OBSERVER DESIGN FOR A CLASS OF NONLINEAR SYSTEMS, WITH APPLICATIONS TO BIOCHEMICAL NETWORKS

BY MADALENA CABRAL FERREIRA CHAVES

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Mathematics

Written under the direction of
Eduardo D. Sontag
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey
May, 2003

ABSTRACT OF THE DISSERTATION

Observer Design for a Class of Nonlinear Systems, with Applications to Biochemical Networks

by Madalena Cabral Ferreira Chaves
Dissertation Director: Eduardo D. Sontag

The design of globally convergent observers, for a given class of nonlinear systems, is the main objective of this work. We start by characterizing this class of nonlinear systems, which provide a model for chemical reaction networks of the Feinberg-Horn-Jackson zero deficiency type, and analyze its dynamics and stability properties. For the system with outputs, a necessary and sufficient condition for detectability is established.

Two explicit observer constructions are proposed in this work, both of which are shown to be globally convergent. The state-estimator design is based on input to state stability (ISS) estimates for the observer driven by the measurements, and the ISS framework together with Lyapunov techniques provide stability and other desirable robustness properties, such as insuring global convergence under sampled outputs, and robustness with respect to noise in the measurements. A notion of robustness with respect to perturbations in the parameters of the system is formulated following ISS ideas, and it is shown that the observers also exhibit robustness in this sense.

Finally, an experiment is reported in which data pertaining to a reversible chemical reaction was collected through an NMR spectrometer. The main observer is implemented and its performance validated against these data. In addition, several numerical tests explore and illustrate the various robustness properties, and a comparison with standard observer constructions (extended Kalman filters, and extended Luenberger observers) is provided.

Acknowledgements

I have very much enjoyed these years spent at Rutgers. It has been a rich and significant experience for me. Throughout graduate courses and research work at the Department of Mathematics, I have had the opportunity to learn about and discuss many topics. I am especially grateful to (the late) Professor Tilla Weinstein and to Professor Peter Landweber (who, as Graduate Director, admitted me to the program), for their help and advice through my first years as a graduate student. I am also very grateful to Professor Stephen Greenfield, for his helpful suggestions and encouragement, and to the members of my committee: Professor Daniel Ocone, Professor Héctor Sussmann and Professor Zoran Gajic, for their useful comments.

Many special thanks are due to Professor Larry Romsfeld and his group (in particular, Justin Mills) at the Department of Chemistry and Chemical Biology, for their collaboration in conducting an experiment that enabled me to test and validate some of my theoretical results.

In the last stage of my research work, I have benefited from many enlightening discussions with Doctor Robert Dinerstein, from Aventis Pharmaceuticals. He has shown me many motivating biological systems related to my work, and pointed out various leading problems and directions to follow. I am most grateful to Doctor Dinerstein for many discussions and for his interest in my work.

Above all, I infinitely thank my advisor, Professor Eduardo Sontag, for his constant support, patience and generosity! It has been a great pleasure and a privilege to work with Professor Sontag. Throughout our many meetings and discussions, I have learned about mathematics and control theory, and have shared Professor Sontag's enthusiasm and insight. I am most grateful for his encouragement and also for his suggestion of such an up-to-date research topic for my thesis, which opens up many interesting and promising directions for future work.

Finally, I wish to acknowledge several sources of support. I thank Fundação para a Ciência e a Tecnologia, and Fundação Calouste Gulbenkian, both in Lisbon, Portugal: their invaluable financial support allowed me to be involved full time in my thesis research work.

I also wish to thank Aventis Pharmaceuticals and the BioMaPS Institute at Rutgers University, for their additional financial support.

Dedication

Para o Miguel

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

Given a system Σ, with a set of measurements, or outputs y, a typical goal is to control, regulate, or monitor the dynamics of the system. To achieve this goal, a full knowledge of the internal state variables of Σ is often desired.

A *state-estimator*, or *observer*, for Σ, is another system that takes as inputs the measurements y and provides estimates of the state variables of Σ. Given general dynamical systems, the existence and design of observers is one of the most challenging problems in control theory.

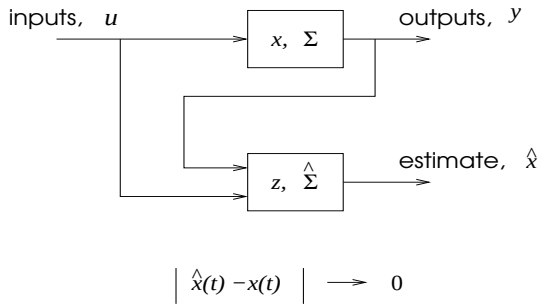


Figure 1.1: A schematic state-estimator.

In approaching the problem of designing a state-estimator, the first question to be asked should be: given the information available on the system, consisting of the dynamics and the measurements, is it reasonable to expect that, through some appropriate "method", a "good estimate" of the state variables may eventually be obtained? This property reflects a necessary condition for the existence of an observer, and it is usually called *detectability*. (It is easy to find examples of systems for which no observer may be expected to give an estimate of the system's internal states, because there is not enough information available: as a simple illustration, consider $\dot{x}_1 = 0$, $\dot{x}_2 = 1$ with outputs $y = x_2$.)

If the system is detectable, then it makes sense to address a second question: that of *existence and construction* of an observer, and then analyzing its convergence.

Thirdly, the rapidity of *convergence* and various *robustness properties* (the performance of the observer when there are noisy measurements, or small perturbations to the system's parameters, etc.), are interesting and desirable from the point of view of practical implementation.

Establishing generally applicable detectability conditions, and then investigating the question of existence and construction of an observer, and proving its convergence, is

not at all a trivial task and is an open and active area of research. For a discussion of detectability and its possible characterizations, see for instance Krener [31], and more recently in the style of *output-to-state stability* (OSS) see Sontag and Wang [51], or in the style of *input-output-to-state stability* (IOSS) see Krichman, Sontag and Wang [34].

There are many contributions to the question of observers for nonlinear systems. Many examples and several theoretical approaches can be found in the literature (just to cite a few examples: [4], [7], [9], [11], [23], [32], [33], [54], [56], [57]). We would like to, very briefly, recall some of the most notable and relevant observer constructions, and hopefuly motivate the work to be developed in the following chapters.

For linear systems, the work of Luenberger provides a complete and very beautiful solution to the observer problem: the system

$$\dot{x} = Ax + Bu, \qquad y = Cx \tag{1.1}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{p\times n}$ is *detectable* if

$$Cx(t) \equiv 0 \;\Rightarrow\; x(t) \to 0,$$

for all solutions $x(\cdot)$ of the system without inputs, $\dot{x} = Ax$. Detectability of the system is equivalent to

$$\text{rank} \begin{pmatrix} \lambda I - A \\ C \end{pmatrix} = n, \quad \text{for all } \lambda \in \mathbb{C},\ \text{Re } \lambda \geq 0,$$

and we also say that the pair (A, C) is detectable.

The system (1.1) admits an observer if and only if it is detectable and, in this case, there is always an observer of the form

$$\dot{z} = Az + Bu - L(y - Cz),$$

where L is any matrix such that $A + LC$ is a Hurwitz matrix (i.e., all its eigenvalues have negative real parts). Under the detectability condition, such a matrix L always exists, as a consequence of the Pole-Shifting Theorem. Then it is easy to see that $|z(t) - x(t)| \to 0$ as $t \to +\infty$, since the error $e(t) = z(t) - x(t)$ satisfies the equation

$$\dot{e} = (A + LC)e.$$

(For a reference see, for instance, [46].)

For nonlinear systems, there is no straighforward or general solution to the problem of designing state-estimators, even though the idea of the Luenberger observer may be explored in several ways, i.e., take a copy of the system and add a correction term, preferably using the measurements/data available. The difficulty now lies in proving the convergence of the observer since, for general nonlinear systems, the error dynamics cannot be defined as an autonomous differential equation, and even if this is so, it isn't usually globally asymptotically stable.

To overcome this obstacle, several different approaches exist, typically involving regularity assumptions, or restrictions to particular classes of systems. Some of the directions pursued in the observer design literature include: the idea of finding a change of state coordinates that tranforms the nonlinear system into a linear one and then, for the new system, constructing a Luenberger-like observer ([9], [32], [33], [23], [57]); establishing regularity conditions on the systems, in order to obtain an observer with a desired error convergence ([11], [56]); establishing general Lyapunov like conditions for the existence of an observer ([54]); making use of particular characteristics of the system (such as bounds on the nonlinearities as in [4]). An alternative, more empirical, approach is to construct an observer based on the linearization of the system along its trajectory: this design is known as an extended Kalman filter, since it is a deterministic, continuous-time, adaptation of the work of Kalman on the filtering problem (there are many references on deterministic Kalman filtering, but see for instance [35], [46]). Extended Kalman filters are not guaranteed to work for nonlinear systems (one expects that locally they should converge, but there is no proof of this fact, in general), but in many practical cases they have been very successful.

The observers based on a change of state coordinates offer perhaps some of the more satisfactory theoretical approaches to observer design for general systems, and have received some attention, so we will next briefly review some of those constructions.

In many approaches, the existence and construction of an observer is based on the property of *observability* of the system (rather than detectability). Roughly, observability is related to the possibility of actual reconstruction of the internal states of the system at any time t, from (past and future) outputs, while detectability is related to the possibility of producing an estimate of the internal state of the system at any time t from (past) outputs — such estimates should become closer to the correct values for large t.

More precisely, the system Σ is *observable* if any two different states x_1 and x_2 are distinguishable, i.e., if there exists a time T and a control w (admissible for both states) such that: applying the control w to system Σ when its internal state is either x_1 or x_2, at $t = T_0$, yields distinct outputs at the (later) time $T \geq T_0$.

The concept of observability has long been explored and studied by many authors, including Hermann and Krener ([26]), Sontag ([42], [43]), Sussmann ([53]), Gauthier and Kupka ([24]).

However, in the present work, it is not our purpose to pursue the discussion of observability. We will adopt a notion of detectability (to be discussed in Chapter 3) which is, in some sense, similar to the one for linear systems.

In order to present some examples of state-estimators, we will introduce a simple characterization of observability. Consider the general system with outputs (but no inputs):

$$\dot{\xi} = f(\xi), \quad y = h(\xi). \tag{1.2}$$

where $\xi \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}^p$. For simplicity, consider the system with

single outputs ($p = 1$). This system is said to be *(locally) observable* at ξ_0 if the matrix

$$O_{\mathrm{N}}(\xi) = \begin{pmatrix} \nabla\, h(\xi) \\ \nabla\, L_f h(\xi) \\ \vdots \\ \nabla\, L_f^{n-1} h(\xi) \end{pmatrix} \tag{1.3}$$

satisfies the *observability rank condition at* ξ_0:

$$\text{rank } O_{\mathrm{N}}(\xi_0) = n. \tag{1.4}$$

The notation is as follows: $\nabla\, h = (\partial h/\partial \xi_1, \ldots, \partial h/\partial \xi_n)$ and $L_f\, h = \langle \nabla h, f \rangle$. $O_{\mathrm{N}}(\xi)$ is usually called the *observability matrix* (for more details see, for instance, [26] and [30]).

Essentially, this condition means that

$$\phi(\xi) = \begin{pmatrix} h(\xi) \\ L_f h(\xi) \\ \vdots \\ L_f^{n-1} h(\xi) \end{pmatrix}$$

is a valid change of coordinates in a neighborhood of ξ_0, $\mathcal{N}(\xi_0)$.

Note that, in the linear case, the matrix O_{N} reduces to the well known *observability matrix*

$$O(A, C) := \begin{pmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{pmatrix},$$

and if the rank condition holds then one says that *the pair* (A, C) *is observable.* For linear systems, the concept of detectability is equivalent to that of *asymptotic observability.*

In [32], Krener and Isidori (1983) suggested an observer that applies to nonlinear systems of a very special form: the observer is based on the existence of a nonlinear change of state coordinates, that transforms the original nonlinear system into a linear system with nonlinear output injection. This approach was later followed by Krener and Respondek, in [33], who generalized this theory to systems with multiple inputs and multiple outputs. The same idea was also developed independently by Bestle and Zeitz in [9].

Given the nonlinear system with outputs (1.2) one would like to find out if it could possibly have arised from a linear system with *nonlinear output injection*

$$\dot{x} = Ax + \varphi(y), \quad y = Cx \tag{1.5}$$

followed by a nonlinear change of state coordinates

$$\xi = \xi(x),$$

and where the pair (A, C) is observable. If (1.2) happens to be the result of such a process, and if one can find out the form of the functions φ and $\xi = \xi(x)$, then the problem of observing (1.2) is transformed into the problem of observing (1.5) and an observer can be easily constructed for the latter system, as in the Luenberger case for linear systems:

$$\dot{z} = Az + \varphi(y) - L(y - Cz),$$

where L is such that $A + LC$ is a Hurwitz matrix.

In [32], it is shown that, under the assumptions:

1. for a point of interest ξ_0, and a neighborhood $\mathcal{N}(\xi_0)$, the observability rank condition (1.4) holds at each $\xi \in \mathcal{N}$;

2. the vector field defined by

$$L_g\, L_f^k\, h = \begin{cases} 0, & 0 \leq k < n-1, \\ 1, & k = n-1 \end{cases}$$

satisfies

$$[g, \mathrm{ad}_f^k\, g] = 0, \quad k = 1, 3, \ldots, 2n-3,$$

(where $[f, g]$ is the Lie bracket of f, g and $\mathrm{ad}_f\, g := [f, g]$, $\mathrm{ad}_f^2\, g = [f, [f, g]]$, etc.)

there exists a (local) change of state coordinates, that tranforms the nonlinear system (1.2) to the form (1.5).

A different approach was proposed by Gauthier, Hammouri and Othman (1992), in the paper [23]. The authors proposed an observer which can be viewed as a Luenberger type observer, and also relies on the existence of a suitable change of variables. For simplicitly, consider the single output case, $p = 1$. The assumptions are

1. the map $F : \Omega \to \mathbb{R}^n$ given by $\xi \mapsto \phi(\xi)$ is a diffeomorphism from Ω onto $F(\Omega)$, where Ω is an open, connected, relatively compact subset of $\mathbb{R}^n$. Thus the system (1.2) can be transformed into

$$\dot{x} = \begin{pmatrix} x_2 \\ \vdots \\ x_n \\ \varphi(x) \end{pmatrix} := \tilde{F}(x), \qquad y = x_1, \tag{1.6}$$

where φ is given by $\varphi(x) = L_f^n\, h\, (\phi^{-1}(x))$;

2. the map φ can be extended to all of $\mathbb{R}^n$ as a C^∞, globally Lipschitz function.

If 1 and 2 hold, then the following is an observer for (1.6)

$$\dot{z} = \tilde{F}(z) + [S_\infty]^{-1}\, C'\, (y - Cz), \quad z(0) \in \mathbb{R}^n. \tag{1.7}$$

where $C = (1, 0, \ldots, 0)$ and S_∞ is the solution of the algebraic equation

$$0 = -\theta S_\infty - A' S_\infty - S_\infty A + C'C$$

for θ large enough, and A is the anti-shift operator $(A)_{ij} = \delta_{i,j-1}$.

The general state-space Ω, accounts for a physical set, ideally a forward invariant set for the system, in which case the observer is proved to have exponential convergence: if the set Ω is positively invariant for (1.2), then the solution of the composite system (1.6) and (1.7) satisfies

$$|z(t) - x(t)| \leq c \, \exp\left(\frac{\theta}{3}\, t\right) |z(0) - x(0)|,$$

for all $t \geq 0$, and all $x(0) \in \Omega$, $z(0) \in \mathbb{R}^n$.

Along the same lines, Ciccarella, Dalla Mora and Germani (1993), in [11] proposed an observer which is designed directly for system (1.2), i.e., a change of coordinates is not required. The assumptions on (1.2) are very similar to the ones in paper [23]:

1. the matrix $O_{\mathrm{N}}(\xi)$ has full rank for all $\xi \in \mathbb{R}^n$;
2. the function $\varphi(x) := L_f^n h(\phi^{-1}(x))$ satisfies a condition of the form

$$|\varphi(x_1) - \varphi(x_2)| \leq \gamma |x_1 - x_2|^\delta, \quad \forall\, x_1, x_2 \in \mathbb{R}^n$$

for some $\gamma > 0$ and $\delta \in (0, 1]$.

Under these two assumptions, the following is an observer for system (1.2)

$$\dot{z} = f(z) + [Q(z)]^{-1} k \, (y - h(z)), \quad z(0) \in \mathbb{R}^n,$$

(where $Q(z) := O_{\mathrm{N}}(z)$, and $k \in \mathbb{R}^n$ is a finite gain vector), in the sense that

$$\begin{aligned} \text{if } \delta \in (0,1): &\qquad \lim_{t \to +\infty} |z(t) - \xi(t)| \leq \varepsilon, \quad \forall\, \varepsilon > 0, \quad \forall\, z(0) \in \mathbb{R}^n, \\ \text{if } \delta = 1: &\qquad \lim_{t \to +\infty} |z(t) - \xi(t)| = 0, \quad \forall\, z(0) \in \mathbb{R}^n. \end{aligned}$$

The methods described above all have the advantage of allowing an extension for multiple-input/multiple-output systems. But, as a disadvantage, these approaches involve very strong regularity assumptions, and the extension to the multiple output case is likely to involve lenghty computations. Nevertheless, within their scope, these observer constructions can be applied with success: for example, in [23], the observer (1.7) is applied to a simple bioreactor. In this application, the importance of proving the results for a more general state-space Ω (rather then the whole of $\mathbb{R}^n$) is made clear, since the bioreactors are positive systems and it is possible to identify a physical, forward invariant set, with respect to which exponential stability of the composite system is obtained.

Bioreactors and other biotechnological processes have long been a large field of application of control theory ([5], [18], [29]) and in particular of observer theory ([6], [7], [17], [39]). A bioreactor consists basically of a tank where several biological reactions occur simultaneously: essentially, there is a population of microorganisms (the biomass) that grows by feeding on a substrate, while some other intermediate products may also be present (examples of such processes are yeast growth and lactic fermentation). There are many variations to these processes, but typically one wishes to monitor and control the concentrations of microorganisms, substrate and the other byproducts.

Extensive work has been done on the subject of state-estimators for these processes and the main directions of research in this subject, as well as the current methodologies, are compiled and summarized in the book [6] by Bastin and Dochain.

This brings us to the standard and widely known extended Luenberger observers and Kalman filters, typically used in these biotechnological processes, and illustrated in [6]. These standard constructions are not guaranteed to work for nonlinear systems (one expects that locally they should converge, but there is no proof of this fact, in general), but it is often the case that they perform adequately in specific examples, and hence their practical success. Extended Kalman filters (continuous-time, deterministic) (for a reference see, for instance, [35], [46]) are based on the linearization of the dynamics f and output map h of nonlinear systems (1.2), and have the form

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

where the gain $L(z)$ is to be found in such a way that (at least locally) $|\xi(t) - z(t)| \to 0$ as $t \to +\infty$.

The gain $L(z)$ is computed online by

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

where $P \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, which is a solution to the following Riccati differential equation:

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

$R \in \mathbb{R}^{p\times p}$ and $Q \in \mathbb{R}^{n\times n}$ are two symmetric cost matrices, respectively positive definite and positive semidefinite, and

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

are the Jacobians of f and h evaluated at the point z. The Kalman filters are very "user friendly" in the sense that, by varying the cost matrices R and Q appropriately, users may assign less weight to measurements which are known to be noisy, as well as adjust the effect of errors in the initial position on the state estimate. However, in some cases they can fail to produce good estimates, as will be illustrated later on with some examples.

Besides the bioreactors mentioned above, many other biological/chemical systems may be represented by networks of reactions among several biological/chemical complexes, where the concentrations of some of these complexes can be monitored and

measured. The evolution of the various concentrations along time typically exhibits nonlinear behavior. Such networks of reactions can be viewed as dynamical systems, and its behavior modeled by a set of (nonlinear) differential equations.

The study of biological systems is an extremely active and rich field of research. As measurement methods and techniques become more advanced and sophisticated, biologists search for new approaches, including mathematically more varied techniques, that may help them deal with the immense amounts of data becoming available, provide more accurate and truthful models for their systems, and hopefully be useful in the study of questions such as predicting the response of a system.

State-estimators, as a concept, certainly have a great potencial to help dealing with and analyzing increasingly more complex models, that typically involve more variables than available measurements. A very appealing feature to biologists is that of using the available measurements to provide estimates of the unmeasured variables, and thus helping with the reconstruction, monitoring and control of the system. This is the case in several 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 general observer constructions described above, despite their satisfactory characteristics, have disadvantages as well and, typically, observers for nonlinear systems are mostly designed on a "case by case" basis. It is thus a natural starting point to focus on a general mathematical model for biological/chemical networks, and study the possibility of constructing state-estimators for such a model.

In this thesis we will explicitly construct observers for the class of systems which describe reaction networks of the type introduced and studied by Feinberg, Horn, and Jackson in [19, 20, 21, 27, 28]. Although the mathematical models that we have chosen to study may be, in some ways, considered restrictive from the biological/chemical point of view (due mostly to the fact that it models closed systems that can be described by weakly reversible networks, as will be discussed in Chapter 2), nevertheless we believe that it still accomodates quite a wide variety of significant biological systems, including many models for enzymatic mechanisms ([41]), a model for T-cell receptor signal transduction ([38]), and receptor–ligand interactions and G-protein coupled receptor activity in cyclic signaling pathways ([10, 36, 40, 55]). Several examples are provided throughout the text. The latter cyclic models, involving G-protein coupled receptors, have recently attracted some attention (see [40, 55]): the analysis and simulations of such models may be useful in the identification of possible new drugs and drug design.

As output maps, we take a subset of state variables or, more generally, monomials in the state variables. These can, in practice, be associated to measured reaction rates, and thus have a real physical meaning, as well as representing quantities that may be measured, for instance, through the energy released or consumed in the reaction. The existence of a forward invariant set (these networks are positive systems, and indeed we prove that the positive orthant is forward invariant) plays an important role in this work. Detectability and observers are meaningful in this forward invariant set and a

precise definition of detectability is formulated with respect to this set.

As a first step, we state and prove a necessary and sufficient theoretical condition for detectability. Under the hypothesis that the system is detectable, we present two full-state observers that are guaranteed to converge globally, and in addition exhibit some desirable robustness properties. Both observers are of the "Luenberger type", since they consist of a copy of the system plus a term proportional to the difference between real and estimated measurements.

The global convergence of the observer, as well as its robustness, may be obtained as an application of *input-to-state stability* (ISS) techniques. An underlying general framework for control systems, the concept of input-to-state stablility was first introduced by Sontag in [44] and then further characterized and investigated in [45, 49, 47] among many other research articles. This concept has been expanded and related notions have been developed (such as ISS with respect to compact sets [50], integral ISS [2, 3], or input to output stability [52]).

In this thesis, we will adapt the ISS notion to deal with positive systems, relative equilibria and possibly constrained inputs, and give a Lyapunov characterization very similar to the one in [49]. In a similar way to the results just cited (and to be made precise later), given any point $\bar{x} \in \mathbb{R}^n_{>0}$, a set $\mathbb{U} \subset \mathbb{R}^p$ and any $\bar{u} \in \mathbb{U}$, the system $\dot{z} = f^*(z, u)$ is input-to-state stable with input value set $\mathbb{U}$ (with respect to the points $\bar{x}, \bar{u}$), if $z(t) \in \mathbb{R}^n_{>0}$ for all $t \geq 0$ and there exist functions $\beta \in \mathcal{KL}$ and $\varphi \in \mathcal{K}_\infty$ such that

$$|z(t) - \bar{x}| \leq \beta(|z(0) - \bar{x}|, t) + \varphi(\|u - \bar{u}\|)$$

for all $t \geq 0$ and all $u \in \mathbb{U}$.

Interestingly, both observers can be derived from a system of the form

$$\dot{z} = f(z) + C'(u - H(z))$$

where f is the dynamics of our nonlinear model of chemical networks, H is a function of the output map, and C is a constant matrix to be defined in Chapter 3. This system is shown to satisfy a *semi-global* ISS property, and the observers are obtained by letting $u(t) = H(x(t))$. For the main observer we have that $H(z) = h(z)$, while an alternative observer can be constructed by letting H be the vector of the logarithms of the measurements $h_1, \ldots, h_p$. As a consequence of the ISS estimate, the observers exhibit robustness with respect to noisy measurements and we also prove convergence under sampled outputs.

Another very important property is the robustness to perturbations in the networks' constant parameters: while dealing with biochemical systems it is not uncommon that the (constant) parameters suffer small perturbations (for instance, depending on the experimental set up). The input-to-state stability estimates provide a "robustness" property (to be stated and proved in Chapter 5) in the sense that small perturbations in the parameters of the system lead to correspondingly small errors in the observer's estimates.

The extension of the observers for the system with inputs is not pursued in this thesis. We also do not deal with system identification or parameter estimation: that is, it is assumed that the system's model and its parameters are known (though we do study the performance of the observer under perturbations to the known parameters).

We next give a brief outline of the material developed in each chapter.

In Chapter 2 the general model for Feinberg-Horn-Jackson zero-deficiency chemical networks is introduced and some background provided: stability properties, characterization of steady-states, and other results are recalled. Many of these results were studied and proved by Sontag in the paper [48], however we will present an alternative proof of the main stability result for these systems.

In Chapter 3 a notion of detectability is introduced and a necessary and sufficient theoretical condition for the system to be detectable is stated and proved. A practical and intuitive way to check whether the system is detectable is illustrated. Also in Chapter 3 the input-to-state stability (ISS) property is adapted to deal with positive systems, relative equilibria and constrained inputs. A Lyapunov characterization, similar to the one in [49], is also given.

Chapter 4 introduces the first (main) observer and its global convergence is proved, using Lyapunov methods and ISS techniques. The robustness to output noise is discussed and we show that sampled outputs of the form "sample and hold" do not affect the convergence of the trajectory of observer to the correct values. It is also shown how to assign different weights to each component of the output, and thus possibly increase the rate of convergence of the estimator.

Chapter 5 proposes a notion of robustness with respect to the parameters of the system for observers, and goes on to explore this robustness for the main observer, where the parameters are the reaction rate constants of the network. This robustness can be quantified by an input-to-state stability type of estimate, and it uses the continuity of the steady-states of the system on these parameters. Viewing the steady-states of the system as a map on the space of parameters, we will show that this map is real analytic, using the Perron-Frobenius Theorem and other techniques for irreducible matrices.

In Chapter 6 an alternative observer is analyzed, which uses the logarithms of the outputs as its input data. This observer is also proved to be globally convergent. In terms of performance, the alternative observer seems to have substantially slower convergence characteristics but this is, however, traded for more robustness to perturbations in measurements. This alternative observer also exhibits robustness with respect to parameters in the sense discussed in Chapter 5.

Chapter 7 illustrates the theoretical results of the previous chapters with simulations, where the observers are applied to the case of the McKeithan model for T-cell receptor signal transduction, and also to a dimer model for receptor–ligand interactions. The simulations test the effect of noise, the effect of unknown inputs acting on states, or state drift, and the performance of the observer under sampled outputs. A comparison between the Luenberger observer, an extended Kalman filter and our main observer is also given here.

Chapter 8 reports an experiment (performed at the L. Romsted Laboratory, in the Department of Chemistry and Chemical Biology at Rugters) where a simple reversible chemical reaction, involving four species, was monitored in an NMR spectrometer. The evolution of the four concentrations was followed and measured along time. The performance of our main observer is tested against this complete information about the system's trajectories.

Finally, Appendix A provides a separation principle for the "system + observer" composite, for the case of output maps for which at least one of the measurements is one of the reaction rates. The remaining other appendices collect various technical results.

Chapter 2

Feinberg-Horn-Jackson Zero-Deficiency Chemical Networks

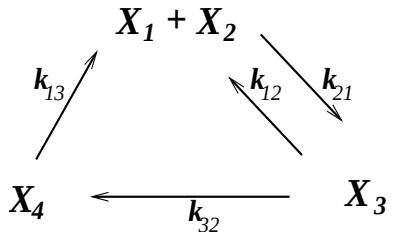


Figure 2.1: Example of a F-H-J 0-deficiency network.

The main research developed in this thesis concerns a certain type of systems that provide a mathematical model for a specific class of chemical reaction networks. The networks of Feinberg-Horn-Jackson zero-deficiency type consist of a family of chemical reactions among various species, characterized by a few specific properties that will be summarized below and basically include: a weak reversibility property, a closed reactor, constant temperature. These types of networks have been extensively studied, by the three authors after which the networks are named, in, for instance, [19, 20, 21, 27, 28], and lately, from the point of view of control theory, by Sontag in [48].

One of the simplest examples (see [18]) is the synthesis of ethyl tert-butyl ether ($C_6H_{14}O$), represented by X_3, from isobutene (C_4H_8) and ethanol (C_2H_6O), represented by X_1 and X_2:

$$X_1 + X_2 \rightleftharpoons X_3.$$

Another example is the process

$$X_1 + X_2 \rightleftharpoons X_3 + X_4,$$

where X_1, X_2, X_3, X_4 represent methanol (CH_4O), acetic acid ($C_2H_4O_2$), water (H_2O) and methyl acetate ($C_3H_6O_2$), respectively. This is a reversible reaction that can be reproduced in a test tube and monitored in an NMR spectrometer, and which was actually the choice for an experiment (conducted at the Department of Chemistry and Chemical Biology, by L. Romsted's group) whereby some of the results presented in this thesis were tested and validated (see Chapter 8).

Yet another Feinberg-Horn-Jackson network is the general enzymatic mechanism with uncompetitive inhibitor (see [41]), consisting of one enzyme E, one substrate S,

one product P and an uncompetitive inhibitor I (Q and R are intermediate complexes):

$$S + E \rightleftharpoons Q \rightleftharpoons P + E, \quad Q + I \rightleftharpoons R,$$

an example to be studied in more detail in Section 4.6.

As a working example, consider the weakly reversible network of reactions among four chemical species, X_1, X_2, X_3 and X_4, depicted in Figure 2.1.

This network is reversible, in the sense that there is a (directed) chemical pathway joining every pair of complexes. The species are assumed to form a homogeneous mixture, which may be contained in a reactor tank (such as a test tube), and is maintained at a constant temperature. The reactor is closed, in the sense that no exchanges (inflows or outflows) with the exterior are permitted.

The evolution, along time, of the concentrations of the four species can be described by a set of differential equations

$$\begin{aligned} \dot{x}_1 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 \\ \dot{x}_2 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 \\ \dot{x}_3 &= k_{21}x_1x_2 - (k_{12} + k_{32})x_3 \\ \dot{x}_4 &= k_{32}x_3 - k_{13}x_4. \end{aligned} \tag{2.1}$$

where mass-action kinetics is used, i.e., the reaction rates are proportional to the species' concentrations. Thus $k_{21}x_1x_2$ is the rate associated to the reaction $X_1 + X_2 \to X_3$, meaning that the rate of *production* of X_3 is proportional to the amount of molecules of both X_1 and X_2 in the mixture; and conversely, X_1 and X_2 are *eliminated* at the same rate at which X_3 is produced ($-k_{21}x_1x_2$).

As further motivation, and by way of relating the mathematical results to current relevant biological issues, we point out that system (2.1) is also a model for kinetic proofreading in T-cell receptor signal transduction, as proposed by McKeithan in 1995 (see [38]). This model greatly motivated the paper [48], which in its turn was a starting point for the work developed in this thesis. Many of our results will be illustrated with simulations concerning the McKeithan model, so we will give a very brief overview of this model and its biological meaning.

Consider that, as part of our immune system, a T-cell exhibits high sensitivity and high selectivity in the recognition and discrimination between foreign and self antigens (in a simplistic view, these may be looked on as "noxious" or "inocuous" antigens) in our body. The receptors of a T-cell (X_1), bind with the antigen presenting cell (X_2) and successively form other complexes (X_3,X_4,X_5,...), through several energy-requiring steps (typically involving phosphorylations).

The main idea behind the kinetic proofreading model is that these successive complexes may be, from a chemical point of view, more or less stable and either dissociate (thus decaying back into X_1, X_2), or continue the sequence of transformations. Antigens that form more stable complexes, and go through all the transformations steps without dissociating, cause the T-cell to generate signaling in recognition of a foreign antigen (in the Figure, signaling would occur as a high concentration of X_4 was achieved). Other

antigens (short-lived) dissociate at intermediate steps and do not cause the T-cell to emit a signal.

Hence the kinetic proofreading hypothesis: through a sequence of transformations, a T-cell has higher sensitivity to discriminate between foreign and self antigens, thus avoiding over signaling and "mistakes" due to a "good" antigen. The model may represent either "good" or "bad" antigens by assigning different values to the rate constants k_i — for instance, a "good" antigen would have higher dissociating rates k_{12}, k_{13} than transformation rates k_{32}.

Returning to the general model, we will show in this chapter that the dynamics of the system will, eventually, come to a steady-state. At this steady-state, the concentrations of each species achieve a constant value,

$$x_1(\infty) = \bar{x}_1,\ x_2(\infty) = \bar{x}_2,\ x_3(\infty) = \bar{x}_3,\ x_4(\infty) = \bar{x}_4.$$

The values of the concentrations at steady-state will, of course, depend on the initial concentrations and, in general, one may not expect to have access to direct measurements of all the concentrations $x_i(t)$ which, asymptotically, give all the values $\bar{x}_i$. In the next chapters, one of the main goals is to construct (when and if possible) a state-estimator for these systems. State-estimators, also called *observers*, compute estimates of the internal states of the system ($x_i(t)$, $i = 1, \ldots, n$), using data provided by measurement probes, or partial state information. Thus through an observer we will obtain estimates for the steady-state values, $\bar{x}_i$.

Throughout the text, we will refer to each individual reactant or product as a *species* and to the set of reactants or the set of products in a reaction as a *complex*. The network depicted on Figure 2.1, consists of four species (X_1, X_2, X_3, X_4), which are organized into three complexes ($X_1 + X_2$, X_3, X_4).

In all our analysis, one of the most important properties of the networks we are studying is their *weak reversibility* in the sense that there is a directed pathway connecting each complex to every other complex or, in other words, each complex is in turn a reactant in some reaction i and a product in another reaction j. When only two complexes are involved (eg., $Y_1 \rightleftharpoons Y_2$) this means that the reaction is reversible, but in general, as in the example of Figure 2.1, it is not required that all reactions are reversible. More precisely, a Feinberg-Horn-Jackson chemical network may consist of one or more weakly reversible blocks (the enzymatic mechanism above has two reversible blocks). Each of these blocks is also known as a "linkage class" or "connected component".

To summarize, the essential properties of Feinberg-Horn-Jackson zero-deficiency chemical networks are:

- Homogeneous mixture of n species, organized into m complexes
- Constant temperature
- Closed reactor (no inflows or outflows)
- Network consists of l weakly reversible blocks, or "linkage classes"

- Kinetics of the mass-action type (or more general, to be described)
- Zero-deficiency: $\delta = m - l - \dim \mathcal{D} = 0$, where $\mathcal{D}$ is the stoichiometric space associated with the network, to be described below.

In this chapter, the general mathematical model is described and several important properties — stability, characterization of steady-states, convergence — are studied. In the paper [48], Sontag discusses and surveys several stability and other control-theoretic results applicable to such systems, some of which will be recalled here. This paper developed some very useful results and formalism for dealing with these systems, which will be frequently referred to here. In addition, some new characterizations of the steady-states are given and a new direct proof of global asymptotic stability is also provided.

2.1 The Mathematical Model

Abstractly, we will consider n species, forming m complexes, under the conditions just specified, and consider a dynamical model for the time evolution of the concentrations of each species

$$x(t) = (x_1(t), x_2(t), \ldots, x_n(t))' \in \mathbb{R}^n.$$

Our model allows a general formulation of the reaction rates, such as mass-action kinetics (monomials in the concentrations x_i, as above), but also other forms as described below:

$$\dot{x} = f(x) := \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} \theta_1(x_1)^{b_{1j}} \theta_2(x_2)^{b_{2j}} \ldots \theta_n(x_n)^{b_{nj}} (b_i - b_j), \tag{2.2}$$

where $m \leq n$ is the number of complexes, each complex being represented by a column vector $b_j \in \mathbb{R}^n$ with entries $b_{1j}, b_{2j}, \ldots, b_{nj}$, which are nonnegative integers. It is assumed that

$$B := (b_1, b_2, \ldots, b_m)$$

has rank m and that none of its rows vanishes.

The constants a_{ij} are all nonnegative and define the reaction rate constants (also refered to as kinetic constants). If the reaction "$b_j \to b_i$" is present in the network, then $a_{ij} \neq 0$, otherwise $a_{ij} = 0$. The matrix

$$A = (a_{ij})$$

is assumed to be *irreducible*. By an irreducible matrix we mean:

$$A \in \mathbb{R}^{m \times m} : \quad (A + I)^k > 0, \quad \text{for some power } k$$

where I is the identity matrix and the inequality $A > 0$ (respectively, $A \geq 0$) means that every entry of the matrix on the left hand side is positive (respectively, nonnegative).

Moreover, observe that each complex b_i can be identified with a *vertex* in a directed graph Gr. The reactions "$b_j \to b_i$" can be identified as the *edges* of the graph Gr. Observe also that $a_{ij} \neq 0$ whenever the reaction (or edge) "$b_j \to b_i$" is present in the network and $a_{ij} = 0$ if the reaction is not taking place. Therefore, the matrix A_d generated from A by letting

$$(A_d)_{ij} = 1, \quad \text{if } a_{ij} \neq 0$$
$$(A_d)_{ij} = 0, \quad \text{if } a_{ij} = 0,$$

is in fact the *adjacency matrix* of the directed graph Gr, since $(A_d)_{ij} = 1$ if the edge $b_j \to b_i$ exists and $(A_d)_{ij} = 0$ otherwise. Equivalently, we also say that $Gr = Gr(A)$ is the incidence graph of A.

The graph $Gr(A)$ is said to be *strongly connected* if, for any pair of vertices b_i, b_j, there exists a (directed) path from b_i to b_j. So, note that

$$A \text{ is irreducible} \iff Gr(A) \text{ is strongly connected.}$$

In the language of Feinberg et. al., each "linkage class" corresponds to one strongly connected component in the graph $Gr(A)$. Irreducibility of A amounts to saying that $Gr(A)$ consists of only one strongly connected component and hence a a restriction to "single linkage class" systems. This restriction can be removed, as explained in [48], 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 4.6.

Each map $\theta_i : \mathbb{R} \to [0, +\infty)$,

(a) is locally Lipschitz, i.e. for each compact set $K \subset \mathbb{R}$, there exists a constant $c > 0$ such that
$$|\theta_i(r) - \theta_i(s)| \leq c|r - s|, \quad \forall\, r, s \in K;$$

(b) $\theta_i(0) = 0$;

(c) $\int_0^1 |\ln \theta_i(r)| dr < \infty$;

(d) its restriction to $\mathbb{R}_{\geq 0}$ is strictly increasing and onto the set $[0, \sigma_i)$, where $0 < \sigma_i \leq +\infty$.

Before stating the last condition that the functions θ_i should satisfy, let us introduce the following vector functions:

$$\rho^{[n]}(x) = (\ln \theta_1(x_1), \ldots, \ln \theta_n(x_n))' \quad \text{and} \quad \mathrm{Exp}^{[n]}(v) = (e^{v_1}, \ldots, e^{v_n})'$$

defined on $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n$, respectively. (From now on, we will drop the superscript n of $\rho^{[n]}$ and $\mathrm{Exp}^{[n]}$, since its value is usually clear from the context.)

Since each θ_i (restricted to $\mathbb{R}_{>0}$) is onto the set $(0, \sigma_i)$, then each function $\rho_i = \ln \theta_i$ (for the restriction of θ_i to $R_{>0}$) is onto $(-\infty, \bar{\rho}_i)$ with $\bar{\rho}_i = \ln \sigma_i$. Then ρ_i has an inverse function, which is onto $\mathbb{R}_{>0}$:

$$\rho_i^{-1} : (-\infty, \bar{\rho}_i) \to \mathbb{R}_{>0}.$$

Each function θ_i should also satisfy:

(e) For any given constant p

$$\lim_{t \to \ln \sigma_i} \int_a^t \rho_i^{-1}(s)\, ds - pt = +\infty,$$

for any $a < \ln \sigma_i$.

Note that, for any constant p, there exists $t_0 \in (-\infty, \bar{\rho}_i)$ such that $\rho_i^{-1}(s) > p + 1$ for all $s \geq t_0$. Therefore, when $\sigma_i = +\infty$, condition (e) always holds.

Example 2.1.1 In the simplest case, $\theta(r) = |r|$ (the one that will be eventually adopted in the following chapters), and the function $f : \mathbb{R}^n \to \mathbb{R}^n$ becomes of the mass-action kinetics form

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

This is the form of the equations, in the motivating McKeithan model, where $n = 4$ and $m = 3$, and $X_1 + X_2$, X_3 and X_4 are represented by the vectors b_1, b_2 and b_3, respectively:

$$b_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \text{and} \quad b_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

The irreducible matrix A determines the reactions and reaction rate constants:

$$A = \begin{pmatrix} 0 & k_{12} & k_{13} \\ k_{21} & 0 & 0 \\ 0 & k_{32} & 0 \end{pmatrix},$$

where $a_{12} = k_{12}$ means that there is a reaction that produces complex 1 from complex 2, with rate constant k_{12}. Note that the diagonal elements of A play no role and may be set to zero.

Example 2.1.2 The last condition holds when $\theta_i(r) = |r|$, since this function is unbounded. Another possible form for θ_i is the usually called Michaelis-Menten kinetics:

$$\theta_i(r) = \frac{r}{k + r}$$

with $k > 0$. Then conditions (a)-(e) are satisfied:

$$\begin{aligned}\int_0^1 |\ln \theta_i(r)| dr &= \int_0^1 \ln(k+r) - \ln r \; dr \\ &= [(r+k)\ln(k+r) - r - (r\ln r - r)]_0^1 = +\infty\end{aligned}$$

This function θ_i is onto $[0,1)$ and ρ_i is onto $(-\infty, 0)$. The inverse function is

$$\rho_i^{-1}(s) = \frac{k\, e^s}{1-e^s}$$

and

$$\lim_{t \to \ln \sigma_i} \int_a^t \frac{k\, e^s}{1-e^s} ds - pt = \lim_{t \to \ln \sigma_i} [-k \ln(1-e^s)]_a^t = +\infty, \ \forall a.$$

Since the system (2.2) describes the evolution of *concentrations*, which are physically positive quantities, we will be interested only on those trajectories which evolve in the *positive orthant*

$$\mathbb{R}^n_{>0} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i > 0 \ \text{ for all } \ i\},$$

even though (2.2) is defined on all of $\mathbb{R}^n$, It is easy to verify (cf. [48] and below) that $\mathbb{R}^n_{>0}$ is a forward invariant set for (2.2).

2.2 The Stoichiometric Space and Positive Classes

An important object associated with a Feinberg-Horn-Jackson network is its stoichiometric space, given by

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

The parallel translates of this space constitute invariant manifolds for the system (2.2): for any $d \in \mathcal{D}^\perp$,

$$\langle d, f(x)\rangle = 0 \ \Rightarrow \ \langle d, \dot{x}\rangle = 0 \ \Rightarrow \ \langle d, x(t)\rangle = \langle d, x(0)\rangle, \ \forall t > 0,$$

that is, given any initial condition $x(0) = x_0 \in \mathbb{R}^n_{\geq 0}$, the trajectory of system (2.2) evolves in a *class* defined (uniquely) by the space $\mathcal{D}$ and the initial state x_0.

Following [48], define the *classes* of the system (2.2) by:

$$\mathcal{S} = (p + \mathcal{D}) \cap \mathbb{R}^n_{\geq 0} = \{p + d : d \in \mathcal{D}\} \cap \mathbb{R}^n_{\geq 0}, \quad p \in \mathbb{R}^n_{\geq 0}.$$

Note that the classes $\mathcal{S}$ depend on the matrix B but not on A, and that, given an initial condition $x_0 \in \mathbb{R}^n_{\geq 0}$, the system (2.2) will stay for all t in the class $\mathcal{S} = (x_0 + \mathcal{D}) \cap \mathbb{R}^n_{\geq 0}$. If $\mathcal{S} \cap \mathbb{R}^n_{>0} \neq \emptyset$, $\mathcal{S}$ is said to be a *positive class*.

The *deficiency* of a chemical reaction mechanism is defined as

$$\delta = m - l - \dim \mathcal{D}$$

where l is number of "linkage classes" or weakly reversible blocks of the network. The networks we are studying here always satisfy $\delta = 0$. For this reason, in the literature, they are called *zero-deficiency chemical networks*. However, the restrictions introduced by the expression $\delta = 0$ are incorporated into our model through the requirements on the matrices B and A. So, throughout this work, the concept of deficiency is not usually referred to.

Another important object associated with (2.2) is the set of equilibria:

$$E := \{x \in \mathbb{R}^n_{\geq 0} : f(x) = 0\}.$$

The set of strictly positive equilibria is denoted by $E_+ = E \cap \mathbb{R}^n_{>0}$. In the work of Feinberg and Horn & Jackson (where they deal with the case of $\theta_i(r) = |r|$ for every i), it has been proved that in each positive class $\mathcal{S}$ there exists a unique positive equilibrium, $\{\bar{x}\} = \mathcal{S} \cap E_+$, and that $\bar{x}$ is locally asymptotically stable relative to $\mathcal{S}$. In this chapter we will present a proof of a corresponding result for the more general maps θ_i as desribed above. For this reason, it will be useful to identify the positive classes by their positive equilibrium, and thus give another representation of each positive class:

$$\mathcal{S}_{\bar{x}} = \{x \in \mathbb{R}^n_{\geq 0} : \langle v_i, x\rangle = \langle v_i, \bar{x}\rangle, i = 1, \ldots, n-m+1\}$$

where $\{v_1, \ldots, v_{n-m+1}\}$ forms a basis of $\mathcal{D}^\perp$.

The set of elements of E which have at least one coordinate equal to zero (the *boundary equilibria*) will be denoted by E_0. It is interesting to note that the boundary equilibria do not depend on the matrix A, but only on the matrix B. Indeed, in [48] the following characterization is proved.

Proposition 2.2.1 (see Proposition VI.3 in [48]) For any system (2.2), and for an arbitrary $x \in \mathbb{R}^n_{\geq 0}$,

$$x \in E_0 \;\Leftrightarrow\; \theta_1(x_1)^{b_{1j}}\theta_2(x_2)^{b_{2j}}\ldots\theta_n(x_n)^{b_{nj}} = 0, \quad \forall\, j \in \{1, \ldots, m\}.$$

The sufficiency of this condition is obvious, since clearly $f(x) = 0$ when x satisfies this condition. The important part of Proposition 2.2.1 is its claim that *every* term in (2.2) vanishes at a boundary equilibrium. The proof uses the rank condition on B as well as irreducibility of A.

Example 2.2.2 In the McKeithan example above, the stoichiometric space is:

$$\mathcal{D} = \text{span}\{(1, 1, -1, 0), (1, 1, 0, -1)\}.$$

The positive classes are characterized by

$$x_1 + x_3 + x_4 = \alpha, \quad x_2 + x_3 + x_4 = \beta$$

for each pair of *positive constants* α, and β.

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 $r \geq 0$. Note that *none of these boundary equilibria is contained in any positive class* (see below), since either $\bar{x}_2 + \bar{x}_3 + \bar{x}_4 = 0$ or $\bar{x}_1 + \bar{x}_3 + \bar{x}_4 = 0$, thus violating either $\beta > 0$ or $\alpha > 0$, respectively.

2.3 A Stability Theorem

We now state the main stability result for networks of FHJ type, which will be often referred to in this thesis. This result is also proved in [48], but a more direct, alternative proof, will be given here which holds for systems (2.2) with the general requirements (a)-(e) together with conditions (2.3) and (2.5), to be stated below.

In particular, the proof of the *existence* of positive equilibria for the system (2.2) in the case where the maps θ_i are onto $[0, \sigma_i)$ with $\sigma_i < +\infty$ was not fully explored and characterized in [48]. We do so now, by stating condition (2.5) which turns out to be necessary and also sufficient for the existence of positive equilibria.

Theorem 1 Consider the system $\dot{x} = f(x)$ given in (2.2), and assume that both conditions (2.3) and (2.5) hold. Then, for each positive class $\mathcal{S}$ there exists a (unique) state $\bar{x} = \bar{x}_{\mathcal{S}} \in \mathbb{R}^n_{>0}$ which is a globally asymptotically stable point relative to $\mathcal{S}$, i.e., for each $x_0 \in \mathcal{S}$, 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 $t > 0$.

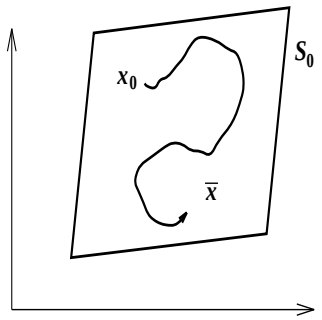


Figure 2.2: Global asymptotic stability with respect to a class.

