted to take values in a set $\mathbb{U}$ of $\mathbb{R}^p$. For $u : [0, +\infty) \to \mathbb{R}^p$ and any fixed $\bar{u} \in \mathbb{R}^n$, denote

$$\|u - \bar{u}\| := \text{ess.}\sup.\{|u(t) - \bar{u}| : t \geq 0\}.$$

Definition 2.2 A system $\dot{z} = f^*(z, u)$, with input-value set $\mathbb{U}$, evolving on a state space $\mathcal{X}$ which is an open subset of $\mathbb{R}^n$ containing $\mathbb{R}^n_{>0}$, is $\mathbb{R}^n_{>0}$*-(forward) invariant* (respectively, $\mathbb{R}^n_{\geq 0}$*-(forward) invariant*) if, for each initial state $z(0) \in \mathbb{R}^n_{>0}$ (respectively, $z(0) \in \mathbb{R}^n_{\geq 0} \cap \mathcal{X}$) and each $\mathbb{U}$-valued input $u(\cdot)$, the corresponding maximal solution of $\dot{z} = f^*(z, u)$ as a differential equation in $\mathcal{X}$, which is defined on an interval $J_{z(0),u} = [0, t_{\max})$, has values $z(t) \in \mathbb{R}^n_{>0}$ (respectively, $z(t) \in \mathbb{R}^n_{\geq 0} \cap \mathcal{X}$) for all $t \in J_{z(0),u}$.

The system is $\mathbb{R}^n_{>0}$*-(forward) complete* if it is $\mathbb{R}^n_{>0}$-(forward) invariant and, for each $z(0) \in \mathbb{R}^n_{>0}$ and $\mathbb{U}$-valued input $u(\cdot)$, $J_{z(0),u} = [0, +\infty)$.

The system is $\mathbb{R}^n_{\geq 0}$*-(forward) complete* if it is $\mathbb{R}^n_{\geq 0}$-(forward) invariant and, for each $z(0) \in \mathbb{R}^n_{\geq 0} \cap \mathcal{X}$ and $\mathbb{U}$-valued input $u(\cdot)$, $J_{z(0),u} = [0, +\infty)$.

For the following definitions, fix points $\bar{x} \in \mathbb{R}^n_{>0}$ and $\bar{u} \in \mathbb{U}$, where $\mathbb{U}$ is a subset of $\mathbb{R}^p$.

Definition 2.3 A system $\dot{z} = f^*(z, u)$, evolving on a state space $\mathcal{X}$ which is an open subset of $\mathbb{R}^n$ containing $\mathbb{R}^n_{>0}$, is *[semi-gobal] input-to-state stable with input-value set* $\mathbb{U}$ (with respect to the point $\bar{x}$ and the input $\bar{u}$) if it is $\mathbb{R}^n_{>0}$-complete and if for every compact set $F \subset \mathcal{X}$ there exist a function $\beta = \beta_F$ of class $\mathcal{KL}$ and a function $\varphi = \varphi_F$ of class $\mathcal{K}_\infty$ such that, for each $\mathbb{U}$-valued input $u(\cdot)$, and each initial condition $z_0 \in F \cap \mathbb{R}^n_{>0}$, it holds that

$$|z(t) - \bar{x}| \leq \beta(|z_0 - \bar{x}|, t) + \varphi(\|u - \bar{u}\|) \tag{9}$$

for all $t \geq 0$ such that $z(s) \in F$ for all $s \in [0, t]$.

If the same functions β, φ are valid for every compact subset F of $\mathcal{X}$, then the system is *input-to-state stable with input-value set* $\mathbb{U}$.

To study the stability properties of the system $\dot{z} = f^*(z, u)$ a Lyapunov-type technique is used, and the following definition is needed.

Definition 2.4 An *[semi-global] ISS-Lyapunov function with respect to the point* $\bar{x}$ *and input* $\bar{u}$, for the system $\dot{z} = f^*(z, u)$ with inputs in $\mathbb{U} \subseteq \mathbb{R}^p$, evolving on a state space $\mathcal{X}$ which is an open subset of $\mathbb{R}^n$ containing $\mathbb{R}^n_{>0}$, is a continuous function $V : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$, whose restriction to $\mathbb{R}^n_{>0}$ is continuously differentiable, which satisfies:

(i) For $z \in \mathbb{R}^n_{\geq 0}$, $V(z) \geq 0$ and $V(z) = 0 \Leftrightarrow z = \bar{x}$.

(ii) The set $\{z \in \mathbb{R}^n_{\geq 0} : V(z) \leq L\}$ is compact, for each positive constant L.

(iii) For each compact subset F of the state space $\mathcal{X}$, there exist two functions $\alpha = \alpha_F, \gamma = \gamma_F \in \mathcal{K}_\infty$ such that

$$\nabla V(z)\ f^*(z,u) \leq -\alpha(|z-\bar{x}|) + \gamma(|u-\bar{u}|)$$

for all $u \in \mathbb{U}$ and $z \in F \cap \mathbb{R}^n_{>0}$. If the same function γ is valid for every compact $F \subset \mathcal{X}$, then one says that V is *γ-uniform on $\mathbb{R}^n_{>0}$*.

If the functions α, γ given in (iii) may be chosen independently of the compact $F \subset \mathcal{X}$, then the function V is an *ISS-Lyapunov function with respect to the point $\bar{x}$ and input $\bar{u}$.*

Remark 2.5 This definition differs slightly from other definitions of "ISS-Lyapunov" functions, such as given in [20]. The difference is in the fact that here the function V is only required to be differentiable in the set $\mathbb{R}^n_{>0}$, and it is not required to satisfy a decrease condition except at positive vectors. Observe that V is not proper when restricted to the positive orthant: it remains finite as the boundary of $\mathbb{R}^n_{>0}$ is approached.

Remark 2.6 For a function V defined as above, there always exist $\mathcal{K}_\infty$ functions, ν_1, ν_2, such that

$$\nu_1(|z-\bar{x}|) \leq V(z) \leq \nu_2(|z-\bar{x}|) \tag{10}$$

for all $z \in \mathbb{R}^n_{\geq 0}$ (see [22]).

Next, we introduce our candidate ISS-Lyapunov function. Fix an $\bar{x} \in E_+$, and define the following function: $V : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$:

$$V(z) = \sum_{i=1}^{n} \bar{x}_i\ g\left(\frac{z_i}{\bar{x}_i}\right) = \sum_{i=1}^{n} \bar{x}_i \left[\frac{z_i}{\bar{x}_i} \ln \frac{z_i}{\bar{x}_i} + 1 - \frac{z_i}{\bar{x}_i}\right] \tag{11}$$

where $g : \mathbb{R}_{>0} \to \mathbb{R}$, $g(r) = r \ln r + 1 - r$, with the convention that $g(0) = 1$. The function V is continuously differentiable on $\mathbb{R}^n_{>0}$ and has the following properties:

(i) It is positive definite on $\mathbb{R}^n_{\geq 0}$ with respect to $\bar{x}$, i.e., $V(z) \geq 0$ and $V(z) = 0 \Leftrightarrow z = \bar{x}$: indeed, $g'(r) = \ln r$ so each $g(r)$ strictly increases on $(1, +\infty)$ and strictly decreases on $(0,1)$. But since $g(1) = 0$, it must be that $g(r) > 0$ for all $r \in \mathbb{R}_{>0}$. As a sum of nonnegative quantities, $V(z)$ is also nonnegative and, moreover,

$$V(z) = 0 \Leftrightarrow g\left(\frac{z_i}{\bar{x}_i}\right) = 0\, \forall\, i \Leftrightarrow z_i = \bar{x}_i\, \forall\, i,$$

that is, $V(z) = 0 \Leftrightarrow z = \bar{x}$.

(ii) For each constant $L > 0$ the set $\{z \in \mathbb{R}^n_{\geq 0} : V(z) \leq L\}$ is a compact subset of $\mathbb{R}^n_{\geq 0}$: $V(z) \leq L$ implies $g(z_i/\bar{x}_i) \leq L$ for all i; continuity of g on $[0, +\infty)$ (since $\lim_{w \to 0} g(w) = 1$) implies that z_i stays in a compact interval of $\mathbb{R}_{\geq 0}$ and hence z stays in a compact subset of $\mathbb{R}^n_{\geq 0}$.

In the next lemma, we show that the existence of an ISS-Lyapunov function (according to Definition 2.4) for a given system, implies that the trajectories of that system satisfy an ISS estimate.

Lemma 2.7 Consider an $\mathbb{R}^n_{>0}$-invariant system $\dot{z} = f^*(z,u)$, with input-value set $\mathbb{U}$. Fix a state $\bar{x} \in \mathbb{R}^n_{>0}$ and input value $\bar{u} \in \mathbb{U}$. Suppose that there is some [semi-global] ISS-Lyapunov funtion V with respect to $\bar{x}$ and $\bar{u}$. Assume that either: (a) the system is $\mathbb{R}^n_{>0}$-complete, or (b) the state-space $\mathcal{X}$ contains $\mathbb{R}^n_{\geq 0}$ and V is γ-uniform on $\mathbb{R}^n_{>0}$.

Then, the system is [semi-global] input-to-state stable with input-value set $\mathbb{U}$ (with respect to the same $\bar{x}$ and $\bar{u}$).

Proof. This proof is very similar to what is done in the case of the usual definition of an ISS system (see [20], for instance). Fix any compact set $F \subset \mathcal{X}$. According to the definition, V satisfies an estimate of the form

$$\nabla V(z)\ f^*(z,u) \leq -\alpha(|z-\bar{x}|) + \gamma(|u-\bar{u}|),$$

for each $z \in F \cap \mathbb{R}^n_{>0}$ and each u in the input value set $\mathbb{U}$, where $\alpha = \alpha_F, \gamma = \gamma_F \in \mathcal{K}_\infty$.

