Lecture Notes in Control and Information Sciences 473

Nicolas Petit *Editor*

Feedback Stabilization of Controlled Dynamical Systems

In Honor of Laurent Praly

LNCIS

Springer

Lecture Notes in Control and Information Sciences

Volume 473

About this Series

This series aims to report new developments in the fields of control and information sciences—quickly, informally and at a high level. The type of material considered for publication includes:

1. Preliminary drafts of monographs and advanced textbooks
2. Lectures on a new field, or presenting a new angle on a classical field
3. Research reports
4. Reports of meetings, provided they are
    (a) of exceptional interest and
    (b) devoted to a specific topic. The timeliness of subject material is very important.

More information about this series at http://www.springer.com/series/642

Nicolas Petit
Editor

Feedback Stabilization of Controlled Dynamical Systems

In Honor of Laurent Praly

Springer

Editor
Nicolas Petit
Centre Automatique et Systèmes (CAS)
MINES ParisTech, PSL Research University
Paris Cedex 06
France

ISSN 0170-8643 ISSN 1610-7411 (electronic)
Lecture Notes in Control and Information Sciences
ISBN 978-3-319-51297-6 ISBN 978-3-319-51298-3 (eBook)
DOI 10.1007/978-3-319-51298-3

Library of Congress Control Number: 2017931053

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

During the 20 years that I have spent working in the close vicinity of Laurent Praly at the Centre Automatique et Systèmes, first as a student, then as colleague, and eventually as his boss, I feel that I have come to understand his philosophy.

Throughout the years, which have recently seen the emergence of a strong trend in project-oriented research and fashionable topics, the style of Laurent's scientific work has remained unchanged. You do not change something that works. While this might appear to explain such consistency reasonably well, the underlying rationale is much more sound.

Laurent belongs to the long tradition of French engineers with a solid background in mathematics. To his colleagues and friends, Laurent's approach to problems is uniquely his own, and thus it is like a signature: he considers the possible usefulness of the intended results, finds obvious counterexamples that would discourage most people from trying to prove the general properties, reformulates the whole question in the most concise and clear manner, and eventually delivers the mathematical proof. This process does not take him days or weeks, but rather it takes him tens of minutes. The rest of his time is simply spent in the careful intellectual construction of the perfect proof: the shortest, with the fewest assumptions and the least elaborate arguments.

This book celebrates the outstanding career of Laurent, which some people may not be aware began in the field of industrial research and development during the early days of adaptive control and model predictive control at the ADERSA company headed by J. Richalet. This rich journey surely explains how he has become such an iconic ideas-man, puzzle-solver, and pusher of frontiers, as is often heard about Laurent.

Together with Laurent, who is certainly too modest to put his name to such a statement, I hope that this book will encourage young generations, as they begin their careers in academia or in industry, to work toward long-term goals. Indeed, in my own experience, having finally isolated the hard theoretical question at the core of an industrial problem after months of difficult work, I usually found that Laurent had already worked on it and he could give me several solutions, ranging from (in his own words) "totally useless" to "stupid", or sometimes "possibly working", In fact, the truth was usually much more than that, and Laurent's was the definitive answer to the problem.

The works and achievements of Laurent Praly should encourage researchers to address difficult fundamental problems, while always remaining open to new approaches, emerging ideas, and new people. Laurent Praly has given much help to many individuals throughout his career. It is a great honor for his colleagues at the Centre Automatique et Systèmes at MINES ParisTech to see contributions from these individuals in this book.

Paris Nicolas Petit
March 2017

Synopsis

This book is a tribute to Prof. Laurent Praly on the occasion of his sixtieth birthday.

Laurent Praly has been an active member of the control community for over 35 years. Throughout his sustained and influential scientific career, he has developed several breakthrough results and contributed towards the foundation of nonlinear control theory.

His full commitment to solve problems of true engineering value with advanced tools that he and his co-workers have developed and his dedication to teach this material at various levels has served as an enlightening example for a generation of colleagues. His lectures, at numerous international workshops and schools, have inspired a great number of students and researchers. His role as mentor to aspiring researchers has nurtured numerous junior researchers and helped many graduate students.

The volume collects contributions written by prominent individuals of the control community. Each author in this list was chosen among researchers who have worked with Laurent Praly, shared ideas or have had the honor of being his Ph.D. students. The contributions presented in this volume address a rich collection of topics: emerging theories, advanced applications, and theoretical concepts. The diversity of the areas covered provides another evidence of the global impact of Laurent Praly in our community.

The reader will find renewed and unified visions on the numerous problems that Prof. Laurent Praly has been working on in his prolific carrier: adaptive control, output feedback and observers, stability, and stabilization. His main contributions are the central points of this book.

The book is organized in three sections. The first section covers the field of adaptive control where Laurent Praly started his carrier. The second section gathers contributions on stabilization and output feedback, which is the topic of the second half of Laurent Praly's carrier. Finally, the third section presents emerging research built on Laurent Praly's scientific legacy.

Nicolas Petit

Contents

Part I Adaptive Control

Part II Stabilization and Output Feedback

Part III New Perspectives

Part I
Adaptive Control

Chapter 1
Lyapunov Functions Obtained from First Order Approximations

Vincent Andrieu

Abstract We study the construction of Lyapunov functions based on first order approximations. We first consider the (transverse) local exponential stability of an invariant manifold and largely rephrase [3]. We show how to construct a Lyapunov function with this framework that characterizes this local stability property. We then consider global stability of an equilibrium point, and show that the first order approximation along solutions of the system allows to construct a global Lyapunov function. This result can be regarded as a new inverse Lyapunov theorem arising from Riemannian metric.

Notation:

- For a vector in $\mathbb{R}^n$ and a matrix in $\mathbb{R}^{n\times n}$, $|\cdot|$ means the usual 2 norm.
- For a positive definite matrix, P, $\mu_{\max}\{P\}$ and $\mu_{\min}\{P\}$ are, respectively, the largest and smallest eigenvalues.

1.1 Introduction

The use of Lyapunov functions in the study of the stability of solutions or invariant sets of dynamical systems has a long history. It can be traced back to Lyapunov himself who has introduced this concept in his dissertation in 1892 (see [18] for an English translation). The primary objective of a Lyapunov function is to analyze the behavior of trajectories of dynamical systems and how this behavior is preserved after perturbations. However, this tool is also very efficient to synthesize control algorithms, such as stabilizing control laws, regulators, and asymptotic observers (see for example [14, 15, 21, 26]).

V. Andrieu (✉)
CNRS UMR 5007 LAGEP, Université Lyon 1, Lyon, France
e-mail: vincent.andrieu@gmail.com

V. Andrieu
Fachbereich C - Mathematik und Naturwissenschaften,
Bergische Universität Wuppertal, Wuppertal, Germany

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_1

Hence, the study of converse Lyapunov theorems have received a considerable attention from the nonlinear control community. One of the first major contributions to the problem of a Lyapunov function existence can be attributed to Massera [19]. This result has been improved over the years (see for example [16, 20]) and Teel-Praly who established an existence theorem for a Lyapunov functions in a very general framework in [27]. However, despite the development of a very complete theory to infer Lyapunov function existence, its construction in practice appears to be a very difficult task.

On another hand, using a first order approximation to analyze the local stability of the origin of a nonlinear system is the most commonly used approach. Indeed, a first order analysis deals intrinsically with linear systems tools and it provides a simple way to construct local Lyapunov functions for a nonlinear system.

In this note, the *linearization approach* is extended in two directions. The first extension is the case in which the stability studied concerns a simple manifold rather than an equilibrium. This extension was first published by [1, 3] and we briefly rephrase these results. The second extension is to show that when dealing with equilibrium points, a global property may be characterized from first order approximations *along solutions*.

In order to introduce these results and aiming at allowing to get a full grip on the key points of the approach the following simpler framework is first considered and some very classical results are rephrased in the following paragraph.

Consider a nonlinear dynamical system defined on $\mathbb{R}^{n_e}$ with the origin as equilibrium.

$$\dot{e} = F(e)\ ,\ F(0) = 0\ , \tag{1.1}$$

with state e in $\mathbb{R}^{n_e}$, and a C^1 vector field $F : \mathbb{R}^{n_e} \to \mathbb{R}^{n_e}$. Solutions initiated from e in $\mathbb{R}^{n_e}$ evaluated at time t are denoted $E(e,t)$.

The origin of system (1.1) is said to be locally exponentially stable (LES) if there exist three positive real numbers k, λ, and r such that

$$|E(e,t)| \leq k \exp(-\lambda t)|e|\ ,\ \forall (e,t) \in \mathbb{R}^{n_e} \times \mathbb{R}_{\geq 0}\ ,\ |e| \leq r\ . \tag{1.2}$$

As well-known, LES of the origin of system (1.1) can be checked from the first order approximation around “0”. Indeed, it is well-known (see [15, Theorem 4.15, p. 165]) that LES of the origin of (1.1) is equivalent with exponential stability of the origin of the linear dynamical system defined in $\mathbb{R}^{n_e}$ as

$$\dot{\tilde{e}} = \frac{\partial F}{\partial e}(0)\tilde{e}\ . \tag{1.3}$$

Constructing a Lyapunov function for the linear system (1.3) is relatively simple. If $\frac{\partial F}{\partial e}(0)$ is Hurwitz, and given a positive definite matrix Q in $\mathbb{R}^{n_e \times n_e}$, then the matrix P in $\mathbb{R}^{n_e \times n_e}$,

$$P = \int_0^{+\infty} \exp\left(\frac{\partial F}{\partial e}(0)s\right)^\top Q \exp\left(\frac{\partial F}{\partial e}(0)s\right) ds\,, \tag{1.4}$$

is well defined, positive definite, and satisfies the Lyapunov algebraic equality

$$\frac{\partial F}{\partial e}(0)^\top P + P\frac{\partial F}{\partial e}(0) = -Q. \tag{1.5}$$

Equation (1.5) implies that the mapping $\widetilde{e} \mapsto \widetilde{e}^\top P\widetilde{e}$ is a Lyapunov function for system (1.3), since it yields along its trajectories $\dot{\overbrace{\widetilde{e}^\top P\widetilde{e}}} = -\widetilde{e}^\top Q\widetilde{e}.$.

Moreover, the quadratic function $V(e) = e^\top Pe$ is also a Lyapunov function for the nonlinear system (1.1) since

$$\dot{\overbrace{e^\top Pe}} = 2e^\top PF(e) = e^\top \left[-Q + 2\int_0^1 P \underbrace{\left[\frac{\partial F}{\partial e}(se) - \frac{\partial F}{\partial e}(0)\right]}_{\text{small if } |e| \text{ small}} ds\right] e\,,$$

holds along its trajectories. This implies that there exists $r > 0$ and $\lambda > 0$ such that, for all e such that $|e| \leq r$, $\dot{\overbrace{e^\top Pe}} < -\lambda e^\top Pe.$. This characterizes local exponential stability of the origin of (1.1).

In rephrasing of the simplest framework, the following assertions have been obtained:

Assertion 1 Exponential stability for the origin of the nonlinear system implies exponential stability for the origin of the linearized system.

Assertion 2 Exponential stability of the origin of the linearized system can be characterized by a quadratic Lyapunov function.

Assertion 3 The Lyapunov function associated with the linearized system may be used directly on the origin of the nonlinear system to characterize its stability.

In Sect. 1.2, we show that this is also the case when considering exponential stability of a simple invariant manifold, based on [3]. This allows the introduction of a Lyapunov function that characterizes the local exponential stability of an invariant manifold.

Section 1.3 considers global properties. We show that these three assertions also hold when considering the global attractivity of an equilibrium. Finally, in Sect. 1.4, we discuss some difficulties faced when considering the global stability of an invariant manifold. This gives a glimpse of the results obtained in [2].

1.2 Local Transverse Exponential Stability of a Manifold

1.2.1 Transverse Local Uniform Exponential Stability

Throughout this section, rather than (1.1), we consider a system.

$$\dot{e} = F(e, x) \ , \ \dot{x} = G(e, x) \ , \ F(0, x) = 0 \ , \tag{1.6}$$

where e is in $\mathbb{R}^{n_e}$, x is in $\mathbb{R}^{n_x}$ and the functions $F : \mathbb{R}^{n_e} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_e}$ and $G : \mathbb{R}^{n_e} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ are C^2. We denote $(E(e, x, t), X(e, x, t))$ as the (unique) solution which goes through (e, x) in $\mathbb{R}^{n_e} \times \mathbb{R}^{n_x}$ at time $t = 0$. It is assumed that these solutions are defined for all positive times, i.e., the system is forward complete.

For this system, $\mathscr{E} = \{(e, x), e = 0\}$ is an invariant manifold. The purpose of this section is to show that the properties that characterize the exponential stability of an equilibrium as discussed in Sect. 1.1 (i.e., Assertions 1–3) remain valid when considering the stability of this manifold.

The local exponential stability of an equilibrium becomes the local exponential stability of the transverse manifold.

Definition 1.1 (*Transversal uniform local exponential stability (TULES-NL)*) System (1.6) is forward complete and there exist strictly positive real numbers r, k, and λ such that, for all (e_0, x_0, t) in $\mathbb{R}^{n_e} \times \mathbb{R}^{n_x} \times \mathbb{R}_{\geq 0}$ with $|e_0| \leq r$,

$$|E(e_0, x_0, t)| \leq k|e_0| \exp(-\lambda t) \ . \tag{1.7}$$

That is, System (1.6) is said to be TULES-NL if manifold $\mathscr{E} := \{(e, x) : e = 0\}$ is exponentially stable for system (1.6), locally in e and uniformly in x.

1.2.2 Assertion 1: Exponential Stability of a Linearized System

As discussed in Sect. 1.1, a linearized system around the invariant manifold must first be considered. In this case, the system is defined as

$$\dot{\widetilde{e}} = \frac{\partial F}{\partial e}(x)\widetilde{e} \ , \ \dot{x} = \tilde{G}(x) \ , \tag{1.8}$$

where $\tilde{G}(x) = G(0, x)$.

If we wish to show that Assertion 1 also holds in this context, we need to establish that manifold $\tilde{\mathscr{E}} := \{(x, \widetilde{e}) : \widetilde{e} = 0\}$ is exponentially stable for the linearized system transverse to $\mathscr{E}$ (i.e., System (1.8)). The following proposition, proved in [3], shows this is indeed the case.

Proposition 1.1 ([3] Assertion 1 holds) *If TULES-NL holds and there exist positive real numbers ρ, μ, and c such that, for all x in $\mathbb{R}^{n_x}$,*

$$\left|\frac{\partial F}{\partial e}(0,x)\right| \leq \mu \ , \ \left|\frac{\partial G}{\partial x}(0,x)\right| \leq \rho \tag{1.9}$$

and, for all (e,x) in $B_e(kr) \times \mathbb{R}^{n_x}$,

$$\left|\frac{\partial^2 F}{\partial e \partial e}(e,x)\right| \leq c \ , \ \left|\frac{\partial^2 F}{\partial x \partial e}(e,x)\right| \leq c \ , \ \left|\frac{\partial G}{\partial e}(e,x)\right| \leq c \ , \tag{1.10}$$

then System (1.8) is forward complete and there exist strictly positive real numbers $\widetilde{k}$ and $\tilde{\lambda}$, such that any solution $(\widetilde{E}(\widetilde{e}_0, x_0, t), X(x_0, t))$ of the transversally linear system (1.8) satisfies, for all $(\widetilde{e}_0, x_0, t)$ in $\mathbb{R}^{n_e} \times \mathbb{R}^{n_x} \times \mathbb{R}_{\geq 0}$,

$$|\widetilde{E}(\widetilde{e}_0, x, t)| \leq \widetilde{k} \exp(-\tilde{\lambda} t) |\widetilde{e}_0| \ . \tag{1.11}$$

The proof of this proposition given in [3] is based on the comparison between a given e-component of a solution $\widetilde{E}(\widetilde{e}_0, x_0, t)$ of (1.8) with pieces of e-component of solutions $E(\widetilde{e}_i, x_i, t - t_i)$ of solutions of (1.6) where $\widetilde{e}_i, x_i$ are sequences of points in $\{(\widetilde{E}(\widetilde{e}_0, x_0, t), X(x_0, t)), t \in \mathbb{R}_{\geq 0}\}$. Thanks to the bounds (1.9) and (1.10), it is possible to show that $\widetilde{E}$ and E remain sufficiently closed so that $\widetilde{E}$ inherits the convergence property of the solution E.

In [3], the exponential stability of manifold $\tilde{\mathscr{E}} := \{(x, \widetilde{e}) : \widetilde{e} = 0\}$ of the linearized system transversal to $\mathscr{E}$ in (1.8) is called UES-TL.

1.2.3 Lyapunov Matrix Inequality

The $\widetilde{e}$ components of System (1.8) is a parametrized time varying linear system. Hence, the solutions $\widetilde{E}(e, x, t)$, can be expressed

$$\widetilde{E}(\widetilde{e}, x, t) = \boldsymbol{\Phi}(x, t)\widetilde{e} \ ,$$

where $\boldsymbol{\Phi}$ is the transition matrix defined as a solution to the $\mathbb{R}^{n_e \times n_e}$ dynamical system

$$\dot{\overbrace{\boldsymbol{\Phi}(x,t)}} = \frac{\partial F}{\partial e}(0, \widetilde{X}(\widetilde{x}, t))\boldsymbol{\Phi}(\widetilde{x}, t) \ , \ \boldsymbol{\Phi}(\widetilde{x}, 0) = I \ .$$

An important point that has to be noticed is that, due to Eq. (1.11), each element of the (matrix) time function $t \mapsto \boldsymbol{\Phi}(x, t)$ is in $L^2([0, +\infty))$. Consequently, for all positive definite matrices Q in $\mathbb{R}^{n_e}$, the matrix function

$$P(x)=\lim_{T\to+\infty}\int_0^T \Phi(x,s)^\top Q\Phi(x,s)ds \tag{1.12}$$

is well defined.

By computing the Lie derivative of the matrix P given in (1.12), it is possible to show that this one satisfies a particular partial differential equation which shows that this function may be used to construct a quadratic Lyapunov function of the linearized system.

Proposition 1.2 ([3] Assertion 2 holds) *Assume UES-TL holds, i.e., there exist* $\widetilde{k}$ *and* $\tilde{\lambda}$ *such that any solution* $(\widetilde{E}(\widetilde{e}_0, x_0, t), X(x_0, t))$ *of the transversally linear system (1.8) satisfies, (1.11). Assume also that there exists a positive real number* μ *such that*

$$\left|\frac{\partial F}{\partial e}(0,x)\right| \leq \mu \qquad \forall x \in \mathbb{R}^{n_x}\ . \tag{1.13}$$

Then for all positive definite matrices Q*, there exists a continuous function* $P: \mathbb{R}^{n_x} \to \mathbb{R}^{n_e \times n_e}$ *and strictly positive real numbers* $\underline{p}$ *and* $\overline{p}$ *such that* P *has a derivative* $\mathfrak{d}_{\tilde{G}}P$ *along* $\tilde{G}$ *in the sense*

$$\mathfrak{d}_{\tilde{G}}P(\widetilde{x}) \ := \ \lim_{h\to 0}\frac{P(\widetilde{X}(\widetilde{x},h)) - P(\widetilde{x})}{h}\ , \tag{1.14}$$

and, for all $\widetilde{x}$ *in* $\mathbb{R}^{n_x}$,

$$\mathfrak{d}_{\tilde{G}}P(\widetilde{x}) + P(\widetilde{x})\frac{\partial F}{\partial e}(0,\widetilde{x}) + \frac{\partial F}{\partial e}(0,\widetilde{x})'P(\widetilde{x}) \leq -Q\ , \tag{1.15}$$

$$\underline{p}\,I \leq P(\widetilde{x}) \leq \overline{p}\,I\ . \tag{1.16}$$

The time derivative of $(\widetilde{e}, x) \mapsto \widetilde{e}^\top P(x)\widetilde{e}$ along the solution of system (1.8) yields

$$\dot{\overbrace{\widetilde{e}^\top P(x)\widetilde{e}}} = -\widetilde{e}^\top Q\widetilde{e}\ .$$

Hence, $(\widetilde{e}, x) \mapsto \widetilde{e}^\top P(x)\widetilde{e}$ is a Lyapunov function for the $\widetilde{e}$ component of the linearized system (1.8). In other words, Assertion 2 remains valid when considering transverse exponential stability.

Assumption (1.13) is used to show that P satisfies the left inequality in (1.16). Nevertheless this inequality holds without (1.13) provided the function $s \mapsto \left|\frac{\partial \widetilde{E}}{\partial \widetilde{e}}(0,\widetilde{x},s)\right|$ does not go too fast to zero.

1.2.4 Construction of a Lyapunov Function

We may define a Lyapunov function that characterizes local exponential stability of $\mathscr{E}$ from P obtained above.

Proposition 1.3 ([3] Assertion 3 holds) *If ULMTE holds and there exist positive real numbers η and c such that, for all (e,x) in $B_e(\eta)\times\mathbb{R}^{n_x}$,*

$$\left|\frac{\partial P}{\partial x}(x)\right| \leq c\ , \tag{1.17}$$

$$\left|\frac{\partial^2 F}{\partial e \partial e}(e,x)\right| \leq c\ ,\ \left|\frac{\partial^2 F}{\partial x \partial e}(e,x)\right| \leq c\ ,\ \left|\frac{\partial G}{\partial e}(e,x)\right| \leq c\,, \tag{1.18}$$

then TULES-NL holds.

This is a direct consequence of $V(e,x) = e'P(x)e$ being a Lyapunov function. Bounds (1.17) and (1.18) are used to show that, with Eq. (1.15), the time derivative of this Lyapunov function is negative in a (uniform) tubular neighborhood of manifold $\{(e,x), e=0\}$.

In conclusion, Propositions 1.1, 1.2 and 1.3 show that Assertions 1–3, obtained in the analysis of local exponential stability of an equilibrium remain valid in the context of local exponential stability of a transverse manifold. In [3] the previous framework has been employed as a design tool in different contexts:

- To construct a Lyapunov function that characterizes exponential incremental stability.
- To show that a detectability property introduced in [24] is a necessary condition for the existence of an exponential full order observer.
- To derive necessary and sufficient conditions to achieve synchronization (see also in [4, 5]).

All results so far concern local properties. The following section considers global properties of an equilibrium point. Similar strategy allows to construct global Lyapunov functions.

1.3 Global Stability Properties

1.3.1 Local Exponential Stability and Global Attractivity

In Sect. 1.2 we studied the case of local asymptotic stability of a manifold or an equilibrium point. In this Section, another property is assumed: the global attractivity of the origin. Consider again system (1.1), and assume that for all e in $\mathbb{R}^{n_e}$,

$$\lim_{t\to+\infty} |E(e,t)| = 0\ . \tag{1.19}$$

Global attractivity in combination with local asymptotic stability of the origin implies that the system is globally and asymptotically stable. However, it is not globally exponentially stable in the usual sense (see [15, Definition 4.5 p. 150]) . Nevertheless, the following property can be obtained.

Proposition 1.4 *Assume the origin of (1.1) is LES and globally attractive. Then there exist a positive real number λ and a continuous strictly increasing function $k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, such that*

$$|E(e,t)| \leq k(|e|)\exp(-\lambda t)|e| \ . \tag{1.20}$$

Proof The origin being LES, there exist three positive real numbers λ_1, k_1 and r_1 such that equality (1.2) holds. Consider the mapping $c : \mathbb{R}^{n_e} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by

$$c(e,t) = \frac{|E(e,t)|}{|e|\exp(-\lambda t)} \ ,$$

where $0 < \lambda < \lambda_1$. Since inequality (1.2) holds, this is a continuous function. Moreover, the global attractivity and the LES of the origin of system (1.1) implies that

$$\lim_{t \to +\infty} c(e,t) = 0 \ , \ \forall e \in \mathbb{R}^{n_e} \ .$$

Consider the function $\bar{c} : [r_1, +\infty) \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$, defined as

$$\bar{c}(s) = \sup_{r_1 \leq |e| \leq s, t \geq 0} \{c(e,t)\} \ .$$

We first show this function takes finite values for all s. Assume this is not the case for a given s, i.e., $\bar{c}(s) = +\infty$. This implies that there exists a sequence $(e_i, t_i)_{i \in \mathbb{N}}$ with $r_1 \leq |e_i| \leq s$ such that $c(e_i, t_i) \geq i$. However, since $(e_i)_{i \in \mathbb{N}}$ is a sequence in a compact set, it is possible to extract a sub-sequence $(e_{i_j})_{j \in \mathbb{N}}$ such that $e_{i_j} \to e^*$ with $r_1 \leq |e^*| \leq s$, which implies $t_{i_j} \to +\infty$. Moreover, from global attractivity of the origin, there exists t^* such that $|E(e^*, t^*)| \leq \frac{r_1}{2}$. Continuity of the solutions implies that there exists j^* such that $|E(e_{i_j}, t^*)| \leq r_1$ for $j > j^*$. Without loss of generality, we may assume that $t_{i_j} \geq t^*$ for $j > j^*$, and LES of the origin implies for all $j > j^*$,

$$i_j < c(e_{i_j}, t_{i_j}) = \frac{|E(e_{i_j}, t_{i_j})|}{|e_{i_j}|\exp(-\lambda t)} \leq \frac{k_1 \exp\left(-\lambda_1(t_{i_j} - t^*)\right)|E(e_{i_j}, t^*)|}{|e_{i_j}|\exp(-\lambda t_{i_j})} \leq \frac{ks\exp(\lambda_1 t^*)}{r_1} \ .$$

Hence, we have a contradiction, and so, all $s \geq r_1$, $\bar{c}(s)$ is bounded and increasing. Therefore it is possible to select $k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ as any continuous function such that

$$k(s) \geq \begin{cases} k_1 & s \leq r_1 \\ \bar{c}(s) & s \geq r_1 \end{cases} \ .$$

It is clear from its definition that (1.20) is satisfied. □

Example 1.1 A very simple example of such a property is the scalar system:

$$\dot{e} = -\frac{e}{1+e^2} \; . \tag{1.21}$$

Solutions of this ordinary differential equation satisfy

$$E(e,t)^2 \exp\left(E(e,t)^2\right) = e^2 \exp(e^2)\exp(-2t) \; , \; \forall e \in \mathbb{R} \; .$$

Which implies

$$E(e,t)^2 \leq E(e,t)^2 \exp\left(E(e,t)^2\right) \leq e^2 \exp(e^2)\exp(-2t) \; ,$$

and global attractivity and LES of the origin of (1.21) hold since (1.20) is satisfied with $k(s) = s\exp\left(\frac{1}{2}s^2\right)$ and $\lambda = 1$.

1.3.2 Global Lyapunov Functions Based on First Order Approximations

1.3.2.1 Assertion 1: Stability of the Origin of the Linearized System Along the Solutions

A natural question is whether if LES and global attractivity of the origin can be characterized from a first order approximation. In contrast to the local study of Sect. 1.1, the linearized system around the equilibrium cannot describe solutions away from the origin. Hence, the linearized system along all solutions must be considered.

Assuming that F is C^1 everywhere, the linearized system along trajectories is

$$\dot{\widetilde{e}} = \frac{\partial F}{\partial e}(e)\widetilde{e} \; , \; \dot{e} = F(e) \; , \tag{1.22}$$

with $(e, \widetilde{e})$ in $\mathbb{R}^{n_e} \times \mathbb{R}^{n_e}$. This system is also called the *lifted* system in [10] or the *variational system* in [9].

The $\widetilde{e}$-components of this system may be expressed as

$$\underbrace{\dot{\widetilde{e}} = \frac{\partial F}{\partial e}(0)\widetilde{e}}_{\text{(LES)}\Rightarrow\text{goes exp. to zero}} \quad + \quad \underbrace{\left[\frac{\partial F}{\partial e}(e) - \frac{\partial F}{\partial e}(0)\right]}_{\text{(Glob. Attract.)}\Rightarrow\text{ goes to zero}} \; \widetilde{e} \tag{1.23}$$

The following proposition shows that if the e components go exponentially to zero, then so do the $\widetilde{e}$ components.

Proposition 1.5 (Assertion 1 globally) *Let F be C^1 in $\mathbb{R}^{n_e}$ and C^2 around the origin. Assume the origin of (1.1) is locally exponentially stable and globally attractive, then* *there exist a positive real number* $\widetilde{\lambda}$ *and a strictly increasing function* $\widetilde{k} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$*, such that*

$$|\widetilde{E}(e,t)| \leq \widetilde{k}(|e|)\exp(-\widetilde{\lambda}t)|\widetilde{e}| \,. \tag{1.24}$$

Proof Since the origin is locally exponentially stable, we can define P as in (1.4). From the algebraic Lyapunov equation (see (1.5)), the following equality holds

$$\begin{aligned}\dot{\overbrace{\widetilde{e}^\top P \widetilde{e}}} &= -\widetilde{e}^\top Q \widetilde{e} + 2\widetilde{e}^\top P \left[\frac{\partial F}{\partial e}(e) - \frac{\partial F}{\partial e}(0)\right]\widetilde{e}\,,\\ &\leq \left[-\frac{\mu_{\min}\{Q\}}{\mu_{\max}\{P\}} + \gamma(e)\right]\widetilde{e}^\top P \widetilde{e}\,,\end{aligned} \tag{1.25}$$

along the solution of system (1.1), where $\gamma : \mathbb{R}^{n_e} \to \mathbb{R}_{\geq 0}$ is the continuous function defined as

$$\gamma(e) = 2\frac{\mu_{\max}\{P\}}{\mu_{\min}\{P\}}\left|\frac{\partial F}{\partial e}(e) - \frac{\partial F}{\partial e}(0)\right|^2 \,.$$

Since F is C^2 around the origin, γ is locally Lipschitz around the origin. Hence, there exist two positive real number r and L such that

$$\gamma(e) \leq L|e| \,,\ \forall |e| \leq r \,. \tag{1.26}$$

From the Grönwall lemma, Eq. (1.25) implies

$$\begin{aligned}|\widetilde{E}(\widetilde{e},e,t)| &\leq \sqrt{\frac{\widetilde{E}(\widetilde{e},e,t)^\top P \widetilde{E}(\widetilde{e},e,t)}{\mu_{\min\{P\}}}}\,,\\ &\leq \sqrt{\frac{\mu_{\max}\{P\}}{\mu_{\min}\{P\}}}\exp\left(\frac{1}{2}\int_0^t \gamma(E(e,s))ds\right)\exp\left(-\frac{\mu_{\min}\{Q\}}{2\mu_{\max}\{P\}}t\right)|\widetilde{e}| \,.\end{aligned} \tag{1.27}$$

Let t^* be the continuous function defined as

$$t^*(e) = \max\left\{0, \frac{-\ln\left(\frac{r}{k(|e|)|e|}\right)}{\lambda}\right\} \,.$$

Note that if $k(|e|)|e| \leq r$, $t^*(e) = 0$. Moreover, if $k(|e|)|e| > r$, $t^*(e) > 0$,

$$k(|e|)\exp(-\lambda t^*(e))|e| \leq r \,.$$

Hence, due to local exponential stability and global attractivity property of the origin, (1.20) yields for all e,

$$|E\,(e, t^*(e))| \leq k(|e|)\exp\left(-\lambda t^*(e)\right)|e| \leq r\,.$$

Employing (1.20) again, and (1.26), the following inequalities are obtained for $t \geq t^*(e)$:

$$\begin{aligned}\int_0^t \gamma(E(e,s))ds &\leq \int_0^{t^*(e)} \gamma(E(e,s))ds + \int_{t^*(e)}^t \gamma(E(e,s))ds\,,\\ &\leq \int_0^{t^*(e)} \gamma(E(e,s))ds + Lk(|e|)|e| \int_{t^*(e)}^t \exp(-\lambda s)ds\,,\\ &\leq \int_0^{t^*(e)} \gamma(E(e,s))ds + \frac{Lr}{\lambda} := c(e)\,.\end{aligned}$$

This inequality is also true for $t \leq t^*(e)$. Consequently, using the previous approximation in (1.27) the proof ends since (1.24) is obtained with

$$\widetilde{\lambda} = \frac{\mu_{\min}\{Q\}}{2\mu_{\max}\{P\}}\,,\ \widetilde{k}(s) = \sqrt{\frac{\mu_{\max}\{P\}}{\mu_{\min}\{P\}}}\exp\left(\frac{1}{2}\max_{|e|\leq s} c(e)\right)\,.$$

□

Example 1.2 Returning to the previous example in (1.21), the linearized system is given as:

$$\dot{\widetilde{e}} = -\frac{1-e^2}{1+e^2}\,\widetilde{e}\,,$$

which gives

$$\begin{aligned}\left|\widetilde{E}(\widetilde{e}, e, t)\right| &= \exp\left(-t + \int_0^t \frac{2E(e,s)^2}{1+E(e,s)^2}ds\right)|\widetilde{e}|\,,\\ &\leq \exp\left(-t + \int_0^t 2k(|e|)^2\exp(-s)ds\right)|\widetilde{e}|\,,\\ &\leq \exp\left(-t + 2k(|e|)^2(1-\exp(-t))\right)|\widetilde{e}|\,.\end{aligned}$$

This gives (1.24) with

$$\widetilde{k}(s) = \exp\left(2k(s)^2\right) = \exp\left(2s^2\exp\left(s^2\right)\right)\,,\ \widetilde{\lambda} = 1\,.$$

1.3.2.2 Assertion 2: Lyapunov Matrix Inequality

From linearity the $\widetilde{e}$ components of linearized system (1.22) can be expressed as

$$\widetilde{E}(\widetilde{e}, e, t) = \Phi(e, t)\widetilde{e} ,$$

where Φ is the transition matrix, defined as the solution of the $\mathbb{R}^{n_e \times n_e}$ dynamical system

$$\overset{\cdot}{\overbrace{\Phi(e,t)}} = \frac{\partial F}{\partial e}(E(e,t))\Phi(e,t) \; , \; \Phi(e,0) = I \; .$$

An important point is that Eq. (1.24) means each element of the (matrix) time function $t \mapsto \Phi(e, t)$ is in $L^2([0, +\infty))$. Consequently, for all positive definite matrices Q in $\mathbb{R}^{n_e \times n_e}$,

$$P(e) = \lim_{T \to +\infty} \int_0^T \Phi(e, s)^\top Q \Phi(e, s) ds \; , \tag{1.28}$$

is well defined. Moreover, the following proposition holds.

Proposition 1.6 (Assertion 2 globally) *Assume that there exist a function* $(k, \widetilde{k})$ *and positive real numbers* $(\lambda, \underset{\sim}{\lambda})$ *such that (1.20) and (1.24) are satisfied. Then,* $P : \mathbb{R}^{n_e} \to \mathbb{R}^{n_e \times n_e}$ *defined in (1.28) is well defined, continuous, and there exist a nonincreasing function,* $\underline{p}$*, and a nondecreasing function,* $\bar{p}$*, such that*

$$0 < \underline{p}(|e|)I \le P(e) \le \bar{p}(|e|)I \; , \; \forall \; e \in \mathbb{R}^{n_e} \; . \tag{1.29}$$

Moreover,[1]

$$\underbrace{\mathfrak{d}_F P(e) + P(e)\frac{\partial F}{\partial e}(e) + \frac{\partial F}{\partial e}(e)^\top P(e)}_{=L_F P(e)} \le -Q \; , \; \forall \; e \in \mathbb{R}^{n_e} \; . \tag{1.30}$$

Finally, if F *is* C^3 *then* P *is* C^2*.*

Proof From (1.24), for all (e, t) in $\mathbb{R}^{n_e} \times \mathbb{R}_{\geq 0}$

$$|\Phi(e, t)| \le \widetilde{k}(|e|) \exp(-\tilde{\lambda} t) \; .$$

Thus, we may claim that, for all symmetric positive definite matrices, $Q, P : \mathbb{R}^{n_e} \to \mathbb{R}^{n_e \times n_e}$ from (1.28) is well defined, continuous, and satisfies

$$\mu_{\max}\{P(e)\} \le \frac{\widetilde{k}(|e|)^2}{2\tilde{\lambda}} \mu_{\max}\{Q\} = \overline{p}(|e|) \; , \; \forall e \in \mathbb{R}^{n_e} \; .$$

[1] See the notation (1.14).

Let c be a continuous mapping which satisfies

$$\left|\frac{\partial F}{\partial e}(e)\right| \leq c(|e|) \ .$$

Moreover, for all (t,v) in $(\mathbb{R}\times\mathbb{R}^{n_e})$

$$\frac{\partial}{\partial t}\left(v'\,[\boldsymbol{\Phi}(e,t)]^{-1}\right) = -v'\,[\boldsymbol{\Phi}(e,t)]^{-1}\,\frac{\partial F}{\partial e}(E(e,t)) \ .$$

However since from (1.20)

$$\left|\frac{\partial F}{\partial e}(E(e,t))\right| \leq c(|k(e)|\,|e|) \ ,$$

it yields the following estimate

$$\left|v'\boldsymbol{\Phi}(e,t)^{-1}\right| \leq \exp\left(c(k(|e|)|e|)t\right)|v| \ , \ \forall (t,v)\in(\mathbb{R}\times\mathbb{R}^{n_e}) \ .$$

This implies, for all (t,v) in $(\mathbb{R}\times\mathbb{R}^{n_e})$,

$$\begin{aligned}
[v'v]^2 &\leq \left|v'\boldsymbol{\Phi}(e,t)^{-1}\right|^2 |\boldsymbol{\Phi}(e,t)v|^2 \ , \\
&\leq \frac{1}{\mu_{\min}\{Q\}}\left|v'\boldsymbol{\Phi}(e,t)^{-1}\right|^2 v'\boldsymbol{\Phi}(e,t)'Q\boldsymbol{\Phi}(e,t)v \ , \\
&\leq \frac{|v|^2\exp\left(2c(k(|e|)|e|)t\right)}{\mu_{\min}\{Q\}}\, v'\boldsymbol{\Phi}(e,t)'Q\boldsymbol{\Phi}(e,t)v \ .
\end{aligned}$$

Which yields

$$v'\boldsymbol{\Phi}(e,t)'Q\boldsymbol{\Phi}(e,t)v \geq \mu_{\min}\{Q\}\exp\left(-c(k(|e|)|e|)t\right)|v|^2 \ , \ \forall(t,v)\in(\mathbb{R}\times\mathbb{R}^{n_e}) \ .$$

Consequently

$$\underline{p}(|e|) = \frac{\mu_{\min}\{Q\}}{2c(k(|e|)|e|)} \leq \lambda_{\min}\{P(e)\} \qquad \forall\widetilde{e}\in\mathbb{R}^{n_e} \ .$$

Finally, to obtain (1.30), we exploit the semi group property of the solutions. For all $(\widetilde{e},e)$ in $\mathbb{R}^{n_e}\times\mathbb{R}^{n_e}$, and all (t,r) in $\mathbb{R}^2_{\geq 0}$

$$\widetilde{E}(\widetilde{E}(\widetilde{e},e,t),E(e,t),r) = \widetilde{E}(\widetilde{e},e,t+r) \ .$$

Differentiating with respect to $\widetilde{e}$ the previous equality yields

$$\frac{\partial\widetilde{E}}{\partial\widetilde{e}}(\widetilde{E}(\widetilde{e},e,t),E(e,t),r)\frac{\partial\widetilde{E}}{\partial\widetilde{e}}(\widetilde{e},e,t)=\frac{\partial\widetilde{E}}{\partial\widetilde{e}}(\widetilde{e},e,t+r)\ .$$

Hence,

$$\boldsymbol{\Phi}(E(e,t),r)\boldsymbol{\Phi}(e,t)=\boldsymbol{\Phi}(e,t+r)\ .$$

Substituting into the previous equality,

$$e\ :=E(e,h)\ ,\ h\ :=-t\ ,\ s\ :=t+r\ ,$$

for all e in $\mathbb{R}^{n_e}$ and all (s,h) in $\mathbb{R}^2$

$$\boldsymbol{\Phi}(e,s+h)\boldsymbol{\Phi}(E(e,h),-h)=\boldsymbol{\Phi}(E(e,h),s)\ .$$

Consequently,

$$\begin{aligned}P(E(e,h))==&\lim_{T\to+\infty}\int_0^T\boldsymbol{\Phi}(E(e,h),s)'Q\boldsymbol{\Phi}(E(e,h),s)ds\ ,\\ =&\lim_{T\to+\infty}(\boldsymbol{\Phi}(E(e,h),-h))'\left[\int_0^T(\boldsymbol{\Phi}(e,s+h))'\,Q\boldsymbol{\Phi}(e,s+h)ds\right]\\ &\qquad\boldsymbol{\Phi}(E(e,h),-h)\ .\end{aligned}$$

However,

$$\begin{aligned}\lim_{h\to0}\frac{\boldsymbol{\Phi}(E(e,h),-h)-I}{h}&=-\frac{\partial F}{\partial e}(e)\ ,\\ \lim_{h\to0}\frac{\boldsymbol{\Phi}(e,s+h)-\boldsymbol{\Phi}(e,s)}{h}&=\frac{\partial}{\partial s}\left(\boldsymbol{\Phi}(e,s)\right)\ ,\end{aligned}$$

and

$$\int_0^T\frac{\partial}{\partial s}\left(\boldsymbol{\Phi}(e,s)\right)'Q\left(\boldsymbol{\Phi}(e,s)\right)ds+\int_0^T\left(\boldsymbol{\Phi}(e,s)\right)'Q\frac{\partial}{\partial s}\left(\boldsymbol{\Phi}(e,s)\right)ds=$$
$$\boldsymbol{\Phi}(e,T)'Q\boldsymbol{\Phi}(e,T)-Q\ .$$

Since $\lim_T$ and $\lim_h$ commute because $\boldsymbol{\Phi}(e,s)$ exponentially converges to 0, (1.30) is satisfied.

The last assertion of the proposition holds since if F is C^3 then the matrix function $\boldsymbol{\Phi}(e,t)$ is C^2 in e. Moreover, the first and second derivatives of its coefficient also belong to $L^2[0,+\infty)$. □

Example 1.3 Following the scalar example given in (1.21),

$$\underline{p}(e) = \frac{1}{2} \ , \ \overline{p}(e) \leq \widetilde{k}(e)^2 \lim_{T\to+\infty} \int_0^T \exp(-2s)ds = \frac{\exp\left(4e^2\exp(e^2)\right)}{2} \ .$$

1.3.2.3 Assertion 3: Construction of a Lyapunov Function

With P as defined for in (1.28) which Lie derivative satisfies inequality (1.30), it yields that along the solution of linearized system (1.22)

$$\overset{\cdot}{\overbrace{\widetilde{e}^\top P(e)\widetilde{e}}} = -\widetilde{e}^\top Q\widetilde{e} \ .$$

In other words, the mapping $(\widetilde{e}, e) \mapsto \widetilde{e}^\top P(e)\widetilde{e}$ is a global Lyapunov function for the $\widetilde{e}$ components of linearized system (1.28).

However, $e \mapsto e^\top P(e)e$ is not a global Lyapunov function for the nonlinear system (1.1) since

$$\overset{\cdot}{\overbrace{e^\top P(e)e}} = 2e^\top P(e)\left[F(e) - \frac{\partial F}{\partial e}(e)e\right] - e^\top Qe \ ,$$

is negative definite if $F(e) - \frac{\partial F}{\partial e}(e)e$ is small only and there is no guarantee that this is case away from the origin.

Nevertheless, it is still possible to construct a Lyapunov function for system (1.1). Indeed, the matrix function P may be used to define a Riemanian metric on $\mathbb{R}^{n_e}$ which may be used as a Lyapunov function. Precisely, if P is a C^2 function with values that are symmetric matrices satisfying (1.29), then length of any piece-wise C^1 path $\gamma : [s_1, s_2] \to \mathbb{R}^{n_e}$ between two arbitrary points $e_1 = \gamma(s_1)$ and $e_2 = \gamma(s_2)$ in $\mathbb{R}^{n_e}$ is

$$L(\gamma)|_{s_1}^{s_2} = \int_{s_1}^{s_2} \sqrt{\frac{d\gamma}{ds}(\sigma)'P(\gamma(\sigma))\frac{d\gamma}{ds}(\sigma)}\, d\sigma \ . \tag{1.31}$$

Minimizing along all such paths we obtain the distance $d_P(e_1, e_2)$.

From the well established relation between (geodesically) monotone vector field (semi-group generator) (operator) and contracting (non-expansive) flow (semi-group) (see, for example [7, 12, 13, 17]), if P is C^2 and the metric space is complete, this distance between any two solutions of (1.1) exponentially decreases to 0 if (1.30) is satisfied with Q a positive definite symmetric matrix. For a proof, see for example [17, Theorem 1], [13, Theorems 5.7 and 5.33] or [22, Lemma 3.3] (replacing $f(x)$ by $x + hf(x)$).

Thus, a candidate Lyapunov function is the Riemannian distance to the origin. Hence, we introduce $V : \mathbb{R}^{n_e} \to \mathbb{R}_{\geq 0}$

$$V(e) = d_P(e, 0) \,. \tag{1.32}$$

In the following proposition we show that this is indeed a good Lyapunov function candidate and moreover that it admits an upper Dini derivative along the solution of system (1.1) which is negative definite.

Proposition 1.7 (Assertion 3 globally) *Assume F is C^2 and that there exists a C^2 matrix function, P, such that (1.29) and (1.30) hold and $\underline{p}$ satisfies*

$$\lim_{r\to+\infty} \underline{p}(r) r^2 = +\infty \,. \tag{1.33}$$

Then V, defined in (1.32), is a Lyapunov function for system (1.1). More precisely V admits an upper Dini derivative along the solutions of system (1.1) defined as

$$D_F^+ V(e) := \limsup_{h\searrow 0} \frac{V(E(e,h)) - V(e)}{h} \,,$$

which satisfies

$$D_F^+ V(e) \leq -\frac{\mu_{\min}\{Q\}}{\bar{p}(|e|)} V(e) \,.$$

Hence, the origin is locally exponentially stable and globally attractive.

Proof Given an initial point e in $\mathbb{R}^{n_e}$, and a direction v also in $\mathbb{R}^{n_e}$, geodesics are given as solution to the geodesic equation

$$\frac{d^2\gamma_\ell}{ds^2}(s)(s) = \sum_{i,j}^{n} \mathfrak{d}_{ij}^{\ell} \frac{d\gamma_i}{ds}(s) \frac{d\gamma_j}{ds}(s) \,,\ \gamma(0) = e \,,\ \frac{d\gamma}{ds}(0) = v \,, \tag{1.34}$$

where the ($\mathfrak{d}_{ij}^{\ell}$) are Christoffel symbols associated to P which are C^1 if P is C^2. Since the right hand side of (1.34) is C^1 solutions $\left(\gamma(s), \frac{d\gamma}{ds}(s)\right)$ of (1.34) exist at least for small s, and are unique and C^1. Hence, $\gamma(\cdot)$ is C^2 on its domain of existence.

From [24, Lemma A.1] and the assumption of (1.33), it yields that these geodesics can be maximally extended to $\mathbb{R}$. From the Hopf–Rinow Theorem, this implies that the metric space ($\mathbb{R}^{n_e}, P$) is complete. Moreover, for any e in $\mathbb{R}^{n_e}$ there exists $\gamma^* : [0, s_e] \to \mathbb{R}^{n_e}$ a C^2 curve (a geodesic) such that

$$d_P(e, 0) = L(\gamma^*)\big|_0^{s_e} \,.$$

Without loss of generality, it is assumed that the geodesics are normalized, and so

$$\frac{d\gamma^*}{ds}(s)^\top P(\gamma^*(s)) \frac{d\gamma^*}{ds}(s) = 1 \,.$$

Hence V satisfies

$$V(e) = \int_0^{s_e} \sqrt{\frac{d\gamma^*}{ds}(s)^\top P(\gamma^*(s))\frac{d\gamma^*}{ds}(s)}\, ds = \int_0^{s_e} \frac{d\gamma^*}{ds}(s)^\top P(\gamma^*(s))\frac{d\gamma^*}{ds}(s)\, ds = s_e\ .$$

Let us first show that V is a positive definite and proper function. Since $\gamma^*\ :\ [0, s_e]$ is a continuous path from e to zero, this implies that there exists s_0 in $[0, s_e]$ such that

$$|\gamma^*(s_0)| = |e|\ ,\ |\gamma^*(s)| \leq |e|\ ,\ \forall s \in [s_0, s_e]\ . \tag{1.35}$$

Since

$$\begin{aligned} V(e) &= \int_0^{s_e} \sqrt{\frac{d\gamma^*}{ds}(s)^\top P(\gamma^*(s))\frac{d\gamma^*}{ds}(s)}ds\ ,\\ &\geq \int_{s_0}^{s_e} \sqrt{\frac{d\gamma^*}{ds}(s)^\top P(\gamma^*(s))\frac{d\gamma^*}{ds}(s)}ds\ ,\\ &\geq \sqrt{\underline{p}(|e|)}\int_{s_0}^{s_e} \sqrt{\frac{d\gamma^*}{ds}(s)^\top \frac{d\gamma^*}{ds}(s)}ds\ , \end{aligned}$$

and since minimal geodesic for an Euclidean metric are straight lines $s \mapsto \frac{s\gamma^*(s_0)}{s_e - s_0}$, then

$$\int_{s_0}^{s_e} \sqrt{\frac{d\gamma^*}{ds}(s)^\top \frac{d\gamma^*}{ds}(s)}ds \geq \int_{s_0}^{s_e} \sqrt{\frac{\gamma^*(s_0)^\top \gamma^*(s_0)}{(s_e - s_0)^2}}ds = |\gamma^*(s_0)|\ .$$

Hence, from (1.35),

$$V(e) \geq \sqrt{\underline{p}(|e|)}|e|\ .$$

Moreover,

$$\begin{aligned} V(e) &\leq \int_0^{s_e} \frac{e^\top}{s_e} P\left(\frac{se^\top}{s_e}\right)\frac{e}{s_e}ds\ ,\\ &\leq \overline{p}(|e|)\int_0^{s_e} \frac{e^\top}{s_e}\frac{e}{s_e}ds\ ,\\ &\leq \frac{\overline{p}(|e|)}{s_e}|e|^2\ . \end{aligned}$$

Since $V(e) = s_e$, the two previous inequalities imply

$$\sqrt{\underline{p}(|e|)}|e| \leq V(e) \leq \sqrt{\overline{p}(|e|)}|e|\ . \tag{1.36}$$

From (1.33), this implies that V is positive definite and proper.