2.3.1 The "No Boundary Equilibra" Assumption

For the rest of this Chapter, we will make the following assumption:

the system (2.2) 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 $x - \bar{x} \in \text{span}\{b_i - b_j, \quad i, j = 1, \ldots, m\}$ for some $\bar{x} \in \mathbb{R}^n_{>0}$, then $f(x) \neq 0$. Using our notations, one may also write

$$\mathcal{S} \cap E_0 = \emptyset, \quad \text{for each positive class } \mathcal{S}. \tag{2.3}$$

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 biochemical reaction models, and is in particular satisfied in several examples to be discussed in Chapters 7 and 8,

such as the kinetic proofreading model. It is possible to weaken this boundary assumption and still obtain significant (though more restricted) results (using the techniques developed in [48]) but we prefer not to do so in order to streamline the presentation.

2.3.2 Existence of Positive Equilibria

In the case of a map θ_i which is onto $\sigma_i < +\infty$, it is *not always true that positive equilibria exist* in each class. (see Example 2.3.4, below). However, as will be seen in the next section, if there exists at least one positive equilibrium, then there exists a unique positive equilibrium in each positive class, $\mathcal{S}$.

Note that the right hand side of system (2.2) can be written in yet another form:

$$f(x) = B\tilde{A}\theta_B(x) \tag{2.4}$$

where

$$\tilde{A} = A - \text{ diag } \left(\sum_i a_{i1}, \sum_i a_{i2}, \ldots, \sum_i a_{im} \right),$$

i.e., in each diagonal entry, $\tilde{A}$ has the (negative) sum of all the elements of the corresponding column of A, so that each column of $\tilde{A}$ adds up to zero. It is clear that $\tilde{A}$ is also an irreducible matrix, because its graph $Gr(\tilde{A})$ coincides with $Gr(A)$.

We will consider the function $\theta_B : \mathbb{R}^n_{>0} \to \mathbb{R}^m_{>0}$ given by

$$\theta_B(x) = \text{Exp}(B'\rho(x)) = (e^{\langle b_1, \rho(x)\rangle}, e^{\langle b_2, \rho(x)\rangle}, \ldots, e^{\langle b_m, \rho(x)\rangle})'.$$

The next two Lemmas are proved in [48]. The first one is easy to see, as it follows directly from the fact that B has full rank. The second Lemma uses the fact that $\tilde{A}$ is an irreducible matrix, and has a Perron eigenvector associated with the zero eigenvalue.

Lemma 2.3.1 (Lemma V.1 in [48]) A state $\bar{x}$ is an equilibrium if and only if

$$\theta_B(\bar{x}) \in \ker \tilde{A}.$$

Lemma 2.3.2 (Lemma V.2 in [48]) There exists $\bar{y} \in \mathbb{R}^m_{>0} \cap \ker \tilde{A}$ so that

$$(\mathbb{R}^m_{\geq 0} \setminus \{0\}) \cap \ker \tilde{A} = \{\kappa \bar{y}, \kappa > 0\}.$$

From Lemma 2.3.1, it is clear that the following is a necessary and also sufficient condition for the existence of positive equilibria for system (2.2):

$$\ker \tilde{A} \cap \text{im}\, \theta_B \neq \emptyset. \tag{2.5}$$

This condition is not always satisfied, as illustrated in Example 2.3.4: a network is given that does not satisfy (2.5) and doesn't have positive equilibria.

However, this condition is clearly satisfied for the case when the restriction of θ_i to $R_{>0}$ is onto $(0, +\infty)$ for every i. In this case, the map θ_B is, in fact, onto $\mathbb{R}^n_{>0}$ (due to the fact that B is a full rank matrix), so it follows from Lemma 2.3.2 that the intersection of $\ker \tilde{A}$ and $\text{im}\, \theta_B$ is not empty:

Lemma 2.3.3 If the restriction of θ_i to $R_{>0}$ is onto $(0,+\infty)$ for every i, then condition (2.5) holds. In particular, condition (2.5) holds in the case when $\theta_i(r) = |r|$ for every i.

Example 2.3.4 Consider the following network

$$X_1 \underset{a_{12}}{\overset{a_{21}}{\rightleftarrows}} 2X_1 + X_2 \quad \overset{a_{32}}{\to} \quad X_1 + X_3 \quad \overset{a_{13}}{\to} X_1$$

and take

$$\theta_1(x_1) = x_1, \quad \theta_2(x_2) = x_2, \quad \theta_3(x_3) = \frac{x_3}{1+x_3}.$$

Then

$$\theta_B(x) = \begin{pmatrix} x_1 \\ x_1^2 x_2 \\ x_1 x_3/(1+x_3) \end{pmatrix},$$

and its image is the set

$$\operatorname{im} \theta_B = \{(\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3_{>0} : \ \xi_3 < \xi_1\}.$$

The matrices A and $\tilde{A}$ are

$$A = \begin{pmatrix} 0 & a_{12} & a_{13} \\ a_{21} & 0 & 0 \\ 0 & a_{32} & 0 \end{pmatrix} \quad \text{and} \quad \tilde{A} = \begin{pmatrix} -a_{21} & a_{12} & a_{13} \\ a_{21} & -a_{12} - a_{32} & 0 \\ 0 & a_{32} & -a_{13} \end{pmatrix},$$

and the kernel of $\tilde{A}$ is given by

$$\ker \tilde{A} = \ \text{span} \ \left\{ \left(1, \frac{a_{21}}{a_{12}+a_{32}}, \frac{a_{32}}{a_{13}} \frac{a_{21}}{a_{12}+a_{32}} \right)' \right\},$$

so, it is clear that (2.5) may not hold for some values of the reaction constants a_{ij}. In particular, if

$$\frac{a_{32}}{a_{13}} \frac{a_{21}}{a_{12}+a_{32}} \geq 1 \tag{2.6}$$

there will be *no positive equilibria* in any class.

In this example, the positive classes have the form:

$$\mathcal{S} = \{x \in \mathbb{R}^3_{\geq 0} : \ x_1 - x_2 = c\}$$

for each $c \in \mathbb{R}$ (see Figure 2.3 which shows the classes corresponding to $c = 1$, $c = -1$ and $c = 0$). If condition (2.6) is not satisfied, then positive equilibria exist in every class and have the form:

$$\bar{x}_1 = \frac{c}{2} + \frac{1}{2}\sqrt{c^2 + \frac{4a_{21}}{a_{12}+a_{32}}}, \quad \bar{x}_2 = -\frac{c}{2} + \frac{1}{2}\sqrt{c^2 + \frac{4a_{21}}{a_{12}+a_{32}}}$$

and

$$\bar{x}_3 = \frac{a_{32}a_{21}}{a_{13}(a_{12}+a_{32}) - a_{32}a_{21}}\,.$$

The two situations (existence or not of positive equilibria), are illustrated in Figures 2.3 and 2.4, for different choices of the parameters a_{ij}. For Figure 2.3 we have used: $a_{12} = 1$, $a_{13} = 1$, $a_{21} = 0.5$ and $a_{32} = 1$. For these parameters condition (2.6) is not satisfied, and so positive equilibria exist and they all lie in the plane $x_3 = \bar{x}_3$. For Figure 2.4 we have used: $a_{12} = 1$, $a_{13} = 1$, $a_{21} = 3$ and $a_{32} = 1$, and condition (2.6) is satisfied. We can see that, in this case, $\bar{x}_3 < 0$, and thus no positive equilibria exist, since the plane $x_3 = \bar{x}_3$ doesn't intersect any positive class.

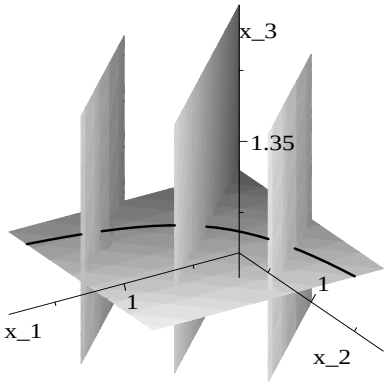


Figure 2.3: Positive classes (vertical planes) and positive equilibria (solid line, in the horizontal plane) for Example 2.3.4.

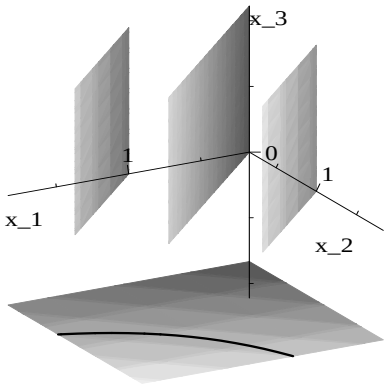


Figure 2.4: Situation where no positive equilibria exist, for Example 2.3.4.

In contrast, consider the case

$$\theta_1(x_1) = x_1, \quad \theta_2(x_2) = x_2, \quad \theta_3(x_3) = x_3.$$

Then

$$\theta_B(x) = \begin{pmatrix} x_1 \\ x_1^2 x_2 \\ x_1 x_3 \end{pmatrix},$$

and its image is $\mathbb{R}^n_{>0}$, so indeed $\ker \tilde{A} \cap \operatorname{im} \theta_B \neq \emptyset$ (Lemma 2.3.3), and positive equilibria exist. The coordinates $\bar{x}_1$ and $\bar{x}_2$ of each equilibrium point are given by the same expressions as above, while coordinate $\bar{x}_3$ is given by

$$\bar{x}_3 = \frac{a_{32}}{a_{13}} \frac{a_{21}}{a_{12} + a_{32}} .$$

In this case, for every choice of the (positive) parameters a_{12}, a_{13}, a_{21} and a_{32}, the value of $\bar{x}_3$ is also positive, and positive equilibria exist in every positive class. Similarly to the situation depicted in Figure 2.3, the horizontal plane $x_3 = \bar{x}_3$ intersects all positive classes.

2.4 Existence and Uniqueness of Equilibria

In order to present our alternative proof of Theorem 1, we start by noting that the following holds for any $x, \bar{x} \in \mathbb{R}^n_{>0}$:

$$\rho(x) - \rho(\bar{x}) \in \mathcal{D}^\perp \;\Leftrightarrow\; (\exists \kappa > 0)\; \theta_B(x) = \kappa \theta_B(\bar{x}). \tag{2.7}$$

To see that this is true, note that $\mathcal{D}$ is equal to $\operatorname{span}\{b_i - b_1 : i = 1, \ldots, m\}$ and

$$\begin{aligned} \rho(x) - \rho(\bar{x}) \in \mathcal{D}^\perp \quad &\Leftrightarrow \quad \langle \rho(x) - \rho(\bar{x}), b_i - b_1 \rangle = 0, \;\forall i \\ &\Leftrightarrow \quad \langle \rho(x), b_i \rangle = \langle \rho(x) - \rho(\bar{x}), b_1 \rangle + \langle \rho(\bar{x}), b_i \rangle, \;\forall i \\ &\Leftrightarrow \quad e^{\langle \rho(x), b_i \rangle} = \kappa\, e^{\langle \rho(\bar{x}), b_i \rangle}, \;\forall i \end{aligned}$$

where $\kappa = e^{\langle \rho(x) - \rho(\bar{x}), b_1 \rangle}$. Given any κ, note that necessarily the last equation says that $\kappa = e^{\langle \rho(x) - \rho(\bar{x}), b_1 \rangle}$, so the reverse implications hold.

In a similar way to Corollary V.3 in [48] we can show the existence of positive equilibria for system (2.2).

Corollary 2.4.1 Assume that (2.5) holds. Then the set of positive equilibria, E_+, is nonempty. Moreover, pick any fixed $\bar{x}$ in E_+. Then, for all $x \in \mathbb{R}^n_{>0}$,

$$x \in E_+ \;\Leftrightarrow\; \rho(x) - \rho(\bar{x}) \in \mathcal{D}^\perp.$$

Proof. By assumption $\ker \tilde{A} \cap \operatorname{im} \theta_B \neq \emptyset$. So, there exists $\bar{x} \in \mathbb{R}^n_{>0}$ such that $\theta_B(\bar{x}) = \bar{z} \in \ker \tilde{A}$ and then Lemma 2.3.1 says that $\bar{x} \in E_+$.

To prove the second part, fix any $\bar{x} \in E_+$. By Lemma 2.3.1, $x \in E_+$ is equivalent to $\theta_B(x) \in \ker \tilde{A}$. By Lemma 2.3.2, $\theta_B(x) = \kappa \theta_B(\bar{x})$, for some $\kappa > 0$. Finally, by (2.7), this is equivalent to $\rho(x) - \rho(\bar{x}) \in \mathcal{D}^\perp$. ∎

The proof of the next Lemma involves the assumption (e) on the maps θ_i. However since the proof can be found in [48], we will not repeat it here.

Lemma 2.4.2 (see Lemma IV.1 in [48]) For each p, w in $\mathbb{R}^n_{>0}$, there exists a unique $x = \varphi(p, w) \in \mathbb{R}^n_{>0}$ such that:

$$x - p \in \mathcal{D} \quad \text{and} \quad \rho(x) - \rho(w) \in \mathcal{D}^\perp.$$

Furthermore, the map $p, w \mapsto \varphi(p, w)$ is of class C^k, if the maps θ_i, restricted to $\mathbb{R}^n_{>0}$ are also of class C^k and satisfy $\theta_i'(y) > 0$ for all $y > 0$.

The uniqueness of a positive equilibrium in each class follows, as long as E_+ is nonempty:

Corollary 2.4.3 Assume that E_+ is nonempty. For each positive class, $\mathcal{S}$, there is a unique element $\bar{x} = \bar{x}_{\mathcal{S}}$ in $E_+ \cap \mathcal{S}$.

Proof. By assumption E_+ is nonempty, so pick any $\bar{x} \in E_+$. Pick any positive class $\mathcal{S}$, and pick an arbitrary $p \in \mathcal{S} \cap \mathbb{R}^n_{>0}$. Applying Lemma 2.4.2 with $w = \bar{x}$, it follows that there is a unique $x = \varphi(p, \bar{x})$ such that $x \in p + \mathcal{D}$ and $\rho(x) - \rho(\bar{x}) \in \mathcal{D}^\perp$. But, from the second part of Corollary 2.4.1, this is equivalent to

$$x \in p + \mathcal{D} \quad \text{and} \quad x \in E_+.$$

Therefore, x is the unique element that both belongs to the class $\mathcal{S}$ and is an equilibrium point in E_+. ∎

2.4.1 A Characterization of E

For any $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)' \in E_+$, introduce the following notation:

$$\pi_j(z, \bar{x}) = \pi_j := \left[\frac{\theta_1(z_1)}{\theta_1(\bar{x}_1)}\right]^{b_{1j}} \left[\frac{\theta_2(z_2)}{\theta_2(\bar{x}_2)}\right]^{b_{2j}} \cdots \left[\frac{\theta_n(z_n)}{\theta_n(\bar{x}_n)}\right]^{b_{nj}},$$

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

where 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, if $z \in \mathbb{R}^n_{>0}$,

$$\pi_j = e^{q_j}.$$

Define the function

$$\Psi : \mathbb{R}^n_{\geq 0} \times \mathbb{R}^n_{>0} \to \mathbb{R}_{\geq 0}$$

given by

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

A very compact characterization of the equilibrium points of (2.2) can be given in terms of this function.

Lemma 2.4.4 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

- ▷ either $\pi_i = 0$ for some $i \in \{1, \ldots, m\}$, hence for all i in this set, which implies, by Proposition 2.2.1 that $z \in E_0$;
- ▷ or all $\pi_i \neq 0$ (which implies all $\theta_k(z_k) > 0$ for all k and hence $z \in \mathbb{R}^n_{>0}$) 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 Corollary 2.4.1, we know this implies $z \in E_+$.

Conversely, if $z \in E$, then

- ▷ either $z \in E_+$, in which case we know from Corollary 2.4.1 that $\rho(z) - \rho(\bar{x}) \in \mathcal{D}^\perp$, and so $q_j = q_i$ for all $i, j \in \{1, \ldots, m\}$ which implies $\pi_j = \pi_i$ for all i, j and therefore $\Psi(z, \bar{x}) = 0$;
- ▷ or $z \in E_0$ and from Proposition 2.2.1 we know that $\pi_j = 0$ for all $j \in \{1, \ldots, m\}$ and therefore $\Psi(z, \bar{x}) = 0$.

∎

2.5 Global Asymptotic Stability

The following definitions are standard in the control theory literature. They are useful in establishing convergence and stability properties.

Definition 2.5.1 A function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of *class* $\mathcal{K}$, if it is continuous, positive definite, strictly increasing and satisfies $\alpha(0) = 0$.

The function α is of *class* $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and in addition satisfies $\alpha(r) \to +\infty$ as $r \to +\infty$.

A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of *class* $\mathcal{KL}$:

(i) for each fixed $t \geq 0$, $\beta(\cdot, t)$ is of class $\mathcal{K}$;

(ii) for each fixed $r \geq 0$, $\beta(r, \cdot)$ is continuous, decreasing and satisfies $\beta(r, t) \to 0$ as $t \to +\infty$.

Example 2.5.2 The function $\alpha(r) = r$ is of class $\mathcal{K}_\infty$, while the function $\alpha(r) = r/(r+1)$ is of class $\mathcal{K}$ but not $\mathcal{K}_\infty$. The function $\beta(r, t) = re^{-t}$ is of class $\mathcal{KL}$.

As in [48], the following function plays a crucial role in our results. Let $\bar{x}$ be any point in E_+ and define

$$V(x, \bar{x}) = \sum_{i=1}^{n} \int_{\bar{x}_i}^{x_i} (\, \rho_i(s) - \rho_i(\bar{x}_i) \,)\, ds. \tag{2.8}$$

This function has the properties:

(i) $V(\cdot, \bar{x})$ is continuous on $\mathbb{R}^n_{\geq 0}$ and differentiable on $\mathbb{R}^n_{>0}$, with

$$\nabla V(x, \bar{x}) = \rho(x) - \rho(\bar{x});$$

(ii) $V(x,\bar{x}) \geq 0$ and $V(x,\bar{x}) = 0$ iff $x = \bar{x}$;

(iii) The set $\{x \in \mathbb{R}^n_{\geq 0} : V(x,\bar{x}) \leq L\}$ is compact for every $\bar{x} \in \mathbb{R}^n_{>0}$ and every $L \geq 0$.

From point (iii), upper and lower bounds for V (see Appendix B) can be constructed:

(iv) There exist functions $\nu_1, \nu_2 \in \mathcal{K}_\infty$ such that

$$\nu_1(|x-\bar{x}|) \leq V(x,\bar{x}) \leq \nu_2(|x-\bar{x}|),$$

for all $x \in \mathbb{R}^n_{\geq 0}$.

Points (i) and (ii) are easy to check. Point (iii) is also very easy: it is enough to show that

$$\lim_{k\to+\infty} V(x^k,\bar{x}) = +\infty.$$

for any sequence $\{x^k\} \subset \mathbb{R}^n_{\geq 0}$ for which $|x^k - \bar{x}| \to +\infty$. To show this, pick one such sequence. Then there exists at least one index $j \in \{1,2,\ldots,n\}$ such that $x^k_j - \bar{x}_j \to +\infty$, so we consider only the term

$$\int_{\bar{x}_j}^{x^k_j} (\,\rho_j(s) - \rho_j(\bar{x}_j)\,)\,ds$$

(which is positive, because each ρ_j is a strictly increasing function). We know that

$$-\infty < \rho_j(s) - \rho_j(\bar{x}_j) < \ln\sigma_j - \rho_j(\bar{x}_j) := 2a.$$

Pick $s_0 \geq \bar{x}_j$ so that

$$s \geq s_0 \;\Rightarrow\; \rho_j(s) - \rho_j(\bar{x}_j) \geq a.$$

Then, for sufficiently large $|x - \bar{x}|$,

$$\int_{\bar{x}_j}^{x^k_j} (\,\rho_j(s) - \rho_j(\bar{x}_j)\,)\,ds \geq \int_{s_0}^{x^k_j} a\,ds = a(x_j - s_0) \to +\infty\,.$$

Therefore, it also follows that $V(x,\bar{x}) \to +\infty$ as $t \to +\infty$ as we wanted to prove. Together with continuity of V, this establishes property (iii).

There are some useful expressions for $\nabla V(x,\bar{x})\; f(x)$. Note that

$$\langle \rho(x) - \rho(\bar{x}), f(x)\rangle = \sum_{i=1}^m \sum_{j=1}^m a_{ij} e^{\langle b_j, \rho(x)\rangle}(q_i - q_j),$$

and we may write (for any i and j)

$$e^{\langle b_j,\rho(x)\rangle} = e^{\langle b_j,\rho(\bar{x})\rangle} e^{q_j}.$$

Also

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}e^{\langle b_j,\rho(\bar{x})\rangle}(e^{q_i}-e^{q_j}) \\
= \;& (e^{q_1},\ldots,e^{q_m})'A\theta_B(\bar{x}) - (e^{q_1},\ldots,e^{q_m})'\ \mathrm{diag}\ (\sum_j a_{j1},\ldots,\sum_j a_{jm})\theta_B(\bar{x}) \\
= \;& (e^{q_1},\ldots,e^{q_m})'\tilde{A}\theta_B(\bar{x}) = 0\,.
\end{aligned}$$

Then we have

$$\langle \rho(x)-\rho(\bar{x}), f(x)\rangle = W(x,\bar{x})$$

where

$$\begin{aligned}
W(x,\bar{x}) \;&:=\; \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}e^{\langle b_j,\rho(\bar{x})\rangle}\ [e^{q_j}(q_i-q_j)-(e^{q_i}-e^{q_j})] \\
&=\; \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}e^{\langle b_j,\rho(\bar{x})\rangle}\ e^{q_j}\ [(q_i-q_j)-(e^{q_i-q_j}-1)]. \qquad (2.9)
\end{aligned}$$

For convenience, we will also introduce the scalar function $\omega : \mathbb{R} \to \mathbb{R}_{\geq 0}$, given by

$$\omega(x) = e^x - 1 - x\,.$$

In terms of ω we have that

$$W(x,\bar{x}) \;=\; -\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}e^{\langle b_j,\rho(\bar{x})\rangle}\ e^{q_j}\ \omega(q_i-q_j). \qquad (2.10)$$

One more Lemma from [48] will be needed, to establish a useful estimate.

Lemma 2.5.3 (Lemma VIII.1 of [48]) Given any irreducible matrix $A = (a_{ij})$, $(a_{ij} \geq 0)$, define the following quadratic function:

$$Q(\eta_1,\ldots,\eta_m) := \sum_{i=1}^{m}\sum_{j=1}^{m}\ a_{ij}(\eta_i-\eta_j)^2. \qquad (2.11)$$

Then, there exists a constant $\kappa = \kappa(A) > 0$ such that

$$Q(q_1,\ldots,q_m) \geq \kappa \sum_{i=1}^{m}\sum_{j=1}^{m}(q_i-q_j)^2 \qquad (2.12)$$

for all $(q_1,\ldots,q_m) \in \mathbb{R}^m$.

Lemma 2.5.4 Let $f(x)$ be defined as in (2.2). There exists a positive constant $\kappa(A)$ and a continuous function $c : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{>0}$ given by

$$c(\xi) = \frac{1}{2}\min_j e^{\langle b_j,\rho(\xi)\rangle}$$

such that, for all $x \in \mathbb{R}^n_{>0}$ and all $\bar{x} \in E_+$:

$$\langle \rho(x)-\rho(\bar{x}), f(x)\rangle \leq -\kappa(A)\,c(\bar{x})\,\Psi(x,\bar{x}). \qquad (2.13)$$

Proof. Note that

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

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, Lemma 2.5.3, provides the positive constant $\kappa(A)$ that depends on A and satisfies that inequality. ∎

2.5.1 Forward-Invariance of $\mathbb{R}^n_{>0}$

Definition 2.5.5 A system $\dot{x} = f(x)$, 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* if, for each initial state $x(0) \in \mathbb{R}^n_{>0}$, the corresponding maximal solution of $\dot{x} = f(x)$, which is defined on an interval $J_{x(0)} = [0, t_{\max})$, has values $x(t) \in \mathbb{R}^n_{>0}$ for all $t \in J_{x(0)}$.

That the system (2.2) is $\mathbb{R}^n_{>0}$-invariant, can be proved as follows. Given an initial condition $x(0) \in \mathbb{R}^n_{>0}$, let $x(t)$ be the maximal solution of (2.2), defined on a (maximal) interval J. Let $\mathcal{I} = [0,+\infty)$.

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

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

Pick one coordinate k such that $x_k(t_0) = 0$. 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:

$$(\check{x}(t), r) = (x_1(t), \ldots, x_{k-1}(t), r, x_{k+1}(t), \ldots, x_n(t)).$$

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:

$$F(t,r) = \begin{cases} f_k(\check{x}(t), r), & t \in [0, t_0) \\ f_k(\check{x}(t_0), r), & t \in [t_0, +\infty). \end{cases}$$

Then, for each fixed t, $F(t,r)$ is locally Lipschitz in r and the Lipschitz constants, $\alpha(t)$, are uniformly bounded (and hence locally integrable as a function of time). In addition, for each fixed r, $F(t,r)$ 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

$$\begin{aligned} f_k(\check{x}, r) &= \sum_{i=1}^{m} \sum_{j \in A_0} a_{ij} \theta_1(x_1)^{b_{1j}} \ldots \theta_{k-1}(x_{k-1})^{b_{(k-1)j}} \theta_{k+1}(x_{k+1})^{b_{(k+1)j}} \ldots \theta_n(x_n)^{b_{nj}} b_{ki} \\ &+ \sum_{i=1}^{m} \sum_{j \in A_+} a_{ij} \theta_1(x_1)^{b_{1j}} \ldots \theta_{k-1}(x_{k-1})^{b_{(k-1)j}} \theta_k(r)^{b_{kj}} \\ &\qquad \theta_{k+1}(x_{k+1})^{b_{(k+1)j}} \ldots \theta_n(x_n)^{b_{nj}} (b_{ki} - b_{kj}) \end{aligned}$$

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

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

(a) the second term is zero since $r = 0$;

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

This proves the claim.

Moreover, notice that, for all $t \leq t_0$, the scalar variable $x_k(t)$ satisfies the initial value problem

$$\begin{aligned} \dot{r} &= F(t,r) \\ r(0) &= x_k(0), \end{aligned}$$

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 $x_k(t)$ solves the equation in that

interval. Then, by Lemma C.0.2, $r(t) > 0$ on $\tilde{J}$ and $x_k(t) = r(t) > 0$ for all $t < t_0$; since both $r(t)$ and $x_k(t)$ are continuous functions, we also have that $x_k(t_0) = r(t_0)$, contradicting the fact that $x_k(t_0) = 0$.

This concludes the proof of $\mathbb{R}^n_{>0}$-invariance.

Corollary 2.5.6 (Corollary VII.7 in [48]) Consider a system (3.1), and pick any $\xi \in \mathbb{R}^n_{\geq 0}$. Then, either $\xi \in E_0$ or ξ belongs to some positive class.

2.5.2 Proof of Theorem 1

Assume that there are no boundary equilibria in each positive class, i.e., $\mathcal{S}_{\bar{x}} \cap E_0 = \emptyset$, for every $\bar{x}$.

Lemma 2.5.7 Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be the function defined in (2.2), and given any $\bar{x} \in E_+$, let V be the function defined in (2.8). Then there exists a continuous positive definite function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $\alpha = \alpha_{\bar{x},A}$, such that,

$$\nabla V(x, \bar{x})\ f(x) \leq -\alpha(V(x, \bar{x}))$$

for all $x \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$.

Proof. Let the function $c : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ and the constant $\kappa(A)$ be as given in Lemma 2.5.4. We will show that the following function, defined from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ is positive definite:

$$\alpha(r) = \inf\{\kappa(A)\, c(\bar{x})\, \Psi(x, \bar{x}) :\ V(x, \bar{x}) = r,\ x \in \mathcal{S}_{\bar{x}}\}.$$

The set $\mathcal{C}_r = \{x \in \mathcal{S}_{\bar{x}} :\ V(x, \bar{x}) = r\}$ is compact, because V is proper and $\mathcal{S}_{\bar{x}}$ is a closed subset of $\mathbb{R}^n_{\geq 0}$. It is clear $r = 0$ implies $\mathcal{C}_r = \{(\bar{x}, \bar{x})\}$, so $\alpha(0) = 0$.

Next, we take any $r > 0$ and show that $\inf\{\Psi(x, \bar{x}) : x \in \mathcal{C}_r\}$ is positive. To get a contradiction, assume that this infimum is zero for some $r > 0$. Then there exists an infinite sequence $\{x^k\}$ such that $\Psi(x^k, \bar{x}) \to 0$. Since $\mathcal{C}_r$ is compact, there exists a converging subsequence: $x^{k_l} \to x_0 \in \mathcal{C}_r$. Then $\Psi(x_0, \bar{x}) = 0$ and by Lemma 2.4.4, $x_0 \in E_0 \cup E_+$. But, under the no boundary equilibrium assumption, $E_0 \cap \mathcal{S}_{\bar{x}} = \emptyset$. So, x_0 cannot be in E_0. By uniqueness of the positive equilibrium in each class, $x_0 = \bar{x}$ which implies $r = V(x_0, \bar{x}) = 0$ and contradicts $r > 0$. Thus, $\alpha(r) > 0$ whenever $r > 0$.

By construction α satisfies

$$\kappa(A)\, c(\bar{x})\, \Psi(x, \bar{x}) \geq \alpha(V(x, \bar{x})) \tag{2.14}$$

for all $x \in \mathcal{S}_{\bar{x}}$. Without loss of generality we may assume that α is continuous on $\mathbb{R}_{\geq 0}$. (Otherwise, it is possible to construct another positive definite function $\tilde{\alpha}$, continuous and satisfying $\alpha(r) \geq \tilde{\alpha}(r)$ for all $r \geq 0$: note that, for every r, there exists κ_r such that $\alpha(s) \geq \kappa_r$ for all $s \in [\frac{r}{2}, 2r]$.)

Finally, by Lemma 2.5.4 we know that

$$\nabla V(x, \bar{x})\ f(x) = \langle \rho(x) - \rho(\bar{x}), f(x) \rangle \leq -\kappa(A)\, c(\bar{x})\, \Psi(x, \bar{x}),$$

for all $x \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$ and so (2.14) gives:

$$\nabla V(x, \bar{x})\ f(x) \leq -\alpha(V(x, \bar{x})),$$

for all $x \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$, and α is positive definite and continuous as wanted. ∎

Corollary 2.5.8 There exists a function $\beta = \beta_{\bar{x},A}$ of class $\mathcal{KL}$ such that for every $q \in \mathcal{S}_{\bar{x}}$, the solution $x(t, q)$ of the initial value problem $\dot{x} = f(x)$, $x(0) = q$, satisfies

$$|x(t, q) - \bar{x}| \leq \beta(|q - \bar{x}|, t),$$

for all $t \geq 0$.

Proof. Let $\alpha = \alpha_{\bar{x},A}$ be the positive definite function given by Lemma 2.5.7. Consider the initial value differential inequality

$$\dot{y} \leq -\alpha(y), \quad y(0) = y_0 \geq 0.$$

Then, by a comparison result such as Lemma 4.4 of [37], there exists a function $\tilde{\beta} = \tilde{\beta}_\alpha$ of class $\mathcal{KL}$ such that, for every solution of this differential inequality,

$$y(t) \leq \tilde{\beta}(y_0, t),$$

for all $t \geq 0$. Using the $\mathcal{K}_\infty$ bounds ν_1, ν_2 of V, define

$$\beta(r, t) = \nu_1^{-1}(\tilde{\beta}(\nu_2(r), t))$$

which is again a $\mathcal{KL}$ function and depends only on $\bar{x}$ and A.

First, given any $q \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$, let $x(t, q)$ be the solution of the initial value problem $\dot{x} = f(x)$, $x(0) = q$, and recall that $x(t, q) \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$ for all $t \geq 0$, since $\mathbb{R}^n_{>0}$ is a forward invariant for the system $\dot{x} = f(x)$ (Section 2.5.1).

Then, we may define $y(t) := V(x(t, q), \bar{x})$, and fom Lemma 2.5.7 the function $y(t)$ satisfies, for all $t \geq 0$, $\dot{y} \leq -\alpha(y)$, $y(0) = V(q, \bar{x})$. Therefore, recalling property (ii) of V, we have

$$\nu_1(|x(t, q) - \bar{x}|) \leq V(x(t, q), \bar{x}) \leq \tilde{\beta}(\nu_2(|q - \bar{x}|), t)$$

for all $t \geq 0$. This finishes the proof of the Corollary for initial conditions $q \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$.

Second, consider any $q \in \mathcal{S}_{\bar{x}} \cap \partial\mathbb{R}^n_{>0}$, and let $\{\xi_k \in \mathbb{R}^n_{>0} : k = 1, 2, \ldots\}$, be a sequence so that $\xi_k \to q$. Then we have

$$|x(t, \xi_k) - \bar{x}| \leq \beta(|\xi_k - \bar{x}|, t),$$

for all $t \geq 0$ and all k. By continuity of solutions of a differential equation on parameters (see [46], for example), and by continuity of β, we finally have

$$|x(t, q) - \bar{x}| \leq \beta(|q - \bar{x}|, t),$$

as we wanted to prove. ∎

Proof of Theorem 1: An alternative proof of Theorem 1 can now be given. Consider the initial value problem

$$\dot{x} = f(x), \quad x(0) = x_0, \quad \text{with} \quad x_0 \in \mathbb{R}^n_{\geq 0}.$$

By Corollary 2.5.6, either $x_0 \in E_0$, in which case $x(t) \equiv x_0$ for all $t \geq 0$, or x_0 belongs to some positive class, say $\mathcal{S}$. In the latter case, let $\bar{x}$ be the unique element in $\mathcal{S} \cap E_+$.

From Corollary 2.5.8, the solution of the initial value problem $\dot{x} = f(x)$, $x(0) = x_0$ satisfies

$$|x(t, x_0) - \bar{x}| \leq \beta(|x_0 - \bar{x}|, t),$$

for all $t \geq 0$, where $\beta \in \mathcal{KL}$.

To prove convergence of the trajectory to the point $\bar{x}$, given any $\varepsilon > 0$ find T so that

$$\beta(|x_0 - \bar{x}|, t) < \varepsilon, \quad \text{for all} \quad t \geq T$$

(such T exists by the property of $\mathcal{KL}$ functions). So we conclude that $x(t, x_0) \to \bar{x}$.

To prove stability with respect to the point $\bar{x}$, given any $\varepsilon > 0$, pick $\delta > 0$ so that

$$\beta(\delta, 0) \leq \varepsilon.$$

Then for any $x_0 \in \mathcal{S}_{\bar{x}}$ with $|x_0 - \bar{x}| < \delta$ we have

$$|x(t, x_0) - \bar{x}| \leq \beta(|x_0 - \bar{x}|, t) \leq \beta(|x_0 - \bar{x}|, 0) \leq \beta(\delta, 0) < \varepsilon.$$

2.6 A Fact About Trajectories Near the Boundary

The next Lemmas give a characterization of the behavior of $\langle \rho(x) - \rho(\bar{x}), f(x) \rangle$ as x approaches the boundary. Since V is not differentiable at the boundary points (and ∇V is not even defined at the boundary), it is not clear what happens to the function $\nabla V(x, \bar{x})\ f(x) = W(x, \bar{x})$ near the boundary. The next Lemmas provide useful insight.

Recall that we defined the nonnegative scalar function $\omega(x) = e^x - 1 - x$, in expression (2.10).

Lemma 2.6.1 There exist functions $\tilde{\alpha}_1, \tilde{\alpha}_2 \in \mathcal{K}_\infty$ such that

$$|a| \leq \tilde{\alpha}_1(e^{|b|}) + \tilde{\alpha}_2(e^b \omega(a - b))$$

for every $a, b \in \mathbb{R}$.

Proof. Consider the function $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$

$$\tilde{\alpha}(r) = \max\{|x| : \omega(x) = r\},$$

which is a $\mathcal{K}_\infty$ function, since $\omega(x) = 0$ iff $x = 0$ and ω is strictly decreasing for $x < 0$ and strictly increasing for $x > 0$ and, moreover, $\lim_{x\to\pm\infty} \omega(x) = +\infty$. Then, for any $a, b \in \mathbb{R}$,

$$|a - b| \leq \tilde{\alpha}(\omega(a - b)) = \tilde{\alpha}(e^{-b}e^{b}\omega(a - b))$$

and now, using first the inequality $rs \leq \frac{1}{2}(r^2 + s^2)$ for any positive numbers r, s, and then the inequality $\tilde{\alpha}(r + s) \leq \tilde{\alpha}(2r) + \tilde{\alpha}(2s)$, valid for any $\mathcal{K}_\infty$ function $\tilde{\alpha}$,

$$\begin{aligned} |a - b| &\leq \tilde{\alpha}(\frac{1}{2}e^{-2b} + \frac{1}{2}e^{2b}\omega(a - b)^2) \\ |a| &\leq |b| + \tilde{\alpha}(e^{-2b}) + \tilde{\alpha}(e^{2b}\omega(a - b)^2). \end{aligned}$$

Finally, since $e^{-2b} \leq e^{2|b|}$ and also $|b| \leq e^{|b|}$, simply let $\tilde{\alpha}_1(r) = r + \tilde{\alpha}(r^2)$ and $\tilde{\alpha}_2(r) = \tilde{\alpha}(r^2)$. ∎

Lemma 2.6.2 Fix any $\bar{x} \in E_+$. Let W be the function defined in (2.9). Assume there exists $M > 0$ so that $|W(x, \bar{x})| \leq M$ for all x that belong to some set $\mathcal{X}_o \subset \mathbb{R}^n_{>0}$. Suppose furthermore, that there exists some j_0 and N_0 satisfying $|q_{j_0}| \leq N_0$, for all $x \in \mathcal{X}_o$.

Then, for each $j \in \{1, \ldots, m\}$, there exist constants N_j such that

$$|q_j| = |\langle b_j, \rho(x) - \rho(\bar{x})\rangle| \leq N_j,$$

for all $x \in \mathcal{X}_o$.

Proof. Without loss of generality, we may assume that $j_0 = 1$. Pick any $j \neq 1$. By irreducibility of A we can pick a sequence of integers, $k(0) = 1$, $k(1) = \bar{1}, k(2) = \bar{2}, \ldots, k(r) = \bar{j}$ and $k(r + 1) = j$ such that

$$a_{\bar{1}1}, a_{\bar{2}\bar{1}}, \ldots, a_{j\bar{j}} \neq 0.$$

Now, given any $s \in \{1, \ldots, r + 1\}$, the fact that $|q_{k(s-1)}| \leq N_{k(s-1)}$ together with the assumption $|W(x, \bar{x})| \leq M$, imply $|q_{k(s)}| \leq N_{k(s)}$, for some $N_{k(s)}$. Indeed, from expression (2.10) it follows that

$$a_{ij}\, e^{\langle b_j, \rho(\bar{x})\rangle}\, e^{q_j}\, \omega(q_i - q_j) < M$$

for all pairs $i, j = 1, \ldots, m$, and therefore

$$e^{q_{k(s-1)}}\omega(q_{k(s)} - q_{k(s-1)}) \leq \frac{M}{a_{k(s),k(s-1)}\, e^{\langle b_{k(s-1)}, \rho(\bar{x})\rangle}} \equiv M_{k(s-1)}.$$

Applying Lemma 2.6.1 with $b = q_{k(s-1)}$ and $a = q_{k(s)}$,

$$\begin{aligned} |q_{k(s)}| &\leq \tilde{\alpha}_1(e^{|q_{k(s-1)}|}) + \tilde{\alpha}_2(e^{q_{k(s-1)}}\omega(q_{k(s)} - q_{k(s-1)})) \\ &\leq \tilde{\alpha}_1(e^{N_{k(s-1)}}) + \tilde{\alpha}_2(M_{k(s-1)}) \equiv N_{k(s)}. \end{aligned}$$

Therefore, since for $s = 0$ we already have $N_{k(0)} = N_0$, induction on s gives $|q_k(s)| \leq N_k(s)$, for all $k(0) = 1, k(1), \ldots, k(r + 1) = j$, for all $x \in \mathcal{X}_o$. Since j is arbitrary, the proof is finished. ∎

Proposition 2.6.3 Let $y \in \partial\mathbb{R}^n_{>0} \setminus E_0$. Let $\{x_k \in \mathbb{R}^n_{>0} : k = 1, 2, \ldots\}$ be any sequence of points in the positive orthant such that

$$x_k \ \to \ y.$$

Then $W(x_k, \bar{x}) \to -\infty$ as $x_k \to y$.

Proof. Assume that the statement is false, i.e.

$$(\exists M > 0)(\forall 0 < k \in \mathbb{N})(\exists x_k \in \mathbb{R}^n_{>0}) \ \ |x_k - y| < \frac{1}{k} \ \text{ and } \ |W(x_k, \bar{x})| \leq M.$$

Let $J_0 = \{j : \pi_j(y, \bar{x}) = 0\}$ and $J_+ = \{j : \pi_j(y, \bar{x}) \neq 0\}$. The set J_0 is nonempty because $y \in \partial\mathbb{R}^n_{>0}$ and the set J_+ is also nonempty because otherwise, by Proposition 2.2.1, $y \in E_0$ which is false.

By irreducibility of A, there exist $\bar{\imath} \in J_0$ and $\bar{\jmath} \in J_+$ such that $a_{\bar{\imath},\bar{\jmath}} \neq 0$. By continuity of the map $\pi_{\bar{\jmath}}$ on $\mathbb{R}^n_{\geq 0}$, $\pi_{\bar{\jmath}}(x_k, \bar{x}) \to \pi_{\bar{\jmath}}(y, \bar{x}) > 0$ and $\pi_{\bar{\imath}}(x_k, \bar{x}) \to \pi_{\bar{\imath}}(y, \bar{x}) = 0$, as $k \to +\infty$. It follows that

$$\begin{aligned} &q_{\bar{\jmath}}(x_k, \bar{x}) \to q_{\bar{\jmath}}(y, \bar{x}) \\ &q_{\bar{\imath}}(x_k, \bar{x}) \to -\infty. \end{aligned} \tag{2.15}$$

Since $q_{\bar{\jmath}}(x_k, \bar{x})$ is convergent, it is bounded, say on $|x - y| < 1$, so there is an $N_0 > 0$ so that $|q_{\bar{\jmath}}(x_k, \bar{x})| \leq N_0$ for all k (large enough). Then, by Lemma 2.6.2, there exists a number $N_1 > 0$ so that $|q_{\bar{\imath}}(x_k, \bar{x})| \leq N_1$ for all k, hence contradicting (2.15). ∎

Corollary 2.6.4 Suppose there exists a sequence $\{x_k \in \mathbb{R}^n_{>0} : k = 1, 2, \ldots\}$ such that $x_k \to y \in \partial\mathbb{R}^n_{>0}$ and that $W(x_k, \bar{x}) \to L$, for some $-\infty < L \leq 0$. Then $y \in E_0$.

Chapter 3
Detectability and Stability Notions

Some basic notions are introduced. In particular, a definition of detectability is adopted, which resembles a steady-state observability condition, and is appropriate for our system. Easy to check necessary and sufficient conditions for the system to be detectable are given and proved.

A useful input-to-state stability (ISS) framework is established, extending the original ISS notion so as to allow one to deal with positive systems, relative equilibria and constrained inputs. A dissipation characterization (similar to the original characterization) of the ISS notion is given in terms of a Lyapunov function.

3.1 The Problem

The main objective of this thesis 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 discussed in the previous Chapter. Several stability and other control-theoretic results applicable to such systems were discussed and surveyed (but without outputs) in the paper [48] (see also [8]). The systems that we study all have the generic form

$$\dot{x} = f(x), \qquad y = h(x), \tag{3.1}$$

with the requirements on f and h specified next.

The function $f : \mathbb{R}^n \to \mathbb{R}^n$ is as defined in Section 2.1, but for simplicity, we restrict ourselves to the case where $\theta_i(r) = |r|$ for all $i = 1, \ldots, n$. In this case, the function f becomes of the mass-action kinetics form

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij} x_1^{b_{1j}} x_2^{b_{2j}} \ldots x_n^{b_{nj}} (b_i - b_j), \tag{3.2}$$

where $m \leq n$ and each b_j is a column vector in $\mathbb{R}^n$ and has entries $b_{1j}, b_{2j}, \ldots, b_{nj}$, which are nonnegative integers. As before, it is assumed that $B := (b_1, b_2, \ldots, b_m)$ 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.

From Chapter 2, we know that the positive orthant, $\mathbb{R}^n_{>0}$ is a forward invariant set for system (3.1). We will be interested only on those trajectories which evolve in $\mathbb{R}^n_{>0}$. For system (3.1), recall that E is the set of all equilibria, i.e., all elements $\bar{x} \in \mathbb{R}^n_{\geq 0}$ such that $f(\bar{x}) = 0$. Likewise, E_+ denotes the subset of E consisting of all strictly positive equilibria, and E_0 denotes the set of all boundary equilibria. Then $E = E_+ \cup E_0$ as a disjoint union. In [48], the following is proved:

Lemma 3.1.1 (Corollaries VII.5 and VII.7 in [48]) 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 $t > 0$.

Furthermore, as stated in Chapter 2, we will assume that no positive class has any boundary equilibria, i.e., $\mathcal{S}_{\bar{x}} \cap E_0 = \emptyset$, for all $\bar{x} \in E_+$.

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 $x_1 x_2$).

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

$$h(x) = \begin{pmatrix} x_1^{c_{11}} x_2^{c_{12}} \cdots x_n^{c_{1n}} \\ \vdots \\ x_1^{c_{p1}} x_2^{c_{p2}} \ldots x_n^{c_{pn}} \end{pmatrix}, \tag{3.3}$$

where

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ & & \ddots & \\ c_{p1} & c_{p2} & \cdots & c_{pn} \end{pmatrix}$$

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.)

Note that the vectorial function $\rho^{[n]} : \mathbb{R}^n_{>0} \to \mathbb{R}^n$ takes the form

$$\begin{aligned} \rho^{[n]}(x) &= (\ln \theta_1(x_1), \ldots, \ln \theta_n(x_n))' \\ &= (\ln x_1, \ldots, \ln x_n)', \end{aligned} \tag{3.4}$$

in the case when $\theta_i(r) = r$ for all $i = 1, \ldots, p$. We will drop the subscript $[n]$ and consider ρ as the (coordinatewise) vectorial logarithmic function.

Also observe that, for $x \in \mathbb{R}^n_{>0}$,

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

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

3.2 Detectability

We will adopt the following definition

Definition 3.2.1 The system (3.1) is *detectable* if, for every two trajectories x and $\tilde{x}$ such that $x(t)$ evolves in $\mathbb{R}^n_{>0}$ and $\tilde{x}(t)$ evolves in $\mathbb{R}^n_{\geq 0}$ and both are defined for all $t \geq 0$,

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

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

By reasons that will presently become clear, this definition is appropriate for our type of systems, which exhibit *multiple equilibria* and have *no inputs*.

However, this is not the strongest possible definition of detectability, because it is not "well posed" enough. In principle, one would want the definition of detectability to also include the property: "$h(x(t)) \approx h(\tilde{x}(t))$ for all t implies $|x(t) - \tilde{x}(t)|$ 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 [34] and [51], can also be studied, although we will not do so in this work.

Definition 3.2.2 By a (full-state) *observer* for (3.1) 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

$$|z(t) - x(t)| \to 0 \;\text{ as }\; t \to +\infty.$$

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 $\mathcal{KL}$ estimates.

Detectability on $\mathbb{R}^n_{>0}$ is, obviously, necessary for the existence of an observer: if the system is not detectable on $\mathbb{R}^n_{>0}$, then the output h doesn't distinguish between trajectories, asymptotically. That is, there exist two trajectories, x and $\tilde{x}$ evolving in $\mathbb{R}^n_{>0}$ and an $\varepsilon > 0$ so that

$$h(x(t)) \equiv h(\tilde{x}(t)) \;\text{ and }\; \exists\{t_k, k = 1, 2, \ldots\} \;\text{ so that }\; |x(t_k) - \tilde{x}(t_k)| > \varepsilon,\ \forall\, k.$$

Then, if an observer exists as in Definition 3.2.2, it will receive the same input ($h(x) \equiv h(\tilde{x})$) from both trajectories, and thus the estimated trajectory z will satisfy

$$|z(t) - x(t)| \to 0 \quad \text{and} \quad |z(t) - \tilde{x}(t)| \to 0,$$

which implies that also $|x(t) - \tilde{x}(t)| \to 0$, contradicting the fact that h doesn't distinguish between trajectories.

Detectability (in view of Lemma 3.1.1 and Theorem 1) consists of detectability on $\mathbb{R}^n_{>0}$ together with the statement that $h(x(t)) \not\equiv h(\xi)$ for every trajectory x evolving in $\mathbb{R}^n_{>0}$ and point $\xi \in E_0$ (since $|x(t) - \xi| \to |\bar{x} - \xi| > 0$ for some $\bar{x} \in E_+$). 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.