From Remark 2.6 there exist two class $\mathcal{K}_\infty$ functions, ν_1, ν_2, such that $\nu_1(|z-\bar{x}|) \leq V(z) \leq \nu_2(|z-\bar{x}|)$, for all $z \in F$.

We define new functions $\chi, \varphi \in \mathcal{K}_\infty$ by $\chi = \alpha^{-1} \circ (2\gamma)$ and $\varphi = \nu_1^{-1} \circ \nu_2 \circ \chi$, and $\beta, \tilde{\beta} \in \mathcal{KL}$ by $\beta(r,t) = \nu_1^{-1}(\tilde{\beta}(\nu_2(r),t))$, where $\tilde{\beta}(r,t)$ is the (unique) solution $y(t)$ of $\dot{y} = -\frac{1}{2}\alpha(\nu_2^{-1}(y)), y(0) = r$. Note that β and φ are independent of F whenever α, γ are independent of F.

Pick any initial condition $z_0 \in F \cap \mathbb{R}^n_{>0}$, and an input $u : [0,+\infty) \to \mathbb{U}$, and consider the corresponding maximal solution $z(t)$, defined on the (maximal) interval J. We first prove the $\mathbb{R}^n_{>0}$-completeness of the system. We have nothing to prove in case (a); if case (b) holds and the system is not complete, i.e., $J = [0, \hat{t}_{\max})$, where $\hat{t}_{\max} < +\infty$, then $\nabla V(z(t))\ f^*(z(t),u) \leq \gamma(\|u-\bar{u}\|) = c$ for all $t \in [0, \hat{t}_{\max})$. Then

$$\frac{d}{dt}V(z(t)) \leq c \ \Rightarrow\ V(z(t)) \leq V(z(0)) + c\,\hat{t}_{\max} = L \quad \forall t \in J$$

and, by property (ii) of Definition 2.4, $z(t)$ belongs to a compact subset of $\mathbb{R}^n_{\geq 0} \subset \mathcal{X}$ for all $t \in J$. Hence J must be $[0,+\infty)$ which is a contradiction.

Define also $t_{\max}$ to be such that $z(t) \in F$ for all $t \in [0, t_{\max}]$. Observe that:

$$|z-\bar{x}| > \chi(|u-\bar{u}|) \Leftrightarrow \alpha(|z-\bar{x}|) > 2\gamma(|u-\bar{u}|) \Rightarrow \nabla V(z)\ f^*(z,u) < -\frac{1}{2}\alpha(|z-\bar{x}|) \tag{12}$$

for all $z \in F \cap \mathbb{R}^n_{>0}$.

For $s = \nu_2(\chi(\|u-\bar{u}\|))$, define the following sublevel set of V: $S = \{\xi \in \mathbb{R}^n_{\geq 0} : V(\xi) \leq s\}$.

Claim. Suppose there exists an instant $\sigma \in I$ such that $z(\sigma) \in S$. Then $z(t) \in S$ for all $\sigma \leq t \leq t_{\max}$.

To see this, argue by contradiction: suppose there exists a $t > \sigma$ (but $t \leq t_{\max}$) and an $\varepsilon > 0$ such that $V(z(t)) > s + \varepsilon$. Let $\tau = \inf\{t \geq \sigma : V(z(t)) \geq s + \varepsilon\}$. Then $z(\tau) \notin S$ which implies $V(z(\tau)) > \nu_2(\chi(\|u-\bar{u}\|))$, and therefore, since $V(z) \leq \nu_2(|z-\bar{x}|)$ and ν_2 is strictly increasing,

$$\nu_2(|z(\tau)-\bar{x}|) > \nu_2(\chi(\|u-\bar{u}\|)) \Longleftrightarrow |z(\tau)-\bar{x}| > \chi(\|u-\bar{u}\|).$$

By (12), $\frac{d}{dt}V(z(t))|_\tau < 0$, implying that $V(z(t_*)) \geq V(z(\tau))$ for some $t_* \in (\sigma,\tau)$, and thus contradicting minimality of τ. So the claim holds.

Now, let $T = \inf\{\sigma : z(\sigma) \in S\}$ (with $T = t_{\max}$ if the trajectory never enters S). We have two cases to consider, for each $0 < t \leq t_{\max}$:

For $t \in (T, t_{\max}]$: $V(z(t)) \leq \nu_2(\chi(\|u-\bar{u}\|))$ implies

$$\nu_1(|z(t)-\bar{x}|) \leq \nu_2(\chi(\|u-\bar{u}\|)) \Longrightarrow |z(t)-\bar{x}| \leq \nu_1^{-1} \circ \nu_2 \circ \chi(\|u-\bar{u}\|).$$

For $t \leq T$: $V(z(t)) \geq \nu_2(\chi(\|u-\bar{u}\|))$, which implies $|z(t)-\bar{x}| \geq \chi(\|u-\bar{u}\|)$ and hence

$$\frac{\mathrm{d}}{\mathrm{d}t}V(z(t)) \leq -\frac{1}{2}\alpha(|z(t)-\bar{x}|) \leq -\frac{1}{2}\alpha[\nu_2^{-1}(V(z(t)))].$$

By a standard comparison principle, there exists a function $\tilde{\beta} \in \mathcal{KL}$ (which depends only on α and ν_2) such that $V(z(t)) \leq \tilde{\beta}(V(z_0), t)$ for all $t < T$. Then

$$|z(t) - \bar{x}| \leq \nu_1^{-1}(\tilde{\beta}(V(z_0), t)) \leq \nu_1^{-1}(\tilde{\beta}(\nu_2(|z_0 - \bar{x}|), t)) := \beta(|z_0 - \bar{x}|, t).$$

Thus, for all $t \in I$,

$$|z(t) - \bar{x}| \leq \max\{\beta(|z_0 - \bar{x}|, t), \varphi(\|u - \bar{u}\|)\} \leq \beta(|z_0 - \bar{x}|, t) + \varphi(\|u - \bar{u}\|)$$

where $\beta = \beta_F \in \mathcal{KL}$ and $\varphi = \varphi_F \in \mathcal{K}_\infty$.

If V is an ISS-Lyapunov function, then α, γ are independent of F, and so are β, φ. ∎

2.2 Proof of Theorem 3

The following formulas will be useful:

$$\pi_j(z, \bar{x}) = \pi_j := \left[\frac{z_1}{\bar{x}_1}\right]^{b_{1j}} \left[\frac{z_2}{\bar{x}_2}\right]^{b_{2j}} \cdots \left[\frac{z_n}{\bar{x}_n}\right]^{b_{nj}},$$

$$q_j(z, \bar{x}) = q_j := \langle b_j, \rho(z) - \rho(\bar{x}) \rangle,$$

where $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)' \in E_+$ and π_j is defined for $z \in \mathbb{R}^n_{\geq 0}$ and q_j is defined for $z \in \mathbb{R}^n_{>0}$. Observe that $\pi_j = e^{q_j}$. Define the function $\mathbb{R}^n_{\geq 0} \times \mathbb{R}^n_{>0} \to \mathbb{R}_{\geq 0}$

$$\Psi(z, \bar{x}) := \sum_{i=1}^{m} \sum_{j=1}^{m} \left(e^{-\pi_i} - e^{-\pi_j}\right)^2.$$

Lemma 2.8 If $\bar{x} \in E_+$, then for all $z \in \mathbb{R}^n_{\geq 0}$:

$$\Psi(z, \bar{x}) = 0 \;\Leftrightarrow\; z \in E.$$

Proof. The function Ψ can be zero only if $\pi_i = \pi_j$ for all $i, j \in \{1, \ldots, m\}$. This can happen if (a) either $\pi_i = 0$ for some $i \in \{1, \ldots, m\}$, hence for all i in this set, which implies that $z \in E_0$; (b) or all $\pi_i \neq 0$ and $e^{q_i} = e^{q_j}$ for all $i, j \in \{1, \ldots, m\}$ which is equivalent to $q_i - q_j = 0$ for all $i, j \in \{1, \ldots, m\}$ and, from (5), we know this implies $z \in E_+$. ∎

Lemma 2.9 Let $c_0, c_1 > 0$ be constants and fix $\bar{x} \in E_+$. Let h be a function of the form (3) such that $\dot{x} = f(x), y = h(x)$ is detectable. Then, the function $\mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$,

$$\mu(z) := c_0 \Psi(z, \bar{x}) + c_1 |h(z) - h(\bar{x})|^2$$

is positive for all $z \in \mathbb{R}^n_{\geq 0} \setminus \{\bar{x}\}$, and $\mu(z) = 0$ if and only if $z = \bar{x}$.

Moreover, given any compact subset $F \subset \mathbb{R}^n_{\geq 0}$, there exists a class $\mathcal{K}_\infty$ function, $\alpha = \alpha_{\bar{x}, F}$ such that

$$\mu(z) \geq \alpha(|z - \bar{x}|)$$

for all $z \in F$.

Proof. Since both terms in μ are nonnegative it is clear that μ can be zero only if both terms are simultaneously zero. By Lemma 2.8, $\Psi(z,\bar{x}) = 0$ iff $z \in E$. Thus we conclude that $\mu(z) = 0$ if and only if $z \in E$ and $h(z) = h(\bar{x})$. But, from the detectability conditions (and because $\bar{x} \in E_+$), we have that: (i) $z \in E_0 \Rightarrow h(z) \neq h(\bar{x})$, so $\mu(z) > 0$; (ii) $z \in E_+$ and $h(z) = h(\bar{x})$ imply $z = \bar{x}$, so that $\mu(z) = 0$ if and only if $z = \bar{x}$.