Let $\varGamma(s,t)$ be the mapping defined by

$$\frac{\partial \varGamma}{\partial t}(s,t) = F(\varGamma(s,t)) \ , \ \varGamma(s,0) = \gamma^*(s) \ .$$

Since F is C^2 and the mapping γ^* is C^2, then $\varGamma$ is C^2. Note that $\varGamma(s,h)$ is a C^2 path such that

$$\varGamma(s_e,h) = E(e,h) \ , \ \varGamma(0,h) = 0 \ .$$

This implies, for all $h \geq 0$

$$V(E(e,h)) \leq \int_0^{s_e} \sqrt{\frac{\partial \varGamma}{\partial s}(s,h)^\top P(\varGamma(s,h)) \frac{\partial \varGamma}{\partial s}(s,h)} ds \ .$$

Thus,

$$D_F^+ V(e) \leq \limsup_{h \to 0} \int_0^{s_e} \frac{\sqrt{\frac{\partial \varGamma}{\partial s}(s,h)^\top P(\varGamma(s,h)) \frac{\partial \varGamma}{\partial s}(s,h)} - \sqrt{\frac{\partial \gamma^*}{\partial s}(s)^\top P(\gamma^*(s)) \frac{\partial \gamma^*}{\partial s}(s)}}{h} ds \ .$$

Hence, with Fatou's lemma, it yields

$$D^+ V(e) \leq \int_0^{s_e} \limsup_{h \to 0} \frac{\sqrt{\frac{\partial \varGamma}{\partial s}(s,h)^\top P(\varGamma(s,h)) \frac{\partial \varGamma}{\partial s}(s,h)} - \sqrt{\frac{\partial \gamma^*}{\partial s}(s)^\top P(\gamma^*(s)) \frac{\partial \gamma^*}{\partial s}(s)}}{h} ds \ .$$

Since the mapping $h \mapsto \sqrt{\frac{\partial \varGamma}{\partial s}(s,h)^\top P(\varGamma(s,h)) \frac{\partial \varGamma}{\partial s}(s,h)}$ is C^1 (since $\varGamma$ and P are C^2), it yields

$$\begin{aligned} D^+ V(e) &\leq \int_0^{s_e} \frac{\partial}{\partial h} \left\{ \sqrt{\frac{\partial \varGamma}{\partial s}(s,\cdot)^\top P(\varGamma(s,\cdot)) \frac{\partial \varGamma}{\partial s}(s,\cdot)} \right\}_{h=0} ds \ , \\ &= - \int_0^{s_e} \frac{1}{2} \frac{\frac{d\gamma^*}{ds}(s)^\top Q \frac{d\gamma^*}{ds}(s)}{\sqrt{\frac{\partial \gamma^*}{\partial s}(s)^\top P(\gamma^*(s)) \frac{\partial \gamma^*}{\partial s}(s)}} ds \ , \\ &\leq -\frac{1}{2} \mu_{\min}\{Q\} \int_0^{s_e} \frac{d\gamma^*}{ds}(s)^\top \frac{d\gamma^*}{ds}(s) ds \ , \end{aligned}$$

where the last inequality employs the fact that the geodesics are normalized. From the Cauchy–Schwartz inequality,

$$D^+ V(e) \leq -\frac{1}{2} \mu_{\min}\{Q\} \left(\int_0^{s_e} \sqrt{\frac{d\gamma^*}{ds}(s)^\top \frac{d\gamma^*}{ds}(s)} ds \right)^2 \ ,$$

and since minimal geodesics for a Euclidean metric are straight lines,

$$\begin{aligned} D^{+}V(e) &\leq -\frac{1}{2}\mu_{\min}\{Q\}\int_0^{s_e}\sqrt{\frac{e}{s_e}^\top\frac{e}{s_e}}ds\ ,\\ &\leq -\frac{\mu_{\min}\{Q\}}{2\sqrt{\overline{p}(|e|)}}\int_0^{s_e}\sqrt{\frac{e}{s_e}^\top P\left(\frac{se}{s_e}\right)\frac{e}{s_e}}ds\ ,\\ &\leq -\frac{\mu_{\min}\{Q\}}{2\sqrt{\overline{p}(|e|)}}V(e)\ . \end{aligned}$$

Together with (1.36), this implies global asymptotic stability of the origin. Since $0 < \underline{p}(0) < \overline{p}(0)$, it also implies that the origin is locally exponentially stable. □

An interesting property of the considered Lyapunov function is that, given two points e_1 and e_2 both in $\mathbb{R}^{n_e}$, if $d_P(e_1, e_2)$ is the Riemmanian distance between these two points and γ^* is the minimal (and normalized) geodesic, this yields, following the previous proof, that there exists s_0 such that

$$|\gamma^*(s_0) - e_2| = |e_1 - e_2|\ ,\ |\gamma^*(s) - e_2| \leq |e_1 - e_2|\ ,\ \forall s \in [s_0, s_2]\ .$$

Thus,

$$d_P(e_1, e_2) \geq \int_{s_0}^{s_2}\sqrt{\frac{d\gamma^*}{ds}(s)^\top P(\gamma^*(s))\frac{d\gamma^*}{ds}(s)ds}\ ,$$

and, for all s in $[s_0, s_2]$, $|\gamma^*(s)| \leq |\gamma^*(s) - e_2| + |e_2| \leq |e_1 - e_2| + |e_2|$. Hence,

$$\begin{aligned} d_P(e_1, e_2) &\geq \sqrt{\underline{p}(|e_1 - e_2| + |e_2|)}\int_{s_0}^{s_2}\sqrt{\frac{d\gamma^*}{ds}(s)^\top\frac{d\gamma^*}{ds}(s)ds}\ ,\\ &\geq \sqrt{\underline{p}(|e_1 - e_2| + |e_2|)}|\gamma^*(s_0) - e_2|\ ,\\ &\geq \sqrt{\underline{p}(|e_1 - e_2| + |e_2|)}|e_1 - e_2|\ . \end{aligned}$$

Moreover,

$$d_P(e_1, e_2) \leq \frac{\overline{p}(|e_1 - e_2| + |e_2|))}{d_P(e_1, e_2)}|e_1 - e_2|^2\ .$$

The two previous inequalities imply

$$\sqrt{\underline{p}(|e_1 - e_2| + |e_2|))}|e_1 - e_2| \leq d_P(e_1, e_2) \leq \sqrt{\overline{p}(|e_1 - e_2| + |e_2|))}|e_1 - e_2|\ ,$$

and

$$D^+_{F,F} d_P(e_1, e_2) \leq -\int_{s_1}^{s_2} \dot{\gamma}^*(s) Q \dot{\gamma}^*(s) ds \leq -\frac{\mu_{\min}\{Q\}}{2\sqrt{\underline{p}(|e_1 - e_2| + |e_2|)}} d_p(e_1, e_2) ,$$

where

$$D^+_{F,F} d_P(e_1, e_2) := \limsup_{h \searrow 0} \frac{d_P(E(e_1, h)), E(e_2, h))}{h} .$$

In other words, there exists a strictly decreasing distance between any two points. Consequently, there is exponential convergence of the euclidean distance between any two trajectories toward zero. Hence, broadly speaking, we have shown that when the origin is locally exponentially stable and globally attractive, there exists a strictly decreasing distance between any two trajectories. However, this convergence is not uniform in e_1 and e_2. This is a strong contrast with the incremental stability previously reported in, for example, [6] or [8]. Note moreover that it is shown in [23] that the asymptotic stability property and incremental stability property are different.

Note that as mentioned in [3], when $\underline{p}$ and $\overline{p}$ are lower and upper bounded by a nonzero constant, respectively, then the convergence is uniform. In this case, the usual definition of incremental stability is recovered.

1.3.2.4 Discussions About the Requirement (1.33)

Requirement (1.33) is essential to ensure that $\mathbb{R}^{n_e}$ endowed with the Riemannian metric, P, is complete and that the obtained Lyapunov function is proper. It imposes that mapping $\underline{p}$ does not vanish too quickly as $|e|$ goes to infinity. Returning to the definition of mapping $\underline{p}$ obtained in the proof of Proposition 1.6, if F is globally Lipschitz then $\underline{p}$ is a constant. In other words, this assumption is trivially satisfied in the globally Lipschitz context.

Another solution to ensure this assumption is satisfied is to modify P to make sure that this one is lower bounded by a positive real number. Indeed, the trajectories of system

$$\dot{e} = \frac{F(e)}{1 + \left|\frac{\partial F}{\partial e}(e)\right|^3} , \dot{\tilde{e}} = \frac{\frac{\partial F}{\partial e}(e)}{1 + \left|\frac{\partial F}{\partial e}(e)\right|^3} \tilde{e}$$

are the same as those of lifted system (1.22) (obtained after time rescaling). Consequently, the origin is globally attractive. Moreover, it is not difficult to show that its origin is also locally exponentially stable. Finally, if F is C^4 then the vector field $e \mapsto \frac{F(e)}{1+\left|\frac{\partial F}{\partial e}(e)\right|^3}$ is C^3. Let $\tilde{\Phi}$ be the transition matrix solution of the following $\mathbb{R}^{n_e \times n_e}$ dynamical system

$$\frac{d}{dt}\tilde{\Phi}\,(e,t) = \frac{\frac{\partial F}{\partial e}(E(e,t))}{1+\left|\frac{\partial F}{\partial e}(E(e,t))\right|^3}\tilde{\Phi}(e,t)\ ,\ \tilde{\Phi}(e,0) = I\ ,$$

where each element of the (matrix) time function $t \mapsto \tilde{\Phi}(e,t)$ is in $L^2([0,+\infty))$. Consequently, for all positive definite matrix Q in $\mathbb{R}^{n_e \times n_e}$,

$$\tilde{P}(e) = \lim_{T\to+\infty} \int_0^T \tilde{\Phi}(e,s)^\top Q \tilde{\Phi}(e,s) ds\ , \tag{1.37}$$

is well defined. With this mapping, the following property may be obtained.

Proposition 1.8 (Lower bounded P) *Assume that there exist function $(k, \widetilde{k})$ and positive real numbers $(\lambda, \widetilde{\lambda})$ such that (1.20) and (1.24) are satisfied. Then, $P : \mathbb{R}^{n_e} \to \mathbb{R}^{n_e \times n_e}$ defined in (1.37) is well defined, continuous, and there exists a positive real number, $\underline{p}$, and a nondecreasing function, $\bar{p}$, such that*

$$0 < \underline{p}I \leq \tilde{P}(e) \leq \bar{p}(|e|)I\ ,\ \forall\ e \in \mathbb{R}^{n_e}\ . \tag{1.38}$$

Moreover,

$$\mathfrak{d}_F\tilde{P}(e) + \tilde{P}(e)\frac{\partial F}{\partial e}(e) + \frac{\partial F}{\partial e}(e)^\top \tilde{P}(e) \leq -Q\left(1+\left|\frac{\partial F}{\partial e}(e)\right|^3\right)\ ,\ \forall\ e \in \mathbb{R}^{n_e}\ . \tag{1.39}$$

Finally, if the vector field F is C^4 then P is C^2.

Proof The proof is similar to the one of Proposition 1.6. For all (e,t) in $\mathbb{R}^{n_e} \times \mathbb{R}_{\geq 0}$ there exists

$$\left|\tilde{\Phi}(e,t)\right| \leq \widetilde{k}(|e|)\exp(-\tilde{\lambda}t)\ .$$

Thus for every symmetric positive definite matrix, Q, $P : \mathbb{R}^{n_x} \to \mathbb{R}^{n_e \times n_e}$ given by (1.28) is well defined, continuous, and satisfies:

$$\mu_{\max}\{P(e)\} \leq \frac{\widetilde{k}(|e|)^2}{2\tilde{\lambda}}\mu_{\max}\{Q\} = \overline{p}(|e|)\ ,\ \forall e \in \mathbb{R}^{n_e}\ .$$

Moreover, for all (t,v) in $(\mathbb{R} \times \mathbb{R}^{n_e})$

$$\frac{\partial}{\partial t}\left(v'\,[\Phi(e,t)]^{-1}\right) = -v'\,[\Phi(e,t)]^{-1}\,\frac{\frac{\partial F}{\partial e}(E(e,t))}{1+\left|\frac{\partial F}{\partial e}(E(e,t))\right|^3}\ .$$

However,

$$\left|\frac{\frac{\partial F}{\partial e}(E(e,t))}{1+\left|\frac{\partial F}{\partial e}(E(e,t))\right|^3}\right| \leq 1\ ,$$

then

$$\left|v'\Phi(e,t)^{-1}\right| \leq \exp(t)\,|v|\ ,\ \forall (t,v)\in(\mathbb{R}\times\mathbb{R}^{n_e})\ ,$$

Following the proof of Proposition 1.6,

$$\underline{p} = \frac{\mu_{\min}\{Q\}}{2} \leq \lambda_{\min}\{P(e)\} \qquad \forall \tilde{e}\in\mathbb{R}^{n_e}\ .$$

The following inequality may also be obtained

$$\mathfrak{d}_{\frac{F}{1+|\frac{\partial F}{\partial e}(e)|^3}}\tilde{P}(e) + \tilde{P}(e)\frac{\frac{\partial F}{\partial e}(e)}{1+|\frac{\partial F}{\partial e}(e)|^3} + \frac{\frac{\partial F}{\partial e}(e)^\top}{1+|\frac{\partial F}{\partial e}(e)|^3}\tilde{P}(e) \leq -Q\ ,\ \forall\ e\in\mathbb{R}^{n_e}\ .$$

Multiplying the former equation by $1+|\frac{\partial F}{\partial e}(e)|^3$ yields the result. □

Since P is lower bounded, we can define a Lyapunov function following Proposition 1.7. Broadly speaking, we have established the following Lyapunov inverse result: *Assuming some regularity on the system, if the origin is locally exponentially stable and globally attractive then there exists a strictly decreasing Lyapunov function given as a Riemannian distance to the origin.*

Of course, the local exponential stability of the origin is essential. Note that in [11], is shown that up to a change of coordinates (which is not a diffeomorphism since it is not smooth at the origin) it is possible to transform any asymptotically stable system in an exponentially stable system. This implies that up to a change of variable, it is always possible to consider a Lyapunov function arising from a Riemannian distance.

1.3.3 Stabilization

We have shown that a linearization approach provides constuction of global Lyapunov functions in the case of local exponential stability and global attractivity of the origin. It may be interesting to know if this type of Lyapunov function may be used in control design.

Consider a controlled nonlinear system on $\mathbb{R}^n$,

$$\dot{w} = f(w) + g(w)u\ , \tag{1.40}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ are smooth vector fields, and u the controlled input is in $\mathbb{R}$.

Our objective is to construct a control $u = \phi(w)$ that achieves local exponential stabilization and global attractivity of the origin. Based on the former analysis, a sufficient condition based on the use of a Riemannian–Lyapunov function may be obtained. However, these assumptions inspired from [9] and [1] are very conservative.

Proposition 1.9 *Assume there exists a mapping $P : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ such that*

1. *P is C^3, satisfies conditions (1.29), (1.33), and there exists a positive real number λ and a positive definite matrix Q such that the following matrix inequality holds*

$$\mathfrak{d}_f P(w) + P(w)\frac{\partial f}{\partial w}(w) + \frac{\partial f}{\partial w}(w)^\top P(w) - \lambda\,|P(w)g(w)|^2 \leq -Q\,,\ \forall\, w \in \mathbb{R}^n\,. \tag{1.41}$$

2. *g is a Killing vector field on $\mathbb{R}^n$ endowed with the Riemannian metric P, i.e., for all w in $\mathbb{R}^n$*

$$L_g P(w) = \mathfrak{d}_g P(w) + P(w)\frac{\partial g}{\partial w}(w) + \frac{\partial g}{\partial w}(w)^\top P(w) = 0\,.$$

3. *There exists a mapping $U : \mathbb{R}^n \to \mathbb{R}$ such that:*

$$\frac{\partial U}{\partial w}(w) = P(w)g(w)^\top\,. \tag{1.42}$$

Then the control law $u = -\lambda U(w)$ achieves local exponential stability and global attractivity of the origin of system (1.40) in closed loop.

Proof The proof follows from Proposition 1.7. The closed loop system may be expressed as

$$F(w) = f(w) - \lambda g(w)U(w)\,.$$

The Lie derivative of the tensor P is

$$L_F P(w) = L_f P(w) - \lambda L_g P(w)U(w) - P(w)g(w)\frac{\partial U}{\partial w}(w)$$

which implies

$$L_F P(w) = L_f P(w) - \lambda\,|P(w)g(w)|^2 \leq -Q\,.$$

From Proposition 1.7, this implies the result. □

Following [3], it is possible to slightly relax these assumptions by introducing a scaling factor, $\alpha(w)$, which multiply g and by rewriting these assumptions accordingly.

1.4 Conclusion and Final Remark

We have shown that a first order approximation leads to the construction of Lyapunov functions that characterize the local exponential stability of a transverse manifold. A global Lyapunov function may be constructed from first order approximation in the context of local exponential stability and global attractivity of an equilibrium point. For this case, one may consider a Riemannian length to the origin as a Lyapunov function.

Some problematic effects arise when considering global transverse exponential stability for a manifold as shown in the following simple example.

Consider the planar system defined on $\mathbb{R}^2$,

$$\dot{e} = -\phi(x)e \ , \ \dot{x} = \mu_x x \ , \ \phi(x) = \lambda + x\sin(x) \ , \tag{1.43}$$

with solution for all t in $\mathbb{R}$,

$$E(e_0, x_0, t) = \exp\left(-\lambda t + \frac{\cos(1) - \cos(e^{\mu_x t}x_0)}{\mu_x}\right) e_0 \ , \ X(e_0, x_0, t) = e^{\mu_x t}x_0 \ .$$

This implies that manifold $\{(e, x), e = 0\}$ is (transversally) locally exponential stable and globally attractive uniformly in x. Indeed, we have for all (e_0, x_0) in $\mathbb{R}$

$$|E((e_0, x_0), t)| \leq \exp\left(\frac{\cos(1) + 1}{\mu_x}\right) \exp(-\lambda t)|e_0| \ .$$

The transversally linear system is

$$\overset{\cdot}{\overbrace{\begin{bmatrix} \widetilde{E} \\ \widetilde{X} \end{bmatrix}}} = \begin{bmatrix} \phi(e^{\mu_x t}x_0) & \phi'(e^{\mu_x t}x_0)E(e_0, x_0, t) \\ 0 & \mu_x \end{bmatrix} \begin{bmatrix} \widetilde{E} \\ \widetilde{X} \end{bmatrix} \ ,$$

which gives (with $\widetilde{E}(t) = \widetilde{E}(\widetilde{e}_0, \widetilde{x}_0, e_0, x_0, t)$)

$$\begin{aligned} \widetilde{E}(t) = \exp\left(\int_0^t \phi(e^{\mu_x s}x_0)ds\right)\widetilde{e}_0 + \int_0^t \exp\left(\int_s^t \phi(e^{\mu_x v}x_0)dv\right) \phi'(e^{\mu_x v}x_0) \\ \times E(w_0, s)e^{\mu_x s}\widetilde{x}_0 ds \ , \\ = \exp\left(\int_0^t \phi(e^{\mu_x s}x_0)ds\right)\left[\widetilde{e}_0 + \int_0^t \phi'(e^{\mu_x s}x_0)e^{\mu_x s}e_0\widetilde{x}_0 ds\right] \ . \end{aligned}$$

Hence, this yields if $x_0 \neq 0$,

$$\widetilde{E}(t) = \exp\left(\int_0^t \phi(e^{\mu_x s}x_0)ds\right)\left[\widetilde{e}_0 + \frac{\phi(e^{\mu_x t}x_0) - \phi(x_0)}{\mu_x}\frac{e_0\widetilde{x}_0}{x_0}\right] \ .$$

With ϕ as previously defined,

$$\widetilde{E}(t) = \exp\left(\frac{\cos(x_0) - \cos(e^{\mu_x t}x_0)}{\mu_x}\right) \times \left[e^{-\lambda t}\widetilde{e}_0 + \frac{e^{(\mu_x-\lambda)t}\sin(e^{\mu_x t}x_0) - \sin(x_0)e^{-\lambda t}}{\mu_x} e_0\widetilde{x}_0\right] .$$

which does not converge to zero if $\lambda < \mu_x$ if for example $e_0 = 1$, $x_0 = 1$, $\widetilde{x}_0 = 1$. Thus study of the linearized systems must be undertaken with care and this implies that Assertion 1 from Sect. 1.1 is no longer valid in this context. More precisely, exponential convergence to the origin of the e dynamics, does not imply that the $\widetilde{e}$ component of the linearized system along solutions converges to zero.

In [2], it has been shown that when the convergence rate to the manifold is larger then the expansion rate in the manifold, Assertion 1 may hold. In this case, it is possible to construct a Lyapunov function based on first order approximation.

The construction of a matrix function which satisfies Eqs. (1.15), (1.30), or (1.41) is a crucial step to make this framework practical. Preliminary results aiming at solving a differential Riccati equation (e.g., (1.41)) are given in [25]. Backstepping-based approaches are also a possible research area (see [28, 29] or [5]). Methods following numerical approximation of the partial differential equation should also be considered.

References

1. Andrieu, V., Jayawardhana, B., Praly, L.: On transverse exponential stability and its use in incremental stability, observer and synchronization. In: Proceedings of the 52nd IEEE Conference on Decision and Control (2013)
2. Andrieu, V., Jayawardhana, B., Praly, L.: Globally transverse exponential stability. Technical report (2015)
3. Andrieu, V., Jayawardhana, B., Praly, L.: Transverse exponential stability and applications. IEEE Trans. Autom. Control (2015)
4. Andrieu, V., Jayawardhana, B., Tarbouriech, S.: Necessary and sufficient condition for local exponential synchronization of nonlinear systems. In: Proceedings of the 54th IEEE Conference on Decision and Control (2015)
5. Andrieu, V., Jayawardhana, B., Tarbouriech, S.: Some results on exponential synchronization of nonlinear systems (2016). arXiv preprint arXiv:1605.09679
6. Angeli, D.: A lyapunov approach to incremental stability properties. IEEE Trans. Autom. Control **47**(3), 410–421 (2002)
7. Brezis, H.: Opérateur maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert, vol. 5. Mathematics Studies (1973)
8. Forni, F., Sepulchre, R.: A differential lyapunov framework for contraction analysis. IEEE Trans. Autom. Control **59**(3), 614–628 (2014)
9. Forni, F., Sepulchre, R., Van Der Schaft, A.J.: On differential passivity of physical systems. In: 2013 IEEE 52nd Annual Conference on Decision and Control (CDC), pp. 6580–6585. IEEE (2013)
10. Gauthier, J.-P., Kupka, I.: Deterministic Observation Theory and Applications. Cambridge University Press (2001)

11. Grüne, L., Sontag, E.D., Wirth, F.R.: Asymptotic stability equals exponential stability, and iss equals finite energy gainif you twist your eyes. Syst. Control Lett. **38**(2), 127–134 (1999)
12. Hartman, P.: Ordinary Differential Equations. Wiley (1964)
13. Isac, G., Németh, S.Z.: Scalar and Asymptotic Scalar Derivatives: Theory and Applications, vol. 13. Springer (2008)
14. Isidori, A.: Nonlinear Control Systems: An Introduction. Springer, New York, Inc., New York, NY, USA (1989)
15. Khalil, H.K.: Nonlinear Systems, 3rd edn. Prentice-Hall (2002)
16. Kurzweil, J.: On the inversion of Lyapunov second theorem on stability of motion. Ann. Math. Soc. Trans. Ser. **2**(24), 19–77 (1956)
17. Lewis, D.C.: Metric properties of differential equations. Am. J. Math. **71**(2), 294–312 (1949)
18. Lyapunov, A.M.: The general problem of the stability of motion. Int. J. Control **55**(3), 531–534 (1992)
19. Massera, J.L.: On Liapunoff s condition of stability. Ann. Math. **50**, 705–721 (1949)
20. Massera, J.L.: Contributions to stability theory. Ann. Math., 182–206 (1956)
21. Praly, L.: Fonctions de Lyapunov, stabilité et stabilisation. Ecole Nationale Supérieure des Mines de Paris (2008)
22. Reich, S.: Nonlinear Semigroups, Fixed Points, and Geometry of Domains in Banach Spaces. Imperial College Press (2005)
23. Rüffer, B.S., van de Wouw, N., Mueller, M.: Convergent systems vs. incremental stability. Syst. Control Lett. **62**(3), 277–285 (2013)
24. Sanfelice, R.G., Praly, L.: Convergence of nonlinear observers on $\mathbb{R}^n$ with a riemannian metric (part i). IEEE Trans. Autom. Control **57**(7), 1709–1722 (2012)
25. Sanfelice, R.G., Praly, L.: Solution of a riccati equation for the design of an observer contracting a riemannian distance. In: 2015 Proceedings of the 54th IEEE Conference on Decision and Control. IEEE (2015)
26. Sepulchre, R., Janković, M., Kokotović, P.V.: Constructive nonlinear control. Communications and Control Engineering Series. Springer (1997)
27. Teel, A.R., Praly, L.: A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions. ESAIM Control Optim. Calc. Var. **5**, 313–367 (2000)
28. Zamani, M., Tabuada, P.: Backstepping design for incremental stability. IEEE Trans. Autom. Control **56**(9), 2184–2189 (2011)
29. Zamani, M., van de Wouw, N., Majumdar, R.: Backstepping controller synthesis and characterizations of incremental stability. Syst. Control Lett. **62**(10), 949–962 (2013)

Chapter 2
A Review on Model Reduction by Moment Matching for Nonlinear Systems

Giordano Scarciotti and Alessandro Astolfi

Abstract The model reduction problem for nonlinear systems and nonlinear time-delay systems based on the steady-state notion of moment is reviewed. We show how this nonlinear description of moment is used to pose and solve the model reduction problem by moment matching for nonlinear systems, to develop a notion of frequency response for nonlinear systems, and to solve model reduction problems in the presence of constraints on the reduced order model. Model reduction of nonlinear time-delay systems is then discussed. Finally, the problem of approximating the moment of nonlinear, possibly time-delay, systems from input/output data is briefly illustrated.

2.1 Introduction

The model reduction problem has been widely studied for the prediction, analysis, and control of a wide class of physical behaviors. For instance, reduced order models are used to simulate or design weather forecast models, very large scale integrated circuits or networked dynamical systems [1]. The model reduction problem consists in finding a simplified description of a dynamical system maintaining at the same time specific properties. For linear system, the problem has been extensively studied exploiting a variety of techniques, some of them based on the singular value decomposition, see, e.g., [2–4] which make use of Hankel operators or, e.g., [5–8] which exploit balanced realizations, and some based on the Krylov projec-

Dedicated to Laurent: a pioneer in the land of control

G. Scarciotti · A. Astolfi (✉)
Imperial College London, London SW7 2AZ, UK
e-mail: a.astolfi@ic.ac.uk

G. Scarciotti
e-mail: gs3610@ic.ac.uk

A. Astolfi
University of Rome “Tor Vergata”, Via Del Politecnico 1, 00133 Rome, Italy

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_2

tion matrices, see, e.g., [9–15], also called moment matching methods. The additional difficulties of the reduction of nonlinear systems carry the need to develop different or “enhanced” techniques. The problem of model reduction for special classes of systems, such as differential-algebraic systems, bilinear systems, and mechanical/Hamiltonian systems has been studied in [16–19]. Energy-based methods have been proposed in [7, 20, 21]. Other techniques, based on the reduction around a limit cycle or a manifold, have been presented in [22, 23]. Model reduction methods based on proper orthogonal decomposition have been developed for linear and nonlinear systems, see, e.g., [24–28]. Finally, note that some computational aspects have been investigated in [23, 26, 29, 30]. In addition, the problem of model reduction of time-delay systems is a classic topic in control theory. The optimal reduction (in the sense of some norm) is listed as an unsolved problem in systems theory in [31] and several results have been given using rational interpolations, see, e.g., [32–34], see also [35–41]. Recent results include model order reduction techniques for linear time-delay systems, see, e.g., [42–44], and for infinite dimensional systems, see, e.g., [45, 46] in which operators are used to provide reduced order models for linear systems. The goal of this chapter is to review the model reduction techniques for nonlinear, possibly time-delay, systems based on the “steady-state” notion of moment. We start introducing the interpolation approach to moment matching, which is how moment matching has been classically interpreted and applied to linear systems. We then move to the steady-state approach introduced in [47]. We present some results on the model reduction problem by moment matching for nonlinear systems, as given in [48], and develop a notion of frequency response for nonlinear systems. These techniques are extended to nonlinear time-delay systems [49] and the problem of obtaining a family of reduced order models matching two (nonlinear) moments is solved for a special class of signal generators. Finally the problem of approximating the moment of nonlinear (time-delay) systems, without solving the partial differential equation that defines it, is presented and solved [50, 51].

Notation. We use standard notation. $\mathbb{R}_{>0}$ denotes the set of positive real numbers; $\mathbb{C}_{<0}$ denotes the set of complex numbers with negative real part; $\mathbb{D}_{<1}$ denotes the set of complex numbers with modulo smaller than one; $\imath$ denotes the imaginary unit. Given a set of delays $\{\tau_j\}$, the symbol $\mathfrak{R}^n_T = \mathfrak{R}^n_T([-T,0],\mathbb{R}^n)$, with $T = \max_j\{\tau_j\}$, indicates the set of continuous functions mapping the interval $[-T,0]$ into $\mathbb{R}^n$ with the topology of uniform convergence [52]. The symbol I denotes the identity matrix, $\sigma(A)$ denotes the spectrum of the matrix $A \in \mathbb{R}^{n\times n}$ and $\otimes$ indicates the Kronecker product. The vectorization of a matrix $A \in \mathbb{R}^{n\times m}$, denoted by $\mathrm{vec}(A)$, is the $nm \times 1$ vector obtained by stacking the columns of the matrix A one on top of the other, namely $\mathrm{vec}(A) = [a_1^\top, a_2^\top, \dots, a_m^\top]^\top$, where $a_i \in \mathbb{R}^n$ are the columns of A and the superscript $\top$ denotes the transposition operator. The superscript $*$ indicates the complex conjugate transposition operator. Let $\bar{s} \in \mathbb{C}$ and $A(s) \in \mathbb{C}^{n\times n}$. Then $\bar{s} \notin \sigma(A(s))$ means that $\det(\bar{s}I - A(\bar{s})) \neq 0$. $\sigma(A(s)) \subset \mathbb{C}_{<0}$ means that for all $\bar{s}$ such that $\det(\bar{s}I - A(\bar{s})) = 0$, $\bar{s} \in \mathbb{C}_{<0}$. $L_f h$ denotes the Lie derivative of the smooth function h along the smooth vector field f, as defined in [53, Chapter 1].

2.2 The Interpolation Approach

In this section we briefly recall the notion of moment and the related model reduction techniques as presented in [1]. We refer to this family of methods as "interpolation-based" methods. The key element to understand this framework is that the moment matching problem is interpreted as a problem of interpolation of points in the complex plane, which has been solved by the Nevanlinna-Pick theory (see, e.g., [54]).

Definition 2.1 Let $\{s_i\}$ be a sequence of distinct points in $Z \subset \mathbb{C}$ and let $\{w_i\}$ be an arbitrary sequence of points in $\mathbb{C}$. Given a space $\mathscr{W}$ of functions on Z, the *interpolation problem* consists in determining a function $W : Z \mapsto \mathbb{C}$ such that $W(s_i) = w_i$, for all $i = 1, \dots, \nu$.

Consider a linear, single-input, single-output, continuous-time, system described by the equations

$$\dot{x} = Ax + Bu, \qquad y = Cx, \tag{2.1}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$ and $C \in \mathbb{R}^{1\times n}$. Let

$$W(s) = C(sI - A)^{-1}B$$

be the associated transfer function and assume that (2.1) is minimal, i.e., controllable and observable. The k-moment of system (2.1) at s_i is defined as the k-th coefficient of the Laurent series expansion of the transfer function $W(s)$ in a neighborhood of $s_i \in \mathbb{C}$ (see [1, Chapter 11]), provided it exists.

Definition 2.2 Let $s_i \in \mathbb{C} \setminus \sigma(A)$. The 0-*moment of system* (2.1) *at* s_i is the complex number $\eta_0(s_i) = W(s_i)$. The *k-moment of system* (2.1) *at* s_i is the complex number

$$\eta_k(s_i) = \frac{(-1)^k}{k!} \left[\frac{d^k}{ds^k} W(s) \right]_{s=s_i},$$

with $k \geq 1$ integer.

In the interpolation approach to moment matching, a reduced order model is such that its transfer function (and, possibly, derivatives of this) takes the same values of the transfer function (and, possibly, derivatives of this) of system (2.1) at s_i. This is graphically represented in Fig. 2.1 in which the magnitude (top) and phase (bottom) of the transfer function of a reduced order model (dashed/red line) matches the respective quantities of a given system (solid/blue line) at the point $s_i = 30\imath$. Since a minimal system can be entirely described by its transfer function, such a system can be effectively reduced using this technique. In this framework, the problem of model reduction by moment matching can be formulated as the problem of finding the correct Petrov-Galerkin projectors $V \in \mathbb{R}^{n\times\nu}$ and $W \in \mathbb{R}^{n\times\nu}$, with $W^*V = I$, such that the model described by the equations

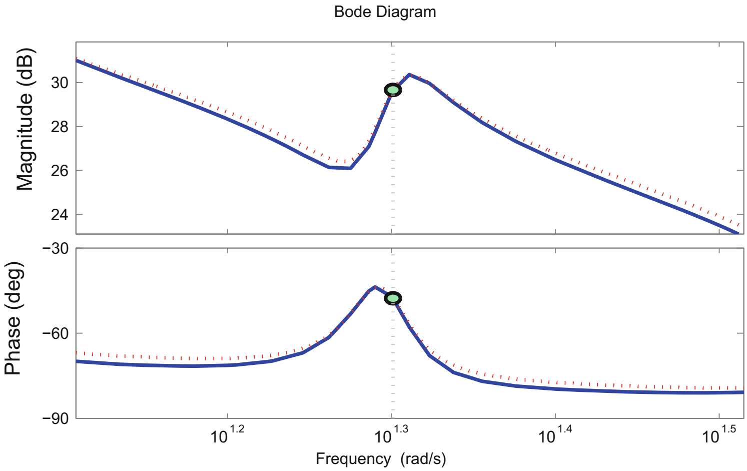


Fig. 2.1 Diagrammatic illustration of the interpolation approach. Magnitude (*top graph*) and phase (*bottom graph*) plot of a given system (*solid/blue line*) and of a reduced order model (*dashed/red line*). The green circle represents the interpolation point

$$\dot{\xi} = F\xi + Gu, \qquad \psi = H\xi, \tag{2.2}$$

with $\xi(t) \in \mathbb{R}^\nu$, $u(t) \in \mathbb{R}$, $\psi(t) \in \mathbb{R}$, $F \in \mathbb{R}^{\nu\times\nu}$, $G \in \mathbb{R}^{\nu\times 1}$, $H \in \mathbb{R}^{1\times\nu}$, and

$$F = W^*AV, \quad G = W^*B, \quad H = CV, \tag{2.3}$$

matches the moments of the given system at a set of points s_i. The problem of model reduction by moment matching using the Petrov-Galerking projectors is thoroughly described in [1] and it is the subject of intensive research, see, e.g., [9–15]. Herein we report a few results which are instrumental for the aims of the chapter. We invite the reader to refer to [1] for additional detail.

Proposition 2.1 [1] *Consider ν distinct points $s_j \in \mathbb{C} \setminus \sigma(A)$, with $j = 1, \dots, \nu$. The transfer function of the reduced order model (2.2), with*

$$V = \left[(s_1 I - A)^{-1}B \cdots (s_\nu I - A)^{-1}B\right] \tag{2.4}$$

a generalized reachability matrix and W any left inverse of V, interpolates the transfer function of system (2.1) at the points s_j, with $j = 1, \dots, \nu$.

Proposition 2.2 [1] *Consider the point $s_0 \in \mathbb{C} \setminus \sigma(A)$. The transfer function of the reduced order model (2.2), with*

$$V = \left[(s_0 I - A)^{-1}B \; (s_0 I - A)^{-2}B \cdots (s_0 I - A)^{-\nu}B\right] \tag{2.5}$$

a generalized reachability matrix and W any left inverse of V, interpolates the transfer function of system (2.1) and its $\nu - 1$ derivatives at the point s_0.

The techniques which result from these propositions are called *rational interpolation methods* by projection, or *Krylov methods*. We note that the matrix W is a free parameter since it has to satisfy only a "mild" constraint, namely that it is a left inverse of V. However, the selection of W such that the reduced order model exhibits specific properties is in general a difficult problem. The results presented to exploit the free parameters of the matrix W play, with different aims, on the possibility of interpolating more, somewhat special, points. The first of these results, which we recall here, provides a method for the so-called two-sided interpolation.

Proposition 2.3 [1] *Consider $s_j \in \mathbb{C} \setminus \sigma(A)$, with $j = 1, \dots, 2\nu$, the generalized reachability matrix*

$$\bar{V} = \left[(s_1 I - A)^{-1} B \cdots (s_\nu I - A)^{-1} B \right], \tag{2.6}$$

and the generalized observability matrix

$$\bar{W} = \left[(s_{\nu+1} I - A^*)^{-1} C^* \cdots (s_{2\nu} I - A^*)^{-1} C^* \right]. \tag{2.7}$$

Assume that $\det(\bar{W}^* \bar{V}) \neq 0$, *then the transfer function of the reduced order model (2.2) with and $V = \bar{V}$ and $W = \bar{W}(\bar{V}^* \bar{W})^{-1}$ interpolates the transfer function of system (2.1) at the points s_j, with $j = 1, \dots, 2\nu$.*

Exploiting this result, the problem of preservation of passivity and stability has been solved in [55, 56], as reported here.

Lemma 2.1 [1] *If the interpolation points in Proposition 2.3 are chosen so that s_j, with $j = 1, \dots, \nu$, are stable spectral zeros, i.e., they are such that $W^*(-s_i) + W(s_i) = 0$, and $s_{j+\nu} = -s_j$, with $j = 1, \dots, \nu$, i.e., the interpolation points are chosen as zeros of the spectral factors and their mirror images, then the projected system is both stable and passive.*

We can now indicate the following drawbacks in the Krylov methods.

- There is no systematic technique to preserve important properties of the system, for instance maintaining prescribed eigenvalues, relative degree, zeros, L_2-gain, or preserving compartmental constraints.
- When a method capable of preserving some of these properties (such as stability and passivity) is presented, it usually implies that specific moments are matched. Hence, the designer cannot chose arbitrary moments. Moreover, there is a lack of system theoretic understanding behind why a particular interpolation point is related to a property like passivity.
- In Lemma 2.1 all the free parameters (the matrix W) are used and no additional property can be preserved.
- Finally, the interpolation-based methods cannot be applied to nonlinear systems (or more general classes of systems), since for these we cannot define a transfer function.

A possible solution to these issues is offered by the "steady-state-based" approach to moment matching. While the first three points are addressed in [48], we focus the rest of the chapter on the last problem: the model reduction of general classes of nonlinear systems.

2.3 The Steady-State Approach

As just observed the interpolation approach cannot be extended to nonlinear systems for which the idea of interpolating points in the complex plane partially loses its meaning (see, however, [57, 58] for some results on the interpolation problem for nonlinear systems). In [48] (see also [14, 59]) a characterization of moment for system (2.1) has been given in terms of the solution of a Sylvester equation as follows.

Lemma 2.2 [48] *Consider system (2.1), $s_i \in \mathbb{C} \setminus \sigma(A)$, for all $i = 1, \dots, \eta$. There exists a one-to-one*[1] *relation between the moments $\eta_0(s_1)$, …, $\eta_{k_1-1}(s_1)$, …, $\eta_0(s_\eta)$, …, $\eta_{k_\eta-1}(s_\eta)$, and the matrix $C\Pi$, where Π is the unique solution of the Sylvester equation*

$$A\Pi + BL = \Pi S, \tag{2.8}$$

with $S \in \mathbb{R}^{\nu\times\nu}$ any non-derogatory[2] *matrix with characteristic polynomial*

$$p(s) = \prod_{i=1}^{\eta}(s - s_i)^{k_i}, \tag{2.9}$$

where $\nu = \sum_{i=1}^{\eta} k_i$, and L is such that the pair (L, S) is observable.

The importance of this formulation, which has resulted in several developments in the area of model reduction by moment matching, see, e.g., [60, 61] and [49–51, 62–66], is that it establishes, through the Sylvester equation (2.8), a relation between the moments and the steady-state response of the output of the system. Before proceeding further we provide a formal definition of steady-state response. With abuse of notation, we indicate the state of a (linear, nonlinear, or more general) dynamical system as $x(t, x_0)$ to highlight the dependency on time and on the initial condition.

Definition 2.3 ([67, 68]) Let $\mathscr{B} \subset \mathbb{R}^n$ and suppose $x(t, x_0)$ is defined for all $t \geq 0$ and all $x_0 \in \mathscr{B}$. The *ω-limit set of the set* denoted by $w(\mathscr{B})$, is the set of all points x for which there exists a sequence of pairs $\{x_k, t_k\}$, with $x_k \in \mathscr{B}$ and $\lim_{k\to\infty} t_k = \infty$ such that $\lim_{k\to\infty} x(t_k, x_k) = x$.

[1] The matrices A, B, C, and the zeros of (2.9) fix the moments. Then, given any observable pair (L, S) with S a non-derogatory matrix with characteristic polynomial (2.9), there exists an invertible matrix $T \in \mathbb{R}^{\nu\times\nu}$ such that the elements of the vector $C\Pi T^{-1}$ are equal to the moments.

[2] A matrix is non-derogatory if its characteristic and minimal polynomials coincide.

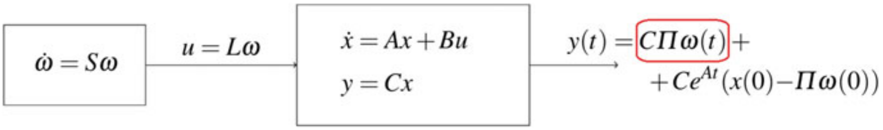

$\dot{\omega} = S\omega$
$u = L\omega$
$\dot{x} = Ax + Bu$
$y = Cx$
$y(t) = \boxed{C\Pi\omega(t)} +$
$+ Ce^{At}(x(0) - \Pi\omega(0))$

Fig. 2.2 Diagrammatic illustration of Theorem 2.1. The term denoting the steady-state response is circled

Definition 2.4 ([67, 68]) Suppose the responses of the system, with initial conditions in a closed and positively invariant set $\mathscr{X}$, are ultimately bounded. A *steady-state response* is any response with initial condition $x_0 \in w(\mathscr{B})$.

Exploiting the notion of steady-state response we can introduce the following result, which is illustrated in Fig. 2.2.

Theorem 2.1 [48] *Consider system (2.1), $s_i \in \mathbb{C} \setminus \sigma(A)$, for all $i = 1, \ldots, \eta$, and $\sigma(A) \subset \mathbb{C}_{<0}$. Let $S \in \mathbb{R}^{\nu\times\nu}$ be any non-derogatory matrix with characteristic polynomial (2.9). Consider the interconnection of system (2.1) with the system*

$$\dot{\omega} = S\omega, \qquad u = L\omega, \tag{2.10}$$

with L and $\omega(0)$ such that the triple $(L, S, \omega(0))$ is minimal. Then there exists a one-to-one relation between the moments $\eta_0(s_1)$, ..., $\eta_{k_1-1}(s_1)$, ..., $\eta_0(s_\eta)$, ..., $\eta_{k_\eta-1}(s_\eta)$, and the steady-state response of the output y of such interconnected system.

Remark 2.1 [69] The minimality of the triple $(L, S, \omega(0))$ implies the observability of the pair (L, S) and the "controllability" of the pair $(S, \omega(0))$. This last condition, called *excitability* of the pair $(S, \omega(0))$, is a geometric characterization of the property that the signals generated by (2.10) are persistently exciting, see [70].

Remark 2.2 By one-to-one relation we mean that the moments are uniquely determined by the steady-state response of $y(t)$ and *vice versa*. Exploiting this fact, in [50] the problem of computing the moments of an unknown linear systems from input/output data has been addressed. Therein an algorithm that, given the signal ω and the output y, retrieves the moments of a system for which the matrices A, B, and C are not known is devised.

The reduction technique based on this notion of moment consists in the interpolation of the steady-state response of the output of the system: a reduced order model is such that its steady-state response is equal to the steady-state response of the output of system (2.1) (provided it exists). Thus, the problem of model reduction by moment matching has been changed from a problem of interpolation of points to a problem of interpolation of signals. The output of the reduced order model has to behave as the output of the original system for a class of input signals, a concept which can be translated to nonlinear systems, time-delay systems, and infinite dimensional systems, [48, 49]. This fact also highlights how important for the moment matching techniques is to let the designer choose the interpolation points, which are related to the class of inputs to the system.

2.4 Model Reduction by Moment Matching for Nonlinear Systems

We can now extend the steady-state description of moment to nonlinear systems.[3] Consider a nonlinear, single-input, single-output, continuous-time system described by the equations

$$\dot{x} = f(x, u), \qquad y = h(x), \tag{2.11}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, f and h smooth mappings, a signal generator described by the equations

$$\dot{\omega} = s(\omega), \qquad u = l(\omega), \tag{2.12}$$

with $\omega(t) \in \mathbb{R}^\nu$, s and l smooth mappings, and the interconnected system

$$\dot{\omega} = s(\omega), \qquad \dot{x} = f(x, l(\omega)), \qquad y = h(x). \tag{2.13}$$

In addition, suppose that $f(0,0) = 0$, $s(0) = 0$, $l(0) = 0$, and $h(0) = 0$. Similarly, to the linear case the interconnection of system (2.11) with the signal generator captures the property that we are interested in preserving the behavior of the system only for *specific* input signals. The following assumptions and definitions provide a generalization of the notion of moment.

Assumption 2.1 The signal generator (2.12) is observable, i.e., for any pair of initial conditions $\omega_a(0)$ and $\omega_b(0)$, such that $\omega_a(0) \neq \omega_b(0)$, the corresponding output trajectories $l(\omega_a(t))$ and $l(\omega_b(t))$ are such that $l(\omega_a(t)) - l(\omega_b(t)) \not\equiv 0$, and Poisson stable[4] with $\omega(0) \neq 0$.

Assumption 2.2 The zero equilibrium of the system $\dot{x} = f(x, 0)$ is locally exponentially stable.

Lemma 2.3 [48] *Consider system (2.11) and the signal generator (2.12). Suppose Assumptions 2.1 and 2.2 hold. Then there is a unique mapping π, locally defined in a neighborhood of $\omega = 0$, which solves the partial differential equation*

$$\frac{\partial \pi}{\partial \omega} s(\omega) = f(\pi(\omega), l(\omega)). \tag{2.14}$$

Remark 2.3 Lemma 2.3 implies that the interconnected system (2.13) possesses an invariant manifold described by the equation $x = \pi(\omega)$.

Definition 2.5 Consider system (2.11) and the signal generator (2.12). Suppose Assumption 2.1 holds. The function $h \circ \pi$, with π solution of equation (2.14), is the *moment of system* (2.11) *at* (s, l).

[3] Note that the results of this section are local.

[4] See [53, Chapter 8] for the definition of Poisson stability.

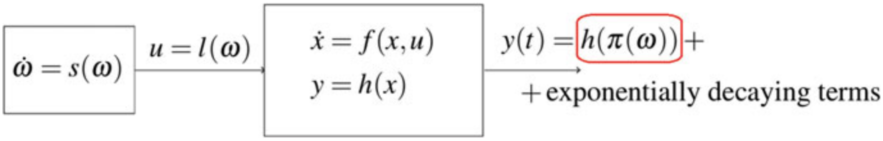

$\dot{\omega} = s(\omega)$
$u = l(\omega)$
$\dot{x} = f(x,u)$
$y = h(x)$
$y(t) = \boxed{h(\pi(\omega))} +$
$+$ exponentially decaying terms

Fig. 2.3 Diagrammatic illustration of Theorem 2.2. The term denoting the steady-state response is circled

Theorem 2.2 [48] *Consider system (2.11) and the signal generator (2.12). Suppose Assumptions 2.1 and 2.2 hold. Then the moment of system (2.11) at (s, l) coincides with the steady-state response of the output of the interconnected system (2.13).*

The result is illustrated in Fig. 2.3 which represents the nonlinear counterpart of Fig. 2.2.

Remark 2.4 [48] If the equilibrium $x = 0$ of the system $\dot{x} = f(x, 0)$ is unstable, it is still possible to define the moment of system (2.11) at (s, l) in terms of the function $h \circ \pi$, provided the equilibrium $x = 0$ is hyperbolic and the system (2.12) is Poisson stable, although it is not possible to establish a relation with the steady-state response of the interconnected system (2.13).

Remark 2.5 [48] While for linear systems it is possible to define k-moments for every $s_i \in \mathbb{C}$ and for any $k \geq 0$, for nonlinear systems it may be difficult, or impossible, to provide general statements if the signal u, generated by system (2.12), is unbounded. Therefore, we assume that the signal generator generates bounded signals. For linear systems this assumption implies that we consider only points $s_i \in \mathbb{C}$ that are distinct and with zero real part.

2.4.1 The Frequency Response of a Nonlinear System

In [48], see also [71, 72], a nonlinear enhancement of the notion of frequency response of a linear system has been derived exploiting the steady-state description of moment. Note that this result is loosely related to the analysis in [66] where a generalization of the phasor transform based on the notion of moment is proposed. Consider system (2.11) and the signal generator (2.12). Let the signal generator (2.12) be such that

$$s(\omega) = \begin{bmatrix} 0 & \bar{\omega} \\ -\bar{\omega} & 0 \end{bmatrix} \omega, \qquad\qquad l(\omega) = \begin{bmatrix} L_1 & L_2 \end{bmatrix} \omega,$$

with $\omega(0) \neq 0$, $\bar{\omega} \neq 0$, and $L_1^2 + L_2^2 \neq 0$. Then, under Assumptions 2.1 and 2.2 the output of the interconnected system (2.13) converges toward a locally well-defined steady-state response, which, by definition, does not depend upon the initial condition $x(0)$. Moreover, such a steady-state response is periodic, hence, if it has the same period of $l(\omega(t))$, it can be written in Fourier series as $h(\pi(\omega(t))) = \sum_{k=-\infty}^{\infty} c_k e^{\imath k\bar{\omega}t}$, with $c_k \in \mathbb{C}$. Consider now the operator $\mathscr{P}_+$ which acts on a Fourier series as follows

$$\mathscr{P}_+\left(\sum_{k=-\infty}^{\infty} c_k e^{\imath k\bar{\omega}t}\right) = \sum_{k=0}^{\infty} \alpha_k e^{\imath k\bar{\omega}t},$$

with $\alpha_k \in \mathbb{C}$. With this operator we can define the frequency response of the nonlinear system (2.11) as

$$F(t, \omega(0), \bar{\omega}) = \frac{\mathscr{P}_+(h(\pi(\omega(t))))}{\mathscr{P}_+(l(\omega(t))}.$$

This function depends upon the frequency $\bar{\omega}$, just as in the linear case, and, unlike the linear case, upon the initial condition $\omega(0)$ of the signal generator and time. Note finally that if the system (2.11) were linear, hence described by the Eq. (2.1), then $F(t, \omega(0), \bar{\omega})$ would be constant with respect to t and equal to $|W(\imath\bar{\omega})|e^{\imath\angle W(\imath\bar{\omega})}$, where $W(s) = C(sI - A)^{-1}B$, $|\cdot|$ indicates the absolute value operator and $\angle$ the phase operator.

2.4.2 Moment Matching

We are now ready to introduce the notion of reduced order model by moment matching for nonlinear systems.

Definition 2.6 [48] Consider the signal generator (2.12). The system described by the equations

$$\dot{\xi} = \phi(\xi, u), \qquad \psi = \kappa(\xi), \tag{2.15}$$

with $\xi(t) \in \mathbb{R}^\nu$, is a *model at* (s, l) *of system* (2.11) if system (2.15) has the same moment at (s, l) as (2.11). In this case, system (2.15) is said to *match* the moment of system (2.11) at (s, l). Furthermore, system (2.15) is a reduced order model of system (2.11) if $\nu < n$.

Lemma 2.4 *Consider system (2.11), system (2.15) and the signal generator (2.12). Suppose Assumptions 2.1 and 2.2 hold. System (2.15)* matches *the moments of (2.11) at* (s, l) *if the equation*

$$\phi(p(\omega), l(\omega)) = \frac{\partial p}{\partial \omega} s(\omega) \tag{2.16}$$

has a unique solution p *such that*

$$h(\pi(\omega)) = \kappa(p(\omega)), \tag{2.17}$$

where π *is the (unique) solution of equation (2.14).*

In other words, we have to determine mappings ϕ, κ, and p such that Eqs. (2.16) and (2.17) hold. We introduce the following assumption to simplify the problem.

Assumption 2.3 There exist mappings κ and p such that $\kappa(0) = 0$, $p(0) = 0$, p is locally continuously differentiable, Eq. (2.17) holds and $\det \left.\frac{\partial p(\omega)}{\partial \omega}\right|_{\omega=0} \neq 0$, i.e., the mapping p possesses a local inverse p^{-1}.

Remark 2.6 [48] Similar to the linear case, Assumption 2.3 holds selecting $p(\omega) = \omega$ and $k(\omega) = h(\pi(\omega))$.

Finally, as shown in [48], the system described by the equations

$$\dot{\xi} = s(\xi) - \delta(\xi)l(\xi) + \delta(\xi)u, \qquad \psi = h(\pi(\xi)), \tag{2.18}$$

where δ is any mapping such that the equation

$$\frac{\partial p}{\partial \omega}s(\omega) = s(p(\omega)) - \delta(p(\omega))l(p(\omega)) + \delta(p(\omega))l(\omega), \tag{2.19}$$

has the unique solution $p(\omega) = \omega$, is a family of *reduced order models of* (2.11) *at* (s, l).

2.4.3 Model Reduction by Moment Matching with Additional Properties

We can determine the conditions on the mapping δ such that the reduced order model satisfies additional properties. The proofs are omitted and can be found in [48].

2.4.3.1 Matching with Asymptotic Stability

Consider the problem of determining a reduced order model (2.18) which has an asymptotically stable zero equilibrium. This problem can be solved if it is possible to select the mapping δ such that the zero equilibrium of the system $\dot{\xi} = s(\xi) - \delta(\xi)l(\xi)$ is locally asymptotically stable. To this end, for instance, it is sufficient that the pair $\left(\left.\frac{\partial l(\xi)}{\partial \xi}\right|_{\xi=0}, \left.\frac{\partial s(\xi)}{\partial \xi}\right|_{\xi=0} \right)$ is observable.

2.4.3.2 Matching with Prescribed Relative Degree

The problem of constructing a reduced order model which has a given relative degree $r \in [1, \nu]$ at some point $\bar{\xi}$ can be solved selecting δ as follows.

Theorem 2.3 [48] *For all $r \in [1, \nu]$ there exists a δ such that system (2.18) has relative degree r at $\bar{\xi}$ if and only if the codistribution*

$$d\mathscr{O}_\nu(\xi) = \operatorname{span}\{dh(\pi(\xi)), \cdots, dL_s^{\nu-1}h(\pi(\xi))\} \tag{2.20}$$

has dimension ν at $\bar{\xi}$.

2.4.3.3 Matching with Prescribed Zero Dynamics

Consider system (2.18) and the problem of determining the mapping δ such that the model has zero dynamics with specific properties. If $\bar{\xi}$ is an equilibrium of system (2.18), the problem is solved selecting δ such that the codistribution (2.20) has dimension ν at $\bar{\xi}$ [48]. Then there is a δ such that the zero dynamics of system (2.18) have a locally exponentially stable equilibrium and there is a coordinate transformation, locally defined around $\bar{\xi}$, such that the zero dynamics are described by the equations

$$\begin{aligned} \dot{z}_1 &= z_2 + \hat{\delta}_1(z)z_1, \\ \dot{z}_2 &= z_3 + \hat{\delta}_2(z)z_1, \\ &\vdots \\ \dot{z}_{\nu-r} &= \hat{f}(z) + \hat{\delta}_{\nu-r}(z)z_1, \end{aligned} \tag{2.21}$$

where the $\hat{\delta}_i$ are free functions and

$$\hat{f}(z) = \tilde{f}(\mathscr{Z})|_{\mathscr{Z}=[0,\ldots,0,z_1,\ldots,z_{\nu-r}]^\top},$$

with $\mathscr{Z} = \Xi(\xi)$ and $\tilde{f}(\mathscr{Z}) = L_s^\nu h(\pi(\Xi^{-1}(\mathscr{Z})))$.

2.4.3.4 Matching with a Passivity Constraint

Consider now the problem of selecting the mapping δ such that system (2.18) is lossless or passive. For such a problem the following fact holds.

Theorem 2.4 [48] *The family of reduced order models (2.18) contains, locally around $\bar{\xi}$, a lossless (passive, respectively) system with a differentiable storage function if there exists a differentiable function V, locally positive definite around $\bar{\xi}$, such that equation*[5]

$$V_\xi s(\xi) = h(\pi(\xi))l(\xi), \quad (V_\xi s(\xi) \le h(\pi(\xi))l(\xi) \text{ respectively}), \tag{2.22}$$

holds locally around $\bar{\xi}$ and

[5] V_ξ and $V_{\xi\xi}$ denote, respectively, the gradient and the Hessian matrix of the scalar function $V : \xi \mapsto V(\xi)$.

$$V_{\xi\xi}(\xi) > 0. \tag{2.23}$$

2.4.3.5 Matching with L_2-gain

We now consider the problem of selecting the mapping δ such that system (2.18) has a given L_2-gain.