Lemma 3.2.3 For system (3.1), detectability is equivalent to:

$$[h(\bar{x}) = h(\bar{z}) \quad \text{and} \quad \bar{x} \in E_+,\, \bar{z} \in E] \quad \Rightarrow \quad \bar{x} = \bar{z} \tag{3.5}$$

and detectability on $\mathbb{R}^n_{>0}$ is equivalent to

$$[h(\bar{x}) = h(\bar{z}) \quad \text{and} \quad \bar{x} \in E_+,\, \bar{z} \in E_+] \quad \Rightarrow \quad \bar{x} = \bar{z}\,. \tag{3.6}$$

Proof. Suppose, first, 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 $h(x(t)) \equiv h(z(t))$ but have distinct limits as $t \to +\infty$, a contradiction. A similar conclusion holds if we assume that the system is detectable on $\mathbb{R}^n_{>0}$, and pick $\bar{x}, \bar{z} \in E_+$ ($\bar{x} \neq \bar{z}$).

Suppose, next, that (3.5) holds and pick any two trajectories $x(\cdot)$ and $z(\cdot)$ evolving in $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n_{\geq 0}$, respectively, such that $h(x(t)) \equiv h(z(t))$. 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 3.1.1, either $z(t) \equiv z(0) \in E_0$, or $z(t) \in \mathbb{R}^n_{>0}$ for all $t > 0$; for this second case, 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 (3.5) says $\bar{x} = \bar{z}$, as we wanted to prove.

Now, assuming (3.6) holds, and that the trajectory $z(\cdot)$ also evolves in $\mathbb{R}^n_{>0}$, then again by Theorem 1, $z(t) \to \bar{z} \in E_+$. So (3.6) implies $\bar{x} = \bar{z}$, meaning that the system is detectable on $\mathbb{R}^n_{>0}$. ∎

Corollary 3.2.4 The system (3.1) 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 3.2.5 A more symmetric detectability condition, corresponding to arbitrary trajectories in Definition 3.2.1, would be "$[h(\bar{x}) = h(\bar{z}) \quad \text{and} \quad \bar{x}, \bar{z} \in E] \quad \Rightarrow \quad \bar{x} = \bar{z}$", 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 4.1.1, the second component of the output satisfies $h_2(x) \equiv 0$ 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 $h(\bar{x}) = h(\bar{z})$ becomes just

$$\rho(\bar{x}) - \rho(\bar{z}) \in \ker C\,.$$

Recall also from Corollary 2.4.1: if $\bar{x} \in E_+$, then, for any $\bar{z} \in \mathbb{R}^n_{>0}$,

$$\rho(\bar{x}) - \rho(\bar{z}) \in \mathcal{D}^\perp \quad \Longleftrightarrow \quad \bar{z} \in E_+\,. \tag{3.7}$$

Theorem 2 The following statements are equivalent:

(a) The system (3.1) with f as in (3.2) and h as in (3.3) 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 (3.6) 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

$$\begin{aligned} \tilde{y} &= \rho(\bar{x}) - y \;\in \mathbb{R}^n, \\ z &= \operatorname{Exp}(\tilde{y}) \in \mathbb{R}^n_{>0}, \end{aligned}$$

so that $\tilde{y} = \rho(z)$. Thus

$$y = \rho(\bar{x}) - \rho(z) \;\text{ with }\; \bar{x} \in E_+ \;\text{ and }\; z \in \mathbb{R}^n_{>0}.$$

By definition of y, $\rho(\bar{x}) - \rho(z)$ is contained both in $\mathcal{D}^\perp$ and in $\ker C$. Condition (3.7) 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 (3.7), 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. ∎

Remark 3.2.6 A useful sufficient condition, on the matrix C, for system (3.1) 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 $i = 1, \ldots, p$ and all $\bar{x} \in E_+$, the condition

$$(\forall\ \bar{z} \in E_0)\ (\exists\ l \in \{1, \ldots, p\})\ h_l(\bar{z}) = 0, \tag{3.8}$$

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

From Proposition 2.2.1 every boundary equilibrium $\bar{z}$ satisfies $\bar{z}_1^{b_{1j}} \bar{z}_2^{b_{2j}} \ldots \bar{z}_n^{b_{nj}} = 0$ for all $j = 1, \ldots, m$. An easy way to satisfy (3.8) 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 Section 3.1, *choose one of the measured quantities to be one of the reaction rates.* For instance,

$$h_l(x) = x_1^{b_{1j}} x_2^{b_{1j}} \ldots x_n^{b_{nj}} \tag{3.9}$$

for some $1 \leq l \leq p$ and some $1 \leq j \leq m$. (In fact, in this case, $h_l(\bar{z}) = 0$ for all $z \in E_0$, where l is the column in question.) Since $\dim \mathcal{D} = m - 1$ and $\operatorname{rank} B = m$, it is very easy to construct C so that both condition (3.8) and $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$ are satisfied, and thus system (3.1) detectable.

Remark 3.2.7 Note that $\dim \mathcal{D} = m - 1$ and that detectability implies $m - 1 + \operatorname{rank} C' \geq n$. In the case the matrix C has rank p, then $\operatorname{rank} C' = p$ and detectability implies $m - 1 + p \geq n$, so that $p \geq n - m + 1 \geq 1$.

3.3 An ISS Property

The definition of an input-to-state stable (ISS) system was introduced in [44]. 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

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

Definition 3.3.1 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 3.3.2 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

$$|z(t) - \bar{x}| \leq \beta(|z_0 - \bar{x}|, t) + \varphi(\|u - \bar{u}\|) \tag{3.10}$$

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 3.3.3 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 3.3.4 This definition differs slightly from other definitions of "ISS-Lyapunov" functions, such as given in [49]. 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 3.3.5 For a function V defined as above, there always exist $\mathcal{K}_\infty$ functions, ν_1, ν_2, such that

$$\nu_1(|z - \bar{x}|) \leq V(z) \leq \nu_2(|z - \bar{x}|) \tag{3.11}$$

for all $z \in \mathbb{R}^n_{\geq 0}$ (see Appendix B for a construction of these bounds).

Next, we introduce our candidate ISS-Lyapunov function. Fix an $\bar{x} \in E_+$, and observe that the function $V : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$, defined in (2.8), in the case of $\theta_i(r) = |r|$ for all $i = 1, \ldots, n$ reduces to

$$V(z, \bar{x}) = \sum_{i=1}^{n} \bar{x}_i \; g\left(\frac{z_i}{\bar{x}_i}\right) = \sum_{i=1}^{n} \bar{x}_i \left[\frac{z_i}{\bar{x}_i} \ln \frac{z_i}{\bar{x}_i} + 1 - \frac{z_i}{\bar{x}_i}\right] \tag{3.12}$$

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 it is easy to see that it does satisfy properties (i) and (ii) of Definition 3.3.3 (cf. Section 2.5 in Chapter 2).

In the next lemma, we show that the existence of an ISS-Lyapunov function (according to Definition 3.3.3) for a given system, implies that the trajectories of that system satisfy an ISS estimate.

Lemma 3.3.6 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 [49], 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 3.3.5 there exist two class $\mathcal{K}_\infty$ functions, ν_1, ν_2, such that

$$\nu_1(|z - \bar{x}|) \leq V(z) \leq \nu_2(|z - \bar{x}|), \; \forall \; z \in F.$$

We define new functions $\chi, \varphi \in \mathcal{K}_\infty$ by

$$\chi = \alpha^{-1} \circ (2\gamma) \quad \text{and} \quad \varphi = \nu_1^{-1} \circ \nu_2 \circ \chi,$$

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 the initial value problem

$$\dot{y} = -\frac{1}{2}\alpha(\nu_2^{-1}(y)),\ y(0) = r.$$

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

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

and, by property (ii) of Definition 3.3.3, $z(t)$ 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:

$$\begin{aligned} & |z - \bar{x}| > \chi(|u - \bar{u}|) \Leftrightarrow \alpha(|z - \bar{x}|) > 2\gamma(|u - \bar{u}|) \\ \Rightarrow\ & \nabla V(z)\ f^*(z, u) < -\frac{1}{2}\alpha(|z - \bar{x}|) \end{aligned} \tag{3.13}$$

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 (3.13), $\frac{d}{dt}V(z(t))|_\tau < 0$, 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 $t \in (T, t_{\max}]$: $V(z(t)) \le \nu_2(\chi(\|u - \bar{u}\|))$ implies

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

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

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

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

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

Thus, for all $t \in I$,

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

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.

∎

Chapter 4
Constructing Observers

Under the assumption that the system $\dot{x} = f(x)$, $y = h(x)$ is detectable, an observer is now explicitly constructed. As stated in Chapter 3, we will only study systems (2.2) in the special case of maps of the form $\theta_i(r) = r$, i.e., systems of the form (3.2). We also only consider systems (3.2) for which every trajectory evolving in $\mathbb{R}^n_{>0}$ converges to a positive equilibrium point, as is the case of systems that have no boundary equilibria in each positive class (as studied in Section 2.3).

The observer is globally convergent and, due to ISS properties, it turns out to be remarkably robust in many ways: with respect to observation noise, unknown (small) inputs, and also output sampling.

In order to streamline the presentation, we prefer to first prove the results for systems with "single linkage" classes (recall from Chapter 2, that this amounts to saying that the matrix A is *irreducible*, while "multiple linkage" corresponds to the matrix A being *block-irreducible*). Then, in Section 4.6, we show how to generalize the results to systems with "multiple linkage" classes.

4.1 An Observer

It has been shown in [48] that the system (3.1) can be steered to any of its equilibrium points in E_+ by applying additive error feedback. Namely, for any specific element $\bar{x} \in E_+$, a set K of indices such that

$$\mathcal{D} + \text{span}\{e_k : k \in K\} = \mathbb{R}^n,$$

where $\{e_k : k = 1, \ldots, n\}$ is the canonical basis of $\mathbb{R}^n$, and any positive constants $\gamma_k, k \in K$, the following system satisfies $\lim_{t\to\infty} |x(t) - \bar{x}| = 0$ for any initial conditions $x(0) \in \mathbb{R}^n_{>0}$:

$$\dot{x} = f(x) + \sum_{k\in K} \gamma_k(\bar{x}_k - x_k)e_k. \tag{4.1}$$

So, a natural starting point in constructing an observer for $\dot{x} = f(x)$, would be to take a copy of it and add a linear feedback term similar to the above:

$$\dot{z} = f(z) + \sum_{k\in K} \gamma_k(x_k - z_k)e_k.$$

We know that eventually the trajectories of the system we want to observe will satisfy $x(t) \to \bar{x}$, for some $\bar{x} \in E_+$, so it seems reasonable to expect that the dynamics of the

z-system will become very close to the dynamics of the modified x-system, (4.1), as time tends to infinity.

To construct such an observer, we need the "right output", that is, a function $h(x)$ that allows the construction of the suitable $\gamma_k(x_k - z_k)e_k$ terms. Thus, this approach will only work when the output map is linear and of a very special form. For instance, if there exists a matrix L such that

$$Ly = LCx = \sum_{k \in K} x_k e_k$$

where K is such that $\mathcal{D} + \text{span}\{e_k : k \in K\} = \mathbb{R}^n$ (for detectability), then the observer suggested above can be constructed (using Ly instead of y as input to the observer).

But in the general case such a matrix L does not exist and a different, but related, construction will be used. The main result is stated next and its proof is given in Section 4.4.

Assume that every trajectory, $x(\cdot)$, of the system $\dot{x} = f(x)$ (when f is given by (3.2)) satisfies: if $x(t) \in \mathbb{R}^n_{>0}$ for every $t > 0$, then $x(t) \to \bar{x} \in E_+$ as $t \to +\infty$. This is the case of systems for which no boundary equilibria exist in any positive class (see Section 2.3 and Theorem 1).

Theorem 3 Consider the system (3.1) and assume that it is detectable. Then the following system, with state space $\mathcal{X} = \mathbb{R}^n$, is an observer for the system (3.1):

$$\dot{z} = f(z) + C'(h(x) - h(z)). \tag{4.2}$$

Example 4.1.1 Consider once again the "McKeithan network" with $n = 4$, the system introduced in Example 2.1.1. The positive classes are characterized by

$$x_1 + x_3 + x_4 = \alpha, \quad x_2 + x_3 + x_4 = \beta$$

for each pair of *positive constants* α, and β, and the set of boundary equilibria of the system is

$$E_0 = \{(r, 0, 0, 0), (0, s, 0, 0) \in \mathbb{R}^4_{\geq 0} : r, s \geq 0\}.$$

I. Suppose that the output is given by $h(x) = (x_1, x_4)'$. Clearly, the matrix C corresponding to this output satisfies both part (c) of Theorem 2 and condition (3.8), since $x_4 =$ for all $x \in E_0$. We can take the following observer:

$$\begin{aligned}
\dot{x}_1 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 + (x_1 - z_1) \\
\dot{x}_2 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 \\
\dot{x}_3 &= k_{21}x_1x_2 - (k_{12} + k_{32})x_3 \\
\dot{x}_4 &= k_{32}x_3 - k_{13}x_4 + (x_4 - z_4).
\end{aligned}$$

II. 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 (3.8)). Then, we can construct the following observer:

$$\begin{aligned}
\dot{x}_1 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 + (x_1x_2^2 - z_1z_2^2) + (x_1x_4 - z_1z_4) \\
\dot{x}_2 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 + 2(x_1x_2^2 - z_1z_2^2) \\
\dot{x}_3 &= k_{21}x_1x_2 - (k_{12} + k_{32})x_3 \\
\dot{x}_4 &= k_{32}x_3 - k_{13}x_4 + (x_1x_4 - z_1z_4).
\end{aligned}$$

The remainder of this section will be devoted to proving this theorem. The basic idea is to study the stability properties of system (3.1) when a certain input is added to the function f, specifically, the system with right-hand side:

$$f^*(z,u) := f(z) + C'(u - h(z)). \tag{4.3}$$

We will show that, for the system thus obtained, an "input to state stability" condition holds. The observer for (3.1) 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.

4.2 Useful Estimates

The ISS property of (4.3) will be established by showing that this system admits V as an ISS-Lyapunov function. In order to do this, appropriate estimates for $\nabla V(z)\, f^*(z,u)$ are essential.

Lemma 4.2.1 Let $c_0, c_1 > 0$ be constants and fix any $\bar{x} \in E_+$. Let h be a function of the form (3.3) such that $\dot{x} = f(x), y = h(x)$ is detectable. Then, the function $\mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$,

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

vanishes only at $z = \bar{x}$.

Moreover, given any compact subset $F \subset \mathbb{R}^n_{\geq 0}$ containing the point $\bar{x}$, 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 2.4.4, $\Psi(z,\bar{x}) = 0$ 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

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

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. ∎

Lemma 4.2.2 For every compact set F in $\mathbb{R}^n_{\geq 0}$, there exists a constant $c_F > 0$ such that for every $z \in F \cap \mathbb{R}^n_{>0}$:

$$-\langle C(\rho(z) - \rho(\bar{x})), h(z) - h(\bar{x})\rangle \leq -c_F \, |h(z) - h(\bar{x})|^2. \tag{4.4}$$

Proof. Recall the form of the function h and using (3.4), 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 (4.4), 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 $c_F = \kappa$. ∎

Proposition 4.2.3 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 (3.12), 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

$$\gamma(r) = c_3 r^2, \text{ with } c_3 = \frac{2}{\theta\lambda},$$

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

$$\begin{aligned} \sigma &= C\varrho, \\ \mu &= h(z) - h(\bar{x}), \\ v &= u - h(\bar{x}), \end{aligned}$$

the expression for $\nabla V(z)\ f^*(z,u)$ becomes

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

where

$$P(z,\bar{x}) = \langle \varrho, f(z)\rangle - (1-\theta)\langle \sigma, \mu\rangle$$

and

$$R(z,u,\bar{x}) = -\theta\langle \sigma,\mu\rangle + \langle \sigma, v\rangle = \sum \sigma_i(-\theta\mu_i + v_i)$$

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

$$\lambda = \min\{h_i(\bar{x})/2 : i = 1,\dots,p\}$$

(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

$$\gamma(r) = c_3 r^2, \quad \text{where} \quad c_3 = \frac{2}{\theta\lambda}.$$

Step 2. Show that $P(z,\bar{x}) \le -\alpha_1(|\rho(z) - \rho(\bar{x})|)$, where α_1 is positive definite. From equation (2.13) 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 $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$

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

where

$$\mathcal{C}_r := \{z \in \mathbb{R}^n_{>0} : |\rho(z) - \rho(\bar{x})| = r\}.$$

This function 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 Proposition.

Step 3. Assume now that $F \subset \mathbb{R}^n_{\geq 0}$ is an arbitrary compact set. Then, using both (2.13) and (4.4), we have that

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

By Lemma 4.2.1, there exists a function $\alpha = \alpha_{\bar{x},F}$, of class $\mathcal{K}_\infty$, such that

$$P(z,\bar{x}) \leq -\alpha(|z - \bar{x}|),$$

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

4.3 Invariance and Completeness

Proposition 4.3.1 Suppose that the system defined by (3.2) and (3.3) is detectable. Consider the system with inputs

$$\dot{z} = f^*(z,u) := f(z) + C'(u - h(z)) \tag{4.5}$$

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 Proposition 4.2.3. Then, the system (4.5) 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 $\mathbb{R}^p_{\geq 0}$-valued input, let $z(t)$ be the maximal solution of (4.5), 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 $\mathbb{R}^p_{\geq 0}$. 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

$$\begin{aligned} f_k^*(\check{z}, x; u) &= \sum_{i=1}^{m} \sum_{j \in A_0} a_{ij} z_1^{b_{1j}} \ldots z_{k-1}^{b_{(k-1)j}} z_{k+1}^{b_{(k+1)j}} \ldots z_n^{b_{nj}} b_{ki} \\ &+ \sum_{i=1}^{m} \sum_{j \in A_+} a_{ij} z_1^{b_{1j}} \ldots z_{k-1}^{b_{(k-1)j}} x^{b_{kj}} z_{k+1}^{b_{(k+1)j}} \ldots z_n^{b_{nj}} (b_{ki} - b_{kj}) \qquad (4.6) \\ &+ \sum_{j=1}^{p} c_{jk}[u_j - h_j(\check{z}, x)], \end{aligned}$$

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

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

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 C.0.2, $x(t) > 0$ 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 (4.5) is also $\mathbb{R}^n_{>0}$-invariant with (the smaller) input value set $\mathbb{U}_\theta$. To prove that (4.5) is semi-global ISS with input value set $\mathbb{U}_\theta$, it is enough, by Lemma 3.3.6, to show that this system admits the function V defined in (3.12) 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 3.3.3 have already been shown in the discussion following formula (3.12). Property (iii) follows from Proposition 4.2.3, as well as the fact that γ may indeed be chosen independently of the set F. ∎

This Proposition shows the that system (4.5) is semi-global ISS with input set $\mathbb{U}_\theta$, which is sufficient for our purposes: as will be seen below, the inputs that we are interested in will eventually be in $\mathbb{U}_\theta$, i.e., for all $t \geq t_0$, $|u(t)| \in \mathbb{U}_\theta$. However, when $t_0 > 0$, we need to worry about the solution of (4.5) for $0 < t < t_0$, that is we need to guarantee that solutions are defined up to t_0. Thus we establish a $\mathbb{R}^n_{\geq 0}$-completeness result for system (4.5) with input-value set $\mathbb{R}^p_{\geq 0}$.

Lemma 4.3.2 Let $f^*(z,u) = f(z) + C'(u - h(z))$ and let V denote the function defined in (3.12). 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 (2.13),

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

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

$$|\ln h_i(z) - \ln h_i(\bar{x})| \leq c_L |h_i(z) - h_i(\bar{x})|,$$

for all z such that $h_i(z) > 1$. So we have,

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

In either case, we may just write

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

whenever $i \in I_+$. Then, with

$$c_{\bar{x},u_{\max}} = p \max\{u_{\max} h_{\max} + 1/e + h_{\max}, 2c_L(u_{\max}^2 + h_{\max}^2)\},$$

it follows that

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

as we wanted to show. ∎

Corollary 4.3.3 Under the assumptions of Proposition 4.3.1, system (4.5) is $\mathbb{R}^n_{\geq 0}$-complete with input-value set $\mathbb{R}^p_{\geq 0}$.

Proof. Suppose that $u(\cdot)$ is an $\mathbb{R}^p_{\geq 0}$-valued input, and first pick any initial condition $z(0) \in \mathbb{R}^n_{>0}$. We already know, from Proposition 4.3.1, that system (4.5) is $\mathbb{R}^n_{>0}$-invariant with input-value set $\mathbb{R}^p_{\geq 0}$. Suppose that the maximal interval of existence would be $[0,t_{\max})$ with $t_{\max} < +\infty$. Put

$$u_{\max} = \text{ess. sup}.\{|u(t) - \bar{u}| : 0 \leq t \leq t_{\max}\}.$$

From Lemma 4.3.2 we have that

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

So

$$V(z(t)) \leq V(z(0)) + c_{\bar{x},u_{\max}} t_{\max}.$$

Since V is proper (property (ii) of Definition 3.3.3), 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 $z(0) \in \mathbb{R}^n_{\geq 0}$, 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

$$\dot{v} = f^*(v, u), \quad v(0) = \xi_k,$$

at time t. For each k, the argument above holds and so

$$V(z^k(t)) \leq V(\xi_k) + c_{\bar{x}, u_{\max}} t_{\max},$$

for all $t < t_{\max}$. By continuity of solutions of differential equations on the initial conditions, we have for each $t \in [0, t_{\max})$,

$$V(z(t)) \leq V(z(0)) + c_{\bar{x}, u_{\max}} t_{\max}.$$

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. ∎

4.4 Proof of Theorem 3

Pick any initial states $x(0) \in \mathbb{R}^n_{>0}$ and $z(0) \in \mathbb{R}^n_{>0}$ of the original system (3.1) 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 $t \geq 0$ and converges to some equilibrium $\bar{x}$ as $t \to +\infty$, we must prove that $z(t)$ is also defined for all $t \geq 0$ and converges to this same $\bar{x}$ as $t \to +\infty$.

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

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

for all $i = 1, \ldots, p$. Let $\mathbb{U}_\theta$ be the set of vectors u such that

$$|u_i - h_i(\bar{x})| \leq \frac{\theta}{2} h_i(\bar{x}).$$

Next, pick $T \geq T_0$ so large that the convergence $x(t) \to \bar{x}$ becomes exponential (such T exists, as shown in [48]). 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 4.3.3 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 Proposition 4.2.3 (where we may pick $\gamma(r) = c_3\, r^2$) and the discussion above imply that

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

with

$$c_4 = c_3 (c_0 c_1 |x(T) - \bar{x}|)^2,$$

and integrating we obtain

$$V(z(t)) \leq V(z(T)) + \frac{c_4}{2c_2} e^{-2c_2 T},$$

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 funtions $\beta = \beta_F$ and $\varphi = \varphi_F$ as given by Definition 3.3.2.

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 4.3.1, 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 4.3.1 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 $|h(x(t)) - h(\bar{x})| \to 0$ 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 $t \to +\infty$, which finishes the proof.

Remark 4.4.1 A different proof, analogous to a separation principle, can be given in the case when at least one of the measurements is a reaction rate, i.e., condition (3.9) is verified. In this case, the system with inputs (4.5) is (*globally*) ISS (this fact is proved in Section 5.2). The proof of this separation principle is given in Appendix A.

4.5 Remarks on the Outputs

4.5.1 Observation Noise

The ISS estimate obtained in Proposition 4.3.1 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

$$u(t) = h(x(t)) + \eta(t)$$

instead of $h(x(t))$, then the same conclusions regarding global existence of trajectories hold, at least provided that:

1. $h(x(t)) + \eta(t)$ is nonnegative;

2. The issuing trajectory of system (4.5) is still confined to some compact subset of $\mathbb{R}^n_{\geq 0}$ (such as for white noise, for instance).

Then the following ISS estimate holds:

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

In addition, for any $\varepsilon > 0$, we can find T large enough so that

$$\varphi(2\|h(x) - h(\bar{x})\|) \leq \varepsilon$$

and

$$\beta(|z(T) - \bar{x}|, t) \leq \varepsilon$$

for all $t \geq T$. So we have (using the triangle inequality and the fact that φ is $\mathcal{K}_\infty$)

$$|z(t) - \bar{x}| \leq 2\varepsilon + \varphi(2\|\eta\|).$$

Thus one obtains an asymptotic estimate on $z(t)$ which is a $\mathcal{K}_\infty$ function of the noise level, and is in particular small when η is small in magnitude.

4.5.2 Sampled Outputs

One of the problems that may arise in the implementation of software to be used with output from chemical reactions is the delay in the measurement of species' concentrations. In sampled observers, it may happen that there are long sampling intervals, in which case one would like to know how this affects the observers and whether they still provide a good estimate of the state of the system.

Since we are dealing with continuous-time systems, and the output measurements are given at regular (and possibly long) time intervals, the available data will be of the form:

$$t_0, y_0 = h(x(t_0)); \quad t_1, y_1 = h(x(t_1)); \ \ldots.$$

Then, one way to interpret the input that actually goes into the observer is to construct a piecewise constant function of the form

$$s(t) = y_i \text{ if } t_i \leq t < t_{i+1}$$

for $i = 0, 1, \ldots$, so that the main observer becomes

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

In particular for the case of the output being sampled at intervals whose length has a lower bound and an upper bound (in fact, typically the sampling intervals have a constant length), we will show that solutions $z(t)$ of this system satisfy (for all t sufficiently large) an ISS estimate

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

for all $t \geq T$, where $\beta \in \mathcal{KL}$ and $\varphi \in \mathcal{K}_\infty$. Since

$$|s(t) - h(\bar{x})| \leq |s(t) - h(x(t))| + |h(x(t)) - h(\bar{x})|,$$

where this last term converges to zero, eventually, we may expect

$$|z(t) - \bar{x}| \lesssim \varphi(\|s - h(x)\|),$$

that is, the estimate provided by the observer will have an error of the order of the difference between the "sampled" and the continuous outputs of the system (using a similar argument to the one in Section 4.5.1). For our kind of systems $h(x(t)) \to h(\bar{x})$, so the function $s(t) - h(x(t))$ will be bounded and will converge to 0 as $t \to +\infty$. From the ISS estimate we may conclude that $z(t)$ still provides a good estimate and asymptotically converges to the correct value.

To see that the ISS estimate holds, note that the observer constructed with the sampled output is a system of the form (4.5) with $u(t) \equiv s(t)$. By Proposition 4.3.1 this system is semi-global ISS with respect to the point $\bar{x}$ and the input $h(\bar{x})$. Thus, to obtain the above (global) ISS estimate, it is enough to show that $z(t)$ evolves in some compact set $F \subset \mathbb{R}^n_{\geq 0}$ for all t.

To do this, first let T be so large that the convergence $x(t) \to \bar{x}$ becomes exponential (such T exists from [48]); there exist constants c_1, c_2 such that

$$|x(t) - \bar{x}| \leq c_1 |x(T) - \bar{x}| e^{-c_2 t}, \ \forall\, t \geq T.$$

Then, using also the fact that $x(t)$ evolves in a compact set for all t, and letting c_0 be a Lipschitz constant for the function h in that compact set,

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

for all $t \geq T$. The sampled output satisfies:

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

for all $t \in [t_i, t_{i+1})$ and $t_i \geq T$. Then, from Proposition 4.2.3, the following holds for each $t \in [t_i, t_{i+1})$:

$$\frac{d}{dt} V(z(t)) \leq c_3 |s(t) - h(\bar{x})|^2 \leq c_4 e^{-2c_2 t_i}$$

where $c_4 = c_3 (c_0 c_1 |x(T) - \bar{x}|)^2$. Suppose that $k, \bar{k}$ are two integers such that $T \in [t_k, t_{k+1})$ and $t \in [t_{\bar{k}}, t_{\bar{k}+1})$. Then

$$[T, t] = [T, t_{k+1}) \cup [t_{k+1}, t_{k+2}) \cup \ldots \cup [t_{\bar{k}}, t].$$

Integration on each of these intervals yields:

$$V(z(t)) \leq V(z(T)) + c_4 \int_T^{t_{k+1}} e^{-2c_2 t_k} + c_4 \sum_{i=k+1}^{\bar{k}-1} \int_{t_i}^{t_{i+1}} e^{-2c_2 t_i} + c_4 \int_{t_{\bar{k}}}^{t} e^{-2c_2 t_{\bar{k}}}.$$

Claim. If the length of each sampling interval satisfies

$$\delta_1 \leq t_{i+1} - t_i \leq \delta_2,$$

then there exists $d > 0$ such that $V(z(t)) \leq V(z(T)) + d$ for all $t \geq T$.

Note that

$$t_i \geq i\delta_1 + t_0 \quad \text{for each} \quad i = 0, 1, \ldots.$$

Then

$$\sum_{i=k}^{\bar{k}} \int_{t_i}^{t_{i+1}} e^{-2c_2 t_i} = \sum_{i=k}^{\bar{k}} e^{-2c_2 t_i}(t_{i+1} - t_i) \leq \delta_2 \sum_{i=k}^{\bar{k}} e^{-2c_2(i\delta_1 + t_0)}.$$

Using the formula for a geometric sum with ratio $e^{-2c_2\delta_1}$ it follows that

$$\sum_{i=k}^{\bar{k}} \int_{t_i}^{t_{i+1}} e^{-2c_2 t_i} \leq \delta_2 e^{-2c_2 t_0} \sum_{i=k}^{\infty} e^{-2c_2 i\delta_1} = \delta_2 e^{-2c_2 t_0} \frac{e^{-2c_2 k\delta_1}}{1 - e^{-2c_2\delta_1}} := d$$

where k is a fixed integer depending only on T and δ_1, δ_2 (in the case $\delta_1 = \delta_2 = \Delta t$, it turns out that $k = \texttt{floor}[T/\Delta t]$), and d is clearly independent of $\bar{k}$. Therefore, for all $t \geq T$, it follows that

$$V(z(t)) \leq V(z(T)) + d,$$

which proves the claim.

It is now clear (by properness of V), that indeed $z(t)$ evolves in a compact set for all $t \geq T$, as we wanted to show.

4.5.3 Weighted Outputs

The observer (4.2) can be slightly modified to the form

$$\dot{z} = f(z) + C'W(h(x) - h(z)) \tag{4.7}$$

where W is any positive definite, diagonal $p \times p$ matrix. This allows more flexibility when choosing the gain matrix, namely, it is possible to assign different weights to each component of $h(x)$. The conclusions of Proposition 4.3.1 still hold for (4.7). Since $W = \text{diag}(w_1, w_2, \ldots, w_p)$ with $w_i > 0$ and $W = S'S$ for

$$S = \text{diag}(\sqrt{w_1}, \sqrt{w_2}, \ldots, \sqrt{w_p}),$$

$\mathbb{R}^n_{>0}$-invariance is guaranteed by the same argument given in that Proposition. Expression (4.6) becomes

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

which is still nonegative when $x = 0$.

The function V in (3.12) is also an ISS-Lyapunov function for (4.7). To check property (iii) we compute (with $\varrho = \rho(z) - \rho(\bar{x})$ and $0 < \theta < 1$):

$$\begin{aligned} \nabla V(z)\ f^*(z,u) &= \langle \varrho, f(z)\rangle + \langle \varrho, C'S'S(u - h(z))\rangle \\ &= \langle \varrho, f(z)\rangle + \langle SC\varrho, S(u - h(z))\rangle \\ &= P(z,\bar{x}) + R(z,u,\bar{x}) \end{aligned}$$

where

$$\begin{aligned} P(z,\bar{x}) &= \langle \varrho, f(z)\rangle - (1-\theta)\langle SC\varrho, S(h(x) - h(z))\rangle \\ R(z,u,\bar{x}) &= \langle SC\varrho, S(u - h(\bar{x})\rangle - \theta\langle SC\varrho, S(h(x) - h(\bar{x}))\rangle. \end{aligned}$$

Using the notation

$$\begin{aligned} \sigma &= SC\varrho, \\ \mu &= S(h(z) - h(\bar{x}), \\ v &= S(u - h(\bar{x})), \end{aligned}$$

steps 1 and 2 of the proof of Proposition 4.2.3 still hold, and we obtain:

$$\begin{aligned} R(z,u,\bar{x}) &\leq c|v|^2 \\ P(z,\bar{x}) &\leq -\tilde{\alpha}_1(|\rho(z) - \rho(\bar{x})|). \end{aligned}$$

The positive definite function $\tilde{\alpha}_1$ is similar to the function α_1 in Proposition 4.2.3. The constant c is similar to c_3 of that Lemma: $c = 2c_L/\theta$, where c_L is a Lipschitz constant of the logarithmic function on $[\lambda, +\infty)$, when $\lambda = \min\{\sqrt{w_i}h_i(\bar{x})/2 : i = 1, \ldots, p\}$.

The completeness result is still valid and the trajectory $z(\cdot)$ still satisfies, for some T large enough and a constant d, $V(z(t)) \leq V(z(T)) + d$ for all $t \geq T$. Hence the ISS estimate holds and the same proof as above shows that (4.7) is an observer for system (3.1).

4.6 Generalization to Systems with "Multiple Linkage Classes"

We now look at the more general case of a system with vector field of the form (3.2) but where the matrix $A = \mathrm{diag}(A_1, \ldots, A_L)$ 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 (3.1) can be written as

$$\dot{x} = f_1(x) + \cdots + f_L(x)$$

where each $f_s(x)$ is computed according to formula (3.2) 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 $Gr(A)$. $Gr(A)$ 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 $m_s - 1$,

$$\mathcal{D} = \mathcal{D}_1 \oplus \cdots \oplus \mathcal{D}_L, \tag{4.8}$$

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

Example 4.6.1 To illustrate this structure, consider a general enzymatic mechanism with uncompetitive inhibitor (see [41]), consisting of one enzyme E, one substrate S, one product P and an uncompetitive inhibitor I (Q and R are intermediate complexes):

$$S + E \underset{k_{12}}{\overset{k_{21}}{\rightleftharpoons}} Q \underset{k_{23}}{\overset{k_{32}}{\rightleftharpoons}} P + E$$

$$Q + I \underset{k_{45}}{\overset{k_{54}}{\rightleftharpoons}} R$$

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 $Gr(A_1)$ 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 species, 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 [41] 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 $A = \text{ diag } (A_1, A_2)$:

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

and

$$A_1 = \begin{pmatrix} 0 & k_{12} & 0 \\ k_{21} & 0 & k_{23} \\ 0 & k_{32} & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & k_{45} \\ k_{54} & 0 \end{pmatrix}$$

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

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

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

$$f_1(x) = \begin{pmatrix} -k_{21}SE + k_{12}Q \\ -k_{23}PE + k_{32}Q \\ -(k_{12}+k_{32})Q + k_{21}SE + k_{23}PE \\ 0 \\ -k_{21}SE - k_{23}PE + (k_{32}+k_{12})Q \\ 0 \end{pmatrix}, \quad f_2(x) = \begin{pmatrix} 0 \\ 0 \\ -k_{54}QI + k_{45}R \\ -k_{45}R + k_{54}QI \\ 0 \\ -k_{54}QI + k_{45}R \end{pmatrix}.$$

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 (3.3). 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 [48]). The main fact to verify is that (3.7) still holds for the general case. But, from [48], 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 of A.) Then, for each s, (3.7) says that, if $\bar{x} \in E_+$, then, for any $\bar{z} \in \mathbb{R}^n_{>0}$,

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

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 $\bar{z} \in \mathbb{R}^n_{>0}$,

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

Thus, Equivalence (3.7) 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 Proposition 2.2.1), and we have $B = (B_1 \ \cdots \ B_L)$.

So, a general system

$$\dot{x} = f_1(x) + \cdots + f_L(x), \quad y = h(x)$$

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 of the form

$$\dot{z} = f_1(z) + \cdots + f_L(z) + C'(h(x) - h(z)). \tag{4.9}$$

To prove convergence of the observer, one may use Proposition 4.3.1, Proposition 4.2.3 and Corollary 4.3.3 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 (4.6) and the corresponding conclusions hold.

To see that Proposition 4.2.3 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 (2.13) 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 $z \in E^s_+ \cup E^s_0$. Hence

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

where the right hand side vanishes only if $z \in (E^1_+ \cap \cdots \cap E^L_+) \cup (E^1_0 \cap \cdots \cap E^L_0)$, i.e., only if $z \in E_+ \cup E_0$. The rest of the proof of Proposition 4.2.3 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 4.3.3 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.

Chapter 5
Robustness with Respect to Parameters

Throughout the previous chapters, we have assumed that the values of the reaction rate constants, a_{ij}, are known for the design of the observers. However, from the biological or the chemical engineering point of view, the values of the constants a_{ij} are determined within a certain error margin.

We would like to investigate the performance of the observer (4.2) under small perturbations of the matrix $A = (a_{ij})$, that is, suppose that the observer is constructed, not with the "real" A, but instead with some "ideal" value A_0. Will the observer still produce a reasonable estimate of the state $x(t)$? Our goal is to establish that, provided the difference $\|A - A_0\|$ is small, the difference $|z(t) - x(t)|$ will also be small, and to provide error estimates.

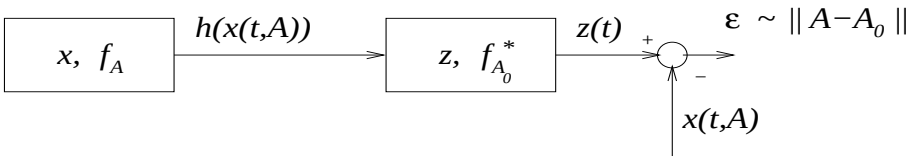


Figure 5.1: Robustness of the state-estimator with respect to parameters.

In Chapter 8 a simple experiment is reported, consisting of a reversible reaction where measurements of all the concentrations are obtained with an NMR spectrometer. This experiment illustrates the usefulness of a state-estimator, as well as the importance of requiring the observers to be robust to small perturbations in the rate constants.

In this Chapter, a definition of a parameter–robust observer is proposed, and it is proved that our observers are robust in this sense. In order to prove this robustness property, we first discuss a result for irreducible matrices, on the dependence of a Perron eigenvector on the entries of the matrix.

Another result established in this chapter is the *global* ISS property for system (4.5) when the outputs include the measurement of a reaction rate (as in Remark 3.2.6). In this case, moreover, a separation principle may be deduced, which is sketched in Appendix A.

5.1 Parameter–Robust Observers

To formalize our concept of robustness, some notation is needed. Let

$$\mathcal{A}_{\geq 0} = \{A \in \mathbb{R}^{m\times m} : A \geq 0 \text{ and } (A+I)^k > 0 \text{ for some power } k\}.$$

Thus $\mathcal{A}_{\geq 0}$ is the set of irreducible $m \times m$ matrices whose entries are nonnegative. The inequality $A \geq 0$ (resp. $A > 0$), means that every entry of the matrix on the left hand side is nonegative (resp. positive). Let $\|A\|_{\text{ecl}}$ denote the matrix norm induced by the vector norm $|\cdot|$ (euclidean norm).

In this Chapter, we will assume that the matrix B (which defines the complexes which form the network) is fixed. Each matrix $A \in \mathcal{A}_{\geq 0}$ characterizes a system of the form (3.1), so let $f_A : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n$ denote the function defined in (3.2) computed with the entries a_{ij} of A:

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

Let

$$\Sigma(A): \qquad \dot{x} = f_A(x), \quad y = h(x),$$

and let $x(t, x_0, A)$ denote the solution of the differential equation $\dot{x} = f_A(x)$, at time t, when the initial condition is $x(0) = x_0 \in \mathbb{R}^n_{>0}$.

Recall also, from Proposition 2.2.1, that the boundary equilibria (E_0) depend only on the matrix B, whereas the positive equilibria (E_+) depend both on B and A. Throughout this Chapter, we assume that no boundary equilibria exist in any positive class: $\mathcal{S} \cap E_0 = \emptyset$, for each positive class $\mathcal{S}$. Then, from Theorem 1, we know that $x(t, x_0, A)$ converges to the positive equilibrium in the class of x_0. So we define $\bar{x}(x_0, A)$ to be unique equilibrium in the same class of x_0. Define

$$E_{A,+} = \{\bar{x}(x_0, A) : x_0 \in \mathbb{R}^n_{>0}\}.$$

and also

$$\mathcal{E} = \bigcup_{A \in \mathcal{A}_{\geq 0}} E_{A,+}$$

to be the set of all such positive equilibrium points.

We will use the following notation, for any function $\varphi : [0, +\infty) \to \mathbb{R}^n$:

$$\|\varphi\|_T := \text{ess.sup. } \{|\varphi(t)| : t \geq T\}$$

(if $T = 0$, we will simply drop the subscript "T").

The concept of "perturbation to the ideal A_0" is now discussed. Given a chemical network, characterized by B and A_0, each nonzero entry of A_0 stands for an existing reaction between two of the complexes. Thus it would be reasonable to ask that a perturbation of the matrix A_0 would mean a perturbation of the nonzero entries of A_0 only, a situation where no alterations to the *structure* of the existing network would be considered. But, surprisingly, the proof of the main robustness result, allows a more general notion: *a matrix A is a perturbation of A_0* if

1. A is irreducible, and
2. the nonzero entries entries of A, as well as the nonzero entries of A_0, belong to a compact interval $[\alpha_0, \alpha^0]$, with $0 < \alpha_0 \leq \alpha^0$, and
3. A is constant along time, that is, we assume the true system may look like $\dot{x} = f_A(x)$, where A is constant, while the design of an observer is based on a *nominal system* $\dot{x} = f_{A_0}(x)$.

So, in some ways, the structure of the network may be modified, i.e., new reactions may be added and existing reactions may be removed, as long as *the irreducibility of the matrix A is not violated*. For example, consider the McKeithan network:

$$A_0 = \begin{pmatrix} 0 & k_{12} & k_{13} \\ k_{21} & 0 & 0 \\ 0 & k_{32} & 0 \end{pmatrix}.$$

Possible perturbations include:

$$A_1 = \begin{pmatrix} 0 & k_{12} - \varepsilon & k_{13} \\ k_{21} - \varepsilon & 0 & 0 \\ 0 & k_{32} + \varepsilon & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & k_{12} & k_{13} \\ k_{21} & 0 & \beta_{23} \\ 0 & k_{32} & 0 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0 & 0 & k_{13} \\ k_{21} & 0 & 0 \\ 0 & k_{32} - \varepsilon & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & k_{12} & 0 \\ 0 & 0 & \beta_{23} \\ \beta_{31} & 0 & 0 \end{pmatrix},$$

where β_{23}, β_{31}, $k_{12} - \varepsilon$, and all nonzero entries, belong to an interval $[\alpha_0, \alpha^0]$, with $0 < \alpha_0 \leq \alpha^0$. For instance, A_2 belongs to the set of perturbations of A_0 (as defined by points 1 – 3), so long as $\beta_{23} = 0$ or $\beta_{23} \in [\alpha_0, \alpha^0]$.

The bounds α_0, α^0 are fixed *a priori*: the estimates of the error $|z(t) - x(t)|$ will depend on these bounds. In addition, the error will be upper bounded mainly by a $\mathcal{K}_\infty$ function of $\|A - A_0\|_{\text{ecl}}$, so that "small $\|A - A_0\|_{\text{ecl}}$" will indeed imply "small error $|z(t) - x(t)|$"; similarly, if the norm $\|A - A_0\|_{\text{ecl}}$ is large one may expect the error to be also large. This is consistent with the intuitive idea that A_4 is a "bad" perturbation of A_0, with $\|A_4 - A_0\|_{\text{ecl}}$ large.

More precisely, we will be interested in subsets of $\mathcal{A}_{\geq 0}$ and subsets of $\mathcal{E}$ of the form stated in the next Definition.

Definition 5.1.1 Let A_0 be any matrix in $\mathcal{A}_{\geq 0}$. A pair of subsets K of $\mathcal{A}_{\geq 0}$ and P of $\mathcal{E}$ is said to satisfy the *property* $\mathcal{P}_0$ (with respect to A_0) if there exist constants $0 < \alpha_0 \leq \alpha_0$ and a point $x_0 \in \mathbb{R}^n_{>0}$ such that the following hold:

(a) $A_0 \in K$,

(b) if $A = (a_{ij}) \in K$, then

$$\text{either } a_{ij} = 0 \quad \text{or} \quad \alpha_0 \leq a_{ij} \leq \alpha^0, \quad \forall\, i, j = 1, \ldots, m, \tag{5.1}$$

(c) P is compact,

(d) $\bar{x}(x_0, A_0) =: \bar{x}_0 \in P$, and

(e) for every $\bar{x} \in P$, $|h_i(\bar{x}_0) - h_i(\bar{x})| \leq \frac{1}{3} h_i(\bar{x}_0)$ for every $i = 1, \ldots, p$.

Observe that, since $\bar{x}_0 \in \mathcal{E} \subset \mathbb{R}^n_{>0}$, then $h_i(\bar{x}_0) \neq 0$ for every i.
It is also useful to define the set

$$Q(P, K) = \{s \in \mathbb{R}^n_{>0} : \bar{x}(s, A) \in P \text{ for some } A \in K\}.$$

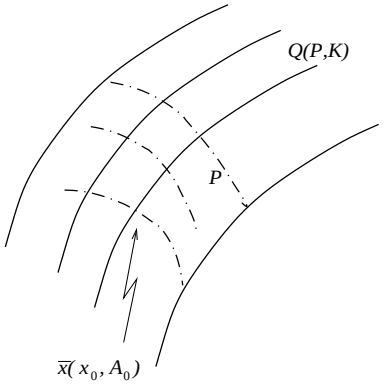


Figure 5.2: The set P is represented by the dash-dotted lines, while the solid lines represent the parallel classes of the system whose unique positive equilibrium is contained in P (i.e., the solid lines represent the set $Q(P, K)$). A reference element $\bar{x}_0$ is shown.

Definition 5.1.2 The system $\dot{z} = g(z, h(x))$ is a *parameter A_0-robust observer* for the system $\Sigma(A_0)$ if, for any pair of subsets $K \subset \mathcal{A}_{\geq 0}$ and $P \subset \mathcal{E}$ that satisfy property $\mathcal{P}_0$, the following hold:

(i) there exist functions $\beta = \beta_{A_0} \in \mathcal{KL}$ and $\varphi = \varphi_P \in \mathcal{K}_\infty$,

(ii) there exist functions $\tilde{\beta} = \tilde{\beta}_{P,K} \in \mathcal{KL}$, and $\zeta = \zeta_{P,K} \in \mathcal{K}$,

(iii) for each compact set $Q_0 \subset Q(P, K)$, there exists $T = T_{K,Q_0,\bar{x}_0}$

such that, for each matrix $A \in K$ and each point $s \in Q_0$ with $\bar{x}(s, A) \in P$, the solution of the extended system

$$\begin{aligned} \dot{x} &= f_A(x), \quad x(0) = s \\ \dot{z} &= g(z, h(x)), \quad z(0) = z_0 \end{aligned}$$

satisfies

$$\begin{aligned} |z(t) - x(t, s, A)| \leq\ & \beta(|z(T) - \bar{x}_0|, t) + \tilde{\beta}(|s - \bar{x}(s, A)|, t) \\ & + \varphi(\|x(\cdot, s, A) - \bar{x}(s, A)\|_T) \\ & + \zeta(\|A_0 - A\|_{\text{ecl}}) + \zeta(|x_0 - s|), \end{aligned}$$

for all $t \geq T$ and all $z_0 \in \mathbb{R}^n_{\geq 0}$.

Remark 5.1.3 The set of initial conditions, $Q(P,K)$, contains whole positive classes (in fact, it contains sets of the form $\mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$ for $\bar{x} \in P$), and thus can be *unbounded*.

Remark 5.1.4 As a particular case, note that if $P \equiv \{\bar{x}_0\}$, and $K \equiv \{A_0\}$ the definition of a parameter A_0-robust observer reduces to that of a full-state observer. Indeed, in this case, the set $Q(P,K)$ coincides with $\mathcal{S}_{\bar{x}_0} \cap \mathbb{R}^n_{>0}$, and we may assume that $x_0 \equiv s$, because in fact $\bar{x}(x_0, A_0) \equiv \bar{x}(s, A_0)$ for all $s \in Q(P,K)$.

Given a system of the form (3.1), Theorem 1 says that, for every $s \in \mathbb{R}^n_{>0}$, $x(t,s,A)$ converges to $\bar{x}(s,A)$, so one may always choose T large enough so that the term $\varphi(\|x(\cdot,s,A) - \bar{x}(s,A)\|_T)$ becomes very small. The terms in β and $\tilde{\beta}$ also become very small, since these are $\mathcal{KL}$ functions. Eventually, the observer's estimates will be dominated only by the differences $\|A_0 - A\|_{\text{ecl}}$ and $|x_0 - s|$. The term $\zeta(|x_0 - s|)$ arises from a technical point: the need to specify a reference equilibrium point, $\bar{x}_0 \in P$, in Definition 5.1.1. This will be used to provide an ISS estimate with respect to this reference point. In general, we may view s as a perturbation of x_0, and (as $\bar{x}(\cdot,\cdot)$ will be shown to be analytic) we may also view $\bar{x}(s,A)$ as a perturbation of $\bar{x}(x_0, A_0)$.