Next, let $F \subset \mathbb{R}^n_{\geq 0}$ be any compact set and set R to be such that the closed ball $|z - \bar{x}| \leq R$ contains the set F. Consider the function $\mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ given by

$$\alpha(r) := \frac{r}{r+1} \min\{\mu(z,\bar{x}) : r \leq |z - \bar{x}| \leq R, z \in \mathbb{R}^n_{\geq 0}\}$$

for all $0 \leq r \leq R$, and $\alpha(r) := \alpha(R)\frac{r}{R}$ for all $r > R$. Since $\mu(z) = 0$ iff $z = \bar{x}$ and since the minimum is taken over a compact set, the function α satisfies $\alpha(0) = 0$ and $\alpha(r) > 0$ for $r > 0$. It is continuous for $0 \leq r < R$ because μ is, and for $R \leq r$ by construction. Also clearly, for $R \leq r$, α is strictly increasing and satisfies $\alpha(r) \to +\infty$ as $r \to +\infty$. For $0 \leq r \leq R$, $\alpha(r)$ is also strictly increasing, as a product of a strictly increasing function and a nondecreasing function. Finally, by construction, $\mu(z) \geq \alpha(|z - \bar{x}|)$ for all F. ∎

We next establish some useful estimates.

Lemma 2.10 Let $f(z)$ be defined as in (2). There exists a positive constant $\kappa(A)$ and a continuous function $c : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ given by $c(\xi) = \frac{1}{2}\min_j e^{\langle b_j,\rho(\xi)\rangle}$ such that, for all $z, \bar{x} \in \mathbb{R}^n_{>0}$:

$$\langle \rho(z) - \rho(\bar{x}), f(z)\rangle \leq -\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}). \tag{13}$$

Proof. Note that

$$\begin{aligned}
\langle \rho(z) - \rho(\bar{x}), f(z)\rangle &= \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} e^{\langle b_j,\rho(\bar{x})\rangle}[e^{q_j}(q_i - q_j) - (e^{q_i} - e^{q_j})] \\
&\leq -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} e^{\langle b_j,\rho(\bar{x})\rangle}(e^{-\pi_i} - e^{-\pi_j})^2 \\
&\leq -\frac{1}{2}\,\kappa(A)\,\min_j e^{\langle b_j,\rho(\bar{x})\rangle}\Psi(z,\bar{x}).
\end{aligned}$$

To justify these inequalities, consider the function, for any fixed $a \in \mathbb{R}$:

$$f_a(r) := e^a(r - a) - (e^r - e^a) + \frac{1}{2}(e^{-e^r} - e^{-e^a})^2$$

and note that it is negative for $r \neq a$, and zero at $r = a$. Indeed, consider its derivative

$$f_a'(r) = e^a - e^r - e^r e^{-e^r}(e^{-e^r} - e^{-e^a})$$

and note that

(i) $|e^{-e^r} - e^{-e^a}| \leq |e^r - e^a|$, because the function e^{-y} is Lipschitz for $y \in [0,+\infty)$ with constant equal to 1;

(ii) $e^r < e^{e^r}$, so $e^r e^{-e^r} < 1$.

From (i) and (ii) it follows that, for all $r, a \in \mathbb{R}$, $r \neq a$,

$$e^r e^{-e^r}|e^{-e^r} - e^{-e^a}| < |e^r - e^a|,$$

meaning that the first term, $e^a - e^r$, always dominates the sign of $f_a'(r)$. Therefore, when $r > a$, $f_a'(r) < 0$ and hence f_a is strictly decreasing on the interval $(a, +\infty)$; when $r < a$, $f_a'(r) > 0$ and hence f_a is strictly increasing on the interval $(-\infty, a)$. This gives the desired result, since $f_a(a) = 0$.

So, let $a = q_j$ and $r = q_i$ and recall that $\pi_i = e^{q_i}$, to obtain the first inequality. For the second inequality, use the irreducibility of the matrix $A = (a_{ij})$ and Lemma 8.1 of [19], to get the positive constant $\kappa(A)$ that depends on A and satisfies that inequality. ∎

Lemma 2.11 For every compact set F in $\mathbb{R}^n_{\geq 0}$, there exists a constant $c_F > 0$ such that for every $z \in F \cap \mathbb{R}^n_{>0}$:

$$-\langle C(\rho(z) - \rho(\bar{x})), h(z) - h(\bar{x})\rangle \leq -c_F\,|h(z) - h(\bar{x})|^2. \tag{14}$$

Proof. Recall the form of the function h and observe that

$$\langle C(\rho(z) - \rho(\bar{x})), h(z) - h(\bar{x})\rangle = \langle \rho(h(z)) - \rho(h(\bar{x})), h(z) - h(\bar{x})\rangle,$$

which in turn is equal to $\sum |\ln(h_i(z)) - \ln(h_i(\bar{x}))||h_i(z) - h_i(\bar{x})|$.

To show (14), we let $M = \max\{h_i(z) : z \in F, i \in \{1, \ldots, p\}\}$ and put $\kappa = 1/M$. For any fixed $a \in (0, M]$, consider the (scalar) function

$$f_a(r) := |\ln r - \ln a| - \kappa |r - a|.$$

We now show that $f_a(r) \geq 0$ for every $0 < r \leq M$. Clearly $f_a(a) = 0$. For $r > a$, $f_a(r) = \ln r - \ln a - \kappa(r - a)$ and $f_a'(r) = \frac{1}{r} - \kappa \geq \frac{1}{M} - \kappa = 0$, so that $f_a(r)$ is increasing for all $r > a$, hence always nonnegative. For $r < a$, $f_a(r) = -\ln r + \ln a + \kappa(r - a)$ and $f_a'(r) = -\frac{1}{r} + \kappa \leq -\frac{1}{a} + \kappa \leq -\frac{1}{M} + \kappa = 0$, so that $f_a(r)$ is decreasing for all $r < a$, hence always nonnegative.

Therefore, taking $r = h_i(z)$ and $a = h_i(\bar{x})$, we obtain $|\ln(h_i(z)) - \ln(h_i(\bar{x}))| \geq \kappa |h_i(z) - h_i(\bar{x})|$ for each i, which gives the desired inequality with $c_F = \kappa$. ∎

Lemma 2.12 Let $f^*(z, u) = f(z) + C'(u - h(z))$ with C such that $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. Let $\bar{x}$ denote any point in E_+, and θ be any real number with $0 < \theta < 1$. Define the following subset of $\mathbb{R}^p$:

$$\mathbb{U}_\theta = \{u \in \mathbb{R}^p : |u_k - h_k(\bar{x})| \leq \frac{\theta}{2} h_k(\bar{x}),\ k = 1, \ldots, p\}.$$

Then, for the function V defined in (11), there exist functions $\alpha_1 = \alpha_{1,\bar{x}}$ positive definite, and $\gamma = \gamma_{\bar{x}}$ of class $\mathcal{K}_\infty$ such that

$$\nabla V(z)\ f^*(z, u) \leq -\alpha_1(|\rho(z) - \rho(\bar{x})|) + \gamma(|u - h(\bar{x})|),$$

for all $z \in \mathbb{R}^n_{>0}$ and all $u \in \mathbb{U}_\theta$. In particular, one may choose $\gamma(r) = c_3 r^2$, with $c_3 = \frac{2}{\theta\lambda}$ where $\lambda = \min\{h_i(\bar{x})/2 : i = 1, \ldots, p\}$.

Furthermore, let F be any compact subset of $\mathbb{R}^n_{\geq 0}$ which contains $\bar{x}$. Then there exists a function $\alpha = \alpha_F$ of class $\mathcal{K}_\infty$ such that (γ is the same as before)

$$\nabla V(z)\ f^*(z, u) \leq -\alpha(|z - \bar{x}|) + \gamma(|u - h(\bar{x})|),$$

for all $z \in F \cap \mathbb{R}^n_{>0}$ and all $u \in \mathbb{U}_\theta$.

Proof. Pick any $\bar{x} \in E_+$ and any $0 < \theta < 1$. We have

$$\nabla V(z)\ f^*(z,u) = \langle \rho(z) - \rho(\bar{x}), f(z)\rangle + \langle \rho(z) - \rho(\bar{x}), C'(u - h(z))\rangle.$$

Using the notation $\varrho = \rho(z) - \rho(\bar{x})$, notice that the second term on the right can be rewritten as:

$$\langle C\varrho, u - h(z)\rangle = \langle C\varrho, u - h(\bar{x})\rangle - \langle C\varrho, h(z) - h(\bar{x})\rangle.$$

Introducing the notation $\sigma = C\varrho$, $\mu = h(z) - h(\bar{x})$ and $v = u - h(\bar{x})$, the expression for $\nabla V(z)\ f^*(z,u)$ becomes

$$\nabla V(z)\ f^*(z,u) = \langle \varrho, f(z)\rangle - \langle \sigma, \mu\rangle + \langle \sigma, v\rangle = P(z,\bar{x}) + R(z,u,\bar{x})$$

where

$$\begin{aligned} P(z,\bar{x}) &= \langle \varrho, f(z)\rangle - (1-\theta)\langle \sigma, \mu\rangle, \\ R(z,u,\bar{x}) &= -\theta\langle \sigma, \mu\rangle + \langle \sigma, v\rangle = \sum \sigma_i(-\theta\mu_i + v_i) \end{aligned}$$

We now bound each of these terms.

Step 1. We show first that $R(z,u,\bar{x}) \le c_3|v|^2$, for some positive constant c_3.

Notice that $\mu_i\sigma_i = (h_i(z) - h_i(\bar{x}))(\ln h_i(z) - \ln h_i(\bar{x})) \ge 0$ for all pairs $h_i(z), h_i(\bar{x})$.

(i) if $\theta|\mu_i| \ge |v_i|$, then immediately $\sigma_i(-\theta\mu_i + v_i) \le 0$.