Theorem 2.5 *[48] The family of reduced order models (2.18) contains, locally around $\bar{\xi}$, a system with L_2-gain not larger than $\ell > 0$, and with a differentiable storage function if there exists a differentiable function V, locally positive definite around $\bar{\xi}$, such that Eq. (2.23) holds and*

$$V_{\xi}s(\xi) + (h(\pi(\xi)))^2 \le \ell^2 l^2(\xi), \tag{2.24}$$

holds locally around $\bar{\xi}$.

2.5 Model Reduction for Nonlinear Time-Delay Systems

Exploiting the steady-state notion of moment an extension of the model reduction method for nonlinear time-delay systems is given. To keep the notation simple we consider, without loss of generality, only delays (discrete or distributed) in the state and in the input, i.e., the output is delay-free. The neutral case is briefly discussed at the end of the section.

2.5.1 Definition of π: Nonlinear Time-Delay Systems

Consider a nonlinear, single-input, single-output, continuous-time, time-delay system described by the equations

$$\begin{aligned} &\dot{x} = f(x_{\tau_0}, \dots, x_{\tau_\varsigma}, u_{\tau_{\varsigma+1}}, \dots, u_{\tau_\mu}), && y = h(x), \\ &x(\theta) = \phi(\theta), && -T \le \theta \le 0, \end{aligned} \tag{2.25}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, $\phi \in \mathfrak{R}^n_T$, $\tau_0 = 0$, $\tau_j \in \mathbb{R}_{>0}$ with $j = 1, \dots, \mu$ and f and h smooth mappings. Consider a signal generator (2.12) and the interconnected system

$$\dot{\omega} = s(\omega), \qquad \dot{x} = f(x_{\tau_0}, \dots, x_{\tau_\varsigma}, l(\omega_{\tau_{\varsigma+1}}), \dots, l(\omega_{\tau_\mu})), \qquad y = h(x). \tag{2.26}$$

Suppose that $f(0, \dots, 0, 0, \dots, 0) = 0$, $s(0) = 0$, $l(0) = 0$ and $h(0) = 0$.

Assumption 2.4 The zero equilibrium of the system $\dot{x} = f(x_{\tau_0}, \ldots, x_{\tau_\varsigma}, 0, \ldots, 0)$ is locally exponentially stable.

Lemma 2.5 [49, 53] *Consider system (2.25) and the signal generator (2.12). Suppose Assumptions 2.1 and 2.4 hold. Then there exists a unique mapping π, locally defined in a neighborhood of $\omega = 0$, which solves the partial differential equation*

$$\frac{\partial \pi}{\partial \omega} s(\omega) = f(\pi(\bar{\omega}_{\tau_0}), \ldots, \pi(\bar{\omega}_{\tau_\varsigma}), l(\bar{\omega}_{\tau_{\varsigma+1}}), \ldots, l(\bar{\omega}_{\tau_\mu})), \tag{2.27}$$

where $\bar{\omega}_{\tau_i} = \Phi^s_{\tau_i}(\omega)$, with $i = 0, \ldots, \mu$, is the flow of the vector field s at $-\tau_i$.

Remark 2.7 Lemma 2.5 implies that the interconnected system (2.26) possesses an invariant manifold, described by the equation $x = \pi(\omega)$. Note that the partial differential equation (2.27) is independent of time (as (2.14) in the delay-free case), e.g., if $s(\omega) = S\omega$ then $\bar{\omega}_{\tau_i} = e^{-S\tau_i}\omega$.

Definition 2.7 Consider system (2.25) and the signal generator (2.12). Suppose Assumption 2.1 holds. The function $h \circ \pi$, with π solution of equation (2.27), is the *moment of system* (2.25) *at* (s, l).

Theorem 2.6 [49] *Consider system (2.25) and the signal generator (2.12). Suppose Assumptions 2.1 and 2.4 hold. Then the moment of system (2.25) at (s, l) coincides with the steady-state response of the output of the interconnected system (2.26).*

2.5.2 Reduced Order Models for Nonlinear Time-Delay Systems

In this section two families of models achieving moment matching are given.

Definition 2.8 Consider system (2.25) and the signal generator (2.12). Suppose Assumption 2.1 and 2.4 hold. Then the system

$$\dot{\xi} = \phi(\xi_{\chi_0}, \ldots, \xi_{\chi_{\hat{\rho}}}, u_{\chi_{\hat{\rho}+1}}, \ldots, u_{\chi_\rho}), \qquad \psi = \kappa(\xi), \tag{2.28}$$

with $\xi(t) \in \mathbb{R}^\nu$, $u(t) \in \mathbb{R}$, $\psi(t) \in \mathbb{R}$, $\chi_0 = 0$, $\chi_j \in \mathbb{R}_{>0}$ with $j = 1, \ldots, \rho$, and ϕ and κ smooth mappings, is a *model of system* (2.25) *at* (s, l) if system (2.28) has the same moment of system (2.25) at (s, l).

Lemma 2.6 *Consider system (2.25) and the signal generator (2.12). Suppose Assumption 2.1 and 2.4 hold. Then the system (2.28) is a model of system (2.25) at (s, l) if the equation*

$$\frac{\partial p}{\partial \omega} s(\omega) = \phi(p(\bar{\omega}_{\chi_0}), \ldots, p(\bar{\omega}_{\chi_{\hat{\rho}}}), l(\bar{\omega}_{\chi_{\hat{\rho}+1}}), \ldots, l(\bar{\omega}_{\chi_\rho})), \tag{2.29}$$

where $\bar{\omega}_{\chi_i} = \Phi^s_{\chi_i}(\omega)$, with $i = 0, \ldots, \rho$, has a unique solution p such that

$$h(\pi(\omega)) = \kappa(p(\omega)), \tag{2.30}$$

where π is the unique solution of (2.27). System (2.28) is a reduced order model of system (2.25) at (s, l) if $\nu < n$, or if $\hat{\rho} < \varsigma$, or if $\rho < \mu$.

Similarly to the delay-free case we use part of the free mappings to obtain a simpler family of models.

Assumption 2.5 There exist mappings κ and p such that $\kappa(0) = 0$, $p(0) = 0$, p is locally continuously differentiable, Eq. (2.30) holds and p has a local inverse p^{-1}.

Consistently with Lemma 2.6, a family of models that achieves moment matching at (s, l) is described by

$$\begin{aligned} \dot{\xi} &= \boldsymbol{\Phi}(\xi, \bar{\xi}_{\chi_1}, \dots, \bar{\xi}_{\chi_{\hat{\rho}}}) + \tfrac{\partial p(\omega)}{\partial \omega}\gamma(\xi_{\chi_1}, \dots, \xi_{\chi_{\hat{\rho}}}) + \tfrac{\partial p(\omega)}{\partial \omega} \textstyle\sum_{j=\hat{\rho}+1}^{\rho} \delta_j(\xi) u_{\chi_j}, \\ \psi &= \kappa(\xi), \end{aligned} \tag{2.31}$$

with

$$\begin{aligned} \boldsymbol{\Phi}(\xi, \bar{\xi}_{\chi_1}, \dots, \bar{\xi}_{\chi_{\hat{\rho}}}) = \Big[\tfrac{\partial p(\omega)}{\partial \omega}(s(\omega) - \gamma(p(\bar{\omega}_{\chi_1}), \dots, p(\bar{\omega}_{\chi_{\hat{\rho}}}))- \\ - \textstyle\sum_{j=\hat{\rho}+1}^{\rho} \delta_j(p(\omega)) l(\bar{\omega}_{\chi_j})) \Big]_{\omega = p^{-1}(\xi)}, \end{aligned}$$

where $\bar{\xi}_{\chi_j} = \left[\bar{\omega}_{\chi_j}\right]_{\omega=p^{-1}(\xi)}$, κ and p are such that Assumption 2.5 holds, p is the unique solution of (2.29) and δ_j and γ are free mappings.
Assumption 2.5 holds with the selection $p(\omega) = \omega$ and $\kappa(\omega) = h(\pi(\omega))$. This yields a family of models described by the equations

$$\begin{aligned} \dot{\xi} &= s(\xi) - \sum_{j=\hat{\rho}+1}^{\rho} \delta_j(\xi) l(\bar{\xi}_{\chi_j}) - \gamma(\bar{\xi}_{\chi_1}, \dots, \bar{\xi}_{\chi_{\hat{\rho}}}) + \gamma(\xi_{\chi_1}, \dots, \xi_{\chi_{\hat{\rho}}}) + \textstyle\sum_{j=\hat{\rho}+1}^{\rho} \delta_j(\xi) u_{\chi_j}, \\ \psi &= h(\pi(\xi)), \end{aligned} \tag{2.32}$$

where δ_j and γ are arbitrary mappings such that Eq. (2.29), namely

$$\begin{aligned} \tfrac{\partial p}{\partial \omega} s(\omega) = s(p(\omega)) - \textstyle\sum_{j=\hat{\rho}+1}^{\rho} \delta_j(p(\omega)) l(p(\bar{\omega}_{\chi_l})) - \gamma(p(\bar{\omega}_{\chi_1}), \dots, p(\bar{\omega}_{\chi_{\hat{\rho}}})) \\ + \textstyle\sum_{j=\hat{\rho}+1}^{\rho} \delta_j(p(\omega)) l(\omega_{\chi_j}) + \gamma(p(\omega_{\chi_1}), \dots, p(\omega_{\chi_{\hat{\rho}}})), \end{aligned}$$

has the unique solution $p(\omega) = \omega$.

The nonlinear model (2.32) has several free design parameters, namely δ_j, γ, χ_j, $\hat{\rho}$ and ρ. We note that selecting $\gamma \equiv 0$, $\hat{\rho} = 0$, $\rho = 1$ and $\chi_1 = 0$ (in this case we define $\delta = \delta_1$), yields a family of reduced order models with no delays. This family coincides with the family (2.18) and all results of Sect. 2.4.3 are directly applicable: the mapping δ can be selected to achieve matching with asymptotic stability, matching with prescribed relative degree, etc. However, note that the choice of eliminating the delays may destroy some important dynamics of the model.

Remark 2.8 The results of this section can be extended to more general classes of time-delay systems provided that, for such systems, the center manifold theory applies. In particular, one can consider the class of neutral differential time-delay systems described by equations of the form

$$\begin{aligned} d(\dot{x}_{\tau_0}, \dots, \dot{x}_{\tau_{\varsigma_1}}) &= f(x_{\tau_{\varsigma_1+1}}, \dots, x_{\tau_{\varsigma_2}}, u_{\tau_{\varsigma_2+1}}, \dots, u_{\tau_\mu}), \\ y &= h(x), \end{aligned} \tag{2.33}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, $\tau_0 = 0$, $\tau_j \in \mathbb{R}_{>0}$ with $j = 1, \dots, \mu$ and d, f, and h smooth mappings. The center manifold theory does not hold for this class of systems for a general mapping d. Specific cases have to be considered and we refer the reader to [73–75] and references therein. Note, however, that for the simple case

$$\begin{aligned} \dot{x} + D\dot{x}_{\tau_1} &= f(x_{\tau_2}, \dots, x_{\tau_{\varsigma_1}}, u_{\tau_{\varsigma_1+1}}, \dots, u_{\tau_\mu}), \\ y &= h(x), \end{aligned} \tag{2.34}$$

with $D \in \mathbb{R}^{n \times n}$, the center manifold theory holds as for standard time-delay systems if the matrix D is such that $\sigma(D) \subset \mathbb{D}_{<1}$.

2.5.3 Exploiting One Delay to Match $h \circ \pi_a$ and $h \circ \pi_b$

In this section we show how to exploit the free parameters to achieve moment matching at two moments $h \circ \pi_a$ and $h \circ \pi_b$ maintaining the same number of equations describing the reduced order model. Consider system (2.25) and, to simplify the exposition, the signal generators described by the linear equation

$$\dot{\omega} = S_a \omega, \qquad\qquad u = L_{ab}\omega, \tag{2.35}$$

Note that, as highlighted in [48], considering the model reduction problem for nonlinear systems when the signal generator is a linear system is of particular interest since the reduced order models have a very simple description, i.e., a family of reduced order models is described by a linear differential equation with a nonlinear output map. This observation holds true also in the case of time-delay systems, namely a nonlinear time-delay system can be approximated by a linear time-delay equation with a nonlinear output map. This structure has two main advantages. Firstly, the selection of the free parameters that achieve additional goals, such as to assign the eigenvalues or the relative degree of the reduced order model, is remarkably simplified. Secondly, the computation of the reduced order model boils down to the computation of the output map $h \circ \pi$. A technique to approximate this mapping is proposed in the next section. As a consequence of this discussion, a reduced order model of system (2.25) at (S_a, L_{ab}) is given by the family

$$\begin{aligned} \dot{\xi} &= F_0\xi + F_1\xi_\chi + G_2 u + G_3 u_\chi, \\ \psi &= \kappa_0(\xi) + \kappa_1(\xi_\chi), \end{aligned} \tag{2.36}$$

with κ_0 and κ_1 smooth mappings, if there exists a unique matrix P_a such that

$$\begin{aligned} F_0 P_a + F_1 P_a e^{-S_a\chi} - P_a S_a &= -G_2 L_{ab} - G_3 L_{ab} e^{-S_a\chi}, \\ h(\pi_a(\omega)) &= \kappa_0(P_a\omega) + \kappa_1(P_a e^{-S_a\chi}\omega), \end{aligned} \tag{2.37}$$

Consider now another signal generator described by the linear equation

$$\dot{\omega} = S_b\omega, \qquad\qquad u = L_{ab}\omega, \tag{2.38}$$

and the problem of selecting F_0, F_1, G_2, G_3, κ_0, and κ_1 such that the reduced order model (2.36) matches the moments of system (2.25) at (S_a, L_{ab}) and (S_b, L_{ab}).

Proposition 2.4 *Let $S_a \in \mathbb{R}^{\nu\times\nu}$ and $S_b \in \mathbb{R}^{\nu\times\nu}$ be two non-derogatory matrices such that $\sigma(S_a) \cap \sigma(S_b) = \emptyset$ and let L_{ab} be such that the pairs (L_{ab}, S_a) and (L_{ab}, S_b) are observable. Let $\pi_a(\omega) = \pi(\omega)$ be the unique solution of (2.27), with $L = L_{ab}$ and $S = S_a$, and let $\pi_b(\omega) = \pi(\omega)$ be the unique solution of (2.27), with $L = L_{ab}$ and $S = S_b$. Then system (2.36) with the selection*

$$\begin{aligned} F_1 &= (S_b - S_a - G_3(e^{-S_b\chi} - e^{-S_a\chi_3}))(e^{-S_b\chi} - e^{-S_a\chi})^{-1}_{,} \\ F_0 &= S_a - G_2 L_{ab} - G_3 L_{ab} e^{-S_a\chi} - F_1 e^{-S_a\chi}, \\ \kappa_0(\omega) &= h(\pi_a(\omega)) - \kappa_1(e^{-S_a\chi}\omega), \end{aligned} \tag{2.39}$$

and k_1 a mapping such that

$$\kappa_1\left(e^{-S_b\chi}\omega\right) - \kappa_1\left(e^{-S_a\chi}\omega\right) = h(\pi_b(\omega)) - h(\pi_a(\omega)),$$

is a reduced order model of the nonlinear time-delay system (2.25) achieving moment matching at (S_a, L_{ab}) and (S_b, L_{ab}), for any G_2 and G_3 such that $s_i \notin \sigma(F_0 + F_1 e^{-s\chi})$, for all $s_i \in \sigma(S_a)$ and $s_i \in \sigma(S_b)$.

Proof As showed in the proof of Proposition 2.1 of [49], F_0 and F_1 solve the two Sylvester equations

$$\begin{aligned} F_0 P_a + F_1 P_a e^{-S_a\chi} - P_a S_a &= -G_2 L_{ab} - G_3 L_{ab} e^{-S_a\chi}, \\ F_0 P_b + F_1 P_b e^{-S_b\chi} - P_b S_b &= -G_2 L_{ab} - G_3 L_{ab} e^{-S_b\chi}, \end{aligned} \tag{2.40}$$

with $P_a = P_b = I$. It remains to determine the mappings κ_0 and κ_1 that solve the matching conditions

$$\begin{aligned} h(\pi_a(\omega)) &= \kappa_0(\omega) + \kappa_1\left(e^{-S_a\chi}\omega\right), \\ h(\pi_b(\omega)) &= \kappa_0(\omega) + \kappa_1\left(e^{-S_b\chi}\omega\right). \end{aligned} \tag{2.41}$$

Solving the first equation with respect to δ_0 and substituting the resulting expression in the second yields

$$\kappa_1\left(e^{-S_b\chi}\omega\right) - \kappa_1\left(e^{-S_a\chi}\omega\right) = h(\pi_b(\omega)) - h(\pi_a(\omega)),$$

from which the claim follows.

The family of linear time-delay systems with nonlinear output mapping characterized in Proposition 2.4 matches the moments $h \circ \pi_a$ and $h \circ \pi_b$ of the nonlinear system (2.25). Note that the matrices G_2 and G_3 remain free parameters and they can be used to achieve the properties discussed in Sect. 2.4.3. For instance, G_2 and G_3 can be used to set both the eigenvalues of F_0 and F_1.

Remark 2.9 Proposition 2.4 can be generalized to $\hat{p} > 1$ delays, obtaining a reduced order model that match $(\hat{p}+1)\nu$ moments. The result can also be generalized to nonlinear generators $s_i(\omega)$ assuming that the flow $\Phi^{s_i}_{\chi_i}(\omega)$ is known for all the delays χ_i and that $\gamma(\xi_{\chi_1}, \ldots, \xi_{\chi_{\hat{p}}})$ in (2.32) is replaced by $\hat{\gamma}_1(\xi_{\chi_1}) + \ldots + \hat{\gamma}_{\hat{p}}(\xi_{\chi_{\hat{p}}})$.

Remark 2.10 The number of delays in (2.25) does not play a role in Proposition 2.4. Thus, this result can be applied to reduce a system with an arbitrary number of delays always obtaining a reduced order model with, for example, two delays. This fact can be taken to the "limit" reducing a system which is not a time-delay system. In other words, a system described by ordinary differential equations can be reduced to a system described by time-delay differential equations with an arbitrary number of delays $\hat{p}$ achieving moment matching at $(\hat{p}+1)\nu$ moments.

2.6 Online Nonlinear Moment Estimation from Data

In this section we solve a fundamental problem for the theory we have presented, namely how to compute an approximation of the moment $h \circ \pi$ when the solution of the partial differential equation (2.13) or (2.27) is not known. Note, first of all, that the results of this section hold indiscriminately for delay-free and time-delay systems. In the following we do not even need to know the mappings f and h. In fact we are going to present a method to approximate the moment $h \circ \pi$ directly from input/output data, namely from $\omega(t)$ and $y(t)$. Note that given the exponential stability hypothesis on the system and Theorem 2.2 (Theorem 2.6 for time-delay systems), the equation

$$y(t) = h(\pi(\omega(t))) + \varepsilon(t), \tag{2.42}$$

where $\varepsilon(t)$ is an exponentially decaying signal, holds for the interconnections (2.13) and (2.26). We introduce the following assumption.

Assumption 2.6 The mapping $h \circ \pi$ belongs to the function space identified by the family of continuous basis functions $\varphi_j : \mathbb{R}^\nu \to \mathbb{R}$, with $j = 1, \dots, M$ (M may be ∞), i.e., there exist $\pi_j \in \mathbb{R}$, with $j = 1, \dots, M$, such that

$$h(\pi(\omega)) = \sum_{j=1}^{M} \pi_j \varphi_j(\omega),$$

for any ω.

Let

$$\begin{aligned} \Gamma &= \left[\pi_1 \ \pi_2 \ \dots \ \pi_N \right], \\ \Omega(\omega(t)) &= \left[\varphi_1(\omega(t)) \ \varphi_2(\omega(t)) \ \dots \ \varphi_N(\omega(t)) \right]^\top, \end{aligned}$$

with $N \le M$. Using a weighted sum of basis functions, Eq. (2.42) can be written as

$$y(t) = \sum_{j=1}^{N} \pi_j \varphi_j(\omega(t)) + e(t) + \varepsilon(t) = \Gamma \Omega(\omega(t)) + e(t) + \varepsilon(t), \tag{2.43}$$

where $e(t) = \sum_{N+1}^{M} \pi_j \varphi_j(\omega(t))$ is the error caused by stopping the summation at N. Consider now the approximation

$$y(t) \approx \sum_{j=1}^{N} \widetilde{\pi}_j \varphi_j(\omega(t)) = \widetilde{\Gamma} \Omega(\omega(t)), \tag{2.44}$$

which neglects the approximation error $e(t)$ and the transient error $\epsilon(t)$. Let $T_k^w = \{t_{k-w+1}, \dots, t_{k-1}, t_k\}$, with $0 \le t_0 < t_1 < \dots < t_{k-w} < \dots < t_k < \dots < t_q$, with $w > 0$ and $q \ge w$, and Γ_k be an on-line estimate of the matrix Γ computed at T_k^w, namely computed at the time t_k using the last w instants of time t_i assuming that $e(t)$ and $\epsilon(t)$ are known. Since this is not the case in practice, define $\widetilde{\Gamma}_k = \left[\widetilde{\pi}_1 \ \widetilde{\pi}_2 \ \dots \ \widetilde{\pi}_N \right]$ as the approximation, in the sense of (2.44), of the estimate Γ_k. Finally, we can compute this approximation as follows.

Theorem 2.7 [64] *Define the time-snapshots* $\widetilde{U}_k \in \mathbb{R}^{w \times N}$ *and* $\widetilde{Y}_k \in \mathbb{R}^w$ *as*

$$\widetilde{U}_k = \left[\Omega(\omega(t_{k-w+1})) \ \dots \ \Omega(\omega(t_{k-1})) \ \Omega(\omega(t_k)) \right]^\top$$

and

$$\widetilde{Y}_k = \left[y(t_{k-w+1}) \ \dots \ y(t_{k-1}) \ y(t_k) \right]^\top .$$

If $\widetilde{U}_k$ *is full rank then*

$$\operatorname{vec}(\widetilde{\Gamma}_k) = (\widetilde{U}_k^\top \widetilde{U}_k)^{-1} \widetilde{U}_k^\top \widetilde{Y}_k, \tag{2.45}$$

is an approximation of the estimate Γ_k.

To ensure that the approximation is well-defined for all k, we give an assumption in the spirit of persistency of excitation.

Assumption 2.7 For any $k \geq 0$, there exist $\bar{K} > 0$ and $\alpha > 0$ such that the elements of T_k^K, with $K > \bar{K}$, are such that

$$\frac{1}{K}\widetilde{U}_k^\top \widetilde{U}_k \geq \alpha I.$$

Note that if Assumption 2.7 holds (see [76] for a similar argument), $\widetilde{U}_k^\top \widetilde{U}_k$ is full rank. The next definition is a direct consequence of the discussion we have carried out.

Definition 2.9 The *estimated moment* of system (2.11) (or system (2.25)) is defined as

$$\widetilde{h \circ \pi}_{N,k}(\omega(t)) = \widetilde{\Gamma}_k \Omega(\omega(t)), \tag{2.46}$$

with $\widetilde{\Gamma}_k$ computed with (2.45).

Equation (2.45) is a classic least-square estimator and an efficient recursive formula can be easily derived.

Theorem 2.8 [64] *Assume that* $\boldsymbol{\Phi}_k = (\widetilde{U}_k^\top \widetilde{U}_k)^{-1}$ *and* $\boldsymbol{\Psi}_k = (\widetilde{U}_{k-1}^\top \widetilde{U}_{k-1} + \omega(t_k)\omega(t_k)^\top)^{-1}$ *are full rank for all* $t \geq t_r$ *with* $t_r \geq t_w$. *Given* $\operatorname{vec}(\widetilde{\Gamma}_r)$, $\boldsymbol{\Phi}_r$ *and* $\boldsymbol{\Psi}_r$, *the least-square estimation*

$$\begin{aligned} \operatorname{vec}(\widetilde{\Gamma}_k) = \operatorname{vec}(\widetilde{\Gamma}_{k-1}) + \boldsymbol{\Phi}_k \omega(t_k)\Big(y(t_k) - \omega(t_k)^\top \operatorname{vec}(\widetilde{\Gamma}_{k-1})\Big) - \\ -\boldsymbol{\Phi}_k \omega(t_{k-w})\Big(y(t_{k-w}) - \omega(t_{k-w})^\top \operatorname{vec}(\widetilde{\Gamma}_{k-1})\Big), \end{aligned} \tag{2.47}$$

with

$$\boldsymbol{\Phi}_k = \boldsymbol{\Psi}_k - \boldsymbol{\Psi}_k \omega(t_{k-w})(I + \omega(t_{k-w})^\top \boldsymbol{\Psi}_k \omega(t_{k-w}))^{-1}\omega(t_{k-w})^\top \boldsymbol{\Psi}_k \tag{2.48}$$

and

$$\boldsymbol{\Psi}_k = \boldsymbol{\Phi}_{k-1} - \boldsymbol{\Phi}_{k-1}\omega(t_k)(I + \omega(t_k)^\top \boldsymbol{\Phi}_{k-1}\omega(t_k))^{-1}\omega(t_k)^\top \boldsymbol{\Phi}_{k-1}. \tag{2.49}$$

holds for all $t \geq t_r$.

Finally, the following result guarantees that the approximation converges to $h \circ \pi$.

Theorem 2.9 [64] *Suppose Assumptions 2.1 (2.1 for time-delay systems), 2.2 (2.4 for time-delay systems), 2.6 and 2.7 hold. Then*

$$\lim_{t \to \infty} \Big(h(\pi(\omega(t))) - \lim_{N \to M} \widetilde{h \circ \pi}_{N,k}(\omega(t)) \Big) = 0.$$

2.7 Conclusion

In this chapter we have reviewed the model reduction technique for nonlinear, possibly time-delay, systems based on the "steady-state" notion of moment. We have firstly recalled the classical interpolation theory and we have then introduced the steady-state-based notion of moment. Exploiting this description of moment the solution of the problem of model reduction by moment matching for nonlinear systems has been given and an enhancement of the notion of frequency response for nonlinear systems has been presented. Subsequently, these techniques have been extended to nonlinear time-delay systems and the problem of obtaining a family of reduced order models matching two moments has been solved for nonlinear time-delay systems. The review is concluded with a recently presented technique to approximate the moment of nonlinear, possibly time-delay, systems, without solving any partial differential equation.

References

1. Antoulas, A.: Approximation of Large-Scale Dynamical Systems. SIAM Advances in Design and Control, Philadelphia, PA (2005)
2. Adamjan, V.M., Arov, D.Z., Krein, M.G.: Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem. Math. USSR Sb. **15**, 31–73 (1971)
3. Glover, K.: All optimal Hankel-norm approximations of linear multivariable systems and their L^{∞}-error bounds. Int. J. Control **39**(6), 1115–1193 (1984)
4. Safonov, M.G., Chiang, R.Y., Limebeer, D.J.N.: Optimal Hankel model reduction for nonminimal systems. IEEE Trans. Autom. Control **35**(4), 496–502 (1990)
5. Moore, B.C.: Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control **26**(1), 17–32 (1981)
6. Meyer, D.G.: Fractional balanced reduction: model reduction via a fractional representation. IEEE Trans. Autom. Control **35**(12), 1341–1345 (1990)
7. Gray, W.S., Mesko, J.: General input balancing and model reduction for linear and nonlinear systems. In: European Control Conference. Brussels, Belgium (1997)
8. Lall, S., Beck, C.: Error bounds for balanced model reduction of linear time-varying systems. IEEE Trans. Autom. Control **48**(6), 946–956 (2003)
9. Kimura, H.: Positive partial realization of covariance sequences. In: Modeling, Identification and Robust Control, pp. 499–513 (1986)
10. Byrnes, C.I., Lindquist, A., Gusev, S.V., Matveev, A.S.: A complete parameterization of all positive rational extensions of a covariance sequence. IEEE Trans. Autom. Control **40**, 1841–1857 (1995)
11. Georgiou, T.T.: The interpolation problem with a degree constraint. IEEE Trans. Autom. Control **44**, 631–635 (1999)
12. Antoulas, A.C., Ball, J.A., Kang, J., Willems, J.C.: On the solution of the minimal rational interpolation problem. In: Linear Algebra and its Applications, Special Issue on Matrix Problems, vol. 137–138, pp. 511–573 (1990)
13. Byrnes, C.I., Lindquist, A., Georgiou, T.T.: A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint. IEEE Trans. Autom. Control **46**, 822–839 (2001)
14. Gallivan, K.A., Vandendorpe, A., Van Dooren, P.: Model reduction and the solution of Sylvester equations. In: MTNS, Kyoto (2006)

15. Beattie, C.A., Gugercin, S.: Interpolation theory for structure-preserving model reduction. In: Proceedings of the 47th IEEE Conference on Decision and Control, Cancun, Mexico (2008)
16. Al-Baiyat, S.A., Bettayeb, M., Al-Saggaf, U.M.: New model reduction scheme for bilinear systems. Int. J. Syst. Sci. **25**(10), 1631–1642 (1994)
17. Lall, S., Krysl, P., Marsden, J.: Structure-preserving model reduction for mechanical systems. Phys. D **184**, 304–318 (2003)
18. Soberg, J., Fujimoto, K., Glad, T.: Model reduction of nonlinear differential-algebraic equations. In: IFAC Symposium Nonlinear Control Systems, Pretoria, South Africa, vol. 7, pp. 712–717 (2007)
19. Fujimoto, K.: Balanced realization and model order reduction for port-Hamiltonian systems. J. Syst. Des. Dyn. **2**(3), 694–702 (2008)
20. Scherpen, J.M.A., Gray, W.S.: Minimality and local state decompositions of a nonlinear state space realization using energy functions. IEEE Trans. Autom. Control **45**(11), 2079–2086 (2000)
21. Gray, W.S., Scherpen, J.M.A.: Nonlinear Hilbert adjoints: properties and applications to Hankel singular value analysis. In: Proceedings of the 2001 American Control Conference, vol. 5, pp. 3582–3587 (2001)
22. Verriest, E., Gray, W.: Dynamics near limit cycles: model reduction and sensitivity. In: Symposium on Mathematical Theory of Networks and Systems, Padova, Italy (1998)
23. Gray, W.S., Verriest, E.I.: Balanced realizations near stable invariant manifolds. Automatica **42**(4), 653–659 (2006)
24. Kunisch, K., Volkwein, S.: Control of the Burgers equation by a reduced-order approach using proper orthogonal decomposition. J. Optim. Theory Appl. **102**(2), 345–371 (1999)
25. Hinze, M., Volkwein, S.: Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: error estimates and suboptimal control. In: Dimension Reduction of Large-Scale Systems. Lecture Notes in Computational and Applied Mathematics, pp. 261–306. Springer (2005)
26. Willcox, K., Peraire, J.: Balanced model reduction via the proper orthogonal decomposition. AIAA J. **40**(11), 2323–2330 (2002)
27. Kunisch, K., Volkwein, S.: Proper orthogonal decomposition for optimality systems. ESAIM Math. Modell. Numer. Anal. **42**(01), 1–23 (2008)
28. Astrid, P., Weiland, S., Willcox, K., Backx, T.: Missing point estimation in models described by proper orthogonal decomposition. IEEE Trans. Autom. Control **53**(10), 2237–2251 (2008)
29. Lall, S., Marsden, J.E., Glavaski, S.: A subspace approach to balanced truncation for model reduction of nonlinear control systems. Int. J. Robust Nonlinear Control **12**, 519–535 (2002)
30. Fujimoto, K., Tsubakino, D.: Computation of nonlinear balanced realization and model reduction based on Taylor series expansion. Syst. Control Lett. **57**(4), 283–289 (2008)
31. Blondel, V.D., Megretski, A.: Unsolved Problems in Mathematical Systems and Control Theory. Princeton University Press (2004)
32. Mäkilä, P.M., Partington, J.R.: Laguerre and Kautz shift approximations of delay systems. Int. J. Control **72**, 932–946 (1999)
33. Mäkilä, P.M., Partington, J.R.: Shift operator induced approximations of delay systems. SIAM J. Control Optim. **37**(6), 1897–1912 (1999)
34. Zhang, J., Knospe, C.R., Tsiotras, P.: Stability of linear time-delay systems: a delay-dependent criterion with a tight conservatism bound. In: Proceedings of the 2000 American Control Conference, Chicago, IL, pp. 1458–1462, June 2000
35. Al-Amer, S.H., Al-Sunni, F.M.: Approximation of time-delay systems. In: Proceedings of the 2000 American Control Conference, Chicago, IL, pp. 2491–2495, June 2000
36. Banks, H.T., Kappel, F.: Spline approximations for functional differential equations. J. Differ. Equ. **34**, 496–522 (1979)
37. Gu, G., Khargonekar, P.P., Lee, E.B.: Approximation of infinite-dimensional systems. IEEE Trans. Autom. Control **34**(6) (1992)
38. Glover, K., Lam, J., Partington, J.R.: Rational approximation of a class of infinite dimensional system i: singular value of hankel operator. Math. Control Circ. Syst. **3**, 325–344 (1990)

39. Glader, C., Hognas, G., Mäkilä, P.M., Toivonen, H.T.: Approximation of delay systems: a case study. Int. J. Control **53**(2), 369–390 (1991)
40. Ohta, Y., Kojima, A.: Formulas for Hankel singular values and vectors for a class of input delay systems. Automatica **35**, 201–215 (1999)
41. Yoon, M.G., Lee, B.H.: A new approximation method for time-delay systems. IEEE Trans. Autom. Control **42**(7), 1008–1012 (1997)
42. Michiels, W., Jarlebring, E., Meerbergen, K.: Krylov-based model order reduction of time-delay systems. SIAM J. Matrix Anal. Appl. **32**(4), 1399–1421 (2011)
43. Jarlebring, E., Damm, T., Michiels, W.: Model reduction of time-delay systems using position balancing and delay Lyapunov equations. Math. Control Signals Syst. **25**(2), 147–166 (2013)
44. Wang, Q., Wang, Y., Lam, E.Y., Wong, N.: Model order reduction for neutral systems by moment matching. Circuits Syst. Signal Process. **32**(3), 1039–1063 (2013)
45. Ionescu, T.C., Iftime, O.V.: Moment matching with prescribed poles and zeros for infinite-dimensional systems. In: American Control Conference, Montreal, Canada, pp. 1412–1417, June 2012
46. Iftime, O.V.: Block circulant and block Toeplitz approximants of a class of spatially distributed systems-An LQR perspective. Automatica **48**(12), 3098–3105 (2012)
47. Astolfi, A.: Model reduction by moment matching, steady-state response and projections. In: Proceedings of the 49th IEEE Conference on Decision and Control (2010)
48. Astolfi, A.: Model reduction by moment matching for linear and nonlinear systems. IEEE Trans. Autom. Control **55**(10), 2321–2336 (2010)
49. Scarciotti, G., Astolfi, A.: Model reduction of neutral linear and nonlinear time-invariant time-delay systems with discrete and distributed delays. IEEE Trans. Autom. Control **61**(6), 1438–1451 (2016)
50. Scarciotti, G., Astolfi, A.: Model reduction for linear systems and linear time-delay systems from input/output data. In: 2015 European Control Conference, Linz, pp. 334–339, July 2015
51. Scarciotti, G., Astolfi, A.: Model reduction for nonlinear systems and nonlinear time-delay systems from input/output data. In: Proceedings of the 54th IEEE Conference on Decision and Control, Osaka, Japan, 15–18 Dec 2015
52. Richard, J.P.: Time-delay systems: an overview of some recent advances and open problems. Automatica **39**(10), 1667–1694 (2003)
53. Isidori, A.: Nonlinear Control Systems. Communications and Control Engineering, 3rd edn. Springer (1995)
54. Doyle, J.C., Francis, B.A., Tannenbaum, A.R.: Feedback Control Theory. Macmillan, New York (1992)
55. Antoulas, A.C.: A new result on passivity preserving model reduction. Syst. Control Lett. **54**(4), 361–374 (2005)
56. Sorensen, D.C.: Passivity preserving model reduction via interpolation of spectral zeros. Syst. Control Lett. **54**(4), 347–360 (2005)
57. Hespel, C., Jacob, G.: Approximation of nonlinear dynamic systems by rational series. Theor. Comput. Sci. **79**(1), 151–162 (1991)
58. Hespel, C.: Truncated bilinear approximants: Carleman, finite Volterra, Padé-type, geometric and structural automata. In: Jacob, G., Lamnabhi-Lagarrigue, F. (eds.) Algebraic Computing in Control. Lecture Notes in Control and Information Sciences, vol. 165, pp. 264–278. Springer (1991)
59. Gallivan, K., Vandendorpe, A., Van Dooren, P.: Sylvester equations and projection-based model reduction. J. Comput. Appl. Math. **162**(1), 213–229 (2004)
60. Dib, W., Astolfi, A., Ortega, R.: Model reduction by moment matching for switched power converters. In: Proceedings of the 48th IEEE Conference on Decision and Control, Held Jointly with the 28th Chinese Control Conference, pp. 6555–6560, Dec 2009
61. Ionescu, T.C., Astolfi, A., Colaneri, P.: Families of moment matching based, low order approximations for linear systems. Syst. Control Lett. **64**, 47–56 (2014)
62. Scarciotti, G., Astolfi, A.: Characterization of the moments of a linear system driven by explicit signal generators. In: Proceedings of the 2015 American Control Conference, Chicago, IL, pp. 589–594, July 2015

63. Scarciotti, G., Astolfi, A.: Model reduction by matching the steady-state response of explicit signal generators. IEEE Trans. Autom. Control **61**(7), 1995–2000 (2016)
64. Scarciotti, G., Astolfi, A.: Data-driven model reduction by moment matching for linear and nonlinear systems. Automatica **79**, 340–351 (2017)
65. Scarciotti, G.: Low computational complexity model reduction of power systems with preservation of physical characteristics. IEEE Trans. Power Syst. **32**(1), 743–752 (2017)
66. Scarciotti, G., Astolfi, A.: Moment based discontinuous phasor transform and its application to the steady-state analysis of inverters and wireless power transfer systems. IEEE Trans. Power Electron. **31**(12), 8448–8460 (2016)
67. Isidori, A., Byrnes, C.I.: Steady-state behaviors in nonlinear systems with an application to robust disturbance rejection. Annu. Rev. Control **32**(1), 1–16 (2008)
68. Scarciotti, G., Astolfi, A.: Model reduction for hybrid systems with state-dependent jumps. In: IFAC Symposium Nonlinear Control Systems, Monterey, CA, USA, pp. 862–867 (2016)
69. Scarciotti, G., Jiang, Z.P., Astolfi, A.: Constrained optimal reduced-order models from input/output data. In: Proceedings of the 55th IEEE Conference on Decision and Control, Las Vegas, NV, USA, pp. 7453–7458, 12–14 Dec 2016
70. Padoan, A., Scarciotti, G., Astolfi, A.: A geometric characterisation of persistently exciting signals generated by autonomous systems. In: IFAC Symposium Nonlinear Control Systems, Monterey, CA, USA, pp. 838–843 (2016)
71. Isidori, A., Byrnes, C.I.: Steady state response, separation principle and the output regulation of nonlinear systems. In: Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, FL, USA, pp. 2247–2251 (1989)
72. Isidori, A., Byrnes, C.I.: Output regulation of nonlinear systems. IEEE Trans. Autom. Control **35**(2), 131–140 (1990)
73. Hale, J.K.: Theory of functional differential equations. Applied Mathematical Sciences Series. Springer Verlag Gmbh (1977)
74. Hale, J.K.: Behavior near constant solutions of functional differential equations. J. Differ. Equ. **15**, 278–294 (1974)
75. Byrnes, C.I., Spong, M.W., Tarn, T.J.: A several complex variables approach to feedback stabilization of linear neutral delay-differential systems. Math. Syst. Theory **17**(1), 97–133 (1984)
76. Bian, T., Jiang, Y., Jiang, Z.P.: Adaptive dynamic programming and optimal control of nonlinear nonaffine systems. Automatica **50**(10), 2624–2632 (2014)

Chapter 3
Event-Triggered Control of Nonlinear Systems: A Small-Gain Approach

Tengfei Liu and Zhong-Ping Jiang

Abstract This chapter studies the event-triggered control problem for nonlinear systems with input-to-state stability (ISS) as the basic notion and the ISS small-gain theorem as a tool. The contribution of this book chapter is twofold. First, an ISS gain condition is proposed for event-triggered control of nonlinear uncertain systems. It is proved that infinitely fast sampling can be avoided with an appropriately designed event-triggering mechanism if the system is input-to-state stabilizable with measurement error as the external input and the resulted ISS gain is Lipschitz on compact sets. No assumption on the existence of known ISS-Lyapunov functions is made in the discussions. Moreover, the forward completeness problem with event-triggered control is studied systematically by ISS small-gain arguments. Self-triggered control designs for systems under external disturbance are also developed in the ISS-based framework. Second, this chapter introduces a new design method for input-to-state stabilization of nonlinear uncertain systems in the strict-feedback form. It is particularly shown that the ISS gain with the measurement error as the input can be designed to satisfy the proposed condition for event-triggered control.

3.1 Introduction

Tremendous efforts have been made for improved performance of control systems. As an alternative to the periodic data-sampling in traditional sampled-data control systems, the aperiodic event-triggered data-sampling depends on the real-time system state, and in this way, takes into account the system behavior between the sampling time instants. Such new data-sampling strategy has been proved to be quite

T. Liu
State Key Laboratory of Synthetical Automation for Process Industries,
Northeastern University, Shenyang, China
e-mail: tfliu@mail.neu.edu.cn; neuralliu@gmail.com

Z.-P. Jiang (✉)
Tandon School of Engineering, New York University, Brooklyn, NY, USA
e-mail: zjiang@nyu.edu

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_3

useful in reducing the waste of computation and communication resources in feedback control systems. Early results in this direction include [5, 15, 31, 40, 51].

Due to the increasing popularity of networked control systems, recent years have seen a renewed interest in event-triggered control of linear and nonlinear systems. Significant contributions have been made to the literature; see, e.g., [3, 4, 6, 8, 12, 18, 19, 37, 42, 49, 56] and the references therein. Specifically, in [4, 19], impulsive control methods are developed to keep the states of first order stochastic systems inside certain thresholds. In [12, 37], prediction of the real-time system state between the sampling time instants is employed to generate the control signal, and the prediction is corrected by data-sampling when the difference between the true state and the predicted state is larger than a threshold signal. In [49], the sampling error of the system state is considered as measurement error, and the system is assumed to be robustly stabilizable in the presence of the measurement error. Then, the event trigger is designed such that the measurement error caused by data-sampling is bounded by a specific threshold (depending on the real-time system state) for convergence of the system state. Reference [38] proposes a universal formula for event-based stabilization of general nonlinear systems affine in the control by extending Sontag's result for continuous-time stabilization [45]. Reference [50] proposes a Lyapunov condition for tracking control of nonlinear systems. The designs have been extended to distributed control [9, 14, 43, 54], decentralized control [6, 8], systems with quantized measurements [11], and periodic event-triggered control [16], to name a few. In the process of event-triggered control, the real-time system state should be continuously monitored. As an alternative, a self-triggered controller computes the control signal as well as the next sampling time instant such that continuous monitoring of system state is not needed [52]. Recent results on self-triggered control can be found in [1, 2, 7, 39, 41, 53, 54]. See also [17] for a literature review and tutorial of event-triggered control,

For practical implementation of event-triggered control, infinitely fast sampling should be avoided, that is, the intervals between all the sampling time instants should be lower bounded by some positive constant [29]. Note that one special case of infinitely fast sampling is that there is an infinite number of sampling time instants converging to a finite time, which is known as the Zeno behavior [13]. In most of the existing results, the events of data-sampling are triggered by comparing the real-time system state and a threshold signal, and the event-triggered control problems are transformed into problems of choosing appropriate threshold signals to avoid infinitely fast sampling.

The notion of input-to-state stability (ISS), invented by Sontag, is a powerful tool to describe the stability property of nonlinear systems with external inputs [46]. For event-triggered control, ISS has been used to describe the influence of data-sampling to control [2, 39, 49]. In this framework, it is usually assumed that the plant has an input-to-state stabilizing controller with the measurement error caused by data-sampling as the external input. The basic idea is to find an appropriate event-trigger such that the influence of data-sampling is attenuated and the closed-loop system augmented with the event-triggered sampling is ISS. A special case is that the system is disturbance-free and asymptotic stability (AS) could be achieved. In the very recent paper [6], the ISS small-gain theorem [23, 32] is applied to guarantee the stability of the overall system composed of interacting ISS subsystems, and a parsimonious event-triggering mechanism is developed to avoid the Zeno behavior.

Based on the recent development of the small-gain methods, this chapter aims to develop a new small-gain approach to event-triggered control of nonlinear systems. By means of small-gain designs, several event-triggered control problems are solved for the first time.

- ISS has been used to describe the influence of the measurement error caused by data-sampling to control performance in event-triggered control systems in [2, 39, 49]. In most of the existing results, a known ISS-Lyapunov function is assumed for event-trigger design. Notice that the construction of ISS-Lyapunov functions for general nonlinear control systems is generally not easy, except for some specific classes of nonlinear systems. We relax this requirement by designing event-triggered controllers without using ISS-Lyapunov functions. With our design, the closed-loop event-triggered control system can be transformed into an interconnected system, the asymptotic stability of which can be guaranteed by using the small-gain theorem.
- Input-to-state stabilization in the presence of measurement errors plays a critical role in the designs for event-triggered control of nonlinear systems. Most of the existing results assume known input-to-state stabilizing controllers a priori. However, it is well known that small measurement error may cause the performance of a nonlinear control system to deteriorate, even if the system with no measurement error is asymptotically stable [10]. Based on our recent results on measurement feedback control [34, 35], this chapter develops a novel design for event-triggered control of nonlinear uncertain systems in the strict-feedback form and output-feedback form. We employ a novel set-valued map design to cover the influence of the measurement error caused by data-sampling, and transform the closed-loop system into a network of ISS subsystems. With the cyclic-small-gain theorem [24, 32], ISS of the closed-loop system with the measurement error as the input is guaranteed, and the influence of the measurement error is explicitly evaluated. More importantly, it is shown that event-triggered control problem is solvable under mild conditions. It should be noted that the design does not require the accurate knowledge of the system dynamics.
- Event-triggered control with partial-state feedback has been studied for linear systems; see, e.g., [8, 30] and also [17] for a recent literature review. However, the corresponding problems with nonlinear systems have not been systematically studied. With the ISS small-gain theorem as a tool, we develop new event-triggered control strategies with partial-state feedback to avoid infinitely fast sampling, and at the same time, achieve asymptotic stabilization. In particular, it is recognized that the decreasing rate of the threshold signal for the event-trigger should be chosen in accordance with the decreasing rate of the closed-loop system. The problem is solved by refining the Lyapunov-based ISS small-gain theorem.

3.2 Preliminaries

In this section, we present some basic notations and review the concept of ISS and its small-gain results that will be used in this chapter.

A function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be positive definite if $\alpha(0) = 0$ and $\alpha(s) > 0$ for $s > 0$. A continuous function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be a class $\mathscr{K}$ function, denoted by $\alpha \in \mathscr{K}$, if it is strictly increasing and $\alpha(0) = 0$; it is said to be a class $\mathscr{K}_\infty$ function, denoted by $\alpha \in \mathscr{K}_\infty$, if it is a class $\mathscr{K}$ function and satisfies $\alpha(s) \to \infty$ as $s \to \infty$. For $\gamma_1, \gamma_2 \in \mathscr{K}$, $\gamma_1 \circ \gamma_2 < \mathrm{Id}$ means $\gamma_1(\gamma_2(s)) < s$ for all $s > 0$. A continuous function $\beta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ is said to be a class $\mathscr{K}\mathscr{L}$ function, denoted by $\beta \in \mathscr{K}\mathscr{L}$, if, for each fixed $t \in \mathbb{R}_+$, function $\beta(\cdot, t)$ is a class $\mathscr{K}$ function and, for each fixed $s \in \mathbb{R}_+$, function $\beta(s, \cdot)$ is decreasing and $\lim_{t\to\infty} \beta(s, t) = 0$.

A function $h : \mathscr{X} \to \mathscr{Y}$ with $\mathscr{X} \subseteq \mathbb{R}^n$ and $\mathscr{Y} \subseteq \mathbb{R}^m$ is said to be Lipschitz continuous, or simply Lipschitz, on $\mathscr{X}$, if there exists a constant $K_h \geq 0$, such that for any $x_1, x_2 \in \mathscr{X}$, $|h(x_1) - h(x_2)| \leq K_h|x_1 - x_2|$. A function $h : \mathscr{X} \to \mathscr{Y}$ with $\mathscr{X} \subseteq \mathbb{R}^n$ being open and connected, and $\mathscr{Y} \subseteq \mathbb{R}^m$ is said to be locally Lipschitz on $\mathscr{X}$, if each $x \in \mathscr{X}$ has a neighborhood $\mathscr{X}_0 \subseteq \mathscr{X}$ such that h is Lipschitz on $\mathscr{X}_0$. A function $h : \mathscr{X} \to \mathscr{Y}$ with $\mathscr{X} \subseteq \mathbb{R}^n$ and $\mathscr{Y} \subseteq \mathbb{R}^m$ is said to be Lipschitz on compact sets, if h is Lipschitz on every compact set $\mathscr{D} \subseteq \mathscr{X}$. Here, it should be noted that the notion of Lipschitz on compact sets is used in [49] for Lyapunov-based event-triggered control design.

3.2.1 *Input-to-State Stability*

For systems with external inputs, the notion of ISS invented by Sontag has been proved to be powerful in evaluating the influence of the external inputs.

Consider system

$$\dot{x} = f(x, u), \tag{3.1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ represents the input, and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a locally Lipschitz function and satisfies $f(0, 0) = 0$. By considering the input u as a function of time, assume that u is measurable and locally essentially bounded.

In [44], the original definition of ISS is given in the "plus" form. For convenience of discussions, we mainly use the equivalent "max"-form definition.

Definition 3.1 System (3.1) is said to be ISS if there exist a $\beta \in \mathscr{K}\mathscr{L}$ and a $\gamma \in \mathscr{K}$ such that for any initial state $x(0) = x_0$ and any measurable and locally essentially bounded u, the solution $x(t)$ satisfies

$$|x(t)| \leq \max\{\beta(|x_0|, t), \gamma(\|u\|_\infty)\} \tag{3.2}$$

for all $t \geq 0$. Here, γ is called the ISS gain of the system.

ISS-Lyapunov functions have been used to formulate the notion of ISS. For system (3.1), the equivalence between ISS and the existence of ISS-Lyapunov functions was originally presented in [47].

Theorem 3.1 *System (3.1) is ISS if and only if it admits a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}_+$, for which*

1. *there exist $\underline{\alpha}, \overline{\alpha} \in \mathscr{K}_\infty$ such that*

$$\underline{\alpha}(|x|) \le V(x) \le \overline{\alpha}(|x|), \quad \forall x, \tag{3.3}$$

2. *there exist a $\gamma \in \mathscr{K}$ and a continuous, positive definite α such that*

$$V(x) \ge \gamma(|u|) \Rightarrow \nabla V(x) f(x, u) \le -\alpha(V(x)), \quad \forall x,\ u. \tag{3.4}$$

A function V satisfying (3.3) and (3.4) is called an ISS-Lyapunov function and γ is called the Lyapunov-based ISS gain.

3.2.2 ISS Small-Gain Theorems

The small-gain theorem developed in [23] has been proved to be very useful in the analysis and design of interconnected nonlinear systems. In this chapter, the systems will be considered as interconnected systems in the event-triggered control designs. Consider system

$$\dot{x}_i = f_i(x, u_i), \quad i = 1, 2 \tag{3.5}$$

where $x = [x_1^T, x_2^T]^T$ with $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$ is the state, $u_1 \in \mathbb{R}^{m_1}$ and $u_2 \in \mathbb{R}^{m_2}$ are external inputs, $f_1 : \mathbb{R}^{n_1+n_2} \times \mathbb{R}^{m_1} \to \mathbb{R}^{n_1}$ and $f_2 : \mathbb{R}^{n_1+n_2} \times \mathbb{R}^{m_2} \to \mathbb{R}^{n_2}$ are locally Lipschitz functions satisfying $f_1(0, 0) = 0$ and $f_2(0, 0) = 0$. For convenience of notations, denote $u = [u_1^T, u_2^T]^T$. By considering u as a function of time, assume that u is measurable and locally essentially bounded.

For $i = 1, 2$, assume that each x_i-subsystem is ISS with x_{3-i} and u_i as inputs. Specifically, for each $i = 1, 2$, there exist $\beta_i \in \mathscr{KL}$ and $\gamma_{i(3-i)}, \gamma_i^u \in \mathscr{K}$ such that for any initial state $x_i(0) = x_{i0}$ and any measurable and locally essentially bounded inputs x_{3-i}, u_i, the solution $x_i(t)$ satisfies

$$|x_i(t)| \le \max\{\beta_i(|x_{i0}|, t), \gamma_{i(3-i)}(\|x_{3-i}\|_\infty), \gamma_i^u(\|u_i\|_\infty)\} \tag{3.6}$$

for all $t \ge 0$.

Reference [23] considers systems in a more general form with the subsystems interconnected by their outputs and assumed to be input-to-output stable (IOS). Here, due to space limitation, we just review the special case in which the subsystems are ISS and are directly interconnected by their states.

Theorem 3.2 *Consider the interconnected system composed of two subsystems in the form of (3.5) satisfying (3.6). The interconnected system is ISS with u as the input if the small-gain condition*

$$\gamma_{12} \circ \gamma_{21} < \mathrm{Id} \tag{3.7}$$

is satisfied.