From now on, in this Chapter, assume that the output maps are such that one of the rows of the matrix C coincides with one of the columns (transposed) of B, for instance (equation (3.9)),

$$h_l(x) = x_1^{b_{1j}} x_2^{b_{1j}} \dots x_n^{b_{nj}} \tag{5.2}$$

for some $1 \leq l \leq p$ and some $1 \leq j \leq m$. This is simply saying that one of the measurements is one of the reaction rates of the network, which seems to be a reasonable choice. For instance, one might be able to monitor the energy released or consumed in the reaction, or tag it with an appropriate reporter. As described in Remark 3.2.6, this condition together with $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$ are sufficient to ensure detectability of the system (3.1) with outputs (3.3). We will prove the following result.

Theorem 4 Let the matrix B be fixed. Let $A_0 \in \mathcal{A}_{\geq 0}$ and let C be such that the system $\Sigma(A_0)$: $\dot{x} = f_{A_0}(x), y = h(x)$ is detectable and $h(x)$ satisfies (5.2). Then the system $\dot{z} = f_{A_0}(z) + C'(h(x) - h(z))$ is an A_0-robust observer for the system $\Sigma(A_0)$.

The difference between the real and ideal values of the equilibrium points will be a major factor in deciding whether an observer is robust, and so whether a reasonable estimate for $x(t,s,A)$ is to be expected. In Section 5.4 we will show how the equilibria and trajectories of the original system behave as the parameter A varies.

Input-to-state stability estimates for the system, $\dot{z} = f_{A_0}(z) + C'(u - h(z))$, in the case when h is of the form (5.2), also play a very important role in establishing Theorem 4, and are stated in Section 5.2.

Remark 5.1.5 Since we are, from now on, restricting the output maps to take the form (5.2), Theorem 4 is a generalization of Theorem 3 *only* for such outputs maps. Note that Theorem 3 applies to the general form of outputs (3.3).

5.2 A (Global) ISS Estimate

In the case of outputs of the form (5.2), we can show a stronger ISS result for the system with inputs

$$\dot{z} = f_A^*(z,u) := f_A(z) + C'(u - h(z)). \tag{5.3}$$

In fact, this system with map h of the form (5.2) is *(globally)* ISS with input set $\mathbb{U}_\theta$ (with respect to $\bar{x}$ and $h(\bar{x})$). For this form of output maps, (5.3) admits V as an ISS-Lyapunov function with respect to the point $\bar{x}$ and the input $h(\bar{x})$ and a stronger version of Proposition 4.2.3 can be proved as follows.

Proposition 5.2.1 Fix any $A \in \mathcal{A}_{\geq 0}$ and any $\bar{x} \in E_{A,+}$. Assume that the map h is such that (5.2) holds and C satisfies $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. Let $0 < \theta < 1$ be arbitrary and define $\mathbb{U}_\theta \subset \mathbb{R}^p$ (as in Proposition 4.2.3) by:

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

Then there exist functions $\alpha, \gamma \in \mathcal{K}_\infty$ such that

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

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

The proof of this Proposition occupies the rest of this Section. Consider

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

where the first part can be written (recall expression (2.9))

$$\nabla V(z,\bar{x})\ f_A(z) = -\sum_{i=1}^m \sum_{j=1}^m a_{ij} e^{\langle b_j, \rho(\bar{x})\rangle} e^{q_j}\, \omega(q_i - q_j)$$

with the notation $q_j = q_j(z,\bar{x}) := \langle b_j, \rho(z) - \rho(\bar{x})\rangle$, and $\omega(x) = e^x - 1 - x$. Since $\omega(x)$ is always positive for $x \neq 0$ and $\omega(0) = 0$, it follows that every term in this double sum is nonpositive.

Write $u - h(z) = u - h(\bar{x}) - (h(z) - h(\bar{x}))$, use the fact that C is the adjoint operator of C' and recall the definition of h and of the vectorial function ρ to get the equality $C\rho(z) = \rho(h(z))$, and finally write:

$$\nabla V(z,\bar{x})\, f_A^*(z) = -\vartheta(z,\bar{x}) + \Upsilon(z,u,\bar{x})$$

where

$$\vartheta(z,\bar{x}) = \sum_{i=1}^m \sum_{j=1}^m a_{ij}\, e^{\langle b_j, \rho(\bar{x})\rangle}\, e^{q_j}\, \omega(q_i - q_j) + (1-\theta)\langle \rho(h(z)) - \rho(h(\bar{x})), h(z) - h(\bar{x})\rangle$$

and

$$\Upsilon(z,u,\bar{x}) = -\theta\langle \rho(h(z)) - \rho(h(\bar{x})), h(z) - h(\bar{x})\rangle + \langle \rho(h(z)) - \rho(h(\bar{x})), u - h(\bar{x})\rangle.$$

We will next show that these functions ϑ, Υ admit $\mathcal{K}_\infty$ bounds that will establish Proposition 5.2.1.

Lemma 5.2.2 There exists a function $\gamma \in \mathcal{K}_\infty$ such that

$$\Upsilon(z, u, \bar{x}) \leq \gamma(|u - h(\bar{x})|)$$

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

The proof of this Lemma is exactly as in Step 1 of Proposition 4.2.3. To produce a lower bound for ϑ, some auxiliary calculations are needed.

Lemma 5.2.3 Assuming that the map h satisfies both condition (5.2) and $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$, the following holds:

$$(\forall\, N > 0)\,(\exists\, R > 0)\ |\rho(z) - \rho(\bar{x})| \geq R \ \Rightarrow\ \vartheta(z, \bar{x}) > N.$$

Proof. We assume that the contrapositive to this statement holds:

$$(\exists\, N_* > 0)\,(\forall\, R > 0)\,(\exists\, z_R \in \mathbb{R}^n_{>0})\ |\rho(z_R) - \rho(\bar{x})| \geq R \ \text{ and } \ \vartheta(z_R, \bar{x}) \leq N_*,$$

and will derive a contradiction. For any R, $\vartheta(z_R, \bar{x}) \leq N_*$ implies

$$\begin{cases} a_{ij}\, e^{\langle b_j, \rho(\bar{x})\rangle}\, e^{q_j}\, \omega(q_i - q_j) \leq N_*, \quad \forall\, i, j = 1, \ldots, m \\ \\ (1 - \theta)\langle \rho(h(z_R)) - \rho(h(\bar{x})), h(z_R) - h(\bar{x})\rangle \leq N_*\,. \end{cases} \tag{5.4}$$

(Since each of these terms is positive, and ϑ is the sum of them all.) On the other hand we have, from Lemma D.0.4,

$$\langle \rho(h(z)) - \rho(h(\bar{x})), h(z) - h(\bar{x})\rangle \geq \mu(|C(\rho(z) - \rho(\bar{x}))|) \tag{5.5}$$

with $\mu \in \mathcal{K}_\infty$. So, by (5.2), (5.4) and (5.5), there is some $j \in \{1, \ldots, m\}$ (which we may assume to be $j = 1$, without loss of generality) such that

$$|q_1| = |\langle b_1, \rho(z_R) - \rho(\bar{x})\rangle| \leq |C(\rho(z_R) - \rho(\bar{x}))| \leq \mu^{-1}\left(\frac{N_*}{1 - \theta}\right)$$

for any R.

Claim. For each $j \in \{1, \ldots, m\}$, there exist constants N_j such that

$$|q_j| = |\langle b_j, \rho(z_R) - \rho(\bar{x})\rangle| \leq N_j,$$

for all R.

This claim follows immediately from Lemma 2.6.2, with $\mathcal{X}_o = \{z_R : R > 0\}$ and $N_0 = \mu^{-1}(N_*/(1 - \theta))$.

Now, define $D \in \mathbb{R}^{(m-1)\times n}$ to be the matrix with transpose given by

$$D' = (b_2 - b_1 \quad \cdots \quad b_m - b_1),$$

and recall that $\{b_j - b_1 : i = 2, \ldots, m\}$ is a basis of the stoichiometric space $\mathcal{D}$. Also, C belongs to $\mathbb{R}^{p\times n}$, with $p = n - (m-1)$, and satisfies $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. It follows that

$$D_0 = \begin{pmatrix} D \\ C \end{pmatrix}$$

is an invertible matrix and

$$|\rho(z) - \rho(\bar{x})|^2 \leq k|D(\rho(z) - \rho(\bar{x}))|^2 + k|C(\rho(z) - \rho(\bar{x}))|^2$$

where $k = \|D_0^{-1}\|^2$.

From (5.4) and (5.5), for any R we have

$$|C(\rho(z_R) - \rho(\bar{x}))| \leq \mu^{-1}(N_*/(1-\theta))$$

and by the claim it is also true that $|q_j - q_1| \leq N_j + N_1$, so

$$|D(\rho(z_R) - \rho(\bar{x}))|^2 = \sum_{j=2}^{m} |q_j - q_1|^2 \leq 4 \max N_j^2.$$

Therefore,

$$R^2 \leq |\rho(z_R) - \rho(\bar{x})|^2 \leq 4k \max N_j^2 + k \left[\mu^{-1}\left(\frac{N_*}{1-\theta}\right)\right]^2$$

which contradicts the fact that R is arbitrarily large. So the Lemma is true. ∎

Lemma 5.2.4 Assume that the map h satisfies both condition (5.2) and $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. There exists a function $\alpha \in \mathcal{K}_\infty$ such that

$$\vartheta(z, \bar{x}) \geq \alpha(|\rho(z) - \rho(\bar{x})|)$$

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

Proof. Consider the function

$$\alpha(r) = \inf\{\vartheta(z, \bar{x}) : |\rho(z) - \rho(\bar{x})| \geq r\}.$$

Note that α is well defined since, given any $r > 0$, pick z_* with $|\rho(z_*) - \rho(\bar{x})| = r$ and put $N = \vartheta(z^*, \bar{x})$ ($N > 0$ because from Lemma 2.5.4 we know that $\vartheta(z_*, \bar{x})$ is always positive for $z \in \mathbb{R}^n_{>0}$, $z \neq \bar{x}$). Then $\alpha(r) \leq N$ and, moreover, from Lemma 5.2.3, get $R \geq r$ (R depends on r, since N does) so that $\vartheta(z, \bar{x}) > N$ for all $|\rho(z) - \rho(\bar{x})| \geq R$. Thus, one can also write

$$\alpha(r) = \inf\{\vartheta(z, \bar{x}) : r \leq |\rho(z) - \rho(\bar{x})| \leq R\},$$

so the infimum is really the minimum of ϑ over a compact subset of $\mathbb{R}^n_{>0}$. Moreover, since ϑ is continuous and strictly positive for all $z \neq \bar{x}$, it follows that $\min \vartheta(z, \bar{x})$ over a compact set which does not include $\bar{x}$ is also strictly positive. Hence $\alpha(r) > 0$ whenever $r > 0$.

To show that $\alpha \in \mathcal{K}_\infty$, we need to check that

(a) α is continuous;

(b) $\alpha(0) = 0$;

(c) $\alpha(r) > \alpha(s)$ for all $r > s$;

(d) $\alpha(r) \to +\infty$ as $r \to +\infty$.

Property (b) holds because the set $|\rho(z) - \rho(\bar{x})| \geq 0$ includes $z = \bar{x}$. To see that (d) holds, apply Lemma 5.2.3 to conclude that

$$(\forall\, N > 0)\,(\exists\, R > 0)\ |\rho(z) - \rho(\bar{x})| \geq R \ \Rightarrow\ \inf\{\vartheta(z, \bar{x})\} \equiv \alpha(R) \geq N.$$

Finally, the function is certainly increasing, because the set $\{|\rho(z) - \rho(\bar{x})| \geq r\}$ is strictly contained in $\{|\rho(z) - \rho(\bar{x})| \geq s\}$ for all $r > s$. So, using (d), without loss of generality one may assume that (c) holds (otherwise, one may take, for instance, $\hat{\alpha}(r) = \frac{r}{r+1}\alpha(r)$, which is strictly increasing and satisfies $\hat{\alpha}(r) \leq \alpha(r)$).

Finally, we may also assume that α is continuous since otherwise, properties (c) and (d) allow us to construct a continuous function $\hat{\alpha}$, such that $\hat{\alpha}(r) \leq \alpha(r)$ and still satisfies (b)-(d). ∎

To summarize, Lemma 5.2.2 together with the combination of Lemmas 5.2.4 and D.0.6 provide the proof of Proposition 5.2.1.

The existence of an ISS-Lyapunov function with respect to the point $\bar{x}$ and the input $h(\bar{x})$ for system (5.3) implies, by Lemma 3.3.6, that (5.3) is ISS with input set $\mathbb{U}_\theta$. In other words, system (5.3) is $\mathbb{R}^n_{>0}$-complete and there exist functions $\beta \in \mathcal{KL}$ and $\varphi \in \mathcal{K}_\infty$ such that, for each $z_0 \in \mathbb{R}^n_{>0}$,

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

for all $t \in [0, +\infty)$ and all $u \in \mathbb{U}_\theta$.

We may now extend this estimate to all $z_0 \in \mathbb{R}^n_{\geq 0}$, by taking a sequence of points $\xi_k \in \mathbb{R}^n_{>0}$, with $\xi_k \to z_0$ and letting $z^k(t)$ be the solution, at time t, of the differential equation

$$\dot{v} = f^*_A(v, u), \quad v(0) = \xi_k.$$

Then, for each k the following estimate holds

$$|z^k(t) - \bar{x}| \leq \beta(|\xi_k - \bar{x}|, t) + \varphi(\|u - h(\bar{x})\|)$$

for all $t \in [0, +\infty)$ and all $u \in \mathbb{U}_\theta$. By the continuity of the solution of a differential equation on the initial conditions, and by continuity of the function β, the corresponding estimate holds for $z(t)$.

We conclude with the following strengthening of Proposition 4.3.1:

Proposition 5.2.5 Assume that the map h is such that (5.2) holds and the matrix C is such that $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. Then the system with inputs $\dot{z} = f^*_A(z, u)$ is ISS with input

set $\mathbb{U}_\theta$, with respect to the point $\bar{x}$ and the input $h(\bar{x})$, i.e., the system is $\mathbb{R}^n_{>0}$-complete and there exist functions $\beta \in \mathcal{KL}$ and $\varphi \in \mathcal{K}_\infty$ such that, for each $z_0 \in \mathbb{R}^n_{\geq 0}$,

$$|z(t) - \bar{x}| \leq \beta(|z_0 - \bar{x}|, t) + \varphi(\|u - h(\bar{x})\|) \tag{5.6}$$

for all $t \in [0, +\infty)$ and all $u \in \mathbb{U}_\theta$.

5.3 Irreducible Matrices

In order to analyze the dependence of the equilibrium points $\bar{x} = \bar{x}(s, A)$ as maps $\mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E}$, on s and A, in this section we will first discuss a result about the Perron eigenvectors of irreducible matrices.

Given any matrix $G \in \mathbb{R}^{m \times m}$, with entries g_{ij}, define another matrix by

$$M_G = \left(\frac{1}{1 + \sum g_{ii}^2} G + I \right)^{m-1}.$$

Also define

$$\phi(G) = \left(1 + \sum_{i=1}^{m} g_{ii}^2 \right)^{-1},$$

so that $M_G = (\phi(G)G + I)^{m-1}$. By construction, the diagonal entries of $\phi(G)G + I$ are positive.

Introduce the following subset of $\mathbb{R}^{m \times m}$

$$\mathcal{G} = \{G \in \mathbb{R}^{m \times m} : M_G > 0 \text{ and } \bar{1}\, G = 0\},$$

where the inequality means that every entry of the matrix on the left hand side is strictly positive, and $\bar{1}$ is the row vector $(1\ 1 \ldots 1)$.

Let $\mathcal{G}_{\geq 0}$ be the set of all the irreducible matrices which have $\bar{1}\, G = 0$, *nonnegative off-diagonal* entries and *arbitrary diagonal* entries. Note that:

$$\mathcal{G}_{\geq 0} = \{G \in \mathcal{G} : G \text{ has nonnegative off-diagonal entries }\}.$$

The set $\mathcal{G}$ may be seen as an open subset of the $m^2 - m$ dimensional linear subspace $\{G : \bar{1}\, G = 0\}$ of $\mathbb{R}^{m \times m}$.

For each $G \in \mathcal{G}$ observe that

$$\bar{1} M_G = \bar{1}\, (\phi(G)\, G + I)^{m-1} = \bar{1}$$

because $\bar{1}\, G = 0$ and $\bar{1}(\phi(G)\, G + I) = \bar{1}$. So, any nonnegative eigenvector, $v \in \mathbb{R}^n_{\geq 0}$, of the matrix M_G must correspond to the eigenvalue $\mu = 1$ since

$$M_G v = \mu v \ \Rightarrow\ \bar{1}(M_G v) = \bar{1}(\mu v) \ \Leftrightarrow\ \bar{1} v = \mu \bar{1} v,$$

and $\bar{1}v$ is a positive scalar (since $v \neq (0, \ldots, 0)'$, by definition of eigenvector).

Since, by definition, M_G is irreducible and has all entries positive, by the Perron-Frobenius Theorem we know that the spectral radius of M_G, $\sigma(M_G)$, is an eigenvalue of M_G, of algebraic (and hence geometric) multiplicity one. Moreover, an eigenvector associated to $\sigma(M_G)$ can be chosen to have all entries strictly positive. Any such eigenvector is usually called a Perron eigenvector of M_G (and any two of these are positive multiples of each other). But, as we have just seen, any positive eigenvector of M_G corresponds to the eigenvalue $\mu = 1$, so we have

$$\sigma(M_G) \equiv 1, \quad \forall\ G \in \mathcal{G}.$$

Define $v_{\mathrm{P}} : \mathcal{G} \to \mathbb{R}^m_{>0}$ to be the map that assigns to each $G \in \mathcal{G}$, the unique Perron eigenvector of M_G, which has its first coordinate equal to 1:

$$v_{\mathrm{P}} = \begin{pmatrix} 1 \\ w_{\mathrm{P}} \end{pmatrix}$$

for some $w_{\mathrm{P}} \in \mathbb{R}^{m-1}_{>0}$.

By a *rational function everywhere defined on* $\mathcal{G}$ we mean a function $\psi : \mathcal{G} \to \mathbb{R}^m$ for which every coordinate is a quotient $\psi_i = p_{\mathrm{num}} p_{\mathrm{den}}^{-1}$ of two polynomial functions (on the entries of G) $p_{\mathrm{num}}, p_{\mathrm{den}} : \mathbb{R}^{m\times m} \to \mathbb{R}$ such that $p_{\mathrm{den}}(G) \neq 0$ for all $G \in \mathcal{G}$.

Proposition 5.3.1 The map v_{P} is a rational function on $\mathcal{G}$.

Proof. For each $G \in \mathcal{G}$, by abuse of notation, write v_{P} for $v_{\mathrm{P}}(G)$. We will also drop the subscript and let $M = M_G$, for simplicity. We have:

$$Mv_{\mathrm{P}} = \sigma(M)v_{\mathrm{P}} \ \Leftrightarrow\ (M - I)v_{\mathrm{P}} = 0.$$

The matrix $M - I$ has rank $m-1$ because $\sigma(M) = 1$ is a simple root of the characteristic polynomial of M. Put

$$M - I = (N_1\ N)$$

where N_1 is the first column of $M - I$ and N is the remaining $m \times (m-1)$ matrix, and notice that

$$(N_1\ N)\begin{pmatrix} 1 \\ w_{\mathrm{P}} \end{pmatrix} = 0 \ \Leftrightarrow\ Nw_{\mathrm{P}} = -N_1.$$

Claim. The matrix N has full rank.

Suppose the claim is false. Then there exists an element u in the kernel of N, and one can write $N(w_{\mathrm{P}} + u) = -N_1$. But if this is true, then it also holds that

$$(M - I)\begin{pmatrix} 1 \\ w_{\mathrm{P}} + u \end{pmatrix} = 0$$

which implies $w_{\mathrm{P}} + u = w_{\mathrm{P}}$, because v_{P} is in fact the unique vector with first coordinate equal to 1 in the kernel of $M - I$. So $u \equiv 0$, which proves the claim.

It follows that $\det(N'N) \neq 0$ for every G, and applying the Moore-Penrose pseudoinverse of N yields

$$v_{\mathrm{P}} = \begin{pmatrix} 1 \\ w_{\mathrm{P}} \end{pmatrix} = \begin{pmatrix} 1 \\ -(N'N)^{-1}N'N_1 \end{pmatrix},$$

where N and N_1 are defined from $M = M_G$, as above. This shows that v_{P} is a rational function on $\mathcal{G}$.

■

For every $G \in \mathcal{G}$, the Perron eigenvector of M_G, v_{P}, is also an eigenvector of the matrix G, corresponding to the 0 eigenvalue with multiplicity one. This fact follows from two observations:

1. $\ker(G) \neq \emptyset$, so $\exists v \in \mathbb{R}^m \setminus \{0\}$ such that $Gv = 0$.

   This is because $\bar{1}G = 0$, which means that the rows of G are linearly dependent, and thus rank $G \leq m - 1$.

2. Any v such that $Gv = 0$, satisfies $v \in$ span $\{v_{\mathrm{P}}\}$.

   This follows from

   $$(\phi(G)\, G + I)v = v \Rightarrow (\phi(G)\, G + I)^{m-1}v = M_G\, v = v$$

   and hence $v \in$ span $\{v_{\mathrm{P}}\}$, since $\sigma(M_G) = 1$ is an eigenvalue of M_G of multiplicity one.

Therefore, the kernel of G has dimension 1 and is given by:

$$\ker(G) = \text{ span } \{v_{\mathrm{P}}(G)\}. \tag{5.7}$$

5.4 Dependence of the System on the Parameters

In this section we will study the regularity properties of the map

$$\bar{x} : \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E}$$

given by

$$(x_0, A) \mapsto \bar{x}(x_0, A),$$

where "$\bar{x}(x_0, A)$" is as defined in Section 5.1, the unique positive equilibrium of the system $\dot{x} = f_A(x)$ in the same class as x_0.

For each fixed A, the smoothness of $\bar{x}$ with respect to x_0 has already been established by Sontag in [48], a result we also stated as Lemma 2.4.2: it follows by picking any equilibrium $\bar{z} \in E_{A,+}$ and observing that $\bar{x}(x_0, A) = \varphi(x_0, \bar{z})$ for all $x_0 \in \mathbb{R}^n_{>0}$, where $\varphi \in C^k$ $(k = 0, 1, \ldots, \infty, \omega)$, when the restriction to $\mathbb{R}^n_{\geq 0}$ of the maps $\theta_i : \mathbb{R}^n \to \mathbb{R}^n_{>0}$ is C^k and satisfies $\theta_i'(r) > 0$.

We will establish that, for the case of the maps $\theta_i(r) = |r|$, $\bar{x}(\cdot, \cdot)$ is real analytic (C^ω) on $\mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0}$.

5.4.1 Continuity of Equilibrium Points

For each fixed $A \in \mathcal{A}_{\geq 0}$, recall the alternative expression (2.4) for f_A:

$$f_A(x) = B\tilde{A}\theta_B(x)$$

where

$$\theta_B(x) = \begin{pmatrix} x_1^{b_{11}} x_2^{b_{21}} \dots x_n^{b_{n1}} \\ x_1^{b_{12}} x_2^{b_{22}} \dots x_n^{b_{n2}} \\ \vdots \\ x_1^{b_{1m}} x_2^{b_{2m}} \dots x_n^{b_{nm}} \end{pmatrix} = \mathrm{Exp}[B'\rho(x)],$$

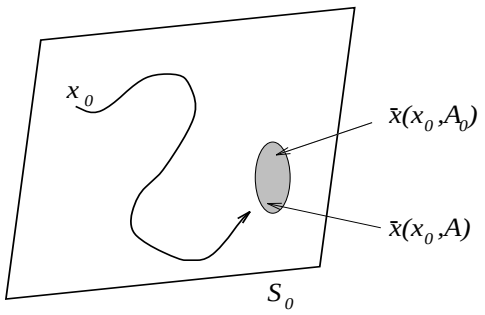
and

$$\tilde{A} = A + \begin{pmatrix} -\sum_{i=1}^m a_{i1} & 0 & \cdots & 0 \\ 0 & -\sum_{i=1}^m a_{i2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -\sum_{i=1}^m a_{im} \end{pmatrix}.$$

(Recall that we assumed that all the diagonal entries of A are zero, since their value does not enter in the computations of the vector field f_A.)

In this way, to each matrix $A \in \mathcal{A}_{\geq 0}$, we associate a matrix $\tilde{A} \in \mathcal{G}_{\geq 0}$: clearly, $\bar{1}\tilde{A} = 0$ and so $\tilde{A} \in \mathcal{G}_{\geq 0}$.

Theorem 5 The map $\bar{x} : \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E} \subset \mathbb{R}^n_{>0}$ given by $(x_0, A) \mapsto \bar{x}(x_0, A)$ is real-analytic.



Proof. A function f, defined on an open set $\mathcal{V}$, is real analytic if it admits a power series expansion on a neighborhood of each point of $\mathcal{V}$. If, as in our case, the set $\mathcal{V}$ is not open, then the function f is still called real analytic if admits an extension to a real analytic function on a neighborhood of $\mathcal{V}$ (see [46]). This is what we will show for the map $\bar{x}$.

For each A consider the matrix $\tilde{A} \in \mathcal{G}_{\geq 0}$, constructed from A as indicated above. Then, from (5.7), $\ker \tilde{A} =$ span $\{v_{\mathrm{P}}(\tilde{A})\}$.

It now follows from Lemma 2.3.1 that each equilibrium $\bar{x} \in E_{A,+}$ is characterized by

$$\theta_B(\bar{x}) = c\, v_{\mathrm{P}}(\tilde{A}) \quad \Leftrightarrow \quad B'\rho(\bar{x}) = \rho(c\, v_{\mathrm{P}}(\tilde{A})),$$

where c is a positive constant.

Claim. For each A, the element $\bar{z}(A) \in \mathbb{R}^n_{>0}$ given by

$$\bar{z}(A) = \mathrm{Exp}[B(B'B)^{-1}\rho(v_{\mathrm{P}}(\tilde{A}))],$$

is an equilibrium point in $E_{A,+}$.

To prove the claim, note that B' has full column rank, so $B'B$ is an invertible matrix and the formula gives

$$\rho(\bar{z}(A)) = B(B'B)^{-1}\rho(v_{\mathrm{P}}(\tilde{A}))$$

or, equivalently,

$$B'\rho(\bar{z}(A)) = B'B(B'B)^{-1}\rho(v_{\mathrm{P}}(\tilde{A})) = \rho(v_{\mathrm{P}}(\tilde{A})).$$

Now, by Proposition 5.3.1, the map v_{P} is a rational function on $\mathcal{G}$ and, furthermore, $v_{\mathrm{P}}(G) \in \mathbb{R}^m_{>0}$. The functions $\mathrm{Exp}(\cdot)$ and $\rho(\cdot)$ are analytic on $\mathbb{R}^n$ and $\mathbb{R}^n_{>0}$, respectively, so it follows that the map

$$\mathsf{m}_1 : \mathcal{G} \to \mathbb{R}^n_{>0}$$

given by

$$G \mapsto \mathrm{Exp}[B(B'B)^{-1}\rho(v_{\mathrm{P}}(G))]$$

is also an analytic map. Note that the entries of $\tilde{A}$ are linear combinations of the entries of A. So the restriction of m_1 to the set $\mathcal{G}_{\geq 0}$ has its image contained in the set $\mathcal{E}$, and coincides with the map

$$\tilde{A} \mapsto A(\tilde{A}) \mapsto \bar{z}(A).$$

(Denoting by $A(\tilde{A})$ the matrix which coincides with A on the off-diagonal entries and has zero in its diagonal.)

Next, from Lemma 2.4.2, recall that, for each $q, w \in \mathbb{R}^n_{>0}$, $x = \varphi(q, w)$ is the unique element of $\mathbb{R}^n_{>0}$ satisfying

$$x - q \in \mathcal{D} \quad \text{and} \quad \rho(x) - \rho(w) \in \mathcal{D}^{\perp}, \tag{5.8}$$

and the map

$$(q, w) \mapsto \varphi(q, w)$$

is real analytic (in the case of the maps $\theta_i(r) = |r|$).

Let $q = x_0$ and $w = \mathsf{m}_1(G)$. We may now conclude that the map

$$\mathsf{m}_2 : \mathbb{R}^n_{>0} \times \mathcal{G} \to \mathbb{R}^n_{>0}$$

given by

$$(x_0, G) \mapsto \varphi(x_0, \mathsf{m}_1(G))$$

is again real analytic. The restriction of m_2 to $\mathbb{R}^n_{>0} \times \mathcal{G}_{\geq 0}$ coincides with the map

$$\mathbb{R}^n_{>0} \times \mathcal{G}_{\geq 0} \to \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E}$$

given by

$$(x_0, \tilde{A}) \mapsto (x_0, A(\tilde{A})) \mapsto \varphi(x_0, \bar{z}(A)).$$

Since $\bar{z} \in E_{A,+}$, the statement (5.8) is equivalent, by Corollary 2.4.1, to

$$x - x_0 \in \mathcal{D} \quad \text{and} \quad x \in E_{A,+},$$

and therefore, $x = \varphi(x_0, \bar{z}(A))$ is the unique element that both belongs to the class of x_0 and to the equilibria set $E_{A,+}$: in other words,

$$\bar{x}(x_0, A) \equiv \varphi(x_0, \bar{z}(A)).$$

So, we finally conclude that, as a restriction of m_2, the map $\bar{x} : \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E}$ is real analytic. ∎

Using the modulus of continuity of a function $F : \mathcal{X} \to \mathbb{R}^n$ on a set $\mathcal{B} \subset \mathcal{X}$

$$\zeta(r) = \sup\{|F(\xi_1) - F(\xi_2)| : \|\xi_1 - \xi_2\|_{\mathcal{X}} \leq r, \quad \xi_1, \xi_2 \in \mathcal{B}\},$$

one can show:

Corollary 5.4.1 For each compact set $P \subset \mathcal{E}$ and each set $K \subset \mathcal{A}_{\geq 0}$, of the form (5.1), there exists a function $\zeta = \zeta_{P,K}$ of class $\mathcal{K}$ such that

$$|\bar{x}(x_1, A_1) - \bar{x}(x_2, A_2)| \leq \zeta(|x_1 - x_2|) + \zeta(\|A_1 - A_2\|_{\text{ecl}})$$

for all $x_1, x_2 \in Q(P, K)$ and all $A_1, A_2 \in K$.

Proof. Given any point $x \in Q(P, K)$, let $\mathcal{C}_x = x + \mathcal{D}$ be the corresponding stoichiometric class (note that $\mathcal{C}_x$ is not restricted to $\mathbb{R}^n_{\geq 0}$). Consider the element of minimum norm in each class

$$v \in \mathcal{C}_x, \ |v| \leq |x + d|, \quad \forall\, d \in \mathcal{D},$$

and define the set

$$\mathcal{M} := \{v \in \mathbb{R}^n : \ x \in P, \ v \in \mathcal{C}_x, \text{ and } |v| \leq |x + d|, \quad \forall\, d \in \mathcal{D}\ \}.$$

Consider also the distance between two classes

$$\text{dist}(\mathcal{C}_x, \mathcal{C}_z) = \inf\{|p - q| : \ p \in \mathcal{C}_x, \ q \in \mathcal{C}_z\}.$$

Claim 1. The element of minimum norm in each class is unique. For any $x, z \in Q(P, K)$, let $\{v\} = \mathcal{M} \cap \mathcal{C}_x$ and $\{w\} = \mathcal{M} \cap \mathcal{C}_z$. Then

$$|v - w| = \ \text{dist}(\mathcal{C}_x, \mathcal{C}_z).$$

Claim 2. The set $\mathcal{M}$ is compact.

The claims will be proved below.

We will show that there exists a continuous function $\Gamma : \mathcal{M} \to \mathbb{R}^n_{>0}$, with

$$\Gamma(v) \in \mathcal{C}_x \text{ if } v \in \mathcal{C}_x,$$

and such that

$$\begin{aligned} |\bar{x}(x_1, A_1) - \bar{x}(x_2, A_2)| &= |\bar{x}(\Gamma(w_1), A_1) - \bar{x}(\Gamma(w_2), A_2)| \\ &\leq \zeta(|w_1 - w_2|) + \zeta(\|A_1 - A_2\|_{\text{ecl}}) \\ &\leq \zeta(|x_1 - x_2|) + \zeta(\|A_1 - A_2\|_{\text{ecl}}) \end{aligned}$$

where w_1 (resp., w_2) denotes the minimum norm element in the class $\mathcal{C}_{x_1}$ (resp,. $\mathcal{C}_{x_2}$), and the function $\zeta = \zeta_{P,K}$ is the modulus of continuity of the continuous function $\bar{x}(\Gamma(\cdot), \cdot)$ on the compact set $\mathcal{M} \times K$. The last inequality follows from Claim 1 and from the fact that ζ is a class $\mathcal{K}$ function.

In order to construct the continuous function Γ, first pick a number $\nu > 0$ such that, for any $\bar{x} \in P$, the closure of the ball of radius ν centered at $\bar{x}$, $\mathcal{B}_\nu(\bar{x})$, is contained in $\mathbb{R}^n_{>0}$. The collection of these balls is an open cover of P, so by compactness we may pick a finite subcover:

$$\mathcal{B}_\nu(\bar{x}_1), \ldots, \mathcal{B}_\nu(\bar{x}_J).$$

For each $\bar{x}_l$, let v_l denote the minimum norm element in the class $\mathcal{C}_{\bar{x}_l}$. Similarly, the balls $\mathcal{B}_\nu(v)$, $v \in \mathcal{M}$ form an open cover of $\mathcal{M}$, and we may pick a finite subcover:

$$\mathcal{B}_\nu(v_{J+1}), \ldots, \mathcal{B}_\nu(v_L).$$

Completing each of these subcovers (by choosing each v_l, $l = 1, \ldots, J$, to be the minimum element norm in the class $\mathcal{C}_{\bar{x}_l}$, and $\bar{x}_l$, $l = J+1, \ldots, L$, to be the positive equilibrium in the class with minimum norm element v_l), we may say that

$$P \subset \bigcup_{l=1}^{L} \mathcal{B}_\nu(\bar{x}_l) \quad \text{and} \quad \mathcal{M} \subset \bigcup_{l=1}^{L} \mathcal{B}_\nu(v_l).$$

For each $l = 1, \ldots, L$, notice that the translation:

$$v \mapsto v + \bar{x}_l - v_l$$

maps the ball $\mathcal{B}_\nu(v_l)$ into the ball $\mathcal{B}_\nu(\bar{x}_l)$ (since v_l maps to $\bar{x}_l$).

Now, using a linear combination of the maps $v \to v + \bar{x}_l - v_l$, define $\Gamma : \mathcal{M} \to \mathbb{R}^n_{>0}$ by

$$\begin{aligned} \Gamma(v) &= \frac{1}{\sum_{l=1}^{L} \delta_l(v)} \sum_{l=1}^{L} (v + \bar{x}_l - v_l)\, \delta_l(v) \\ &= v + \frac{1}{\sum_{l=1}^{L} \delta_l(v)} \sum_{l=1}^{L} (\bar{x}_l - v_l)\, \delta_l(v), \end{aligned}$$

where

$$\delta_l(v) = \begin{cases} 1, & \text{if } v \in \mathcal{B}_\nu(v_l) \\ 0, & \text{otherwise.} \end{cases}$$

The image of Γ is indeed contained in $\mathbb{R}^n_{>0}$ since: the balls $\mathcal{B}_\nu(v_l)$ cover $\mathcal{M}$ and, for each l for which $v \in \mathcal{B}_\nu(v_l)$, the term $v + \bar{x}_l - v_l$ in contained in $\mathcal{B}_\nu(\bar{x}_l) \subset \mathbb{R}^n_{>0}$, as we saw above. Thus, $\Gamma(v)$, as a linear combination of elements of $\mathbb{R}^n_{>0}$ is also in $\mathbb{R}^n_{>0}$. Furthermore,

$$\Gamma(v) - v = \frac{1}{\sum_{l=1}^{L} \delta_l(v)} \sum_{l=1}^{L} (\bar{x}_l - v_l)\delta_l(v) \ \in \mathcal{D},$$

so $\Gamma(v)$ and v are in the same class.

To see that the map Γ is continuous at each $v \in \mathcal{M}$, consider all (say k) the balls that contain v:

$$v \in \bigcap_{i=1}^{k} \mathcal{B}_\nu(v_{l_i}),$$

so that

$$\Gamma(v) = v + \frac{1}{k} \sum_{i=1}^{k} (\bar{x}_{l_i} - v_{l_i}).$$

Now, for a sufficiently small radius $\mu > 0$, there exists another ball, centered at v, so that

$$\mathcal{B}_\mu(v) \subset \bigcap_{i=1}^{k} \mathcal{B}_\nu(v_{l_i}),$$

and, for every $w \in \mathcal{B}_\mu(v)$, we have

$$\Gamma(w) = w + \frac{1}{k} \sum_{i=1}^{k} (\bar{x}_{l_i} - v_{l_i}).$$

Therefore, given any ε, we may take $\delta = \min\{\mu, \varepsilon\}$, to obtain

$$|v - w| < \delta \quad \Rightarrow \quad |\Gamma(v) - \Gamma(w)| \equiv |v - w| < \varepsilon.$$

So Γ is continuous on $\mathcal{M}$, as we wanted to show.

We now prove Claim 1. Let $\{d^{(1)}, \ldots, d^{(m-1)}\}$ form a basis of $\mathcal{D}$. Then, finding the element v with minimun norm is equivalent to minimizing the expression

$$|x + a_1 d^{(1)} + \cdots + a_{m-1} d^{(m-1)}|^2 = |x + Da|^2$$

with respect to the variables $a = (a_1, \ldots, a_{m-1})$, where $D = (d^{(1)} \ \cdots \ d^{(m-1)})$ is an $n \times (m-1)$ matrix with full rank. The minimization problem yields:

$$D'(x + Da) = 0. \tag{5.9}$$

But, since D has full rank, $D'D$ is an invertible matrix, so the solution of this linear system is unique and given by

$$a = -(D'D)^{-1}D'x.$$

Therefore, $v = x - D(D'D)^{-1}D'x$ is the unique minimun norm element of $\mathcal{C}_x$.

To prove the second part of this claim, observe that we want to minimize the expression

$$|x + Da - (z - Db)|^2 = |x - z + D(a - b)|^2$$

among all possible vectors a, b. But this is exactly the same problem as before, and its solution is therefore

$$a - b = -(D'D)^{-1}D'(x - z).$$

This shows that

$$\mathrm{dist}(\mathcal{C}_x, \mathcal{C}_z) = |x - D(D'D)^{-1}D'x - (z - D(D'D)^{-1}D'z)| \equiv |v - w|,$$

which proves Claim 1.

Finally, we prove Claim 2. The set $\mathcal{M}$ is bounded since, for a fixed $d_0 \in \mathcal{D}$,

$$|v| \leq \max\{|\bar{x}| : \ \bar{x} \in P\} + |d_0|, \quad \forall\ v \in \mathcal{M},$$

and it is also closed since, given any sequence $\{v_k\} \subset \mathcal{M}$ so that $v_k \to v$, let $\bar{x}_k$ be the unique positive equilibria in the class of v_k. Since $\bar{x}_k \in P$ for every k, by compactness of P, pick a converging subsequence $\bar{x}_{k_i} \to \bar{x} \in P$. Then (using the triangle inequality repeatedly)

$$\begin{aligned} |v| &\leq |v - v_{k_i}| + |v_{k_i}| \leq |v - v_{k_i}| + |\bar{x}_{k_i} + d| \\ &\leq |v - v_{k_i}| + |\bar{x}_{k_i} - \bar{x}| + |\bar{x} + d| \end{aligned}$$

for any $d \in \mathcal{D}$. By convergence of the two sequences, it follows that, for any $\varepsilon > 0$,

$$|v| \leq 2\,\varepsilon + |\bar{x} + d|,$$

for any $d \in \mathcal{D}$, which implies that v is the minimum norm element in the class of $\bar{x} \in P$, i.e., $v \in \mathcal{M}$, as we wanted to show. ∎

5.4.2 Uniform Bounds

For a fixed matrix A, the convergence of system

$$\dot{x} = f_A(x), \quad x(0) = s$$

to the point $\bar{x}(s, A)$ was proved using the Lyapunov function (3.12), $V : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$, defined in terms of a (constant) $\bar{x} \in \mathbb{R}^n_{>0}$

$$V(x, \bar{x}) = \sum_{i=1}^{n} \bar{x}_i \left[\frac{x_i}{\bar{x}_i} \ln \frac{x_i}{\bar{x}_i} + 1 - \frac{x_i}{\bar{x}_i} \right].$$

Recall that, for each fixed $\bar{x}$, there are two class $\mathcal{K}_\infty$ functions ν_1, ν_2 such that (equation (3.11))

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

for all $x \in \mathbb{R}^n_{\geq 0}$. In a similar way, for each compact set $P \subset \mathcal{E}$, one can find two class $\mathcal{K}_\infty$ functions $\nu_1 = \nu_{1,P}, \nu_2 = \nu_{2,P}$ such that this property is satisfied for every $\bar{x} \in P$, as well. Consider

$$\nu_1(r) = \inf\{V(x,\bar{x}) : |x-\bar{x}| \geq r,\ x \in \mathbb{R}^n_{\geq 0},\ \bar{x} \in P\}$$

and

$$\nu_2(r) = r + \max\{V(x,\bar{x}) : |x-\bar{x}| \leq r,\ x \in \mathbb{R}^n_{\geq 0},\ \bar{x} \in P\}.$$

(A similar construction is detailed in Appendix B.)

To summarize, for each matrix $A \in \mathcal{A}_{\geq 0}$ and each $\bar{x} \in E_{A,+} \cap P$ we have:

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

(ii) The set $\{x \in \mathbb{R}^n_{\geq 0} : V(x,\bar{x}) \leq L\}$ is compact, for each positive constant L. In particular, there exist two functions $\nu_{1,P}, \nu_{2,P} \in \mathcal{K}_\infty$ such that

$$\nu_1(|x-\bar{x}|) \leq V(x,\bar{x}) \leq \nu_2(|x-\bar{x}|)$$

for all $x \in \mathbb{R}^n_{\geq 0}$.

(iii) Recall from Lemma 2.5.4 that

$$\nabla V(x,\bar{x})\ f_A(x) \leq -\frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m a_{ij}\, e^{\langle b_j, \rho(\bar{x})\rangle}\ \left(e^{-\pi_i} - e^{-\pi_j}\right)^2.$$

for all $x \in \mathbb{R}^n_{>0}$; the expression on the right hand side is zero if and only if $x \equiv \bar{x}$.

Given any $A \in \mathcal{A}_{\geq 0}$, suppose that A_1 is a matrix with entries $a^1_{ij} = 1$ if $a_{ij} > 0$ and $a^1_{ij} = 0$ if $a_{ij} = 0$. Then, for the expression in (iii), we can write:

$$\nabla V(x,\bar{x})\ f_A(x) \leq -\frac{1}{2}\min_{a_{ij}>0}\{a_{ij}\}\ \min_j e^{\langle b_j, \rho(\bar{x})\rangle}\ \sum_{i=1}^m \sum_{j=1}^m a^1_{ij}\ \left(e^{-\pi_i} - e^{-\pi_j}\right)^2.$$

Since A_1 is irreducible, we may apply Lemma 2.5.3 to conclude that there exists a positive constant k such that

$$\sum_{i=1}^m \sum_{j=1}^m a^1_{ij}\ \left(e^{-\pi_i} - e^{-\pi_j}\right)^2 \geq k\ \sum_{i=1}^m \sum_{j=1}^m \left(e^{-\pi_i} - e^{-\pi_j}\right)^2.$$

Thus we have a reformulation of Lemma 2.5.4:

Lemma 5.4.2 There exists a positive constant k, a continuous function $c : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{>0}$ and a function $\kappa : \mathcal{A}_{\geq 0} \to \mathbb{R}_{\geq 0}$ given by

$$c(\xi) = \frac{1}{2} \min_j e^{\langle b_j, \rho(\xi) \rangle} \quad \text{and} \quad \kappa(A) = k \, \min\{a_{ij} : a_{ij} \neq 0, \ i, j = 1, \ldots, m\}$$

such that, given any matrix $A \in \mathcal{A}_{\geq 0}$:

$$\nabla V(x, \bar{x}) \ f_A(x) \leq -\kappa(A) \, c(\bar{x}) \, \Psi(x, \bar{x}), \tag{5.10}$$

for all $x \in \mathbb{R}^n_{>0}$ and any element $\bar{x} \in E_{A,+}$.

Consider a set $K \subset \mathcal{A}_{\geq 0}$ of the form (5.1) described in Definition 5.1.1 (property $\mathcal{P}_0$):

$$A = (a_{ij}) \in K \quad \Rightarrow \quad \text{either } a_{ij} = 0 \quad \text{or} \quad \alpha_0 \leq a_{ij} \leq \alpha^0, \ \ \forall \, i, j = 1, \ldots, m$$

where α_0 and α^0 are positive constants. It follows that

$$\kappa(A) \geq \alpha_0 k, \ \ \forall \ A \in K.$$

Lemma 5.4.3 Let $P \subset \mathcal{E}$ be a compact set and $K \subset \mathcal{A}_{\geq 0}$ be a set of the form (5.1) Then there exists a continuous positive definite function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $\alpha = \alpha_{P,K}$, such that, given any $\bar{x} \in P$ and any $A \in K$ with $\bar{x} \in E_{A,+}$,

$$\nabla V(x, \bar{x}) \ f_A(x) \leq -\alpha(V(x, \bar{x}))$$

for all $x \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$.

Proof. Let $c : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{>0}$ and $\kappa : \mathcal{A}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be the functions given in Lemma 5.4.2. We will show that the following function, defined from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$ is positive definite:

$$\alpha(r) = \inf\{\kappa(A) \, c(\bar{x}) \, \Psi(x, \bar{x}) : A \in K, \bar{x} \in P, x \in \mathcal{C}_r \cap \mathcal{S}_{\bar{x}}\}$$

where

$$\mathcal{C}_r = \{x \in \mathbb{R}^n_{\geq 0} : V(x, \bar{x}) = r, \ \text{for any } \bar{x} \in P\}.$$

First, we show that $\mathcal{C}_r$ is a compact subset of $\mathbb{R}^n_{\geq 0}$:

- ▷ Closed: let x^k be a sequence in $\mathcal{C}_r$ converging to a point $x \in \mathbb{R}^n_{\geq 0}$. We must show that $x \in \mathcal{C}_r$. For each x^k there exist $\bar{x}^k \in P$ with $V(x^k, \bar{x}^k) = r$. Since P is compact, the sequence $\{\bar{x}^k\}$ has a converging subsequence, $\bar{x}^{k_l} \to \bar{x}$. By continuity of V, $V(x^{k_l}, \bar{x}^{k_l}) \to V(x, \bar{x})$, hence $V(x, \bar{x}) = r$. So $x \in \mathcal{C}_r$ as wanted.

- ▷ Bounded: let $x \in \mathcal{C}_r$. Then $\nu_{1,P}(|x - \bar{x}|) \leq r$ for all $\bar{x} \in P$, which implies $|x| \leq \nu^{-1}_{1,P}(r) + |\bar{x}|$. So, with $M = \max\{|\bar{x}| : \bar{x} \in P\}$, $\mathcal{C}_r$ is contained in the closed ball of radius $\nu^{-1}_{1,P}(r) + M$ centered at the origin.

Next we show that the set

$$\tilde{\mathcal{C}}_r = \{(x, \bar{x}) \in \mathbb{R}^n_{\geq 0} \times \mathbb{R}^n_{>0} : x \in \mathcal{C}_r \cap \mathcal{S}_{\bar{x}}, \bar{x} \in P\}$$

is also compact: since it is clearly a subset of the compact set $\mathcal{C}_r \times P$, it is enough to show that $\tilde{\mathcal{C}}_r$ is closed. Given a sequence of points $(x^k, \bar{x}^k) \in \tilde{\mathcal{C}}_r$ converging to a point $(x, \bar{x}) \in \mathcal{C}_r \times P$, we need to show that also $x \in \mathcal{S}_{\bar{x}}$. But, by definition of a class we have

$$\langle v_i, x^k - \bar{x}^k \rangle = 0, \quad \forall\, i = 1, \ldots, n - m + 1,$$

(where $\{v_1, \ldots, v_{n-m+1}\}$ is a basis of $\mathcal{D}^\perp$) and by continuity it follows that $\langle v_i, x - \bar{x} \rangle = 0$ for all such i, that is, x belongs to the class $\mathcal{S}_{\bar{x}}$, as wanted.