(ii) if $\theta|\mu_i| < |v_i|$, then $\sigma_i(-\theta\mu_i + v_i) \le 2|\sigma_i||v_i| \le \frac{2c_L}{\theta}|v_i|^2$ where the last inequality follows from the bounds on u:

$$|v_i| = |u_i - h_i(\bar{x})| \le \frac{\theta}{2}h_i(\bar{x}) \quad \Rightarrow \quad |h_i(z) - h_i(\bar{x})| = |\mu_i| \le \frac{1}{\theta}|v_i| \le \frac{1}{2}h_i(\bar{x}),$$

so that $h_i(z) \ge h_i(\bar{x})/2$, and with a Lipschitz constant, c_L, of the logarithmic function on $[\lambda, +\infty)$, when $\lambda = \min\{h_i(\bar{x})/2 : i = 1, \ldots, p\}$ (for instance, $c_L = 1/\lambda$):

$$|\sigma_i| = |\ln h_i(z) - \ln h_i(\bar{x})| \le c_L|h_i(z) - h_i(\bar{x})| = c_L|\mu_i|.$$

In either case, $R(z,u,\bar{x}) \le \frac{2c_L}{\theta}\sum |v_i|^2$ as wanted. We may take $\gamma(r) = \frac{2}{\theta\lambda}r^2$.

Step 2. Show that $P(z,\bar{x}) \le -\alpha_1(|\rho(z) - \rho(\bar{x})|)$, where α_1 is positive definite. From equation (13) and from the form of h, it follows that

$$P(z,\bar{x}) \quad \le \quad -\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}) - (1-\theta)\langle \rho(h(z)) - \rho(h(\bar{x})), h(z) - h(\bar{x})\rangle.$$

Then $P(z,\bar{x})$ is clearly either negative or zero. Recall that, for $z \in \mathbb{R}^n_{>0}$, the first term is zero only when $z \in E_+$ and the second term is zero only when $h(z) = h(\bar{x})$: thus, from the detectability condition it follows that $P(z,\bar{x})$ may be zero only when $z = \bar{x}$.

Consider the function $\mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$

$$\alpha_1(r) = \inf\left\{\kappa(A)c(\bar{x})\Psi(z,\bar{x}) + (1-\theta)\langle \rho(h(z)) - \rho(h(\bar{x})), h(z) - h(\bar{x})\rangle :\ z \in \mathcal{C}_r\right\},$$

where $\mathcal{C}_r := \{z \in \mathbb{R}^n_{>0} : |\rho(z) - \rho(\bar{x})| = r\}$. This function is continuous, has $\alpha_1(0) = 0$ and is strictly positive for all $r > 0$ (by the previous discussion and since $\mathcal{C}_r$ defines a compact subset of $\mathbb{R}^n_{>0}$, because, for all i, $\ln z_i \to \pm\infty$ if $z_i \to +\infty$ or $z_i \to 0$), and satisfies the desired inequality for every $z \in \mathbb{R}^n_{>0}$. Steps 1 and 2 establish the first part of the Lemma.

Step 3. Assume now that $F \subset \mathbb{R}^n_{\ge 0}$ is an arbitrary compact set. Then, using both (13) and (14), we have that

$$P(z,\bar{x}) \quad \le \quad -\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}) - (1-\theta)c_F\, |h(z) - h(\bar{x})|^2.$$

By Lemma 2.9, there exists a function $\alpha = \alpha_{\bar{x},F}$, of class $\mathcal{K}_\infty$, such that $P(z,\bar{x}) \le -\alpha(|z - \bar{x}|)$, for all $z \in \mathbb{R}^n_{>0} \cap F$. Since the estimate obtained in step 1 is valid for all $z \in \mathbb{R}^n_{>0}$, putting these together proves the second part of the Lemma. ∎

Proposition 2.13 Suppose that the system defined by (2) and (3) is detectable. Consider the system with inputs

$$\dot{z} = f^*(z,u) := f(z) + C'(u - h(z)) \tag{15}$$

with state-space $\mathcal{X} = \mathbb{R}^n$. Then, the system is $\mathbb{R}^n_{>0}$-invariant with input-value set $\mathbb{R}^p_{\geq 0}$.

Furthermore, let θ be any real number with $0 < \theta < 1$, and pick any fixed state $\bar{x} \in E_+$. Let $\mathbb{U}_\theta$ be the subset of $\mathbb{R}^p$ defined in Lemma 2.12. Then, the system (15) is semi-global ISS with input value set $\mathbb{U}_\theta$ (with respect to the point $\bar{x}$ and the input $\bar{u} = h(\bar{x})$).

Proof. The proof of the first statement, namely that the system is $\mathbb{R}^n_{>0}$-invariant with input-value set $\mathbb{R}^p_{\geq 0}$, is fairly routine, and it proceeds as follows.

Given an initial condition $z(0) \in \mathbb{R}^n_{>0}$, and an $\mathbb{U}_\theta$-valued input, let $z(t)$ be the maximal solution of (15), defined on a (maximal) interval J. Let $\mathcal{I} = [0, +\infty)$.

Assume that one of the coordinates becomes ≤ 0 at some instant and define

$$t_0 = \inf\{t \in J : z_k(t) = 0 \text{ for some } 1 \leq k \leq n\}.$$

Pick one coordinate k such that $z_k(t_0) = 0$. We reorder variables, singling out this coordinate, and look at the time-dependent differential equation that results by fixing the remaining $n-1$ variables. It is useful for that purpose to introduce the following notation:

$$(\check{z}(t), x) = (z_1(t), \ldots, z_{k-1}(t), x, z_{k+1}(t), \ldots, z_n(t)).$$

In addition, we wish to see the obtained scalar equation as well-defined for all t, not just $t \leq t_0$. So we construct a new function $F : \mathcal{I} \times \mathbb{R} \to \mathbb{R}$ as follows:

$$F(t,x) = \begin{cases} f_k^*(\check{z}(t), x; u(t)), & t \in [0, t_0) \\ f_k^*(\check{z}(t_0), x; u_0), & t \in [t_0, +\infty) \end{cases}$$

where u_0 is any fixed element of $\mathbb{U}_\theta$. Then, for each fixed t, $F(t,x)$ is locally Lipschitz in x and the Lipschitz constants, $\alpha(t)$, are uniformly bounded (and hence locally integrable as a function of time). In addition, for each fixed x, $F(t,x)$ is measurable as a function of time. Thus the standard existence and uniqueness conditions apply.

Claim. $F(t,0) \geq 0$ for almost all $t \in \mathcal{I}$.

To prove this, write

$$\begin{aligned} f_k^*(\check{z}, x; u) &= \sum_{i=1}^{m} \sum_{j \in A_0} a_{ij} z_1^{b_{1j}} \ldots z_{k-1}^{b_{(k-1)j}} z_{k+1}^{b_{(k+1)j}} \ldots z_n^{b_{nj}} b_{ki} \\ &+ \sum_{i=1}^{m} \sum_{j \in A_+} a_{ij} z_1^{b_{1j}} \ldots z_{k-1}^{b_{(k-1)j}} x^{b_{kj}} z_{k+1}^{b_{(k+1)j}} \ldots z_n^{b_{nj}} (b_{ki} - b_{kj}) \\ &+ \sum_{j=1}^{p} c_{jk} [u_j - h_j(\check{z}, x)], \end{aligned} \tag{16}$$

where $A_0 = \{j : b_{kj} = 0\}$ and $A_+ = \{j : b_{kj} > 0\}$.

For $x = 0$ and $t \in \mathcal{I}$:

(a) the third term is nonnegative since we are assuming that $c_{ji} \geq 0$ and $u_j \geq 0$ for all i, j, and because $h_j(\check{z}, x) = z_1^{c_{j1}} \ldots x^{c_{jk}} \ldots z_n^{c_{jn}}$, so either $c_{jk} = 0$, or $c_{jk} > 0$ and $h_j(\check{z}, 0) = 0$.

(b) the second term is zero since $x = 0$;

(c) the first term is nonnegative since, by definition of t_0, we are evaluating at $z_i = z_i(t) \geq 0$, for all i and $t \leq t_0$, and $z_i = z_i(t_0)$ for $t > t_0$.

This proves the claim.

Moreover, notice that, for all $t \leq t_0$, the scalar variable $z_k(t)$ satisfies the initial value problem $\dot{x} = F(t,x)$, $x(0) = z_k(0)$, where $F(t,0) \geq 0$ for all $t \geq 0$. Solutions of this initial value problem exist on an open interval $\tilde{J}$, and this interval contains $[0,t_0]$ because $z_k(t)$ solves the equation in that interval. Then, by Lemma A.1, $x(t) > 0$ on $\tilde{J}$ and $z_k(t) = x(t) > 0$ for all $t < t_0$; since both $x(t)$ and $z_k(t)$ are continuous functions, we also have that $z_k(t_0) = x(t_0)$, contradicting the fact that $z_k(t_0) = 0$.

This concludes the proof of $\mathbb{R}^n_{>0}$-invariance with input-value set $\mathbb{R}^p_{\geq 0}$. This implies that (15) is also $\mathbb{R}^n_{>0}$-invariant with (the smaller) input value set $\mathbb{U}_\theta$. To prove that (15) is semi-global ISS with input value set $\mathbb{U}_\theta$, it is enough, by Lemma 2.7, to show that this system admits the function V defined in (11) as a semi-global ISS-Lyapunov function with respect to the state $\bar{x}$ and the input $h(\bar{x})$, with V γ-uniform on $\mathbb{R}^n_{>0}$. Properties (i) and (ii) of Definition 2.4 have already been shown in the discussion following formula (11). Property (iii) follows from Lemma 2.12, as well as the fact that γ may indeed be chosen independently of the set F. ∎

Lemma 2.14 Let $f^*(z,u) = f(z) + C'(u - h(z))$ and let V denote the function defined in (11). For each fixed $\bar{x} \in E_+$, and each constant $u_{\max} \geq 0$, there exists a constant $c_{\bar{x},u_{\max}}$ such that

$$\nabla V(z)\ f^*(z,u) \leq c_{\bar{x},u_{\max}}, \quad \forall\, z \in \mathbb{R}^n_{>0}, \quad \forall\, u \in [0,u_{\max}]^p.$$

Proof. Pick any $\bar{x} \in E_+$ and any nonnegative $u_{\max}$. Then, using estimate (13),

$$\begin{aligned}
\nabla V(z)\ f^*(z,u) &= \langle \rho(z) - \rho(\bar{x}), f(z)\rangle + \langle \rho(z) - \rho(\bar{x}), C'(u - h(z))\rangle \\
&\leq -\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}) + \langle \rho(z) - \rho(\bar{x}), C'(u - h(z))\rangle \\
&\leq \langle C(\rho(z) - \rho(\bar{x})), u - h(z)\rangle := \sum_{i=1}^{p} s_i(z,u,\bar{x})
\end{aligned}$$

where $s_i(z,u,\bar{x}) = (\ln h_i(z) - \ln h_i(\bar{x}))(u_i - h_i(z))$.