Remark 3.1 It is obvious that Theorem 3.2 still holds if one of the subsystems, say the x_2-subsystem, satisfies (3.6) with $\beta_2 = 0$. This is the case where the x_2-subsystem is memoryless. Also, due to causality, if (3.6) holds for $t \in [0, T_{\max})$ with $0 < T_{\max} \leq \infty$, then with (3.7) satisfied, the solution $x(t)$ of interconnected system (3.5) satisfies (3.2) for $t \in [0, T_{\max})$.

Remark 3.2 It is interesting to note that the Lyapunov formulation of the ISS small-gain Theorem 3.2 was developed in [22] based on the equivalence of ISS and ISS-Lyapunov function.

Recently, the small-gain result in [23] has been significantly generalized to address problems arising from large-scale systems in [24]. With the new result called cyclic-small-gain theorem, the IOS property of a large-scale system composed of IOS subsystems can be tested by checking the composition of the IOS gains along every simple cycle of the network interconnection structure. The Lyapunov formulation of the ISS cyclic-small-gain theorem has been developed in [32].

Consider system

$$\dot{x}_i = f_i\left(x, u_i\right), \quad i = 1, \ldots, N \tag{3.8}$$

where $x = \left[x_1^T, \ldots, x_N^T\right]^T$ with $x_i \in \mathbb{R}^{n_i}$ is the state, $u_i \in \mathbb{R}^{m_i}$ represents the external input of the x_i-subsystem, and each $f_i : \mathbb{R}^{n+m_i} \to \mathbb{R}^{n_i}$ with $n = \sum_{j=1}^{N} n_j$ is a locally Lipschitz function satisfying $f_i(0,0) = 0$. The external input $u = \left[u_1^T, \ldots, u_N^T\right]^T$ is a measurable and locally essentially bounded function from $\mathbb{R}_+$ to $\mathbb{R}^m$ with $m = \sum_{i=1}^{N} m_i$. Denote $f(x,u) = [f_1^T(x,u_1), \ldots, f_N^T(x,u_N)]^T$.

Assume that for $i = 1, \ldots, N$, each x_i-subsystem admits a continuously differentiable ISS-Lyapunov function $V_i : \mathbb{R}^{n_i} \to \mathbb{R}_+$ satisfying

1. there exist $\underline{\alpha}_i, \overline{\alpha}_i \in \mathscr{K}_\infty$ such that

$$\underline{\alpha}_i(|x_i|) \leq V_i(x_i) \leq \overline{\alpha}_i(|x_i|), \quad \forall x_i; \tag{3.9}$$

2. there exist $\gamma_{ij} \in \mathscr{K} \cup \{0\}$ $(j = 1, \ldots, N, j \neq i)$ and $\gamma_{ui} \in \mathscr{K} \cup \{0\}$ such that

$$\begin{aligned} &V_i(x_i) \geq \max_{j \neq i} \left\{\gamma_{ij}(V_j(x_j)), \gamma_{ui}(|u_i|)\right\} \\ &\Rightarrow \nabla V_i(x_i) f_i(x,u_i) \leq -\alpha_i(V_i(x_i)), \quad \forall x,\ u_i \end{aligned} \tag{3.10}$$

where α_i is a continuous and positive definite function.

The functions γ_{ij}, γ_i^u are known as the ISS gains of the subsystems. The following theorem presents a cyclic-small-gain condition to guarantee the ISS property of the large-scale system (3.8) with state x and input $u = [u_1^T, \ldots, u_N^T]^T$.

Theorem 3.3 [24] *Consider the large-scale system (3.8). Assume each x_i-subsystem admits an ISS-Lyapunov function V_i satisfying (3.9) and (3.10). Then, the large-scale nonlinear system (3.8) is ISS if for $r = 2, \ldots, N$,*

$$\gamma_{i_1 i_2} \circ \gamma_{i_2 i_3} \circ \ldots \circ \gamma_{i_r i_1} < \mathrm{Id} \tag{3.11}$$

where $1 \leq i_k \leq N$ and $i_k \neq i_{k'}$ if $k \neq k'$ for $1 \leq k \leq r$.

By considering the subsystems (3.8) as vertices and the gains as the weights of the directed connections between the subsystems, the interconnection structure of the large-scale nonlinear system can be represented with a system digraph. Condition (3.11) is called *cyclic-small-gain condition* and means that the composition of the ISS gains along every simple cycle in the large-scale nonlinear system is less than the identity function Id.

For the ISS gains γ_{ij}'s $(1 \leq i \leq N, j \neq i)$ satisfying condition (3.11), according to [22, Lemma A.1], we can find $\mathscr{K}_\infty$ functions $\hat{\gamma}_{ij}$'s $(1 \leq i \leq N, j \neq i)$ which are continuously differentiable on $(0, \infty)$ and slightly larger than the corresponding γ_{ij}'s such that condition (3.11) still holds by replacing the γ_{ij}'s with the $\hat{\gamma}_{ij}$'s. Motivated by the ISS-Lyapunov function construction in [22], a locally Lipschitz ISS-Lyapunov function can be constructed for the large-scale system (3.8) as

$$V(x) = \max_{i=1,\ldots,n} \{\sigma_i(V_i(x_i))\} \tag{3.12}$$

where σ_i's are specific compositions of the $\hat{\gamma}_{(\cdot)}$'s [32].

The influence of the external input u can be represented as

$$\theta(u) = \max_{i=1,\ldots,n} \{\sigma_i \circ \gamma_i^u(|u_i|)\}. \tag{3.13}$$

Denote $f(x, u) = [f_1^T(x, u_1), \ldots, f_N^T(x, u_N)]^T$. With the Lyapunov-based ISS cyclic-small-gain theorem presented in [32], we have

$$V(x) \geq \theta(u) \Rightarrow \nabla V(x) f(x, u) \leq -\alpha(V(x)) \quad \text{a.e.} \tag{3.14}$$

with α being a continuous and positive definite function.

In this chapter, the ISS small-gain result given by Theorem 3.2 will be used in Sect. 3.3 to develop a new ISS gain condition to avoid infinitely fast sampling for event-triggered control of nonlinear systems. Theorem 3.2 will also be used for the designs to handle external disturbances in Sect. 3.4. In Sect. 3.5, we will consider nonlinear uncertain systems in the strict-feedback form. Through a recursive design, we will design a controller by transforming the system into a large-scale system composed of ISS subsystems, and then use Theorem 3.3 to guarantee the ISS of the closed-loop system. By means of (3.12), an ISS-Lyapunov function will be constructed to evaluate the influence of the measurement error caused by data-sampling. In Sect. 3.6, the ISS-Lyapunov function will be employed to evaluate the converging rates of closed-loop event-triggered systems and to design event-triggered control laws with partial-state feedback.

3.3 An ISS Gain Condition for Event-Triggered Control

An event-triggered control system is a sampled-data system in which the sampling time instants are determined by events depending on the real-time system state. An event-triggered state-feedback control system is generally in the following form:

$$\dot{x}(t) = f(x(t), u(t)), \tag{3.15}$$

$$u(t) = v(x(t_k)), \quad t \in [t_k, t_{k+1}),\ k \in \mathbb{S}, \tag{3.16}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control input, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a locally Lipschitz function representing system dynamics, $v : \mathbb{R}^n \to \mathbb{R}^m$ is a locally Lipschitz function representing the control law. It is assumed that $f(0, v(0)) = 0$. The time sequence $\{t_k\}_{k\in\mathbb{S}}$ is determined online based on the measurement of the real-time system state. If there is an infinite number of sampling time instants, then $\mathbb{S} = \mathbb{Z}_+$; otherwise, $\mathbb{S}$ is in the form of $\{0, \dots, k^*\}$ with $k^* \in \mathbb{Z}_+$ being the last sampling time instant. For convenience of notations, denote $t_{k^*+1} = \infty$.

Define

$$w(t) = x(t_k) - x(t), \quad t \in [t_k, t_{k+1}),\ k \in \mathbb{S} \tag{3.17}$$

as the measurement error caused by data-sampling, and rewrite

$$u(t) = v(x(t) + w(t)). \tag{3.18}$$

Then, by substituting (3.18) into (3.15), we have

$$\dot{x}(t) = f(x(t), v(x(t) + w(t))) := \bar{f}(x(t), x(t) + w(t)). \tag{3.19}$$

If $w(t)$ is not adjustable, then the event-triggered control problem is reduced to the measurement feedback control problem. The basic idea of event-triggered control is to adjust $w(t)$ online with an appropriately designed data-sampling strategy, to realize asymptotic convergence of $x(t)$, if possible. In this chapter, this problem is treated as a robust control problem. The block diagram of the system is shown in Fig. 3.1.

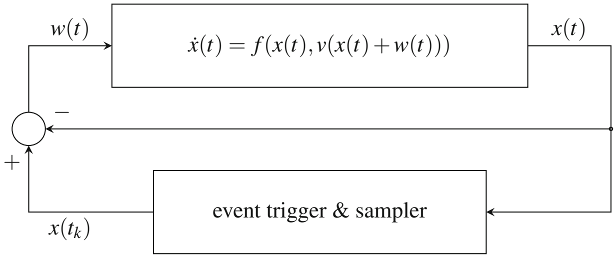


Fig. 3.1 Event-triggered control problem as a robust control problem

Considering the equivalence between ISS and robust stability [47] (see Remark 3.1), we develop an ISS-based approach for event-triggered control. In this section, we consider the systems which are input-to-state stabilizable when the measurement errors caused by data-sampling is considered as the input.

Assumption 3.1 System (3.19) is ISS with w as the input, that is, there exist $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ such that for any initial state $x(0)$ and any piecewise continuous, locally essentially bounded w, it holds that

$$|x(t)| \leq \max\left\{\beta(|x(0)|, t), \gamma(\|w\|_\infty)\right\} \tag{3.20}$$

for all $t \geq 0$.

According to Theorem 3.2, under Assumption 3.1, if the event-trigger is designed such that $|w(t)| \leq \rho(|x(t)|)$ for all $t \geq 0$ with $\rho \in \mathcal{K}$ satisfying

$$\rho \circ \gamma < \mathrm{Id}, \tag{3.21}$$

then $x(t)$ asymptotically converges to the origin. Based on this idea, the event-trigger considered in this chapter can be defined as follows: if $x(t_k) \neq 0$, then

$$t_{k+1} = \inf\left\{t > t_k \,:\, \rho(|x(t)|) - |x(t) - x(t_k)| = 0\right\}. \tag{3.22}$$

The data-sampling event is not triggered if for some specific $k^* \in \mathbb{Z}_+$, $x(t_{k^*}) = 0$ or $\left\{t > t_{k^*} \,:\, H(x(t), x(t_{k^*})) = 0\right\} = \emptyset$. In this case, $\mathbb{S}$ is in the form of $\{0, \ldots, k^*\}$ with k^* being the last sampling time instant, and we set $t_{k^*+1} = T_{\max}$ with $0 < T_{\max} \leq \infty$ such that $x(t)$ is right maximally defined on $[0, T_{\max})$. Note that, under the assumption of $f(0, v(0)) = 0$, if $x(t_k) = 0$, then $u(t) = v(x(t_k)) = 0$ keeps the system state at the origin for all $t \in [t_k, \infty)$.

With the event-trigger proposed above, given t_k and $x(t_k) \neq 0$, t_{k+1} is the first time instant after t_k such that

$$\rho(|x(t_{k+1})|) - |x(t_{k+1}) - x(t_k)| = 0. \tag{3.23}$$

Since $\rho(|x(t_k)|) - |x(t_k) - x(t_k)| = \rho(|x(t_k)|) > 0$ for any $x(t_k) \neq 0$ and $x(t)$ is continuous on the time-line,

$$\rho(|x(t)|) - |x(t) - x(t_k)| > 0 \tag{3.24}$$

holds for $t \in [t_k, t_{k+1})$, $k \in \mathbb{S}$. Recall the definition of $w(t)$ in (3.17). Property (3.24) implies that

$$|w(t)| \leq \rho(|x(t)|) \tag{3.25}$$

holds for $t \in \bigcup_{k \in \mathbb{S}} [t_k, t_{k+1})$.

For physical realization of (3.25) with event-triggered sampling, infinitely fast sampling should be avoided, that is, $\inf_{k \in \mathbb{S}} \{t_{k+1} - t_k\} > 0$. A special case is the Zeno behavior, with which, $\mathbb{S} = \mathbb{Z}_+$ and $\lim_{k \to \infty} t_k < \infty$.

The objective of this section is to propose an event-triggered sampling strategy to avoid infinitely fast sampling and at the same time to realize asymptotic stabilization.

We first present a technical lemma, which will be used for the proof of the main result of this section.

Lemma 3.1 *For any $a, b \in \mathbb{R}$, if there exist a $\theta \in \mathscr{K}$ and a constant $c \geq 0$ such that*

$$|a - b| \leq \max\{\theta \circ (\mathrm{Id} + \theta)^{-1}(|a|), c\}, \tag{3.26}$$

then $|a - b| \leq \max\{\theta(|b|), c\}$.

Proof We first consider the case of $\theta \circ (\mathrm{Id} + \theta)^{-1}(|a|) \geq c$, which together with (3.26) implies

$$|a - b| \leq \theta \circ (\mathrm{Id} + \theta)^{-1}(|a|). \tag{3.27}$$

In this case, $|a| - |b| \leq \theta \circ (\mathrm{Id} + \theta)^{-1}(|a|)$, and thus, $(\mathrm{Id} - \theta \circ (\mathrm{Id} + \theta)^{-1})(|a|) \leq |b|$. Note that $\mathrm{Id} - \theta \circ (\mathrm{Id} + \theta)^{-1} = (\mathrm{Id} + \theta) \circ (\mathrm{Id} + \theta)^{-1} - \theta \circ (\mathrm{Id} + \theta)^{-1} = (\mathrm{Id} + \theta)^{-1}$. Then, we have $|a| \leq (\mathrm{Id} + \theta)(|b|)$. By using (3.27) again, it can be achieved that $|a - b| \leq \theta(|b|)$. Next, we consider the case of $\theta \circ (\mathrm{Id} + \theta)^{-1}(|a|) < c$. Clearly, from (3.26), it follows that $|a - b| \leq c$. Therefore, Lemma 3.1 is proved.

Theorem 3.4 presents a condition on the ISS gain γ to find a ρ for the event-trigger (3.22) to avoid infinitely fast sampling and asymptotically stabilize the closed-loop system at the origin.

Theorem 3.4 *Consider the event-triggered control system (3.19) with locally Lipschitz $\bar{f}$ satisfying $\bar{f}(0, 0) = 0$ and w defined in (3.17). If Assumption 3.1 is satisfied with a γ being Lipschitz on compact sets, then one can find a $\rho \in \mathscr{K}_\infty$ such that*

- *ρ satisfies (3.21), and*
- *ρ^{-1} is Lipschitz on compact sets.*

Moreover, with the sampling time instants triggered by (3.22), it can always be guaranteed that

$$\inf_{k\in\mathbb{S}}\{t_{k+1}-t_k\}>0 \tag{3.28}$$

and, for any specific initial state $x(0)$, the system state $x(t)$ satisfies

$$|x(t)|\le\breve{\beta}(|x(0)|,t) \tag{3.29}$$

with $\breve{\beta}\in\mathscr{KL}$, for all $t\ge0$.

Proof Using (3.20) and (3.25), by Theorem 3.2, there exists a $\breve{\beta}\in\mathscr{KL}$ such that

$$|x(t)|\le\breve{\beta}(|x(0)|,t) \tag{3.30}$$

for all $t\in\bigcup_{k\in\mathbb{S}}[t_k,t_{k+1})$. Now, we prove (3.28) and

$$\bigcup_{k\in\mathbb{S}}[t_k,t_{k+1})=[0,\infty). \tag{3.31}$$

With a $\gamma\in\mathscr{K}$ being Lipschitz on compact sets, one can always find a $\bar{\gamma}\in\mathscr{K}_\infty$ being Lipschitz on compact sets such that $\bar{\gamma}>\gamma$. By choosing $\rho=\bar{\gamma}^{-1}$, we have $\rho\circ\gamma=\bar{\gamma}\circ\gamma<\bar{\gamma}\circ\bar{\gamma}^{-1}<\mathrm{Id}$, and $\rho^{-1}=\bar{\gamma}$ is Lipschitz on compact sets.

Along each trajectory of the closed-loop system, for each $k\in\mathbb{S}$ with state $x(t_k)$ at time instant t_k, define

$$\Theta_1(x(t_k))=\left\{x\in\mathbb{R}^n\,:\,|x-x(t_k)|\le\rho\circ(\mathrm{Id}+\rho)^{-1}(|x(t_k)|)\right\}, \tag{3.32}$$

$$\Theta_2(x(t_k))=\left\{x\in\mathbb{R}^n\,:\,|x-x(t_k)|\le\rho(|x|)\right\}. \tag{3.33}$$

Then, the lower bound of $t_{k+1}-t_k$ can be considered as the minimum time needed for $x(t)$ starting at $x(t_k)$ to go outside $\Theta_2(x(t_k))$. By directly using Lemma 3.1, it can be proved that $\Theta_1(x(t_k))\subseteq\Theta_2(x(t_k))$. An illustration with $x=[x_1,x_2]^T\in\mathbb{R}^2$ is given in Fig. 3.2. Now, we estimate the minimum time needed for $x(t)$ starting at $x(t_k)$ to go outside $\Theta_1(x(t_k))$.

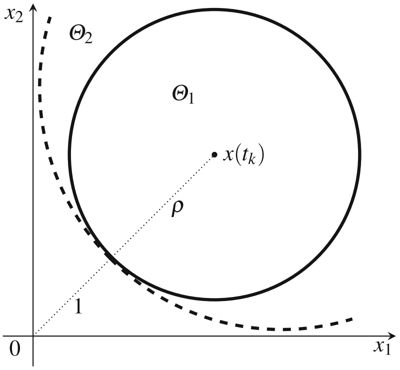
Given a $\rho\in\mathscr{K}_\infty$ such that ρ^{-1} is Lipschitz on compact sets, it can be proved that $\rho^{-1}+\mathrm{Id}$ is Lipschitz on compact sets and there exists a continuous, positive function $\breve{\rho}:\mathbb{R}_+\to\mathbb{R}_+$ such that

$$(\rho^{-1}+\mathrm{Id})(s)\le\breve{\rho}(s)s:=\hat{\rho}(s) \tag{3.34}$$

for $s\in\mathbb{R}_+$. Note that $\breve{\rho}(s)s=\hat{\rho}(s)$ implies $s=\left(\breve{\rho}\circ\hat{\rho}^{-1}(s)\right)\hat{\rho}^{-1}(s)$. We have

$$\hat{\rho}^{-1}(s)=\frac{s}{\breve{\rho}\circ\hat{\rho}^{-1}(s)}:=\bar{\rho}(s)s. \tag{3.35}$$

Fig. 3.2 An illustration of $\Theta_1(x(t_k)) \subseteq \Theta_2(x(t_k))$



Here, $\bar{\rho} : \mathbb{R}_+ \to \mathbb{R}_+$ is a continuous and positive function as $\breve{\rho}$ is continuous and positive. Then, by using (3.34) and (3.35), we have $\rho \circ (\mathrm{Id} + \rho)^{-1}(s) = (\rho^{-1} + \mathrm{Id})^{-1}(s) \geq \hat{\rho}^{-1}(s) = \bar{\rho}(s)s$. This means, if

$$|x - x(t_k)| \leq \bar{\rho}(|x(t_k)|)|x(t_k)| \tag{3.36}$$

then $x \in \Theta_1(x(t_k))$.

Note that for the locally Lipschitz $\bar{f}$ as defined in (3.19), there exists a continuous and positive function $L_{\bar{f}}$ such that

$$\begin{aligned} |\bar{f}(x, x(t_k))| &= |\bar{f}(x - x(t_k) + x(t_k), x(t_k))| \\ &\leq L_{\bar{f}} \left(|[x^T - x^T(t_k), x^T(t_k)]^T| \right) \\ &\quad \times |[x^T - x^T(t_k), x^T(t_k)]^T|. \end{aligned} \tag{3.37}$$

If moreover $|x - x(t_k)| \leq \bar{\rho}(|x(t_k)|)|x(t_k)|$, then there exists a continuous and positive function $\bar{L}$ such that

$$|\bar{f}(x, v(x(t_k)))| \leq \bar{L}(|x(t_k)|)|x(t_k)|. \tag{3.38}$$

Thus, the minimum time $T_k^{\min}$ needed for the state of the closed-loop system starting at $x(t_k)$ to go outside the region $\Theta_1(x(t_k))$ can be estimated by

$$T_k^{\min} \geq \frac{\bar{\rho}(|x(t_k)|)|x(t_k)|}{\bar{L}(|x(t_k)|)|x(t_k)|} = \frac{\bar{\rho}(|x(t_k)|)}{\bar{L}(|x(t_k)|)}, \tag{3.39}$$

which is well defined and strictly larger than zero for any $x(t_k)$. Since $\Theta_1(x(t_k)) \subseteq \Theta_2(x(t_k))$ and $x(t)$ is continuous on the time-line, the minimum interval needed for the state starting at $x(t_k)$ to go outside $\Theta_2(x(t_k))$ is not less than $T_k^{\min}$. Thus, with (3.39) and (3.30), it is achieved that

$$T_k^{\min} \geq \min\left\{ \frac{\bar{\rho}(|x|)}{\bar{L}(|x|)} : |x| \leq \breve{\beta}(|x(0)|, 0) \right\} \tag{3.40}$$

for all $k \in \mathbb{S}$. Note that the right-hand side of (3.40) only depends on $x(0)$ and is strictly positive. Property (3.28) is proved.

Property (3.31) is proved by considering the following two cases:

- $\mathbb{S} = \mathbb{Z}_+$. In this case, (3.31) can be directly proved.
- $\mathbb{S} = \{0, \ldots, k^*\}$ with $k^* \in \mathbb{Z}_+$. In this case, since $t_{k^*+1} = T_{\max}$, we have $\bigcup_{k \in \mathbb{S}} [t_k, t_{k+1}) = [0, T_{\max})$, and thus (3.30) holds for all $t \in [0, T_{\max})$. This implies that $x(t)$ is defined for all $t \in [0, \infty)$, i.e., $T_{\max} = \infty$.

This ends the proof.

Remark 3.3 Theoretically, a system is ISS if and only if it has an ISS-Lyapunov function. To use the recent results in [29, 49, 55], one may need to assume the existence of known ISS-Lyapunov functions for the system (3.19) with w as the input. However, even if a nonlinear system has been designed to be ISS, the construction of an ISS-Lyapunov function may not be straightforward. On the other hand, given an ISS-Lyapunov function, one can easily determine the ISS characteristics of a system. By using the relationship between ISS and robust stability, the study in this section shows that a known ISS-Lyapunov function may not be necessary for event-triggered control.

Remark 3.4 In [49], the existence of an ISS-Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}_+$ is assumed for the system composed of (3.15) and (3.18) with w as the input, i.e., $\nabla V(x) f(x, v(x + w)) \leq -\alpha(|x|) + \gamma(|w|)$ with $\alpha \in \mathscr{K}_\infty$ and $\gamma \in \mathscr{K}$. Under this assumption, the event-trigger can be designed to satisfy (3.23) with $\rho \in \mathscr{K}_\infty$ such that $\alpha^{-1} \circ \gamma \circ \rho < \mathrm{Id}$ and ρ^{-1} is Lipschitz on compact sets. The design in this chapter also requires that ρ^{-1} is Lipschitz on compact sets, and is in accordance with the result in [49].

The proof of Theorem 3.4 naturally leads to a self-triggered sampling strategy, which does not continuously monitor the trajectory of $x(t)$.

Theorem 3.5 *Consider system (3.19) with locally Lipschitz $\bar{f}$ satisfying $\bar{f}(0, 0) = 0$ and w defined in (3.17). If Assumption 3.1 is satisfied with a γ being Lipschitz on compact sets, then there exist continuous and positive functions $\bar{\rho}, \bar{L} : \mathbb{R}_+ \to \mathbb{R}_+ \backslash \{0\}$ such that with the self-triggered sampling strategy*

$$t_{k+1} = \frac{\bar{\rho}(|x(t_k)|)}{\bar{L}(|x(t_k)|)} + t_k, \quad k \in \mathbb{Z}_+, \tag{3.41}$$

the system state satisfies (3.29) for all $t \geq 0$.

Proof Following the proof of Theorem 3.4, property (3.39) still holds for the self-triggered control system, and the self-triggered sampling strategy guarantees that $x(t) \in \Theta_2(x(t_k))$ for $t \in [t_k, t_{k+1})$, $k \in \mathbb{Z}_+$, which means that (3.25) holds for all $t \geq 0$. Property (3.29) can then be proved by directly using Theorem 3.2.

Example 3.1 Assumption 3.1 can be readily satisfied by designing a linear controller for the linear time-invariant system $\dot{x} = Ax + Bu$ with $x \in \mathbb{R}^n$ as the state and $u \in \mathbb{R}^m$ as the control input, if the system is controllable. One can find a K such that $A - BK$ is Hurwitz and design $u = -K(x + w)$ with w being the measurement error caused by data-sampling. Then, $\dot{x} = Ax - BK(x + w) = (A - BK)x - BKw$. With initial state $x(0)$, the solution of the system is $x(t) = e^{(A-BK)t}x(0) - \int_0^t e^{(A-BK)(t-\tau)}BKw(\tau)d\tau$ for $t \geq 0$. It can be verified that $x(t)$ satisfies property (3.20) with $\beta(s,t) = (1 + 1/\delta)|e^{(A-BK)t}|s$ and $\gamma(s) = (1 + \delta)\left(\int_0^\infty |e^{(A-BK)\tau}BK|d\tau\right)s$, where δ can be selected as any positive constant. Then, $\beta \in \mathscr{K}\mathscr{L}$ and $\gamma \in \mathscr{K}$. Moreover, γ is Lipschitz on compact sets.

3.4 Event-Triggered Control and Self-Triggered Control in the Presence of External Disturbances

Theorems 3.4 and 3.5 do not address the presence of external disturbances. In this section, we consider the systems with external disturbances taking the form:

$$\dot{x}(t) = f(x(t), u(t), d(t)) \tag{3.42}$$

where $d(t) \in \mathbb{R}^{n_d}$ represents the external disturbances, and the other variables are defined as in (3.15). It is assumed that d is piecewise continuous and bounded.

With w defined in (3.17) as the measurement error, the control law (3.16) can be rewritten as (3.18). By substituting (3.18) into (3.15), we have

$$\begin{aligned} \dot{x}(t) &= f(x(t), v(x(t) + w(t)), d(t)) \\ &:= \bar{f}(x(t), x(t) + w(t), d(t)). \end{aligned} \tag{3.43}$$

Corresponding to Assumption 3.1 for the disturbance-free case, we make the following assumption for system (3.43).

Assumption 3.2 System (3.43) is ISS with w and d as the inputs, that is, there exist $\beta \in \mathscr{K}\mathscr{L}$ and $\gamma, \gamma^d \in \mathscr{K}$ such that for any initial state $x(0)$ and any piecewise continuous, bounded w and d, it holds that

$$|x(t)| \leq \max\left\{\beta(|x(0)|, t), \gamma(\|w\|_\infty), \gamma^d(\|d\|_\infty)\right\} \tag{3.44}$$

for all $t \geq 0$.

Under Assumption 3.2, if the event-trigger is still capable of guaranteeing (3.24) with $\rho \in \mathscr{K}$ such that $\rho\circ\gamma < \mathrm{Id}$. Then, by directly using Theorem 3.2, we can prove that

$$|x(t)| \leq \max\left\{\breve{\beta}(|x(0)|, t), \breve{\gamma}^d(\|d\|_\infty)\right\} \tag{3.45}$$

with $\breve{\beta} \in \mathcal{KL}$ and $\breve{\gamma}^d \in \mathcal{K}$. As x converges to the origin, the upper bound of $|w(t)| = |x(t_k) - x(t)|$ converges to zero according to (3.24). However, due to the presence of the external disturbance d, the function of system dynamics $f(x(t), v(x(t) + w(t)), d(t))$ may not converge to zero as x converges to the origin. This means that the inter-sample period $t_{k+1} - t_k$ could be arbitrarily small.

3.4.1 Event-Triggered Sampling with ε Modification

Inspired by the recent result [8], we modify the event-trigger (3.22) as

$$t_{k+1} = \inf \left\{ t > t_k : \max\{\rho(|x(t)|), \varepsilon\} - |x(t) - x(t_k)| = 0 \right\} \tag{3.46}$$

with $\rho \in \mathcal{K}$ satisfying $\rho \circ \gamma < \mathrm{Id}$ and constant $\varepsilon > 0$. The modified event-trigger guarantees

$$|x(t) - x(t_k)| < \max\{\rho(|x(t)|), \varepsilon\} \tag{3.47}$$

for $t \in [t_k, t_{k+1})$, $k \in \mathbb{S}$. With Theorem 3.2, there exist $\breve{\beta} \in \mathcal{KL}$ and $\breve{\gamma}, \breve{\gamma}^d \in \mathcal{K}$ such that

$$|x(t)| \leq \max \left\{ \breve{\beta}(|x(0)|, t), \breve{\gamma}(\varepsilon), \breve{\gamma}^d(\|d\|_\infty) \right\} \tag{3.48}$$

for all $t \geq 0$. It should be noted that, with $\varepsilon > 0$, the function ρ^{-1} is no longer required to be Lipschitz on compact sets. This result is summarized by Theorem 3.6 without proof.

Theorem 3.6 *Consider the event-triggered control system (3.43) with locally Lipschitz $\bar{f}$ and w defined in (3.17). If Assumption 3.2 is satisfied, with the sampling time instants triggered by (3.46), for any specific initial state $x(0)$, the system state $x(t)$ satisfies (3.48) for all $t \geq 0$, with $\breve{\beta} \in \mathcal{KL}$ and $\breve{\gamma}, \breve{\gamma}^d \in \mathcal{K}$, and the inter-sample periods are lower bounded by a positive constant.*

For such event-triggered control system, even if $d \equiv 0$, only practical convergence can be guaranteed, that is, $x(t)$ can only be guaranteed to converge to within a neighborhood of the origin $\{x \in \mathbb{R}^n : |x| \leq \breve{\gamma}(\varepsilon)\}$. It should be mentioned that ε can be made arbitrarily small. In the next section, we present a self-triggered sampling mechanism to overcome this obstacle, under the assumption of an a priori known upper bound of $\|d\|_\infty$.

3.4.2 Self-Triggered Sampling

In this section, we show that if an upper bound of $\|d\|_\infty$ is known a priori, then we can design a self-triggered sampling mechanism such that $x(t)$ is practically steered to within a neighborhood of the origin with size depending solely on $\|d\|_\infty$. Moreover, if $d(t)$ converges to zero, then $x(t)$ asymptotically converges to the origin.

Assumption 3.3 There is a known constant $B^d \geq 0$ such that

$$\|d\|_\infty \leq B^d. \tag{3.49}$$

Lemma 3.2 presents a property of locally Lipschitz functions and will be used in the following design procedure.

Lemma 3.2 *For any locally Lipschitz function $h : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_m} \to \mathbb{R}^p$ satisfying $h(0, \ldots, 0) = 0$ and any $\varphi_1, \ldots, \varphi_m \in \mathcal{K}_\infty$ with $\varphi_1^{-1}, \ldots, \varphi_m^{-1}$ being Lipschitz on compact sets, there exists a continuous, positive, and nondecreasing function L_h : $\mathbb{R}_+ \to \mathbb{R}_+$ such that $|h(z_1, \ldots, z_m)| \leq L_h\left(\max_{i=1,\ldots,m}\{|z_i|\}\right) \max_{i=1,\ldots,m}\{\varphi_i(|z_i|)\}$ for all z, where $z = [z_1^T, \ldots, z_m^T]^T$.*

Proof For a locally Lipschitz h satisfying $h(0, \ldots, 0) = 0$, one can always find a continuous, positive, and nondecreasing function $L_{h0} : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$|h(z_1, \ldots, z_m)| \leq L_{h0}\left(\max_{i=1,\ldots,m}\{|z_i|\}\right) \max_{i=1,\ldots,m}\{|z_i|\} \tag{3.50}$$

for all z.

Define $\breve{\varphi}(s) = \max_{i=1,\ldots,m}\{\varphi_i^{-1}(s)\}$ for $s \in \mathbb{R}_+$. Then, $\breve{\varphi} \in \mathcal{K}_\infty$. Since $\varphi_1^{-1}, \ldots,$ φ_m^{-1} are Lipschitz on compact sets, $\breve{\varphi}$ is Lipschitz on compact sets.

From the definition, one has

$$\begin{aligned}
\breve{\varphi}\left(\max_{i=1,\ldots,m}\{\varphi_i(|z_i|)\}\right) &= \max_{i=1,\ldots,m}\{\breve{\varphi}\circ\varphi_i(|z_i|)\} \\
&\geq \max_{i=1,\ldots,m}\{\varphi_i^{-1}\circ\varphi_i(|z_i|)\} \\
&= \max_{i=1,\ldots,m}\{|z_i|\}.
\end{aligned} \tag{3.51}$$

With $\breve{\varphi}$ being Lipschitz on compact sets, there exists a continuous, positive, and nondecreasing function $L_{\breve{\varphi}} : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\breve{\varphi}\left(\max_{i=1,\ldots,m}\{\varphi_i(|z_i|)\}\right) \leq L_{\breve{\varphi}}\left(\max_{i=1,\ldots,m}\{\varphi_i(|z_i|)\}\right) \max_{i=1,\ldots,m}\{\varphi_i(|z_i|)\}. \tag{3.52}$$

Lemma 3.2 is proved by substituting (3.51) and (3.52) into (3.50), and defining a continuous, positive, and nondecreasing L_h such that for all z,

$$L_h\left(\max_{i=1,\ldots,m}\{|z_i|\}\right) \geq L_{h0}\left(\max_{i=1,\ldots,m}\{|z_i|\}\right) L_{\breve{\varphi}}\left(\max_{i=1,\ldots,m}\{\varphi_i(|z_i|)\}\right). \tag{3.53}$$

Assume that $\bar{f}$ is locally Lipschitz and $\bar{f}(0,0,0) = 0$. Then, with Lemma 3.2, for any specific $\chi, \chi^d \in \mathscr{K}_\infty$ with $\chi^{-1}, \left(\chi^d\right)^{-1}$ being Lipschitz on compact sets, one can find a continuous, positive, and nondecreasing $L_{\bar{f}}$ such that

$$|\bar{f}(x+w,x,d)| \leq L_{\bar{f}}\left(\max\left\{|x|,|w|,|d|\right\}\right)\max\left\{\chi(|x|),|w|,\chi^d(|d|)\right\} \tag{3.54}$$

for all x, w, d.

By choosing $\chi, \chi^d \in \mathscr{K}_\infty$ with $\chi^{-1}, \left(\chi^d\right)^{-1}$ being locally Lipchitz, the self-triggered sampling mechanism is designed as

$$t_{k+1} = \frac{1}{L_{\bar{f}}\left(\max\left\{\bar{\chi}(|x(t_k)|), \bar{\chi}^d(B^d)\right\}\right)} + t_k, \quad k \in \mathbb{Z}_+ \tag{3.55}$$

where $\bar{\chi}(s) = \max\{\chi(s), s\}$ and $\bar{\chi}^d(s) = \max\{\chi^d(s), s\}$ for $s \in \mathbb{R}_+$.

Theorem 3.7 provides the main result of this section.

Theorem 3.7 *Consider the event-triggered control system (3.43) with locally Lipschitz $\bar{f}$ satisfying $\bar{f}(0,0,0) = 0$ and w defined in (3.17). If Assumption 3.2 holds with a γ being Lipschitz on compact sets, then one can find a $\rho \in \mathscr{K}_\infty$ such that*

- *ρ satisfies*

$$\rho\circ\gamma < \mathrm{Id}, \tag{3.56}$$

- *ρ^{-1} is Lipschitz on compact sets.*

Moreover, under Assumption 3.3, by choosing $\chi = \rho\circ(\mathrm{Id}+\rho)^{-1}$ and $\chi^d \in \mathscr{K}_\infty$ with $\left(\chi^d\right)^{-1}$ being Lipschitz on compact sets for the self-triggered sampling mechanism (3.55), for any specific initial state $x(0)$, the system state $x(t)$ satisfies

$$|x(t)| \leq \max\{\breve{\beta}(|x(0)|,t), \breve{\gamma}\circ\chi^d(\|d\|_\infty), \breve{\gamma}^d(\|d\|_\infty)\} \tag{3.57}$$

for all $t \geq 0$, with $\breve{\beta} \in \mathscr{KL}$ and $\breve{\gamma}, \breve{\gamma}^d \in \mathscr{K}$, and the inter-sample periods are lower bounded by a positive constant.

Proof Note that $\chi = \rho\circ(\mathrm{Id}+\rho)^{-1}$ implies $\chi^{-1} = \mathrm{Id} + \rho^{-1}$. If ρ^{-1} is Lipschitz on compact sets, then χ^{-1} is Lipschitz on compact sets. Also note that $\left(\chi^d\right)^{-1}$ is chosen to be Lipschitz on compact sets.

We first prove that the self-triggered sampling mechanism guarantees that

$$|x(t) - x(t_k)| \leq \max\{\chi(|x(t_k)|), \chi^d(\|d\|_\infty)\} \tag{3.58}$$

for $t \in [t_k, t_{k+1})$.

By taking the integration of both the sides of (3.43), one has

$$x(t) - x(t_k) = \int_{t_k}^{t} \bar{f}(x(t_k) + w(\tau), x(t_k), d(\tau))d\tau, \tag{3.59}$$

and thus,

$$|x(t) - x(t_k)| \le \int_{t_k}^{t} |\bar{f}(x(t_k) + w(\tau), x(t_k), d(\tau))|d\tau. \tag{3.60}$$

Denote $\Omega(x(t_k), \|d\|_\infty)$ as the region of x such that $|x - x(t_k)| \le \max\{\chi(|x(t_k)|), \chi^d(\|d\|_\infty)\}$. Then, the minimum time needed for $x(t)$ to go outside the region $\Omega(x(t_k), \|d\|_\infty)$ can be estimated by

$$\begin{aligned}
&\frac{\max\{\chi(|x(t_k)|), \chi^d(\|d\|_\infty)\}}{C(x(t_k), \|d\|_\infty)} \\
\ge & \frac{\max\{\chi(|x(t_k)|), \chi^d(\|d\|_\infty)\}}{L_{\bar{f}}\left(\max\{\bar{\chi}(|x(t_k)|), \bar{\chi}^d(\|d\|_\infty)\}\right) \max\{\chi(|x(t_k)|), \chi^d(\|d\|_\infty)\}} \\
= & \frac{1}{L_{\bar{f}}\left(\max\{\bar{\chi}(|x(t_k)|), \bar{\chi}^d(\|d\|_\infty)\}\right)} \\
\ge & \frac{1}{L_{\bar{f}}\left(\max\{\bar{\chi}(|x(t_k)|), \bar{\chi}^d(B^d)\}\right)}
\end{aligned} \tag{3.61}$$

where $\bar{\chi}(s) = \max\{\chi(s), s\}$ and $\bar{\chi}^d(s) = \max\{\chi^d(s), s\}$ for $s \in \mathbb{R}_+$, and

$$\begin{aligned}
C(x(t_k), \|d\|_\infty) = \max\Big\{ & |\bar{f}(x(t_k) + w, x(t_k), d)| : \\
& |w| \le \max\{\chi(|x(t_k)|), \chi^d(\|d\|_\infty)\}, \\
& |d| \le \|d\|_\infty \Big\}.
\end{aligned}$$

Thus, the proposed self-triggered sampling mechanism (3.55) guarantees (3.58). With Lemma 3.1, (3.58) implies

$$|w(t)| = |x(t) - x(t_k)| \le \max\{\rho(|x(t)|), \chi^d(\|d\|_\infty)\} \tag{3.62}$$

for $t \in [t_k, t_{k+1})$, $k \in \mathbb{Z}_+$. With $\rho \circ \gamma < \mathrm{Id}$, using Theorem 3.2, one can prove property (3.57). This ends the proof of Theorem 3.7.

3.5 Event-Triggered Control of Nonlinear Uncertain Systems

To realize event-triggered control of nonlinear systems by using the results in Sects. 3.3 and 3.4, control laws should first be designed to guarantee ISS with respect to the measurement errors caused by data-sampling. Moreover, the ISS gain corresponding to the measurement error should be Lipschitz on compact sets to avoid infinitely fast sampling. This is usually not trivial for nonlinear systems. This section proposes a new nonlinear control design method for event-triggered control of nonlinear uncertain systems transformable into the strict-feedback form [27].

Consider nonlinear system:

$$\dot{x}_i = x_{i+1} + \Delta_i(\bar{x}_i, d), \quad i = 1, \ldots, n-1 \tag{3.63}$$

$$\dot{x}_n = u + \Delta_n(\bar{x}_n, d) \tag{3.64}$$

where $[x_1, \ldots, x_n]^T := x$ is the state, $u \in \mathbb{R}$ is the control input, Δ_i's for $i = 1, \ldots, n$ are uncertain locally Lipschitz functions, $d \in \mathbb{R}^{n_d}$ is the external disturbance, and $\bar{x}_i := [x_1, \ldots, x_i]^T$. It is assumed that d is piecewise continuous and bounded on the time line.

The following assumption is made on the functions Δ_i in system (3.63) and (3.64).

Assumption 3.4 For each $i = 1, \ldots, n$, there exists a $\psi_{\Delta_i} \in \mathscr{K}_\infty$ being Lipschitz on compact sets such that for all $\bar{x}_i$,

$$|\Delta_i(\bar{x}_i, d)| \leq \psi_{\Delta_i}(|[\bar{x}_i^T, d^T]^T|). \tag{3.65}$$

Remark 3.5 For a locally Lipschitz function Δ_i, if $\Delta_i(0, 0) = 0$, then there always exists a $\psi_{\Delta_i} \in \mathscr{K}_\infty$ being Lipschitz on compact sets such that (3.65) holds. Specifically, ψ_{Δ_i} can be chosen such that $\psi_{\Delta_i}(s) = \max_{|[\bar{x}_i^T, d^T]^T| \leq s} |\Delta_i(\bar{x}_i)| + \varepsilon s$ for $s \in \mathbb{R}_+$, where ε can be an arbitrarily small positive constant.

The basic idea of the control design is to transform the closed-loop system into a large-scale system composed of n ISS subsystems, and use the cyclic-small-gain theorem [24, 32] to guarantee the ISS of the closed-loop system. In this procedure, the measurement error should be carefully handled such that the corresponding ISS gain is Lipschitz on compact sets. In Sect. 3.5.1, we present the basic form of the proposed control law to transform the closed-loop system into a network of ISS subsystems. Then, we fine tune the ISS gains in Sect. 3.5.2 to guarantee the ISS of the closed-loop system, and moreover, to satisfy the ISS gain condition to avoid infinitely fast sampling in event-triggered control.

For convenience of discussions, denote $w = [w_1, \ldots, w_n]^T$. In the design, we assume that for each $i = 1, \ldots, n$, w_i is piecewise continuous and bounded. Denote $w_i^\infty = \|w_i\|_\infty$ for $i = 1, \ldots, n$, $\bar{w}_i^\infty = [w_1^\infty, \ldots, w_i^\infty]^T$ and $w^\infty = \bar{w}_n^\infty$. Also denote $d^\infty = \|d\|_\infty$.

3.5.1 Control Design

For nonlinear systems in the form of (3.63) and (3.64), if there is no measurement error, one may design a control law in the form of

$$\breve{p}_1^* = \breve{\kappa}_1(x_1) \tag{3.66}$$

$$\breve{p}_i^* = \breve{\kappa}_i(x_i - \breve{p}_{i-1}^*), \quad i = 2, \dots, n-1 \tag{3.67}$$

$$u = \breve{\kappa}_n(x_n - \breve{p}_{n-1}^*) \tag{3.68}$$

with appropriately chosen nonlinear functions $\breve{\kappa}_k$ for $k = 1, \dots, n$. Then, the achievement of the control objective can be guaranteed by checking the stability property of the closed-loop system with new state variables defined as

$$e_1 = x_1, \tag{3.69}$$

$$e_i = x_i - \breve{p}_{i-1}^*, \quad i = 2, \dots, n. \tag{3.70}$$

In the presence of measurement errors, we propose a control law in the form of

$$p_1^* = \kappa_1(x_1 + w_1), \tag{3.71}$$

$$p_i^* = \kappa_i(x_i + w_i - p_{i-1}^*), \quad i = 2, \dots, n-1, \tag{3.72}$$

$$u = \kappa_n(x_n + w_n - p_{n-1}^*). \tag{3.73}$$

where the κ_i are appropriately chosen functions. Clearly, control law (3.71)–(3.73) uses the measurements $x_i + w_i$ for $i = 1, \dots, n$.

In this case, because of the discontinuity of w caused by data-sampling, the $e_2, \dots, e_n$ defined by state transformation (3.69) and (3.70) are discontinuous. This may lead to difficulties in stability analysis. Instead, we employ a modified state transformation by using set-valued maps to cover the influence of the measurement errors. Define

$$S_1(\bar{x}_1, \bar{w}_1^\infty) = \{\kappa_1(x_1 + a_1 w_1^\infty) : |a_1| \le 1\} \tag{3.74}$$

$$\begin{aligned} S_i(\bar{x}_i, \bar{w}_i^\infty) = \{\kappa_i(x_i + a_i w_i^\infty - p_{i-1}) :& \\ |a_i| \le 1, p_{i-1} \in S_{i-1}(\bar{x}_{i-1}, \bar{w}_{i-1}^\infty)\},& \\ i = 2, \dots, n.& \end{aligned} \tag{3.75}$$

It can be directly checked that $p_i^* \in S_i(\bar{x}_i, \bar{w}_i^\infty)$ for $i = 1, \dots, n-1$ and $u \in S_n(\bar{x}_n, \bar{w}_n^\infty)$.

The new state variables are defined as

$$e_1 = x_1 \tag{3.76}$$

$$e_i = \vec{d}(x_i, S_{i-1}(\bar{x}_{i-1}, \bar{w}_{i-1}^\infty)), \quad i = 2, \dots, n, \tag{3.77}$$

where, for each $i = 1, \dots, n$, $S_i : \mathbb{R}^i \times \mathbb{R}^i \rightsquigarrow \mathbb{R}$ is an appropriately designed set-valued map to cover the influence of the measurement errors, and $\vec{d}(z, \Omega) := z - \arg\min_{z' \in \Omega}\{|z - z'|\}$ for any $z \in \mathbb{R}$ and any compact $\Omega \subset \mathbb{R}$. In the following procedure, we verify the validity of the control law (3.71)–(3.73) by showing that the closed-loop system with $e = [e_1, \dots, e_n]^T$ is ISS with w as the input.

For convenience of notations, denote $w_0^\infty = \bar{w}_0^\infty = 0$, $e_{n+1} = 0$, $\bar{e}_i = [e_1, \ldots, e_i]^T$ for $i = 1, \ldots, n+1$ and $e = \bar{e}_n$. Lemma 3.3 shows that, with the set-valued map design, each e_i-subsystem for $i = 1, \ldots, n$ can be represented by a differential inclusion and can be rendered ISS with $V_i(e_i) = |e_i|$ as an ISS-Lyapunov function by appropriately choosing κ_i.

Lemma 3.3 *Consider the nonlinear system (3.63) and (3.64) satisfying Assumption 3.4. With the transformation (3.76) and (3.77) and the control law (3.71)–(3.73), by choosing each $\kappa_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, \ldots, n$ to be continuously differentiable, odd and strictly decreasing, when $e_i \neq 0$, each e_i-subsystem can be represented by a differential inclusion as*

$$\dot{e}_i \in S_i(\bar{x}_i, \bar{w}_i^\infty) + \Phi_i(\bar{x}_i, \bar{w}_{i-1}^\infty, e_{i+1}, d). \tag{3.78}$$

Moreover, with specific $\kappa_1, \ldots, \kappa_{i-1}$, for any $\gamma_{e_i}^{e_k}, \gamma_{e_i}^{w_k} \in \mathcal{K}_\infty$ $(k = 1, \ldots, i-1)$ with their inverse functions being Lipschitz on compact sets, any $\gamma_{e_i}^{e_{i+1}}, \gamma_{e_i}^{d} \in \mathcal{K}_\infty$ with their inverse functions being Lipschitz on compact sets, and any constant $0 < c_i < 1$, one can find a κ_i such that the e_i-subsystem is ISS with $V_i(e_i) = |e_i|$ as an ISS-Lyapunov function satisfying

$$\begin{aligned} V_i(e_i) \geq \max_{k=1,\ldots,i-1} & \left\{ \begin{array}{l} \gamma_{e_i}^{e_k}(V_k(e_k)), \gamma_{e_i}^{e_{i+1}}(V_{i+1}(e_{i+1})), \\ \gamma_{e_i}^{w_k}(w_k^\infty), \gamma_{e_i}^{w_i}(w_i^\infty), \gamma_{e_i}^{d}(d^\infty) \end{array} \right\} \\ \Rightarrow \max_{f_i \in F_i(\bar{x}_i, \bar{w}_i^\infty, e_{i+1}, d)} & \nabla V_i(e_i) f_i \leq -\ell_i(V_i(e_i)) \quad a.e. \end{aligned} \tag{3.79}$$

where $\gamma_{e_i}^{w_i}(s) = s/c_i$ for $s \in \mathbb{R}_+$, and $F_i(\bar{x}_i, \bar{w}_i^\infty, e_{i+1}, d) = S_i(\bar{x}_i, \bar{w}_i^\infty) + \Phi_i(\bar{x}_i, \bar{w}_{i-1}^\infty, e_{i+1}, d)$.

The proof of Lemma 3.3 is not proved in this chapter due to the space limitation. The interested reader may consult the gain assignment lemmas in [20, 35] for the proof.

3.5.2 ISS Cyclic-Small-Gain Synthesis

With Lemma 3.3, we can transform the system (3.63) and (3.64) into a network of ISS e_i-subsystems. The interconnection structure of the e-system is shown in Fig. 3.3.

With Lemma 3.3, for each $i = 2, \ldots, n$, with $\kappa_1, \ldots, \kappa_{i-1}$ designed, one can design κ_i such that the following conditions are satisfied at the same time:

(a) the interconnection ISS gains $\gamma_{e_i}^{e_k} \in \mathcal{K}_\infty$ for $1 \leq k \leq i-1$ satisfy the cyclic-small-gain condition [24, 32]:

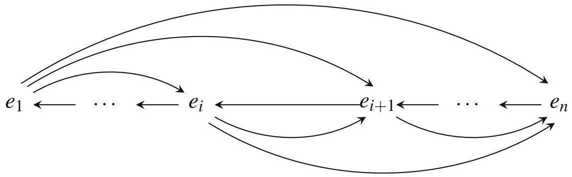


Fig. 3.3 The interconnection structure of the *e*-system

$$
\begin{aligned}
\gamma_{e_1}^{e_2}\circ\gamma_{e_2}^{e_3}\circ\gamma_{e_3}^{e_4}\circ\cdots\circ\gamma_{e_{i-1}}^{e_i}\circ\gamma_{e_i}^{e_1} &< \mathrm{Id},\\
\gamma_{e_2}^{e_3}\circ\gamma_{e_3}^{e_4}\circ\cdots\circ\gamma_{e_{i-1}}^{e_i}\circ\gamma_{e_i}^{e_2} &< \mathrm{Id},\\
&\vdots\\
\gamma_{e_{i-1}}^{e_i}\circ\gamma_{e_i}^{e_{i-1}} &< \mathrm{Id},
\end{aligned}
\tag{3.80}
$$

(b) for each $i = 1, \ldots, n-1$, $\gamma_{e_i}^{e_{i+1}}$ is Lipschitz on compact sets, and
(c) for each $i = 1, \ldots, n$, $\gamma_{e_i}^{w_1}, \ldots, \gamma_{e_i}^{w_i}$ are Lipschitz on compact sets.

The ISS of the *e*-system is guaranteed with the satisfaction of condition (3.80). Conditions (b) and (c) are needed to fulfill the requirement for event-triggered control, as shown in the proof of Theorem 3.8.

The main result in this section is summarized in Theorem 3.8.

Theorem 3.8 *Consider nonlinear uncertain system (3.63) and (3.64) satisfying Assumption 3.4. By choosing $\kappa_1, \ldots, \kappa_n$ according to Lemma 3.3 such that the ISS gains satisfy conditions (a)–(c), one can design a control law in the form of (3.71)–(3.73) to make the closed-loop system ISS. Specifically, there exist $\beta \in \mathscr{KL}$ and $\gamma, \gamma^d \in \mathscr{K}$ such that for any initial state $x(0)$ and any piecewise continuous and bounded w and d,*

$$|x(t)| \leq \max\{\beta(|x(0)|, t), \gamma(|w^\infty|), \gamma^d(d^\infty)\} \tag{3.81}$$

holds for all $t \geq 0$. Moreover, γ can be designed to be Lipschitz on compact sets.

Proof For specific w^∞, with Lemma 3.3, the closed-loop system has been transformed into a large-scale system of ISS e_i-subsystems. With the cyclic-small-gain condition (3.80) satisfied, by using the technique in [32], we construct an ISS-Lyapunov function for the closed-loop system as

$$V(e) = \max_{i=1,\ldots,n}\{\sigma_i(V_i(e_i))\} \tag{3.82}$$

with $\sigma_1 = \mathrm{Id}$ and $\sigma_i = \hat{\gamma}_{e_1}^{e_2}\circ\hat{\gamma}_{e_2}^{e_3}\circ\cdots\circ\hat{\gamma}_{e_{i-1}}^{e_i}$ for $i = 2, \ldots, n$ where $\hat{\gamma}_{e_k}^{e_{k+1}} \in \mathscr{K}_\infty$ for $k = 1, \ldots, n-1$ are chosen such that

- each $\hat{\gamma}_{e_k}^{e_{k+1}}$ is slightly larger than $\gamma_{e_k}^{e_{k+1}}$,
- both $\hat{\gamma}_{e_k}^{e_{k+1}}$ and $\left(\hat{\gamma}_{e_k}^{e_{k+1}}\right)^{-1}$ are Lipschitz on compact sets, and
- the cyclic-small-gain condition (3.80) is still satisfied with the $\gamma_{e_k}^{e_{k+1}}$ replaced by $\hat{\gamma}_{e_k}^{e_{k+1}}$ for $k = 1, \ldots, n-1$.