It is clear $r = 0$ implies $\tilde{\mathcal{C}}_r = \{(\bar{x}, \bar{x}) : \bar{x} \in P\}$, so $\alpha(0) = 0$, because $\Psi(\bar{x}, \bar{x}) = 0$. Notice that $\kappa(A) \geq \alpha_0 > 0$ on K and $c(\cdot)$ is continuous on $\mathbb{R}^n_{\geq 0}$, so $c(\cdot)$ will have a strictly positive minimum value on the compact set K.

Now, we take any $r > 0$ and show that $\inf\{\Psi(x, \bar{x}) : (x, \bar{x}) \in \tilde{\mathcal{C}}_r\}$ is positive. To get a contradiction, assume that this infimum is zero for some $r > 0$. Then there exists an infinite sequence $(x^k, \bar{x}^k)$ such that $\Psi(x^k, \bar{x}^k) \to 0$. Since $\tilde{\mathcal{C}}_r$ is compact, there exists a converging subsequence: $(x^{k_l}, \bar{x}^{k_l}) \to (x_0, \bar{x}_0) \in \tilde{\mathcal{C}}_r$. Then $\Psi(x_0, \bar{x}_0) = 0$ and by Lemma 2.4.4, $x_0 \in E_0 \cup E_{A_0,+}$ for some $A_0 \in \mathcal{A}_{\geq 0}$. But, under the no boundary equilibrium assumption, $E_0 \cap \mathcal{S}_{\bar{x}} = \emptyset$ for all $\bar{x}$. So, by uniqueness of the positive equilibrium in each class, $x_0 = \bar{x}_0$ which implies $r = V(x_0, \bar{x}_0) = 0$ and contradicts $r > 0$. Thus, $\alpha(r) > 0$ whenever $r > 0$.

Without loss of generality we may assume that α is continuous on $\mathbb{R}_{\geq 0}$. (Otherwise, it is possible to construct another positive definite function $\tilde{\alpha}$, continuous and satisfying $\alpha(r) \geq \tilde{\alpha}(r)$ for all $r \geq 0$: note that, for every R, there exists a_R such that $\alpha(r) \geq a_R$ for all $r \in [R/2, 2R]$.)

Finally, by construction, α satisfies:

$$\kappa(A)\, c(\bar{x})\, \Psi(x, \bar{x}) \geq \alpha(V(x, \bar{x}))$$

for all $A \in K$, all $\bar{x} \in P \cap E_{A,+}$, and all $x \in \mathcal{S}_{\bar{x}}$, and in particular, for all $x \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$. Combining this with equation (5.10) finishes the proof. ∎

Corollary 5.4.4 Given any compact set $P \subset \mathcal{E}$ and any set $K \subset \mathcal{A}_{\geq 0}$ of the form (5.1), there exists a function $\tilde{\beta} = \tilde{\beta}_{P,K}$ of class $\mathcal{KL}$ such that for any pair $\bar{x} \in P$, $A \in K$ and all $s \in Q(P, K)$ with $\bar{x} = \bar{x}(s, A)$

$$|x(t, s, A) - \bar{x}(s, A)| \leq \tilde{\beta}(|s - \bar{x}(s, A)|, t),$$

for all $t \geq 0$.

Proof. Pick any compact set $P \subset \mathcal{E}$ and any set $K \subset \mathcal{A}_{\geq 0}$ of the form (5.1), and let $\alpha = \alpha_{P,K}$ be the continuous, positive definite function given by Lemma 5.4.3. Consider the initial value problem

$$\dot{y} \leq -\alpha(y), \quad y(0) = y_0.$$

where $y_0 \in \mathbb{R}^n_{\geq 0}$. Then by a comparison result such as Lemma 4.4 of [37], there exists a function $\beta = \beta_\alpha$ of class $\mathcal{KL}$ such that

$$y(t) \leq \beta(y_0, t), \quad \text{for all} \quad t \geq 0.$$

Using the functions $\nu_1 = \nu_{1,P}$, $\nu_2 = \nu_{2,P} \in \mathcal{K}_\infty$ define

$$\tilde{\beta}(r, t) = \nu_1^{-1}(\beta(\nu_2(r), t))$$

which is again a $\mathcal{KL}$ function and depends only on P and K.

Now, pick any $\bar{x} \in P$ and $A \in K$. Let $s \in Q(P, K)$ be such that $\bar{x} = \bar{x}(s, A)$. Recall that $x(t, s, A)$ is the unique solution of the initial value problem $\dot{x} = f_A(x)$, $x(0) = s$, and we know that $x(t, s, A) \in \mathcal{S}_{\bar{x}} \cap \mathbb{R}^n_{>0}$ for all $t \geq 0$ (since the set $\mathbb{R}^n_{>0}$ is forward invariant for the system $\dot{x} = f_A(x)$, as shown in Section 2.5.1).

Define

$$y(t) := V(x(t, s, A), \bar{x}(s, A)).$$

From Lemma 5.4.3 the function $y(t)$ satisfies

$$\dot{y} \leq -\alpha(y), \quad y(0) = V(s, \bar{x}(s, A)), \quad \text{for all} \quad t \geq 0.$$

Therefore, recalling property (ii) of V, we have

$$\nu_1(|x(t, s, A) - \bar{x}(s, A)|) \leq V(x(t, s, A), \bar{x}(s, A)) \leq \beta(\nu_2(|s - \bar{x}(s, A)|), t)$$

for all $t \geq 0$, which gives the desired result. ∎

Lemma 5.4.5 Let $r_0 > 0$ be any real number. Given any compact set $P \subset \mathcal{E}$ and any set $K \subset \mathcal{A}_{\geq 0}$, of the form (5.1), and given any compact set $Q_0 \subset Q(P, K)$, there exists $T = T_{K, Q_0, r_0} > 0$ such that, for every $\bar{x} \in P$, $A \in K$ and $s \in Q_0$ such that $\bar{x} = \bar{x}(s, A)$, the following hold

$$\begin{aligned} |x(t, s, A) - \bar{x}(s, A)| &\leq 1 \\ |h_i(x(t, s, A)) - h_i(\bar{x}(s, A))| &\leq r_0, \end{aligned}$$

for all $t \geq T$ and all $i = 1, \ldots, p$.

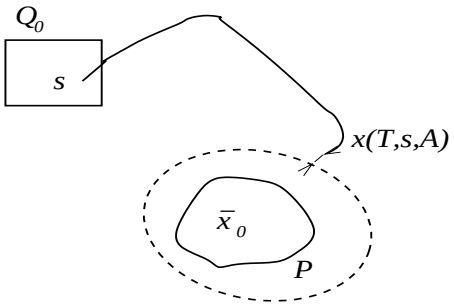
Proof. Let $P \subset \mathcal{E}$, $K \subset \mathcal{A}_{\geq 0}$ and $Q_0 \subset Q(P, K)$ be given sets as in the statement of the Lemma. Let $\tilde{\beta}$ be the $\mathcal{KL}$ function given in Corollary 5.4.4. Put

$$M = \max_{s \in Q_0, A \in K} |s - \bar{x}(s, A)|,$$

and let $T_1 = T_1(M)$ be so that $\tilde{\beta}(M, t) \leq 1$ for all $t \geq T_1$. Consider the compact set

$$\mathcal{B}_1 = \{x \in \mathbb{R}^n_{\geq 0} : |x| \leq 1 + \max_{\bar{x} \in P} |\bar{x}| \} \tag{5.11}$$

and let c_1 be a Lipschitz constant for the function h in the set $\mathcal{B}_1$.



Next, pick $T = T_{K,Q_0,r_0} \geq T_1$, so that

$$\tilde{\beta}(M,t) \leq \frac{r_0}{c_1},$$

for all $t \geq T$.

Now, pick any $\bar{x} \in P$, $A \in K$ and $s \in Q_0$ with $\bar{x} = \bar{x}(s,A)$. For all $t \geq T_1$, from Corollary 5.4.4 it follows that

$$|x(t,s,A) - \bar{x}(s,A)| \leq \tilde{\beta}(M,t) \leq 1$$

and so $x(t,s,A) \in \mathcal{B}_1$ for all $t \geq T_1$.

We then have, using the fact that h is Lipschitz on $\mathcal{B}_1$,

$$|h_i(x(t,s,A)) - h_i(\bar{x}(s,A))| \leq c_1|x(t,s,A) - \bar{x}(s,A)| \leq c_1\tilde{\beta}(M,t) \leq r_0$$

for all $t \geq T$, which finishes the proof. ∎

5.5 Proof of Theorem 4

Given a chemical network characterized by matrices B as in Section 3.1 and $A_0 \in \mathcal{A}_{\geq 0}$, let C be such that the system $\Sigma(A_0)$: $\dot{x} = f_{A_0}(x)$, $y = h(x)$ is detectable. We will next show that $\dot{z} = f_{A_0}(z) + C'(h(x) - h(z))$ provides an A_0-robust observer for $\Sigma(A_0)$.

Pick any pair of sets $P \subset \mathcal{E}$ and $K \subset \mathcal{A}_{\geq 0}$, which satisfy property $\mathcal{P}_0$, so that: P is compact, K is of the form (5.1), $A_0 \in K$, $\bar{x}_0 = \bar{x}(x_0, A_0) \in P$ for some $x_0 \in \mathbb{R}^n_{>0}$, and $|h_i(\bar{x}_0) - h_i(\bar{x})| \leq 1/2\ h_i(\bar{x}_0)$, for every $i = 1, \ldots, p$.

Let $\tilde{\beta} = \tilde{\beta}_{P,K} \in \mathcal{KL}$ be as in Corollary 5.4.4 and $\zeta = \zeta_{P,K} \in \mathcal{K}$ as in Corollary 5.4.1. Pick any compact subset $Q_0 \subset Q(P,K)$ and let $T = T_{K,Q_0,\bar{x}_0}$ be the number given by Lemma 5.4.5, when

$$r_0 = \frac{1}{7} \min_{1 \leq i \leq p} h_i(\bar{x}_0).$$

For any $A \in K$ and any $s \in Q_0$ with $\bar{x}(s,A) \in P$, Lemma 5.4.5, together with part (e) of property $\mathcal{P}_0$ and triangle inequality, imply

$$|h_i(x(t,s,A) - h_i(\bar{x}_0)| \leq \frac{10}{21} h_i(\bar{x}_0)$$

for all $t \geq T$, and all $i = 1, \ldots, p$. Let $\theta = 20/21$ and observe that $h(x(t,s,A)) \in \mathbb{U}_\theta$ for all $t \geq T$.

For the system with inputs $\dot{z} = f_{A_0}(z) + C'(u - h(z))$ and input set $\mathbb{U}_\theta$, let $\beta = \beta_{A_0} \in \mathcal{KL}$ and $\varphi = \varphi_{A_0} \in \mathcal{K}_\infty$ be as in Proposition 5.2.5.

Consider the extended system

$$\begin{aligned} \dot{x} &= f_A(x) \\ \dot{z} &= f_{A_0}(z) + C'(h(x) - h(z)) \end{aligned}$$

and let $(x(t,s,A), z(t))$ be its solution at time t, with initial condition $(s, z(0)) \in Q_0 \times \mathbb{R}^n_{\geq 0}$.

Applying Proposition 5.2.5 with $u(t) \equiv h(x(t+T,s,A))$ and $z(T)$ as initial condition, yields

$$|z(t) - \bar{x}_0| \leq \beta(|z(T) - \bar{x}_0|, t) + \varphi(\|h(x(\cdot,s,A)) - h(\bar{x}_0)\|_T)$$

for all $t \geq T$.

Now, using the triangle inequality

$$|z(t) - x(t,s,A)| \leq |z(t) - \bar{x}_0| + |\bar{x}_0 - \bar{x}(s,A)| + |\bar{x}(s,A) - x(t,s,A)|$$

together with the ISS estimate, Corollary 5.4.1 and Corollary 5.4.4 obtain

$$\begin{aligned} |z(t) - x(t,s,A)| \quad \leq \quad & \beta(|z(T) - \bar{x}_0|, t) + \tilde{\beta}(|s - \bar{x}(s,A)|, t) \\ & + \varphi(\|h(x(\cdot,s,A)) - h(\bar{x}_0)\|_T) \\ & + \zeta(\|A_0 - A\|_{\text{ecl}}) + \zeta(|x_0 - s|) \end{aligned}$$

for all $t \geq T$.

Let c_1 be a Lipschitz constant of the function h on the compact set $\mathcal{B}_1$ defined in (5.11). Note that $x(t,s,A) \in \mathcal{B}_1$ for $t \geq T$ and also $P \subset \mathcal{B}_1$. Then, since φ is $\mathcal{K}_\infty$, we have

$$\begin{aligned} \varphi(\|h(x(\cdot,s,A)) - h(\bar{x}_0)\|_T) \quad &\leq \quad \varphi(c_1 \|x(\cdot,s,A) - \bar{x}_0\|_T) \\ &\leq \quad \varphi(2c_1 \|x(\cdot,s,A) - \bar{x}(s,A)\|_T) \\ &\quad + \varphi(2c_1 \|\bar{x}(s,A) - \bar{x}_0\|_T). \end{aligned}$$

Using Corollary 5.4.1 once more,

$$\varphi(2c_1 \|\bar{x}(s,A) - \bar{x}_0\|_T) \leq \varphi(4c_1\, \zeta(|x_0 - s|)) + \varphi(4c_1\, \zeta(\|A_0 - A\|_{\text{ecl}}))$$

and renaming the functions

$$\varphi(2c_1 r) \rightsquigarrow \varphi(r)$$

and

$$\varphi(4c_1 \zeta(r)) + \zeta(r) \rightsquigarrow \zeta(r)$$

(note that $\varphi(4c_1\zeta(\cdot))$ is still of class $\mathcal{K}$ and depends on P and K), the desired estimate for the difference $|z(t) - x(t,s,A)|$ follows.

Chapter 6
An Alternative Observer

In this Chapter we construct an alternative observer, which works under the same necessary and sufficient conditions for detectability. In terms of performance, the second observer seems to have substantially slower convergence characteristics but, on the other hand, it exhibits more "robustness" to additive disturbances in the measurements. This robustness can be quantified by an input-to-state stability type of estimate, and we prove that the second observer is ISS with respect to a far wider class of external disturbances than the first (more efficient) observer.

Among other interesting features, the observer state never becomes negative, even under arbitrary disturbances, which is a natural requirement on physical grounds (since the state being observed is always nonnegative). This nonnegativity is enforced by a "barrier feedback" reminiscent of optimization techniques.

6.1 An Alternative Observer

We now use logarithms of the outputs as input to the observer. Since the system under observation is $\mathbb{R}^n_{>0}$-invariant, it follows that $h(x)$ has strictly positive entries and thus the logarithms of the outputs are well defined. Recalling our definition of the vectorial function ρ (3.4), we will use the convenient notation

$$H(x) := [H_1(x)\ H_2(x) \dots H_p(x)]' = C\rho(x) = \rho(h(x)) \tag{6.1}$$

where $H_i(x) = c_{i1} \ln x_1 + \cdots + c_{in} \ln x_n$.

In a similar way to the observer described in Chapter 4, the alternative observer will be obtained from the more general system

$$\dot{z} = f^*(z,u) := f(z) + C'(u - H(z)). \tag{6.2}$$

However, some care is needed to establish the state-space for (6.2). We will next introduce this state-space $\mathcal{X}$, which will be the largest subset of $\mathbb{R}^n$ where solutions of this new system (6.2) are defined. Note that, for any vector $u \in \mathbb{R}^p$,

$$C'(u - H(z)) = \sum_{k=1}^{n} \sum_{i=1}^{p} c_{ik}(u_i - H_i(z))e_k,$$

where $\{e_k : k = 1, \dots, n\}$ is the canonical basis of $\mathbb{R}^n$, so we may introduce the set of indices

$$\mathtt{Ind} = \{k : \exists i,\ 1 \le i \le p \text{ with } c_{ik} \neq 0\},$$

which is the complement of all coordinates for which $f_k^*(z,u) \equiv f_k(z)$ (i.e., all k such that the term $\ln z_k$ appears in f_k^* are contained in K). When considering the system (6.2), we view it as evolving on the following state-space:

$$\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$$

where $\mathcal{X}_k = \mathbb{R}$ if $k \notin \texttt{Ind}$ and $\mathcal{X}_k = (0, +\infty)$ if $k \in \texttt{Ind}$. Observe that "the term $\ln z_i$ appears anywhere in f^* if and only if it appears in f_i^*", so the largest open set where f^* is well defined is indeed $\mathcal{X}$.

As in Chapter 4, assume that every trajectory, $x(\cdot)$, of the system $\dot{x} = f(x)$ (when f is given by (3.2)) satisfies: if $x(t) \in \mathbb{R}^n_{>0}$ for every $t > 0$, then $x(t) \to \bar{x} \in E_+$ as $t \to +\infty$. This is the case of systems for which no boundary equilibria exist in any positive class (see Section 2.3 and Theorem 1).

Theorem 6 Consider the system with outputs (3.1), and assume that it is detectable. Then the following system, with state-space $\mathcal{X}$, is an observer for system (3.1):

$$\dot{z} = f(z) + C'(\rho(h(x)) - \rho(h(z))). \tag{6.3}$$

The observer for (3.1) is obtained from (6.2) by letting the input be $u(t) = H(x(t))$ and it can be written

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

Example 6.1.1 Consider again the system in Example (4.1.1). Suppose that the output is given by $h(x) = (x_1 x_2^2, x_1 x_4)'$. Then $H(x) = (\ln x_1 + 2\ln x_2, \ln x_1 + \ln x_4)$ and we can construct the following observer:

$$\begin{aligned}
\dot{x}_1 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 \\
&\quad +(\ln x_1 + 2\ln x_2 - \ln z_1 - 2\ln z_2) + (\ln x_1 + \ln x_4 - \ln z_1 - \ln z_4) \\
\dot{x}_2 &= -k_{21}x_1x_2 + k_{12}x_3 + k_{13}x_4 + 2(\ln x_1 + 2\ln x_2 - \ln z_1 - 2\ln z_2) \\
\dot{x}_3 &= k_{21}x_1x_2 - (k_{12} + k_{32})x_3 \\
\dot{x}_4 &= k_{32}x_3 - k_{13}x_4 + (\ln x_1 + \ln x_4 - \ln z_1 - \ln z_4).
\end{aligned}$$

The proof of this theorem involves a technique similar to the one used in the proof of Theorem 3, by studying the stability properties of the more general system (6.2) and showing that it satisfies an input to state stability condition.

6.2 $\mathbb{R}^n_{>0}$-Invariance

We first prove an algebraic result concerning f^*.

Lemma 6.2.1 Let $f^*(z,u) = f(z) + C'(u - H(\bar{x}))$ and assume C is such that $\mathcal{D} +$ im $C' = \mathbb{R}^n$. Let $\bar{x}$ denote any point in E_+ and V be the function defined in (3.12).

Then 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 $u \in \mathbb{R}^p$. Moreover, one may pick $\gamma(r) = r^2/2$.

Furthermore, let F be any compact subset of $\mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ 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 $u \in \mathbb{R}^p$.

Proof. Pick any $\bar{x} \in E_+$. We have (using the notation $\varrho = \rho(z) - \rho(\bar{x})$):

$$\begin{aligned}
\nabla V(z)\ f^*(z,u) &= \langle \rho(z) - \rho(\bar{x}), f(z)\rangle + \langle \rho(z) - \rho(\bar{x}), C'(u - H(z))\rangle \\
&= \langle \varrho, f(z)\rangle + \langle C\varrho, u - H(z)\rangle \\
&= \langle \varrho, f(z)\rangle + \langle C\varrho, u - H(\bar{x})\rangle - \langle C\varrho, C\varrho\rangle
\end{aligned}$$

where the definition $H(z) = C\rho(z)$ was used, as well as the equality

$$u - H(z) = u - H(\bar{x}) + H(\bar{x}) - H(z).$$

Now, using result (2.13) and then Cauchy-Schwartz and the inequality

$$ab \leq \frac{1}{2}(a^2 + b^2)$$

we have

$$\begin{aligned}
\nabla V(z)\ f^*(z,u) &\leq -\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}) - |C\,\varrho|^2 + \langle C\varrho, u - H(\bar{x})\rangle \\
&\leq -\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}) - \frac{1}{2}|C\,\varrho|^2 + \frac{1}{2}|u - H(\bar{x})|^2
\end{aligned}$$

for all $z \in \mathbb{R}^n_{>0}$ and all $u \in \mathbb{R}^p$. Now, consider the function $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$:

$$\alpha_1(r) = \inf\left\{\kappa(A)c(\bar{x})\Psi(z,\bar{x}) + \frac{1}{2}|C\,(\rho(z) - \rho(\bar{x}))|^2 :\ z \in \mathcal{C}_r\right\},$$

where

$$\mathcal{C}_r := \{z \in \mathbb{R}^n_{>0} : |\rho(z) - \rho(\bar{x})| = r\}.$$

Note that $\mathcal{C}_r$ 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 $z_i \to 0$. 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 $H(z) = H(\bar{x})$ (or, equivalently, $h(z) = h(\bar{x})$): thus, from the detectability condition it follows that $\alpha_1(r)$ may be zero only when $z = \bar{x}$, i.e., when $r = 0$. Therefore, the function α_1 is continuous, has $\alpha_1(0) = 0$ and is strictly positive for all $r > 0$, and satisfies the inequality

$$-\kappa(A)\, c(\bar{x})\, \Psi(z,\bar{x}) - \frac{1}{2}|C\,\varrho|^2 \leq -\alpha_1(|\rho(z) - \rho(\bar{x})|),$$

for every $z \in \mathbb{R}^n_{>0}$. Letting $\gamma(r) = \frac{1}{2}r^2$, the first part of the Lemma is established.

For the second part, assume that $F \subset \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ is an arbitrary compact set. Observe that

$$|C\,(\rho(z) - \rho(\bar{x}))|^2 = |\rho(h(z)) - \rho(h(\bar{x}))|^2 = \sum_{i=1}^{p} |\ln h_i(z) - \ln h_i(\bar{x})|^2.$$

Let

$$M = \max\{h_i(z) : z \in F,\, i \in \{1, \ldots p\}\}$$

and put $c_F = 1/M$. For any fixed $a \in (0, M]$ the analysis of the function

$$f_a(r) = |\ln r - \ln a| - c_F|r - a|$$

(using the same arguments as in the proof of Lemma 4.2.2) shows that $f_a(r) \geq 0$ for every $0 < r \leq M$. Thus:

$$\begin{aligned} -\kappa(A)\, c(\bar{x})\, \Psi(z, \bar{x}) - \frac{1}{2}|C\,\varrho|^2 &\leq -\kappa(A)\, c(\bar{x})\, \Psi(z, \bar{x}) - \frac{1}{2}c_F|h(z) - h(\bar{x})|^2 \\ &\leq -\alpha(|z - \bar{x}|) \end{aligned}$$

for all $z \in \mathbb{R}^n_{>0} \cap F$, where $\alpha = \alpha_{\bar{x},F}$ is a $\mathcal{K}_\infty$ function given by Lemma 4.2.1.

Finally, with the same function γ, we obtain

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

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

The next lemma illustrates the "barrier" effect introduced by the logarithmic function $H(z)$.

Lemma 6.2.2 Consider the system (6.2) with state-space $\mathcal{X}$. Let $u(\cdot)$ be any $\mathbb{R}^p$-valued input with $\|u\| < +\infty$. Assume $z(t)$ is a solution of (6.2) defined on a (maximal) interval J, with $z(0) \in \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$. Suppose that $z(t)$ evolves in a compact subset $\tilde{\mathcal{C}}$ of $\mathbb{R}^n_{\geq 0}$. Then there exists another compact set $\mathcal{C}$, contained in $\mathcal{X} \cap \mathbb{R}^n_{\geq 0}$, and such that $z(t) \in \mathcal{C}$ for every $t \in J$.

Proof. Let $b \in \mathbb{R}$ be a (strictly) positive constant such that $\tilde{\mathcal{C}} \subset [0, b]^n$, and so $z(t) \in [0, b]^n$ for all $t \in J$. Choose strictly positive constants ε_k for each $k \in \mathtt{Ind}$ such that:

$$f^*_k(z, u) = f_k(z) + \sum_{i=1}^{p} c_{ik}u_i - \sum_{i=1}^{p}\sum_{j \neq k} c_{ik}c_{ij} \ln z_j - \sum_{i=1}^{p} c_{ik}^2 \ln z_k > 0,$$

whenever each u_i takes values in $[-\|u\|, \|u\|]$, z is a vector in $[0, b]^n \cap \mathcal{X}$ and $z_k < 2\varepsilon_k$. It is possible to make this choice of ε_k, because we are assuming $c_{ij} \geq 0$ for all i, j and we know that u is bounded and also that $0 \leq z_j \leq b$ which imply the existence of positive constants a_1, a_2 and a_3 such that:

$$|f_k(z)| < a_1, \quad \sum_{i=1}^{p} c_{ik}|u_i| < a_2, \quad \text{and} \quad \sum_{i=1}^{p}\sum_{j \neq k} c_{ik}c_{ij} \ln z_j < a_3.$$

Thus, it is enough to take ε_k such that

$$\ln(2\varepsilon_k) \leq -\frac{1}{\sum_{i=1}^{p} c_{ik}^2}(a_1 + a_2 + a_3).$$

(Note that $\sum_{i=1}^{p} c_{ik}^2 \neq 0$ precisely because $k \in \mathtt{Ind}$.) If the value of ε_k thus obtained is greater than the initial condition $z_k(0)$, we should rechoose ε_k to get $\varepsilon_k \leq z_k(0)$.

Now, each coordinate function $z_k(t)$, for $k \in \mathtt{Ind}$ and $t \in J$ satisfies an initial value problem of the form

$$\begin{aligned} \dot{x} &= F(t,x) \\ x(0) &= z_k(0) > 0, \end{aligned}$$

where

$$F(t,x) := f_k^*(z_1(t), \ldots, z_{k-1}(t), x, z_{k+1}(t), \ldots, z_n(t), u(t))$$

is locally Lipschitz on $\mathbb{R}_{>0}$ for each fixed t and measurable on J for each fixed x, and moreover $F(t,x) > 0$ whenever $x < 2\varepsilon_k$. Then Lemma C.0.3 applies and we conclude that

$$z_k(t) > \varepsilon_k, \quad \forall\, t \in J.$$

Finally, setting $\varepsilon_k = 0$ if $k \notin \mathtt{Ind}$, we have that trajectories of (6.2) stay inside the set

$$\mathcal{C} = [\varepsilon_1, b] \times \cdots \times [\varepsilon_n, b].$$

for all t in J, where $\mathcal{C}$ is indeed a compact subset of the state space $\mathcal{X}$. ∎

Proposition 6.2.3 Assume system (3.1), defined by (3.2) and (3.3), to be detectable. Then the system (6.2) is $\mathbb{R}^n_{>0}$-invariant with input-value set $\mathbb{R}^p$.

Furthermore, pick any $\bar{x} \in E_+$ and let κ be a positive constant. Define $\mathbb{U}_\kappa$ to be the subset of $\mathbb{R}^p$ consisting of vectors u which satisfy $|u| \leq \kappa$. Then, the system (6.2) is semi-global ISS with input-value set $\mathbb{U}_\kappa$ (with respect to $\bar{x}$ and the input $\bar{u} = H(\bar{x})$).

Proof. Note that, for each input $u : [0, +\infty) \to \mathbb{R}^p$, $f^*(\cdot, u)$ is of class C^1 and locally Lipschitz on $\mathcal{X}$ for each fixed u, and $f^*(z, \cdot)$ is measurable and locally integrable on $[0, +\infty)$ for each fixed z. Then, given an initial condition $z_0 \in \mathcal{X}$ and a fixed input u as above, a unique maximal solution of (6.2) exists on a maximal interval, J. Observe that f^* does not extend to $z_k = 0$ and $k \in \mathtt{Ind}$, but while $z_k(t) \in \mathcal{X}$ no difficulties exist. Then J is not empty, and we will show that $J = [0, +\infty)$.

Step 1. To prove $\mathbb{R}^n_{>0}$-invariance (with input-value set $\mathbb{R}^p$), pick any initial condition $z_0 \in \mathbb{R}^n_{>0}$. We only need to show that, for each coordinate k with $k \notin \mathtt{Ind}$,

$$z_k(t) > 0, \quad \forall\, t \in J,$$

since all the others satisfy this condition by definition of $\mathcal{X}$.

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

$$t_0 = \inf\{t \in J : z_k(t) = 0 \text{ for some } k \notin \texttt{Ind}\}.$$

Pick any coordinate k such that $z_k(t_0) = 0$.

Note that for every $k \notin \texttt{Ind}$, $f_k^*(z,u) \equiv f_k(z)$, so using the notation

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

we can construct a new function $F : [0,\infty) \times \mathbb{R} \to \mathbb{R}$:

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

(arbitrary u_0 in $\mathbb{U}_\kappa$). Then standard existence and uniqueness results apply to $\dot{x} = F(t,x)$.

Claim. $F(t,0) \geq 0$ for almost all $t \in [0, +\infty)$.

To prove this, write

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

where $A_0 = \{j : b_{kj} = 0\}$ and $A_+ = \{j : b_{kj} > 0\}$. For $x = 0$ and $t \in [0, +\infty)$:

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

(b) the first term is nonnegative since, by definition of t_0, $z_i(t) \geq 0$, for all $i \neq k$ and $t \leq t_0$, and either

$$F(t,0) = f_k(\check{z}(t), 0; u(t)) \quad \text{for } t < t_0,$$

or

$$F(t,0) = f_k(\check{z}(t_0), 0; u(t_0)) \quad \text{for } t \geq t_0.$$

This proves the claim.

Now, notice that the initial value problem

$$\dot{x} = F(t,x), \quad x(0) = z_k(0)$$

has a solution $x(t)$ in some open interval, $\tilde{J}$, which contains $[0, t_0]$, and moreover, for all $t < t_0$, that solution coincides exactly with the trajectory of the k-coordinate: $x(t) \equiv z_k(t)$. Also, by continuity of both $x(t)$ and $z_k(t)$, it must be that $x(t_0) = z_k(t_0)$.

From Lemma C.0.2 (with $\mathcal{X} = \mathbb{R}$ and $J_{x_0} = \tilde{J}$), it follows that $x(t) > 0$ for all $t \in \tilde{J}$, and hence also $z_k(t) > 0$ for all $t \leq t_0$, which contradicts the fact that $z_k(t_0) = 0$.

To conclude, $z_k(t) > 0$ for all $k \notin \texttt{Ind}$ and all $t \in J$, and thus the solution $z(t)$ never leaves the set $\mathbb{R}^n_{>0}$ in the maximal interval J, proving $\mathbb{R}^n_{>0}$-invariance with input-value set $\mathbb{R}^p$.

Step 2. Now assume (to get a contradiction) that the interval J is finite, $J = [0, t_{\max})$, $t_{\max} < +\infty$. From Lemma 6.2.1, and from boundedness of u, we know that the derivative of $V(z(t))$ satisfies

$$\frac{d}{dt}V(z(t)) \leq \frac{1}{2}|u - H(\bar{x})|^2 \leq \kappa^2 + |H(\bar{x})|^2 = c,$$

so

$$V(z(t)) \leq V(0) + \int_0^t c\, dt \;\leq\; V(0) + c\, t_{\max}.$$

Since $V^{-1}([0, V(0)+c\,t_{\max}])$ is compact, this inequality implies that $z(t)$ remains bounded for all $t \in J$. (This does not yet establish that $J = [0, +\infty)$, since our state-space is $\mathcal{X}$, not $\mathbb{R}^n$; trajectories could conceivably approach the boundary of the positive orthant for some coordinates in Ind.)

But, by Lemma 6.2.2, there exists another compact set $\mathcal{C} \subset \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ so that $z(t) \in \mathcal{C}$ for all $t \in J$. Since $\mathcal{C}$ is a compact subset of the domain of definition, $\mathcal{X}$, and since $f^*(z, u)$ (when seen as an explicit function of time through $u = u(t)$), is defined for all $t \in [0, +\infty)$, it follows by standard ODE arguments (e.g., Proposition C.3.6 in [46]) that the maximal interval is $J = [0, +\infty)$, contradicting the existence of a finite $t_{\max}$. This proves $\mathbb{R}^n_{>0}$-completeness with input-value set $\mathbb{U}_\kappa$.

Step 3. Finally, notice that the function (3.12) is in fact a semi-global ISS-Lyapunov function with respect to the point $\bar{x}$ and the input $H(\bar{x})$ for the system (6.2), because all three properties of Definition 3.3.3 are satisfied ((i) and (ii) we have previously shown and (iii) follows immediately from Lemma 6.2.1).

It follows from Lemma 3.3.6 that (6.2) is semi-global ISS with input-value set $\mathbb{U}_\kappa$ (with respect to the point $\bar{x}$ and the input $H(\bar{x})$). ∎

Remark 6.2.4 Proposition 6.2.3 showed that the given system is semi-global ISS with respect to bounded inputs, for each possible bound on inputs. It would be more interesting (although, not needed for our purposes) to know if the same system is ISS with respect to arbitrary inputs. That is, it might be possible to pick the comparison functions in the ISS definition in a manner independent of the bound κ. It is remarked in [3] that "input semiglobal" ISS (ISS with respect to bounded inputs) indeed implies ISS. However, the proof given in that paper is for the more standard notion of ISS, for systems evolving in $\mathbb{R}^n$, while we have generalized the notion to deal with systems with positive states. Thus the proof does not apply.

6.3 Proof of Theorem 6

Pick any initial states $x(0) \in \mathbb{R}^n_{>0}$ and $z(0) \in \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ of the original system (3.1) 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 $t \to +\infty$.

The rest of the argument is completely analogous to what was done for the other observer. Since we know that $x(t)$ is defined for all $t \geq 0$ and converges to some equilibrium $\bar{x}$ as $t \to +\infty$, we must prove that $z(t)$ is also defined for all $t \geq 0$ and converges to this same $\bar{x}$ as $t \to +\infty$.

Since $x(t)$ converges to $\bar{x}$, so does $H(x(t))$ converge to $H(\bar{x})$. Let κ be a constant such that $|H(x(t))| \leq \kappa$ for all t. Let $\mathbb{U}_\kappa$ be the set of vectors u with norm less than or equal to κ, so that $H(x(t)) \in \mathbb{U}_\kappa$ for every $t \geq 0$.

Pick $T \geq 0$ so large that the convergence $x(t) \to \bar{x}$ becomes exponential (such T exists, as shown in [48]). 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 $x(t)$.

Proposition 6.2.3 shows that, whenever $z(0) \in \mathbb{R}^n_{>0}$, the solution $z(t)$ exists for all $t \geq 0$ and satisfies $z(t) \in \mathbb{R}^n_{>0}$.

Claim 1. There exists a constant $d > 0$ such that for all $z(0) \in \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ 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 6.2.1 (where we may pick $\gamma(r) = r^2/2$) and the discussion above imply

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

with

$$c_4 = \frac{1}{2}(c_0 c_1 |x(T) - \bar{x}|)^2,$$

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 $z(0) \in \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$, 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.

At this point we can prove that the observer system is $\mathbb{R}^n_{\geq 0}$-complete, so that for each $z(0) \in \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ the maximal interval of definition of the solution is $J = [0, +\infty)$. To prove $\mathbb{R}^n_{\geq 0}$-completeness, suppose, to get a contradiction, that $J = [0, t_{\max})$ with $t_{\max} < +\infty$. Then claim 1 says that $z(t) \in V^{-1}([0, V(z(T)) + d])$ for all $t \in J$, and this is a compact subset of $\mathbb{R}^n_{\geq 0}$ by properness of V. Now, by Lemma 6.2.2, there exists another compact set $\mathcal{C} \subset \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$ such that $z(t) \in \mathcal{C}$ for all $t \in J$. Since $\mathcal{C}$ is a compact subset of the state space, this contradicts $t_{\max} < +\infty$.

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 3.3.2.

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 6.2.3, immediately gives the ISS estimate with those functions β, φ.

Next take the more general case $z(0) \in \mathcal{X} \cap \mathbb{R}^n_{\geq 0}$. 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

$$\dot{v} = f^*(v, u), \quad v(0) = \xi_k,$$

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 6.2.3 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 convergence is established in a routine ISS fashion: given any $\varepsilon > 0$, pick $T_1 \geq T$ such that

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

where

$$\|H(x) - H(\bar{x})\|_{T_1} = \text{ess. sup}.\{|H(x(t)) - H(\bar{x})| : t \geq T_1\}.$$

Next, pick $T_2 \geq T_1$ such that

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

Then, for all $t \geq T_2$, $|z(t) - \bar{x}| \leq \varepsilon$. Therefore, $z(t) \to \bar{x}$ as $t \to +\infty$ as wanted.

6.4 Robustness with Respect to Disturbances

In Sections 4.2 and 6.1, we have considered the general models with inputs

$$\dot{z} = f(z) + C'(u - h(z)) \tag{6.4}$$

and

$$\dot{z} = f(z) + C'(u - H(z)) \tag{6.5}$$

which (besides providing us with observers when $u(t) \equiv h(x(t))$ or $u(t) \equiv H(x(t))$) may be useful in questions regarding the control of the class of chemical reactions described in the Introduction.

We have shown in Proposition 6.2.3 that, as long as u remains in a set $\mathbb{U}_\kappa$ for all times, trajectories of (6.5) are well defined for all t and satisfy an input-to-state stability condition.

The system (6.4) also exhibits an ISS stability property, but only with respect to a more restricted set $\mathbb{U}$, namely, $\|u\|$ has to be uniformly bounded by a fixed constant depending on $\bar{x}$ which, in particular, requires every coordinate of u to be nonnegative.

In fact, if it happens that $u(t)$ becomes less than an (arbitrary) amount $-\varepsilon$ on some interval of time (in the case of a systematic negative disturbance or some negative offset in a control, for instance due to leakage outflows in a chemical reactor), then solutions to (6.4) may blow up in finite time. This problem can be seen in the following simple example. Take the network $X_1 + X_2 \rightleftharpoons X_3$ with $\mathcal{D} = \text{span}\{(1,1,-1)'\}$ and rate constants equal to 1, for simplicity. Let

$$C = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and consider an input $u(t) = (u_1(t), u_2(t))'$ with $2(u_1(t) + u_2(t)) < -\varepsilon$ for some $\varepsilon > 0$ and all t. Then (6.4) becomes

$$\begin{aligned} \dot{x}_1 &= -x_1x_2 + x_3 + 2(u_1 - x_1^2) \\ \dot{x}_2 &= -x_1x_2 + x_3 \\ \dot{x}_3 &= x_1x_2 - x_3 + 2(u_2 - x_3^2). \end{aligned}$$

Now look at the variable $x_1 + x_3$:

$$\dot{x}_1 + \dot{x}_3 = 2(u_1 + u_2) - (x_1^2 + x_3^2) \leq -\varepsilon - \frac{1}{2}(x_1 + x_3)^2.$$

Viewing this as a differential inequality, we may conclude that $(x_1 + x_3)(t) \leq w(t)$ where $w(t)$ is the solution to $\dot{w} = -\varepsilon - \frac{1}{2}w^2$, $w(0) = (x_1 + x_3)(0)$, i.e.,

$$(x_1 + x_3)(t) \leq -\sqrt{\varepsilon} \tan\left(\frac{\sqrt{\varepsilon}}{2}t - a_0\right)$$

where

$$a_0 = \arctan\left(\frac{(x_1 + x_3)(0)}{\sqrt{\varepsilon}}\right) > 0.$$

Therefore, for a suitably chosen initial condition, $w(t)$ blows up to $-\infty$ as t tends to $\frac{\sqrt{\varepsilon}}{2}(a_0 + \frac{\pi}{2})$. Hence, since $(x_1 + x_3)(t) \leq w(t)$, $x_1 + x_3$ also has to blow up to $-\infty$ in finite time, and this can happen even if all states started positive.

In contrast, model (6.5), allows each coordinate of u to have negative values, for an arbitrarily large interval of time, and only requires that $|u(t)| < \kappa$ for all t, for some (fixed but arbitratry) positive constant, κ.

6.5 Robustness with Respect to Parameters

With arguments very similar to those used in Section 5.5, one can show that the alternative observer is also A_0-robust, and in a slightly stronger sense (see Remark 6.5.3, below).

Theorem 7 Let the matrix B be fixed. For each $A_0 \in \mathcal{A}_{\geq 0}$, let C be such that the system $\Sigma(A_0)$: $\dot{x} = f_{A_0}(x)$, $y = h(x)$ is detectable and $h(x)$ satisfies (3.9).

Then the system $\dot{z} = f_{A_0}(z) + C'(\rho(h(x)) - \rho(h(z)))$ is an A_0-robust observer for system $\Sigma(A_0)$.

As before, this Theorem follows from the ISS property of system (6.2) and the continuity of the map $\bar{x}(\cdot,\cdot)$. The latter doesn't depend on the observer itself, only on the original system, so all the results of Section 5.4 still hold. For the former, given any point $\bar{x} \in E_{A,+}$ one can show that (6.2) admits V as an ISS-Lyapunov function with respect to the point $\bar{x}$ and input $\rho(h(\bar{x}))$. Indeed, consider

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

where the last term can be written as (using first Cauchy-Schwartz, and then the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$ for positive numbers)

$$\begin{aligned} & \langle C(\rho(z) - \rho(\bar{x})), u - \rho(h(z))\rangle \\ = \ & \langle C(\rho(z) - \rho(\bar{x})), u - \rho(h(\bar{x}))\rangle - \langle C(\rho(z) - \rho(\bar{x})), \rho(h(z)) - \rho(h(\bar{x}))\rangle \\ \leq \ & \frac{1}{2}|C(\rho(z) - \rho(\bar{x}))|^2 + \frac{1}{2}|u - \rho(h(\bar{x}))|^2 - |C(\rho(z) - \rho(\bar{x}))|^2, \end{aligned}$$

so that

$$\nabla V(z,\bar{x})\ f_A^*(z,u) \leq -\vartheta(z,\bar{x}) + \Upsilon(z,u,\bar{x})$$

with

$$\vartheta(z,\bar{x}) = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}\, e^{\langle b_j, \rho(\bar{x})\rangle}\, e^{q_j}\, \omega(q_i - q_j) + \frac{1}{2}|C(\rho(z) - \rho(\bar{x}))|^2$$

and

$$\Upsilon(z,u,\bar{x}) = \frac{1}{2}|u - \rho(h(\bar{x}))|^2.$$

Since $\frac{1}{2}|C(\rho(z) - \rho(\bar{x}))|^2$ is certainly a $\mathcal{K}_\infty$ function of $|C(\rho(z) - \rho(\bar{x}))|$, Lemma 5.2.2 (with $\gamma(r) = 1/2\ r^2$) and Lemma 5.2.4 are still valid and so the main ISS property follows:

Lemma 6.5.1 Assume that the map h is such that (3.9) holds and C satisfies $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. Then there exist functions $\alpha, \gamma \in \mathcal{K}_\infty$ such that

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

for all $z \in \mathbb{R}^n_{>0}$ and all $u \in \mathbb{R}^p$. In particular, one can choose $\gamma(r) = \frac{1}{2}r^2$.

Given any $\kappa > 0$, define the subset of $\mathbb{R}^p$

$$\mathbb{U}_\kappa = \{u \in \mathbb{R}^p : |u| \leq \kappa\}.$$

From Lemma 6.2.2, it follows that the system $\dot{z} = f_A^*(z, u)$ is $\mathbb{R}^n_{>0}$-complete with input set $\mathbb{U}_\kappa$. As in Section 5.2, we may now state a similar result to Proposition 5.2.5:

Proposition 6.5.2 Assume that the map h is such that (3.9) holds, and the matrix C is such that $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n$. Then the system with inputs $\dot{z} = f_A^*(z, u)$ is ISS with input set $\mathbb{U}_\kappa$, with respect to the point $\bar{x}$ and the input $\rho(h(\bar{x}))$, i.e., there exist functions $\beta \in \mathcal{KL}$ and $\varphi \in \mathcal{K}_\infty$ such that

$$|z(t) - \bar{x}| \leq \beta(|z(0) - \bar{x}|, t) + \varphi(\|u - \rho(h(\bar{x}))\|) \tag{6.6}$$

for all $z \in \mathbb{R}^n_{\geq 0}$ and all $u \in \mathbb{U}_\kappa$.

Now the proof of Theorem 7 is very similar to that of Theorem 4. However, note that (6.2) is ISS with an input set $\mathbb{U}_\kappa$, for arbitrary κ, so we may set the number κ to be such that the signal $\rho[h(x(t))] \in \mathbb{U}_\kappa$, for all $t \geq 0$. Since $x(\cdot)$ evolves in the positive orthant and converges to a (positive) steady state, for all t, $x(t)$ belongs to some compact set $\mathcal{C} \subset \mathbb{R}^n_{>0}$, so one may take

$$\kappa := \max\{|\rho(h_i(x))| : x \in \mathcal{C},\ i = 1, \ldots, m\}.$$

Remark 6.5.3 Another advantage of this alternative observer is that, due to the fact that the ISS property holds with respect to a larger input set $\mathbb{U}_\kappa$, it is not necessary to ask for size restrictions on the compact set of equilibria, P. In other words, condition (e) of Definition 5.1.1, is not required in the proof of A_0-robustness for the alternative observer. In fact, the compact $P \subset \mathcal{E}$ can have an arbitrary large size (as well as $K \subset \mathcal{A}_{\geq 0}$), and the A_0-robustness estimate holds, provided the pair of sets P, K satisfy conditions (a)–(d) of Definition 5.1.1.

Chapter 7
Numerical Tests and Simulations

To numerically test the performance of our observers, we carried out some simulations to explore their responses in several situations. Some of the simulations illustrate the properties predicted by the ISS estimate such as robustness with respect to noisy measurements, the convergence of the observer under sampled outputs and the important robustness with respect to perturbations in the kinetic constants. The effect of weighting the outputs on the rapidity of convergence is also illustrated.

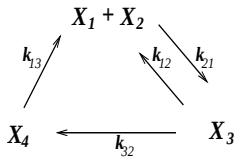
For chemical networks, other circumstances that are likely to occur may include the existence of unknown inputs acting in the system we wish to observe, or perturbations of a periodic nature in one or more kinetic constants. The response of the observer in these two cases is exemplified.

Finally, we compare the performance of our main observer with those of an extended Kalman filter and an extended Luenberger-type observer.

For these simulations, we will use essentially two different models, which will be fully described and detailed, as working examples.

In addition, a (theoretical) result is sketched, that deals with the stability of our systems in the case when boundary equilibrium points exist in positive classes. This result is applicable to our systems under certain general hypotheses.

7.1 The McKeithan T-cell Signal Transduction Model



As a first working example, we choose the T-cell signal transduction model ([38]) depicted above, and described in more detail in Chapter 2. As we have already seen in Examples 2.1.1 and 4.1.1, this system is characterized by

$$A = \begin{pmatrix} 0 & k_{12} & k_{13} \\ k_{21} & 0 & 0 \\ 0 & k_{32} & 0 \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The stoichiometric space and its orthogonal space are given by

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

$$\mathcal{D}^\perp = \text{span}\{(1,0,1,1),(0,1,1,1)\}.$$

Recall that the *positive classes* of the system are defined as

$$\mathcal{S} = (x_0 + \mathcal{D}) \cap \mathbb{R}^n_{\geq 0}, \quad \text{with} \quad \mathcal{S} \cap \mathbb{R}^n_{>0} \neq \emptyset$$

for $x_0 \in \mathbb{R}^n_{\geq 0}$, so in this example, the positive classes $\mathcal{S}_\alpha$ are characterized by a 2-tuple of *positive constants* $\alpha = (\alpha_1, \alpha_2)$, as

$$x_1 + x_3 + x_4 = \alpha_1, \quad x_2 + x_3 + x_4 = \alpha_2.$$

The set of boundary equilibria of the system is of the form (as determined by Proposition 2.2.1:

$$E_0 = \{(r_1,0,0,0)', (0,r_2,0,0)' \in \mathbb{R}^4_{\geq 0} : r_1, r_2 \geq 0\}.$$

Observe that points of the form $(r_1,0,0,0)'$ imply $\alpha_2 = 0$ and points of the form $(0,r_2,0,0)'$ imply $\alpha_1 = 0$. In a class corresponding to $\alpha_2 = 0$ (respectively, $\alpha_1 = 0$), it follows that $x_2 \equiv x_3 \equiv x_4 \equiv 0$ (respectively, $x_1 \equiv x_3 \equiv x_4 \equiv 0$), so the class is not positive, since it does not intersect the positive orthant. Thus it is clear that no boundary equilibria exist in any positive class.

Theorem 1 applies, so we know that, for any $x_0 \in \mathbb{R}^n_{\geq 0}$, the trajectory of the system $\dot{x} = f(x)$, $x(0) = x_0$, converges to a point $\bar{x} \in E_+$.

For the simulations in this Chapter, we took the output to be

$$h(x) = \begin{pmatrix} x_1 x_2^2 \\ x_1 x_4 \end{pmatrix},$$

as in Example 4.1.1, where

$$C = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

so that $\mathcal{D} + \text{im}\, C' = \mathbb{R}^4$. Note also that $h_2(E_0) = 0$, while $h(E_+) \in \mathbb{R}^2_{>0}$ so, by Theorem 2 and Corollary 3.2.4, the system with outputs is detectable.