Now, let $h_{\max}$ be a constant depending only on $\bar{x}$ such that, $\max_{i=1,\ldots,p} |\ln h_i(\bar{x})| \leq h_{\max}$.

For each fixed z, define the following finite, disjoint sets of integers

$$I_+ = I_+(z) = \{i : h_i(z) > 1\} \quad \text{and} \quad I_- = I_-(z) = \{i : h_i(z) \leq 1\}.$$

Clearly $I_+ \cup I_- = \{1,\ldots,p\}$, and for each $i \in I_-$,

$$\begin{aligned}
u_i \ln h_i(z) &\leq 0 \\
|u_i \ln h_i(\bar{x})| &\leq u_{\max} |\ln h_i(\bar{x})| \leq u_{\max} h_{\max} \\
|h_i(z) \ln h_i(z)| &\leq \frac{1}{e} \\
|h_i(z) \ln h_i(\bar{x})| &\leq |\ln h_i(\bar{x})| \leq h_{\max},
\end{aligned}$$

so that, for the corresponding ith term in the above sum:

$$\begin{aligned}
s_i(z,u,\bar{x}) &= u_i \ln h_i(z) - u_i \ln h_i(\bar{x}) - h_i(z) \ln h_i(z) + h_i(z) \ln h_i(\bar{x}) \\
&\leq 0 + u_{\max} h_{\max} + \frac{1}{e} + h_{\max}.
\end{aligned}$$

On the other hand, for each $i \in I_+$, s_i can be decomposed into two terms

$$-(\ln h_i(z) - \ln h_i(\bar{x}))(h_i(z) - h_i(\bar{x})) + (\ln h_i(z) - \ln h_i(\bar{x}))(u_i - h_i(\bar{x}))$$

the first of which is always negative. Since $h_i(z) > 1$ there is a Lipschitz constant $c_L = c_L(\bar{x})$ such that $|\ln h_i(z) - \ln h_i(\bar{x})| \le c_L |h_i(z) - h_i(\bar{x})|$ for all z such that $h_i(z) > 1$. So we have,

$$s_i(z,u,\bar{x}) \le \begin{cases} 0, & \text{if } |u_i - h_i(\bar{x})| \le |h_i(z) - h_i(\bar{x})| \\ 2c_L|u_i - h_i(\bar{x})|^2, & \text{if } |u_i - h_i(\bar{x})| > |h_i(z) - h_i(\bar{x})|. \end{cases}$$

In either case, we may just write

$$s_i(z,u,\bar{x}) \le 2c_L(u_{\max}^2 + h_{\max}^2)$$

whenever $i \in I_+$. Then, with $c_{\bar{x},u_{\max}} = p\max\{u_{\max}h_{\max} + 1/e + h_{\max}, 2c_L(u_{\max}^2 + h_{\max}^2)\}$,

$$\nabla V(z)\ f^*(z,u) \le c_{\bar{x},u_{\max}}, \quad \forall\, z \in \mathbb{R}^n_{>0}, \quad \forall\, u \in [0, u_{\max}]^p,$$

as we wanted to show. ∎

We also have the following $\mathbb{R}^n_{\ge 0}$-completeness result:

Corollary 2.15 Under the assumptions of Proposition 2.13, system (15) is $\mathbb{R}^n_{\ge 0}$-complete with input-value set $\mathbb{R}^p_{\ge 0}$.

Proof. Suppose that $u(\cdot)$ is an $\mathbb{R}^p_{\ge 0}$-valued input, and first pick any initial condition $z(0) \in \mathbb{R}^n_{>0}$. We already know, from Proposition 2.13, that system (15) is $\mathbb{R}^n_{>0}$-invariant with input-value set $\mathbb{R}^p_{\ge 0}$. Suppose that the maximal interval of existence would be $[0, t_{\max})$ with $t_{\max} < +\infty$. Put $u_{\max} = \text{ess.sup}.\{|u(t) - \bar{u}| : 0 \le t \le t_{\max}\}$.

From Lemma 2.14 we have that

$$\frac{d}{dt}V(z(t)) = \nabla V(z)\ f^*(z,u) \le c_{\bar{x},u_{\max}},\ \forall t < t_{\max}\,.$$

So $V(z(t)) \le V(z(0)) + c_{\bar{x},u_{\max}} t_{\max}$. Since V is proper (property (ii) of Definition 2.4), we conclude that $z(t)$ belongs to a compact subset of the state space $\mathbb{R}^n$, a contradiction with $t_{\max} < \infty$.

More generally, let $z(0) \in \mathbb{R}^n_{\ge 0}$, and suppose that $z(t)$ is defined on the maximal interval $[0, t_{\max})$. Let ξ_k, $k = 1, 2, \ldots$ be a sequence in $\mathbb{R}^n_{>0}$ with $\xi_k \to z(0)$. Denote by $z^k(t)$ the solution of the initial value problem $\dot{v} = f^*(v,u)$, $v(0) = \xi_k$, at time t. For each k, the argument above holds and so $V(z^k(t)) \le V(\xi_k) + c_{\bar{x},u_{\max}} t_{\max}$, for all $t < t_{\max}$. By continuity of solutions of differential equations on the initial conditions, we have for each $t \in [0, t_{\max})$, $V(z(t)) \le V(z(0)) + c_{\bar{x},u_{\max}} t_{\max}$. Therefore, if $t_{\max}$ is finite, we again conclude that $z(t)$ belongs to a compact subset of the state space $\mathbb{R}^n$, a contradiction. ∎

Proof of Theorem 3: Pick any initial states $x(0) \in \mathbb{R}^n_{>0}$ and $z(0) \in \mathbb{R}^n_{\ge 0}$ of the original system (1) and the observer, respectively. We let $w(\cdot) = (x(\cdot), z(\cdot))$ be the maximal trajectory of the composite system

$$\begin{aligned} \dot{x} &= f(x) \\ \dot{z} &= f(z) + C'(h(x) - h(z)), \end{aligned}$$

which we also write as $\dot{w} = g(w)$, with initial condition $(x(0), z(0))$. We need to show that $w(t) = (x(t), z(t))$ is defined for all $t > 0$, and $|z(t) - x(t)| \to 0$ as $t \to +\infty$.

Since we know that $x(t)$ is defined for all $t \ge 0$ and converges to some equilibrium $\bar{x}$ as $t \to +\infty$, we must prove that $z(t)$ is also defined for all $t \ge 0$ and converges to this same $\bar{x}$ as $t \to +\infty$.

Fix θ to be any constant such that $0 < \theta < 1$ and (since $x(t)$ converges) let T_0 be such that

$$t \geq T_0 \quad \Rightarrow |h_i(x(t)) - h_i(\bar{x})| \leq \frac{\theta}{2} h_i(\bar{x})$$

for all $i = 1, \ldots, p$. Let $\mathbb{U}_\theta$ be the set of vectors u such that $|u_i - h_i(\bar{x})| \leq \theta h_i(\bar{x})/2$.

Next, pick $T \geq T_0$ so large that the convergence $x(t) \to \bar{x}$ becomes exponential (such T exists, as shown in [19]). Then, for all $t \geq T$, $x(t)$ evolves in a compact set and, letting c_0 be a Lipschitz constant for the function h in this compact,

$$|h(x(t)) - h(\bar{x})| \leq c_0 |x(t) - \bar{x}| \leq c_0 c_1 e^{-c_2 t} |x(T) - \bar{x}|$$

where $c_1, c_2 > 0$ are constants that quantify the convergence of $x(t)$.

Corollary 2.15 shows that the solution $z(t)$ exists for all $t \geq 0$ and satisfies $z(t) \in \mathbb{R}^n_{\geq 0}$ and, in particular, by $\mathbb{R}^n_{>0}$-invariance, $z(t) \in \mathbb{R}^n_{>0}$ if $z(0) \in \mathbb{R}^n_{>0}$.

Claim 1. There exists a constant $d > 0$ such that for all $z(0) \in \mathbb{R}^n_{\geq 0}$ the trajectory z satisfies

$$V(z(t)) \leq V(z(T)) + d, \quad \forall t \geq T.$$

We first take the case $z(0) \in \mathbb{R}^n_{>0}$. Observe that the first part of Lemma 2.12 (where we may pick $\gamma(r) = c_3 r^2$) and the discussion above imply (with $c_4 = c_3 (c_0 c_1 |x(T) - \bar{x}|)^2$)

$$\frac{d}{dt} V(z(t)) \leq c_4 e^{-2c_2 t}$$

and integrating, $V(z(t)) \leq V(z(T)) + \frac{c_4}{2c_2} e^{-2c_2 T}$ for all $t \geq T$. We let $d = \frac{c_4}{2c_2} e^{-2c_2 T}$ (which indeed does not depend on z).