With conditions (a)–(c) satisfied, such $\hat{\gamma}_{e_k}^{e_{k+1}}$'s for $k = 1, \ldots, n-1$ exist. Then, the functions σ_i and σ_i^{-1} for $i = 1, \ldots, n$ are Lipschitz on compact sets.

From the definition of V in (3.82), one has

$$\underline{\alpha}(|e|) \leq V(e) \leq \overline{\alpha}(|e|) \tag{3.83}$$

where $\underline{\alpha}(s) = \min_{i=1,\ldots,n} \sigma_i(s/n)$ and $\overline{\alpha}(s) = \max_{i=1,\ldots,n} \sigma_i(s)$ for $s \in \mathbb{R}_+$. With the $\hat{\gamma}_{e_k}^{e_{k+1}}$ chosen above, both $\underline{\alpha}$ and $\overline{\alpha}$ are of class $\mathcal{K}_\infty$ and Lipschitz on compact sets.

The influence of the measurement errors w_i for $i = 1, \ldots, n$ can be described by

$$\theta = \max_{i=1,\ldots,n} \left\{ \sigma_i \left(\max_{k=1,\ldots,i} \left\{ \gamma_{e_i}^{w_k}(w_k^\infty) \right\} \right), \sigma_i \circ \gamma_{e_i}^d(d^\infty) \right\}. \tag{3.84}$$

According to the Lyapunov-based cyclic-small-gain theorem [32], it holds that

$$V(e) \geq \theta \Rightarrow \max_{f \in F(x,w^\infty,e,d)} \nabla V(e) f \leq -\alpha(V(e)), \quad \text{a.e.} \tag{3.85}$$

where

$$F(x, w^\infty, e, d) = \begin{bmatrix} F_1(\bar{x}_1, \bar{w}_1^\infty, e_2, d) \\ \vdots \\ F_n(\bar{x}_n, \bar{w}_n^\infty, e_{n+1}, d) \end{bmatrix}^T.$$

Note that $e_{n+1} = 0$.

Define $\gamma_0(s) = \max_{i=1,\ldots,n} \left\{ \sigma_i \left(\max_{k=1,\ldots,i} \left\{ \gamma_{e_i}^{w_k}(s) \right\} \right) \right\}$ and $\gamma_0^d(s) = \max_{i=1,\ldots,n} \left\{ \sigma_i \circ \gamma_{e_i}^d(s) \right\}$ for $s \in \mathbb{R}_+$. Then, $\gamma_0, \gamma_0^d \in \mathcal{K}_\infty$, and γ_0 is Lipschitz on compact sets. With property (3.85), there exists a $\beta_0 \in \mathcal{KL}$ such that

$$V(e(t)) \leq \max\{\beta_0(V(e(0)), t), \gamma_0(|w^\infty|), \gamma_0^d(d^\infty)\} \tag{3.86}$$

for all $t \geq 0$, which together with (3.83) implies

$$|e(t)| \leq \max \left\{ \underline{\alpha}^{-1} \circ \beta_0 \left(\overline{\alpha}(|e(0)|), t \right), \underline{\alpha}^{-1} \circ \gamma_0(|w^\infty|), \underline{\alpha}^{-1} \circ \gamma_0^d(d^\infty) \right\} \tag{3.87}$$

for all $t \geq 0$.

According to the definition of e_i in (3.77), one has

$$|e_i| \leq |x_i - \kappa_{i-1}(e_{i-1})| \leq |x_i| + |\kappa_{i-1}(e_{i-1})| \tag{3.88}$$

for $i = 2, \ldots, n$, where κ_{i-1} has been chosen to be continuously differentiable. Also note that $e_1 = x_1$. Then, one can find an $\alpha_x \in \mathcal{K}_\infty$ being Lipschitz on compact sets such that

$$|e| \leq \alpha_x(|x|). \tag{3.89}$$

Also from the definition of e_i for $i = 1, \ldots, n$ in (3.76) and (3.77), one has

$$|x_1| = |e_1|, \tag{3.90}$$

$$|x_i| \leq \max \left\{ \begin{array}{l} |\max S_{i-1}(\bar{x}_{i-1}, \bar{w}_i^\infty) + e_i|, \\ |\min S_{i-1}(\bar{x}_{i-1}, \bar{w}_i^\infty) - e_i| \end{array} \right\}, \ i = 2, \ldots, n. \tag{3.91}$$

Due to the continuous differentiability of the κ_i's used for the definition of the set-valued maps S_i's, there exist $\alpha_e, \alpha_w \in \mathscr{K}_\infty$ being Lipschitz on compact sets such that

$$|x| \leq \max\{\alpha_e(|e|), \alpha_w(|w^\infty|)\}. \tag{3.92}$$

By substituting (3.89) and (3.92) into (3.87), one achieves (3.81) by defining

$$\beta(s,t) = \alpha_e \circ \underline{\alpha}^{-1} \circ \beta_0 \left(\overline{\alpha} \circ \alpha_x(s), t \right), \tag{3.93}$$

$$\gamma(s) = \max \left\{ \alpha_e \circ \underline{\alpha}^{-1} \circ \gamma_0(s), \alpha_w(s) \right\}, \tag{3.94}$$

$$\gamma^d(s) = \alpha_e \circ \underline{\alpha}^{-1} \circ \gamma_0^d(s) \tag{3.95}$$

for $s, t \in \mathbb{R}_+$. It can be verified that $\beta \in \mathscr{KL}$ and $\gamma, \gamma^d \in \mathscr{K}_\infty$. Since the design of the control law does not depend on w^∞, (3.81) holds for all w^∞ and d^∞. This proves the ISS of the closed-loop system with w and d as the inputs.

As α_e, $\underline{\alpha}^{-1}$, γ_0, and α_w are Lipschitz on compact sets, γ is Lipschitz on compact sets.

Remark 3.6 Input-to-state stabilization plays a central role in several recent results on ISS-based event-triggered control [39, 49]. However, there have not been many previous published results on input-to-state stabilization (and more generally, robust control) of nonlinear systems with measurement errors. In [10], a class of backstepping controllers was developed with set-valued maps and "flattened" Lyapunov functions such that the closed-loop system is ISS with the measurement error as the input. Reference [21] considered nonlinear uncertain systems composed of two subsystems, one is ISS and the other one is input-to-state stabilizable. In [28], it was found that, for general nonlinear control systems under persistently acting disturbances, the existence of smooth Lyapunov functions is equivalent to the existence of (possibly discontinuous) feedback stabilizers which are robust to small measurement errors and small additive external disturbances. However, these results study the general measurement feedback control problem, and may not be directly applicable to event-triggered control as the ISS gain condition may not be satisfied with the designs. It is our belief that the techniques developed in this chapter should be useful for event-based control of other classes of nonlinear systems.

Remark 3.7 For specific w^∞, we have constructed an ISS-Lyapunov function V for the closed-loop system with e as the state. However, the Lyapunov-based event-triggered control design may not be applicable, as the definition of each e_i depends on $\bar{w}_{i-1}^{\infty}$ for $i = 2, \ldots, n$ (cf. (3.77)), which represents the bounds of the measurement errors and is unavailable. This highlights the necessity of the ISS gain condition for event-triggered control without using ISS-Lyapunov functions.

3.6 Event-Trigger Design for Interconnected Systems

In this section, we study an event-trigger design problem for interconnected nonlinear systems. The objective is to develop an ISS gain condition for event-triggered control without infinitely fast sampling.

3.6.1 Problem Formulation

We consider the general case in which a well-designed control system is assumed to be in the form of

$$\dot{z}(t) = h(z(t), x(t), w(t)) \tag{3.96}$$

$$\dot{x}(t) = f(x(t), z(t), w(t)) \tag{3.97}$$

where $[z^T, x^T]^T$ with $z \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ is the system state, $w \in \mathbb{R}^n$ represents the measurement error of x caused by data-sampling, $h : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ represents the system dynamics with $h(0, 0, 0) = 0$ and $f(0, 0, 0) = 0$. Here, z is considered to be unavailable to event-trigger design.

The measurement error $w(t)$ caused by data-sampling is defined as

$$w(t) = x(t_k) - x(t), \quad t \in [t_k, t_{k+1}),\ k \in \mathbb{S} \tag{3.98}$$

where $\{t_k\}_{k \in \mathbb{S}}$ is the sequence of the sampling time instants with $\mathbb{S}$ being the set of the indices of all the sampling time instants. If there is a finite number of sampling time instants, then $\mathbb{S} = \{0, 1, \ldots, k^*\}$ with k^* being the index of the last sampling time instant; otherwise, $\mathbb{S} = \mathbb{Z}_+$. For convenience of notations, if $\mathbb{S} = \{0, 1, \ldots, k^*\}$, we denote $t_{k^*+1} = \infty$.

In event-triggered control, the sampling time instants are generated by the event-trigger as

$$t_{k+1} = \inf_{t>t_k} \{|w(t)| = \mu(t)\}, \quad k \in \mathbb{S} \tag{3.99}$$

where $\mu(t)$ is called threshold signal, and $\mathbb{S}$ is the set of the indices of all the sampling time instants.

It can be directly observed that $w(t)$ is always bounded by the threshold signal, i.e.,

$$|w(t)| \le \mu(t) \tag{3.100}$$

for all $t \ge 0$.

Then, the problem of event-triggered control is reduced to the problem of finding an appropriate threshold signal $\mu(t)$ for the event-trigger such that

- (z, x) asymptotically converges to the origin, and
- for any specified $(z(0), x(0))$, there exists a $\Delta t > 0$ such that

$$t_{k+1} - t_k \ge \Delta t \tag{3.101}$$

for $k \in \mathbb{S}$.

Here, the second objective is to avoid infinitely fast sampling, that is, the sequence $\{t_k\}_{k\in\mathbb{S}}$ has infinite number of elements, i.e., $\mathbb{S} = \mathbb{Z}_+$ and at the same time $\lim_{k\to\infty}(t_{k+1} - t_k) = 0$. Note that a special case is that $\lim_{k\to\infty} t_k < \infty$, which is known as the Zeno behavior.

We assume that both the z-subsystem and the x-subsystem are ISS. More precisely, we make the following assumption on the Lyapunov-based ISS properties of the subsystems.

Assumption 3.5 Both the z-subsystem and the x-subsystem are ISS with locally Lipschitz ISS-Lyapunov functions $V_z : \mathbb{R}^m \to \mathbb{R}_+$ and $V_x : \mathbb{R}^n \to \mathbb{R}_+$ satisfying

$$\underline{\alpha}_z(|z|) \le V_z(z) \le \overline{\alpha}_z(|z|) \tag{3.102}$$

$$V_z(z) \ge \max\{\gamma_z^x(V_x(x)), \gamma_z^w(|w|)\}$$
$$\Rightarrow \nabla V_z(z) h(z, x, w) \le -\alpha_z(V_z(z)), \quad \text{a.e.} \tag{3.103}$$

$$\underline{\alpha}_x(|x|) \le V_x(x) \le \overline{\alpha}_x(|x|) \tag{3.104}$$

$$V_x(x) \ge \max\{\gamma_x^z(V_z(z)), \gamma_x^w(|w|)\}$$
$$\Rightarrow \nabla V_x(x) f(x, z, w) \le -\alpha_x(V_x(x)), \quad \text{a.e.} \tag{3.105}$$

where $\underline{\alpha}_z, \overline{\alpha}_z, \underline{\alpha}_x, \overline{\alpha}_x \in \mathscr{K}_\infty$ and $\gamma_z^x, \gamma_z^w, \gamma_x^z, \gamma_x^w \in \mathscr{K} \cup \{0\}$.

We employ Example 3.2 to show how a event-triggered control system can be transformed into the form of (3.96) and (3.97) satisfying Assumption 3.5. The system in Example 3.2 is also used to show the necessity of a threshold signal which does not decrease exponentially.

Example 3.2 Consider system

$$\dot{z}(t) = -z^3(t) \tag{3.106}$$

$$\dot{x}(t) = u(t) + z(t) \tag{3.107}$$

where $z \in \mathbb{R}$ and $x \in \mathbb{R}$ are the state variables, $u \in \mathbb{R}$ is the control input. We consider the case where only x is available for feedback.

We employ feedback control law:

$$u(t) = -x(t_k), \quad t \in [t_k, t_{k+1}), \quad k \in \mathbb{S} \tag{3.108}$$

where $\{t_k\}_{k\in\mathbb{S}}$ represents the sequence of sampling time instants.

By using (3.98) and (3.108), we have

$$\dot{x}(t) = -x(t) - w(t) + z(t). \tag{3.109}$$

Thus, the controlled system composed of (3.106) and (3.109) is in the form of (3.96) and (3.97) with $h(z, x, w) = -z^3$ and $f(x, z, w) = -x - w + z$.

To verify the satisfaction of Assumption 3.5, we define $V_z(z) = |z|$ and $V_x(x) = |x|$. Clearly, V_z and V_x are locally Lipschitz. It can be directly checked that V_z and V_x satisfy (3.102) and (3.104), respectively, with $\underline{\alpha}_z, \overline{\alpha}_z, \underline{\alpha}_x, \overline{\alpha}_x = \mathrm{Id}$. Also, direct calculation yields:

$$\nabla V_z(z)h(z, x, w) = -|z|^3 = -V_z^3(z) \quad \text{a.e.} \tag{3.110}$$

$$\begin{aligned} \nabla V_x(x)f(x, z, w) &\le -|x| + |w| + |z| \\ &= -V_x(x) + V_z(z) + |w| \\ &\le -V_x(x) + 2\max\left\{V_z(z), |w|\right\} \quad \text{a.e.} \end{aligned} \tag{3.111}$$

Then, property (3.111) implies

$$\begin{aligned} &V_x(x) \ge 4\max\left\{V_z(z), |w|\right\} \\ \Rightarrow\ &\nabla V_x(x)f(x, z, w) \le -0.5V_x(x) \quad \text{a.e.} \end{aligned} \tag{3.112}$$

Thus, properties (3.103) and (3.105) are also satisfied with $\gamma_z^x(s) = 0$, $\gamma_z^w(s) = 0$, $\alpha_z(s) = s^3$, $\gamma_x^z(s) = 4s$, $\gamma_x^w(s) = 4s$, and $\alpha_x(s) = 0.5s$ for $s \in \mathbb{R}_+$.

Under Assumption 3.5, the interconnected system (3.96) and (3.97) is ISS with w as the input if the small-gain condition [22, 23] is satisfied:

$$\gamma_x^z \circ \gamma_z^x < \mathrm{Id}. \tag{3.113}$$

Then, with a direct application of the asymptotic gain property of ISS, if $w(t)$ asymptotically converges to the origin, then $(z(t), x(t))$ converges to the origin. See [48] for the original discussions of the asymptotic gain property.

In [14, 43], the event-triggered control problem is studied in the context of distributed control and exponentially converging threshold signals are used. Based on this idea, we may consider the following threshold signal for our problem:

$$\mu(t) = \mu(0)e^{-ct} \tag{3.114}$$

for $t \geq 0$, with initial state $\mu(0) > 0$ and constant $c > 0$. Or equivalently, $\mu(t)$ is the solution of the initial value problem

$$\dot{\mu}(t) = -c\mu(t). \tag{3.115}$$

However, the discussion above neglects the issue with infinitely fast sampling. An exponentially converging $\mu(t)$ may lead to infinitely fast sampling. Consider Example 3.3.

Example 3.3 Consider the system composed of (3.106) and (3.109), which is in the form of (3.96) and (3.97) with $h(z, x, w) = -z^3$ and $f(x, z, w) = -x - w + z$. It is shown in Example 3.2 that the system satisfies Assumption 3.5. Moreover, the system is a cascade connection of the z-subsystem and the x-subsystem, and thus the small-gain condition (3.113) is satisfied automatically. We show that for some initial states, one cannot find a constant c for the exponentially converging threshold signal (3.114) to avoid infinitely fast sampling.

Now, we show that, for any specific $z(0), x(0), \mu(0)$ and constant c, there exist constants m_1, m_2, m_3, c^* such that

$$|f(x(t), w(t), z(t))| \geq m_1 e^{-t} + m_2 e^{-c^* t} - m_3 e^{-ct} \tag{3.116}$$

for all $t \geq 0$. Moreover, there exist $z(0), x(0), \mu(0)$ such that

$$m_1 \geq 0,\ m_2 > 0,\ m_2 \geq 2m_3,\ 2c^* \leq c, \tag{3.117}$$
$$x(t) > z(t) > 0, \tag{3.118}$$
$$f(x(t), w(t), z(t)) < 0 \tag{3.119}$$

for all $t \geq 0$.

Define $\upsilon = x - z$ and $v = z^3$. Then,

$$\begin{aligned} \dot{\upsilon}(t) &= \dot{x}(t) - \dot{z}(t) \\ &= -\upsilon(t) - w(t) + z^3(t) \\ &\geq -\upsilon(t) - \mu(t) + v(t), \end{aligned} \tag{3.120}$$

and

$$\dot{v}(t) = 3z^2(t)\dot{z}(t) = -3z^5(t) = -3v^{\frac{5}{3}}(t). \tag{3.121}$$

We consider the case of $v(0) > 0$. In this case, $v(t)$ is strictly decreasing and $v(t) \leq v(0)$ for all $t \geq 0$. Then, one can find a $c^* > 0$ such that

$$v(t) \geq v(0)e^{-c^*t} := \breve{v}(t) \tag{3.122}$$

for all $t \geq 0$.

Define $\upsilon^*(t)$ as the solution of the initial value problem

$$\dot{\upsilon}^*(t) = -\upsilon^*(t) - \mu(t) + \breve{v}(t) \tag{3.123}$$

with initial condition $\upsilon^*(0) = \upsilon(0)$. Then, a direct application of the comparison principle yields:

$$\upsilon(t) \geq \upsilon^*(t) \tag{3.124}$$

for all $t \geq 0$

With $\mu(t)$ defined in (3.114) and $\breve{v}(t)$ defined above, we have

$$\begin{aligned}
\upsilon^*(t) &= \upsilon(0)e^{-t} + \int_0^t e^{-(t-\tau)}\left(-\mu(\tau) + \breve{v}(t)\right)d\tau \\
&= \upsilon(0)e^{-t} + e^{-t}\int_0^t e^{\tau}\left(-\mu(0)e^{-c\tau} + \breve{v}(0)e^{-c^*\tau}\right)d\tau \\
&= \left(\upsilon(0) + \frac{\mu(0)}{1-c} - \frac{v(0)}{1-c^*}\right)e^{-t} + \frac{v(0)}{1-c^*}e^{-c^*t} - \frac{\mu(0)}{1-c}e^{-ct}.
\end{aligned} \tag{3.125}$$

Thus,

$$\begin{aligned}
|f(x(t), w(t), z(t))| &= |-x(t) - w(t) + z(t)| \\
&= |\upsilon(t) + w(t)| \\
&\geq \upsilon(t) - \mu(t) \\
&\geq \left(\upsilon(0) + \frac{\mu(0)}{1-c} - \frac{v(0)}{1-c^*}\right)e^{-t} \\
&\quad + \frac{v(0)}{1-c^*}e^{-c^*t} - \frac{\mu(0)}{1-c}e^{-ct} - \mu(0)e^{-ct} \\
&:= m_1e^{-t} + m_2e^{-c^*t} - m_3e^{-ct}
\end{aligned} \tag{3.126}$$

for all $t \geq 0$. Moreover, there exist $z(0) > 0, x(0) > 0, \mu(0) > 0$ such that

$$m_1 \geq 0,\ m_2 > 0,\ m_2 \geq 2m_3,\ 2c^* \leq c. \tag{3.127}$$

In this case, it is directly checked that

$$|f(x(t), w(t), z(t))| \geq v(t) - \mu(t) \geq \frac{m_2}{2} e^{-c^* t} > 0 \tag{3.128}$$

for all $t \geq 0$.

Recall $v(t) = x(t) - z(t)$. With $z(0) > 0$ and $\mu(0) > 0$, $z(t) > 0$, and $\mu(t) > 0$ for all $t \geq 0$. Then, we have

$$x(t) > z(t) > 0 \tag{3.129}$$

for all $t \geq 0$.

With $v(t) - \mu(t) > 0$ given by (3.128), we also have

$$\begin{aligned} f(x(t), w(t), z(t)) &= -x(t) - w(t) + z(t) \\ &= -v(t) - w(t) \\ &\leq -v(t) + \mu(t) < 0 \end{aligned} \tag{3.130}$$

for all $t \geq 0$.

Properties (3.116) and (3.117) together imply

$$|f(x(t), w(t), z(t))| \geq \frac{m_2}{2} e^{-c^* t} \tag{3.131}$$

for all $t \geq 0$.

Given t_k, we give an estimate of the upper bound of $\delta t_k = t_{k+1} - t_k$. With property (3.119), we have

$$\begin{aligned} \mu(t_{k+1}) &= \left| \int_{t_k}^{t_{k+1}} f(x(\tau), w(\tau), z(\tau)) d\tau \right| \\ &= \int_{t_k}^{t_{k+1}} |f(x(\tau), w(\tau), z(\tau))| d\tau. \end{aligned} \tag{3.132}$$

Then, by using (3.114) and (3.131), we have

$$\mu(0) e^{-c(t_k + \delta t_k)} \geq \frac{m_2}{2} e^{-c^*(t_k + \delta t_k)} \delta t_k, \tag{3.133}$$

which implies

$$\delta t_k e^{c^*(t_k+\delta t_k)} \leq \frac{2\mu(0)}{m_2} \tag{3.134}$$

due to $2c^* \leq c$.

Recall property (3.118). Suppose $\mathbb{S} = \{0, 1, \ldots, k^*\}$ with k^* being a positive integer. Then, one can find a $t^* \geq t_{k^*}$ such that

$$z(t) \leq \frac{1}{2}x(t_{k^*}) \tag{3.135}$$

and thus

$$\dot{x}(t) = -x(t_{k^*}) + z(t) \leq -\frac{1}{2}x(t_{k^*}) \tag{3.136}$$

for all $t \geq t^*$. This implies $\lim_{t\to\infty} x(t) = -\infty$ and contradicts with property (3.118). Thus, $\mathbb{S} = \mathbb{Z}_+$.

Now, we consider the following two cases. If $\lim_{k\to\infty} t_k < \infty$, then Zeno behavior occurs. If $\lim_{k\to\infty} t_k = \infty$, then property (3.134) implies $\lim_{k\to\infty} \delta t_k = 0$. In any case, infinitely fast sampling happens.

Note that we assume that the system dynamics are known in Example 3.3. The problem would be more complicated for nonlinear uncertain systems.

From the discussions in Example 3.3, it can be observed that the problem is caused by the nonlinearity z^3 of the z-subsystem. The signal $z(t)$ does not converge to the origin exponentially. Intuitively, the exponential convergence of $\mu(t)$ is too fast compared with the converging rate of $|f(x(t), z(t), w(t))|$, which depends on $z(t)$.

To overcome the limitation of the exponentially decreasing threshold signal, we consider threshold signals generated by more general dynamic systems

$$\dot{\mu}(t) = -\Omega(\mu(t)) \tag{3.137}$$

with $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ being a positive definite function and initial condition $\mu(0) > 0$. Clearly, if $\Omega(s)$ is in the form of ks with constant $k > 0$, then the $\mu(t)$ defined by (3.137) is reduced to the one defined by (3.115).

Under Assumption 3.5, we develop a condition on the ISS gains of the subsystems under which event-triggered control can be realized without infinitely fast sampling.

We consider the interconnected system composed of the z-subsystem (3.96), the x-subsystem (3.97), and the μ-subsystem (3.137). Recall that the measurement error w satisfies (3.100). Under Assumption 3.5, if $w(t)$ is well defined for all $t \geq 0$ and the small-gain condition (3.113) is satisfied, then the interconnected system is asymptotically stable at the origin.

Moreover, we can construct a Lyapunov function for the interconnected system:

$$V_0(z, x, \mu) = \max\left\{\hat{\gamma}_x^z(V_z(z)), V_x(x), \hat{\gamma}_x^w(\mu), \hat{\gamma}_x^z \circ \hat{\gamma}_z^w(\mu)\right\}. \tag{3.138}$$

If $\gamma_{(\cdot)}^{(\cdot)}$ is nonzero, then the corresponding $\hat{\gamma}_{(\cdot)}^{(\cdot)}$ in (3.138) is chosen such that $\hat{\gamma}_{(\cdot)}^{(\cdot)} \in \mathscr{K}_\infty$ and it is continuously differentiable on $(0, \infty)$ and slightly larger than its corresponding $\gamma_{(\cdot)}^{(\cdot)}$; if $\gamma_{(\cdot)}^{(\cdot)} = 0$, then $\hat{\gamma}_{(\cdot)}^{(\cdot)} = 0$. Moreover, $\hat{\gamma}_x^z$ satisfies $\hat{\gamma}_x^z \circ \gamma_z^x < \mathrm{Id}$. See, e.g., [32] for the Lyapunov-based ISS cyclic-small-gain theorem for interconnected nonlinear systems.

Define $\bar{\gamma}_x^w(s) = \max\{\hat{\gamma}_x^w(s), \hat{\gamma}_x^z \circ \hat{\gamma}_z^w(s)\}$ for $s \in \mathbb{R}_+$. Clearly, $\bar{\gamma}_x^w$ is a $\mathscr{K}_\infty$ function and is continuously differentiable on $(0, \infty)$. It is a standard result that

$$\begin{aligned} V(z, x, \mu) &= \left(\bar{\gamma}_x^w\right)^{-1} (V_0(z, x, \mu)) \\ &= \max\left\{ \left(\bar{\gamma}_x^w\right)^{-1} \circ \hat{\gamma}_x^z(V_z(z)), \left(\bar{\gamma}_x^w\right)^{-1} (V_x(x)), \mu \right\} \\ &:= \max\left\{\sigma_z(V_z(z)), \sigma_x(V_x(x)), \mu\right\} \end{aligned} \tag{3.139}$$

is also a Lyapunov function of the interconnected system.

It is shown in the following discussions that, to guarantee (3.101), the decreasing rate of $\mu(t)$ should be chosen in accordance with the decreasing rate of $V(z(t), x(t), \mu(t))$, which is studied in Sect. 3.6.2.

3.6.2 Decreasing Rate of $V(z(t), x(t), \mu(t))$

According to the definition of V in (3.139), the decreasing rate of $V(z(t), x(t), \mu(t))$ depends the decreasing rates of $V_z(z(t))$, $V_x(x(t))$, and $\mu(t)$. Lemma 3.4 gives a condition on Ω under which the decreasing rate of $V(z(t), x(t), \mu(t))$ equals the decreasing rate of $\mu(t)$.

Lemma 3.4 *Consider the interconnected system composed of (3.96), (3.97), and (3.137). Under Assumption 3.5, if (3.113) is satisfied, and if Ω is chosen to be positive definite and satisfies*

$$\Omega(s) \le \min\left\{\partial\sigma_z(\sigma_z^{-1}(s))\alpha_z(\sigma_z^{-1}(s)), \partial\sigma_x(\sigma_x^{-1}(s))\alpha_x(\sigma_x^{-1}(s))\right\} \tag{3.140}$$

for all $s > 0$ with σ_z and σ_x defined in (3.139), then for any $V(z(0), x(0), \mu(0))$,

$$V(z(t), x(t), \mu(t)) \le \eta(t) \tag{3.141}$$

holds for all $t \ge 0$, where $\eta(t)$ is the solution of the initial value problem

$$\dot{\eta}(t) = -\Omega(\eta(t)) \tag{3.142}$$

with initial condition $\eta(0) = V(z(0), x(0), \mu(0))$.

Proof Define $\bar{V}_z(z) = \sigma_z(V_z(z))$ and $\bar{V}_x(x) = \sigma_x(V_x(x))$. Since $\sigma_z, \sigma_x \in \mathscr{K}_\infty$, $\bar{V}_z(z)$, and $\bar{V}_x(x)$ are also ISS-Lyapunov functions of the z-subsystem and the x-subsystem, respectively. Based on (3.103) and (3.105), direct calculation yields:

$$\begin{aligned} &\bar{V}_z(z) \geq \max\{\bar{V}_x(x), |w|\} \\ \Rightarrow &\nabla \bar{V}_z(z) h(z, x, w) \leq -\bar{\alpha}_z(\bar{V}_z(z)) \quad \text{a.e.} \qquad (3.143) \\ &\bar{V}_x(x) \geq \max\{\bar{V}_z(z), |w|\} \\ \Rightarrow &\nabla \bar{V}_x(x) f(x, z, w) \leq -\bar{\alpha}_x(\bar{V}_x(x)) \quad \text{a.e.} \qquad (3.144) \end{aligned}$$

where $\bar{\alpha}_z(s) = \partial\sigma_z(\sigma_z^{-1}(s))\alpha_z(\sigma_z^{-1}(s))$ and $\bar{\alpha}_x(s) = \partial\sigma_x(\sigma_x^{-1}(s))\alpha_x(\sigma_x^{-1}(s))$ for $s \in \mathbb{R}_+$.

Now, we prove

$$D^+ V(z(t), x(t), \mu(t)) \leq -\Omega\left(V(z(t), x(t), \mu(t))\right) \qquad (3.145)$$

for all $t \geq 0$, where D^+ represents the upper right-hand derivative and is defined by

$$D^+ v(t) = \limsup_{h \to 0^+} \frac{v(t+h) - v(t)}{h} \qquad (3.146)$$

for continuous signal $v(t)$.

For convenience of notations, define

$$\mathbb{T}(z, x, \mu) = \left\{\bar{V}_z(z), \bar{V}_x(x), \mu\right\}. \qquad (3.147)$$

Then,

$$V(z, x, \mu) = \max \mathbb{T}(z, x, \mu). \qquad (3.148)$$

Thus,

$$D^+ V(z, x, \mu) = \max\left\{D^+\theta : \theta \in \mathbb{T}(z, x, \mu), \theta = V(z, x, \mu)\right\}. \qquad (3.149)$$

By using (3.143) and (3.144), for any $\theta \in \mathbb{T}(z, x, \mu)$ satisfying $\theta = V(z, x, \mu)$, we have

$$D^+\theta \leq -\bar{\alpha}_\theta(\theta) \qquad (3.150)$$

where

$$\bar{\alpha}_\theta = \begin{cases} \bar{\alpha}_z, & \text{if } \theta = \bar{V}_z(z); \\ \bar{\alpha}_x, & \text{if } \theta = \bar{V}_x(x); \\ \Omega, & \text{if } \theta = \mu. \end{cases} \qquad (3.151)$$

If condition (3.140) is satisfied, then

$$\Omega(s) \leq \min\left\{\bar{\alpha}_z(s), \bar{\alpha}_x(s)\right\} \tag{3.152}$$

for all $s > 0$. In this case, we can replace the $\bar{\alpha}_\theta$ in (3.150) with Ω. Property (3.145) is proved.

Then, property (3.142) can be by directly applying the comparison principle; see, e.g., [26, Lemma 3.4]. This ends the proof of Lemma 3.4.

Remark 3.8 Although Lemma 3.4 only considers the special case with two ISS subsystems and one asymptotically stable subsystem, the converging rate result can be easily extended for large-scale dynamic networks, as long as the cyclic-small-gain condition is satisfied. Due to space limitation, this chapter focuses on the issues closely related to event-triggered control design, and the extension of the convergence rate result is not provided.

3.6.3 Event-Trigger Design

The main result of this section is given in Theorem 3.9.

Theorem 3.9 *Consider the interconnected system composed of (3.96), (3.97), and (3.137) with Assumption 3.5 and (3.113) satisfied. Then, for any specific initial state of the system, there exists a $\Delta t > 0$ such that (3.101) holds for all $k \in \mathbb{S}$ if*

- *Ω is chosen to be positive definite and Lipschitz on compact sets, and satisfies (3.140),*
- *there exists a constant $\Delta > 0$ such that $\Omega(s)/s$ exists and is nondecreasing for $s \in (0, \Delta]$, and*
- *$\left(\sigma_z \circ \underline{\alpha}_z\right)^{-1}$ and $\left(\sigma_x \circ \underline{\alpha}_x\right)^{-1}$ are Lipschitz on compact sets.*

Proof Due to the positive definiteness of Ω, the $\mu(t)$ generated by (3.137) satisfies

$$0 \leq \mu(t) \leq \mu(0) \tag{3.153}$$

for all $t \geq 0$. Moreover, since Ω is chosen to be Lipschitz on compact sets, there exists a constant $\bar{c} > 0$ such that

$$\Omega(s) \leq \bar{c}s \tag{3.154}$$

for $0 \leq s \leq \mu(0)$, and thus

$$\dot{\mu}(t) = -\Omega(\mu(t)) \geq -\bar{c}\mu(t) \tag{3.155}$$

along the trajectory of μ with initial state $\mu(0)$. A direct application of the comparison principle implies

$$\mu(t+\delta) \geq \mu(t)e^{-\bar{c}\delta} \tag{3.156}$$

for all $\delta, t \geq 0$.

For any given t_k, we prove the lower boundedness of $t_{k+1} - t_k := \delta t_k$. By using the event-trigger (3.99), we have

$$\begin{aligned}\mu(t_{k+1}) &= |x(t_{k+1}) - x(t_k)| \\ &= \left|\int_{t_k}^{t_{k+1}} f(x(\tau), z(\tau), \mu(\tau))d\tau\right| \\ &\leq \int_{t_k}^{t_{k+1}} |f(x(\tau), z(\tau), \mu(\tau))|\, d\tau.\end{aligned} \tag{3.157}$$

If the conditions for Theorem 3.9 are satisfied, then the conditions for Lemma 3.4 are satisfied. Thus, the Lyapunov function V defined in (3.139) has property (3.141) with η generated by (3.142). Due to the positive definiteness of Ω, for any initial condition $V(z(0), x(0), \mu(0))$,

$$V(z(t), x(t), \mu(t)) \leq V(z(0), x(0), \mu(0)) \tag{3.158}$$

for all $t \geq 0$, and moreover, there exists a finite time instant $T^* \geq 0$ such that

$$V(z(t), x(t), \mu(t)) \leq \Delta \tag{3.159}$$

for all $t \geq T^*$.

In the following procedure, we consider the cases of $t_k \leq T^*$ and $t_k > T^*$ separately.

Case 1: $t_k \leq T^*$. Property (3.158) means that there exists a finite $\Delta_s > 0$ depending on the initial state such that

$$\left|[z^T(t), x^T(t), \mu(t)]^T\right| \leq \Delta_s \tag{3.160}$$

for all $t \geq 0$. Thus, there exists a Δ_f such that

$$|f(z(t), x(t), \mu(t))| \leq \Delta_f \tag{3.161}$$

for all $t \geq 0$. Then, by also using properties (3.156) and (3.157), we have

$$\begin{aligned}\mu(0)e^{-\bar{c}(t_k+\delta t_k)} &\leq \int_{t_k}^{t_{k+1}} |f(x(\tau), z(\tau), \mu(\tau))|\, d\tau \\ &\leq (t_{k+1} - t_k)\Delta_f = \delta t_k \Delta_f, \end{aligned} \tag{3.162}$$

i.e.,

$$\delta t_k e^{\bar{c}(t_k+\delta t_k)} \geq \frac{\mu(0)}{\Delta_f}. \tag{3.163}$$

In the case of $t_k \leq T^*$, it is concluded that

$$\delta t_k e^{\bar{c}(T^*+\delta t_k)} \geq \frac{\mu(0)}{\Delta_f}. \tag{3.164}$$

Case 2: $t_k > T^*$. Consider an $\eta(t)$ defined by

$$\dot{\eta}(t) = -\Omega(\eta(t)) \tag{3.165}$$

for $t > T^*$ with $\eta(T^*) = V(z(T^*), x(T^*), \mu(T^*))$. Then, by using Lemma 3.4, we have $V(z(t), x(t), \mu(t)) \leq \eta(t)$ for $t > T^*$. Also, by using the definition of V in (3.139), we have $V(z(t), x(t), \mu(t)) \geq \mu(t)$ for all $t \geq 0$.

Thus,

$$\mu(t) \leq V(z(t), x(t), \mu(t)) \leq \eta(t) \tag{3.166}$$

for $t > T^*$.

With a similar reasoning as for (3.156), it can be proved that the $\eta(t)$ defined by (3.165) is strictly positive for all $t > T^*$.

Define

$$k_\mu = \frac{\eta(T^*)}{\mu(T^*)}. \tag{3.167}$$

Then, according to (3.166), $k_\mu \geq 1$. We prove that

$$\eta(t) \leq k_\mu \mu(t) \tag{3.168}$$

for all $t > T^*$.

Since $\Omega(s)/s$ is nondecreasing for all $s \in (0, \Delta]$,

$$\frac{\Omega(\eta)}{\eta} \geq \frac{\Omega\left(\eta/k_\mu\right)}{\eta/k_\mu}, \tag{3.169}$$

which implies $\Omega(\eta)/k_\mu \geq \Omega(\eta/k_\mu)$ for $\eta \in (0, \Delta]$. Then, by using (3.165), we have

$$\frac{1}{k_\mu}\dot{\eta}(t) = -\frac{1}{k_\mu}\Omega(\eta(t)) \le -\Omega\left(\frac{1}{k_\mu}\eta(t)\right) \tag{3.170}$$

for $t > T^*$. Property (3.168) can then be proved by using the comparison principle for $\eta(t)/k_\mu$ and $\mu(t)$.

If $\left(\sigma_z \circ \underline{\alpha}_z\right)^{-1}$ and $\left(\sigma_x \circ \underline{\alpha}_x\right)^{-1}$ are Lipschitz on compact sets, then one can find constants $k_z, k_x > 0$ such that

$$\eta(t) \ge V(z(t), x(t), \mu(t)) \ge \max\left\{k_z(|z(t)|), k_x(|x(t)|)\right\}. \tag{3.171}$$

for all $t > T^*$. Then, with property (3.168), we have

$$\mu(t) \ge \frac{1}{k_\mu} \max\left\{k_z(|z(t)|), k_x(|x(t)|)\right\} \tag{3.172}$$

for all $t > T^*$.

By using the locally Lipschitz property of f, there exists a constant $k_f > 0$ such that

$$|f(z, x, \mu)| \le k_f \max\{|z|, |x|, \mu\} \tag{3.173}$$

for all (z, x, μ) satisfying $V(z, x, \mu) \le V(z(T^*), x(T^*), \mu(T^*))$.

Then, properties (3.157), (3.172), and (3.173) together imply

$$\begin{aligned}
\mu(t_{k+1}) &\le (t_{k+1} - t_k)k_f \max_{t_k \le \tau \le t_{k+1}} \{|z(\tau)|, |x(\tau)|, \mu(\tau)\} \\
&\le (t_{k+1} - t_k)k_f \max_{t_k \le \tau \le t_{k+1}} \left\{k_\mu \mu(\tau)/k_z, k_\mu \mu(\tau)/k_x, \mu(\tau)\right\} \\
&\le \delta t_k k_f \max\left\{k_\mu/k_z, k_\mu/k_x, 1\right\} \mu(t_k).
\end{aligned} \tag{3.174}$$

Also note that (3.156) means

$$\mu(t_{k+1}) \ge e^{-\bar{c}\delta t_k}\mu(t_k). \tag{3.175}$$

Then, we have

$$e^{-\bar{c}\delta t_k} \le \delta t_k k_f \max\left\{k_\mu/k_z, k_\mu/k_x, 1\right\}, \tag{3.176}$$

i.e.,

$$\delta t_k e^{\bar{c}\delta t_k} \ge k_f \max\left\{k_\mu/k_z, k_\mu/k_x, 1\right\}. \tag{3.177}$$

The lower boundedness of δt_k is guaranteed by (3.164) and (3.177) for the two cases, respectively. This ends the proof of Theorem 3.9.

Remark 3.9 The first two conditions listed in Theorem 3.9 are for Ω. Given specific $\partial\sigma_z(s)\alpha_z(\sigma_z^{-1}(s))$ and $\partial\sigma_x(s)\alpha_x(\sigma_x^{-1}(s))$, one can always find a Ω to satisfy the first two conditions. In the following section, we show how to realize the third condition by appropriately choosing control laws for event-triggered output-feedback control of a class of nonlinear uncertain systems.

Example 3.4 The infinitely fast sampling problem arising in Example 3.3 can be readily solved by Theorem 3.9. By using the V_z and V_x defined in Example 3.2, we choose $\hat{\gamma}_x^z(s) = 5s$, $\hat{\gamma}_z^w(s) = 0$, and $\hat{\gamma}_x^w(s) = 5s$ for $s \in \mathbb{R}_+$. According to (3.138), we define

$$V_0(z, x, \mu) = \max\{5V_z(z), V_x(x), 5\mu\}. \tag{3.178}$$

By choosing $\bar{\gamma}_x^w(s) = 5s$ for $s \in \mathbb{R}_+$, we have

$$V(z, x, \mu) = \max\{V_z(z), V_x(x)/5, \mu\}. \tag{3.179}$$

Thus, $\sigma_z(s) = s$ and $\sigma_x(s) = s/5$ for $s \in \mathbb{R}_+$.

It can be verified that $\left(\sigma_z \circ \underline{\alpha}_z\right)^{-1}(s) = s$ and $\left(\sigma_x \circ \underline{\alpha}_x\right)^{-1}(s) = 2s$ for $s \in \mathbb{R}_+$ are Lipschitz on compact sets.

We choose

$$\begin{aligned} \Omega(s) &= \min\left\{\partial\sigma_z(s)\alpha_z(\sigma_z^{-1}(s)), \partial\sigma_x(s)\alpha_x(\sigma_x^{-1}(s))\right\} \\ &= \min\{s^3, s/2\} \end{aligned} \tag{3.180}$$

for $s \in \mathbb{R}_+$. Then, Ω is positive definite and Lipschitz on compact sets, and satisfies (3.140). Also, $\Omega(s)/s$ exists and is non-decreasing for $s \in (0, \infty)$. Intuitively, compared with the $\Omega(s)$ defined in (3.180), the decreasing rate of $\mu(t)$ with $\Omega(s) = cs$ with $c > 0$ is too fast for the nonlinear system.

Remark 3.10 A special case is that both the z-subsystem and the x-subsystem are linear and there ISS-Lyapunov functions are in the quadratic form. In this case, Assumption 3.5 can be modified with $\underline{\alpha}_z(s) = \underline{a}_z s^2$, $\overline{\alpha}_z(s) = \overline{a}_z s^2$, $\underline{\alpha}_x(s) = \underline{a}_x s^2$, $\overline{\alpha}_x(s) = \overline{a}_x s^2$, $\gamma_z^x(s) = b_z^x s$, $\gamma_z^w(s) = b_z^w s^2$, $\gamma_x^z(s) = b_x^z s$, $\gamma_x^w(s) = b_x^w s^2$, $\alpha_z(s) = a_z s$, and $\alpha_x(s) = a_x s$ for $s \in \mathbb{R}_+$ with $\underline{a}_z, \overline{a}_z, \underline{a}_x, \overline{a}_x, b_z^x, b_z^w, b_x^z, b_x^w, a_z, a_z$ being positive constants.

For the linear case, the small-gain condition (3.113) is equivalent to $b_x^z b_z^x < 1$. Assume that the small-gain condition is satisfied. Then, there exists an $\varepsilon > 0$ such that $(b_x^z + \varepsilon)b_z^x < 1$. We choose $\hat{\gamma}_x^z(s) = (b_x^z + \varepsilon)s := \hat{b}_x^z s$, $\hat{\gamma}_x^w(s) = (b_x^w + \varepsilon)s^2 := \hat{b}_x^w s^2$, $\hat{\gamma}_z^w(s) = (b_z^w + \varepsilon)s := \hat{b}_z^w s^2$ for $s \in \mathbb{R}_+$. Then, $\bar{\gamma}_x^w$ can be written in the form $\bar{\gamma}_x^w(s) = \bar{b}_x^w s^2$. It can be calculated that $\sigma_z(s) = \sqrt{\hat{b}_x^z s/\bar{b}_x^w}$ and $\sigma_x(s) = \sqrt{s/\bar{b}_x^w}$. Then, the right-hand side of (3.140) equals $\min\left\{a_z/\hat{b}_x^z, a_x\right\}\bar{b}_x^w s/2$. Thus, one can find a positive constant c such that $\Omega(s) = cs$ satisfies (3.140).

Remark 3.11 With w being the measurement error caused by data-sampling, the event-trigger design in this section considers the systems which are input-to-state stabilizable with w as the input. One of the related results can be found in [21], which considers nonlinear systems composed of two subsystems, one is ISS and the other one is input-to-state stabilizable with respect to the measurement disturbance. In [21], the ISS of the closed-loop system is guaranteed by using the ISS small-gain theorem [22, 23].

A further research direction is to design event-triggered output-feedback controllers for nonlinear uncertain systems, by transforming the closed-loop system into the form of (3.96) and (3.97).

3.6.4 An Extension

In the discussions above, we consider $\mu(t)$ to be generated by system (3.137) with the system dynamics depending solely on $\mu(t)$. A more general case is that $\mu(t)$ is generated by a system

$$\dot{\mu}(t) = \bar{\Omega}(\mu(t), x(t)) \tag{3.181}$$

with $\bar{\Omega} : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ being an appropriately chosen function and $\mu(0) > 0$. In this case, the structure of the interconnected system composed of (3.96), (3.97), and (3.181) is shown in Fig. 3.4.

Under Assumption 3.5, the basic idea is to choose $\bar{\Omega}$ such that

- for all $\mu \in \mathbb{R}_+$ and all $x \in \mathbb{R}^n$,

$$\bar{\Omega}(\mu, x) \geq -\Omega(\mu) \tag{3.182}$$

  where Ω is chosen to satisfy the conditions given in Theorem 3.9;
- system (3.181) is ISS with μ as the state and x as the input, and moreover,

$$\mu \geq \chi_\mu^x(|x|) \Rightarrow \bar{\Omega}(\mu, x) \leq -\alpha_\mu(\mu) \tag{3.183}$$

  where α_μ is a continuous, positive definite function and χ_μ^x is a $\mathscr{K}$ function and satisfies

$$\gamma_x^w \circ \chi_\mu^x \circ \underline{\alpha}_x^{-1} < \mathrm{Id}. \tag{3.184}$$

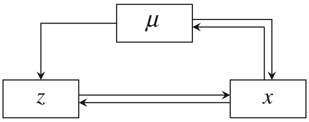


Fig. 3.4 The structure of the interconnected system composed of (3.96), (3.97), and (3.181)

In this case, by using the comparison principle, we have $\mu(t) > 0$ for all $t \geq 0$ if $\mu(0) > 0$. Then, the system composed of subsystems (3.96), (3.97), and (3.181) is an interconnection of ISS subsystems. Conditions (3.113) and (3.184) form the cyclic-small-gain condition for the interconnected system. If (3.113) and (3.184) are satisfied, then the interconnected system is ISS.

With $\mu(t)$ generated by (3.181), one can still guarantee the boundedness of $\mu(t)$ for $t \geq 0$. And with a similar reasoning as for (3.155), one can find a $\bar{c} > 0$ such that

$$\dot{\mu}(t) \geq -\bar{c}\mu(t) \tag{3.185}$$

for all $t \geq 0$. Then, the validity of (3.181) can be proved in the same way as in the proof of Theorem 3.9. Note that in this case, Lemma 3.4 should also be generalized with (3.142) replaced by

$$\dot{\eta}(t) = \bar{\Omega}(\eta(t), x(t)). \tag{3.186}$$

A detailed proof is not given here due to space limitation.

One realization of the $\bar{\Omega}$ in (3.181) to fulfill the requirement of (3.182) and (3.183) is

$$\bar{\Omega}(\mu, x) = -\Omega(\mu) + \pi\left(\chi_0 \circ \chi_\mu^x(|x|) - \chi_0(\mu)\right) \tag{3.187}$$

where Ω is chosen to satisfy the conditions given in Theorem 3.9, χ_0 can be any $\mathscr{K}$ function, and $\pi : \mathbb{R} \to \mathbb{R}_+$ is defined as

$$\pi(r) = \begin{cases} \chi_\pi(r), & \text{if } r \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{3.188}$$

where χ_π can be any $\mathscr{K}$ function. Clearly, with such design, condition (3.183) is satisfied with $\alpha_\mu = \Omega$.

3.6.5 A Simulation Example

We use a simulation to verify the theoretical results. Consider the system given in Example 3.2. The design of an event-trigger (3.99) with the threshold signal $\mu(t)$ generated by (3.137) is given in Example 3.4. We choose initial conditions: $z(0) = -0.2$, $x(0) = 0.2$, and $\mu(0) = 0.5$.

Figure 3.5 shows the convergence of $x(t)$ and the convergence of $w(t)$ bounded by the threshold signal $\mu(t)$. The control signal $u(t)$ and the inter-sampling times $\delta t_k = t_k - t_{k-1}$ during the event-triggered control process is shown in Fig. 3.6. According to the simulation, the minimal inter-sampling time for $0 \leq t \leq 500$ is 1.028.

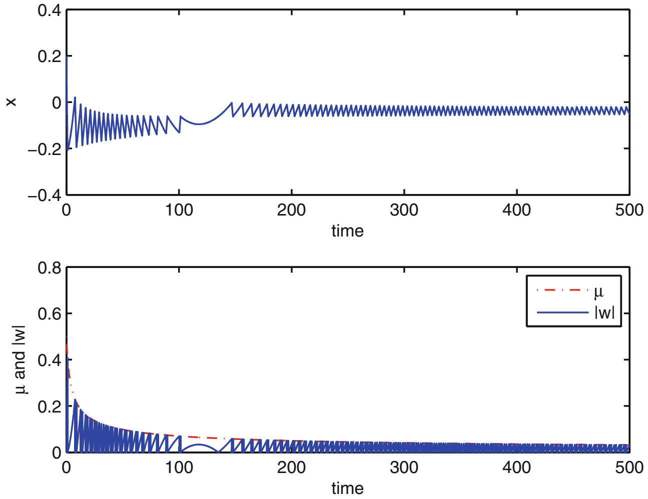


Fig. 3.5 The trajectories of x and w with $\Omega(\mu) = \min\left\{\mu^3, \mu/2\right\}$

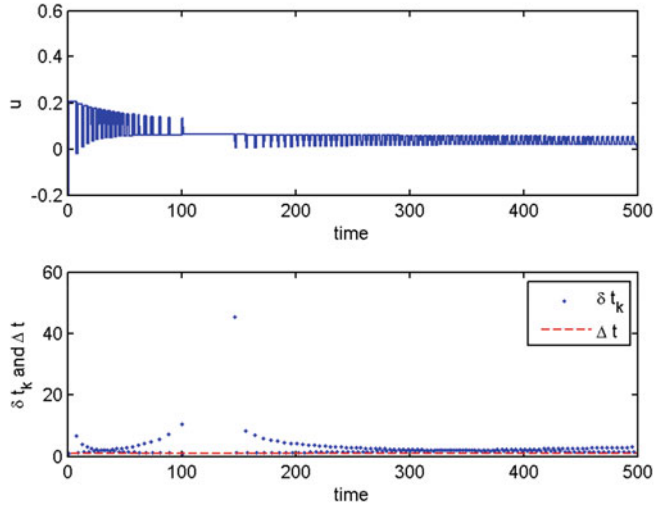


Fig. 3.6 The control signal u and the inter-sampling times $\delta t_k = t_k - t_{k-1}$ with $\Omega(\mu) = \min\left\{\mu^3, \mu/2\right\}$. The minimal inter-sampling time during the period of simulation is 1.028

For comparison, we also consider event-triggers with exponentially decreasing threshold signals. Figure 3.7 shows the inter-sampling times during the control process with $\Omega(\mu) = \mu/2$ and $\Omega(\mu) = \mu/30$, respectively. Clearly, the minimal inter-sampling time cannot be guaranteed to be strictly positive even if $\mu(t)$ decreases very slowly.

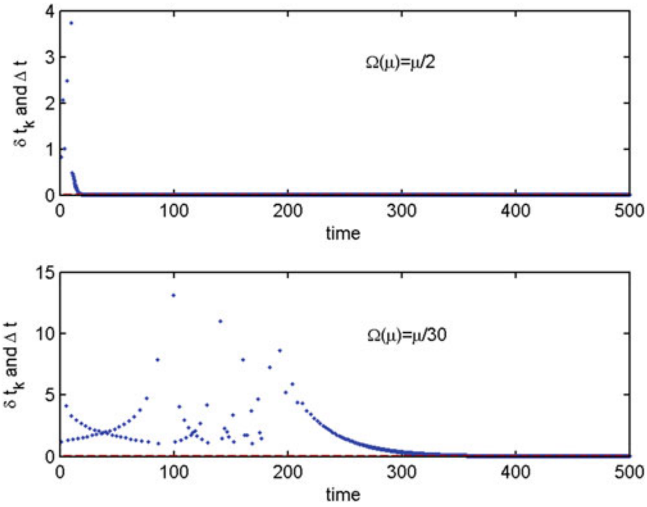


Fig. 3.7 The inter-sampling times $\delta t_k = t_k - t_{k-1}$ with $\Omega(\mu) = \mu/2$ and $\Omega(\mu) = \mu/30$, respectively

3.7 Conclusions

This chapter has studied the event-triggered control problem for nonlinear uncertain systems based on ISS and the nonlinear small-gain theorem. By considering the problem as a robust control problem, we have developed a new ISS gain condition for event-trigger design to avoid infinitely fast sampling. No assumption on the existence of known ISS-Lyapunov functions is made. The basic idea has also been extended to the systems influenced by external disturbances. Through a well-designed self-triggered sampling strategy, input-to-state stabilization of nonlinear systems can be realized by only using the nonperiodic sampled-data. Moreover, asymptotic stabilization can be achieved by means of a self-triggered control algorithm if the external disturbances decay to zero. This chapter has also contributed a new nonlinear control design method for event-triggered control of nonlinear uncertain systems transformable into the strict-feedback form. Event-triggered control of nonlinear uncertain systems with partial-state feedback has also been studied in this chapter. In particular, the event-trigger design problem for the systems that are transformable into an interconnection of two ISS subsystems is solved for the first time. It is shown that infinitely fast sampling can be avoided by considering the threshold signal to be generated by an asymptotically stable system. Based on this result, a more general class of event-triggers with the threshold signals depending on the real-time system state has also been proposed.