Unless otherwise specified, the constants were taken to be:

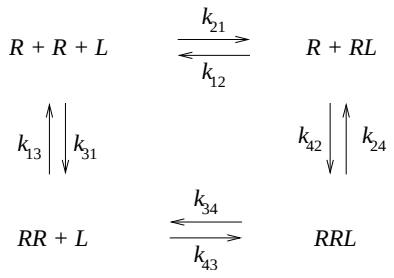
$$k_{21} = 6, \quad k_{12} = 0.5, \quad k_{13} = 7, \quad k_{32} = 1,$$

and the initial conditions for system (3.1) and for the observers were:

$$x(0) = (1,3,3,2)', \quad z(0) = (2,25,20,1)'.$$

7.2 A Receptor–Ligand Dimer Model

As a second working example, we choose a receptor–ligand model, in the style of [10]. In this model, as well as the usual receptor–ligand binding ($R + L \to RL$), the ligand is also allowed to bind to a "dimer" state of the receptor, which consists of a pair of linked receptors (RR).



For this system we have five species $x = (R, RR, L, RL, RRL)'$, and four complexes, defined by the matrix:

$$B = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The nonzero entries of the matrix A are as depicted in the figure.

The stoichiometric space and its orthogonal space are given by:

$$\begin{aligned} \mathcal{D} &= \text{ span } \{(1,1,0,-1,0)', (2,1,0,0,-1)', (2,0,-1,0,0)'\}, \\ \mathcal{D}^{\perp} &= \text{ span } \{(0,0,1,1,1)', (1,2,0,1,2)'\}. \end{aligned}$$

so the positive classes $\mathcal{S}_\alpha$ are characterized by a 2-tuple of positive constants $\alpha = (\alpha_1, \alpha_2)$ as follows:

$$[L] + [RL] + [RRL] = \alpha_1, \quad [R] + 2[RR] + [RL] + 2[RRL] = \alpha_2.$$

The boundary equilibria (see Proposition 2.2.1) are determined by

$$[R]^2[L] = 0, \quad [R][RL] = 0, \quad [RRL] = 0, \quad [RR][L] = 0$$

and so

$$E_0 = \{(0,0,r_3,r_4,0)', (0,r_2,0,r_4,0)', (r_1,r_2,0,0,0)' : \ r_1, r_2, r_3, r_4 \in [0,+\infty)\}.$$

We note that in this example, exactly one boundary equilibrium exists in each positive class. Indeed,

1. a point $(0,0,r_3,r_4,0)'$ implies $r_4 = \alpha_2$ and $r_3 + r_4 = \alpha_1$, so, classes characterized by two positive scalars $0 < \alpha_2 \leq \alpha_1$, contain the boundary equilibrium $(0,0,\alpha_1 - \alpha_2, \alpha_2, 0)'$;

2. a point $(0,r_2,0,r_4,0)'$ implies $r_4 = \alpha_1$ and $2r_2 + r_4 = \alpha_2$, so, classes characterized by two positive scalars $0 < \alpha_1 \leq \alpha_2$, contain the boundary equilibrium $(0,(\alpha_2 - \alpha_1)/2, 0, \alpha_1, 0)'$;

3. a point $(r_1, r_2, 0, 0, 0)'$ implies $\alpha_1 = 0$, but, in classes characterized by $\alpha_1 = 0$, it follows that $L \equiv 0$, $[RL] \equiv 0$ and $[RRL] \equiv 0$, so the intersection of the class with the positive orthant is empty, hence it is not a positive class. No positive class contains an equilibrium of this form;

4. finally, note that, for classes defined by $\alpha_1 = \alpha_2 > 0$, the equilibria in 1. and 2. coincide.

Even though Theorem 1 doesn't apply to such an example, and despite the existence of boundary equilibria, we will still be able to conclude that any trajectory of the system (corresponding to an initial condition not in E_0 in the interior of a class) converges to some $\bar{x} \in \mathbb{R}^n_{>0}$. This is done in Section 7.2.1, below.

Under these conditions we may still apply our observer, since its construction and proof of convergence use only the fact that the system's trajectory converges to a positive equilibrium point, when starting from an interior initial condition.

In all simulations relative to the dimer model, we have used the output

$$h(x) = \begin{pmatrix} [R]^2[L] \\ [RR] \end{pmatrix},$$

where

$$C = \begin{pmatrix} 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

so it is clear that $\mathcal{D} + \operatorname{im} C' = \mathbb{R}^5$, and note that $h_1(E_0) = 0$ (while $h_1(E_+) > 0$), so the system with outputs is detectable.

Unless otherwise specified the matrix of reaction rate constants was

$$A = \begin{pmatrix} 0 & 0.5 & 1 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 1 \\ 0 & 1.5 & 0.5 & 0 \end{pmatrix}, \tag{7.1}$$

and the initial conditions for the system were taken to be

$$[R](0) = 6,\ [RR](0) = 0.1,\ [L](0) = 11,\ [RL](0) = 0.1 \ \text{ and } \ [RRL](0) = 0.1, \tag{7.2}$$

while for the observer (unless otherwise indicated) all the concentrations were set initially to 1.

In all the Figures related to the dimer model, the `error` represents the vector norm $|x(t) - z(t)|$, where $x(t)$ is the solution of the nominal system and $z(t)$ is the solution of the observer at time t.

7.2.1 Instability of Boundary Equilibria

Consider a general Feinberg-Horn-Jackson chemical network, $\dot{x} = f(x)$, with a boundary equilibria set E_0, and positive equilibria set E_+, suppose that the intersection of some

positive class $\mathcal{S}$ with E_0 is nonempty. In this case, under some hypotheses (which are satisfied, for instance, by the dimer model), we may show that trajectories starting in the interior of $\mathcal{S}$, converge to the unique $\bar{x} \in \mathcal{S} \cap E_+$, by looking at the stable/unstable manifolds at a point $p \in \mathcal{S} \cap E_0$. We sketch this proof next.

Define

$$\mathcal{S}_0 := \mathcal{S} \cap E_0$$

and for any class $\mathcal{S}$, let

$$f_{\text{red}} : \mathbb{R}^{n-m+1}_{\geq 0} \to \mathbb{R}^{n-m+1}$$

be the reduced order vector field (which may be obtained by eliminating $m-1$ variables from the system $\dot{x} = f(x)$, by using the class equations). For any $\bar{z} \in \mathcal{S}_0$, let

$$\begin{aligned} W^s_{\bar{z}} &:= \text{ stable manifold at } \bar{z}, \\ L^s_{\bar{z}} &:= \text{ stable subspace of the linearization of } f_{\text{red}} \text{ at } \bar{z}, \end{aligned}$$

and consider the following standard notion of *tangent cone* to a subset A of an Euclidean space, at $\bar{z} \in A$:

$$T_{\bar{z}}A := \left\{ \lim_{i \to +\infty} \frac{1}{t_i} (p_i - \bar{z}) : \ p_i \in A \text{ and } p_i \to \bar{z} \text{ as } t_i \searrow 0 \right\}.$$

Define

$$\begin{aligned} I_{\bar{z}} &:= \{i \in \{1, \ldots, n\} : \ \bar{z}_i = 0\}; \\ \mathbb{R}^n_{\bar{z}} &:= \{x \in \mathbb{R}^n : \ x_i \geq 0, \ \forall\, i \in I_{\bar{z}}\}. \end{aligned}$$

Consider the following hypotheses:

(H1) $\mathcal{S}_0$ is a discrete set;

(H2) for all $\bar{z} \in \mathcal{S}_0$, $L^s_{\bar{z}} \bigcap \mathbb{R}^n_{\bar{z}} \bigcap \mathcal{S} = \{\bar{z}\}$;

(H3) the equilibria $\bar{z} \in \mathcal{S}_0$ are all hyperbolic with respect to the class $\mathcal{S}$ (i.e., the linearization of f_{red} at $\bar{z}$ has eigenvalues with *real part different from zero*).

Proposition 7.2.1 If the hypotheses (H1), (H2) and (H3) hold for the system $\dot{x} = f(x)$, then every trajectory starting in the interior of $\mathcal{S}$ converges to the unique positive equilibrium $\bar{x}_{\mathcal{S}}$.

Proof. Given any $x_0 \in \mathcal{S} \cap \mathbb{R}^n_{>0}$, let $x(t)$ be the solution of the system $\dot{x} = f(x)$, with $x(0) = x_0$. Since $x(t) \to \mathcal{S}_0 \bigcup \{\bar{x}_{\mathcal{S}}\}$, and this set is discrete, we know that $x(\cdot)$ converges to an equilibrium. So we only need to show that $x(t) \not\to \bar{z}$ for all $\bar{z} \in \mathcal{S}_0$. To get a contradiction, assume that $x(t) \to \bar{z}$ for some $\bar{z} \in \mathcal{S}_0$. By definition of the stable manifold $W^s_{\bar{z}}$, this means that

$$x(t) \in W^s_{\bar{z}} \ \text{ for all } \ t \geq 0.$$

On the other hand, also

$$x(t) \in \mathcal{S} \quad \text{for all} \quad t \geq 0.$$

In general, assume that there is a sequence of points $\{x^{(k)}\}$, contained in a closed set A and another point $\bar{z} \in A$ such that $x^{(k)} \to \bar{z}$, but $x^{(k)} \neq \bar{z}$ for all k. Let $\nu > 0$ be any constant with $\nu \neq |\bar{z}|$. Then let

$$v^{(k)} := \frac{1}{r_k}\left(x^{(k)} - \bar{z}\right)$$

for all k, with $r_k := |x^{(k)} - \bar{z}|$. Since each $v^{(k)}$ has norm ν (i.e., the sequence $v^{(k)}$ is contained in the compact set $\{x \in \mathbb{R}^n : |x| = \nu\}$), we can assume, taking a subsequence if necessary, that

$$v^{(k)} \to v$$

for some v (also of norm ν). Now, because $x^{(k)} \in A$ and $r_k \to 0$, by definition, v belongs to $T_{\bar{z}}A$, the tangent cone to A at the point $\bar{z}$.

In our case, letting

$$x^{(k)} = x(t_k), \quad \text{for any sequence} \quad t_k \nearrow +\infty,$$

and in turn taking $A \equiv \mathcal{S}$ and $A \equiv W^s_{\bar{z}}$, we conclude that there is some nonzero vector v satisfying

$$v \in T_{\bar{z}}W^s_{\bar{z}} \bigcap T_{\bar{z}}\mathcal{S}.$$

Now, note that

$$T_{\bar{z}}W^s_{\bar{z}} = L^s_{\bar{z}} \quad \text{and} \quad T_{\bar{z}}\mathcal{S} = T_{\bar{z}}\mathbb{R}^n_{\geq 0} \bigcap \mathcal{S} = \mathbb{R}^n_{\bar{z}} \bigcap \mathcal{S}.$$

So, $v \in L^s_{\bar{z}} \bigcap \mathbb{R}^n_{\bar{z}} \bigcap \mathcal{S}$, which is $\{\bar{z}\}$ by hypothesis (H2). This leads to a contradiction, since $|v| = \nu \neq \bar{z}$. ∎

We next verify that the dimer model described above does satisfy the three hypotheses of Proposition 7.2.1:

(H1) that each $\mathcal{S}_0$ is a discrete set follows from points 1-4 in Section 7.2. In fact, each $\mathcal{S}_0$ consists of one element only;

(H2) for the initial conditions (7.2), the class $\mathcal{S}$ is given by the 2-tuple $(11.2, 6.5)$, and the corresponding boundary equilibrium is $\bar{z} = (0, 0, 4.7, 6.5, 0)'$. Taking kinetic constants to be given by (7.1) (and eliminating $[RR]$ and $[RL]$ from the system), the linearization around the equilibrium point $\bar{z}$ of the reduced system, has eigenvalues and eigenvectors:

$$\begin{aligned}
&\lambda_1 = 0.9223, \quad w_1 = (0.1543, 0.8493, 0.5049)' \\
&\lambda_2 = -8.4861 + i\,3.8650, \quad w_2 = (-0.4559 + 0.3740i, 0.1275 - 0.1590i, 0.7815)' \\
&\lambda_3 = -8.4861 - i\,3.8650, \quad w_3 = (-0.4559 - 0.3740i, 0.1275 + 0.1590i, 0.7815)',
\end{aligned}$$

showing that the stable subspace $L^s_{\bar{z}} =$ span $\{w_2, w_3\}$ doesn't contain any positive vector so its intersection with $\mathcal{S}$ is $\{\bar{z}\}$;

(H3) since $\lambda_1 > 0$ and $\mathrm{Re}(\lambda_{2,3}) < 0$, the only boundary equilibrium in the class is hyperbolic.

7.3 Measurement Noise and Unknown Inputs

The numerical tests described in this Section all concern the T-cell signal transduction model. Both the main and alternative observer are simulated for this system. In Figure 7.2, the convergence of both observers for this example is shown, in the ideal situation with no perturbations.

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

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

and $\eta(t)$ is an $\mathbb{R}^2$-valued vector white noise. In view of the Section 4.5.1, solutions of this system exist for all $t \geq 0$, provided that $h(x(t)) + \eta(t)$ is nonnegative; and the observer should provide estimates close to the true state as long as the magnitude of $\eta(t)$ is small and its average over time is zero. Thus we chose $x(0)$ so that $x_i(t) \geq 2$ and $|\eta(t)| < 2$ for all t and all i.

With output noise the alternative observer becomes

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

so the noise appears inside the logarithmic function. But under the conditions above we know that this system has a solution for all t (by Proposition 6.2.3).

In Figure 7.3 we can see that the trajectories of the four coordinates of our two observers exhibit small magnitude perturbations as a result of the output noise. The alternative observer, although slower in the convergence, has smoothed out the noise effect.

In another simulation, the model (3.1) 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 is shown in Figure 7.1. 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 observers are still the ones for the nominal system, with no disturbance.) Note how the observers catch-up after the "delta" disturbance, and also track (with a small lag) the limit cycle into which the observed system trajectories converge (Figure 7.4).

7.4 Sampled and Weighted Outputs, Parameter–Robustness

The numerical tests described in this Section all concern the main observer applied to the receptor–ligand dimer model.

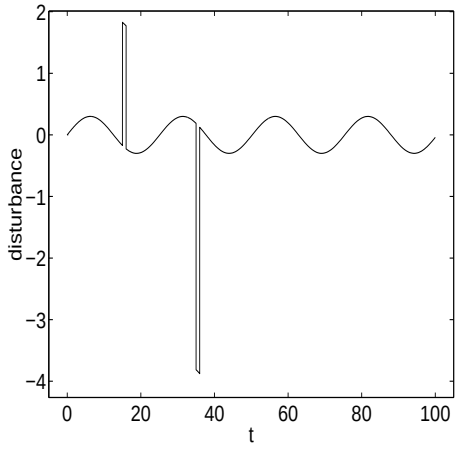


Figure 7.1: The disturbance added to the system.

According to Section 4.5.3, the outputs may be multiplied by a diagonal matrix. In Figure 7.7 the effect of different weights, of increasing magnitude, is to also increase the rapidity of convergence to correct values. However, numerical tests with other models, have also shown that simply increasing the weights doesn't always imply faster convergence (see Figure 8.7).

As shown in Section 4.5.2, the main observer is still convergent under sampling at regular time intervals. In this simulation the observer has the form

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

where

$$s(t) = h(x(t_i)) \quad \text{whenever} \quad t_i \leq t < t_{i+1}$$

and $t_i = i\Delta t$, $i = 0, 1, \ldots$, represent the times at which a new measurement is received.

Figures 7.8 and 7.9 provide an illustration of the sampled output described above, when the sampling interval was $\Delta t = 0.1$ and the corresponding response of the observers. As expected, the trajectories of the observers converge to the trajectory of the system. Each time a "fresh" measurement is received by the observers, these make a quick recovery towards the correct signal.

Following the robustness result developed in Chapter 5, we perturbed the kinetic constants of the nominal system while constructing the observer with the ideal parameters (7.1). Figure 7.10 shows the response of the observer under random perturbations of the parameters within $\pm 15\%$. The perturbed matrix A is

$$A = \begin{pmatrix} 0 & 0.4391 & 0.93770 & 0 \\ 0.4506 & 0 & 0 & 0.4350 \\ 0.4969 & 0 & 0 & 0.9753 \\ 0 & 1.7035 & 0.5381 & 0 \end{pmatrix}$$

and a verification shows that the corresponding linearization of the reduced system still satisfies hypotheses (H1)-(H3).

As can be seen in the Figure, weighting the outputs to increase the rapidity of convergence may give better results. The errors have stabilized at a value of ≈ 1.19 in the case when the outputs are not weighted, against a value of ≈ 0.54 in the case when the outputs are weighted by $W = \text{ diag } (5, 5)$.

Another situation is ilustrated in Figure 7.11. One of the kinetic constants was perturbed with a periodic pattern (unknown to the observer) according to

$$k_{42} = k_{42}(t) = 1.5 + 0.2 \cos(0.9t) + 0.1 \cos(0.2t).$$

The observer still provides a reasonable estimate (final error is of magnitude approximately 0.1), and we note that (at least for some coordinates) it appears to follow the oscillating pattern, while the estimate tends to be an average of the oscillating values. A perturbation with larger periods tends to be more easily tracked by the observer.

7.5 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 they perform adequately in specific examples, and hence their practical success. In 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.

7.5.1 Extended Kalman Filter

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 matrix, which is a 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.

To solve the Ricatti differential equation, we took $R = I_{p\times p}$ and $Q = I_{n\times n}$ (the identity matrices in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively), and the initial condition $P(0) = I_{n\times n}$.

7.5.2 A Luenberger-type Observer

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.)

For the T-cell signal transduction model, 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 $-0.48 \pm 0.08i$.

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 $\bar{x}$. Figure 7.5 shows the local convergence of the observers. Comparing the behavior of the three observers in Figure 7.6, shows that the performance of EKF and Luenberger are clearly inferior to that of our observer.

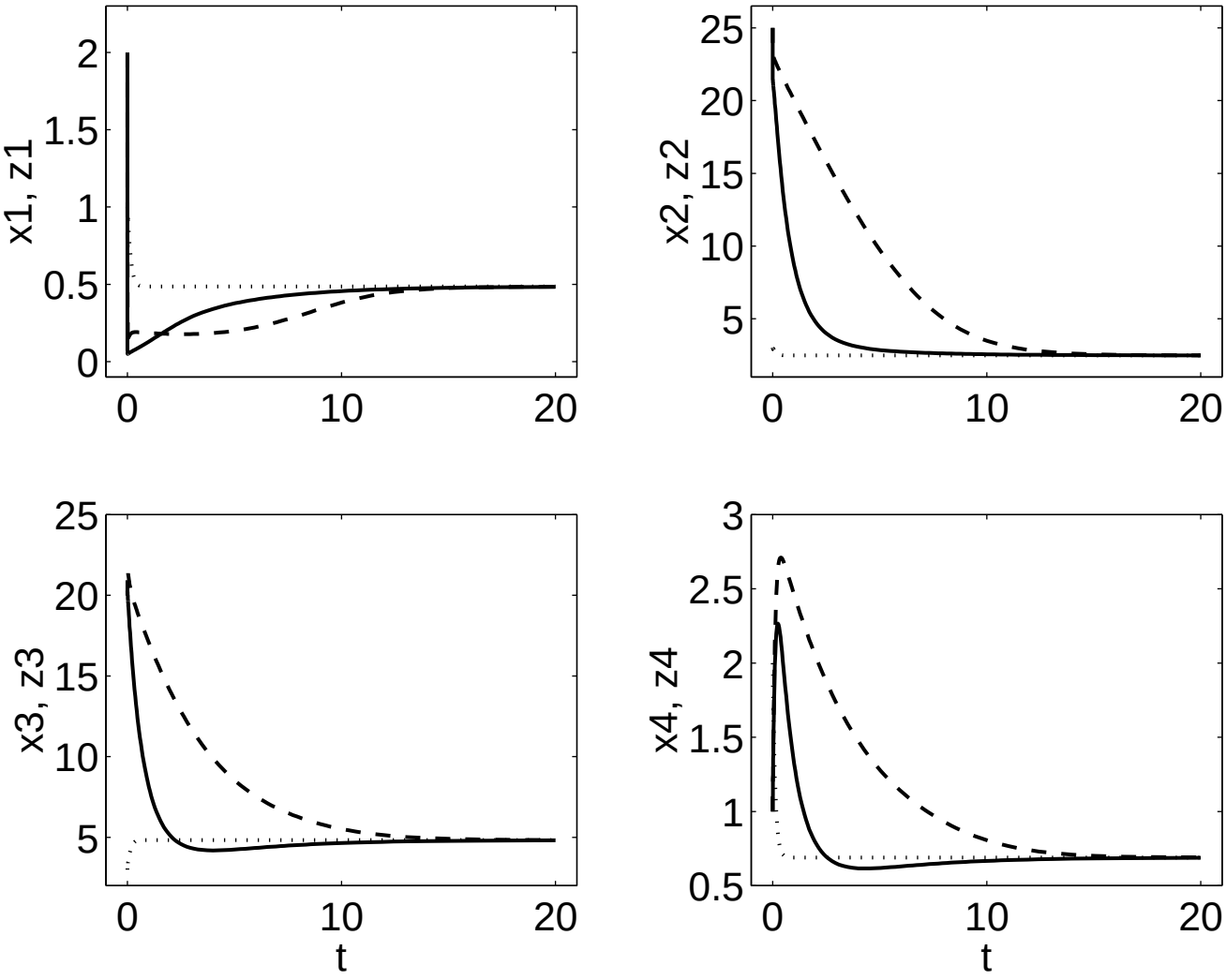


Figure 7.2: The trajectories of the system and observers without noise and without unknown inputs. The dotted line corresponds to the T-cell signal transduction model, the solid line corresponds to our main observer, and the dashed line to the alternative observer (where the logarithm of the output is used).

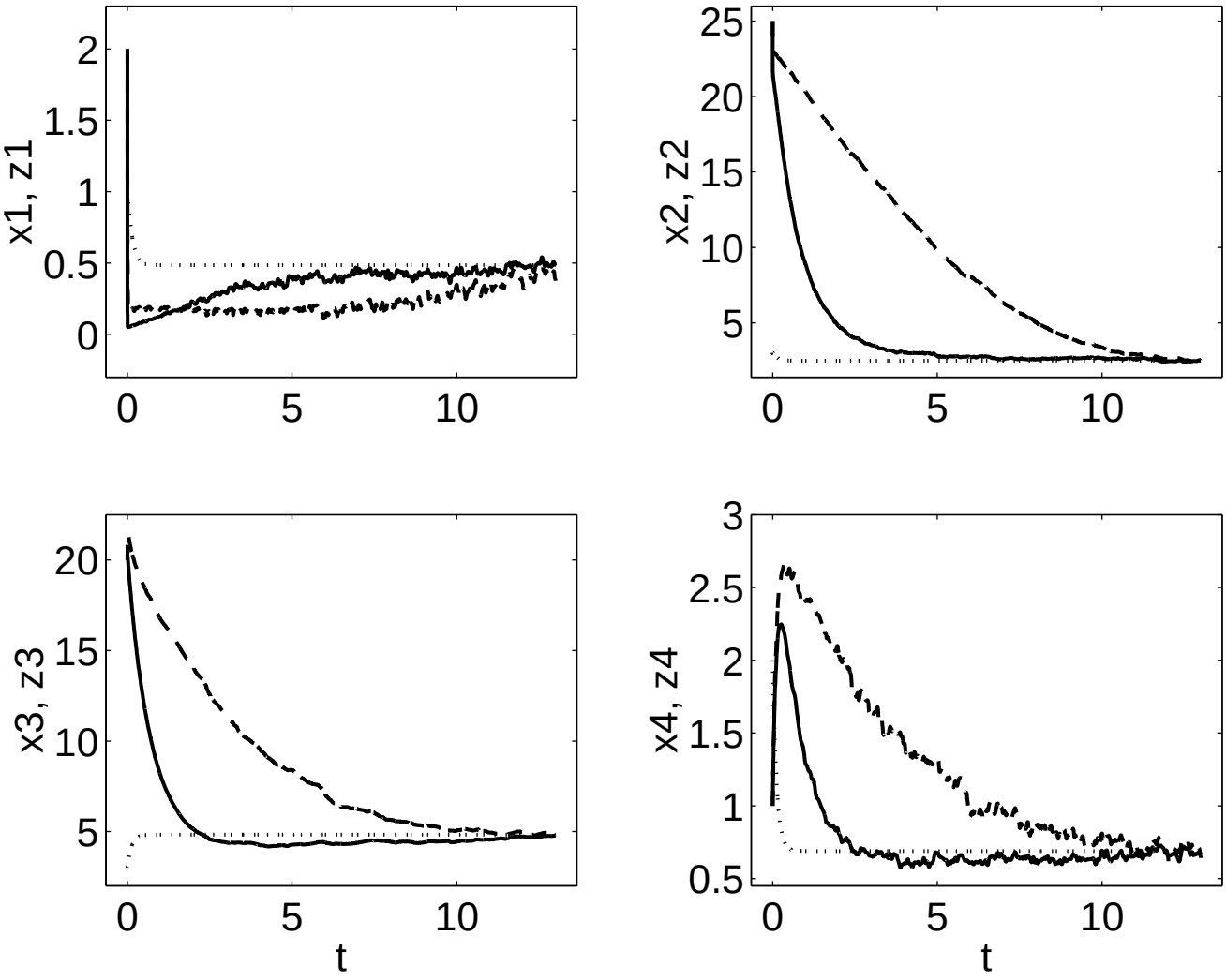


Figure 7.3: The trajectories of the system (dotted line) and our main observer (solid line) and alternative observer (dashed line) in the presence of observation noise.

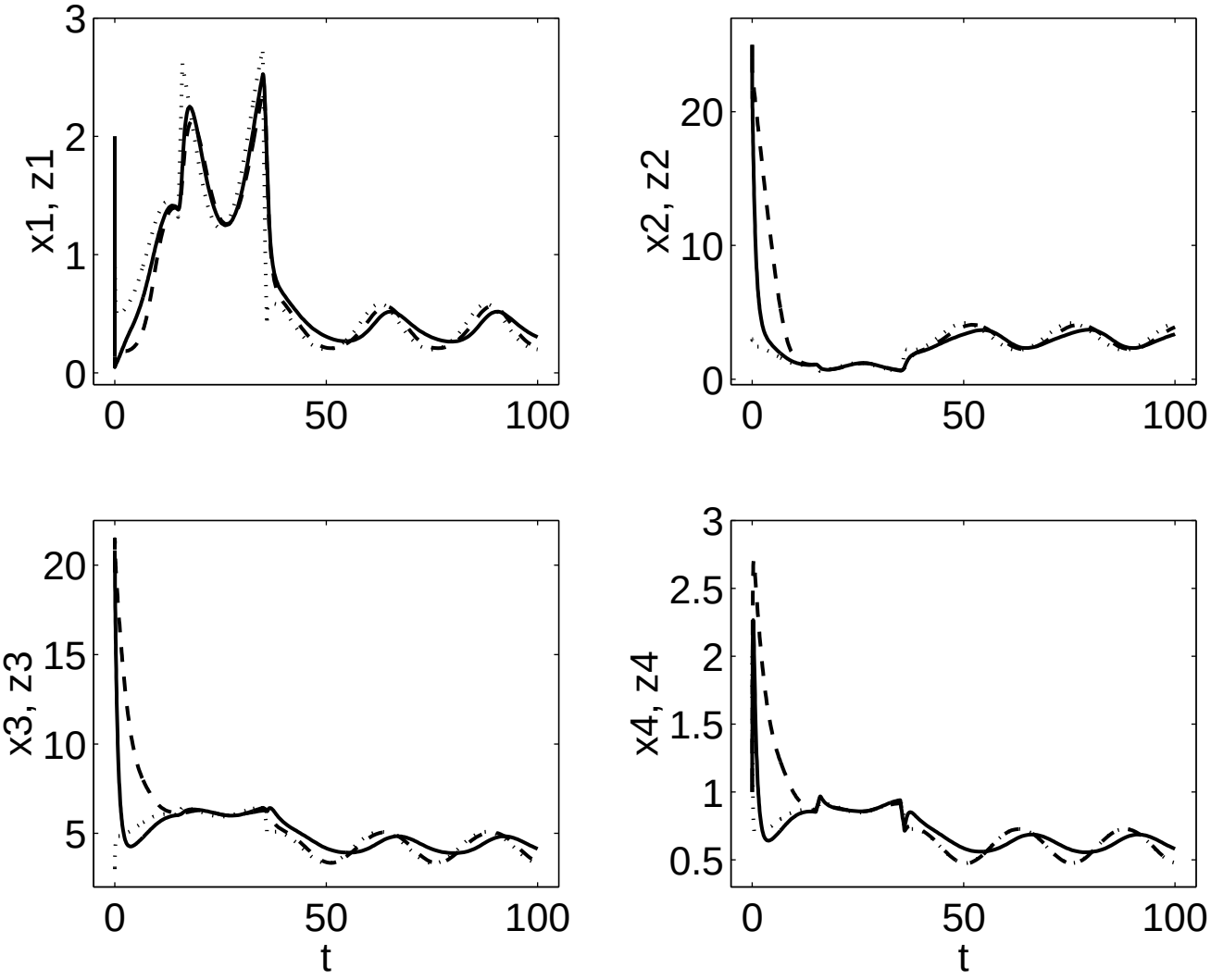


Figure 7.4: The effect of a disturbance on the coordinates of the system (dotted line), main observer (solid line) and alternative observer (dashed line).

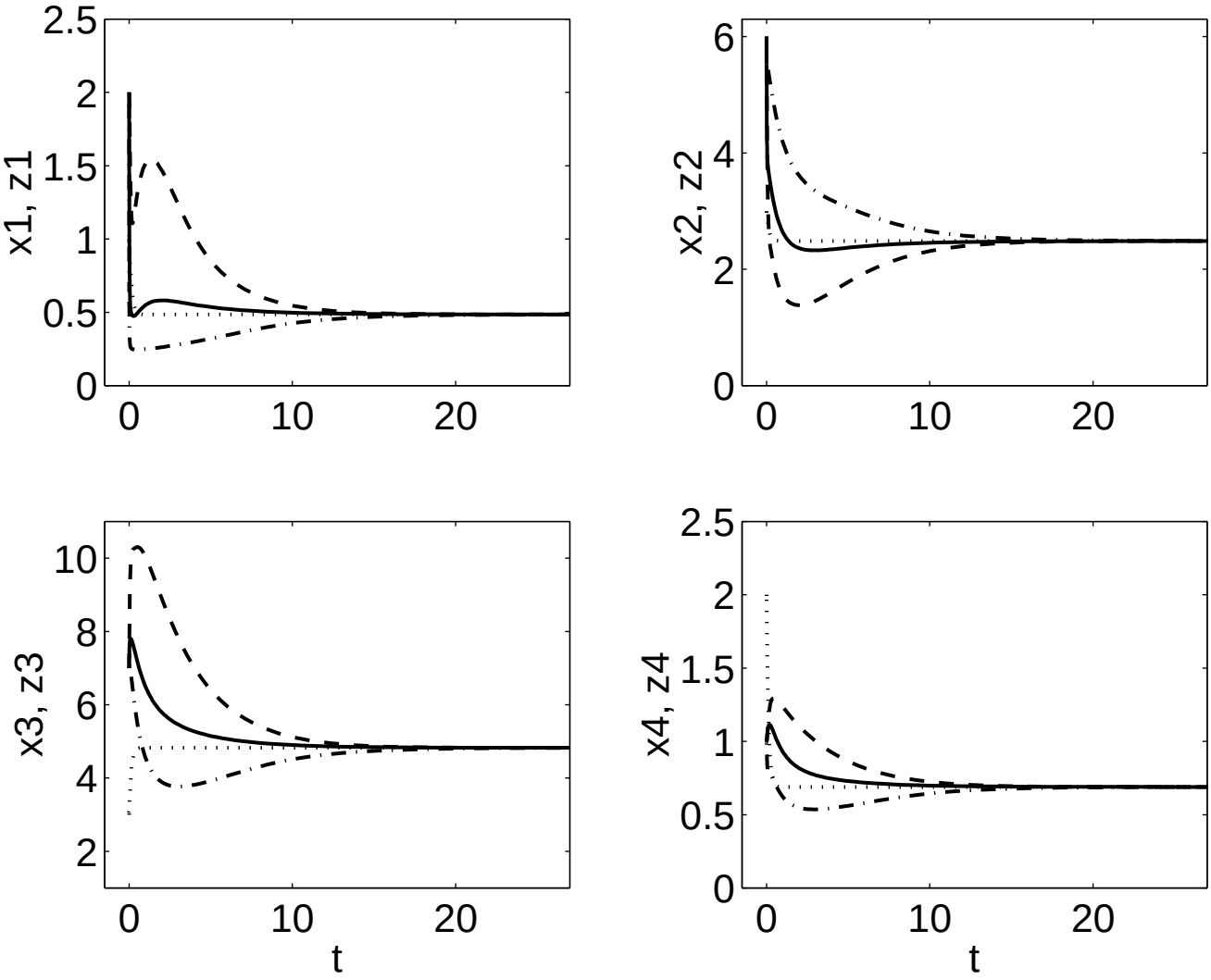


Figure 7.5: Comparison with standard observers. Local convergence with initial condition $z(0) = (2, 6, 7, 1)'$. The trajectories of the system (dotted line), our observer (solid line), a Lbg(dashed line) and an EKF (dash-dotted line) are shown against time.

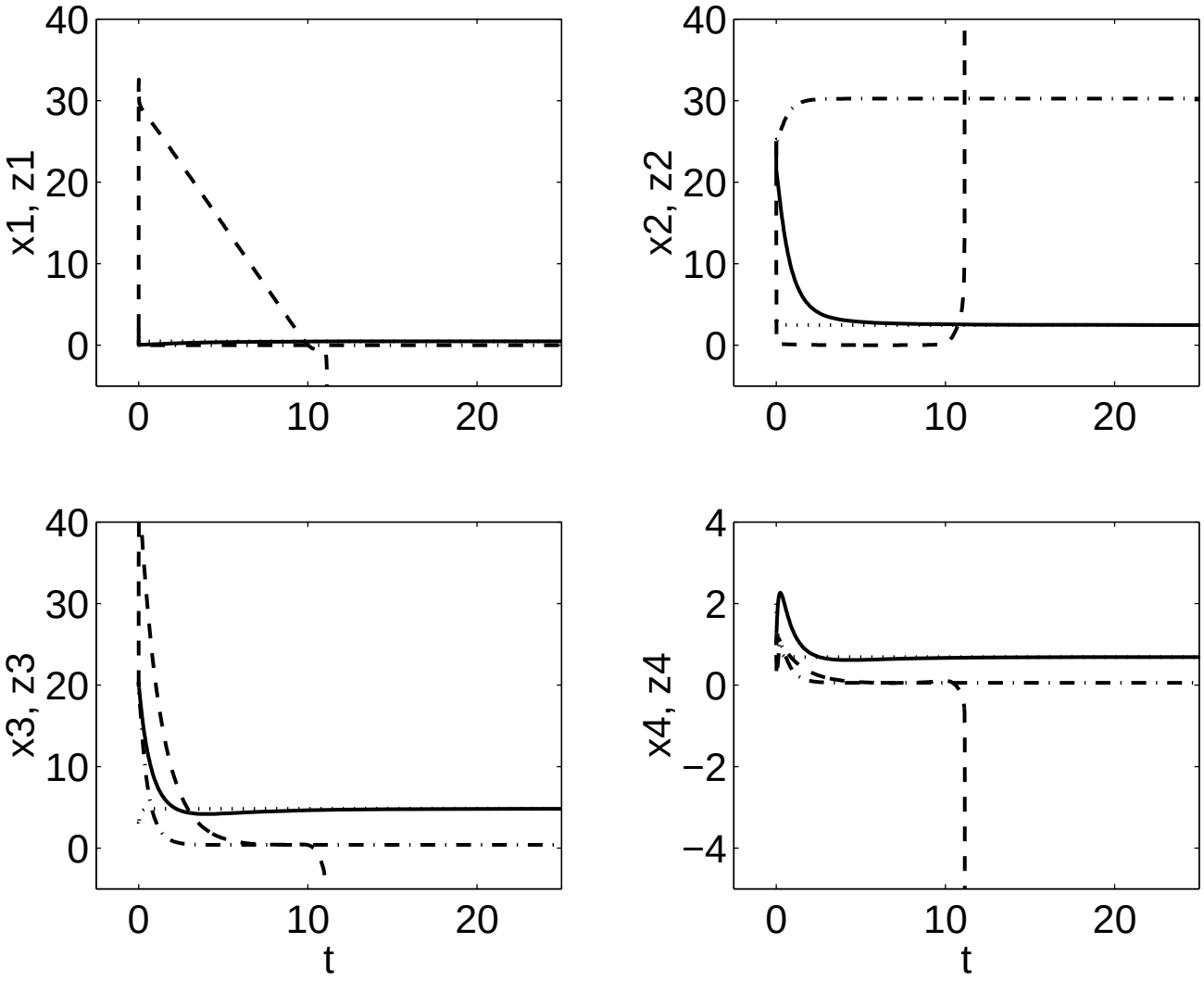


Figure 7.6: Comparison with standard observers. The trajectories of the system (dotted line), our observer (solid line), a Lbg(dashed line) and an EKF (dash-dotted line) are shown against time. Lbg and EKF diverge for an initial condition $z(0) = (2, 25, 20, 1)'$.

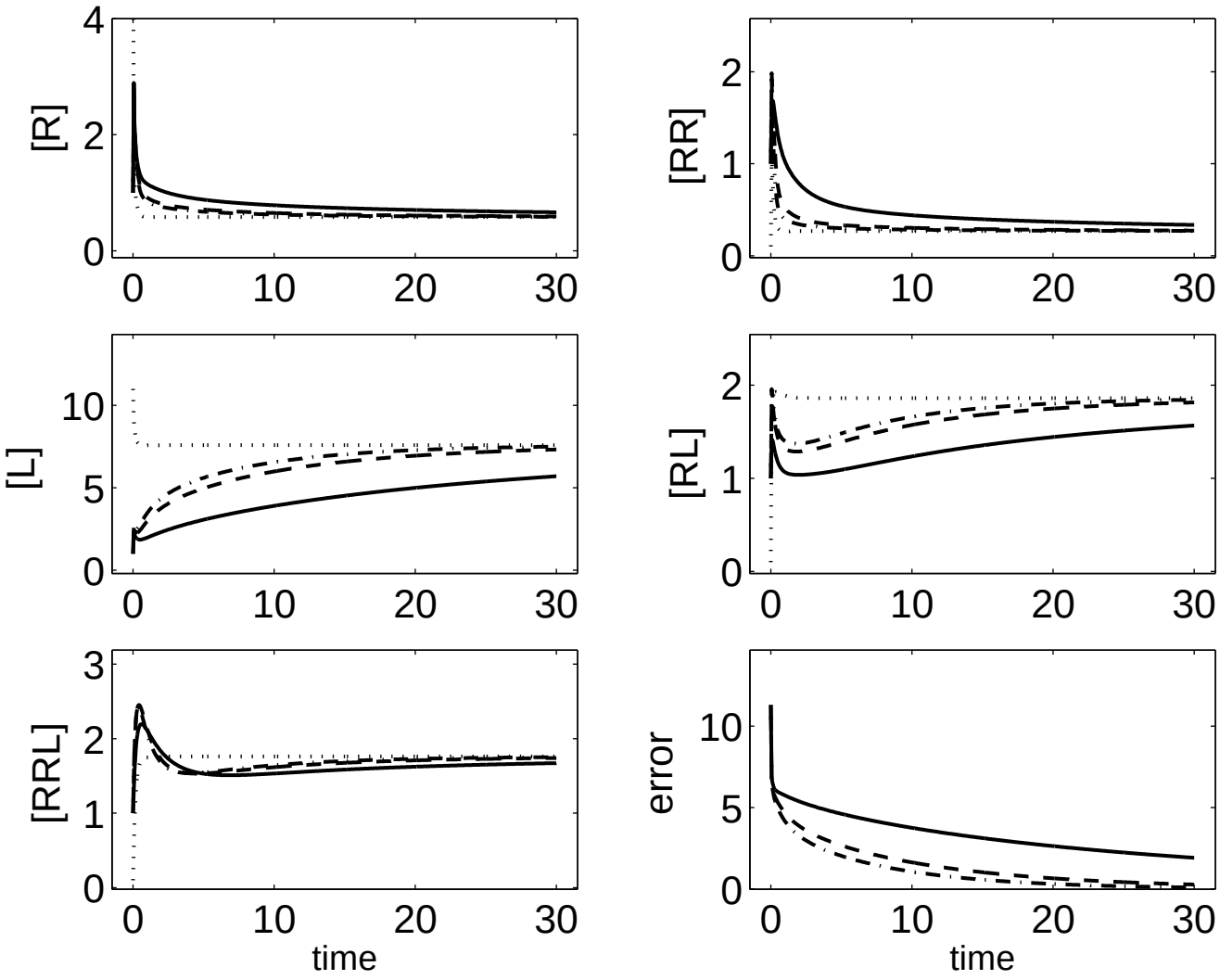


Figure 7.7: Effect of weighted outputs: (i) $W = \text{diag}\,(1,1)$ (solid line); (ii) $W = \text{diag}\,(5,5)$ (dashed line); (iii) $W = \text{diag}\,(10,10)$ (dash-dotted line). The dotted lines represent the trajectory of the receptor–ligand dimer model.

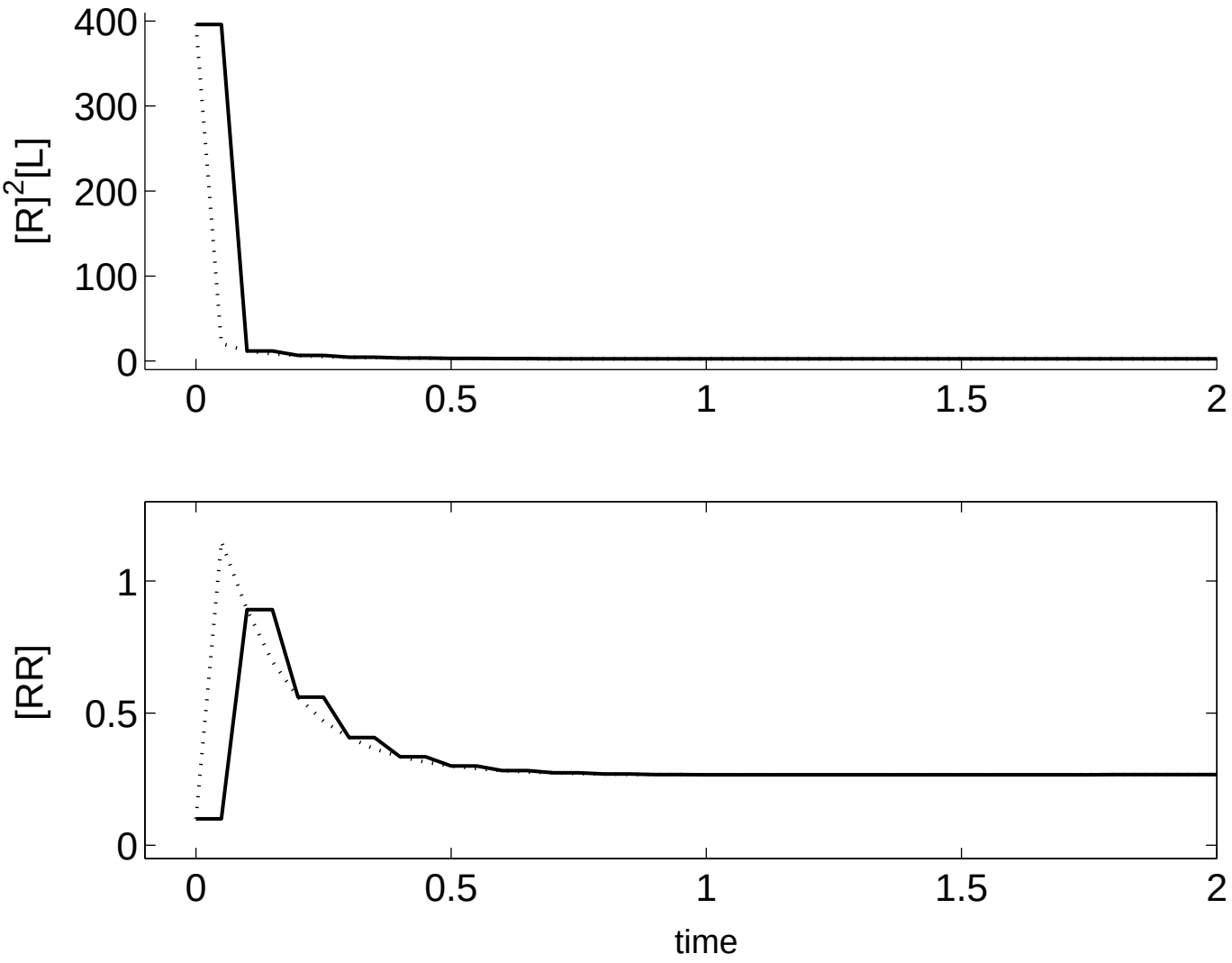


Figure 7.8: Sample & hold outputs (solid line). The sampling interval is $\Delta t = 0.1$.

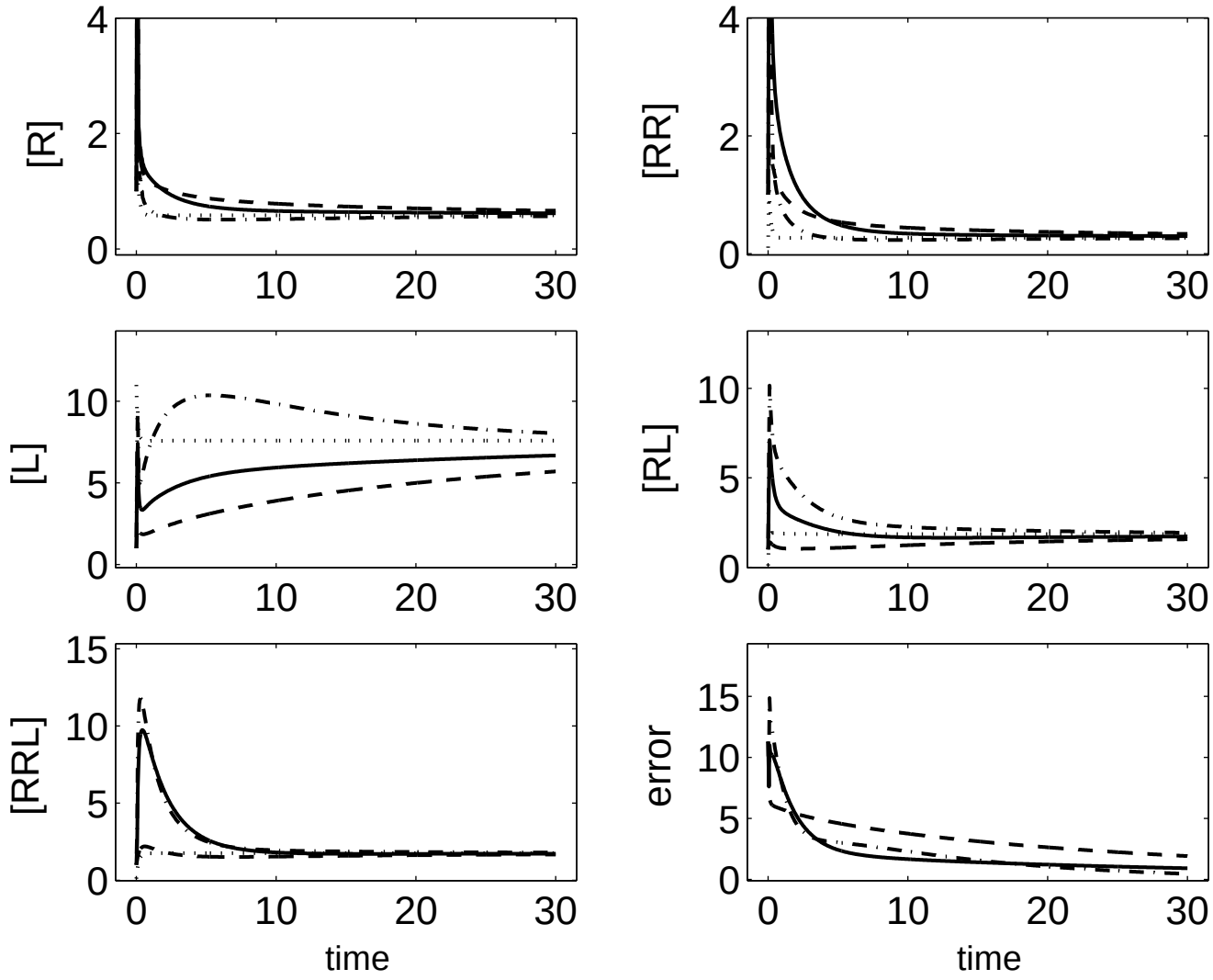


Figure 7.9: Convergence of the observer under sampled outputs. Shown are the trajectories of the system (dotted line), the trajectories of the observer when receiving continuous outputs (dashed line), the trajectories of the system when receiving the sample & hold output (solid line), and the effect of weighting ($W = \text{diag}\,(5,5)$) together with sampled & hold outputs (dash-dotted line).