In the general case $z(0) \in \mathbb{R}^n_{\geq 0}$, we let ξ_k, $k = 1, 2, \ldots$ be a sequence in $\mathbb{R}^n_{>0}$ with $\xi_k \to z(0)$. Denote by $z^k(t)$ the solution of the initial value problem $\dot{v} = f^*(v, u)$, $v(0) = \xi_k$, at time t. For each k, $V(z^k(t)) \leq V(\xi_k) + d$, for all $t \geq T$. By continuity of solutions of differential equations on the initial conditions, taking limits we have $V(z(t)) \leq V(z(T)) + d$, for all $t \geq T$. The claim holds.

Claim 2. For each trajectory $z(\cdot)$, there exist functions $\beta \in \mathcal{KL}$, $\varphi \in \mathcal{K}_\infty$, such that

$$|z(t) - \bar{x}| \leq \beta(|z(T) - \bar{x}|, t) + \varphi(\|h(x) - h(\bar{x})\|).$$

for all $t \geq T$.

To see this, pick any trajectory $z(\cdot)$ and put

$$F = \{x : V(x) \leq \nu_2(|z(T)| + 1) + d\}$$

which is a compact set, by properness of V. Pick functions $\beta = \beta_F$ and $\varphi = \varphi_F$ as given by Definition 2.3.

First take the case $z(0) \in \mathbb{R}^n_{>0}$: claim 1 shows that $z(t) \in F$ for all $t \geq T$. Proposition 2.13, applied with $u(t) = h(x(t+T))$ and $\bar{u} = h(\bar{x})$ (note that $h(x(t)) \in \mathbb{U}_\theta$ for all $t \geq T$), immediately gives the ISS estimate with those functions β, φ.

Next take the more general case $z(0) \in \mathbb{R}^n_{\geq 0}$. Let ξ_k, $k = 1, 2, \ldots$ be a sequence in $\mathbb{R}^n_{>0}$ with $\xi_k \to z(0)$. Denote by $z^k(t)$ the solution of the initial value problem $\dot{v} = f^*(v, u)$, $v(0) = \xi_k$, at time t. Without loss of generality, by claim 1 we can conclude that $z^k(t) \in F$ for all k, for all $t \geq T$ (because $|z^k(T)| \leq |z(T)| + 1$, for all k). So for each k, Proposition 2.13 says that $|z^k(t) - \bar{x}| \leq \beta(|z^k(T) - \bar{x}|, t) + \varphi(\|h(x) - h(\bar{x})\|)$ holds for all $t \geq T$. Taking limits as $k \to +\infty$, shows that the same ISS estimate holds for $z(\cdot)$.

Now, given any $\varepsilon > 0$, let $T_1 \geq T$ be such that

$$\varphi(\|h(x) - h(\bar{x})\|_{T_1}) < \frac{\varepsilon}{2}$$

where $\|h(x) - h(\bar{x})\|_{T_1} = \text{ess. sup}\,.\{|h(x(t)) - h(\bar{x})| : t \geq T_1\}$ (such T_1 exists because the difference $|h(x(t)) - h(\bar{x})|$ goes to 0 as $t \to +\infty$).

Next, choose $T_2 \geq T_1$ such that

$$\beta(|z(T_1) - \bar{x}|, t) < \frac{\varepsilon}{2}, \quad \forall t \geq T_2.$$

Then, rechoosing T (if necessary) to be larger than T_2 we have that, for all $t \geq T$, $|z(t) - \bar{x}| \leq \varepsilon$. Therefore, $z(t) \to \bar{x}$ as $t \to +\infty$, as wanted. ∎

Remark 2.16 The observer (7) can be slightly modified to the form

$$\dot{z} = f(z) + C'W(h(x) - h(z)) \tag{17}$$

where W is any positive definite, diagonal $p \times p$ matrix. (See more details in [22].)

2.3 A Remark on Observation Noise

The ISS estimate obtained for the observer allows us to conclude that the observer is robust with respect to small observation noise. We sketch this next. If the input to the observer is $u(t) = h(x(t)) + n(t)$ instead of $h(x(t))$, then the same conclusions hold regarding global existence of trajectories (at least, provided that $h(x(t)) + n(t)$ is nonnegative). In addition, for large t (so $h(x(t)) \approx h(\bar{x}) = \bar{u}$), the term $\varphi(\|u - \bar{u}\|)$ in the ISS estimate becomes approximately $\varphi(\|n\|)$; thus one obtains an asymptotic estimate on $z(t)$ which is a $\mathcal{K}$ function of the noise level, and is in particular small when n is small in magnitude.

3 Generalization to Systems with "Multiple Linkage Classes"

We now look at the more general case of a system with vector field of the form (2) but where the matrix $A = \text{diag}(A_1, \ldots, A_L)$ is block diagonal, and each A_s (of size m_s) is itself irreducible and has nonnegative entries (or, at least, there exists a permutation matrix, P, such that PAP^{-1} has that diagonal form). The matrix B is also partitioned into $B = [B_1 \cdots B_L]$, where each B_s is of dimension $n \times m_s$ and, since the assumption that B has full rank is still valid, each B_s itself has full rank, m_s ($m_1 + \cdots + m_L = m$). The system (1) can be written as $\dot{x} = f_1(x) + \cdots + f_L(x)$ where each $f_s(x)$ is computed according to formula (2) using A_s and B_s.

The number L is called the number of "linkage classes" and denotes the (smallest) number of connected components of the incidence graph $G(A)$. $G(A)$ is the graph whose nodes are the integers $\{1, \ldots, m\}$ and for which there is an edge $j \to i$, iff $a_{ij} > 0$.

To each connected component there corresponds a space

$$\mathcal{D}_s = \text{span}\ \{b_i - b_j : b_i, b_j \text{ are columns of } B_s\}.$$

The assumptions on the B_s imply that each space $\mathcal{D}_s$ has dimension $m_s - 1$,

$$\mathcal{D} = \mathcal{D}_1 \oplus \cdots \oplus \mathcal{D}_L, \tag{18}$$

(direct sum), and so $\dim\mathcal{D} = (m_1 - 1) + \cdots + (m_L - 1) = m - L$.

Example 3.1 To illustrate this structure, consider a general enzymatic mechanism with uncompetitive inhibitor, consisting of one enzyme E, one substrate S, one product P and an uncompetitive inhibitor I (Q and R are intermediate complexes):

$$\begin{array}{ll} S+E \stackrel{k_1}{\to} Q, & Q \stackrel{k_{-1}}{\to} S+E, \\ Q \stackrel{k_2}{\to} P+E, & P+E \stackrel{k_{-2}}{\to} Q, \\ Q+I \stackrel{k_3}{\to} R, & R \stackrel{k_{-3}}{\to} Q+I. \end{array}$$

There are two linkage classes, $L = 2$:

(i) the first class consisting of the complexes $S+E$, $P+E$ and Q;

(ii) the second class consisting of the complexes $Q+I$ and R.

For class (i), $S+E$, $P+E$ and Q are the three nodes of $G(A_1)$ and, due to the reversibility of the reactions, it is possible to "connect" any two of these nodes through a path in the graph. The same is true for class (ii).

Notice that the same complex, e.g., Q in the above example, may belong to different connected components (otherwise the problem could be reduced to two completely independent "single linkage" problems).

In [15] it is shown that this system does not admit boundary equilibria in any positive class.

Let $x = (S, P, Q, R, E, I)'$. Then $B = [B_1 \;\; B_2]$ and $A = \operatorname{diag}(A_1, A_2)$:

$$B_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

and

$$A_1 = \begin{bmatrix} 0 & 0 & k_{-1} \\ 0 & 0 & k_2 \\ k_1 & k_{-2} & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & k_{-3} \\ k_3 & 0 \end{bmatrix}.$$

The space $\mathcal{D}$ is given by $\mathcal{D}_1 + \mathcal{D}_2$:

$$\mathcal{D}_1 = \text{span}\,\{(1,-1,0,0,0,0)', (1,0,-1,0,1,0)'\}$$
$$\mathcal{D}_2 = \text{span}\,\{(0,0,1,-1,0,1)'\}$$

and the function f is given by $f_1 + f_2$:

$$f_1(x) = \begin{bmatrix} k_{-1}Q - k_1 SE \\ -k_{-2}PE + k_2 Q \\ -k_{-1}Q + k_1 SE + k_{-2}PE - k_2 Q \\ 0 \\ k_{-1}Q - k_1 SE - k_{-2}PE + k_2 Q \\ 0 \end{bmatrix}, \quad f_2(x) = \begin{bmatrix} 0 \\ 0 \\ k_{-3}R - k_3 QI \\ -k_{-3}R + k_3 QI \\ 0 \\ k_{-3}R - k_3 QI \end{bmatrix}.$$

For a general "multiple linkage" system we may consider output maps of the same form as before, i.e., monomials in the state variables, as in (3). These monomials may include any of the variables "x_i", and they have the same interpretation as before: either representing the concentration of some of the substances (as in the case of x_1), or being proportional to some reaction rate (as in the case of $x_1^3 x_4$, etc.).

The necessary and sufficient detectability condition given in Theorem 2 is still valid (Theorem 1 generalizes, as sketched in [19]). The main fact to verify is that (5) still holds for the general case. But, from [19], we know that for an interior point $\bar{x}$, $f(\bar{x}) = 0$ if and only if $f_s(\bar{x}) = 0$ for each $s = 1, \ldots, L$. (That is, $\bar{x}$ is an equilibrium of the entire system if and only if it is an equilibrium of every system $\dot{x} = f_s(x)$; this nontrivial fact follows from the block irreducibility property.) Then, for each s, (5) says that, if $\bar{x} \in E_+$, then, for any $\bar{z} \in \mathbb{R}^n_{>0}$,

$$\rho(\bar{x}) - \rho(\bar{z}) \in \mathcal{D}_s^{\perp} \iff \bar{z} \in E_+^s\,,$$

where E_+^s is the set of interior equilibria of $\dot{x} = f_s(x)$ (so $E_+ = E_+^1 \cap \ldots \cap E_+^L$). Equivalently, if $\bar{x} \in E_+$, then, for any $\bar{z} \in \mathbb{R}^n_{>0}$,

$$\rho(\bar{x}) - \rho(\bar{z}) \in \mathcal{D}_1^{\perp} \cap \ldots \cap \mathcal{D}_L^{\perp} = \mathcal{D}^{\perp} \iff \bar{z} \in E_+\,.$$

Thus, Equivalence (5) holds for $L > 1$ as well.