Based on the achievements in this chapter, several related topics may be studied in the future:

- Event-triggered control of nonlinear systems with quantized and/or delayed measurements. In a networked control systems, data-sampling and quantization usually coexist. Recently, we have developed small-gain methods for quantized control of nonlinear systems [36]. In the quantized control results, we use ISS gains to represent the influence of quantization error, while in this chapter, we employ an ISS gain to represent the influence of data-sampling. This creates an opportunity to develop a unified framework for event-triggered and quantized control of nonlinear systems. Time-delays also arise from networked control systems. Note that the recent paper [11] has studied event-triggered control for linear systems with quantization and delays. Based on the recent theoretical achievements for nonlinear systems with time-delays [25], it is of interest to study the event-triggered control problem for nonlinear systems by taking into account the effects of time-delays.
- Distributed event-triggered control. The idea of small-gain designs also bridge event-triggered control and our recent distributed control results. In [33], it is shown that a distributed control problem for nonlinear uncertain systems can ultimately be transformed into a stability problem of a network of ISS subsystems. By integrating the idea in this chapter, distributed control could be realized through event-triggered information exchange. Note that such idea has been implemented for linear systems [9, 14, 43, 54].

References

1. Almeida, J., Silvestre, C., Pascoal, A.: Self-triggered state feedback control of linear plants under bounded disturbances. In: Proceedings of the 49th IEEE Conference on Decision and Control, pp. 7588–7593 (2010)
2. Anta, A., Tabuada, P.: To sample or not to sample: self-triggered control for nonlinear systems. IEEE Trans. Autom. Control **55**, 2030–2042 (2010)
3. Arzen, K.-E.: A simple event-based pid controller. In: Proceedings of the 1999 IFAC World Congress, pp. 423–428 (1999)
4. Åström, K.J., Bernhardsson, B.M.: Comparison of Riemann and Lebesgue sampling for first order stochastic systems. In: Proceedings of the 41th IEEE Conference on Decision and Control, pp. 2011–2016 (2002)
5. Ciscato, D., Martiani, L.: On increasing sampling efficiency by adaptive sampling. IEEE Trans. Autom. Control **12**, 318 (1967)
6. De Persis, C., Saile, R., Wirth, F.: Parsimonious event-triggered distributed control: a zeno free approach. Automatica **49**, 2116–2124 (2013)
7. Di Benedetto, M., Di Gennaro, S., D'Innocenzo, A.: Digital self triggered robust control of nonlinear systems. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference, pp. 1674–1679 (2011)
8. Donkers, T., Heemels, M.: Output-based event-triggered control with guaranteed $\mathscr{L}_\infty$-gain and improved and decentralized event-triggering. IEEE Trans. Autom. Control **57**, 1362–1376 (2012)
9. Fan, Y., Feng, G., Wang, Y., Song, C.: Distributed event-triggered control of multi-agent systems with combinational measurements. Automatica **49**, 671–675 (2013)
10. Freeman, R.A., Kokotović, P.V.: Robust Nonlinear Control Design: State-space and Lyapunov Techniques. Birkhäuser, Boston (1996)

11. Garcia, E., Antsaklis, P.J.: Model-based event-triggered control for systems with quantization and time-varying network delays. IEEE Trans. Autom. Control **58**, 422–434 (2013)
12. Gawthrop, P.J., Wang, L.B.: Event-driven intermittent control. Int. J. Control **82**, 2235–2248 (2009)
13. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid dynamical systems: robust stability and control for systems that combine continuous-time and discrete-time dynamics. IEEE Control Syst. Mag. **4**, 28–93 (2009)
14. Guinaldo, M., Dimarogonas, D.V., Johansson, K.H., Sánchez, J., Dormido, S.: Distributed event-based control strategies for interconnected linear systems. IET Control Theory Appl. **7**, 877–886 (2013)
15. Gupta, S.: Increasing the sampling efficiency for a control system. IEEE Trans. Autom. Control **8**, 263–264 (1963)
16. Heemels, W.P.M.H., Donkers, M.C.F., Teel, A.R.: Periodic event-triggered control for linear systems. IEEE Trans. Autom. Control **58**, 847–861 (2013)
17. Heemels, W.P.M.H., Johansson, K.H., Tabuada, P.: An introduction to event-triggered and self-triggered control. In: Proceedings of the 51st IEEE Conference on Decision and Control, pp. 3270–3285 (2012)
18. Heemels, W.P.M.H., Sandee, J.H., Van Den Bosch, P.P.J.: Analysis of event-driven controllers for linear systems. Int. J. Control **81**, 571–590 (2008)
19. Henningsson, T., Johannesson, E., Cervin, A.: Sporadic event-based control of first-order linear stochastic systems. Automatica **44**, 2890–2895 (2008)
20. Jiang, Z.P., Mareels, I.M.Y.: A small-gain control method for nonlinear cascade systems with dynamic uncertainties. IEEE Trans. Autom. Control **42**, 292–308 (1997)
21. Jiang, Z.P., Mareels, I.M.Y., Hill, D.J.: Robust control of uncertain nonlinear systems via measurement feedback. IEEE Trans. Autom. Control **44**, 807–812 (1999)
22. Jiang, Z.P., Mareels, I.M.Y., Wang, Y.: A Lyapunov formulation of the nonlinear small-gain theorem for interconnected systems. Automatica **32**, 1211–1214 (1996)
23. Jiang, Z.P., Teel, A.R., Praly, L.: Small-gain theorem for ISS systems and applications. Math. Control Signals Syst. **7**, 95–120 (1994)
24. Jiang, Z.P., Wang, Y.: A generalization of the nonlinear small-gain theorem for large-scale complex systems. In: Proceedings of the 7th World Congress on Intelligent Control and Automation, pp. 1188–1193 (2008)
25. Karafyllis, I., Jiang, Z.P.: Stability and Stabilization of Nonlinear Systems. Springer, London (2011)
26. Khalil, H.K.: Nonlinear Systems, 3rd edn. Prentice-Hall, NJ (2001)
27. Krstić, M., Kanellakopoulos, I., Kokotović, P.V.: Nonlinear and Adaptive Control Design. Wiley, NY (1995)
28. Ledyaev, Y.S., Sontag, E.D.: A Lyapunov characterization of robust stabilization. Nonlinear Anal. **37**, 813–840 (1999)
29. Lemmon, M.D.: Event-triggered feedback in control, estimation, and optimization. In: Bemporad, A., Heemels, M., Johansson, M. (eds.) Networked Control Systems. Lecture notes in control and information sciences, pp. 293–358. Springer, Berlin (2010)
30. Li, L., Lemmon, M.: Weakly coupled event triggered output feedback system in wireless networked control systems. Discrete Event Dynamic Systems. **24**, 247–260 (2014)
31. Liff, A., Wolf, J.: On the optimum sampling rate for discrete time modeling of continuous-time systems. IEEE Trans. Autom. Control **11**, 288–290 (1966)
32. Liu, T., Hill, D.J., Jiang, Z.P.: Lyapunov formulation of ISS cyclic-small-gain in continuous-time dynamical networks. Automatica **47**, 2088–2093 (2011)
33. Liu, T., Jiang, Z.P.: Distributed formation control of nonholonomic mobile robots without global position measurements. Automatica. **49**, 592–600 (2013)
34. Liu, T., Jiang, Z.P., Hill, D.J.: Robust control of nonlinear strict-feedback systems with measurement errors. In: Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference, pp. 2034–2039 (2011)

35. Liu, T., Jiang, Z.P., Hill, D.J.: a), Decentralized output-feedback control of large-scale nonlinear systems with sensor noise. Automatica **48**, 2560–2568 (2012)
36. Liu, T., Jiang, Z.P., Hill, D.J.: b), A sector bound approach to feedback control of nonlinear systems with state quantization. Automatica **48**, 145–152 (2012)
37. Lunze, J., Lehmann, D.: A state-feedback approach to event-based control. Automatica **46**, 211–215 (2010)
38. Marchand, N., Durand, S., Castellanos, J.F.G.: A general formula for event-based stabilization of nonlinear systems. IEEE Trans. Autom. Control **58**, 1332–1337 (2013)
39. Mazo Jr., M., Anta, A., Tabuada, P.: An ISS self-triggered implementation of linear controllers. Automatica **46**, 1310–1314 (2010)
40. Mitchell, J., McDaniel Jr., W.: Adaptive sampling technique. IEEE Trans. Autom. Control **14**, 200–201 (1969)
41. Nowzari, C., Cortes, J.: Self-triggered coordination of robotic networks for optimal deployment. In: Proceedings of the 2011 American Control Conference, pp. 1039–1044 (2011)
42. Postoyan, R., Tabuada, P., Nesic, D.: A framework for the event-triggered stabilization of nonlinear systems. IEEE Trans. Autom. Control **60**, 983–996 (2015)
43. Seyboth, G.S., Dimarogonas, D.V., Johansson, K.H.: Event-based broadcasting for multi-agent average consensus. Automatica **49**, 249–252 (2013)
44. Sontag, E.D.: Smooth stabilization implies coprime factorization. IEEE Trans. Autom. Control **34**, 435–443 (1989)
45. Sontag, E.D.: Mathematical Control Theory: Deterministic Finite Dimensional Systems. Springer, Berlin (1998)
46. Sontag, E.D.: Input to state stability: basic concepts and results. In: Nistri, P., Stefani, G. (eds.) Nonlinear and Optimal Control Theory, pp. 163–220. Springer, Berlin (2007)
47. Sontag, E.D., Wang, Y.: On characterizations of the input-to-state stability property. Syst. Control Lett. **24**, 351–359 (1995)
48. Sontag, E.D., Wang, Y.: New characterizations of input-to-state stability. IEEE Trans. Autom. Control **41**, 1283–1294 (1996)
49. Tabuada, P.: Event-triggered real-time scheduling of stabilizing control tasks. IEEE Trans. Autom. Control **52**, 1680–1685 (2007)
50. Tallapragada, P., Chopra, N.: On event triggered tracking for nonlinear systems. IEEE Trans. Autom. Control **58**, 2343–2348 (2013)
51. Tomovic, R., Bekey, G.: Adaptive sampling based on amplitude sensitivity. IEEE Trans. Autom. Control **11**, 282–284 (1966)
52. Velasco, M., Marti, P., Fuertes, J.M.: The self triggered task model for real-time control systems. In: Proceedings of 24th IEEE Real-Time Systems Symposium, pp. Work–in–Progress Session (2003)
53. Wang, X., Lemmon, M.D.: Self-triggered feedback control systems with finite-gain $\mathscr{L}_2$ stability. IEEE Trans. Autom. Control **54**, 452–467 (2009)
54. Wang, X., Lemmon, M.D.: a), Event-triggering in distributed networked control systems. IEEE Trans. Autom. Control **56**, 586–601 (2011)
55. Wang, X., Lemmon, M.D.: b), On event design in event-triggered feedback systems. Automatica **47**, 2319–2322 (2011)
56. Yook, J., Tilbury, D., Soparkar, N.: Trading computation for bandwidth: reducing communication in distributed control systems using state estimators. IEEE Trans. Control Syst. Technol. **10**, 503–518 (2002)

Part II
Stabilization and Output Feedback

Chapter 4
An ODE Observer for Lyapunov-Based Global Stabilization of a Bioreactor Nonlinear PDE

Iasson Karafyllis and Miroslav Krstic

Abstract We solve the stabilization problem of a neutrally stable nonlinear hyperbolic PDE with in-domain actuation, which models the population dynamics in a bioreactor, where the "spatial variable" is not the physical space but the age of the microorganisms being grown. The control challenges arise from (1) the structure of the plant dynamics in which the full state of the system gets recirculated back from the PDE domain to the inlet (birth) boundary condition, (2) the fact that control (harvesting rate across the entire age range) multiplies the state, (3) the fact that the state (population density distributed by age) must be kept nonnegative, and (4) the fact that the renewal kernel (the birth rate at different ages) is unknown. We find a nonlinear infinite-dimensional transformation which reveals that the system's relative degree is one and that its zero dynamics are autonomous and exponentially stable, which we prove using a Lyapunov–Krasovskii functional. We take advantage of this structure and achieve stabilization of a desired measured population density under a saturated harvesting input and using a finite-dimensional observer-based feedback, where the observer estimates the harvesting rate set-point, which depends on the unknown renewal kernel (birth rate).

4.1 Introduction

Many areas of nonlinear control would not be what they are-and some would not exist-without the ideas and techniques generated by Laurent Praly. From adaptive robust control and adaptive stabilization, to recursive methodologies such as

I. Karafyllis
Department of Mathematics, National Technical University of Athens,
Zografou Campus, 15780 Athens, Greece
e-mail: iasonkar@central.ntua.gr

M. Krstic (✉)
Departament of Mechanical and Aerospace Engineering, University of California,
San Diego CA 92093-0411, USA
e-mail: krstic@ucsd.edu

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_4

backstepping and forwarding, to robust stabilization approaches, to nonlinear observers and output feedback-Laurent has seen further ahead than anyone in this richly talented field on countless occasions. As someone unparalleled in generously sharing his ideas, Laurent's legacy exists not only in his papers but also in the papers of numerous other researchers who were fortunate enough to meet him or even to read his papers.

This chapter is an example of results that would not have come into existence without the ways of thinking about nonlinear control, adaptive control, and output feedback stabilization that Laurent has either instilled or inspired with his work. The fact that our chapter deals with a problem modeled by a nonlinear partial differential equation, in a particular application area, testifies to the generality of concepts through which Laurent has influenced us and many other researchers.

Our chapter deals with continuous-time age structured population dynamics which are modeled by the so-called McKendrick–von Foerster equation (see [3–5] and the references therein), which is a first order hyperbolic Partial Differential Equation (PDE) with a nonlocal boundary condition.

Optimal control problems for age-structured models have been studied (see [3, 8, 24] and the references therein) while the ergodicity theorem (see [11, 12]) has proved an important tool for the study of the dynamics of continuous-time age structured models (see also [25] for a study of the existence of limit cycles).

The study of feedback control problems was initiated in the recent work [17], where a sampled-data output feedback stabilizer was designed for the global stabilization of an equilibrium age profile for an age-structured chemostat model. Just as in other chemostat feedback control problems described by Ordinary Differential Equations (ODEs; see [9, 13–15, 20, 21]), the dilution rate was selected to be the control input while the output was a weighted integral of the age distribution function. The assumed output functional form was chosen because it is an appropriate form for the expression of the measurement of the total concentration of the microorganism in the bioreactor or for the expression of any other measured variable (e.g., light absorption) that depends on the amount (and its distribution) of the microorganism in the bioreactor. The main idea for the solution of the feedback control problem in [17] was the transformation of the first order hyperbolic PDE to an Integral Delay Equation (IDE; see [16]) and the application of the strong ergodic theorem. This feature differentiated the work in [17] from recent works on feedback control problems for first order hyperbolic PDEs (see [1, 2, 6, 7, 16, 19]).

This work studies the global stabilization problem of an equilibrium age profile for an age-structured chemostat model by means of a continuously applied feedback stabilizer (instead of sampled-data feedback stabilizer as in [17]). Moreover, the present work does not assume knowledge of the equilibrium value of the dilution rate, which was an important assumption in [17]. In other words, the model is completely unknown. A family of observer-based (dynamic), output feedback laws is proposed: the equilibrium value of the dilution rate is estimated by the observer. Moreover, the dilution rate (control input) takes values in a prespecified bounded interval and consequently input constraints are taken into account. The main idea for the solution of the feedback control problem is the transformation of the PDE to an ODE and an IDE.

However, instead of simply designing a continuously applied, dynamic, output feedback law that guarantees global asymptotic stability of an equilibrium age profile, the present work has an additional goal: the explicit construction of a family of Control Lyapunov Functionals (CLFs) for the age-structured chemostat model. Therefore, the present work avoids the application of the strong ergodic theorem (which does not give formulas for Lyapunov functionals) and provides/uses novel stability results on linear IDEs, which are of independent interest. In fact, the newly developed results, provide a Lyapunov-like proof of the scalar, strong ergodic theorem for special cases of the integral kernel. Stability results for linear IDEs similar to those studied in this work have been also studied in [22].

Since the state of the chemostat model is the population density of a particular age at a given time, the state of the chemostat PDE is nonnegative valued. Accordingly, the desired equilibrium profile (a function of the age variable) is positive-valued. So the state space of this PDE system is the positive orthant in a particular function space. We pursue *global* stabilization of the positive equilibrium profile in such a state space. This requires a novel approach and even a novel formulation of stability estimates in which the norm of the state at the desired equilibrium is zero but takes the infinite value not only when the population density (of some age) is infinite but also when it is zero, i.e., we infinitely penalize the population death (the so-called “washout”), as we should. Our main idea in this development is a particular logarithmic transformation of the state, which penalizes both the overpopulated and underpopulated conditions, with an infinite penalty on the washout condition.

The structure of the paper is described next. In Sect. 4.2, we describe the chemostat stabilization problem in a precise way and we provide the main result of the paper (Theorem 4.2.1). Sect. 4.3 provides useful existing results for the uncontrolled PDE, while Sect. 4.4 is devoted to the presentation of stability results on linear IDEs, which allow us to construct CLFs for the chemostat problem. Section 4.5 presents a result (similar to Theorem 4.2.1), which uses a reduced order observer instead of a full-order observer. The concluding remarks of the paper are given in Sect. 4.6. All proofs are omitted.

Notation. Throughout this paper we adopt the following notation.

- For a real number $x \in \mathfrak{R}$, $[x]$ denotes the integer part of $x \in \mathfrak{R}$. $\mathfrak{R}_+$ denotes the interval $[0, +\infty)$.
- Let U be an open subset of a metric space and $\Omega \subseteq \mathfrak{R}^m$ be a set. By $C^0(U;\ \Omega)$, we denote the class of continuous mappings on U, which take values in Ω. When $U \subseteq \mathfrak{R}^n$, by $C^1(U;\ \Omega)$, we denote the class of continuously differentiable functions on U, which take values in Ω. When $U = [a,b) \subseteq \mathfrak{R}$ (or $U = [a,b] \subseteq \mathfrak{R}$) with $a < b$, $C^0([a,b);\ \Omega)$ (or $C^0([a,b];\ \Omega)$) denotes all functions f: $[a,b) \to \Omega$ (or f: $[a,b] \to \Omega$), which are continuous on (a,b) and satisfy $\lim\limits_{s \to a^+}(f(s)) = f(a)$ (or $\lim\limits_{s \to a^+}(f(s)) = f(a)$ and $\lim\limits_{s \to b^-}(f(s)) = f(b)$). When

$U=[a,b)\subseteq \Re$, $C^1([a,b);\ \Omega)$ denotes all functions $f:[a,b)\to\Omega$ which are continuously differentiable on (a,b) and satisfy $\lim\limits_{s\to a^+}(f(s))=f(a)$ and $\lim\limits_{h\to 0^+} h^{-1}(f(a+h)-f(a))=\lim\limits_{s\to a^+} f'(s)$.

- K_∞ is the class of all strictly increasing, unbounded functions $a\in C^0(\Re_+;\Re_+)$, with $a(0)=0$.
- For any subset $S\subseteq\Re$ and for any $A>0$, $PC^1([0,A];S)$ denotes the class of all functions $f\in C^0([0,A];S)$ for which there exists a finite (or empty) set $B\subset(0,A)$ such that: (i) the derivative $f'(a)$ exists at every $a\in(0,A)\backslash B$ and is a continuous function on $(0,A)\backslash B$, (ii) all meaningful right and left limits of $f'(a)$ when a tends to a point in $B\cup\{0,A\}$ exist and are finite.

4.2 Problem Description and Main Result

Consider the age-structured chemostat model:

$$\frac{\partial f}{\partial t}(t,a)+\frac{\partial f}{\partial a}(t,a)=-(\mu(a)+D(t))f(t,a),\ \text{for}\, t>0,\ a\in(0,A) \tag{4.2.1}$$

$$f(t,0)=\int_0^A k(a)f(t,a)da,\ \text{for}\, t\geq 0 \tag{4.2.2}$$

where $D(t)\in[D_{\min},D_{\max}]$ is the dilution rate, $D_{\max}>D_{\min}>0$ are constants, $A>0$ is a constant and $\mu:[0,A]\to\Re_+$, $k:[0,A]\to\Re_+$ are continuous functions with $\int_0^A k(a)da>0$. System (4.2.1), (4.2.2) is a continuous age-structured model of a microbial population in a chemostat. The function $\mu(a)\geq 0$ is called the mortality function, the function $f(t,a)$ denotes the density of the population of age $a\in[0,A]$ at time $t\geq 0$ and the function $k(a)\geq 0$ is the birth modulus of the population. The boundary condition (4.2.2) is the renewal condition, which determines the number of newborn individuals $f(t,0)$. Finally, $A>0$ is the maximum reproductive age. Physically meaningful solutions of (4.2.1), (4.2.2) are only the nonnegative solutions, i.e., solutions satisfying $f(t,a)\geq 0$, for all $(t,a)\in\Re_+\times[0,A]$.

We assume that there exists $D^*\in(D_{\min},D_{\max})$ such that $1=\int_0^A k(a)\exp\left(-D^*a-\int_0^a\mu(s)ds\right)da$. This assumption is necessary for the existence of an equilibrium point for the control system (4.2.1), (4.2.2), which is different from the identically zero function. Any function of the form:

$$f^*(a) = M\exp\left(-D^*a - \int_0^a \mu(s)ds\right), \text{ for } a \in [0, A] \tag{4.2.3}$$

where $M \geq 0$ being an arbitrary constant, is an equilibrium point for the control system (4.2.1), (4.2.2) with $D(t) \equiv D^*$. Notice that there is a continuum of equilibria.

The measured output of the control system (4.2.1), (4.2.2) is given by the equation:

$$y(t) = \int_0^A p(a)f(t,a)da, \text{ for } t \geq 0 \tag{4.2.4}$$

where $p: [0, A] \to \mathfrak{R}_+$ is a continuous function with $\int_0^A p(a)da > 0$. Notice that the case $p(a) \equiv 1$ corresponds to the total concentration of the microorganism in the chemostat.

Let $y^* > 0$ be an arbitrary constant (the set point) and let $f^*(a)$ be the equilibrium age profile given by (4.2.3) with $M = y^*\left(\int_0^A p(a)\exp\left(-D^*a - \int_0^a \mu(s)ds\right)da\right)^{-1}$. Consider the dynamic feedback law given by

$$\begin{aligned} \dot{z}_1(t) &= z_2(t) - D(t) - l_1\left(z_1(t) - \ln\left(\frac{y(t)}{y^*}\right)\right) \\ \dot{z}_2(t) &= -l_2\left(z_1(t) - \ln\left(\frac{y(t)}{y^*}\right)\right) \\ z(t) &= (z_1(t), z_2(t))' \in \mathfrak{R}^2 \end{aligned} \tag{4.2.5}$$

And

$$D(t) = \min\left(D_{\max}, \max\left(D_{\min}, z_2(t) + \gamma\ln\left(\frac{y(t)}{y^*}\right)\right)\right) \tag{4.2.6}$$

where $l_1, l_2, \gamma > 0$ are constants. Next consider solutions of the initial-value problem (4.2.1), (4.2.2), (4.2.4), (4.2.5), (4.2.6) with initial condition $(f_0, z_0) \in \tilde{X} \times \mathfrak{R}^2$, where $\tilde{X}$ is the set $\tilde{X} = \left\{f \in PC^1([0, A]; (0, +\infty)) : f(0) = \int_0^A k(a)f(a)da\right\}$. By a solution of the initial-value problem (4.2.1), (4.2.2), (4.2.4), (4.2.5), (4.2.6) a initial condition $(f_0, z_0) \in \tilde{X} \times \mathfrak{R}^2$, we mean a pair of mappings $f \in C^0([0, T) \times [0, A]; (0, +\infty))$, $z \in C^1([0, T); \mathfrak{R}^2)$, where $T > 0$, which satisfies the following properties:

(i) $f \in C^1\left(D_f;(0,+\infty)\right)$, where $D_f = \{ (t,a) \in (0,T) \times (0,A) : (a-t) \notin B \cup \{0,A\} \}$ and $B \subseteq (0,A)$ is the finite (possibly empty) set where the derivative of $f_0 \in \tilde{X}$ is not defined or is not continuous,

(ii) $f_t \in \tilde{X}$ for all $t \in [0,T)$, where $(f_t)(a) = f(t,a)$ for $a \in [0,A]$,

(iii) Equations (4.2.4), (4.2.5), (4.2.6) hold for all $t \in [0,T)$,

(iv) $\frac{\partial f}{\partial t}(t,a) + \frac{\partial f}{\partial a}(t,a) = -(\mu(a) + D(t))f(t,a)$ holds for all $(t,a) \in D_f$, and

(v) $z(0) = z_0 = (z_{1,0}, z_{2,0})$, $f(0,a) = f_0(a)$ for all $a \in [0,A]$.

The mapping $[0,T) \ni t \to (f_t, z(t)) \in \tilde{X} \times \mathfrak{R}^2$ is called the *solution of the closed-loop system (4.2.1), (4.2.2), (4.2.4) with (4.2.5), (4.2.6)* and initial condition $(f_0, z_0) \in \tilde{X} \times \mathfrak{R}^2$ defined for $t \in [0,T)$.

Define the functional $\Pi\colon C^0([0,A];\mathfrak{R}) \to \mathfrak{R}$ by means of the equation

$$\Pi(f) := \frac{\int_0^A f(a)\left(\int_a^A k(s)\exp\left(\int_s^a (\mu(l) + D^*)dl\right)ds\right)da}{\int_0^A ak(a)f^*(a)da} \tag{4.2.7}$$

and assume that the following technical assumption holds for the nonnegative function $\tilde{k}(a) := k(a)\exp\left(-D^*a - \int_0^a \mu(s)ds\right)$ that satisfies $\int_0^A \tilde{k}(a)da = 1$ (recall that $1 = \int_0^A k(a)\exp\left(-D^*a - \int_0^a \mu(s)ds\right)da$):

(A) *There exists a constant* $\lambda > 0$ *such that* $\int_0^A \left|\tilde{k}(a) - r\lambda\int_a^A \tilde{k}(s)ds\right|da < 1$, *where* $r := \left(\int_0^A a\tilde{k}(a)da\right)^{-1}$ *and* $\tilde{k}(a) := k(a)\exp\left(-D^*a - \int_0^a \mu(s)ds\right)$ *for all* $a \in [0,A]$.

We are now ready to state the main result of the present work.

Theorem 4.2.1 *Consider the age-structured chemostat model (4.2.1), (4.2.2) with* $k \in PC^1([0,A];\mathfrak{R}_+)$ *under Assumption (A). Then for every* $f_0 \in \tilde{X}$ *and* $z_0 \in \mathfrak{R}^2$ *there exists a unique solution of the closed-loop (4.2.1), (4.2.2), (4.2.4) with (4.2.5), (4.2.6) and initial condition* $(f_0, z_0) \in \tilde{X} \times \mathfrak{R}^2$. *Furthermore, there exist a constant* $L > 0$ *and a function* $\rho \in K_\infty$ *such that for every* $f_0 \in \tilde{X}$ *and* $z_0 \in \mathfrak{R}^2$ *the unique solution of the closed-loop (4.2.1), (4.2.2), (4.2.4) with (4.2.5), (4.2.6) and initial condition* $(f_0, z_0) \in \tilde{X} \times \mathfrak{R}^2$ *is defined for all* $t \geq 0$ *and satisfies the following estimate:*

$$\max_{a\in[0,A]}\left(\left|\ln\left(\frac{f(t,a)}{f^*(a)}\right)\right|\right)+|z_1(t)|+\left|z_2(t)-D^*\right|\leq$$
$$\exp\left(-\frac{L}{4}t\right)\rho\left(\max_{a\in[0,A]}\left(\left|\ln\left(\frac{f_0(a)}{f^*(a)}\right)\right|\right)+|z_{1,0}|+\left|z_{2,0}-D^*\right|\right),\text{ for all } t\geq 0 \tag{4.2.8}$$

Moreover, let $p_1,p_2>0$ *be a pair of constants satisfying* $(2+l_1p_1-2l_2p_2)^2<8l_1p_1-4l_2p_1^2$, $p_1^2<4p_2$. *Then the continuous functional* $W\colon\mathfrak{R}^2\times C^0([0,A];(0,+\infty))\to\mathfrak{R}_+$ *defined by:*

$$W(z,f):=(\ln(\Pi(f)))^2+G\sqrt{Q(z,f)}+\beta\,Q(z,f) \tag{4.2.9}$$

where $\beta\geq 0$ *is an arbitrary constant,*

$$Q(z,f):=\frac{M}{2}\left(\frac{\max\limits_{a\in[0,A]}\left(\exp(-\sigma a)\left|\frac{f(a)-\Pi(f)f^*(a)}{f^*(a)}\right|\right)}{\min\left(\Pi(f),\ \min\limits_{a\in[0,A]}\left(\frac{f(a)}{f^*(a)}\right)\right)}\right)^2$$
$$+(z_1-\ln(\Pi(f)))^2-p_1(z_1-\ln(\Pi(f)))(z_2-D^*)+p_2(z_2-D^*)^2 \tag{4.2.10}$$

$\sigma>0$ *is a sufficiently small constant and* $M,G>0$ *are sufficiently large constants, is a Lyapunov functional for the closed-loop system (4.2.1), (4.2.2), (4.2.4) with (4.2.5), (4.2.6), in the sense that every solution* $(f_t,z(t))\in\tilde{X}\times\mathfrak{R}^2$ *of the closed-loop system (4.2.1), (4.2.2), (4.2.4) with (4.2.5), (4.2.6) satisfies the inequality for all* $t\geq 0$:

$$\limsup_{h\to 0^+}\left(h^{-1}(W(z(t+h),f_{t+h})-W(z(t),f_t))\right)$$
$$\leq -L\frac{W(z(t),f_t)}{1+\sqrt{W(z(t),f_t)}} \tag{4.2.11}$$

As remarked in the Introduction, the main result of the present work does not only provide formulas for dynamic output feedback stabilizers that guarantee global asymptotic stability of the selected equilibrium age profile, but also provides explicit formulas for a family of CLFs for system (4.2.1), (4.2.2). Indeed, the continuous functional $W\colon\mathfrak{R}^2\times C^0([0,A];(0,+\infty))\to\mathfrak{R}_+$ defined by (4.2.9), (4.2.10) is a CLF for system (4.2.1), (4.2.2).

Remark 4.2.2

(i) The family of feedback laws (4.2.5), (4.2.6) (parameterized by $l_1,l_2,\gamma>0$) guarantees global asymptotic stabilization of every selected equilibrium age profile. Moreover, the feedback law (4.2.5), (4.2.6) achieves a global exponential convergence rate (see estimate (4.2.8)), in the sense that estimate

(4.2.8) holds for all physically meaningful initial conditions ($f_0 \in \tilde{X}$). As indicated in the Introduction, the logarithmic penalty in (4.2.8) penalizes both the overpopulated and underpopulated conditions, with an infinite penalty on zero density for some age. The state converges to the desired equilibrium profiles from all positive initial conditions, but not from the zero-density initial condition, which itself is an equilibrium (population cannot develop from a "dead" initial state).

(ii) The feedback law (4.2.5), (4.2.6) is a dynamic output feedback law. The subsystem (4.2.5) is an observer that primarily estimates the equilibrium value of the dilution rate D^*. The observer (4.2.5) is a highly reduced order, since it estimates only two variables, the afore-mentioned constant D^* and the scalar functional of the infinite-dimensional state, $\Pi(f)$, introduced in (4.2.7). All the remaining infinitely many states are not estimated. This is the key achievement of our paper-attaining stabilization without the estimation of nearly the entire infinite-dimensional state and proving this result in an appropriately constructed transformed representation of that unmeasured infinite-dimensional state.

(iii) The family of feedback laws (4.2.5), (4.2.6) does not require knowledge of the mortality function of the population, the birth modulus of the population and the maximum reproductive age of the population. It does not require the knowledge of the equilibrium value of the dilution rate D^* (as the sampled-data controller in [17] did). Instead, the equilibrium value of the dilution rate D^* is estimated by the observer state $z_2(t)$ (see estimate (4.2.8)).

(iv) The feedback law (4.2.5), (4.2.6) can work with arbitrary input constraints. The only condition that needs to be satisfied is that the equilibrium value of the dilution rate D^* must satisfy the input constraints, i.e., $D^* \in (D_{\min}, D_{\max})$, which is a reasonable requirement (otherwise the selected equilibrium age profile is not feasible).

(v) The parameters $l_1, l_2, \gamma > 0$ can be used by the control practitioner for the optimal tuning of the controller (4.2.5), (4.2.6): the selection of the values of these parameters affects the value of the constant $L > 0$ that determines the exponential convergence rate. Since the proof of Theorem 4.2.1 is constructive, useful formulas showing the dependence of the constant $L > 0$ on the parameters $l_1, l_2, \gamma > 0$ are established in the proof of Theorem 4.2.1.

(vi) It should be noted that for every pair of constants $l_1, l_2 > 0$ it is possible to find constants $p_1, p_2 > 0$ satisfying $(2 + l_1 p_1 - 2 l_2 p_2)^2 < 8 l_1 p_1 - 4 l_2 p_1^2$, $p_1^2 < 4 p_2$. Indeed, for every $l_1, l_2 > 0$ the matrix $\begin{bmatrix} -l_1 & 1 \\ -l_2 & 0 \end{bmatrix}$ is a Hurwitz matrix. Consequently, there exists a positive definite matrix $\begin{bmatrix} 1 & -p_1/2 \\ -p_1/2 & p_2 \end{bmatrix}$ so that the matrix

$$\begin{bmatrix} -2l_1+l_2p_1 & 1+l_1p_1/2-l_2p_2 \\ 1+l_1p_1/2-l_2p_2 & -p_1 \end{bmatrix} =$$
$$\begin{bmatrix} -l_1 & -l_2 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & -p_1/2 \\ -p_1/2 & p_2 \end{bmatrix}+\begin{bmatrix} 1 & -p_1/2 \\ -p_1/2 & p_2 \end{bmatrix}\begin{bmatrix} -l_1 & 1 \\ -l_2 & 0 \end{bmatrix}$$

is negative definite. This implies the inequalities $p_1^2 < 4p_2$ and $(2+l_1p_1-2l_2p_2)^2 < 8l_1p_1-4l_2p_1^2$.

(vii) The main idea for the construction of the feedback law (4.2.5), (4.2.6) is the transformation of the PDE problem (4.2.1), (4.2.2) into a system that consists of an ODE and an IDE along with the logarithmic output transformation $Y(t)=\ln\left(\frac{y(t)}{y^*}\right)$. The transformations are presented in Fig. 4.1 and are

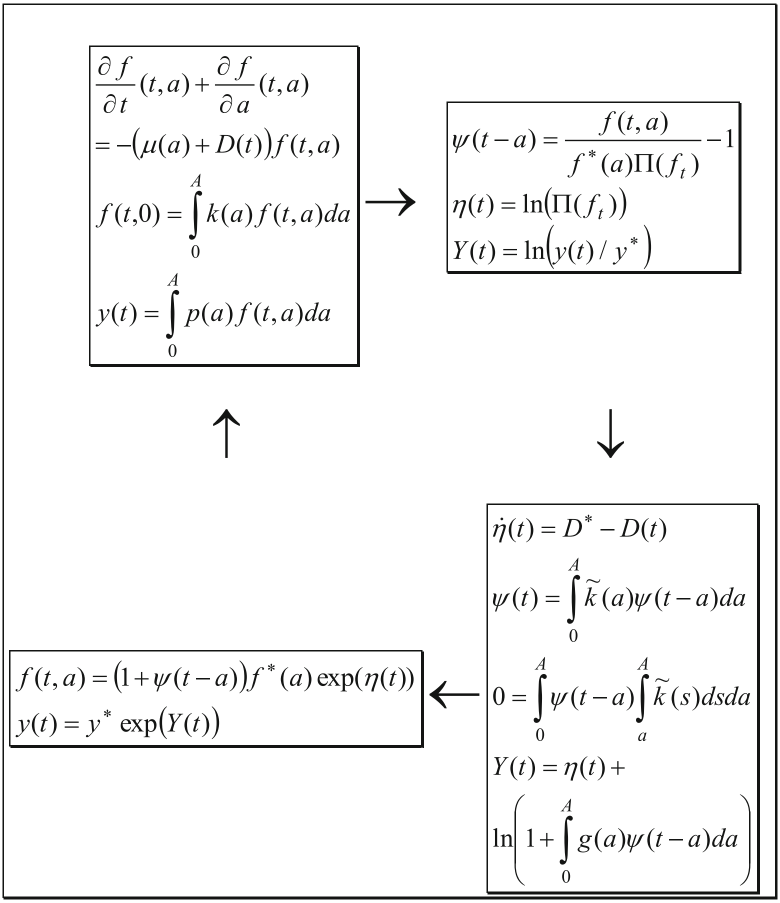

$\frac{\partial f}{\partial t}(t,a)+\frac{\partial f}{\partial a}(t,a)=-\big(\mu(a)+D(t)\big)f(t,a)$
$f(t,0)=\int_0^A k(a)f(t,a)da$
$y(t)=\int_0^A p(a)f(t,a)da$
$\psi(t-a)=\frac{f(t,a)}{f^*(a)\Pi(f_t)}-1$
$\eta(t)=\ln\big(\Pi(f_t)\big)$
$Y(t)=\ln\big(y(t)/y^*\big)$
$\dot{\eta}(t)=D^*-D(t)$
$\psi(t)=\int_0^A \tilde{k}(a)\psi(t-a)da$
$0=\int_0^A \psi(t-a)\int_a^A \tilde{k}(s)dsda$
$Y(t)=\eta(t)+\ln\left(1+\int_0^A g(a)\psi(t-a)da\right)$
$f(t,a)=\big(1+\psi(t-a)\big)f^*(a)\exp(\eta(t))$
$y(t)=y^*\exp\big(Y(t)\big)$

Fig. 4.1 The transformation of the PDE (4.2.1) with boundary condition given by (4.2.2) to an IDE and an ODE and the inverse transformation

exploited rigorously in the proof of Theorem 4.2.1. Fig. 4.1 also shows that the full-order observer (4.2.5) is actually an observer for the system $\dot{\eta}(t) = D^*(t) - D(t)$, $\dot{D}^*(t) = 0$.

Theorem 4.2.1 assumes that the birth modulus of the population satisfies Assumption (A). This is not an assumption that is needed for the establishment of the exponential estimate (4.2.8). Estimate (4.2.8) could have been established without Assumption (A) by means of the strong ergodic theorem. The role of Assumption (A) is crucial for the establishment of the CLF, given by (4.2.9), (4.2.10). However, since Assumption (A) demands a specific property for the function $\tilde{k}(a) := k(a)\exp\left(-D^*a - \int\limits_0^a \mu(s)ds\right)$ that involves the (unknown) equilibrium value of the dilution rate D^*, the verification of the validity of Assumption (A) becomes an issue. The following proposition provides useful sufficient conditions for Assumption (A).

Proposition 4.2.3 *Let* $\tilde{k} \in C^0([0,A];\mathfrak{R})$ *be a function that satisfies the following assumption:*

(B) *The function* $\tilde{k} \in C^0([0,A];\mathfrak{R})$ *satisfies* $\tilde{k}(a) \geq 0$ *for all* $a \in [0,A]$ *and* $\int\limits_0^A \tilde{k}(a)da = 1$. *Moreover, there exists* $\varepsilon > 0$ *such that the set* $S_\varepsilon = \left\{ a \in [0,T] : \tilde{k}(a) \leq \varepsilon \right\}$, *where* $T := \sup\left\{a \in [0,A] : \tilde{k}(a) > 0\right\}$, *has Lebesgue measure* $|S_\varepsilon| < (2r)^{-1}$, *where* $r := \left(\int\limits_0^A a\tilde{k}(a)da\right)^{-1}$.

Then for every $\lambda \in [0, r^{-1}\varepsilon]$ *it holds that* $\int\limits_0^A \left|\tilde{k}(a) - r\lambda \int\limits_a^A \tilde{k}(s)ds\right| da \leq 1 - \lambda(1 - 2r|S_\varepsilon|)$.

Proposition 4.2.3 shows that Assumption (A) is valid for a function that satisfies Assumption (B). On the other hand, we know that Assumption (B) holds for every function $\tilde{k} \in C^0([0,A];\mathfrak{R}_+)$ satisfying $\int\limits_0^A \tilde{k}(a)da = 1$ and having only a finite number of zeros in the interval $[0,A]$. Since $\tilde{k}(a) := k(a)\exp\left(-D^*a - \int\limits_0^a \mu(s)ds\right)$, we can be sure that Assumption (A) necessarily holds *for all birth moduli* $k \in C^0([0,A];\mathfrak{R}_+)$ *of the population with only a finite number of zeros in the interval* $[0,A]$, *no matter what the equilibrium value of the dilution rate* D^* *is and no matter what the mortality function* $\mu : [0,A] \to \mathfrak{R}_+$ *is.*

The basic tool for the proof of the main result is the transformation shown in Fig. 4.1. The main idea comes from the recent work [16]: the transformation of a first order hyperbolic PDE to an IDE. However, if we applied the results of [16] in a straightforward way, then we would end up with the following IDE:

$$v(t)=\int_0^A k(a)\exp\left(-\int_0^a \mu(s)ds\right)\exp\left(-\int_{t-a}^t D(s)ds\right)v(t-a)da, \tag{4.2.12}$$

where $v(t)=f(t,0)$ and $f(t,a)=\exp\left(-\int_0^a \mu(s)ds\right)\exp\left(-\int_{t-a}^t D(s)ds\right)v(t-a)$. However, the IDE is input-dependent. Instead, we would like to describe the effect of the control input in a more convenient way: this is achieved by introducing one more state

$$\eta(t)=\ln(\Pi(f_t)), \tag{4.2.13}$$

where Π is given by (4.2.7). The evolution of $\eta(t)$ is described by the ODE $\dot{\eta}(t)=D^*-D(t)$. Then we are in a position to obtain the transformation

$$\psi(t-a)=\frac{f(t,a)}{f^*(a)\Pi(f_t)}-1,\ \text{for all}\ (t,a)\in\mathfrak{R}_+\times[0,A] \tag{4.2.14}$$

which decomposes the dynamics of (4.2.12) to the input-independent dynamics of the IDE $\psi(t)=\int_0^A \tilde{k}(a)\psi(t-a)da$ evolving on the subspace described by the equation $\int_0^A \psi(t-a)\int_a^A \tilde{k}(s)dsda=0$ and the input-dependent ODE $\dot{\eta}(t)=D^*-D(t)$. After achieving this objective, the next step is the stability analysis of the zero solution of the IDE $\psi(t)=\int_0^A \tilde{k}(a)\psi(t-a)da$: this is exactly the point where the strong ergodic theorem or the results on linear IDEs are used.

4.3 The Uncontrolled PDE

Let $A>0$ be a constant and let $\mu:[0,A]\to\mathfrak{R}_+$, $k:[0,A]\to\mathfrak{R}_+$ be continuous functions with $\int_0^A k(a)da>0$. Consider the initial-value PDE problem:

$$\frac{\partial z}{\partial t}(t,a)+\frac{\partial z}{\partial a}(t,a)=-\mu(a)z(t,a),\ \text{for}\ t>0,\ a\in(0,A) \tag{4.3.1}$$

$$z(t,0)=\int_0^A k(a)z(t,a)da,\ \text{for}\ t\geq 0 \tag{4.3.2}$$

with initial condition $z(0,a)=z_0(a)$ for all $a\in[0,A]$. The following existence and uniqueness result follows directly from Proposition 2.4 in [11] and Theorems 1.3–1.4 on pages 102–104 in [23]:

Lemma 4.3.1 (existence/uniqueness):
For each absolutely continuous function $z_0\in C^0([0,A];\mathfrak{R})$ *with* $z_0(0)=\int_0^A k(a)z_0(a)da$ *, there exists a unique function* $z\colon[0,+\infty)\times[0,A]\to\mathfrak{R}$ *with* $z(0,a)=z_0(a)$ *for all* $a\in[0,A]$ *that satisfies: (a) For each* $t\geq 0$, *the function* z_t *defined by* $(z_t)(a)=z(t,a)$ *for* $a\in[0,A]$ *is absolutely continuous and satisfies* $z_t(0)=\int_0^A k(a)z_t(a)da$ *for all* $t\geq 0$, *(b) the mapping* $\mathfrak{R}_+\ni t\to z_t\in L^1([0,A];\mathfrak{R})$ *is continuously differentiable, and (c) Eq. (4.3.1) holds for almost all* $t>0$ *and* $a\in(0,A)$ *Moreover, if* $z_0(a)\geq 0$ *for all* $a\in[0,A]$ *then* $z(t,a)\geq 0$, *for all* $(t,a)\in\mathfrak{R}_+\times[0,A]$.

The function $z\colon[0,+\infty)\times[0,A]\to\mathfrak{R}$ is called the solution of (4.3.1), (4.3.2). When additional regularity properties hold then the solution of (4.3.1), (4.3.2) satisfies the properties shown by the following lemma.

Lemma 4.3.2 (regularity/relation to IDEs):
If $k\in PC^1([0,A];\mathfrak{R}_+)$, *then for every* $z_0\in PC^1([0,A];\mathfrak{R})$ *satisfying* $z_0(0)=\int_0^A k(a)z_0(a)da$ *the function* $z\colon[0,+\infty)\times[0,A]\to\mathfrak{R}$ *from Lemma 4.3.1 is* C^1 *on*

$$S=\{(t,a)\in(0,+\infty)\times(0,A)\colon(a-t)\notin B\cup\{0,A\}\}$$

where B is the finite (or empty) set where the derivative of z_0 *is not defined, satisfies (4.3.1) on S and Eq. (4.3.2) for all* $t\geq 0$. *Also,*

$$z(t,a)=\exp\left(-\int_0^a \mu(s)ds\right)v(t-a),\ \text{for all}\ (t,a)\in\mathfrak{R}_+\times[0,A] \tag{4.3.3}$$

where $v\in C^0([-A,+\infty);\mathfrak{R})\cap C^1((0,+\infty);\mathfrak{R})$ *is the unique solution of the Integral Delay Equation (IDE):*

$$v(t)=\int_0^A k(a)\exp\left(-\int_0^a \mu(s)ds\right)v(t-a)da,\ \text{for}\ t\geq 0 \tag{4.3.4}$$

with initial condition $v(-a)=\exp\left(\int_0^a \mu(s)ds\right)z_0(a)$, *for all* $a\in(0,A]$.

Lemma 4.3.2 is obtained by integration on the characteristic lines of (4.3.1). The solution $v \in C^0([-A, +\infty); \Re) \cap C^1((0, +\infty); \Re)$ of the IDE (4.3.4) is obtained as the solution of the delay differential equation

$$\dot{v}(t) = \tilde{k}(0)v(t) - \tilde{k}(A)v(t-A) + \int_0^A \frac{d\tilde{k}}{da}(a)v(t-a)da \tag{4.3.5}$$

where $\tilde{k}(a) := k(a)\exp\left(-\int_0^a \mu(s)ds\right)$ for $a \in [0, A]$. The differential Eq. (4.3.5) is obtained by formal differentiation of the IDE (4.3.4) and its solution satisfies (4.3.4) (the verification requires integration by parts).

It is straightforward to show that the function $f(D) = \int_0^A k(a)\exp\left(-Da - \int_0^a \mu(s)ds\right)da$ is strictly decreasing with $\lim_{D \to +\infty} f(D) = 0$ and $\lim_{D \to -\infty} f(D) = +\infty$. Therefore, there exists a unique $D^* \in \Re$ such that

$$1 = \int_0^A k(a)\exp\left(-D^*a - \int_0^a \mu(s)ds\right)da. \tag{4.3.6}$$

Equation (4.3.6) is the Lotka-Sharpe condition [4]. The following strong ergodicity result follows from the results of Sect. 4.3 in [12] and Proposition 3.2 in [11]:

Theorem 4.3.3 (scalar strong ergodic theorem):

Let $D^* \in \Re$ be the unique solution of (4.3.6). Then, there exist constants $\varepsilon > 0$, $K \geq 1$ such that for every absolutely continuous function $z_0 \in C^0([0, A]; \Re)$ with $z_0(0) = \int_0^A k(a)z_0(a)da$, the corresponding solution $z : [0, +\infty) \times [0, A] \to \Re$ of (4.3.1), (4.3.2) satisfies for all $t \geq 0$:

$$\begin{gathered}\int_0^A \left| \exp\left(\int_0^a \mu(s)ds\right) z(t,a) - \exp\left(D^*(t-a)\right)\Phi(z_0) \right| da \\ \leq K\exp\left((D^* - \varepsilon)t\right) \int_0^A \exp\left(\int_0^a \mu(s)ds\right) |z_0(a)| da\end{gathered} \tag{4.3.7}$$

where $\Phi : L^1([0, A]; \Re) \to \Re$ is the linear continuous functional defined by:

$$\Phi(z_0):=\frac{\int_0^A z_0(a)\int_0^A k(s)\exp\left(\int_s^a(\mu(l)+D^*)dl\right)dsda}{\int_0^A ak(a)\exp\left(-\int_0^a(\mu(l)+D^*)dl\right)da} \tag{4.3.8}$$

4.4 Results on Linear Integral Delay Equations

Consider the system described by the following linear IDE:

$$x(t)=\int_0^A \varphi(a)x(t-a)da \tag{4.4.1}$$

where $x(t)\in\Re$, $A>0$ is a constant and $\varphi\in C^0([0,A];\Re)$.

For every $x_0\in C^0([-A,0];\Re)$ with $x_0(0)=\int_0^A\varphi(a)x_0(-a)da$ there exists a unique function $x\in C^0([-A,+\infty);\Re$ that satisfies (4.4.1) for $t\geq 0$ and $x(-a)=x_0(-a)$ for all $a\in[0,A]$. This function is called the solution of (4.4.1) with initial condition $x_0\in C^0([-A,0];\Re)$. The solution is obtained as the solution of the neutral delay equation $\frac{d}{dt}\left(x(t)-\int_0^A\varphi(a)x(t-a)da\right)=0$ (Theorem 1.1 on page 256 in [10] guarantees the existence of a unique function $x\in C^0([-A,+\infty);\Re)\cap C^1((0,+\infty);\Re)$ that satisfies $\frac{d}{dt}\left(x(t)-\int_0^A\varphi(a)x(t-a)da\right)=0$ for $t\geq 0$ and $x(-a)=x_0(-a)$ for all $a\in[0,A]$).

Therefore, the IDE (4.4.1) defines a dynamical system on $X=\left\{x\in C^0([-A,0];\Re):x(0)=\int_0^A\varphi(a)x(-a)da\right\}$ with state $x_t\in X$, where $(x_t)(-a)=x(t-a)$ for all $a\in[0,A]$.

The first result of this section provides useful bounds for the solution of (4.4.1) with non-negative kernel. Notice that the following lemma allows discontinuous solutions of (4.4.1).

Lemma 4.4.1 *Let $\varphi\in C^0([0,A];\Re_+)$ be a given function with $\int_0^A\varphi(a)da\geq 1$ and consider the IDE (4.4.1). Let $\delta>0$ be an arbitrary constant with $\int_0^\delta\varphi(a)da<1$. Then for every $x_0\in L^\infty([-A,0);\Re])$ the solution $x\in L^\infty_{loc}([-A,+\infty);\Re)$ of (4.4.1)* *with initial condition* $x(a) = x_0(a)$ *for* $a \in [-A, 0)$ *exists for all* $t \geq 0$ *and satisfies for all* $t \geq 0$ *the following inequality:*

$$\min\left(\inf_{-A \leq a \leq 0}(x_0(a)), \left(\frac{L-c}{1-c}\right)^{1+h^{-1}t} \inf_{-A \leq a \leq 0}(x_0(a)) \right) \leq \inf_{-A \leq a \leq 0}(x(t+a))$$
$$\leq \sup_{-A \leq a \leq 0}(x(t+a)) \leq \max\left(\left(\frac{L-c}{1-c}\right)^{1+h^{-1}t} \sup_{-A \leq a \leq 0}(x_0(a)), \sup_{-A \leq a \leq 0}(x_0(a)) \right) \tag{4.4.2}$$

where $h := \min(\delta, A-\delta)$, $L := \int\limits_0^A \varphi(a)da \geq 1$, $c = \int\limits_0^\delta \varphi(a)da \geq 1$.

A direct consequence of Lemmas 4.4.1 and 4.3.2 is that if $k \in PC^1([0,A]; \mathfrak{R}_+)$, then for every $z_0 \in PC^1([0,A]; \mathfrak{R})$ satisfying $z_0(0) = \int\limits_0^A k(a) z_0(a) da$ and $z_0(a) > 0$ for all $a \in [0,A]$, the corresponding solution of (4.3.1), (4.3.2) satisfies $z(t,a) > 0$, for all $(t,a) \in \mathfrak{R}_+ \times [0,A]$. To see this, notice that if $\int\limits_0^A k(a) \exp\left(-\int\limits_0^a \mu(s)ds \right) da \geq 1$ then we may apply Lemmas 4.3.2 and 4.4.1 directly for the IDE (4.3.4). On the other hand, if $\int\limits_0^A k(a) \exp\left(-\int\limits_0^a \mu(s)ds \right) da < 1$ then we define $x(t) = \exp(pt)v(t)$ for all $t \geq -A$, where $p > 0$. It follows that $x(t) = \int\limits_0^A k(a) \exp\left(pa - \int\limits_0^a \mu(s)ds \right) x(t-a) da$ for $t \geq 0$ and that $\int\limits_0^A k(a) \exp\left(pa - \int\limits_0^a \mu(s)ds \right) da \geq 1$ for $p > 0$ sufficiently large.