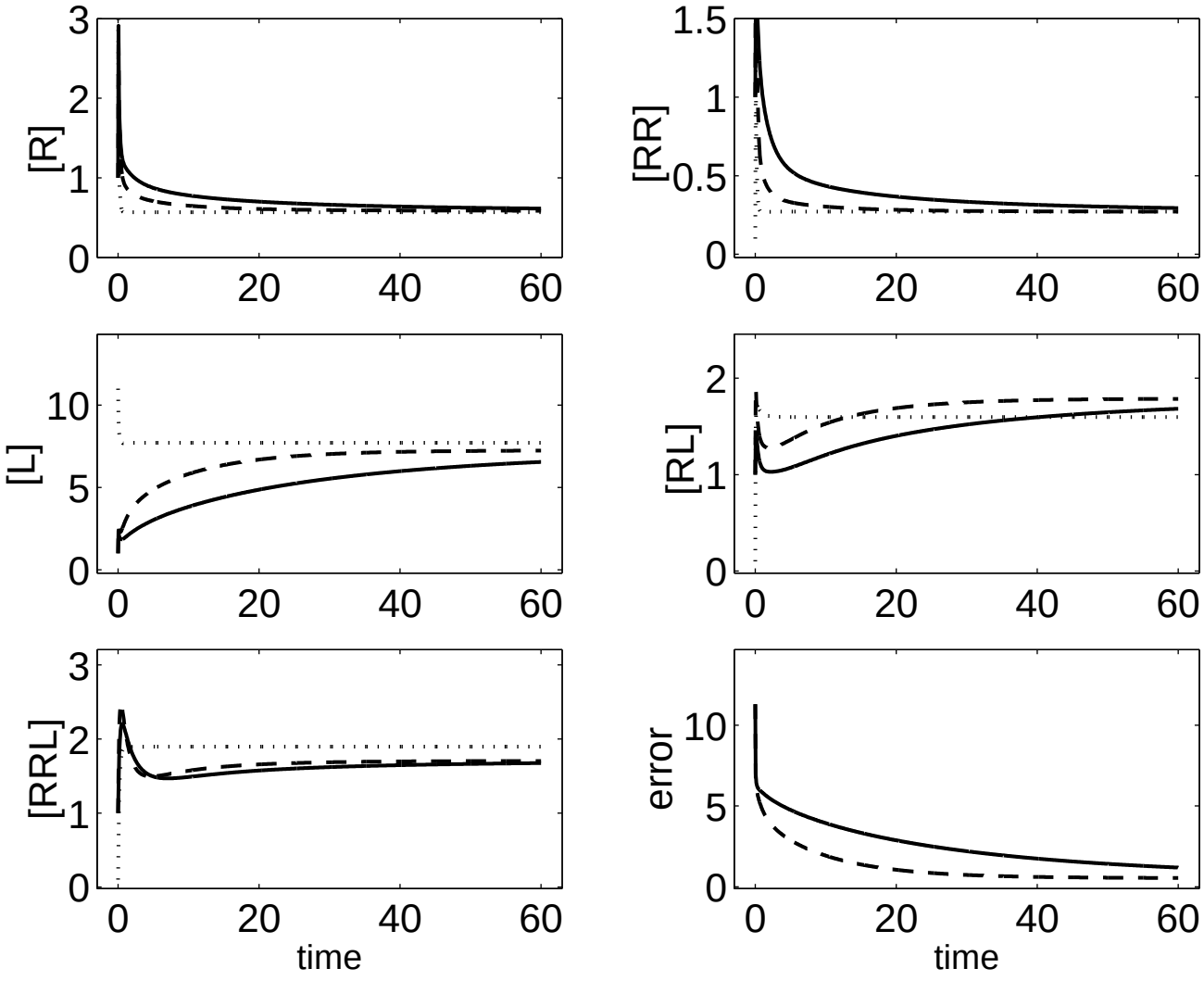


Figure 7.10: Effect of random perturbations of $\pm 15\%$ in the kinetic constants. Shown are the trajectories of the (perturbed) system (dotted line), the trajectories of the observer with $W = \text{diag}\,(1,1)$ (solid line), and the trajectories of the observer with $W = \text{diag}\,(5,5)$ (dashed line).

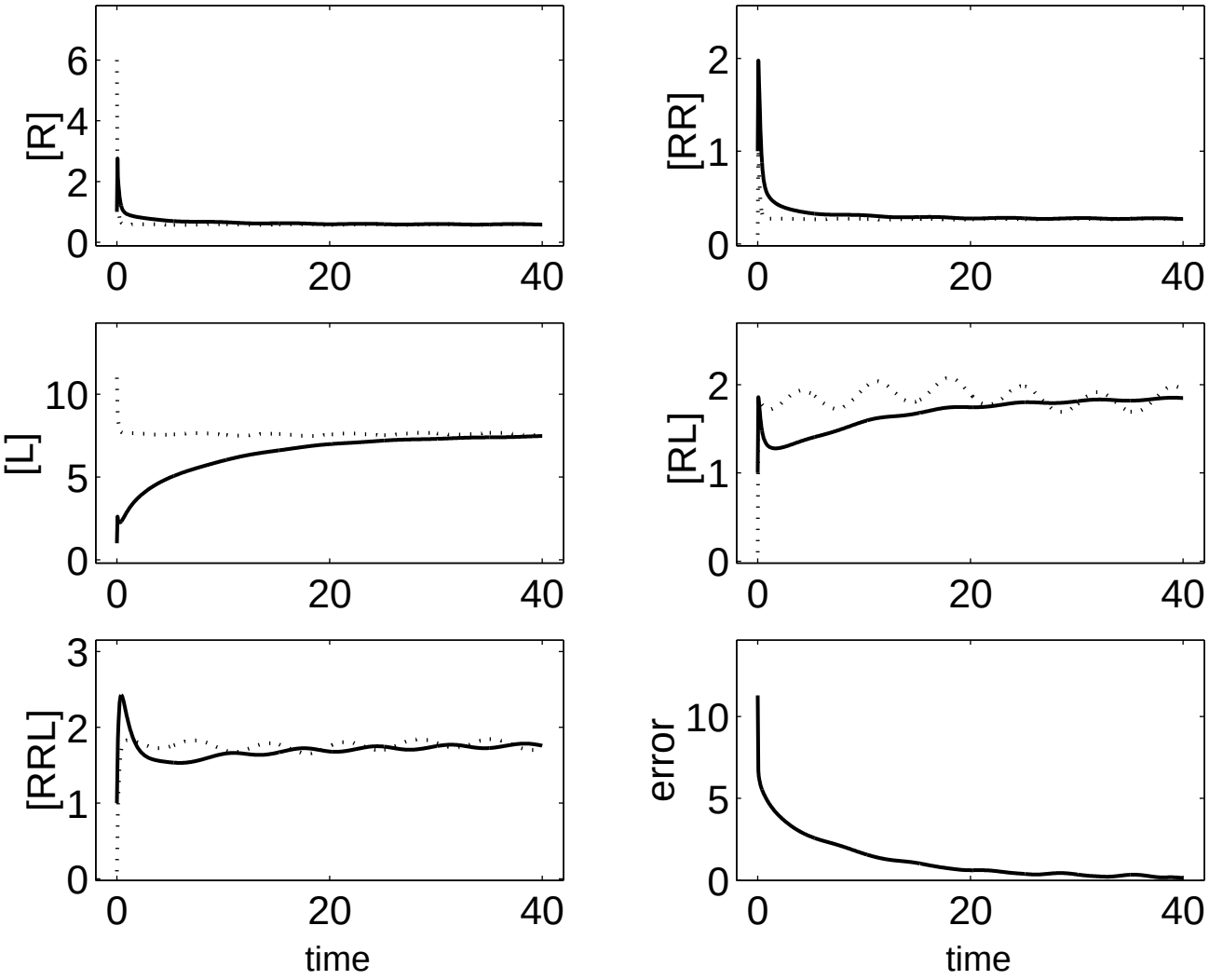


Figure 7.11: Effect of a periodic pattern for the perturbations on the kinetic constant k_{32}. Shown are the trajectories of the (perturbed) system (dotted line) and the trajectories of the observer with $W = \text{diag}\,(5, 5)$ (solid line).

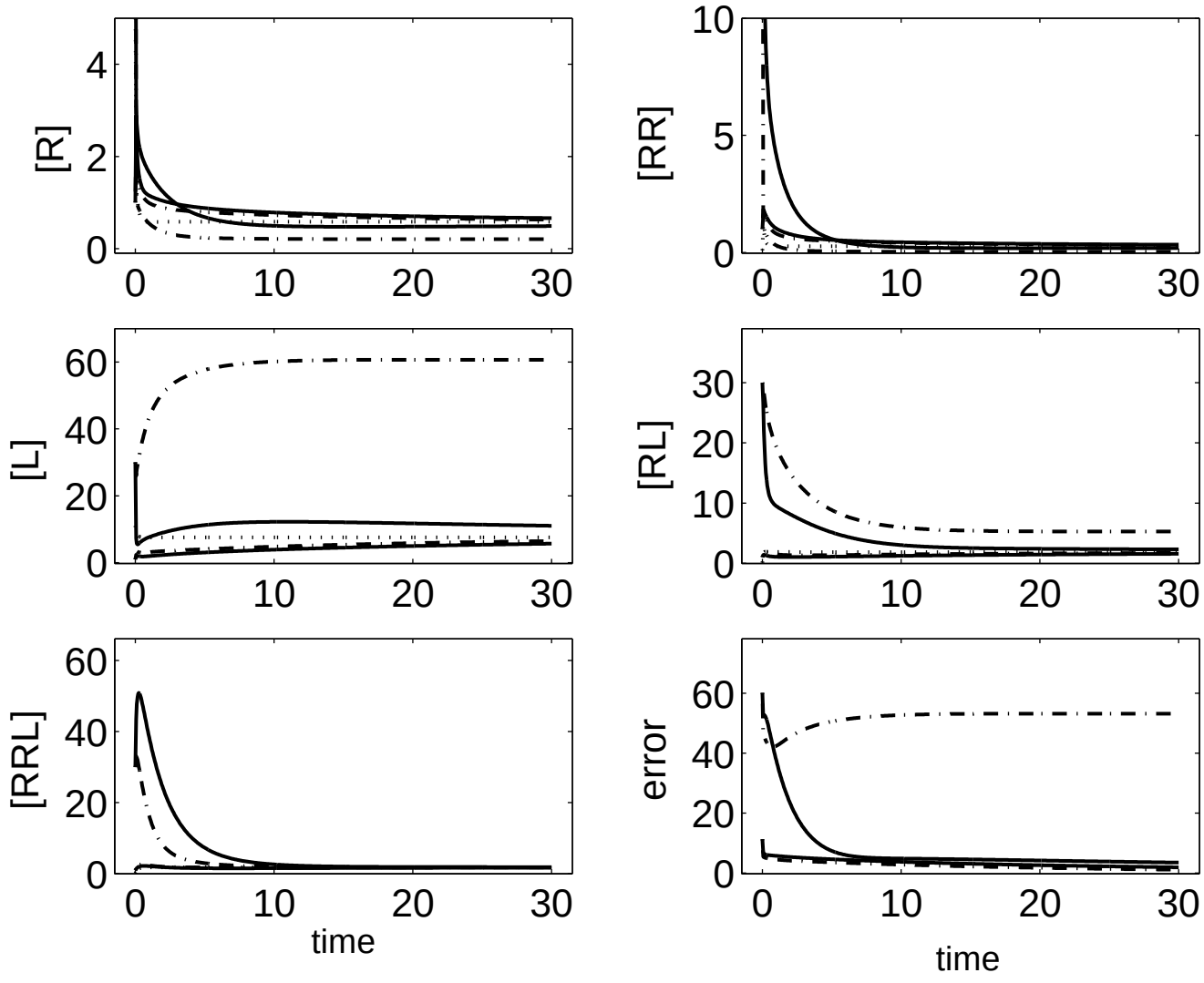


Figure 7.12: Comparison between our main observer (solid line) and an EKF (dash-dotted line). Local convergence for $z(0) = (1, 1, 1, 1, 1)'$, and divergence of the EKF for $z(0) = (30, 30, 30, 30, 30)'$.

Chapter 8
An Experiment using NMR Spectroscopy

The (main) observer constructed and analyzed in previous chapters is now further tested with data obtained from a simple experiment, performed at L. Romsted's Laboratory at the Department of Chemistry and Chemical Biology of Rutgers University. In this experiment, an NMR spectrometer was used to monitor and measure, along time, the concentrations of all species involved in the following reversible reaction:

$$\text{Methanol} + \text{Acetic acid} \xrightarrow{k_+} \text{Water} + \text{Methyl acetate} \tag{8.1}$$

$$\text{Methanol} + \text{Acetic acid} \xleftarrow{k_-} \text{Water} + \text{Methyl acetate}\ . \tag{8.2}$$

This chemical network involves four distinct species and two complexes, is clearly weakly reversible, and its dynamics may be modeled by mass-action kinetics, according to (3.2).

The experiment was performed under the conditions itemized in Chapter 2, namely, the temperature was maintained constant, the mixture in the test tube was homogeneous and there were no exchanges with the exterior.

From the collected data, as a first step, the mathematical model was calibrated and values for the kinetic constants k_+, k_- were obtained.

With the goal of validating our observer, several possible sets of measurements (or output maps) were considered, and the corresponding data fed into the observer as sampled and hold signals. The response of the observer was then compared against the actual data from the NMR readings. The effect of perturbations in the kinetic constants and the effect of output weight assignment were numerically analysed.

Finally, the performance of our observer was also compared to that of an extended Kalman filter.

8.1 The Chemical Network

Let us represent each species as follows:

$$\begin{aligned} X_1 &= \text{Methanol}\ \ (CH_4O), \\ X_2 &= \text{Acetic acid}\ \ (C_2H_4O_2), \\ X_3 &= \text{Water}\ \ (H_2O), \\ X_4 &= \text{Methyl acetate}\ \ (C_3H_6O_2), \end{aligned}$$

and let $x_i(t)$ be the concentration of species i at time t. The chemical network consisting of reactions (8.1) and (8.2) may be characterized by the matrices

$$A = \begin{pmatrix} 0 & k_- \\ k_+ & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

where the matrix A is irreducible and the matrix B has full rank, with none of its rows vanishing. The dynamics of (8.1)+8.2 may be described by the following system of ordinary differential equations

$$\begin{aligned} \frac{dx_1}{dt} &= -k_+ x_1 x_2 + k_- x_3 x_4 \\ \frac{dx_2}{dt} &= -k_+ x_1 x_2 + k_- x_3 x_4 \\ \frac{dx_3}{dt} &= k_+ x_1 x_2 - k_- x_3 x_4 \\ \frac{dx_4}{dt} &= k_+ x_1 x_2 - k_- x_3 x_4. \end{aligned} \tag{8.3}$$

The stoichiometric space is given by

$$\mathcal{D} = \text{ span } \{b_i - b_j : i, j = 1, \ldots, 2\} = \text{ span } \{(1, 1, -1, -1)'\},$$

with

$$\mathcal{D}^\perp = \text{ span } \{(1, 0, 1, 0)', (1, 0, 0, 1)', (0, 1, 1, 0)'\},$$

and recall that the *positive classes* of the system are defined as

$$\mathcal{S} = (x_0 + \mathcal{D}) \cap \mathbb{R}^n_{\geq 0}$$

for $x_0 \in \mathbb{R}^n_{\geq 0}$, and satisfy

$$\mathcal{S} \cap \mathbb{R}^n_{>0} \neq \emptyset.$$

For system (8.3), the positive classes are characterized by a triple of positive constants $(\alpha_1, \alpha_2, \alpha_3)$:

$$\begin{aligned} x_1 + x_3 &= \alpha_1 \\ x_1 + x_4 &= \alpha_2 \\ x_2 + x_3 &= \alpha_3. \end{aligned}$$

Note that

$$x_2 + x_4 = (x_2 + x_3) + (x_1 + x_4) - (x_1 + x_3) = \alpha_2 + \alpha_3 - \alpha_1,$$

so that the constants must satisfy

$$\alpha_2 + \alpha_3 > \alpha_1. \tag{8.4}$$

In the case $\alpha_2 + \alpha_3 = \alpha_1$, $x_2 + x_4 \equiv 0$ implies $x_2 \equiv 0 \equiv x_4$ and the corresponding class doesn't intersect $\mathbb{R}^n_{>0}$, so it is not a positive class. Likewise, when either of the constants α_1, α_2 or α_3 is zero, the triple doesn't define a positive class.

The boundary equilibria (i.e., those steady states for which at least one of the species vanishes) are given by Proposition 2.2.1

$$x_1 x_2 = 0, \quad \text{and} \quad x_3 x_4 = 0,$$

so

$$\begin{aligned} E_0 &= \{\bar{z} \in \partial\mathbb{R}^n_{\geq 0} : \ f(\bar{z}) = 0\} \\ &= \{(r_1, 0, r_3, 0)', (r_1, 0, 0, r_4)', (0, r_2, r_3, 0)', (0, r_2, 0, r_4)' : \ r_1, r_2, r_3, r_4 \in [0, +\infty)\}. \end{aligned}$$

We can now verify that no boundary equilibria exist in any positive class, that is,

$$\mathcal{S} \cap E_0 = \emptyset, \quad \text{for each positive class } \mathcal{S}.$$

Observe that, points of the form

1. $(0, r_2, 0, r_4)'$ imply $\alpha_1 = 0$,
2. $(0, r_2, r_3, 0')$ imply $\alpha_2 = 0$,
3. $(r_1, 0, 0, r_4)'$ imply $\alpha_3 = 0$,
4. $(r_1, 0, r_3, 0)'$ imply $\alpha_2 + \alpha_3 - \alpha_1 = 0$,

which, in the light of the discussion above, means that in any of these four cases the corresponding class $\mathcal{S}$ does not intersect $\mathbb{R}^n_{>0}$, i.e., $\mathcal{S}$ is not a positive class. We conclude that, in fact, no boundary equilibria may exist in a positive class.

The conditions of Theorem 1 are thus satisfied and we have that, for every initial condition $x_0 \in \mathbb{R}^n_{>0}$, the solution of the system differential equations (8.3) converges to the unique (positive) equilibrium, $\bar{x}$, in the same class of x_0.

8.2 The Output Maps

From Theorem 2 we know that, in order to satisfy the detectability on $\mathbb{R}^n_{>0}$ condition, the matrix C should satisfy

$$\mathcal{D} + \operatorname{im} C' = \mathbb{R}^n,$$

so, for the present example, it is necessary that rank $C = 3$. The following outputs maps certainly satisfy this condition:

$$h^{(1)}(x) = \begin{pmatrix} x_1 \\ x_3 \\ x_4 \end{pmatrix}, \ h^{(2)}(x) = \begin{pmatrix} x_1 x_2 \\ x_1 x_4 \\ x_4 \end{pmatrix}, \ h^{(3)}(x) = \begin{pmatrix} x_1 \\ x_2 x_4 \\ x_3 x_4 \end{pmatrix} . \tag{8.5}$$

Furthermore, note that, for any of these output maps,

$$h^{(l)}(E_0) \in \partial \mathbb{R}^p_{\geq 0},$$

and since, on the other hand,

$$h^{(l)}(E_+) \in \mathbb{R}^p_{>0}$$

it follows that

$$h^{(l)}(\bar{x}) \neq h^{(l)}(\bar{z}), \quad \text{whenever} \quad \bar{x} \in E_+ \text{ and } \bar{z} \in E_0.$$

Therefore, by Corollary 3.2.4, for each $l = 1, 2, 3$, the system (8.3) together with the output $y = h^{(l)}(x)$ is detectable.

In addition, for the outputs $h^{(2)}, h^{(3)}$, at least one of the measurements is one of the reaction rates (x_1x_2 and x_3x_4, respectively), so the robustness results developed in Chapter 5 apply.

The data from the experiment consist of NMR spectrometer readings of the concentration of each of the four species at regular time intervals, so the inputs into the observer will be piecewise constant signals, constructed as discussed in Section 4.5.2. In that Section it was proved that, for these sampled outputs, the observer still converges to the correct values.

8.3 NMR Spectroscopy

The reaction network (8.1), (8.2) was selected because it is a fairly simple, well known process that may be reproduced under isothermic and closed tank conditions (there are no exchanges with the exterior). Furthermore, the process could be analysed in an NMR spectrometer and, *direct*, accurate measurements of all four concentrations (methanol, acetic acid, water and methyl acetate) obtained at regular time intervals.

In an NMR spectrometer the sample (consisting of a homogeneous mixture of the species) is exposed to an external magnetic field which interacts with the nuclear magnetic dipole moments. An outline of the working principle of NMR (Nuclear Magnetic Resonance) is given next (for reference, see [22], chapter 35 and also [16]).

Some elements (hydrogen, carbon, etc.) have nuclei with a nonzero *nuclear spin angular momentum*, $\hbar\mathbf{I}$ ($\hbar$ is the Planck constant divided by 2π). This is a quantized property that satisfies $|\mathbf{I}|^2 = I(I+1)$ where I is an integral or half integral number and *depends on the type of nuclei.* The component of $\mathbf{I}$ along any particular axis, say I_z, is allowed to take values in the range: $-I, -I+1, \ldots, I-1, I$. For example, protons have spins $\pm 1/2$.

For nuclei with a nonzero nuclear spin angular momentum, the *nuclear spin magnetic momentum* is

$$\mu = \gamma \hbar \mathbf{I},$$

where γ is a constant depending on the type of nuclei. Each value of the nuclear spin angular momentum (I_z) corresponds to a different state of the atom. In the presence

of an external constant magnetic field along the z-axis $\mathbf{B}_0 = (0, 0, B_0)$, the atoms with nonzero nuclear spin magnetic moment align themselves with the direction of $\mathbf{B}_0$ (just as magnets will align themselves in the same N-S direction), so $\mu = \gamma \hbar I_z$. Under this constant magnetic field, to each state corresponds a distinct energy level given by

$$E = \gamma \, \hbar \, I_z \, B_0.$$

Transitions between adjacent energy levels may occur: from a higher level to a lower level (or, respectively, from a lower to a higher level) with the emission (respectively, absorption) of a quantum of energy $\Delta E = \pm \gamma \hbar B_0$, of frequency

$$\omega_0 = \gamma \, B_0.$$

So $\Delta E = \pm \hbar \omega_0$, where ω_0 may be seen as the frequency at which the particle precesses around the z-axis due to the magnetic field $\mathbf{B}_0$, and is called the *Larmor frequency*. For magnetic fields, the differences between adjacent energy levels are relatively low (ω_0 is in the radiofrequency band) and, as a result, the lifetime of the excited states can be fairly large, which accounts for the high resolution of the NMR spectroscopy.

In an NMR experiment, the sample is permanently exposed to a constant magnetic field, $\mathbf{B}_0$, and also to short pulses of an oscillating magnetic field, applied in a transversal direction to the constant field: $\mathbf{B}_1 = (B_x(t), B_y(t), 0)$. The atoms are excited by these short bursts of radiofrequency and transitions between adjacent energy levels will occur. According to the Boltzmann distribution, there will be slightly more atoms in the lower level states than in higher levels and this difference will induce an internal magnetic field in the sample that is detected by the NMR device. The *net magnetic moment* of the sample is defined by

$$\mathbf{M} = \sum \mu_i.$$

and, under a constant magnetic field, $\mathbf{M}$ aligns itself with $\mathbf{B}_0$, i.e., $\mathbf{M} = (0, 0, M_0)$. When the oscillating magnetic field pulse is turned on, M will acquire components on the x-y plane. When the oscillating pulse is turned off, the net magnetic moment of the sample is well described by the *Bloch equations*:

$$\begin{aligned} \frac{dM_x}{dt} &= \gamma \, (\mathbf{M} \times \mathbf{B}_0)_x - \frac{M_x}{T_2} \\ \frac{dM_y}{dt} &= \gamma \, (\mathbf{M} \times \mathbf{B}_0)_y - \frac{M_y}{T_2} \\ \frac{dM_z}{dt} &= \gamma \, (\mathbf{M} \times \mathbf{B}_0)_z - \frac{M_z - M_0}{T_1}, \end{aligned} \tag{8.6}$$

where T_1 and T_2 are relaxation times deduced from physical properties of the nuclei. Suppose, for simplicity, that the net magnetic moment of sample is due to only one type of nuclei (i.e., same γ all over the sample). Since $\mathbf{B}_0 = (0, 0, B_0)$, the Bloch equations become

$$\frac{dM_x}{dt} = \gamma \, M_y B_0 - \frac{M_x}{T_2}$$

$$\frac{dM_y}{dt} = -\gamma\, M_x B_0 - \frac{M_y}{T_2}$$
$$\frac{dM_z}{dt} = -\frac{M_z - M_0}{T_1},$$

and the solutions satisfy (recall $\omega_0 = \gamma B_0$)

$$M_x(t) + iM_y(t) = a\,(\cos\omega_0 t + i\sin\omega_0 t)\,e^{-\frac{t}{T_2}}, \quad M_z(t) = M_0 - b\,e^{-\frac{t}{T_1}},$$

for some positive constants a, $b \leq M_0$. So, the magnetic field $\mathbf{M}$ will tend to readjust itself with the constant external field $\mathbf{B}_0$: $M_z \to M_0$ with a relaxation time T_1 and $M_x, M_y \to 0$ with a relaxation time T_2. The decaying signals M_x and M_y are detected by the NMR device, and the Fourier transform of these signals will produce the usual NMR spectrum, showing a number of peaks that correspond to the frequencies ω_0.

Since γ is characteristic of the type of nuclei, so is ω_0, hence each type of nuclei will have its own signature peaks. Usually, the position of the peaks in an NMR spectrum are shown relative to a reference substance. This is due to the *chemical shift*: the spin magnetic moment of a nucleus is affected by the magnetic moments of the "environment" surrounding that nucleus (e.g., the sample consists of a very large amount of atoms of solvent, an "environment" in which the much smaller concentrations of atoms of the reactant species are evolving). This leads to a shift in the frequencies ω_0 of each species, relative to the solvent. An NMR spectrometer is usually callibrated to the reference solvent and a table of the relative positions of peaks can be constructed. Thus a species consisting of different nuclei can also be identified by several peaks, corresponding to the different nuclei. A peak is detected at ω_0 because this is the frequency at which more transitions occur — the resonance frequency. However, the peak has a slight dispersion because some transitions will still be detected at frequencies $\omega_0 \pm \epsilon$, ϵ very small, since not all the atoms will respond to exactly the same frequency.

A simple rule says that *the area below the peaks is proportional to the amount of the given species* in the sample.

Thus, through NMR spectroscopy it is possible to simultaneously and independently obtain measurements of the concentrations of all species along time. With this full information about the reaction and assuming that the mass-action kinetics system (8.3) is a good model for (8.1), we could first compute an estimate and error bounds for the kinetic constants, using some reliable straightforward method, and then test the state-estimator by using several forms of output, as indicated above, and comparing the observer's estimates with the actual data.

8.4 Results

8.4.1 The Experiment

An initial amount of each of the four species and the solvent (Table 8.1) were homogeneously mixed in an NMR tube.

The temperature was maintained constant at 20.0 °C. The sampling interval was 13 minutes and the reaction was followed during a total of 6h30m, i.e., 30 data points

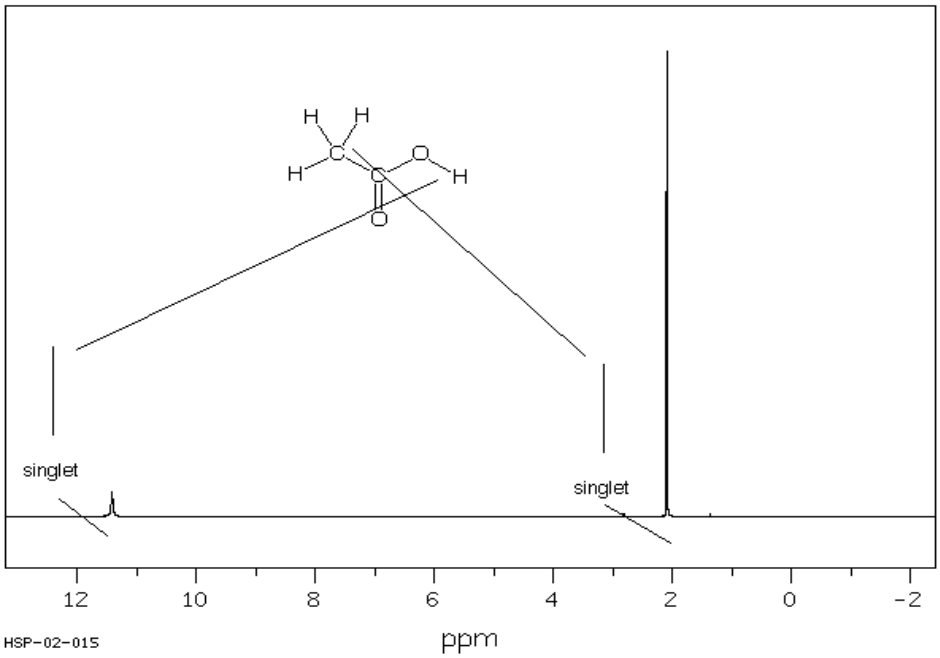


Figure 8.1: An NMR spectrum corresponding to acetic acid. (Adapted from the websites: Spectral Data Base Systems, `http://www.aist.go.jp/RIODB/SDBS/`, and University of Birmingham's `http://www.chem.bham.ac.uk/schools/compounds.htm.`)

Species	Initial Conc.
Methanol	0.705 M
Acetic Acid	0.506 M
Sulfuric Acid	0.026 M
1,3,5 Trimethoxybenzene	0.235 M
Acetonitrile-d_3	19.2 M

Table 8.1: The initial conditions of the experiment.

were collected for each of the four species. These points consist of the values of the integration of the area below the peaks given by the NMR spectrometer. This area is proportional to the concentration of the corresponding species in the sample at the given time, so it is correct to interpret each data point $x_k(t_i)$ as a measure (given in terms of "intensity" units) of the concentration of the species k at time t_i. Let us represent these data points in the form

$$\begin{aligned} \text{Methanol} &: \quad x_1(t_0), x_1(t_1), \ldots, x_1(t_{29}) \\ \text{Acetic acid} &: \quad x_2(t_0), x_2(t_1), \ldots, x_2(t_{29}) \\ \text{Water} &: \quad x_3(t_0), x_3(t_1), \ldots, x_3(t_{29}) \\ \text{Methyl acetate} &: \quad x_4(t_0), x_4(t_1), \ldots, x_4(t_{29}) \end{aligned}$$

where $\Delta t = 13$ and $t_i = i\Delta t$.

Some disappointing facts: it is clear from the data that the reaction has not reached

steady state after the 6h30m total running time. Since the theoretical results guarantee asymptotic convergence, the running time was insufficient to show the actual convergence of the observer. The methanol concentration follows a pattern of small amplitude oscillations which is not possible to account for. Unfortunately, the data available from the experiment prevents any error estimates for the accuracy of the measurements, and this may also affect the computation of the kinetic constants.

8.4.2 The Kinetic Constants

A simple approach to estimate the values of k_-, k_+ is to use the four equations (8.3) (which are linear in the parameters k_-, k_+) to find the least squares estimate for the kinetic constants. To do this, we approximate the derivatives $\frac{dx_1}{dt}$, $\frac{dx_2}{dt}$, $\frac{dx_3}{dt}$ and $\frac{dx_4}{dt}$ from the data using central differences or forward, backward at the extremes:

$$\begin{aligned}
\frac{dx_k}{dt}(t_0) &= \frac{x_k(t_1) - x_k(t_0)}{\Delta t}, \\
\frac{dx_k}{dt}(t_i) &= \frac{x_k(t_{i+1}) - x_k(t_{i-1})}{2\Delta t}, \\
\frac{dx_k}{dt}(t_{29}) &= \frac{x_k(t_{29}) - x_k(t_{28})}{\Delta t},
\end{aligned}$$

for each $i = 1, \ldots, 28$ and for $k = 1, \ldots, 4$. Then from the equations (8.3) we obtain linear systems of the form $Y = XK$ where $Y \in \mathbb{R}^{30}$, $X \in \mathbb{R}^{30\times 2}$ and $K = \begin{pmatrix} k_+ \\ k_- \end{pmatrix}$. The least squares estimator, $\hat{K}$ for the kinetics constants satisfies

$$\|Y - X\hat{K}\|^2 \leq \|Y - XK\|^2$$

for every $K \in \mathbb{R}^2$. All these computations were performed in Maple.

The poor readings of the methanol and acetic acid concentrations do not provide a reliable value for the kinetic constants (in fact, the value of k_- obtained from the least squares fit of the $\frac{dx_1}{dt}$ and $\frac{dx_2}{dt}$ equations is negative, which has obviously no physical significance). The readings of the water and methyl acetate concentrations look more reliable and, dropping the first 10 data points, fairly decent values for k_+ and k_- are obtained from the $\frac{dx_3}{dt}$ and $\frac{dx_4}{dt}$ equations. The results are summarized in Table 8.2.

Least Squares fit to	k_+	k_-
$\frac{dx_1}{dt} = -k_+x_1x_2 + k_-x_3x_4$	0.00023	-0.0003
$\frac{dx_2}{dt} = -k_+x_1x_2 + k_-x_3x_4$	0.00034	-0.00011
$\frac{dx_3}{dt} = k_+x_1x_2 - k_-x_3x_4$	0.00039	0.000008
$\frac{dx_4}{dt} = k_+x_1x_2 - k_-x_3x_4$	0.00063	0.00011
Average (last two eqs.)	0.00051	0.00006

Table 8.2: The kinetic constants.

8.4.3 The State-Estimator

The state-estimator is basically a set of non-autonomous differential equations, with the explicit time-dependence coming from the output $h(x(t))$. In order to be consistent with real time working conditions, and take into account the fact that a fresh measurement is received every 13 minutes, the output fed into the observer is really of the form "sample & hold", as discussed in Section 4.5.2:

$$\tilde{h}(t) := h(x(t_i)), \quad \text{for} \quad t_i \leq t < t_{i+1}.$$

A code was implemented in Matlab to solve the system of differential equations

$$\dot{z} = f(z) + C'(\tilde{h}(t) - h(z)) \tag{8.7}$$

using one of its built-in ODE solvers. Several outputs were tested (among which $h^{(1)}$, $h^{(2)}$, $h^{(3)}$), and the trajectories $z_1, \ldots, z_4$ were plotted against the data $x_1, \ldots, x_4$ to check the performance of the state-estimator.

An extended Kalman filter (as described in Section 7.5) was also implemented in Matlab, to compare with our estimator. The input into this EKF was also the signal $\tilde{h}(t)$.

In order to best evaluate the performance of our estimator, an "ideal system" and corresponding observer

$$\begin{aligned} \frac{dx^{th}}{dt} &= f(x^{th}), \quad x^{th}(0) = x(0), \\ \frac{dz^{th}}{dt} &= f(z^{th}) + C'(h(x^{th}(t)) - h(z^{th})), \quad z^{th}(0) = z(0), \end{aligned} \tag{8.8}$$

were simulated, in parallel with the simulation described above. The theoretical solution $x^{th}(t)$ represents the ideal evolution of the concentrations in the sample (assuming that the reactions follow exactly the mass action-kinetics model). The outputs $h(x^{th}(t))$ (continuous, as opposed to sampled), are the input into the observer. One expects that the rapidity of convergence of $|z^{th}(t) - x^{th}(t)|$ to zero may be an indicator of the rapidity of convergence of $|z(t) - x(t_i)|$.

In every Figure, the values of the concentrations z_k are given in the same "intensity" units as the data points, and one time unit corresponds to 13 minutes of real time. Unless otherwise indicated, the average values of k_+, k_- were adopted in the simulations.

To compare more consistently the performances under the different outputs, $h^{(1)}$, $h^{(2)}$ and $h^{(3)}$, the initial state of the observer was set (unless otherwise stated) to $z(0) = (6, 6, 60, 6)'$ in all simulations depicted.

8.5 Discussion

The first item to be noted is the reasonable agreement between the data and the theoretical model (8.3), after the calibration of the kinetic constants (Figure 8.2).

The response of our observer was then tested using several different outputs, among which are the examples shown in (8.5). The fact that the kinetic constants were determined with fairly large error does not seem to have a negative effect on the convergence of the observer's estimates to the (neighborhood of the) correct values and, as predicted by the robustness result developed in Chapter 5, an error proportional to the order of magnitude of the error in the kinetic constants may be expected (Figure 8.3). In this Figure, the estimates yielded by the observer (implemented with different sets of parameters k_+, k_-) differ (in the worst case) by approximately 20%.

In all the figures it is apparent that our state-estimator tends to follow closely the signal which it receives, and is typically stable with respect to irregular patterns in the signal $\tilde{h}$ (for instance, the pattern of oscillations in the methanol and acetic acid concentrations are, in general, followed by the corresponding estimates of the observer). If we interpret this pattern of oscillations as due to noisy measurements, the input-to-state stability property may be invoked to account for the stability of our state-estimator (see Section 4.5.1).

It is to be noted that, in all cases, the performance of the observer under the sampled outputs is clearly consistent with the performance of the observer under the "ideal" continuous outputs, from the comparison between the solutions of (8.7) and (8.8).

However, from the figures, the rapidity of convergence is not always very striking: it can be quite fast in some cases and for well chosen initial conditions $z(0)$, but it can also be quite slow. This apparent discrepancy in the performance may be accounted for by the relatively short running time of the experiment, i.e., the reactions have clearly not reached steady-state yet, and also because the observer is asymptotic, in the sense that $z(t)$ should be expected to be close to the correct values $x(t_i)$ only as t becomes large. In any case, of course, a choice of initial conditions $z(0)$ of the observer close to the initial values $x(0)$ of the data leads to a faster convergence of $z(\cdot)$ to the correct values.

We also note that (not very surprisingly) our state-estimator yields better results in the case of outputs such as $h^{(3)}$. That is, outputs $h = (h_1, h_2, h_3)'$ for which *all* the variables x_1, x_2, x_3 and x_4 appear in at least one of the products h_1, h_2 or h_3 (Figure 8.6). Note that, in these cases, *all* the equations dz_k/dt have a nonzero correction term which will increase the rapidity of convergence. Moreover, for these cases, a certain control over the rapidity of convergence is possible, by weigthing the output, i.e., multiplying each measurement h_1, h_2 and h_3 by an appropriate positive constant w_1, w_2 and w_3, as discussed in Section 4.5.3 (Figure 8.7).

Another interesting item is the comparison with an extended Kalman filter (EKF). In most cases the performances are similar, however our estimator produces the best results when the measured quantities are the reaction rates x_1x_2, x_3x_4 or other products x_1x_3, x_1x_4, etc. (Figures 8.5 and 8.6) as opposed to EKF which yields the best results in the case of linear outputs (Figure 8.4).

8.6 Conclusions from the Experiment

The performance of our state-estimator (8.7) under the experiment's sampled data is in very good agreement with the "expected" performance (that of system (8.8), which receives the ideal continuous outputs from the nominal system), during all the running time of the experiment. However, the global convergence of the state-estimator to the correct values cannot be fully checked, due to the short running time of the experiment.

The robustness with respect to perturbations in the kinetic constants (discussed in Chapter 5) can be verified in these tests, so far as the estimates produced by the observer – when this is constructed with different parameters – are within approximately 20% of each other.

The stability with respect to perturbations in the outputs is also apparent in the performance of our state-estimator.

Finally, our state-estimator's performance is, locally, comparable to that of a standard construction (EKF). While in some cases the EKF seems to have better performance, on the other hand, our state-estimator shows better results globally, as well as for nonlinear outputs.

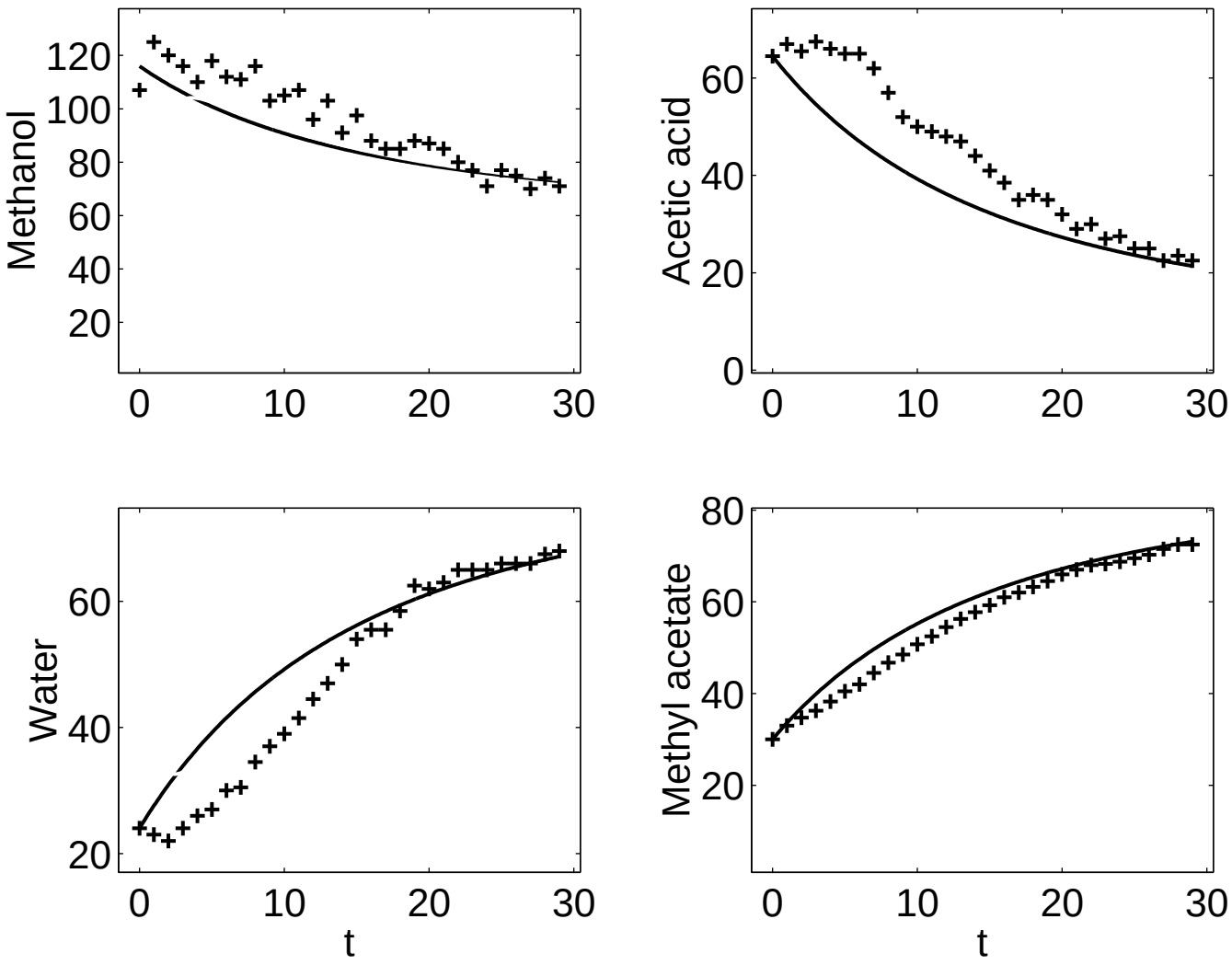


Figure 8.2: The data fits the theoretical model.

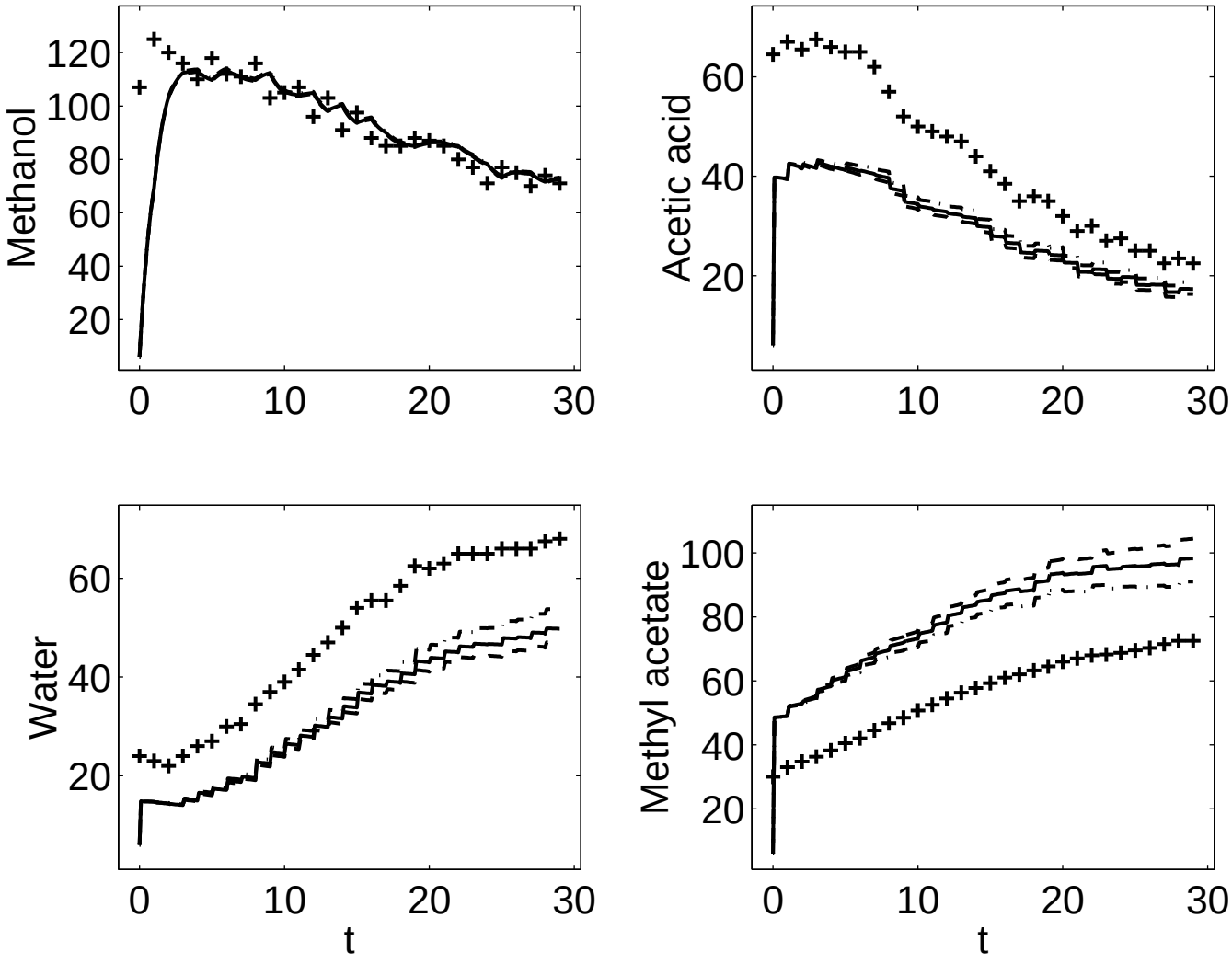


Figure 8.3: Robustness with respect to errors in the kinetics constants. Output is $h^{(3)}$ and $z(0) = (6, 6, 6, 6)'$. Shown are the trajectories of our estimator for 3 cases: (i) $k_+ = 0.00051$, $k_- = 0.00006$ (solid line); (ii) $k_+ = 0.0006$, $k_- = 0.000008$ (dashed line); (iii) $k_+ = 0.00039$, $k_- = 0.00011$ (dash-dotted line).