It is also true that $E_0 = E_0^1 \cap \ldots \cap E_0^L$, since each $x \in E_0^s$ is characterized by $x_1^{b_{1j}} x_2^{b_{2j}} \ldots x_n^{b_{nj}} = 0$ for all j such that b_j that is a column of B_s (see Lemma 6.2 in [19]), and we have $B = [B_1 \ \cdots \ B_L]$.

So, a general system $\dot{x} = f_1(x) + \cdots + f_L(x)$, $y = h(x)$ is detectable if and only if the matrix C of the (exponents of the) output map satisfies either condition (c) or (d) in Theorem 2 as well as $h(\bar{x}) \neq h(\bar{z})$ whenever $\bar{x} \in E_+$ and $\bar{z} \in E_0$. Note that more "linkage classes" mean more information is needed in order for the system to be detectable. For a n-dimensional system, the space $\mathcal{D}$ has dimension $m - L$ and detectability implies that the matrix C must have rank $p = n - (m - L)$. As we have seen, a single linkage class requires $p = n - m + 1$, whereas multiple linkage classes require $p = n - m + L$.

In the example above, for detectability of $\dot{x} = f_1(x) + f_2(x)$, $y = h(x)$, C will need to have rank 3. The following output would be a suitable choice: $h(x) = (S^2Q,\, RI^2,\, E)'$.

For a detectable "multiple linkage" system, an observer for $\dot{x} = f_1(x) + \cdots + f_L(x)$ is again of the form

$$\dot{z} = f_1(z) + \cdots + f_L(z) + C'(h(x) - h(z)). \tag{19}$$

To prove convergence of the observer, one may use Proposition 2.13, Lemma 2.12 and Corollary 2.15 as before, after checking some points.

The $\mathbb{R}^n_{>0}$-invariance argument is unchanged since the particular forms of A and B still imply that (16) and the corresponding conclusions hold.

To see that Lemma 2.12 holds, we must analyse the term

$$\nabla V(z)\ f(z) = \nabla V(z)\ f_1(z) + \cdots + \nabla V(z)\ f_L(z).$$

For each $s = 1, \ldots, L$, estimate (13) holds, so

$$\nabla V(z)\ f_s(z) = \langle \rho(z) - \rho(\bar{x}), f_s(z) \rangle \leq -\kappa(A_s)\, c(\bar{x}) \sum_{i,j \dashv B_s} (e^{-\pi_i} - e^{-\pi_j})^2$$

where $i, j \dashv B_s$ means that only the columns b_i, b_j of B_s are present in the sum. The right hand side of this inequality vanishes whenever $z \in E_+^s \cup E_0^s$. Hence

$$\nabla V(z)\ f(z) \ \leq\ -\sum_{s=1}^{L} \kappa(A_s)\, c(\bar{x}) \sum_{i,j \dashv B_s} (e^{-\pi_i} - e^{-\pi_j})^2,$$

where the right hand side vanishes only if $z \in (E_+^1 \cap \cdots \cap E_+^L) \cup (E_0^1 \cap \cdots \cap E_0^L)$, i.e., only if $z \in E_+ \cup E_0$. The rest of the proof of Lemma 2.12 is unchanged, since the form of the observer is the same as before. Thus, given any compact set $F \subset \mathbb{R}^n_{\geq 0}$, the function V satisfies an estimate

$$\nabla V(z)\ f^*(z,u) \leq -\alpha(|z-\bar{x}|) + c_3|u-h(\bar{x})|^2$$

for every $z \in F \cap \mathbb{R}^n_{>0}$ and $u \in \mathbb{U}_\theta$, where $\alpha \in \mathcal{K}_\infty$ and the set of inputs $\mathbb{U}_\theta$, and the constant c_3 are as in the Lemma. Corollary 2.15 is still valid, as follows from this estimate.

Finally, since the trajectory of the original system will still converge exponentially after a given time, the proof of Theorem 3 is the same as before.

4 Some Simulations

4.1 Robustness of the Observers

To numerically test robustness of our observers, we carried out some simulations to explore their responses in two cases: existence of noise in the output measurements and unknown inputs acting in the system we wish to observe.

As a working example we choose $\dot{x} = f(x)$ to be the network proposed by McKeithan in [13] and took the output to be $h(x) = (x_1x_2^2, x_1x_4)'$, as in Example 2.1. The constants were taken to be: $k = 6, k_3 = 0.5, k_4 = 7, \beta_3 = 1$. In all simulations in this section, the initial conditions for system (1) were $x(0) = (1,3,3,2)'$ and for the observers we took $z(0) = (2,25,20,1)'$.

In one simulation, white noise was added to the outputs, so that the equation for the main observer becomes

$$\dot{z} = f(z) + C'(h(x(t)) + n(t) - h(z))$$

and $n(t)$ is an $\mathbb{R}^2$-valued vector white noise. In view of the Section 2.3, solutions of this system exist for all $t \geq 0$, provided that $h(x(t)) + n(t)$ is nonnegative; and the observer should provide estimates close to the true state as long as the magnitude of $n(t)$ is small. Thus we chose $x(0)$ so that $x_i(t) \geq 2$ and $|n(t)| < 2$ for all t and all i.

In Figure 1, left side, we can see that the trajectories of the four coordinates of our main observer exhibit small magnitude perturbations as a result of the output noise.

In another simulation, the model (1) was perturbed by a disturbance consisting of a periodic signal and two "delta" functions. The equation for the model is

$$\dot{x} = f(x) + d(t)$$

where $d(t) = (d_1(t), 0, 0, 0)'$ and $d_1(t) = 0.3\sin(t/4) + 2\,[15 < t < 16] - 4\,[35 < t < 36]$ ($j(t) = a\,[m < t < M]$ means that $j(t) = a$ if $m < t < M$ and $j(t) = 0$ otherwise). The function d_1 was chosen so that (for the same initial condition $x(0)$ as above) $x_i(t) > 0$ for all i and all t. (The observer is still the one for the nominal system, with no disturbance.) Note how our first observer catches-up after the "delta" disturbance, and also tracks (with a small lag) the limit cycle into which the observed system trajectories converge (figure 1, at right).

4.2 Comparison with Standard Observers

In the chemical reactor literature, observers are typically constructed using an extended Kalman filter (EKF), or, less often, a Luenberger type observer (Lbg) for a linearized system. Neither of these approaches is guaranteed to work for nonlinear systems, but it is often the case that

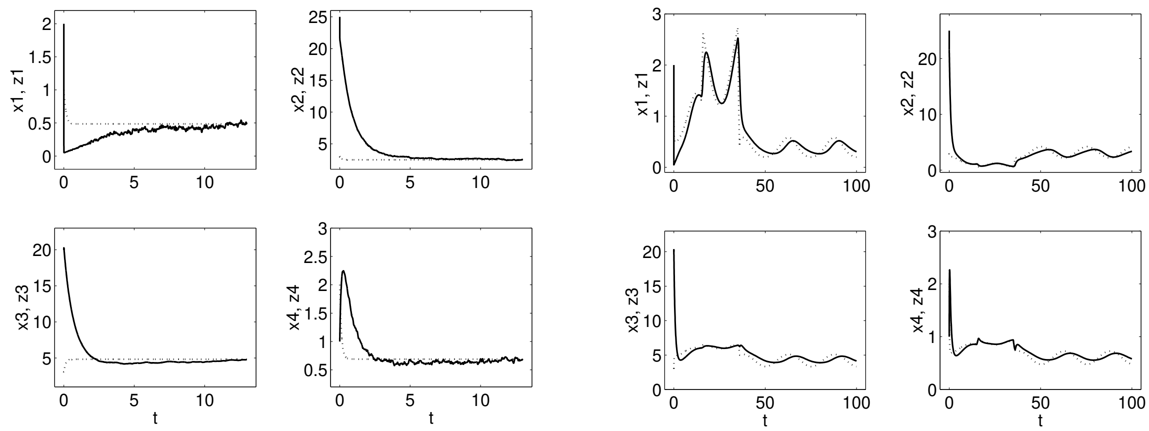

Figure 1: Left: the effect of output noise. Right: the effect of an unknown disturbance acting on the the system. The trajectories of the system (dotted line) and our observer (solid line) are shown against time.

they perform adequately in specific examples, and hence their practical success. In the second part of this section, we compare the performance of our main observer with those of an EKF observer and of a Lbg observer. Our purpose in doing so is to illustrate that, even for very simple examples, these standard techniqes can fail in a major way, while our observer is, as predicted by the theory, convergent to the right estimate.

These two standard constructions both have the form

$$\dot{z} = f(z) + L(z)(h(x) - h(z)),$$

where the gain $L(z)$ is to be found in such a way that (at least locally) $|x(t) - z(t)| \to 0$ as $t \to +\infty$. Let us briefly review these constructions.

We consider the linearized dynamics of the error $e = x - z$, around the origin $e = 0$:

$$\dot{e} \doteq [F(z(t)) - L(z(t))H(z(t))]e,$$

where

$$F(z) = Df(e+z)|_{e=0} \quad \text{and} \quad H(z) = Dh(e+z)|_{e=0}$$

are the Jacobians of f and h evaluated at the point z.