Another direct consequence of Lemmas 4.4.1 and 4.3.2 is that if $k \in PC^1([0,A]; \mathfrak{R}_+)$, then the quantity $\frac{f(t,a)}{f^*(a)\Pi(f_t)} - 1$ appearing in the right-hand side of the transformation (4.2.14) is only a function of $t - a$ (and thus (4.2.14) is a valid transformation). Indeed, it is straightforward to verify that for every piecewise continuous function $D : \mathfrak{R}_+ \to [D_{\min}, D_{\max}]$ and for every $f_0 \in PC^1([0,A]; (0, +\infty))$ with $f_0(0) = \int\limits_0^A k(a) f_0(a) da$, the solution of (4.2.1), (4.2.2) with $f(0,a) = f_0(a)$ for $a \in [0,A]$, corresponding to input $D : \mathfrak{R}_+ \to [D_{\min}, D_{\max}]$ satisfies $f(t,a) = z(t,a) \exp\left(-\int\limits_0^t D(s)ds \right)$ for all $(t,a) \in \mathfrak{R}_+ \times [0,A]$, where $z : [0, +\infty) \times [0,A] \to \mathfrak{R}$ is the solution of (4.3.1), (4.3.2) with same initial condition $z(0,a) = f_0(a)$ for $a \in [0,A]$. Using (4.2.7), (4.2.3), (4.3.3), equation

$f(t,a)=z(t,a)\exp\left(-\int_0^t D(s)ds\right)$ and the fact that $\Pi(f_t)>0$ for all $t\geq 0$ (implied by Lemma 4.4.1), we obtain:

$$\frac{f(t,a)}{f^*(a)\Pi(f_t)}=\frac{v(t-a)\exp(-D^*(t-a))\int_0^A w\tilde{k}(w)dw}{\int_{t-A}^t v(l)\exp(-D^*l)\left(\int_{t-l}^A \tilde{k}(s)ds\right)dl},\text{ for all }(t,a)\in\mathfrak{R}_+\times[0,A]$$

where $\tilde{k}(a)=k(a)\exp\left(-D^*a-\int_0^a\mu(s)ds\right)$ for $a\in[0,A]$. Notice that (4.3.4) implies that $\frac{d}{dt}\int_{t-A}^t v(l)\exp(-D^*l)\left(\int_{t-l}^A\tilde{k}(s)ds\right)dl=0$ and consequently, the quantity $\frac{f(t,a)}{f^*(a)\Pi(f_t)}-1$ is a function of $t-a$.

Next, we state the strong ergodic theorem (Theorem 4.3.3) in terms of the IDE (4.4.1). To this goal, we define the operator

$$G: C^0([-A,0];\mathfrak{R})\to C^0([0,A];\mathfrak{R})$$

for every $v\in C^0([-A,0];\mathfrak{R})$ by the relation $(Gv)(a)=v(-a)$ for all $a\in[0,A]$.

If $\mu\in C^0([0,A];\mathfrak{R}_+)$, $k\in PC^1([0,A];\mathfrak{R}_+)$ satisfy (4.3.6) for certain $D^*\in\mathfrak{R}$, then it follows from Lemma 3.2 and Theorem 4.3.3 that there exist constants $K,\varepsilon>0$ such that for every $z_0\in PC^1([0,A];\mathfrak{R})$ satisfying $z_0(0)=\int_0^A k(a)z_0(a)da$, the unique solution of the IDE (4.3.4) with initial condition $v(-a)=\exp\left(\int_0^a\mu(s)ds\right)z_0(a)$ for all $a\in[0,A]$ satisfies for all $t\geq 0$ the following estimate:

$$\int_0^A\left|v(t-a)-\exp\left(D^*(t-a)\right)\Phi(z_0)\right|da\leq K\exp\left((D^*-\varepsilon)t\right)\int_0^A|v(-a)|da\quad(4.4.3)$$

The above property can be rephrased without any reference to the PDE: for every $\bar{k}\in PC^1([0,A];\mathfrak{R}_+)$ with $1=\int_0^A\bar{k}(a)\exp(-D^*a)da$ there exist constants $K,\varepsilon>0$ such that for every $v_0\in C^0([-A,0];\mathfrak{R})$ with $v_0(0)=\int_0^A\bar{k}(a)v_0(-a)da$ and $(Gv_0)\in PC^1([0,A];\mathfrak{R})$, the unique solution of the IDE $v(t)=\int_0^A\bar{k}(a)v(t-a)da$ with initial condition $v(-a)=v_0(-a)$, for all $a\in[0,A]$ satisfies (4.4.3) for all $t\geq 0$.

Using the transformation $x(t)=\exp(-D^{*}t)v(t)$, for all $t\geq -A$, we obtain a "one-to-one" mapping of solutions of the IDE $v(t)=\int\limits_{0}^{A}\bar{k}(a)v(t-a)da$ to the solutions of the IDE (4.4.1) with $\varphi(a):=\bar{k}(a)\exp(-D^{*}a)$ for all $a\in[0,A]$. Moreover, estimate (4.4.3) implies the following estimate for all $t\leq 0$:

$$\int\limits_{0}^{A}|x(t-a)-P(x_0)|da\leq K\,\exp(-\varepsilon t)\exp\left(D^{*}A\right)\int\limits_{0}^{A}|x(-a)|da$$

Therefore, we are in a position to conclude that the following property holds: for every $\varphi\in PC^{1}([0,A];\mathfrak{R}_{+})$ with $1=\int\limits_{0}^{A}\varphi(a)da$ there exist constants $\widetilde{K},\varepsilon>0$ such that for every $x_0\in C^{0}([-A,0];\mathfrak{R})$ with $x_0(0)=\int\limits_{0}^{A}\varphi(a)x_0(-a)da$ and $(Gx_0)\in PC^{1}([0,A];\mathfrak{R})$, the unique solution of the IDE (4.4.1) with initial condition $x(-a)=x_0(-a)$, for all $a\in[0,A]$ satisfies the following estimate for all $t\geq 0$

$$\int\limits_{0}^{A}|x(t-a)-P(x_0)|da\leq\widetilde{K}\exp(-\varepsilon t)\int\limits_{0}^{A}|x(-a)|da \tag{4.4.4}$$

where the functional $P\colon C^{0}([-A,0];\mathfrak{R})\to\mathfrak{R}$ is defined by means of the equation

$$P(x)=r\int\limits_{0}^{A}x(-a)\int\limits_{a}^{A}\varphi(s)dsda \tag{4.4.5}$$

and $r:=\left(\int\limits_{0}^{A}a\varphi(a)da\right)^{-1}$. Using this property, we obtain the following corollary, which is a restatement of the strong ergodic theorem (Theorem 4.3.3) in terms of IDEs and the L^{∞} norm (instead of the L^{1} norm). Recall that $X=\left\{x\in C^{0}([-A,0];\mathfrak{R}):x(0)=\int\limits_{0}^{A}\varphi(a)x(-a)da\right\}$.

Corollary 4.4.2 *Suppose that $\varphi\in PC^{1}([0,A];\mathfrak{R}_{+})$ with $1=\int\limits_{0}^{A}\varphi(a)da$. Then there exist constants $M,\sigma>0$ such that for every $x_0\in X$ with $(Gx_0)\in PC^{1}([0,A];\mathfrak{R})$, the unique solution of the IDE (4.4.1) with initial condition $x(-a)=x_0(-a)$ for all $a\in[0,A]$ satisfies the following estimate for all $t\geq 0$:*

$$\max_{-A\leq\theta\leq 0}(|x(t+\theta)-P(x_0)|)\leq M\exp(-\sigma t)\max_{-A\leq a\leq 0}(|x_0(a)|) \tag{4.4.6}$$

The problem with Corollary 4.4.2 is that it does not provide a Lyapunov-like functional, which allows the derivation of the important property (4.4.6). Moreover, it does not provide information about the magnitude of the constant $\sigma>0$. In order to construct a Lyapunov-like functional and obtain information about the magnitude of the constant $\sigma>0$, we need some technical results. The first result deals with the exponential stability of the zero solution for (4.4.1). Notice that the proof of the exponential stability property is made by means of a Lyapunov functional.

Lemma 4.4.3 *Suppose that* $\int_0^A|\varphi(a)|da<1$. *Then* $0\leq X$ *is globally exponentially stable for (4.4.1). Moreover, the functional* $V:X\to\mathfrak{R}_+$ *defined by* $V(x):=\max_{a\in[0,A]}(\exp(-\sigma a)|x(-a)|)$, *where* $\sigma>0$ *is a constant that satisfies* $\int_0^A|\varphi(a)|\exp(\sigma a)da<1$, *satisfies the differential inequality:*

$$\lim_{h\to 0^+}\sup(h^{-1}(V(x_{t+h})-V(x_t)))\leq-\sigma V(x_t),\ \text{for all}\ t\leq 0 \tag{4.4.7}$$

for every solution of (4.4.1).

Lemma 4.4.3 is useful because we next construct Lyapunov functionals of the form used in Lemma 4.4.3. However, we are mostly interested in kernels $\varphi\in C^0([0,A];\mathfrak{R})$ with non-negative values that satisfy $\int_0^A\varphi(a)da$. We show next that even for this specific case, it is possible to construct a Lyapunov functional on an invariant subspace of the state space $X=\{x\in C^0([-A,0];\mathfrak{R}):x(0)=\int_0^A\varphi(a)x(-a)da\}$. We next introduce a technical assumption.

(H1) *The function* $\varphi\in C^0([0,A];\mathfrak{R})$ *satisfies* $\varphi(a)\geq 0$ *for all* $a\in[0,A]$ *and* $\int_0^A\varphi(a)da=1$. *Moreover, there exists* $\lambda>0$ *such that* $\int_0^A\left|\varphi(a)-r\lambda\int_a^A\varphi(s)ds\right|da<1$, *where* $r:=\left(\int_0^A a\varphi(a)da\right)^{-1}$.

The following result provides the construction of a Lyapunov functional for system (4.4.1) under assumption (H1).

Theorem 4.4.4 *Consider system (4.4.1), where* $\varphi\in C^0([0,A];\mathfrak{R}_+)$ *satisfies assumption (H1). Let* $\lambda=0$ *be a real number for which* $\int_0^A\left|\varphi(a)-r\lambda\int_a^A\varphi(s)ds\right|da<1$, *where* $r:=\left(\int_0^A a\varphi(a)da\right)^{-1}$. *Define the functional* $V:X\to\mathfrak{R}_+$ *by means of the equation:*

$$V(x) := \max_{a \in [0,A]} \left(\exp(-\sigma a) |x(-a) - P(x)| \right) \tag{4.4.8}$$

where $\sigma > 0$ is a real number for which $\int_0^A \left| \varphi(a) - r\lambda \int_a^A \varphi(s)ds \right| \exp(\sigma a)da < 1$ and $P : X \to \Re$ is the functional defined by (4.4.5). Then the following relations hold

$$P(x_t) = P(x_0), \text{ for all } t \geq 0 \tag{4.4.9}$$

$$\limsup_{h \to 0^+} \left(h^{-1} (V(x_{t+h}) - V(x_t)) \right) \leq -\sigma V(x_t), \text{ for all } t \geq 0 \tag{4.10}$$

for every solution of (4.4.1).

Remark 4.4.5 Theorem 4.4.4 is a Lyapunov-like version of the scalar strong ergodic theorem (compare with Corollary 4.4.2) for kernels that satisfy assumption (H1). Corollary 4.4.2 does not allow us to estimate the magnitude of the constant $\sigma > 0$ that determines the convergence rate. On the other hand, Theorem 4.4.4 allows us to estimate $\sigma > 0$: the Comparison Lemma on page 85 in [18] and differential inequality (4.10) guarantee that $V(x_t) \leq \exp(-\sigma t) V(x_0)$ for all $t \geq 0$ and for every solution of (4.4.1). Using (4.4.9), definition (4.4.8) and the previous estimate, we can guarantee that

$$\begin{aligned} \max_{a \in [0,A]} (|x(t-a) - P(x_t)|) &= \max_{a \in [0,A]} (|x(t-a) - P(x_0)|) \\ &\leq \exp(-\sigma(t-A)) \max_{a \in [0,A]} (|x(-a) - P(x_0)|), \text{ for all } t \geq 0 \end{aligned}$$

Therefore, bounds for $\sigma > 0$ can be computed in a straightforward way using the inequality $\int_0^A \left| \varphi(a) - r\lambda \int_a^A \varphi(s)ds \right| \exp(\sigma a)da < 1$ (e.g., an allowable value for $\sigma > 0$ is $-A^{-1} \ln \left(\int_0^A \left| \varphi(a) - r\lambda \int_a^A \varphi(s)ds \right| da \right)$). Moreover, Corollary 4.4.2 does not provide a Lyapunov-like functional for Eq. (4.4.1). However, the cost of these features is the loss of generality: while Corollary 4.4.2 holds for all kernels $\varphi \in PC^1([0,A]; \Re_+)$ that satisfy $\varphi(a) \geq 0$ for all $a \in [0,A]$ and $\int_0^A \varphi(a)da = 1$, Theorem 4.4.4 holds only for kernels that satisfy Assumption (H1).

Theorem 4.4.4 can allow us to guarantee exponential stability for the zero solution of (4.4.1), when the state evolves in certain invariant subsets of the state space. This is shown in the following result.

Corollary 4.4.6 *Consider system (4.4.1), where $\varphi \in C^0([0,A]; \Re_+)$ satisfies assumption (H1). Let $\lambda > 0$ be a real number for which*

$\int_0^A \left|\varphi(a) - r\lambda \int_a^A \varphi(s)ds\right| da < 1$, *where* $r = \left(\int_0^A a\varphi(a)da\right)^{-1}$. *Let* $P: X \to \Re$ *be the functional defined by (4.4.1). Define the functional* $W: X \to \Re_+$ *by means of the equation:*

$$W(x) := \max_{a \in [0,A]} (\exp(-\sigma a)|x(-a)|) \tag{4.4.11}$$

where $\sigma > 0$ *is a real number for which* $\int_0^A \left|\varphi(a) - r\lambda \int_a^A \varphi(s)ds\right| \exp(\sigma a)da < 1$. *Let* $S \subseteq X$ *be a positively invariant set for system (4.4.1) and let* $C: S' \to [\kappa, +\infty)$, *where* $\kappa > 0$ *is a constant and* $S' \subseteq C^0([-A, 0]; \Re)$ *is an open set with* $S \subset S'$, *be a continuous functional that satisfies*

$$\lim_{h \to 0^+} \sup\left(h^{-1}(C(x_{t+h}) - C(x_t))\right) \le 0 \tag{4.4.12}$$

for every $t \ge 0$ *and for every solution* $x(t) \in \Re$ *of (4.4.1) with* $x_t \in S$. *Then for every* $x_0 \in S$ *with* $P(x_0) = 0$ *and for every* $b \in K_\infty \cap C^1([0, +\infty); \Re_+)$, *the following hold for the solution* $x(t) \in \Re$ *of (4.4.1) with initial condition* $x_0 \in S$:

$$\begin{aligned} &\lim_{h \to 0^+} \sup\left(h^{-1}(C(x_{t+h})b(W(x_{t+h})) - C(x_t)b(W(x_t)))\right) \\ &\quad \le -\sigma C(x_t)b'(W(x_t))W(x_t), \text{ for all } t \ge 0 \end{aligned} \tag{4.4.13}$$

$$P(x_t) = 0, \text{ for all } t \ge 0 \tag{4.4.14}$$

4.5 Using a Reduced Order Observer

Instead of using the full-order observer (4.2.5) of the system $\dot{\eta}(t) = D^*(t) - D(t)$, $\dot{D}^*(t) = 0$, one can think of the possibility of using a reduced order observer that estimates the equilibrium value of the dilution rate D^*. Such a dynamic, output feedback law will be given by the equations:

$$\dot{z}(t) = -l_1 l_2^{-1} z(t) + l_1^2 l_2^{-1} \ln\left(\frac{y(t)}{y^*}\right) - l_1 D(t) \quad , \quad z(t) \in \Re \tag{4.5.1}$$

and

$$D(t) = \min\left(D_{\max}, \max\left(D_{\min}, -l_2^{-1} z(t) + (\gamma + l_1 l_2^{-1}) \ln\left(\frac{y(t)}{y^*}\right)\right)\right) \tag{4.5.2}$$

where $l_1, l_2, \gamma > 0$ are constants. In such a case, a solution of the initial-value problem (4.2.1), (4.2.2), (4.2.4) with (4.5.1), (4.5.2) with initial condition $(f_0, z_0) \in \tilde{X} \times \Re$, where $\tilde{X} = \left\{ f \in PC^1([0,A];(0,+\infty)) : f(0) = \int\limits_0^A k(a)f(a)da \right\}$, is a pair of mappings $f \in C^0([0,T) \times [0,A];(0,+\infty))$, $z \in C^1([0,T);\Re)$, where $T > 0$, which satisfies the following properties:

(i) $f \in C^1(D_f;(0,+\infty))$, where $D_f = \{ (t,a) \in (0,T) \times (0,A) : (a-t) \notin B \cup \{0,A\} \}$ and $B \subseteq (0,A)$ is the finite (possibly empty) set where the derivative of $f_0 \in \tilde{X}$ is not defined or is not continuous,
(ii) $f_t \in \tilde{X}$ for all $t \in [0,T)$, where $(f_t)(a) = f(t,a)$ for $a \in [0,A]$,
(iii) Equations (4.2.4), (4.5.1), (4.5.2) hold for all $t \in [0,T)$,
(iv) $\frac{\partial f}{\partial t}(t,a) + \frac{\partial f}{\partial a}(t,a) = -(\mu(a) + D(t))f(t,a)$ holds for all $(t,a) \in D_f$, and
(v) $z(0) = z_0$, $f(0,a) = f_0(a)$ for all $a \in [0,A]$.

The mapping $[0,T) \ni t \to (f_t, z(t)) \in \tilde{X} \times \Re$ is called the *solution of the closed-loop system (4.2.1), (4.2.2), (4.2.4) with (4.5.1), (4.5.2)* and initial condition $(f_0, z_0) \in \tilde{X} \times \Re$ defined for $t \in [0,T)$.

For the reduced order observer case, we are in a position to prove, exactly in the same way of proving Theorem 4.2.1, the following result.

Theorem 4.5.1 *Consider the age-structured chemostat model (4.2.1), (4.2.2) with $k \in PC^1([0,A];\Re_+)$ under Assumption (A). Then for every $f_0 \in \tilde{X}$ and $z_0 \in \Re$ there exists a unique solution of the closed-loop (4.2.1), (4.2.2), (4.2.4) with (4.5.1), (4.5.2) and initial condition $(f_0, z_0) \in \tilde{X} \times \Re$. Furthermore, there exist a constant $L > 0$ and a function $\rho \in K_\infty$ such that for every $f_0 \in \tilde{X}$ and $z_0 \in \Re$ the unique solution of the closed-loop (4.2.1), (4.2.2), (4.2.4) with (4.5.1), (4.5.2) and initial condition $(f_0, z_0) \in \tilde{X} \times \Re$ is defined for all $t \geq 0$ and satisfies the following estimate:*

$$\begin{aligned} &\max_{a \in [0,A]} \left(\left| \ln\left(\frac{f(t,a)}{f^*(a)} \right) \right| \right) + \left| z(t) + l_2 D^* \right| \leq \\ &\exp\left(-\frac{L}{4} t \right) \rho\left(\max_{a \in [0,A]} \left(\left| \ln\left(\frac{f_0(a)}{f^*(a)} \right) \right| \right) + \left| z_0 + l_2 D^* \right| \right), \text{ for all } t \geq 0 \end{aligned} \tag{4.5.3}$$

Moreover, the continuous functional $W: \Re \times C^0([0,A];(0,+\infty)) \to \Re_+$ defined by:

$$W(z,f) := (\ln(\Pi(f)))^2 + G\sqrt{Q(z,f)} + \beta\, Q(z,f) \tag{4.5.4}$$

where $\beta \geq 0$ is an arbitrary constant,

$$Q(z,f):=\left(z-l_1\ln(\Pi(f))+l_2D^*\right)^2+\frac{M}{2}\left(\frac{\max\limits_{a\in[0,A]}\left(\exp(-\sigma a)\left|\frac{f(a)-\Pi(f)f^*(a)}{f^*(a)}\right|\right)}{\min\left(\Pi(f),\ \min\limits_{a\in[0,A]}\left(\frac{f(a)}{f^*(a)}\right)\right)}\right)^2, \tag{4.5.5}$$

$\Pi: C^0([0,A];\mathfrak{R})\to\mathfrak{R}$ is given by (4.2.7), $\sigma>0$ is a sufficiently small constant and $M,G>0$ are sufficiently large constants, is a Lyapunov functional for the closed-loop system (4.2.1), (4.2.2), (4.2.4) with (4.5.1), (4.5.2), in the sense that every solution $(f_t,z(t))\in\tilde{X}\times\mathfrak{R}$ of the closed-loop system (4.2.1), (4.2.2), (4.2.4) with (4.5.1), (4.5.2) satisfies the inequality:

$$\begin{aligned}&\lim_{h\to0^+}\sup\left(h^{-1}\left(W(z(t+h),f_{t+h})-W(z(t),f_t)\right)\right)\\&\le -L\,\frac{W(z(t),f_t)}{1+\sqrt{W(z(t),f_t)}},\ \text{for all}\ t\ge0\end{aligned} \tag{4.5.6}$$

The family of dynamic, bounded, output feedback laws (*4.5.1*), (*4.5.2*) presents the same features as the family (*4.2.5*), (*4.2.6*). The only difference lies in the dimension of the observer.

4.6 Concluding Remarks

Age-structured chemostats present challenging control problems for first order hyperbolic PDEs that require novel results. We studied the problem of stabilizing an equilibrium age profile in an age-structured chemostat, using the dilution rate as the control. We built a family of dynamic, bounded, output feedback laws that ensures stability under arbitrary physically meaningful initial conditions. Our control does not require knowledge of the model, in contrast to the sampled-data feedback law proposed in [17] that required knowledge of the equilibrium value of the dilution rate. We also provided a family of CLFs for the age-structured chemostat model. The construction of the CLF was based on novel stability results on linear IDEs, which are of independent interest. The newly developed results provide a Lyapunov-like proof of the scalar, strong ergodic theorem for special cases of the integral kernel.

Since the growth of the microorganism may sometimes depend on the concentration of a limiting substrate, it would be useful to solve the stabilization problem for an enlarged system that has one PDE for the age distribution, coupled with one ODE for the substrate (as proposed in [25], in the context of studying limit cycles with constant dilution rates instead of a control). This is going to be the topic of our future research.

References

1. Bastin, G., Coron, J.-M.: On Boundary Feedback Stabilization of Non-Uniform Linear 2x2 Hyperbolic Systems Over a Bounded Interval. Syst. Control Lett. **60**, 900–906 (2011)
2. Bernard, P., Krstic, M.: Adaptive Output-Feedback Stabilization of Non-Local Hyperbolic PDEs. Automatica **50**, 2692–2699 (2014)
3. Boucekkine, R., Hritonenko, N., Yatsenko, Y.: Optimal Control of Age-structured Populations in Economy, Demography, and the Environment (Google eBook), (2013)
4. Brauer, F., Castillo-Chavez, C.: Mathematical Models in Population Biology and Epidemiology. Springer-Verlag, New York (2001)
5. Charlesworth, B.: Evolution in Age-structured Populations, 2nd edn, Cambridge University Press, (1994)
6. Coron, J.-M., Vazquez, R., Krstic, M., Bastin, G.: Local Exponential H2 Stabilization of a 2x2 Quasilinear Hyperbolic System Using Backstepping. SIAM Journal of Control and Optimization **51**, 2005–2035 (2013)
7. Di Meglio, F., Vazquez, R., Krstic, M.: Stabilization of a System of $n + 1$ Coupled First-Order Hyperbolic Linear PDEs with a Single Boundary Input. IEEE Trans. Autom. Control **58**, 3097–3111 (2013)
8. Feichtinger, G., Tragler, G., Veliov, V.M.: Optimality Conditions for Age-Structured Control Systems. J. Math. Anal. Appl. **288**(1), 47–68 (2003)
9. Gouze, J.L., Robledo, G.: Robust Control for an Uncertain Chemostat Model, Int. J. Robust Nonlinear Control. **16**(3), 133–155, (2006)
10. Hale, J.K., Lunel, S.M.V.: Introduction to Functional Differential Equations. Springer-Verlag, New York (1993)
11. Inaba, H., A.: Semigroup approach to the strong ergodic theorem of the multistate stable population process, Math. Popul. Stud. **1**(1), 49–77, (1988)
12. Inaba, H.: Asymptotic properties of the inhomogeneous Lotka-von foerster system, Math. Popul. Stud. **1**(3), 247–264, (1988)
13. Karafyllis, I., Kravaris, C., Syrou, L., Lyberatos, G.: A vector lyapunov function characterization of input-to-state stability with application to robust global stabilization of the chemostat, Eur. J Control. **14**(1), 47–61, (2008)
14. Karafyllis, I., Kravaris, C., Kalogerakis, N.: Relaxed Lyapunov Criteria for Robust Global Stabilization of Nonlinear Systems. Int. J. Control **82**(11), 2077–2094 (2009)
15. Karafyllis, I., Jiang, Z.-P.: A New Small-Gain Theorem with an Application to the Stabilization of the Chemostat. Int. J. Robust Nonlinear Control **22**(14), 1602–1630 (2012)
16. Karafyllis, I., Krstic,M.: On the relation of delay equations to first-order hyperbolic partial differential equations. ESAIM; Control, Optim. Calc.Var. **20**(3), 894–923, (2014)
17. Karafyllis, I., Malisoff, M., Krstic,M.: Sampled-data feedback stabilization of Age-structured chemostat models.In; Proceedings of the American Control Conference, Chicago, IL, U.S.A., pp. 4549–4554, (2015)
18. Khalil, H.K.: Nonlinear systems, 2nd edn, Prentice-Hall, (1996)
19. Krstic, M., Smyshlyaev, A.: Backstepping boundary control for first-order hyperbolic PDEs and application to systems with actuator and sensor delays, Syst. Control Lett. **57**(9), 750–758, (2008)
20. Mazenc, F., Malisoff, M., Harmand, J.: Stabilization and robustness analysis for a chemostat model with two species and monod growth rates via a lyapunov approach. In: Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, (2007)
21. Mazenc, F., Malisoff, M., Harmand, J.: Further results on stabilization of periodic trajectories for a chemostat with two species, *IEEE* Trans. Autom.Control. **53**(1), 66–74, (2008)
22. Melchor-Aguilar, D.: Exponential Stability of Some Linear Continuous Time Difference Systems. Syst. Control Lett. **61**, 62–68 (2012)
23. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Springer-Verlag, New York (1983)

24. Sun, B.: Optimal control of age-structured population dynamics for spread of universally fatal diseases II, Appl.Anal. Int. J. **93**(8), 1730–1744, (2014)
25. Toth, D., Kot, M.: Limit Cycles in a Chemostat Model for a Single Species with Age Structure. Math. Biosci. **202**, 194–217 (2006)

Chapter 5
From Pure State and Input Constraints to Mixed Constraints in Nonlinear Systems

Willem Esterhuizen and Jean Lévine

Abstract We survey the results on the problem of pure/mixed state and input constrained control, with multidimensional constraints, for finite dimensional nonlinear differential systems with focus on the so-called *admissible set* and its boundary. The admissible set is the set of initial conditions for which there exist a control and an integral curve satisfying the constraints for all time. Its boundary is made of two disjoint parts: the subset of the state constraint boundary on which there are trajectories pointing towards the interior of the admissible set or tangentially to it; and a *barrier*, namely a semipermeable surface which is constructed via a generalized minimum-like principle with nonsmooth terminal conditions. Comparisons between pure state constraints and mixed ones are presented on a series of simple academic examples.

5.1 Introduction

Though constrained systems, namely with restrictions on the control and the state, are present in many applications due to actuator limitations and obstacles, they are not generally studied on their own and are more often studied in the context of optimal control or differential games [8]. We focus here on a fully qualitative approach, i.e., without any optimisation framework where the aim is the construction of the set of initial conditions such that the system variables can satisfy the constraints for all time, called *admissible set*, and we show how to compute its boundary. Other approaches based on flow computation, or Lyapunov functions, or other variants, may be found in [1, 2, 11–14, 16–20].

We first review the results of [6] for pure state and input constraints (Sect. 5.2) and present a simple example of double integrator. In a second part (Sect. 5.3), we

W. Esterhuizen · J. Lévine (✉)
CAS, Mathématiques et Systèmes, MINES-ParisTech, PSL Research University,
35 rue Saint-Honoré, 77300 Fontainebleau, France
e-mail: jean.levine@mines-paristech.fr

W. Esterhuizen
e-mail: willem.esterhuizen@mines-paristech.fr

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_5

review their extension to mixed constraints (see [7]) and show, on the double integrator example, how mixed constraints may modify the previously presented behavior. Then another simple example of a spring system is presented in two versions with different mixed constraints and again, we compare their consequences on the respective solutions.

5.2 Recalls on Pure State and Input Constrained Systems

The material of this section is a summary of [6]. We consider the constrained nonlinear system

$$\dot{x} = f(x, u), \tag{5.1}$$

$$x(t_0) = x_0, \tag{5.2}$$

$$u \in \mathcal{U}, \tag{5.3}$$

$$g_i\big(x(t)\big) \leq 0 \quad \forall t \in [t_0, \infty), \quad \forall i \in \{1, \dots, p\} \tag{5.4}$$

where $x(t) \in \mathbb{R}^n$. $\mathcal{U}$ is the set of Lebesgue measurable functions from $[t_0, \infty)$ to U, where U is a compact convex subset of $\mathbb{R}^m$, and not a singleton.

The *constraint set* is defined by

$$G \triangleq \{x \in \mathbb{R}^n \,:\, g_i(x) \leq 0, i = 1, \dots, p\}$$

The notation $g(x) \mathring{=} 0$ indicates that there exists an $i \in \{1, \dots, p\}$ such that x satisfies $g_i(x) = 0$ and $g_j(x) \leq 0$ for all $j \in \{1, \dots, p\}$, and $\mathbb{I}(x)$ denotes the set of all indices $i \in \{1, \dots, p\}$ such that $g_i(x) = 0$. Also, $g(x) \prec 0$ (resp. $g(x) \preceq 0$) indicates that $g_i(x) < 0$ (resp. $g_i(x) \leq 0$) for all $i \in \{1, \dots, p\}$.

The sets

$$G_0 \triangleq \{x \in \mathbb{R}^n \,:\, g(x) \mathring{=} 0\}, \qquad G_- \triangleq \{x \in \mathbb{R}^n \,:\, g(x) \prec 0\}. \tag{5.5}$$

are indeed such that $G = G_0 \cup G_-$.

We further assume (see [6])

(A1) f is at least C^2 on $\mathbb{R}^n \times \tilde{U}$ where $\tilde{U}$ in an open subset of $\mathbb{R}^m$, $U \subset \tilde{U}$.
(A2) There exists a positive and finite constant C such that

$$\sup_{u \in U} |x^T f(x, u)| \leq C(1 + \|x\|^2), \quad \text{for all } x$$

(A3) The set $f(x, U)$, called the *vectogram* in [10], is convex for all $x \in \mathbb{R}^n$.
(A4) For each $i = 1, \dots, p$, g_i is an at least C^2 function from $\mathbb{R}^n$ to $\mathbb{R}$,
(A5) the set of points given by $g_i(x) = 0$ defines an $n - 1$ dimensional manifold.

In the sequel we will denote by $x^{(u,x_0)}$ the solution of the differential equation (5.1) with input $u \in \mathcal{U}$ and initial condition x_0, and by $x^{(u,x_0)}(t)$ its solution at time t. We also use the notation x^u and $x^u(t)$ when the initial condition is unambiguous or unimportant.

5.2.1 *The Admissible Set*

Following [6], we define:

Definition 5.1 (*Admissible Set*) We say that the point $\bar{x} \in G$ is *admissible* if, and only if, there exists at least one input function $v \in \mathcal{U}$, such that (5.1)–(5.4) are satisfied for $x_0 = \bar{x}$ and $u = v$. The set of all such $\bar{x}$ is called the *admissible set*:

$$\mathcal{A} \triangleq \{\bar{x} \in G : \exists u \in \mathcal{U},\ g\big(x^{(u,\bar{x})}(t)\big) \leq 0, \forall t \in [t_0, \infty)\}. \tag{5.6}$$

Clearly, if $\bar{x}$ is admissible, any point of the integral curve, $x^{(v,\bar{x})}(t_1)$, $t_1 \in [t_0, \infty)$, with $v \in \mathcal{U}$ as in the above definition, is also an admissible point.

We now recall from [6] the following results:

Proposition 5.1 *Assume that (A1)–(A4) are valid. The set $\mathcal{A}$ is closed.*

Denote by $\partial\mathcal{A}$ the boundary of the admissible set and define

$$[\partial\mathcal{A}]_0 = \partial\mathcal{A} \cap G_0, \quad [\partial\mathcal{A}]_- = \partial\mathcal{A} \cap G_-. \tag{5.7}$$

We indeed have $\partial\mathcal{A} = [\partial\mathcal{A}]_0 \cup [\partial\mathcal{A}]_-$.

5.2.2 *The Barrier*

We next consider the subset $[\partial\mathcal{A}]_-$ of the boundary of the admissible set.

Definition 5.2 The set $[\partial\mathcal{A}]_-$ is called the *barrier* of the set $\mathcal{A}$.

Still following [6], $[\partial\mathcal{A}]_-$ is "fibered" by arcs of integral curves:

Proposition 5.2 *Assume that (A1)–(A4) hold. The barrier $[\partial\mathcal{A}]_-$ is made of points $\bar{x} \in G_-$ for which there exists $\bar{u} \in \mathcal{U}$ and an arc of integral curve $x^{(\bar{u},\bar{x})}$ entirely contained in $[\partial\mathcal{A}]_-$ until it intersects G_0 at a point $x^{(\bar{u},\bar{x})}(\bar{t})$ for some $\bar{t} \in [t_0, +\infty)$.*

Corollary 5.1 (Semi-permeability) *From any point on the boundary $[\partial\mathcal{A}]_-$, there cannot exist a trajectory penetrating the interior of $\mathcal{A}$, denoted by* $\mathsf{int}(\mathcal{A})$, *before leaving G_-.*

The intersection of $\mathsf{cl}([\partial\mathcal{A}]_-)$, the closure of $[\partial\mathcal{A}]_-$, with G_0 is remarkable:

Proposition 5.3 (Ultimate Tangentiality Condition [6]) *Assume that (A1)–(A5) hold and consider $\bar{x} \in [\partial\mathcal{A}]_-$ and $\bar{u} \in \mathcal{U}$ as in Proposition 5.2, i.e., such that $x^{(\bar{u},\bar{x})}(t) \in [\partial\mathcal{A}]_-$ for all t in some time interval until it reaches G_0. Then, there exists a point $z = x^{(\bar{u},\bar{x})}(\bar{t}) \in \mathsf{cl}([\partial\mathcal{A}]_-) \cap G_0$ for some finite time $\bar{t} \geq t_0$ such that*

$$\min_{u\in U}\max_{i\in\mathbb{I}(z)} L_f g_i(z,u) = 0. \tag{5.8}$$

where $L_f g_i(x,u) \triangleq Dg_i(x).f(x,u)$ is the Lie derivative of g_i along the vector field $f(\cdot,u)$ at the point x.

Let $H(x,\lambda,u) = \lambda^T f(x,u)$ denote the Hamiltonian.

Theorem 5.1 (Minimum-like principle [6]) *Under the assumptions of Proposition 5.3, every integral curve $x^{\bar{u}}$ on $[\partial\mathcal{A}]_- \cap \mathsf{cl}(\mathsf{int}(\mathcal{A}))$ and the corresponding control function $\bar{u}$, as in Proposition 5.2, satisfies the following necessary condition.*

There exists a (nonzero) absolutely continuous maximal solution $\lambda^{\bar{u}}$ to the adjoint equation

$$\dot{\lambda}^{\bar{u}}(t) = -\left(\frac{\partial f}{\partial x}(x^{\bar{u}}(t),\bar{u}(t))\right)^T \lambda^{\bar{u}}(t), \quad \lambda^{\bar{u}}(\bar{t}) = \left(Dg_{i^*}(z)\right)^T \tag{5.9}$$

such that the Hamiltonian is minimized

$$\min_{u\in U}\left\{(\lambda^{\bar{u}}(t))^T f(x^{\bar{u}}(t),u)\right\} = (\lambda^{\bar{u}}(t))^T f(x^{\bar{u}}(t),\bar{u}(t)) = 0 \tag{5.10}$$

at every Lebesgue point t of $\bar{u}$ (i.e., for almost all $t \leq \bar{t}$).

In (5.9), $\bar{t}$ denotes the time at which z is reached, i.e., $x^{\bar{u}}(\bar{t}) = z$, with $z \in G_0$ satisfying the ultimate tangentiality condition

$$g_i(z) = 0, \quad i \in \mathbb{I}(z), \quad \min_{u\in U}\max_{i\in\mathbb{I}(z)} L_f g_i(z,u) \triangleq L_f g_{i^*}(z,\bar{u}(\bar{t})) = 0. \tag{5.11}$$

We illustrate this result by the next particularly simple example (double integrator).

5.2.3 Double Integrator, Pure State Constraint

Let us consider the double integrator subjected to a pure state constraint

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad |u| \leq 1, \quad x_1 - 1 \leq 0 \tag{5.12}$$

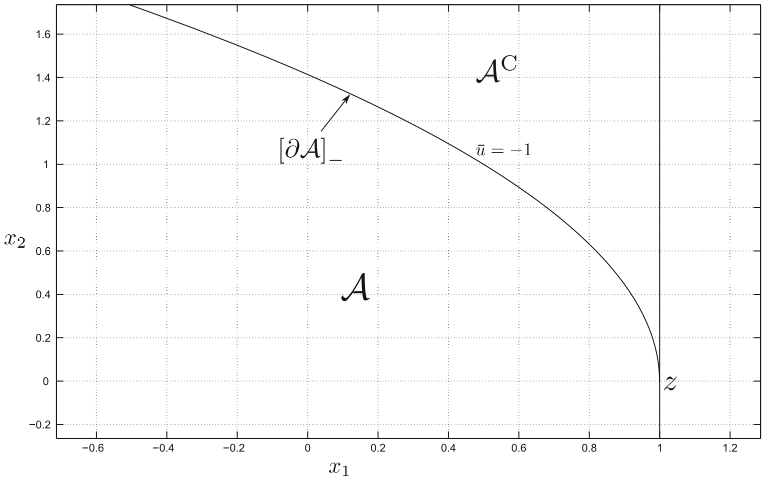


Fig. 5.1 Admissible set and barrier for system (5.12)

The ultimate tangentiality condition reads $\min_{|u|\leq 1} Dg(z).f(z,u) = z_2 = 0$ with $z \triangleq (z_1, z_2) = (x_1^{\bar{u}}(\bar{t}), x_2^{\bar{u}}(\bar{t})) = (1,0)$, $\bar{t}$ indicating the time of tangential arrival on G_0 and $\bar{u}$ denoting the control associated to the barrier trajectory. The costate satisfies

$$\dot{\lambda} = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} \lambda, \quad \lambda^{\bar{u}}(\bar{t}) = (1,0)$$

and we deduce $\lambda_1^{\bar{u}}(t) \equiv 1$ and $\lambda_2^{\bar{u}}(t) = -t + \bar{t} > 0$ for all $t \in (-\infty, \bar{t}]$. We find that the control is given by $\bar{u}(t) = -\mathsf{sign}(\lambda_2(t)) \equiv -1$. Integrating backwards from z with $\bar{u}$ gives the parabola-shaped barrier in Fig. 5.1.

5.3 Dynamical Control Systems with Mixed Constraints

The material of this section is borrowed from [7]. We now consider the following constrained nonlinear system:

$$\dot{x} = f(x,u), \tag{5.13}$$

$$x(t_0) = x_0, \tag{5.14}$$

$$u \in \mathcal{U}, \tag{5.15}$$

$$g_i\big(x(t), u(t)\big) \leq 0 \quad \text{for } a.e.\ t \in [t_0, \infty) \quad i = 1, \dots, p \tag{5.16}$$

where $x(t) \in \mathbb{R}^n$.

As before, $\mathcal{U}$ is the set of Lebesgue measurable functions from $[t_0, \infty)$ to U, with U a given compact convex subset of $\mathbb{R}^m$, expressible as

$$U \triangleq \{u \in \mathbb{R}^m \, : \, \gamma_j(u) \leq 0, j = 1, \dots, r\}$$

with $r \geq m$, the functions γ_j being convex and of class C^2.

Let us stress that the constraints (5.16), called *mixed constraints* [3, 9], depend both on the state and the control. We denote by $g(x, u)$ the vector-valued function whose i-th component is $g_i(x, u)$. As before, by $g(x, u) \prec 0$ (resp. $g(x, u) \preceq 0$) we mean $g_i(x, u) < 0$ (resp. $g_i(x, u) \leq 0$) for all i and by $g(x, u) \overset{\circ}{=} 0$, we mean $g_i(x, u) = 0$ for at least one i.

We define the following sets:

$$G \triangleq \bigcup_{u \in U} \{x \in \mathbb{R}^n \, : \, g(x, u) \preceq 0\} \tag{5.17}$$

$$G_0 \triangleq \{x \in G \, : \, \min_{u \in U} \max_{i \in \{1, \dots, p\}} g_i(x, u) = 0\} \tag{5.18}$$

$$G_- \triangleq \bigcup_{u \in U} \{x \in \mathbb{R}^n \, : \, g(x, u) \prec 0\} \tag{5.19}$$

$$U(x) \triangleq \{u \in U \, : \, g(x, u) \preceq 0\} \quad \forall x \in G. \tag{5.20}$$

Given a pair $(x, u) \in \mathbb{R}^n \times U$, we denote by $\mathbb{I}(x, u)$ the set of indices, possibly empty, corresponding to the "active" mixed constraints, namely

$$\mathbb{I}(x, u) = \{i_1, \dots, i_{s_1}\} \triangleq \{i \in \{1, \dots, p\} \, : \, g_i(x, u) = 0\}$$

and by $\mathbb{J}(u)$ the set of indices, possibly empty, corresponding to the "active" input constraints:

$$\mathbb{J}(u) = \{j_1, \dots, j_{s_2}\} \triangleq \{j \in \{1, \dots r\} \, : \, \gamma_j(u) = 0\}.$$

The integer $s_1 \triangleq \#(\mathbb{I}(x, u)) \leq p$ (resp. $s_2 \triangleq \#(\mathbb{J}(u)) \leq r$) is the number of elements of $\mathbb{I}(x, u)$ (resp. of $\mathbb{J}(u)$). Thus, $s_1 + s_2$ represents the number of "active" constraints, among the $p + r$ constraints, at (x, u).

In addition to (A1)–(A4) of the previous section, we assume

(A6) For all $i = 1, \dots, p$, the mapping $u \mapsto g_i(x, u)$ is convex for all $x \in \mathbb{R}^n$.

(A7) The (row) vectors

$$\left\{ \frac{\partial g_i}{\partial u}(x, u), \frac{\partial \gamma_j}{\partial u}(u) \, : \, i \in \mathbb{I}(x, u), j \in \mathbb{J}(u) \right\} \tag{5.21}$$

are linearly independent at every $(x, u) \in \mathbb{R}^n \times U$ for which $\mathbb{I}(x, u)$ or $\mathbb{J}(u)$ is non empty.[1] We say, in this case, that the point x is *regular* with respect to u (see e.g., [9, 15]).

Given $u \in \mathcal{U}$, we say that an integral curve x^u of Eq. (5.13) defined on $[t_0, T]$ is *regular* if, and only if, at each *Lebesgue* point, or shortly L-point, t of u, $x^u(t)$ is

[1] Note that this implies that $s_1 + s_2 \leq m$, with $s_1 = \#(\mathbb{I}(x, u))$ and $s_2 = \#(\mathbb{J}(u))$.

regular in the aforementioned sense w.r.t. $u(t)$, and, if t is a point of discontinuity of u, $x^u(t)$ is regular in the aforementioned sense w.r.t. $u(t_-)$ and $u(t_+)$, with $u(t_-) \triangleq \lim_{\tau \nearrow t, t \notin I_0} u(\tau)$ and $u(t_+) \triangleq \lim_{\tau \searrow t, t \notin I_0} u(\tau)$, I_0 being a suitable zero measure set of $\mathbb{R}$.

Since system (5.13) is time-invariant, the initial time t_0 may be taken as 0. When clear from the context, "$\forall t$" or "for *a.e* t" will mean "$\forall t \in [0, \infty)$" or "for *a.e.* $t \in [0, \infty)$". Note that throughout this paper *a.e.* is understood with respect to the Lebesgue measure.

5.3.1 The Admissible Set in the Mixed Case: Topological Properties

Definition 5.1 (*Admissible States, Mixed Case*) We say that the point $\bar{x} \in G$ is *admissible* if, and only if, there exists $v \in \mathcal{U}$, such that (5.13)–(5.16) are satisfied for $x_0 = \bar{x}$ and $u = v$:

$$\mathcal{A} \triangleq \{\bar{x} \in G : \exists u \in \mathcal{U},\ g\big(x^{(u,\bar{x})}(t), u(t)\big) \preceq 0, \text{for } a.e.\ t\}. \tag{5.22}$$

As before, any point of the integral curve, $x^{(v,\bar{x})}(t')$, $t' \in [0, \infty)$, is also an admissible point.

We assume that both $\mathcal{A}$ and $\mathcal{A}^{\mathsf{C}}$ contain at least one element to discard the trivial cases $\mathcal{A} = \emptyset$ and $\mathcal{A}^{\mathsf{C}} = \emptyset$.

We use the notations $\mathsf{int}(S)$ (resp. $\mathsf{cl}(S)$) (resp. $\mathsf{co}(S)$) for the interior (resp. the closure) (resp. the closed and convex hull) of a set S.

Proposition 5.4 *Assume that (A1)–(A5) are valid. The set $\mathcal{A}$ is closed.*

5.3.2 Boundary of the Admissible Set (Mixed Case)

5.3.2.1 Geometric Description of the Barrier

As before, we define the barrier as $[\partial\mathcal{A}]_- = \partial\mathcal{A} \cap G_-$.

Proposition 5.5 *Assume (A1)–(A4) and (A6) hold. $[\partial\mathcal{A}]_-$ is made of points $\bar{x} \in G_-$ for which there exists $\bar{u} \in \mathcal{U}$ and an integral curve $x^{(\bar{u},\bar{x})}$ entirely contained in $[\partial\mathcal{A}]_-$ until it intersects G_0, i.e., at a point $z = x^{(\bar{u},\bar{x})}(\tilde{t})$, for some $\tilde{t}$, such that* $\min_{u \in U} \max_{i=1,\ldots,p} g_i(z, u) = 0$.

The "fibered" nature of the barrier thus extends to the mixed case. Note however that G_0 is now modified: it is not defined as the set of x for which there exists $u \in U$ such that $g(x, u) \overset{\circ}{=} 0$ but is given by (5.18). Note that $\tilde{t}$ may be infinite, in which case the barrier does not intersect G_0 as shown in the next double integrator with mixed constraint example.

Corollary 5.2 (Semi-permeability) *Assume (A1)–(A4) and (A6) hold. Then from any point on the boundary $[\partial\mathcal{A}]_-$, there cannot exist a trajectory penetrating the interior of $\mathcal{A}$ before leaving G_-.*

5.3.2.2 Ultimate Tangentiality

We now characterize the intersection of $[\partial\mathcal{A}]_-$ with G_0 at the point z defined in Proposition 5.5. We define

$$\tilde{g}(x) \triangleq \min_{u\in U} \max_{i\in\{1,\ldots,p\}} g_i(x,u). \tag{5.23}$$

Comparing to (5.18) we readily see that $G_0 = \{x \in G : \tilde{g}(x) = 0\}$. According to a result of Danskin [5], $\tilde{g}$ is locally Lipschitz and thus absolutely continuous and almost everywhere differentiable, on every open and bounded subset of $\mathbb{R}^n$.

We now recall basic notions from nonsmooth analysis [4] that are used in the next proposition. Consider $h : \mathbb{R}^n \to \mathbb{R}$ Lipschitz near a given point $x \in \mathbb{R}^n$. The *generalized directional derivative* of h at x in the direction v is defined as follows:

$$h^0(x;v) \triangleq \limsup_{y\to x,\ t\to 0^+} \frac{h(y+tv)-h(y)}{t}. \tag{5.24}$$

We also need to introduce the *generalized gradient* of h at x, labeled $\partial h(x)$. It is well-known that the generalized gradient of a locally Lipschitz function $h : \mathbb{R}^n \to \mathbb{R}$ is the compact and convex set

$$\partial h(x) = \mathsf{co}\{\lim_{i\to\infty} Dh(x_i) : x_i \to x, x_i \notin \Omega_1 \cup \Omega_2\} \tag{5.25}$$

where $Dh(x)$ denotes the row vector $Dh(x)$ at x, Ω_1 is a zero measure set where h is nondifferentiable and Ω_2 is an arbitrary zero measure set.

The relationship between the generalized directional derivative and the generalized gradient is given by

$$h^0(x;v) = \max_{\xi\in\partial h(x)} \xi v. \tag{5.26}$$

Proposition 5.6 (Ultimate Generalized Tangentiality Condition [7]) *Assume (A1)–(A4) and (A6)–(A7) hold. Consider $\bar{x} \in [\partial\mathcal{A}]_-$ and $\bar{u} \in \mathcal{U}$ as in Proposition 5.5, i.e., such that the integral curve $x^{(\bar{u},\bar{x})}(t)$ remains in $[\partial\mathcal{A}]_-$ for all t in some time interval until it reaches G_0 at some finite time $\bar{t} \geq 0$. Then, the point $z = x^{(\bar{u},\bar{x})}(\bar{t}) \in \mathsf{cl}([\partial\mathcal{A}]_-) \cap G_0$, satisfies*

$$0 = \max_{\xi\in\partial\tilde{g}(z)} \xi f(z,\bar{u}(\bar{t})) = \min_{v\in U(z)} \max_{\xi\in\partial\tilde{g}(z)} \xi f(z,v) = \max_{\xi\in\partial\tilde{g}(z)} \min_{v\in U(z)} \xi f(z,v). \tag{5.27}$$

Moreover, if the function $\tilde{g}$ is differentiable at z, then (5.27) *reduces to*

$$0 = L_f \tilde{g}(z, \bar{u}(\bar{t})) = \min_{u \in U(z)} L_f \tilde{g}(z, u). \tag{5.28}$$

Remark 5.1 Note that (5.28) significantly differs from (5.8) on several aspects: in (5.28), $U(z)$ replaces U, where z is such that $\tilde{g}(z) = 0$; moreover, in (5.28), if g_i effectively depends on u for $i \in \mathbb{I}(z)$, $L_f \tilde{g}(z, u)$ is not generally equal to $\max_{i \in \mathbb{I}(z)} L_f g_i(z, u)$.

5.3.3 The Barrier Equation (Mixed Case)

The next necessary conditions are essential to construct the integral curves running along the barrier.

Theorem 5.2 (Minimum-like Principle (Mixed Case) [7]) *Under the assumptions of Proposition 5.6, consider an integral curve $x^{\bar{u}}$ on $[\partial \mathcal{A}]_- \cap \mathsf{cl}(\mathsf{int}(\mathcal{A}))$ and assume that the control function $\bar{u}$ is piecewise continuous. Then $\bar{u}$ and $x^{\bar{u}}$ satisfy the following necessary conditions.*

There exists a nonzero absolutely continuous adjoint $\lambda^{\bar{u}}$ and piecewise continuous multipliers $\mu_i^{\bar{u}} \geq 0$, $i = 1, \ldots, p$, such that

$$\dot{\lambda}^{\bar{u}}(t) = -\left(\frac{\partial f}{\partial x}(x^{\bar{u}}(t), \bar{u}(t))\right)^T \lambda^{\bar{u}}(t) - \sum_{i=1}^{p} \mu_i^{\bar{u}}(t) \frac{\partial g_i}{\partial x}(x^{\bar{u}}(t), \bar{u}(t)) \tag{5.29}$$

with the "complementary slackness condition"

$$\mu_i^{\bar{u}}(t) g_i(x^{\bar{u}}(t), \bar{u}(t)) = 0, \quad i = 1, \ldots, p \tag{5.30}$$

and final conditions

$$\lambda^{\bar{u}}(\bar{t})^T \in \arg \max_{\xi \in \partial \tilde{g}(z)} \xi . f(z, \bar{u}(\bar{t})) \tag{5.31}$$

where $z = x^{\bar{u}}(\bar{t})$ with $\bar{t}$ such that $z \in G_0$, i.e., $\min_{u \in U} \max_{i=1,\ldots,p} g_i(z, u) = 0$, $\partial \tilde{g}(z)$ being the generalized gradient of $\tilde{g}$, defined by (5.23), at z.