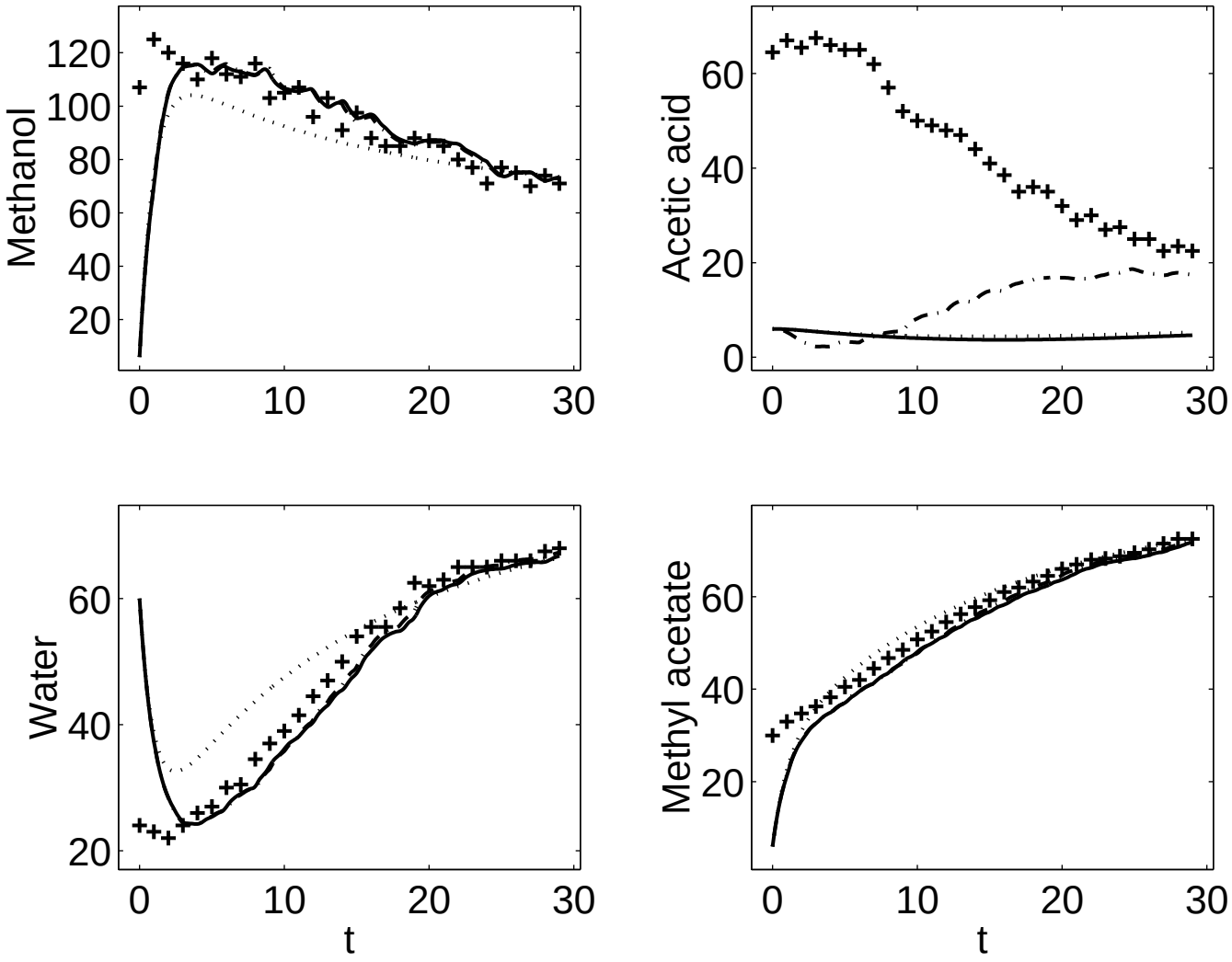


Figure 8.4: The trajectories of our estimator (solid), EKF (dash-dot) and the "theoretical" estimator (dot). The output is $h^{(1)}$.

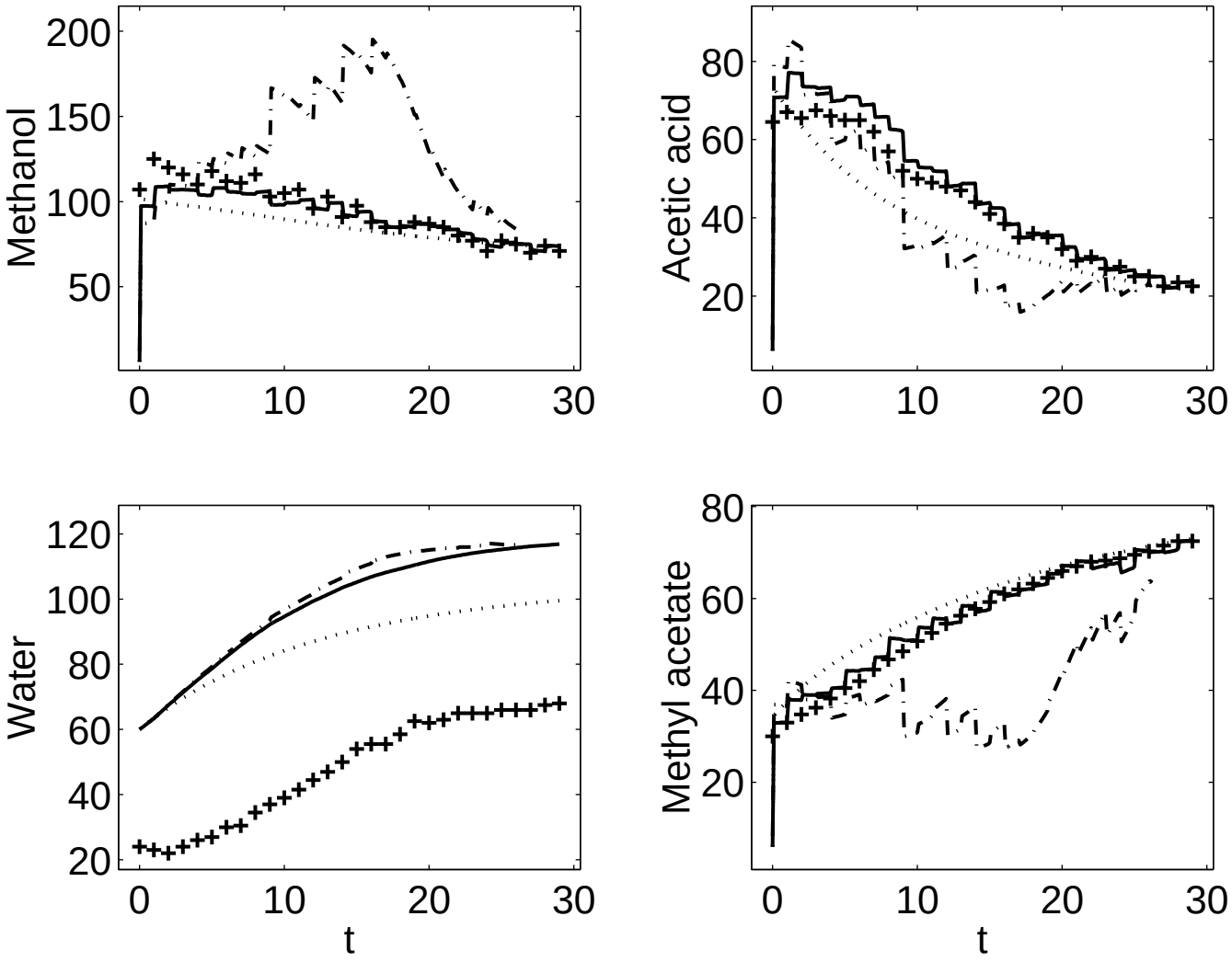


Figure 8.5: The trajectories of our estimator (solid), EKF (dash-dot) and the "theoretical" estimator (dot). The output is $h^{(2)}$.

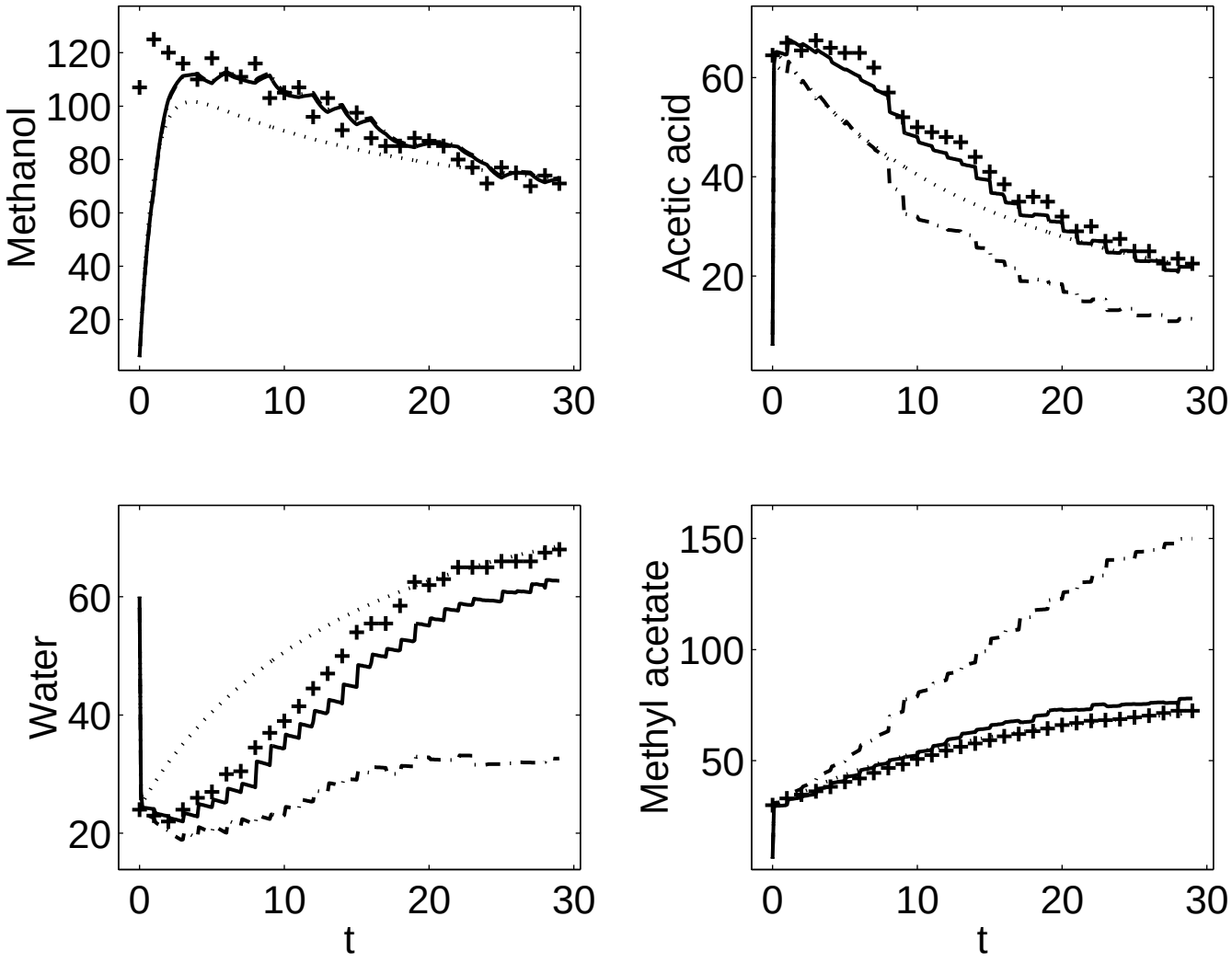


Figure 8.6: The trajectories of our estimator (solid), EKF (dash-dot) and the "theoretical" estimator (dot). The output is $h^{(3)}$.

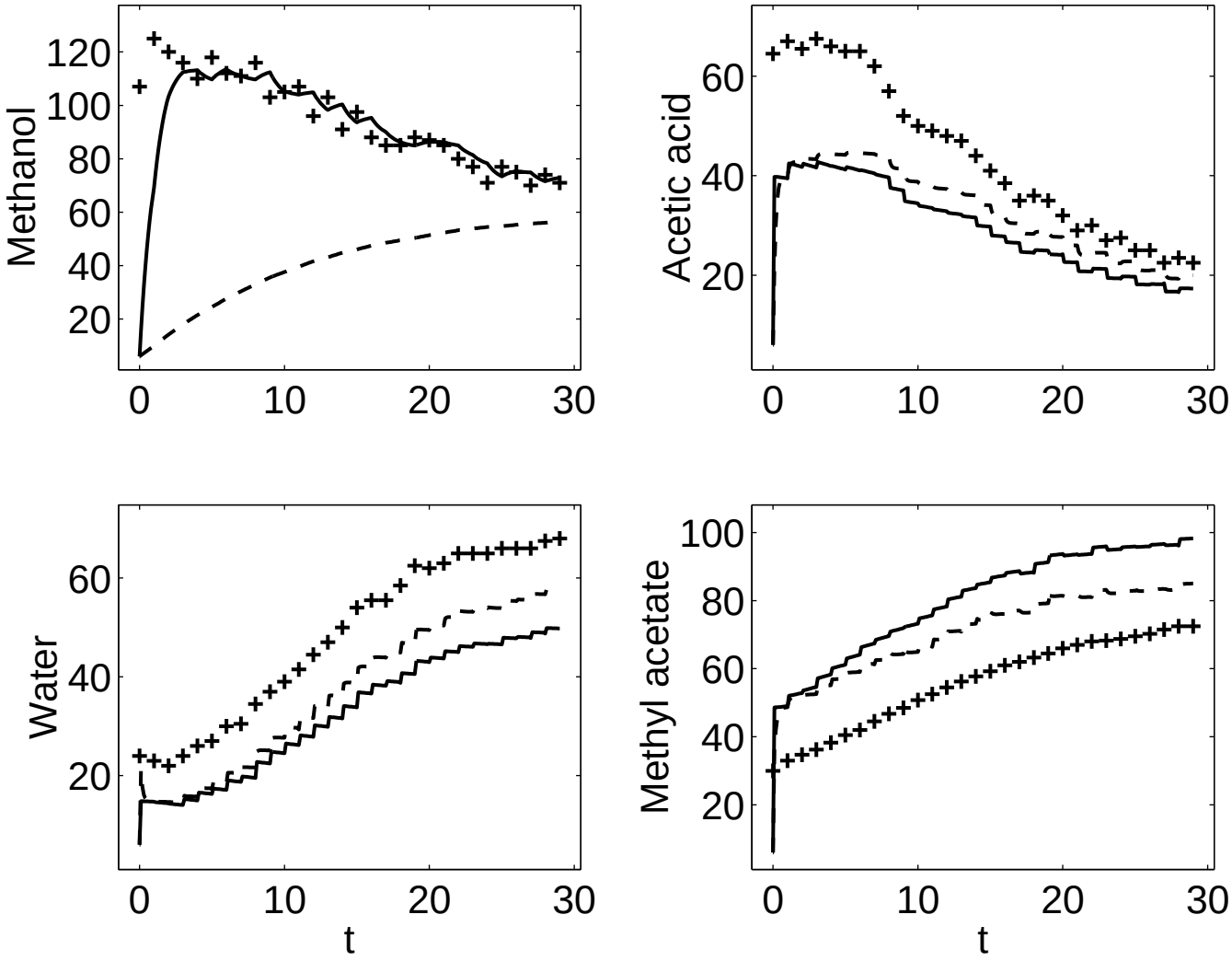


Figure 8.7: Weigthed output $h^{(3)}$, with $z(0) = (6, 6, 6, 6)'$. Shown are the trajectories of our estimator for: (i) $w_1 = 1$, $w_2 = 1$, $w_3 = 1$ (solid line); (ii) $w_1 = 0.04$, $w_2 = 0.1$, $w_3 = 1$ (dashed line).

Appendix A
A Separation Principle

When the output map h is of the form described in Remark 3.2.6, a direct proof of Theorem 3 can be given using a Lyapunov function for the composite system (plant and observer), which is entirely analogous to the "separation theorem" proof used to establish the convergence of Luenberger observers. Let us sketch the procedure here.

Consider the error $\epsilon = x - z$ and look at the composite system in new coordinates x and ϵ instead of x and z:

$$\Sigma_\epsilon : \begin{array}{l} \dot{x} = f(x) \\ \dot{\epsilon} = f(x) - f(x-\epsilon) - C'(h(x) - h(x-\epsilon)) \,. \end{array}$$

Given any $\bar{x}$, consider the following function:

$$V_2(x,\epsilon) = V(x-\epsilon,\bar{x}) = \sum_{i=1}^{n} \bar{x}_i \left[\frac{x_i - \epsilon_i}{\bar{x}_i} \ln \frac{x_i - \epsilon_i}{\bar{x}_i} + 1 - \frac{x_i - \epsilon_i}{\bar{x}_i} \right].$$

Note that $x - \epsilon = z \in \mathbb{R}^n_{\geq 0}$, so from Lemma 5.2.1 we know that there exists a function α of class $\mathcal{K}_\infty$ such that V_2 satisfies the following estimate

$$\frac{d}{dt} V_2(x(t),\epsilon(t)) \leq -\alpha(|x - \epsilon - \bar{x}|) + c_3 |h(x) - h(\bar{x})|^2$$

for all $x = x(t)$ such that $|h_k(x) - h_k(\bar{x})| < \frac{h_k(\bar{x})}{2}$ for all $k = 1, \ldots, p$. (here we are taking the number θ from Lemma 5.2.1 to be equal to 1/2.) Since $x(t) \to \bar{x}$ as $t \to +\infty$ (Theorem 1), there exists some T such that this inequality holds for every $t > T$.

(Indeed this estimate holds since

$$\begin{array}{rcl} \frac{d}{dt} V_2(x(t),\epsilon(t)) & = & \nabla_x V_2(x(t),\epsilon(t)) \;\; \dot{x} \;+\; \nabla_\epsilon V_2(x(t),\epsilon(t)) \;\; \dot{\epsilon} \\ & = & \nabla V(x(t) - \epsilon(t)) \;\; \dot{x} \;+\; \nabla V(x(t) - \epsilon(t)) \;\; \dot{\epsilon} \\ & = & \nabla V(x(t) - \epsilon(t)) \;\; [f(x(t) - \epsilon(t)) + C'(h(x(t)) - h(x(t) - \epsilon(t)))] \\ & = & \nabla V(z(t)) \;\; f^*(z(t)), \end{array}$$

and now we may apply Lemma 5.2.1.)

Since $\epsilon = x - z$ and we have shown that both solutions $x(t)$ and $z(t)$ exist for all $t \geq 0$, we may conclude that a solution $\epsilon(t)$ to the ϵ-equation also exists for all $t \geq 0$ (for a direct proof of this fact, one may use a similar argument to that of Proposition 4.3.1).

Recall that we already have a Lyapunov function, $V_1(x) = \sum_{i=1}^n \bar{x}_i \, g(x_i/\bar{x}_i)$, for the system $\dot{x} = f(x)$, which satisfies an estimate of the form (see [48]):

$$\frac{d}{dt} V_1(x(t)) \leq -c(\bar{x}) \, \delta(x,\bar{x})$$

where

$$\delta(x,\bar{x}) = c_1|x-\bar{x}|^2 + d(|x-\bar{x}|^2)$$

and $d(r) = o(r)$ as $r \to 0$. Pick $\varepsilon < |\bar{x}|/2$ such that

$$|d(|x-\bar{x}|^2)| < \frac{c_1}{2}|x-\bar{x}|^2 \quad \forall \ |x-\bar{x}| < \varepsilon\,.$$

Now choose the time $T = T(\varepsilon)$ so large that

$$|x(t)-\bar{x}| < \varepsilon \quad \text{and} \quad |h(x(t)) - h(\bar{x})| < \frac{h(\bar{x})}{2}, \text{ for all } t > T.$$

So, for $t > T$, we also have that $x(t)$ evolves in the closed ball $B(\bar{x}) = \{a \in \mathbb{R}^n_{\geq 0} : |a-\bar{x}| \leq |\bar{x}|/2\}$, and hence there exists a Lipschitz constant, c_L, (take, for example, $c_L = \max|\frac{\partial h_k}{\partial x_j}(x)|$ with the maximum taken over $j = 1,\ldots,n$ and $k = 1,\ldots,p$ and $x \in B(\bar{x})$) such that

$$|h(x) - h(\bar{x})| \leq c_L|x-\bar{x}|, \quad \forall\ x \in B(\bar{x}).$$

Thus for $t > T$, the following estimates hold:

$$\begin{aligned} \tfrac{d}{dt}V_1(x(t)) &\leq -\tfrac{c_1 c(\bar{x})}{2}|x-\bar{x}|^2, \\ \tfrac{d}{dt}V_2(x(t),\epsilon(t)) &\leq -\alpha(|x-\epsilon-\bar{x}|) + c_3 c_L^2|x-\bar{x}|^2. \end{aligned}$$

This means that

$$W(x,\epsilon) = \frac{2(c_3c_L^2+2)}{c_1c(\bar{x})}V_1(x) + V_2(x,\epsilon)$$

is a Lyapunov function for the system Σ_ϵ, with

$$\frac{d}{dt}W(x(t),\epsilon(t)) \leq -\alpha(|x-\epsilon-\bar{x}|) - 2|x-\bar{x}|^2,$$

for all $t > T$.

It is not difficult to verify that this estimate can be rewritten as

$$\frac{d}{dt}W(x(t),\epsilon(t)) \leq -\tilde{\alpha}(|\epsilon|^2 + |x-\bar{x}|^2)$$

for some other $\tilde{\alpha} \in \mathcal{K}_\infty$. (First note that, with $\alpha_2(c) = c^2$ and using Lemma B.0.1, there exists $\alpha_3 \in \mathcal{K}_\infty$ such that

$$\begin{aligned} \alpha(|x-\epsilon-\bar{x}|) + \alpha_2(|x-\bar{x}|) &\geq \alpha_3(\sqrt{|x-\epsilon-\bar{x}|^2 + |x-\bar{x}|^2}) \\ &\geq \alpha_3\left(\frac{1}{2}|x-\epsilon-\bar{x}| + \frac{1}{2}|x-\bar{x}|\right) \\ &\geq \alpha_3\left(\frac{1}{2}|\epsilon|\right). \end{aligned}$$

And then, using Lemma B.0.1 again,

$$\alpha_3(|\epsilon|/2) + \alpha_2(|x-\bar{x}|) \geq \alpha_3(\sqrt{|\epsilon|^2 + |x-\bar{x}|^2}).)$$

Next, notice that

$$W(x,\epsilon) \leq \frac{2(c_3 c_L^2 + 2)}{c_1 c(\bar{x})} \nu_2(|x - \bar{x}|) + \nu_2(|x - \epsilon - \bar{x}|),$$

where $\nu_2 \in \mathcal{K}_\infty$ is as in (3.11), so immediately

$$W(x,\epsilon) \leq \kappa \nu_2(2\sqrt{|\epsilon|^2 + |x - \bar{x}|^2}) := \tilde{\nu}_2(|\epsilon|^2 + |x - \bar{x}|^2),$$

where

$$\kappa = 1 + \frac{2(c_3 c_L^2 + 2)}{c_1 c(\bar{x})}.$$

Then

$$\frac{d}{dt} W(x(t), \epsilon(t)) \leq -\tilde{\alpha}(\tilde{\nu}_2^{-1}(W(x,\epsilon)))$$

which implies that $W(x(t), \epsilon(t)) \to 0$ as $t \to +\infty$ and so $(x(t), \epsilon(t)) \to (\bar{x}, 0)$, as wanted.

Appendix B
Upper and Lower Bounds for an ISS-Lyapunov Function

For a general ISS-Lyapunov function defined according to Definition 3.3.3, one can show the existence of class $\mathcal{K}_\infty$ bounds, as follows. Let

$$\nu_1(r) = \inf\{V(z) : |z - \bar{x}| \geq r \text{ and } z \in \mathbb{R}^n_{\geq 0}\}$$

and

$$\nu_2(r) = \max\{V(z) : |z - \bar{x}| \leq r \text{ and } z \in \mathbb{R}^n_{\geq 0}\}.$$

The function ν_2 is continuous because V is, nondecreasing and satisfies $\nu_2(0) = 0$. Without loss of generality, we can say that it is strictly increasing (by taking $\nu_2(r) + r$).

The function ν_1 is finite, because for any $r > 0$, we may pick any x_r with $|x_r - \bar{x}| = r$ and consider $L := V(x_r)$. Then

$$\nu_1(r) = \min\{V(z) : |z - \bar{x}| \geq r \text{ and } z \in \mathbb{R}^n_{\geq 0} \text{ and } V(z) \leq L\}$$

(the minimum over this set exists, because it is a compact). Thus ν_1 is also continuous and nondecreasing and $\nu_1(0) = 0$. Without loss of generality, ν_1 can be assumed to be strictly increasing.

In the particular case of the function V defined in (3.12) the functions ν_1, ν_2 may be taken as follows. Consider $v : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ given by:

$$v(a) = \begin{cases} (1-a)\ln(1-a) + a, & 0 \leq a \leq 1 \\ (1+a)\ln(1+a) - a + 2(1 - \ln 2), & 1 \leq a < +\infty. \end{cases}$$

Then v is a $\mathcal{K}_\infty$ function (note that $2(1 - \ln 2)$ is a positive quantity that guarantees continuity of v at $a = 1$).

From (3.12), recall that $g(r) = r \ln r + 1 - r$ for $r \geq 0$. Equivalently,

$$g(1 + a) = (1 + a)\ln(1 + a) - a, \text{ for } a \geq -1.$$

Then

$$g(1+a) \leq \begin{cases} (1-|a|)\ln(1-|a|) + |a|, & |a| \leq 1 \\ (1+a)\ln(1+a) - a + 2(1 - \ln 2), & 1 \leq a < +\infty. \end{cases}$$

So $g(z_i/\bar{x}_i) \leq v(|z_i - \bar{x}_i|/\bar{x}_i)$ and

$$V(z) \leq \sum_{i=1}^{n} \bar{x}_i \, v\left(\frac{|z_i - \bar{x}_i|}{\bar{x}_i}\right).$$

By using the fact that

$$v\left(\frac{|z_i - \bar{x}_i|}{\bar{x}_i}\right) \leq v\left(\frac{|z_i - \bar{x}_i|}{\min \bar{x}_i}\right),$$

and also keeping in mind that $v(a_i) \leq v(\sqrt{a_1^2 + \cdots + a_n^2})$, it follows that

$$V(z) \leq n \max \bar{x}_i \, v(|z - \bar{x}|/\min \bar{x}_i) = \nu_2(|z - \bar{x}|).$$

For the lower bound, consider

$$w(a) = (1 + a)\ln(1 + a) - a, \quad 0 \leq a \leq +\infty$$

and notice that

$$g(1 + a) \geq \begin{cases} (1 + |a|)\ln(1 + |a|) - |a|, & |a| \leq 1 \\ (1 + a)\ln(1 + a) - a, & 1 \leq a < +\infty. \end{cases}$$

So, $g(z_i/\bar{x}_i) \geq w(|z_i - \bar{x}_i|/\bar{x}_i)$ and

$$V(z) \geq \sum_{i=1}^{n} \bar{x}_i w\left(\frac{|z_i - \bar{x}_i|}{\bar{x}_i}\right).$$

Using Lemma B.0.1 repeatedly, we may conclude that V is lower bounded by a $\mathcal{K}_\infty$ function.

The following is easy to prove using $\alpha_3(c) = \min\{\alpha_1(c/2), \alpha_2(c/2)\}$:

Lemma B.0.1 Assume that α_1, α_2 are two $\mathcal{K}_\infty$ functions. Let a and b be positive numbers. Then there exists another $\mathcal{K}_\infty$ function, α_3, such that

$$\alpha_1(a) + \alpha_2(b) \geq \alpha_3(\sqrt{a^2 + b^2}).$$

for all $a, b > 0$.

Appendix C
Some Simple Facts Concerning Invariance

Consider the scalar initial value problem

$$\begin{aligned} \dot{x} &= F(t,x) \\ x(0) &= x_0 \end{aligned} \tag{C.1}$$

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 (C.1) 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 [48]), we have:

Lemma C.0.2 Consider system (C.1) 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 $x(t)>0 \quad\forall t\in J$, i.e., the solution of (C.1) remains positive for all times in J.

Lemma C.0.3 Consider again the scalar initial value problem (C.1), but now with domain $\mathcal{X}=(0,+\infty)$. Assume that the initial condition is positive: $x_0>0$, and that the following property holds, for some $0<\varepsilon<x_0$:

$$\forall t\in\mathbb{R},\quad F(t,x)>0 \quad\text{whenever}\quad x<2\varepsilon. \tag{C.2}$$

Then for all $t\in J$, the solutions of (C.1) satisfy $x(t)>\varepsilon$.

Proof. Suppose that the result is false, and let

$$s=\inf_{t\in J}\{t:x(t)\leq\varepsilon\}.$$

We may assume that $s < \infty$ and $s \in J$, because otherwise the desired result is trivial. In other words, let us suppose that the solution $x(t)$ of (C.1) hits the value ε at some finite instant s in J.

Continuity of solutions gives

$$\exists\, \delta = \delta(\varepsilon) \quad \text{such that} \quad |x(t) - \varepsilon| < \varepsilon/2 \quad \forall\, t \in (s - \delta, s + \delta).$$

By property (C.2), $F(t, x(t)) > 0\ \forall t \in (s - \delta, s + \delta)$ implying that $x(t)$ is strictly increasing on this δ-interval and so $x(s) > x(s - \delta)$. But, then $x(s - \delta) < \varepsilon$, which contradicts the definition of s. ∎

Appendix D
Some auxiliar calculations

Lemma D.0.4 Let $p \leq n$ be integers and $C \in \mathbb{R}^{p\times n}$ be a matrix of full rank p. Let h be the map (3.3), i.e., $h(x) = \mathrm{Exp}(C\rho(x))$. There exists a function α of class $\mathcal{K}_\infty$ such that,

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

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

Proof. Define, for each fixed z, the following finite, disjoint sets of integers

$$I_+ = I_+(z) = \{i : h_i(z) - h_i(\bar{x}) \geq 0\}$$

and

$$I_- = I_-(z) = \{i : h_i(z) - h_i(\bar{x}) < 0\}.$$

Write for each i,

$$\frac{h_i(z) - h_i(\bar{x})}{h_i(\bar{x})} = \ln h_i(z) - \ln h_i(\bar{x}) + g(h_i(z), h_i(\bar{x})).$$

It is not difficult to check that

$$g(h_i(z), h_i(\bar{x})) = \frac{h_i(z)}{h_i(\bar{x})} - \ln \frac{h_i(z)}{h_i(\bar{x})} - 1$$

has a minimum value of 0, achieved whenever $h_i(z) = h_i(\bar{x})$, and otherwise is strictly positive.

Introduce the notation:

$$\begin{aligned} \mu &= h(z) - h(\bar{x}) \\ \varrho &= \rho(z) - \rho(\bar{x}) \\ \sigma &= C(\rho(z) - \rho(\bar{x})) = C\varrho. \end{aligned}$$

Let C_i denote the i-th row of the matrix C and let

$$\frac{\mu_i}{h_i(\bar{x})} = \xi_i + \zeta_i \quad \text{and} \quad \sigma_i = C_i\varrho = \xi_i + \eta_i$$

with

$$\xi_i = \begin{cases} \sigma_i = \ln h_i(z) - \ln h_i(\bar{x}) = C_i\varrho, & \text{if } i \in I_+ \\ \mu_i/h_i(\bar{x}) = (h_i(z) - h_i(\bar{x}))/h_i(\bar{x}), & \text{if } i \in I_- \end{cases}$$

$$\zeta_i = \begin{cases} g(h_i(z), h_i(\bar{x})), & \text{if } i \in I_+ \\ 0, & \text{if } i \in I_- \end{cases}$$

$$\eta_i = \begin{cases} 0, & \text{if } i \in I_+ \\ -g(h_i(z), h_i(\bar{x})), & \text{if } i \in I_-. \end{cases}$$

By construction, $(\xi_i + \zeta_i)(\xi_i + \eta_i) \geq \xi_i^2$, since all cross terms are nonnegative:

$$\xi_i\zeta_i = \begin{cases} g(h_i(z), h_i(\bar{x}))(\ln h_i(z) - \ln h_i(\bar{x})) \geq 0, & \text{if } i \in I_+ \\ 0, & \text{if } i \in I_- \end{cases}$$

$$\xi_i\eta_i = \begin{cases} 0, & \text{if } i \in I_+ \\ -g(h_i(z), h_i(\bar{x}))(h_i(z) - h_i(\bar{x}))/h_i(\bar{x}) \geq 0, & \text{if } i \in I_-, \end{cases}$$

and $\zeta_i\eta_i = 0$ for all i. Therefore, for each $i \in I_+$

$$\begin{aligned} \sigma_i\mu_i &= h_i(\bar{x})(\xi_i + \zeta_i)(\xi_i + \eta_i) \\ &\geq h_i(\bar{x})\xi_i^2 \\ &= h_i(\bar{x})(C_i\varrho)^2, \end{aligned}$$

and for each $i \in I_-$

$$\begin{aligned} \sigma_i\mu_i &= h_i(\bar{x})C_i\varrho\left(\frac{h_i(z) - h_i(\bar{x})}{h_i(\bar{x})}\right) \\ &= h_i(\bar{x})|C_i\varrho|\left|\frac{h_i(z) - h_i(\bar{x})}{h_i(\bar{x})}\right| \\ &= h_i(\bar{x})|C_i\varrho|\left|\frac{h_i(z)}{h_i(\bar{x})} - 1\right| \\ &= h_i(\bar{x})|C_i\varrho|\left|e^{\ln\left(\frac{h_i(z)}{h_i(\bar{x})}\right)} - 1\right| \\ &= h_i(\bar{x})|C_i\varrho|\left|e^{[\ln h_i(z) - \ln h_i(\bar{x})]} - 1\right| \\ &= h_i(\bar{x})|C_i\varrho|\left|e^{-|C_i\varrho|} - 1\right| \end{aligned}$$

since $C_i\varrho = \ln h_i(z) - \ln h_i(\bar{x}) < 0$ for $i \in I_-$.

In summary, with $c_2 = \min\{h_1(\bar{x}), \ldots, h_p(\bar{x})\}$,

$$\langle \sigma, \mu \rangle \geq c_2 \left(\sum_{i \in I_+} |C_i\varrho|^2 + \sum_{i \in I_-} |C_i\varrho| \left(1 - e^{-|C_i\varrho|}\right) \right),$$

and since both functions $a(r) = c_2\, r^2$ and $b(r) = c_2\, r(1 - e^{-r})$ are of class $\mathcal{K}_\infty$, one may apply Lemma B.0.1 repeatedly to obtain $\alpha \in \mathcal{K}_\infty$ such that $\langle \sigma, \mu \rangle \geq \alpha(|C\varrho|)$. ∎

The next lemmas state some simple inequalities which are used throughout the text.

Lemma D.0.5 Let $v, w \in (-1, +\infty)$. Then

$$(\ln(1+v))^2 + (\ln(1+w))^2 \geq \frac{1}{2}\left(\ln(1+\sqrt{v^2+w^2})\right)^2.$$

Proof. First, note that $(\ln(1+v))^2 \geq (\ln(1-v))^2$ for all $-1 < v < 0$ by putting,

$$g(v) = (\ln(1+v))^2 - (\ln(1-v))^2 = (\ln(1+v) + \ln(1-v))(\ln(1+v) - \ln(1-v)).$$

The second factor is clearly negative and the first factor is strictly increasing on the interval $(-1, 0)$ (by computing the derivative) and is 0 when $v = 0$, therefore it must be also negative on this interval, implying that $g(v) \geq 0$ as desired.

Then it holds that

$$(\ln(1+v))^2 \geq (\ln(1+|v|))^2, \quad \forall v \in (-1, +\infty),$$

so it is enough to prove the lemma for all $v, w \in [0, \infty)$. Using the fact that x^2 is a convex function, it follows that

$$\begin{aligned} \frac{1}{2}(\ln(1+v))^2 + \ln(1+w))^2) &\geq \frac{1}{4}(\ln(1+v) + \ln(1+w))^2 \\ &= \frac{1}{4}(\ln(1+v+w+vw))^2. \end{aligned} \tag{D.1}$$

Moreover, since both v and w are positive: $w + v \geq \sqrt{v^2+w^2}$. The logarithm is an increasing function, so

$$\frac{1}{4}(\ln(1+v+w+vw))^2 \geq \frac{1}{4}(\ln(1+\sqrt{v^2+w^2}))^2. \tag{D.2}$$

Putting (D.1) and (D.2) together gives desired result. ∎

Lemma D.0.6 Let $\ell \geq n$ and let M be an $\ell \times n$ matrix of rank n, and consider a fixed compact subset P of $\mathbb{R}^n_{>0}$. Then, there exists a class $\mathcal{K}_\infty$ function α such that $|M(\rho(x) - \rho(a))|^2 \geq \alpha(|x-a|)$ for all $a \in P$ and all $x \in \mathbb{R}^n_{>0}$.

Proof. Since M has maximal rank, it has a left pseudo-inverse, for example $M^\# = (M'M)^{-1}M'$. Then, for any vector $\varrho \in \mathbb{R}^n$,

$$|\varrho| = |M^\# M\varrho| \leq \|M^\#\||M\varrho| \quad \Longrightarrow \quad |M\varrho| \geq \frac{1}{\|M^\#\|}|\varrho|.$$

Also, if $\varrho = \rho(x) - \rho(a)$,

$$\varrho_i = \ln\frac{x_i}{a_i} = \ln\left(1 + \frac{x_i - a_i}{a_i}\right) := \ln(1+v_i).$$

Then, applying Lemma D.0.5 inductively on n we have

$$|\varrho|^2 \geq \frac{1}{2^{n-1}}\left(\ln\left(1+\sqrt{v_1^2+\cdots+v_n^2}\right)\right)^2 \geq \frac{1}{2^{n-1}}\left(\ln\left(1+\frac{|x-a|}{\max a_i}\right)\right)^2.$$

Now, clearly

$$\alpha(w) = \frac{1}{2^{n-1}\|M^\#\|^2}\left(\ln\left(1+\frac{w}{\bar{a}}\right)\right)^2$$

with $\bar{a}$ an upper bound on the magnitudes of the coordinates of points of P, is of class $\mathcal{K}_\infty$, which proves the lemma. ∎

References

[1] F. Albertini and E.D. Sontag, "Continuous control-Lyapunov functions for asymptotically controllable time-varying systems", *International Journal of Control*, **72**(1999), pp. 1630-1641.

[2] D. Angeli, E.D. Sontag, and Y. Wang, "A characterization of integral input-to-state stability," *IEEE Transactions on Automatic Control*, **46**(2000), pp. 1082-1097.

[3] D. Angeli, E.D. Sontag, and Y. Wang, "Further equivalences and semiglobal versions of integral input-to-state stability," *Dynamics and Control* **10**(2000), pp. 127-149.

[4] M. Arcak and P. Kokotović, "Observer-based control of systems with slope-restricted nonlinearities," *IEEE Transactions on Automatic Control*, **46**(2001), pp. 1146-1150.

[5] G. Bastin and J.F. van Impe, "Nonlinear and adaptive control in biotechnology: a tutorial," *European Journal of Control* **1**(1995), pp. 1-37.

[6] G. Bastin and D. Dochain, *On-line Estimation and Adaptive Control of Bioreactors*, Elsevier, Amsterdam, 1990.

[7] G. Bastin and M.R. Gevers, "Stable adaptive observers for nonlinear time-varying systems", *IEEE Transactions on Automatic Control*, **33**(1988), pp. 650-658.

[8] D. Bernstein and S.J. Bhat, "Nonnegativity, reducibility, and semistability of mass action kinetics," in *Proceedings of the IEEE Conference on Decision and Control*, IEEE Publications, Dec. 1999, pp. 2206-2211.

[9] D. Bestle and M. Zeitz, "Canonical form observer design for nonlinear time variable systems," *International Journal of Control* **38**(1983), pp. 419-431.

[10] R.P. Bywater, A. Sørensen, P. Røgen and P.G. Hjorth "Construction of the simplest model to explain complex receptor activation kinetics" *Journal of Theoretical Biology* **218**(2002), pp. 139-147.

[11] G. Ciccarella, M. Dalla Mora and A. Germani, "A Luenberger-like observer for nonlinear systems," *International Journal of Control* **57**(1993), pp. 537-556.

[12] M. Chaves, "A parameter–robust observer as an application of ISS Techniques," in *Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems (MTNS'02)*, South Bend, IN, August 2002.

[13] M. Chaves and E.D. Sontag, "An alternative observer for zero deficiency chemical networks," in *Proceedings of the IFAC Non-Linear Control Systems Design Symposium (NOLCOS'01)*, St. Petersburg, Russia, June 2001, pp. 575-578.

[14] M. Chaves and E.D. Sontag, "Observers for chemical reaction networks," in *Proceedings of the European Control Conference (ECC'01)*, Porto, Portugal, September 2001, pp. 3715-3720.

[15] M. Chaves and E.D. Sontag, "State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero-deficiency type," *European Journal of Control* **8**(2002), pp. 343-359.

[16] C. Cohen-Tannoudji, B. Diu and F. Laloe, *Quantum Mechanics*, Vol. I, J. Wiley & Sons, New York, 1977.

[17] D. Dochain, E. Buyl, and G. Bastin, "Experimental validation of a methodology for on-line state estimation in bioreactors," in *Computer Applications in Fermentation Technology: Modelling and Control of Biotechnological Processes* (N.M. Fish, R.I. Fox, and N.F. Thornhill, eds.), Elsevier, Amsterdam, 1988, pp. 187-194.

[18] J. Fabrice, "Dynamics and Robust Nonlinear PI Control of Stirred Tank Reactors" Thèse de Docteur en Sciences Appliquées, Université Catholique de Louvain, Belgique, 1996.

[19] M. Feinberg, "Chemical reaction network structure and the stabiliy of complex isothermal reactors - I. The deficiency zero and deficiency one theorems," Review Article 25, *Chemical Engineering Science* **42**(1987), pp. 2229-2268.

[20] M. Feinberg, "The existence and uniqueness of steady states for a class of chemical reaction networks," *Archive for Rational Mechanics and Analysis* **132**(1995), pp. 311-370.

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

[22] R. Feynman, R.B. Leigthon and M.L. Sands, *The Feynamn Lectures on Physics*, Vol. II, Addison-Wesley, 1977.

[23] J.P. Gauthier, H. Hammouri and S. Othman "A simple observer for nonlinear systems, applications to bioreactors," *IEEE Transactions on Automatic Control*, **37**(1992), pp. 875-880.

[24] J.P. Gauthier and I.A.K. Kupka, "Observability and observers for nonlinear systems," *SIAM Journal on Control and Optimization*, **32**(1994), pp. 975-994.

[25] P. Hartman, *Ordinary Differential Equations*, John Wiley & Sons, Inc., New York, 1964.

[26] R. Hermann and A.J. Krener, "Nonlinear controllability and observability," *IEEE Transactions on Automatic Control*, **22**(1977), pp. 728-740.

[27] F.J.M. Horn, and R. Jackson, "General mass action kinetics," *Archive for Rational Mechanics and Analysis* **49**(1972), pp. 81-116.

[28] F.J.M. Horn, "The dynamics of open reaction systems," in *Mathematical aspects of chemical and biochemical problems and quantum chemistry (Proceedings of the SIAM-AMS Symposium in Applied Mathematics), New York, 1974.* Providence, RI: American Mathematical Society, 1974, vol. VIII, SIAM-AMS Proceedings, pp. 125-137.

[29] A. Ilchmann and M.F. Weirig, "Modelling of general biotechnological processes," *Mathematical and Computer Modelling of Dynamical Systems* **5**(1999), pp. 152-178.

[30] A. Isidori, *Nonlinear Control Systems*, Third Edition, Springer-Verlag, New York, 1995.

[31] A.J. Krener, "Nonlinear stabilizability and detectability," in *Systems and Networks: Mathematical Theory and Applications*, (U. Helmke, R. Mennicken and J. Saurer, eds.), Akademie Verlag, Berlin, 1994, pp. 231-250.

[32] A.J. Krener and A. Isidori, "Linearization by output injection and nonlinear observers," *Systems & Control Letters* **3**(1983), pp. 47-52.

[33] A.J. Krener, and W. Respondek, "Nonlinear observers with linearizable error dynamics," *SIAM Journal on Control and Optimization* , **23**(1985), pp. 198-216.

[34] M. Krichman, E.D. Sontag, and Y. Wang, "Input-output-to-state stability," *SIAM Journal on Control and Optimization*, **39**(2001), pp. 1874-1928.

[35] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[36] D.A. Lauffenburger and J.J. Linderman, *Receptors: Models for Binding, Trafficking, and Signaling*, Oxford University Press, New York, 1993.

[37] Y. Lin, E.D. Sontag and Y. Wang, "A smooth converse Lyapunov theorem for robust stability," *SIAM Journal on Control and Optimization*, **34**(1996), pp. 124-160.

[38] T.W. McKeithan, "Kinetic proofreading in T-cell receptor signal transduction," *Proceedings of the National Academy of Sciences of the USA* **92**(1995), pp. 5042-5046.

[39] N.M. Pous, A. Rajab, A. Flaus, J.M. Engasser and A. Cheruy "Comparison of estimation methods for biotechnological processes," *Chemical Engineering Science* **43**(1988), pp. 1909-1914.

[40] L.D. Shea, R.R. Neubig, J.J. Linderman, "Timing is everything: the role of kinetics in G protein activation," *Life Sciences*, **68**(2000), pp. 647-658.

[41] D. Siegel and D. MacLean, "Global stability of complex balanced mechanisms," preprint.

[42] E.D. Sontag, "On the observability of polynomial systems, I: Finite–time problems," *SIAM Journal on Control and Optimization*, **17**(1979), pp. 139-151.

[43] E.D. Sontag, "A concept of local observability," *Systems & Control Letters* **5**(1984), pp. 41-47.

[44] E.D. Sontag, "Smooth stabilization implies coprime factorization," *IEEE Transactions on Automatic Control*, **34**(1989), pp. 435-443.

[45] E.D. Sontag, "Remarks on stabilization and input-to-state stability," *Proceedings of the IEEE Conference in Decision and Control, Tampa, Dec. 1989*, IEEE Publications, Dec. 1989, pp. 1376-1378.

[46] E.D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Second Edition, Springer-Verlag, New York, 1998.

[47] E.D. Sontag, "The ISS philosophy as a unifying framework for stability-like behavior," in *Proceedings of the Second Nonlinear Control Network (NCN) Workshop: Nonlinear Control in the Year 2000* , Paris, France, June 2000.

[48] E.D. Sontag, "Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction," *IEEE Transactions on Automatic Control*, **46**(2001), pp. 1028-1047. Errata in *IEEE Transactions on Automatic Control*, **47**(2002), pp. 705.

[49] E.D. Sontag and Y. Wang, "On characterizations of the input-to-state stability property," *Systems & Control letters* **24**(1995), pp. 351-359.

[50] E.D. Sontag and Y. Wang, "On characterizations of input-to-state stability with respect to compact sets," in *Proceedings of the IFAC Non-Linear Control Systems Design Symposium, (NOLCOS '95)*, Tahoe City, CA, June 1995, pp. 226-231.

[51] E.D. Sontag and Y. Wang, "Output-to-state stability and detectability of nonlinear systems," *Systems & Control Letters* **29**(1997), pp. 279-290.

[52] E.D. Sontag and Y. Wang, "Notions of input to output stability," *Systems & Control Letters* **38**(1999), pp. 235-248.

[53] H. Sussmann, "Single-input observability of continuous-time systems," *Mathematical Systems Theory* **12**(1979), pp. 371-393.

[54] J. Tsinias, "Further results on the observer design problem," *Systems & Control Letters* **14**(1990), pp. 411-418.

[55] P.J. Woolf, T.P. Kenakin and J.J. Linderman, "Uncovering biases in high throughput screens of G-protein coupled receptors," *Journal of Theoretical Biology* **208**(2001), pp. 403-418.

[56] X.H. Xia and W.B. Gao, "Nonlinear observer design by observer error linearization," *SIAM Journal on Control and Optimization*, **27**(1989), pp. 199-216.

[57] X.H. Xia and M. Zeitz, "On nonlinear continuous observers," *International Journal of Control* **66**(1997), pp. 943-954.

Vita

Madalena C. Chaves

1990-95 Attended Instituto Superior Técnico, Lisbon, Portugal. Majored in Physics.

1995 Licenciatura (BSc.), Instituto Superior Técnico.

1994-96 Teaching Assistant, Department of Physics, Instituto Superior Técnico, Lisbon, Portugal.

1995-96 Teaching Assistant, Department of Economics, Universidade Lusíada, Lisbon, Portugal.

1996-02 Graduate work in Mathematics, Rutgers, The State University of New Jersey, New Brunswick, New Jersey.

1997-01 Fellow, Fundação para a Ciência e a Tecnologia (Portugal).

2001-03 Teaching Assistant, and Graduate Assistant, Department of Mathematics.

2002 Fellow, Fundação Calouste Gulbenkian.

2001 Chaves, M., and E.D. Sontag, An alternative observer for zero deficiency chemical networks, in *Proceedings of the IFAC Non-Linear Control Systems Design Symposium (NOLCOS'01)*, St. Petersburg, Russia, June 2001, pp. 575-578.

2001 Chaves, M., and E.D. Sontag, Observers for chemical reaction networks, in *Proceedings of the European Control Conference (ECC'01)*, Porto, Portugal, September 2001, pp. 3715-3720.

2002 Chaves, M., and E.D. Sontag, State-estimators for chemical reaction networks of Feinberg-Horn-Jackson zero-deficiency type. *European Journal of Control*, **8**, pp. 343-359.

2002 Chaves, M., A parameter–robust observer as an application of ISS Techniques, in *Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems (MTNS'02)*, South Bend, IN, August 2002.