The gain $L(z)$ for a (continuous) extended Kalman filter is given by

$$L(z(t)) = P(t)H'(z(t))R^{-1},$$

where P is a symmetric positive definite solution to the following Riccati differential equation:

$$\dot{P} = -PH'R^{-1}HP + FP + PF' + Q,$$

and R and Q are two positive definite cost matrices.

A Luenberger type observer is obtained by finding a constant gain L such that the matrix $F(\bar{x}) - LH(\bar{x})$ is Hurwitz. A linearized error equation can also be written as:

$$\dot{e} \doteq [F(x(t)) - LH(x(t))]e$$

(note that the time-dependence of F and H is given in terms of a dependence on the trajectory of the system itself, instead of on the trajectory of the observer). It can be shown that, for initial conditions $x(0)$ and $z(0)$ sufficiently close to $\bar{x}$, this error is asymptotically stable with respect to the origin. (Note that Luenberger observers, at least in their standard formulation, are not a reasonable choice for our example, since their design assumes the knowledge of the equilibrium point around which we are observing. For multi-stable systems such as ours, it makes little sense to assume that this equilibrium is known – in fact, knowing this equilibrium amounts to solving the detectability problem. However, we can still study the behavior of a Luenberger observer, especially since we will show that it does not work even when this additional information is provided.)

With the same working example as above, and same initial conditions $x(0) = (1, 3, 3, 2)'$ for the system, we chose the gain for the Luenberger observer to be the 4×2 matrix with entries $l_{11} = -1$, $l_{42} = 1$ and all others equal to zero. A computation shows that the matrix $F(\bar{x}) - LH(\bar{x})$ is indeed Hurwitz with eigenvalues $-9.8 \pm 3.7i$ and $-0.48 \pm 0.08i$. To solve the Ricatti differential equation, we took $R = I_{2\times 2}$ and $Q = I_{4\times 4}$ (the identity matrices in $\mathbb{R}^2$ and $\mathbb{R}^4$, respectively), and the initial condition $P(0) = I_{4\times 4}$.

The simulations show that both EKF and Lbg converge provided that $z(0)$ is in a sufficiently small neighborhood of $\bar{x}$, but they may diverge when $z(0)$ is away from $\bar{x}$. In Figure 2 the behavior of the three observers is shown, and the performance of EKF and Luenberger are clearly inferior to that of our observer.

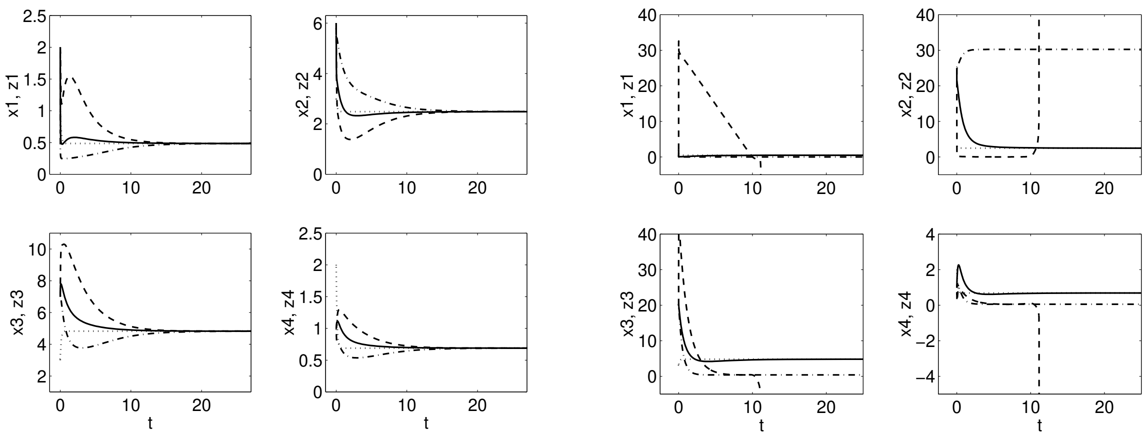

Figure 2: Comparison with standard observers. The trajectories of the system (dotted line), our observer (solid line), a Lbg(dashed line) and an EKF (dash-dotted line) are shown against time. Left: local convergence with $z(0) = (2, 6, 7, 1)'$. Right: Lbg and EKF diverge for $z(0) = (2, 25, 20, 1)'$.

A Appendix

A.1 Some Simple Facts Concerning Invariance

Consider the scalar initial value problem

$$\begin{aligned} \dot{x} &= F(t, x) \\ x(0) &= x_0 \end{aligned} \tag{20}$$

where the function F is assumed to have domain $\mathcal{I} \times \mathcal{X}$, where $\mathcal{X}$ is an open subset of $\mathbb{R}$ and $\mathcal{I} = [0, +\infty)$. Let F be locally Lipschitz in x and measurable in t, more precisely,

(i) For each $a \in \mathcal{X}$ there exists a real number r_a and a locally integrable function $\alpha : \mathbb{R} \to [0, +\infty)$ such that the ball of radius r_a centered at a, $B_{r_a}(a) \subset \mathcal{X}$ and

$$\|F(t,x) - F(t,y)\| \leq \alpha(t)\|x - y\|$$

for each $t \in \mathbb{R}$ and $x, y \in B_{r_a}(a)$.

(ii) For each fixed $a \in \mathcal{X}$, the function $g : \mathcal{I} \to \mathcal{X}$ given by $g(t) := F(t, a)$ is measurable.

For each $x_0 \in \mathcal{X}$ let $J = J_{x_0}$ be the maximal interval of existence of solutions of (20) in forward time. This is an interval of the form $[0, t_{\max})$ with $0 < t_{\max} \leq +\infty$.

Using a standard comparison principle (as done for instance in [19]), we have:

Lemma A.1 Consider system (20) with domain $\mathcal{X} = \mathbb{R}$ and assume further that,

$$x = 0 \quad \Rightarrow \quad F(t, 0) \geq 0 \quad \forall t \in \mathcal{I}.$$

Assume also that the initial condition is positive: $x_0 > 0$. Then $x(t) > 0 \ \ \forall t \in J$, i.e., the solution of (20) remains positive for all times in J.

References

[1] Angeli D, Sontag ED, Wang Y. Further equivalences and semiglobal versions of integral input to state stability. Dynamics and Control 2000; 10: 127-149.

[2] Bastin G, van Impe JF. Nonlinear and adaptive control in biotechnology: A tutorial. European J. Control 1995; 1: 1-37.

[3] Bastin G, Dochain D. On-line Estimation and Adaptive Control of Bioreactors. Elsevier, Amsterdam Oxford New York Tokyo 1990.

[4] Bernstein D, Bhat SJ. Nonnegativity, reducibility, and semistability of mass action kinetics. In: Proc. 1999 IEEE Conference on Decision and Control. IEEE Publications, Dec. 1999, pp 2206-2211.

[5] Dochain D, Buyl E, Bastin G. Experimental validation of a methodology for on-line state estimation in bioreactors. In: Fish NM, Fox RI, Thornhill NF (eds). Computer Applications in Fermentation Technology: Modelling and Control of Biotechnological Processes. Elsevier Applied Science, London New York, 1988, pp 187-194.

[6] Feinberg M. Chemical reaction network structure and the stabiliy of complex isothermal reactors - I. The deficiency zero and deficiency one theorems. Review Article 25, Chemical Engr. Sci. 1987; 42: 2229-2268.

[7] Feinberg M. The existence and uniqueness of steady states for a class of chemical reaction networks. Archive for Rational Mechanics and Analysis 1995; 132: 311-370.

[8] Feinberg M. Lectures on Chemical Reaction Networks. 4.5 out of 9 lectures delivered at the Mathematics Research Center, University of Wisconsin, Fall 1979.

[9] Feinberg M. Mathematical aspects of mass action kinetics. In: Lapidus L, Amundson N (eds). Chemical Reactor Theory: A Review. Prentice-Hall, Englewood Cliffs. 1977, pp 1-78.

[10] Horn FJM, Jackson R. General mass action kinetics. Arch Rational Mech Anal 1972; 49: 81-116.

[11] Horn, FJM. The dynamics of open reaction systems. In: Mathematical aspects of chemical and biochemical problems and quantum chemistry. Proc SIAM-AMS Sympos Appl Math, New York 1974, pp 125-137. SIAM-AMS Proceedings, Vol. VIII, Amer. Math. Soc., Providence, 1974.

[12] Krichman M, Sontag ED, Wang Y. Input-output-to-state stability. SIAM J Control 2001; 39: 1874-1928.

[13] McKeithan TW. Kinetic proofreading in T-cell receptor signal transduction. Proc Natl Acad Sci USA 1995; 92: 5042-5046.

[14] Pous NM, Rajab A, Flaus A, Engasser JM, Cheruy A. Comparison of estimation methods for biotechnological processes. Chem Engr Sci 1988; 43: 1909-1914.

[15] Siegel D, MacLean D. Global Stability of Complex Balanced Mechanisms. Preprint.

[16] Sontag ED. Mathematical Control Theory: Deterministic Finite Dimensional Systems. 2^{nd} edn. Springer-Verlag New York 1998.

[17] Sontag ED. Smooth Stabilization Implies Coprime Factorization. IEEE Trans Autom Control, vol. 34, 1989; 4: 435-443.

[18] Sontag ED. Remarks on stabilization and input-to-state stability. In: Proc IEEE Conf Decision and Control, Tampa. IEEE Publications, Dec. 1989, pp 1376-1378.

[19] Sontag ED. Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction. IEEE Trans Autom Control 2001; 46: 1028-1047.

[20] Sontag ED, Wang Y. On characterizations of the input-to-state stability property. Systems and Control letters 1995; 24: 351-359.

[21] Sontag ED, Wang Y. Output-to-state stability and detectability of nonlinear systems. Systems and Control Letters 1997; 29: 279-290.

[22] See more details in the the preprint math.OC/0012130 (2000) in Los Alamos Archive (http:arXiv.org).