Moreover, at almost every t, the Hamiltonian

$$H(\lambda^{\bar{u}}(t), x^{\bar{u}}(t), u) = \left(\lambda^{\bar{u}}(t)\right)^T f(x^{\bar{u}}(t), u)$$

is minimized over the set $U(x^{\bar{u}}(t))$ and equals zero

$$\begin{aligned} \min_{u \in U(x^{\bar{u}}(t))} \lambda^{\bar{u}}(t)^T f(x^{\bar{u}}(t), u) &= \min_{u \in U} \left[\left(\lambda^{\bar{u}}(t)\right)^T f(x^{\bar{u}}(t), u) + \sum_{i=1}^{p} \mu_i^{\bar{u}}(t) g_i(x^{\bar{u}}(t), u) \right] \\ &= \lambda^{\bar{u}}(t)^T f(x^{\bar{u}}(t), \bar{u}(t)) = 0 \end{aligned} \tag{5.32}$$

Remark 5.2 If $\tilde{g}$ is differentiable at the point z, condition (5.31) indeed reduces to its smooth counterpart, i.e., $\lambda^{\bar{u}}(\bar{t})^T = D\tilde{g}(z)$

Remark 5.3 The assumption that $x^{(\bar{u},\bar{x})} \in [\partial\mathcal{A}]_- \cap \mathsf{cl}(\mathsf{int}(\mathcal{A}))$ means that we possibly miss isolated trajectories which are in $\mathcal{A} \setminus \mathsf{cl}(\mathsf{int}(\mathcal{A}))$. The existence and computation of such trajectories, if they exist, are open questions.

5.4 Examples

5.4.1 Double Integrator, Mixed Constraint

Let us go back to the double integrator introduced in Sect. 5.2.3, the pure state constraint $x_1 \leq 1$ being now replaced by the mixed constraint $x_1 \leq u$

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad |u| \leq 1, \quad x_1 - u \leq 0 \tag{5.33}$$

We will show that this apparently innocuous change dramatically modifies the admissible set and its barrier since, in the mixed case, the latter does not intersect G_0 anymore (compare Figs. 5.1 and 5.2).

We readily get $\tilde{g}(x) = x_1 - 1$ and $G_0 = \{(x_1, x_2) : x_1 = 1\}$. The ultimate tangentiality condition reads $\min_{u\in U(z)} D\tilde{g}(z).f(z,u) = z_2 = 0$, or $z \triangleq (z_1, z_2) = (1, 0)$. Note that, at this point, $U(z) = \{1\}$ is reduced to a single element. The minimal Hamiltonian is given by

$$\min_{u\in U(x)} \lambda_1 x_2 + \lambda_2 u = 0, \quad \text{a.e. } t.$$

Thus:

$$\begin{aligned} &\text{if } \lambda_2(t) < 0, \ \ \bar{u}(t) = 1 \text{ if } x_1 \in]\infty, 1] \\ &\text{if } \lambda_2(t) > 0, \ \ \bar{u}(t) = \begin{cases} x_1 & \text{if } x_1 \in [-1, 1] \\ -1 & \text{if } x_1 \in]-\infty, -1[\end{cases} \\ &\text{if } \lambda_2(t) = 0, \ \ \bar{u}(t) = \text{arbitrary}. \end{aligned}$$

The costate equations are given by

$$\dot{\lambda} = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} \lambda$$

with $\lambda(\bar{t}) = D\tilde{g}(z) = (1, 0)^T$. From here we deduce that $\lambda_2(t) = -t + \bar{t}$ for all $t \in (-\infty, \bar{t}]$, and thus $\lambda_2(t) > 0$ for all $t \in (-\infty, \bar{t}]$. Integrating backwards from the point $z = (1, 0)$, we find that the integral curve immediately leaves G, and so this curve cannot be part of the barrier.

However, let us show that the barrier indeed exists and that it remains in G_- for all time.

When the control $\hat{u}(t) = x_1(t)$ is applied to (5.33), the analytic solution initiating at $t = 0$ from $x_0 \triangleq (x_{1,0}, x_{2,0})$ is given by

$$x_1^{(\hat{u},x_0)}(t) = \frac{x_{1,0} + x_{2,0}}{2} e^t + \frac{x_{1,0} - x_{2,0}}{2} e^{-t}$$
$$x_2^{(\hat{u},x_0)}(t) = \frac{x_{1,0} + x_{2,0}}{2} e^t - \frac{x_{1,0} - x_{2,0}}{2} e^{-t}.$$

It is thus immediately seen that, with this control, the origin is a saddle point, the line $x_1 + x_2 = 0$ being the associated stable manifold, and $x_1 - x_2 = 0$ the unstable one.

We now prove that the line segment $\mathcal{L} \triangleq \{(x_1, x_2) : x_1 + x_2 = 0, -1 \leq x_1 < 1\}$ is a subset of $[\partial\mathcal{A}]_-$.

Clearly, $\mathcal{L}$ is positively invariant and every integral curve starting on it asymptotically approaches the origin. Moreover, $g(x^{(\hat{u},x_0)}(t), \hat{u}(t)) = x_1^{(\hat{u},x_0)}(t) - \hat{u}(t) = 0$ for all t such that $-1 \leq x_1^{(\hat{u},x_0)}(t) < 1$. Let $h(x) \triangleq x_1 + x_2$ and denote $x_i(t) \triangleq x_i^{(\hat{u},x_0)}(t)$, $i = 1, 2$, for simplicity's sake. If at a suitable time t_1, the state satisfies $x(t_1) \in \mathcal{L}$, i.e. with $h(x(t_1)) = 0$, using any other admissible control $v > \hat{u}(t_1) = x_1(t_1)$, with $|v| \leq 1$, we get

$$Dh(x(t_1)).f(x(t_1), v) = x_2(t_1) + v > -x_1(t_1) + x_1(t_1) = 0.$$

Therefore, any other admissible control results in the state entering the set $\mathcal{B} \triangleq \{h(x) = x_1 + x_2 > 0\}$. Moreover, in $\mathcal{B}$, all trajectories are such that h is nondecreasing for all admissible control v: $L_f h(x, v) = x_2 + v > -x_1 + x_1 = 0$ as long as $x_1 \leq 1$, which implies that all trajectories starting from $\mathcal{B}$ cross the constraint $x_1 = 1$ and hence are not admissible, i.e., $\mathcal{B} \subset \mathcal{A}^{\mathsf{C}}$. Moreover, starting from any point in the complement, i.e., such that $x_1 + x_2 \leq 0$, denoted by C in Fig. 5.2, it is straightforward to verify that $\hat{u}$ ensures that the corresponding integral curve remains in G for all time which proves the assertion that $\mathcal{L}$ is a subset of $[\partial\mathcal{A}]_-$.

We now prove that the barrier extends, for $x_2 > 1$, by the integral curve starting backwards from the point $(x_1, x_2) = (-1, 1)$, with the control $\bar{u}(t) \equiv -1$ for all $t \in]-\infty, \bar{t}]$.

By Theorem 5.2, assuming that $\bar{u}$ is piecewise continuous, any trajectory running along the barrier, generated by $\bar{u}$, satisfies Eqs. (5.29), (5.30) and (5.32) with absolutely continuous costate $\lambda^{\bar{u}}$ and piecewise continuous multipliers $\mu^{\bar{u}}$.

Consider the end point of $\mathcal{L}$, denoted by ξ, of coordinates $\xi_1 = -1, \xi_2 = 1$. The set $U(\xi)$ at that point is equal to $[-1, 1]$. By (5.32) we must have

$$\min_{u \in U(\xi)} \lambda^{\bar{u}}(t)^T f(\xi, u) = \min_{u \in [-1,1]} \lambda_1^{\bar{u}}(t) + \lambda_2^{\bar{u}}(t) u = 0$$

and, by continuity of the Hamiltonian on $\mathcal{L}$, since we had $\bar{u} = x_1$, considering the limit of the Hamiltonian for $x \to \xi$, $x \in \mathcal{L}$, we deduce that the costate $(\lambda_1^{\bar{u}}(t), \lambda_2^{\bar{u}}(t))^T$

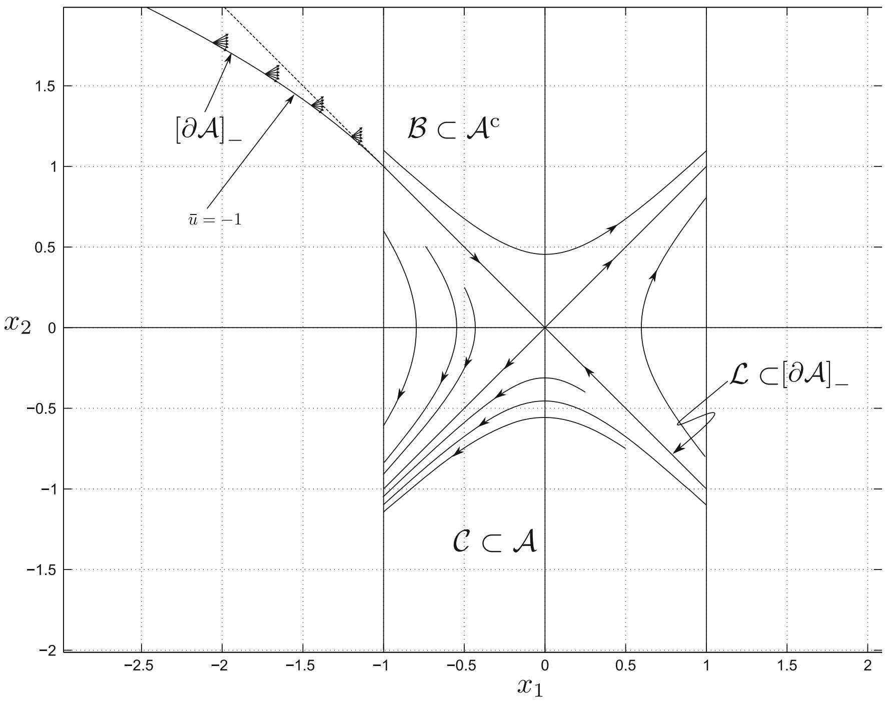


Fig. 5.2 Figure showing some of the sets referred to in Sect. 5.4.1, along with a curve obtained by backward integration from the point $(-1, 1)$ which we have shown to be the backward extension of the barrier

is orthogonal to the vector $(1, -1)^T$, i.e., $\lambda(\bar{t}) = k(1, 1)^T$, with k a positive constant, and the minimizing $\bar{u}$ is thus $\bar{u}(t) = -\mathsf{sign}(\lambda_2(t)) = -1$. Therefore, in $[\partial\mathcal{A}]_- \setminus \mathcal{L}$, since $x_1 < -1$, the constraint $x_1 - u$ is nowhere active and $\mu^{\bar{u}} = 0$ by (5.30). Thus the costate equation reads

$$\dot{\lambda} = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} \lambda, \quad \lambda(\bar{t}) = k(1, 1)$$

from which we deduce that $\lambda_1(t) \equiv k$ and $\lambda_2(t) = -k(t - \bar{t}) + k$, $t \in (-\infty, \bar{t}]$ and $\bar{u}(t) = -\mathsf{sign}(\lambda_2(t)) \equiv -1$. Note that this solution indeed satisfies the piecewise continuous assumption of $\bar{u}$ in Theorem 5.2. The barrier is thus further extended backwards as in Fig. 5.2. We have also included a few of the vectograms along the extension of the barrier in order to emphasize that this is indeed an "extremal" trajectory and that as we approach the point $(-1, 1)$, the vectogram points towards the set $\mathcal{B}$, which we have shown to be a subset of $\mathcal{A}^{\mathsf{C}}$.

5.4.2 Constrained Spring I

Consider the following constrained mass–spring–damper model:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \quad |u| \leq 1, \quad x_2 - u \leq 0$$

where x_1 is the mass's displacement. The spring stiffness is here equal to 2 for a mass equal to 1 and the friction coefficient is equal to 2. u is the force applied to the mass.

We identify $g(x,u) = x_2 - u$, $U = [-1, 1]$ and $\tilde{g}(x) = x_2 - 1$. We also identify the following sets: $G = \{x \in \mathbb{R}^2 : x_2 \leq 1\}$, $G_0 = \{x \in G : x_2 = 1\}$ and $U(x) = \{u \in U : x_2 \leq u \leq 1\}$. Note that if $z \triangleq (z_1, z_2) \in G_0$, i.e. $z_2 = 1$, then $U(z)$ is the singleton $U(z) = \{1\}$.

We have $\partial\tilde{g}(z) = \{(0,1)\} = D\tilde{g}(z)$ ($\tilde{g}$ being indeed differentiable everywhere) and the ultimate tangentiality condition reads

$$\min_{u \in U(z)} D\tilde{g}(z) f(z,u) = 0$$

which gives

$$\min_{u \in U(z)} -2z_1 - 2z_2 + u = -2z_1 - 2 + 1 = 0$$

Thus $z = (-\frac{1}{2}, 1)$.

The final costate $\lambda(\bar{t})$, according to (5.31), which here reduces to (5.28), is given by $\lambda^T(\bar{t}) = D\tilde{g}(z) = (0,1)$.

The Hamiltonian being here $H(x, \lambda, u) = \lambda_1 x_2 + \lambda_2(-2x_1 - 2x_2 + u)$, condition (5.32) reads

$$\min_{x_2 \leq u \leq 1} \lambda_1 x_2 + \lambda_2(-2x_1 - 2x_2 + u) = 0 \tag{5.34}$$

which gives the control $\bar{u}$ associated with the barrier

$$\begin{aligned} &\text{if } \lambda_2(t) < 0, \quad \bar{u}(t) = 1 \\ &\text{if } \lambda_2(t) > 0, \quad \bar{u}(t) = \begin{cases} x_2 & \text{if } x_2 \in]-1, 1] \\ -1 & \text{if } x_2 \in]-\infty, -1] \end{cases} \\ &\text{if } \lambda_2(t) = 0, \quad \bar{u}(t) = \text{arbitrary} \end{aligned}$$

We note from condition (5.29) that if the constraint is active (i.e., $g(x,u) = 0$), the costate differential equation is given by

$$\dot{\lambda}^{\bar{u}} = -\frac{\partial f}{\partial x}^T \lambda^{\bar{u}} - \mu^{\bar{u}} \frac{\partial g}{\partial x} = \begin{pmatrix} 0 & 2 \\ -1 & 2 \end{pmatrix} \lambda^{\bar{u}} - \mu^{\bar{u}} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{5.35}$$

and is otherwise (when $g(x,u) < 0$) given by

$$\dot{\lambda}^{\bar{u}} = -\frac{\partial f}{\partial x}^T \lambda^{\bar{u}} = \begin{pmatrix} 0 & 2 \\ -1 & 2 \end{pmatrix} \lambda^{\bar{u}}. \tag{5.36}$$

Recall that $\lambda_2(\bar{t}) > 0$ and $x_2(\bar{t}) = z_2 = 1 > 0$. Therefore, because λ and x are continuous, $\bar{u}(t) = x_2(t)$ over an interval before $\bar{t}$. We can show that $\bar{u}(t) \neq 1$ over this interval: if $x_2 = 1$ and $u = 1$ over an interval before $\bar{t}$, then we get $\dot{x}_2 = -2x_1 - 2 + 1 = 0$, or $x_1 = -\frac{1}{2}$, meaning that x_1 remains constant over the same interval. Thus $\dot{x}_1 = 0$ for all $t \in]\bar{t} - \eta, \bar{t}]$, $\eta > 0$, which contradicts the fact that $\dot{x}_1 = 1$ over $t \in]\bar{t} - \eta, \bar{t}]$.

Therefore, only the constraint g can be active over an interval before $\bar{t}$, and by (5.32), we obtain μ over this interval

$$\frac{\partial H}{\partial u} + \mu \frac{\partial g}{\partial u} = \lambda_2 - \mu = 0$$

thus $\lambda_2 = \mu$ and the adjoint equation (5.35) reads

$$\dot{\lambda} = \begin{pmatrix} 0 & 2 \\ -1 & 1 \end{pmatrix} \lambda, \quad \forall t \in]\bar{t} - \eta, \bar{t}] \tag{5.37}$$

Let us next analyze the switching condition of $\bar{u}$, or more precisely the change of signum of λ_2. We know that, in an interval $]\bar{t} - \eta, \bar{t}]$ with $\eta > 0$, we have $\lambda_2 > 0$ and we want to characterize η such that $\lambda_2(t) < 0$ for $t \leq \bar{t} - \eta$ and $\lambda_2(\bar{t} - \eta) = 0$. Note that λ_2 cannot vanish over a nonempty open interval since then, according to (5.36) or (5.37), we would also get $\lambda_1 = 0$ which is impossible since the vector λ cannot vanish. Thus, since λ_2 is locally increasing in a neighborhood of $\bar{t} - \eta$, we must have $\dot{\lambda}_2(\bar{t} - \eta) > 0$, which is equivalent to $\lambda_1(\bar{t} - \eta) < 0$. Thus, expressing (5.34) at time $\bar{t} - \eta$, we get $x_2(\bar{t} - \eta) = 0$ and $\bar{u}(t) = 1$ for $t < \bar{t} - \eta$.

As long as λ_2 remains different from zero we keep $\bar{u} = 1$. As seen on Fig. 5.3, x_2 crosses for a second time the x_1 axis and it can be checked that, at this time, λ_2 also vanishes. Therefore, the last section of the barrier is made of the trajectory generated by $\bar{u} = x_2$ from this time.

Remark 5.4 Note that Assumption (A7) does not hold true at the final point z since there are two active constraints for only one control. However, since this condition is violated only at this point, we may conclude by continuity that condition (5.31) still holds.

5.4.3 Constrained Spring II

Consider the same mass–spring–damper system with the same constants as in the previous example, but with a richer constraint

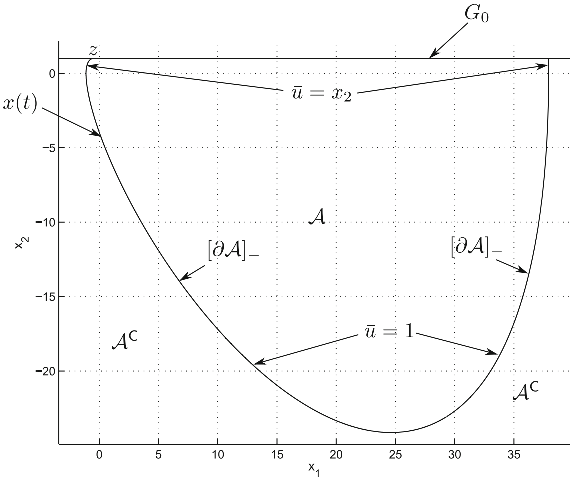


Fig. 5.3 Admissible set of the constrained spring from Sect. 5.4.2

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \quad |u| \leq 1, \quad x_2(x_2 - u) \leq 0 \tag{5.38}$$

We identify $\tilde{g}(x) = x_2^2 - |x_2|$, and $G_0 = \{x : x_2 = 0 \text{ or } x_2 = \pm 1\}$. $\tilde{g}$ is differentiable for $x_2 \neq 0$ and from (5.31) and (5.32) we identify, in same manner as in the previous example, two points of ultimate tangentiality, namely $z = (-\frac{1}{2}, 1)$ along with $\lambda(\bar{t}) = (0, 1)$, and $z = (\frac{1}{2}, -1)$ along with $\lambda(\bar{t}) = (0, -1)$. We defer the treatment of the x_1 axis, which is also in G_0, to the discussion below.

From the minimisation of the Hamiltonian, which is the same as in the previous example except that $U(x)$ now corresponds to $u \geq x_2$ if $x_2 \geq 0$ and $u \leq x_2$ if $x_2 \leq 0$, we find the control $\bar{u}$

$$\begin{aligned}
&\text{if } \lambda_2(t) < 0 \quad \bar{u}(t) = \begin{cases} 1 & \text{if } x_2 \in]0, 1] \\ x_2 & \text{if } x_2 \in]-1, 0[\end{cases} \\
&\text{if } \lambda_2(t) > 0 \quad \bar{u}(t) = \begin{cases} x_2 & \text{if } x_2 \in]0, 1] \\ -1 & \text{if } x_2 \in]-1, 0[\end{cases} \\
&\text{if } \lambda_2(t) = 0 \quad \bar{u}(t) = \text{arbitrary}
\end{aligned}$$

If we now integrate backwards from the points $(-\frac{1}{2}, 1)$ and $(\frac{1}{2}, -1)$ with the control $\bar{u}(t)$ we obtain the barrier as in Fig. 5.4. It turns out that $\bar{u}(t) = x_2(t)$ all along both curves: the reader may easily check that, the necessary condition $\frac{\partial H}{\partial u} + \mu \frac{\partial g}{\partial u} = 0$ yields $\lambda_2 - \mu x_2 = 0$ and, since $\frac{\partial g}{\partial x} = (0, 2x_2 - u)^T$, we get the same adjoint equation as (5.37) when $\bar{u} = x_2$, and conclude that $\lambda_2(t)$ is positive as long as $x_2(t)$ is positive, which implies that $\bar{u} = x_2$, and $\lambda_2(t)$ must be negative as long as $x_2(t)$ is negative, which again implies that $\bar{u} = x_2$, hence the result.

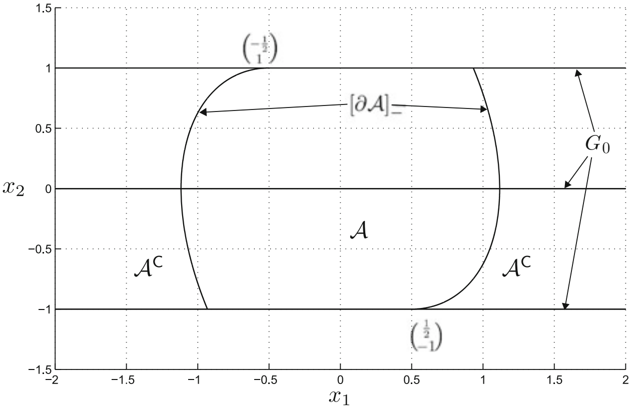


Fig. 5.4 Admissible set of the constrained spring from Sect. 5.4.3

Let us now turn to the x_1 axis, where $\tilde{g} = x_2^2 - |x_2|$ is non-differentiable. For any z on the x_1 axis, we have $U(z) = [-1, 1]$ and $\partial\tilde{g}(z) = \bar{\text{co}}\left((0, -1)^T, (0, 1)^T\right) = \{0\} \times [-1, 1]$ and we must have

$$\min_{u\in[-1,1]} \max_{\xi\in\partial\tilde{g}(\tilde{z})} \xi.f(\tilde{z}, u) = 0 = \min_{u\in[-1,1]} \max_{\xi_2\in[-1,1]} \xi_2(-2x_1 + u) \tag{5.39}$$

For each $-\frac{1}{2} \le z_1 \le \frac{1}{2}$ Eq. (5.39) has a solution given by $\xi = (0, \mathsf{sign}(-2z_1 + u))$ from which we deduce that $\bar{u} = 2z_1$. However, one can directly verify that the integral curves of (5.38) with endpoints in the set $[-\frac{1}{2}, \frac{1}{2}] \times \{0\}$ with the control $u = x_2$ all correspond to admissible curves (integrated backwards) and therefore do not belong to the barrier, but that they make the constraint $g(x^{(\bar{u},\bar{x})}(t), \bar{u}(t))$ equal to 0 for $\bar{u} = x_2$ for all $\bar{x} \in [-\frac{1}{2}, \frac{1}{2}] \times \{0\}$ and for all t. This attests that our conditions are only necessary and far from being sufficient.

Remark 5.5 Note that, as in Sect. 5.4.2, Assumption (A7) does not hold true at the final points $z \in G_0$ since there are two active constraints for only one control. Again, we conclude by a continuity argument that condition (5.31) still holds.

5.5 Conclusion

In this paper, we have demonstrated on elementary examples of systems subject to pure or mixed constraints, the effectiveness of the results obtained in [6, 7], which allowed us to give a complete construction of their barriers and admissible sets. We also pointed out some significant differences in these constructions. In particular, we have shown, in the mixed constrained case, that the barrier does not need to intersect the boundary G_0 of the constraint set; that, according to the feedback nature of the control, due to the state dependence of the control set, the equilibria and their stability could be modified to be repelled from G_0; that the *nonsmooth* version of the necessary ultimate tangentiality condition, though useful, is far from being sufficient; and that Assumption (A7) is, even in simple examples, not everywhere satisfied. Higher dimensional examples are presently under investigation and will be published elsewhere.

References

1. Aubin, J.P.: Viability Theory. Systems and Control Foundations, Birkhäuser (1991)
2. Chutinan, A.C., Krogh, B.H.: Computational techniques for hybrid system verification. IEEE Trans. Autom. Control 64–75 (2003)
3. Clarke, F.H., de Pinho, M.: Optimal control problems with mixed constraints. SIAM J. Control Optim. **48**, 4500–4524 (2010)
4. Clarke, F.H., Ledyaev, Yu.S., Stern, R.J.,Wolenski, P.R.: Nonsmooth Analysis and Control Theory. Springer, New York (1998)
5. Danskin, J.: The Theory of Max-Min. Springer (1967)
6. De Dona, J.A., Lévine, J.: On barriers in state and input constrained nonlinear systems. SIAM J. Control Optim. **51**(4), 3208–3234 (2013)
7. Esterhuizen, W., Lévine, J.: Barriers in nonlinear control systems with mixed constraints (2015). arXiv:1508.01708 [math.OC]
8. Hartl, R.F., Sethi, S.P., Vickson, R.J.: A survey of the maximal principles for optimal control problems with state constraints. SIAM Rev. **37**(2), 181–218 (1995)
9. Hestenes, M.R.: Calculus of Variations and Optimal Control Theory. Wiley (1966)
10. Isaacs, R.: Differential Games. Wiley (1965)
11. Kaynama, S., Maidens, J., Oishi, M., Mitchell, I., Dumont, G.: Computing the viability kernel using maximal reachable sets. In: Proceedings of the 15th ACM HSCC'12, New York, NY, USA, pp. 55–64. ACM (2012)
12. Lhommeau, M., Jaulin, L., Hardouin, L.: Capture basin approximation using interval analysis. Int. J. Adapt. Control Signal Process. **25**(3), 264–272 (2011)
13. Lygeros, J., Tomlin, C., Sastry, S.: Controllers for reachability specifications for hybrid systems. Automatica **35**(3), 349–370 (1999)
14. Mitchell, I.M., Bayen, A.M., Tomlin, C.J.: A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. IEEE Trans. Autom. Control **50**(7), 947–957 (2005)
15. Pontryagin, L., Boltyanskii, V., Gamkrelidze, R., Mishchenko, E.: The Mathematical Theory of Optimal Processes. Wiley (1965)
16. Prajna, S.: Barrier certificates for nonlinear model validation. Automatica **42**(1), 117–126 (2006)
17. Tee, K.P., Ge, S.S., Tay, E.H.: Barrier Lyapunov functions for the control of output-constrained nonlinear systems. Automatica **45**(4), 918–927 (2009)
18. Tomlin, C.J., Lygeros, J., Sastry, S.S.: A game theoretic approach to controller design for hybrid systems. Proc. IEEE **88**(7), 949–970 (2000)
19. Tomlin, C.J., Mitchell, I., Bayen, A.M., Oishi, M.: Computational techniques for the verification of hybrid systems. Proc. IEEE **91**(7), 986–1001 (2003)
20. van der Schaft, A., Schumacher, H.: An Introduction to Hybrid Dynamical Systems. Lecture Notes in Control and Information Sciences, vol. 251. Springer (2000)

Chapter 6
Output Regulation via Low-Power Construction

Daniele Astolfi, Alberto Isidori and Lorenzo Marconi

Abstract The paper deals with the problem of output regulation for the class of nonlinear systems that have a well-defined relative degree and are minimum-phase. The goal is to present a design methodology of the internal model-based regulator that adopts the tools recently proposed in Astolfi and Marconi (IEEE Trans Autom Control 60:3059–3064, 2015) [1] for the design of nonlinear high-gain observers in which the power of the high-gain parameter is raised just up to the order 2 regardless the dimension of the observed system. In this context, we show how to design a high-gain internal model in which the power of the high-gain is raised up to the order two regardless the dimension of the internal model. The same methodology is also used in the dynamic stabilizer in the presence of regulated plants with high relative degree by presenting backstepping and dirty-derivatives observer techniques with limited high-gain power.

6.1 Introduction

High-gain techniques are widely adopted for robust stabilization and observation of nonlinear systems in a global or semiglobal sense. Major contributions to the field have been developed around early 90s in a very florid research period that has definitely seen Laurent Praly as one of the main actors. Among the different contributions it is worth recalling the milestone papers [17, 18] in which very general

D. Astolfi · L. Marconi (✉)
CASY-DEI University of Bologna, Bologna, Italy
e-mail: lorenzo.marconi@unibo.it

D. Astolfi
e-mail: daniele.astolfi@unibo.it

D. Astolfi
MINES ParisTech, PSL Research University, CAS, Paris, France

A. Isidori
Sapienza, Universitá di Roma, Rome, Italy
e-mail: albisidori@dis.uniroma1.it

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_6

tools for robust stabilization of nonlinear systems by state and output feedback are presented. Among the different results presented in the paper, the authors develop a systematic way for stabilising nonlinear systems by output feedback by using a "dirty-derivative" observer, namely a high-gain observer able to practically estimate the time-derivates of the output with an arbitrarily fast estimation error dynamics. High-gain observation tools play, in general, a key role in the problem of stabilization in the large of nonlinear systems by output feedback by separation principle due to the distinguishing feature of quickly recovering the ideal stabilising state-feedback by thus preventing finite escape time. A very general and elegant nonlinear separation principle that relies on high-gain observers can be found in Chap. 6 of [6]. The reader is also linked to the special issue [13] for an updated and reasoned overview on the use of high-gain observers in nonlinear control and observation. One of the main criticisms that is typically raised towards high-gain structures, however, is the fact that the high-gain parameter, by which the speed of convergence is tuned, is powered up to the dimension of the observer. This, in turn, makes high-gain observers very critical in numerical implementation whenever the dimension of the observer is large. In this respect, recently, a new "low-power" design methodology for high-gain observers has been proposed in [1]. The observer structure presented in this paper preserves all the main features of classical high-gain observers in terms of speed of convergence and practical convergence but with the advantage of having the power of the high-gain parameter raised just up to the order 2 regardless the dimension of the observed system, and the (mild) drawback that the dimension of the observer must be increased to the order $2n - 2$ (n being the dimension of the observed system). Besides substantially overtaking the numerical problems that characterize the implementation of classical observers, the new structure has also superior features in terms of sensitivity to measurement noises as detailed in [1]. The use of the new structure in the context of the nonlinear separation principle of [6] can be found in [19].

A research area closely related to the one of stabilization overviewed before is the one of output regulation for nonlinear systems. Compared with a "standard" stabilization problem, the problem of output regulation amounts to making a compact attractor, on which some regulated variables are zero, asymptotically stable. The distinguishing feature of the framework is that the attractor is not invariant for the original uncontrolled plant and has internal dynamics governed by the dynamics of an autonomous exogenous system (the so-called exosystem) whose state is not measurable. This, in turn, asks for the design of regulators that include an appropriate copy of the exosystem dynamics able to make the desired attractor invariant and leads to the celebrated internal model-based design strategy. Starting with the contribution in [10], many improvements have been proposed in the output regulation literature in the last twenty years or so with the aim of making the framework where internal model-based regulators could be systematically designed even more general (see, among others, [7, 15, 16]). To the purpose of the design methodology presented in this chapter, an interesting framework has been proposed in [3] in which the so-called "friend", which is the ideal steady state input able to make the desired attractor invariant, and a certain number of its time derivatives are assumed to fulfill a regression law. The important observation made in [3] is that, in this framework, tools typically adopted in the field of nonlinear high-gain observers can be successfully adopted in order to design internal model-based regulators. This observation opened an interesting research direction in which high-gain tools are adopted not only for the design of the stabilizer but also of the internal model. The same framework has been also taken in [11] in order to design adaptive linear regulators, namely regulators with adaptive mechanisms able to cope with uncertainties in the exosystem. The fact that design methodologies typically used in the design of nonlinear observers could be successfully employed in the design of internal models has been further investigated and developed in [4] in which the theory of adaptive (not necessarily high-gain) nonlinear observers has been proposed for this purpose.

The main criticisms that can be done to the high-gain design methodology for internal models is the same mentioned above in the context of stabilization, namely the fact that the power of the high-gain parameter is raised up to the order of the internal model, with the latter than can be large to eventually have the friend and its time derivatives fulfilling the regression law said before. Motivated by this, in this chapter, we adapt the tools presented in [1] to develop a "low-power" methodology for the design of internal models. The low-power construction is, indeed, used not only for the design of the internal model but also for the design of the high-gain dynamic output feedback stabilizer needed to deal with systems that do not have unitary relative degree. Simulation results are also presented to show the effectiveness of the approach.

6.2 The Framework of Output Regulation by Means of High-Gain Tools

We consider the class of systems in normal form with unitary relative degree described by

$$\begin{aligned}\dot{z} &= f(w, z, e)\\ \dot{e} &= q(w, z, e) + b(w, z, e)u\end{aligned} \tag{6.1}$$

in which $(z, e) \in \mathbb{R}^n \times \mathbb{R}$ is the state, $u \in \mathbb{R}$ is the control input and $w \in \mathbb{R}^\rho$ is a an exogenous variable that, in the context of output regulation, is thought of as generated by an autonomous system (typically referred to as exosystem) of the form

$$\dot{w} \;=\; s(w) \tag{6.2}$$

whose state ranges in a compact *invariant* set $W \subset \mathbb{R}^\rho$. The state component e represents the measured output and the regulation error to be steered to zero. It is assumed that $f(\cdot), q(\cdot), b(\cdot), s(\cdot)$ are smooth enough functions and that the function $b(\cdot)$ is bounded from below, i.e., there exists a strictly positive real numbers $\underline{b}$ such that

$$b(w,z,e) \geq \underline{b} \qquad \forall\,(w,z,e) \in \mathbb{R}^\rho \times \mathbb{R}^n \times \mathbb{R}\,. \tag{6.3}$$

The initial condition of the system (6.1) is assumed to range in an arbitrary but known compact set $Z \times E \subset \mathbb{R}^n \times \mathbb{R}$. Within this framework the problem of output regulation amounts to designing a controller of the form

$$\begin{aligned} \dot{\xi} &= \psi(\xi, e) \\ u &= \gamma(\xi, e) \end{aligned} \tag{6.4}$$

with initial conditions in a compact set Ξ, such that the trajectories of the closed-loop system originating from $W \times Z \times E \times \Xi$ are bounded and

$$\lim_{t\to\infty} e(t) = 0 \tag{6.5}$$

uniformly in the initial conditions. Very often asymptotic regulation is difficult to achieve in a general nonlinear context and it thus makes sense to relax (6.5) into a *practical* regulation objective, namely to ask that $\lim_{t\to\infty} \sup |e(t)| \leq \epsilon$ with ϵ a small positive number.

The previous problem is addressed under a number of assumptions that are customary in the literature of output regulation by means of high-gain tools. The first regards the existence of the solution of the so-called *regulator equations*. In this framework, in particular, we assume the existence of a differentiable function $\pi : \mathbb{R} \to \mathbb{R}^n$ solution of

$$L_s\pi(w) = f(w, \pi(w), 0) \qquad \forall\, w \in W\,.$$

This assumption guarantees that the set $\mathscr{A} \subset W \times \mathbb{R}^n$, defined as

$$\mathscr{A} \;=\; \{(w,z) \in W \times \mathbb{R}^n \;:\; z = \pi(w)\}\,,$$

is invariant for the *zero dynamics* of the system (6.1) with input u and output e that are described by

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w,z,0)\,. \end{aligned} \tag{6.6}$$

The second assumption asks that system (6.1) is also minimum-phase. In our framework the minimum-phaseness assumption is formalized as follows.

Assumption 6.1 (Minimum-phase) *The set $\mathscr{A}$ is asymptotically and locally exponentially stable for the system (6.6) with a domain of attraction of the form $W \times \mathscr{D}$ where $\mathscr{D}$ is an open set of $\mathbb{R}^n$ such that $Z \subset \mathscr{D}$.*

The local exponential stability requirement in the previous assumption is done just for sake of simplicity and it could be easily removed by properly adapting the design of the regulator presented in the following, see, for instance, in [12]. In the design of the regulator solving the problem of output regulation, a crucial role is played by the so-called "friend", which is the function $c : W \to \mathbb{R}$ defined as

$$c(w) \ := \ - \frac{q(w, \pi(w), 0)}{b(w, \pi(w), 0)} \ . \tag{6.7}$$

By bearing in mind (6.1), it turns out that such a function represents the ideal steady-state input needed to keep the regulation error identically zero, namely the control input that must be applied to (6.1) to make the set $\mathscr{A} \times \{0\}$ invariant. In the following construction, we do not assume a specific structure for $c(\cdot)$ as typically done, through the so-called immersion assumption, in most of the work on the subject ([3, 9] and references therein). Rather, the internal model-based regulator designed in the following relies on the knowledge of an integer $d > 0$ and of a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ fulfilling

$$L_s^d c(w) \ = \ \varphi \left(c(w),\, L_s\, c(w),\, \ldots,\, L_s^{d-1}\, c(w) \right) + \nu(w) \qquad \forall\, w \in W \tag{6.8}$$

for some (unknown) function $\nu : W \to \mathbb{R}$. In case the previous relation is fulfilled with $\nu \equiv 0$ asymptotic regulator will be achieved. Practical regulation, with an asymptotic error that is upper bounded by a function of $\sup_{w \in W} \|\nu(w)\|$, is otherwise obtained. The previous framework allows one to regard the parameter d as a degree-of-freedom by which the designer can tradeoff the dimension of the regulator (and thus its complexity) and the bound on the asymptotic error. As a matter of fact, larger values of d allow, in general, to identify a $\varphi(\cdot)$ that makes relation (6.8) fulfilled with a smaller bound of the residual term $|\nu(\cdot)|$, by thus obtaining a regulator able to guarantee smaller asymptotic errors.

In the remaining part of the section, we illustrate the main framework under which a regulator can be designed (see [3]), by highlighting how the theory of nonlinear observers, and in particular the one of high-gain observers, turns out to be useful in the regulator construction. The fact of dealing with regulated plant that are affine in the input suggests to consider regulator structures of the same kind, namely regulators of the form

$$\begin{aligned} \dot{\xi} &= \phi(\xi) + \Psi v \qquad \xi \in \mathbb{R}^m \\ u &= \gamma(\xi) + v \\ v &= -\kappa e \end{aligned} \tag{6.9}$$

where $\phi(\cdot)$ and $\gamma(\cdot)$ are smooth functions, Ψ is a column vector, and κ is a design parameter, all to be designed. The resulting closed-loop system, has a normal form that, having defined the change of variables

$$\xi \ \mapsto \ \chi := \xi - \Psi \int_0^e \frac{1}{b(w, z, s)} ds \, ,$$

reads as

$$\begin{aligned}
\dot{w} &= s(w)\\
\dot{z} &= f(w,z,e)\\
\dot{\chi} &= \phi(\chi) - \Psi\left(\gamma(\chi) + \frac{q(w,z,e)}{b(w,z,e)}\right) + \Delta(w,z,\chi,e)\\
\dot{e} &= q(w,z,e) + b(w,z,e)\gamma(\chi) + b(w,z,e)v + L(w,z,\chi,e)
\end{aligned} \tag{6.10}$$

where $\Delta(\cdot)$ and $L(\cdot)$ are properly defined functions such that $\Delta(w,z,\chi,0) = 0$ and $L(w,z,\chi,0) = 0$ for all $(w,z,\chi) \in W \times \mathbb{R}^n \times \mathbb{R}^m$. This system, regarded as a system with input v and output e, has still unitary relative degree and, as an easy computation shows, zero dynamics described by

$$\begin{aligned}
\dot{w} &= s(w)\\
\dot{z} &= f(w,z,0)\\
\dot{\chi} &= \phi(\chi) - \Psi\left(\gamma(\chi) + \frac{q(w,z,0)}{b(w,z,0)}\right) .
\end{aligned} \tag{6.11}$$

Note that these dynamics have a cascade structure with system (6.6) driving the system with state χ. In the following, we denote by $X \subset \mathbb{R}^m$ the compact set of initial conditions for the new variable χ. The problem of output regulation is then reformulated as a problem of output feedback stabilization of system (6.10). In particular the problem at hand is solved if one is able to prove the existence of a compact set of $\mathbb{R}^\rho \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$, on which the regulation error e is identically zero, that is asymptotically stable for system (6.10) with a domain of attraction containing the set of initial conditions. To this purpose high-gain design paradigms for minimum-phase systems can be successfully adopted ([3]). In particular, the following two requirements play a role in the design of the stabilizer:

(a) there exists a set $\mathscr{B} \subset \mathbb{R}^\rho \times \mathbb{R}^n \times \mathbb{R}^m$ that is asymptotically and locally exponentially stable for system (6.11) with a domain of attraction of the form $W \times \mathscr{D}_e$ with $\mathscr{D}_e \subset \mathbb{R}^n \times \mathbb{R}^m$ an open set fulfilling $Z \times X \subset \mathscr{D}_e$.

(b) the following holds:

$$q(w,z,0) + b(w,z,0)\gamma(\chi) = 0 \quad \forall (w,z,\chi) \in \mathscr{B}\,.$$

Requirement (a), in turn, asks that system (6.10), regarded as a system with input v and output e, is minimum-phase. On the other hand, requirement (b) asks that the coupling term between the zero dynamics (6.11) and the error dynamics is vanishing on $\mathscr{B} \times \{0\}$, namely that the latter set is invariant for (6.10) with $v = 0$. That properties, in turn, make system (6.10) fitting into frameworks of stabilization of minimum-phase nonlinear systems in which the choice $v = -\kappa e$, with κ sufficiently large, succeeds in asymptotically stabilising the set $\mathscr{B} \times \{0\}$. This is formalized in the next theorem whose proof can be found in [12].

Theorem 6.1 *Assume that the requirements* (a) *and* (b) *specified before are fulfilled for some compact set* $\mathscr{B}$. *Then, there exists a* $\kappa^\star > 0$ *such that for all* $\kappa \geq \kappa^\star$ *the* *set $\mathscr{B} \times \{0\}$ is asymptotically and locally exponentially stable for system (6.1)–(6.2) controlled by (6.9) with a domain of attraction of the form $W \times \mathscr{D}_{cl}$ with $\mathscr{D}_{cl} \subset \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$ an open set fulfilling $Z \times X \times E \subset \mathscr{D}_{cl}$.*

The high-gain paradigm is indeed robust in case requirement (a) above is only achieved practically rather than asymptotically. More specifically, requirement (a) above can be relaxed to the requirement (a') specified in the following at the price of achieving just practical instead of asymptotic regulation as claimed in the next Theorem 6.2.

(a') There exists a set $\mathscr{B} \subset \mathbb{R}^\rho \times \mathbb{R}^n \times \mathbb{R}^m$ such that the trajectories of system (6.11) originating from $W \times Z \times X$ fulfill

$$\|(w(t), z(t), \chi(t))\|_{\mathscr{B}} \le \max\{c_1 \exp(-c_2 t)\|(w(0), z(0), \chi(0))\|_{\mathscr{B}}, \ \epsilon\} \tag{6.12}$$

for some positive constants c_1, c_2 and ϵ

Theorem 6.2 *Assume that the requirements* (a') *and* (b) *specified before are fulfilled for some compact set $\mathscr{B}$ and positive constants c_1, c_2 and ϵ. Then, there exist a $\kappa^\star > 0$ and a $c > 0$, such that for all $\kappa \ge \kappa^\star$ the trajectories of the resulting closed-loop (6.1)–(6.2) and (6.9) originating from the compact set of initial conditions $W \times Z \times X \times E$ are bounded and*

$$\lim_{t\to\infty} \sup \|e(t)\| \le \frac{c}{\kappa}\epsilon \, .$$

The previous considerations shift the focus of the design on system (6.11) and, in particular, on the design of the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ fulfilling the requirements (a') and (b). In the next section, the problem in question is solved using the high-gain observer theory.

6.3 Low-Power High-Gain Tools for the Internal-Model Design

The problem of fulfilling requirement (a') and (b) introduced at the end of the previous section is now addressed using design tools that are adopted in the literature of high-gain observers. Our main goal is to show that the "low-power" tools introduced in [1] can be successfully adopted in order to design the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ fulfilling the requirements in question. It is argued that it is known a positive $d > 0$ and a function $\varphi(\cdot)$ fulfilling (6.8) for some (unknown) function $\nu(\cdot)$.

We start by recalling the result presented in [3] in which the standard high-gain tools typically used for observer design are shown to be successful for the regulation purposes. To this end, let the dimension of the regulator (6.9) be taken as $m = d$ and, by bearing in mind the definition in (6.7), let $\tau : W \to \mathbb{R}^d$ be defined as

$$\tau(w) := \left(c(w)\ L_s c(w)\ \ldots\ L_s^{d-1} c(w)\right)^T ,$$

and the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ be taken as

$$\phi(\xi) := \begin{pmatrix} \xi_2 \\ \vdots \\ \xi_{i+1} \\ \vdots \\ \varphi_s(\xi) \end{pmatrix} , \qquad \Psi := \begin{pmatrix} \ell a_1 \\ \vdots \\ \ell^i a_i \\ \vdots \\ \ell^d a_d \end{pmatrix} , \qquad \gamma(\xi) := \xi_1 \tag{6.13}$$

where ℓ is a design parameter, the a_i's are coefficients of an Hurwitz polynomial, and $\varphi_s(\cdot)$ is a bounded function that agrees with $\varphi(\cdot)$ on $\tau(W)$. Then, we have the following proposition whose proof can be obtained by slightly generalizing the arguments in [3].

Proposition 6.1 *Let $c(\cdot)$ in (6.7) be fulfilling (6.8) and let the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ be taken as in (6.13). Then there exist a $\ell^\star > 0$ such that for all $\ell \geq \ell^\star$ requirements* (a') *and* (b) *of Sect. 6.2 are fulfilled with*

$$\mathscr{B} = \{(w, z, \chi) \in W \times \mathbb{R}^n \times \mathbb{R}^d , \quad z = \pi(w) , \ \chi = \tau(w)\}$$

and the ϵ in (6.12) of the form

$$\epsilon = \frac{r}{\ell^d} \sup_{w \in W} \|\nu(w)\|$$

with r a positive number.

By joining the result of Theorem 6.2 and the previous proposition it is then immediately concluded that there exists a $\kappa^\star$ (dependent on ℓ) such that for all $\kappa \geq \kappa^\star$ the regulator (6.9) with $(\phi(\cdot), \Psi, \gamma(\cdot))$ taken as in (6.13) guarantees that the trajectories of the closed-loop systems originating from the given compact sets are bounded and

$$\lim_{t\to\infty} \sup \|e(t)\| \leq \frac{r'}{\kappa \ell^d} \sup_{w \in W} \|\nu(w)\| \tag{6.14}$$

for some positive constant r'. In particular, if the integer d and the function $\varphi(\cdot)$ can be taken so that relation (6.8) is fulfilled with $\nu(\cdot) = 0$, the proposed controller guarantees asymptotic regulation. Otherwise, just practical regulation is achieved with the bound on the asymptotic error that can be arbitrarily decreased by increasing κ or ℓ^d.

The main criticism that can be raised to the previous control structure is that the power of the high-gain parameter ℓ is raised up to the order d in the expression of $\phi(\cdot)$ and Ψ. This, in turn, makes the practical implementation of the regulator very hard if the value of d is large. In order to overtake this problem, we present an high-gain design methodology, recently proposed in [1] for the design of nonlinear observers, in which the power of the high-gain parameter is raised just up to the order 2 regardless the dimension of the regulator, at the price of increasing the dimension of the regulator to the order $2d-2$. More precisely, set

$$m = 2d - 2$$

and, let

$$\phi(\xi) := \begin{pmatrix} \phi_1(\xi) \\ \phi_2(\xi) \\ \vdots \\ \phi_{d-1}(\xi) \end{pmatrix}, \quad \Psi := \begin{pmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_{d-1} \end{pmatrix}, \quad \gamma(\xi) := \xi_{1,1}\,, \tag{6.15}$$

where

$$\xi = \mathrm{col}(\xi_1, \ldots, \xi_{d-1}) \in \mathbb{R}^{2d-2}\,, \quad \xi_i = (\xi_{i,1}, \xi_{i,2})^T \in \mathbb{R}^2\,,$$

the functions $\phi_i : \mathbb{R}^{2d-2} \to \mathbb{R}^2$, $i = 1, \ldots, d-1$, are defined as

$$\phi_1(\xi) := \begin{pmatrix} \xi_{1,2} \\ \xi_{2,2} \end{pmatrix}, \quad \phi_i(\xi) := \begin{pmatrix} \xi_{i,2} + \ell\, a_{i,1}\,(\xi_{i-1,2} - \xi_{i,1}) \\ \xi_{i+1,2} + \ell^2\, a_{i,2}\,(\xi_{i-1,2} - \xi_{i,1}) \end{pmatrix}, \tag{6.16}$$

for $i = 2, \ldots, d-2$,

$$\phi_{d-1}(\xi) := \begin{pmatrix} \xi_{d-1,2} + \ell\, a_{d-1,1}\,(\xi_{d-2,2} - \xi_{d-1,1}) \\ \varphi_s(\Gamma\xi) + \ell^2\, a_{d-1,2}\,(\xi_{d-2,2} - \xi_{d-1,1}) \end{pmatrix} \tag{6.17}$$

in which

$$\Gamma := \mathrm{blkdiag}\left(\begin{pmatrix}1 & 0\end{pmatrix} \begin{pmatrix}1 & 0\end{pmatrix} \cdots \begin{pmatrix}1 & 0\end{pmatrix} \begin{pmatrix}1 & 1\end{pmatrix}\right)\,, \tag{6.18}$$

$(a_{i,1}, a_{i,2})$, $i = 1, \ldots, d-1$, are coefficients to be appropriately chosen, and the vectors Ψ_i, $i = 1, \ldots, d-1$ are defined as

$$\Psi_1 := \begin{pmatrix} \ell\, a_{1,1} \\ \ell^2\, a_{1,2} \end{pmatrix}, \quad \Psi_i := \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad i = 2, \ldots, d-1\,.$$

It will be shown that the previous choice of the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ makes the requirements (a') and (b) fulfilled provided that the coefficients $(a_{i,1}, a_{i,2})$, $i = 1, \ldots, d-1$, are appropriately chosen and ℓ is taken sufficiently large. As far as the design of the coefficients $(a_{i,1}, a_{i,2})$ is concerned, they must be chosen in such a way that the block tri-diagonal matrix M defined as

$$M := \begin{pmatrix} L_1 & N & 0 & \dots & \dots & 0 \\ Q_2 & L_2 & N & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & Q_j & L_j & N & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & Q_{d-2} & L_{d-2} & N \\ 0 & \dots & \dots & \dots & 0 & Q_{d-1} & L_{d-1} \end{pmatrix} \tag{6.19}$$

in which

$$L_i := \begin{pmatrix} -a_{i,1} & 1 \\ -a_{i,2} & 0 \end{pmatrix}, \quad Q_i := \begin{pmatrix} 0 & a_{i,1} \\ 0 & a_{i,2} \end{pmatrix}, \quad i = 1, \dots, d-1, \quad N := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

is Hurwitz. To this purpose a procedure for systematically designing those parameters is presented in Appendix 6.7.1. The main result is then detailed in the next proposition in which we refer to the function $\tau_e : W \to \mathbb{R}^{2d-2}$ defined as

$$\tau_e(w) := \begin{pmatrix} \tau_{e,1}(w) \\ \tau_{e,2}(w) \\ \vdots \\ \tau_{e,d-1}(w) \end{pmatrix}, \quad \tau_{e,i} := \begin{pmatrix} L_s^{i-1} c(w) \\ L_s^{i} c(w) \end{pmatrix}, \quad i = 1, \dots, d-1\,.$$

Proposition 6.2 *Let $c(\cdot)$ in (6.7) be fulfilling (6.8) and let the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ be taken as in (6.15)–(6.17) with the coefficients $(a_{i,1}, a_{i,2})$, $i = 1, \dots, d-1$, fixed so that the matrix M in (6.19) is Hurwitz. Then there exist a $\ell^\star > 0$ such that for all $\ell \geq \ell^\star$ requirements* (a') *and* (b) *of Sect. 6.2 are fulfilled with*

$$\mathscr{B} = \{(w, z, \chi) \in W \times \mathbb{R}^n \times \mathbb{R}^{2d-2}, \quad z = \pi(w)\,, \ \chi = \tau_e(w)\}$$

and the ϵ in (6.12) of the form

$$\epsilon = \frac{r}{\ell^d} \sup_{w \in W} \|\nu(w)\|$$

with r a positive number.

The proof of this proposition is presented in Appendix 6.7.2. In view of Theorem 6.2, similarly to the standard high-gain design presented in the first part of the section, the regulator (6.9) with $(\phi(\cdot), \Psi, \gamma(\cdot))$ obtained from (6.15)–(6.17) guarantees asymptotic regulation if the function $\varphi(\cdot)$ fulfills (6.8) with $\nu(\cdot) = 0$. Otherwise just practical regulation is achieved with a bound on the asymptotic error that can be arbitrarily decreased by increasing κ or ℓ^d. With respect to the previous case, however, the remarkable feature of the proposed regulator is that the high-gain parameter ℓ is powered just up to the order 2 by making the design possible even in presence of large values of d. Note, in particular, that the asymptotic gain relating the term $\nu(\cdot)$ to the regulation error is still proportional to $1/\ell^d$ notwithstanding the regulator is implemented only with terms proportional to ℓ and ℓ^2.

6.4 Dealing with Higher Relative Degree with Low-Power Tools

The analysis in the previous section has shown how to design internal model-based regulators for systems of the form (6.1) under the mild assumptions of unitary relative degree and minimum-phase. In case of systems with relative degree higher than one, standard tools proposed in the literature can be used to reduce the problem to a relative degree-one scenario to which the same design methodology presented in the Sects. 6.2 and 6.3 can be adopted. In particular, the approach that is typically pursued in literature relies on a two-phase design procedure. In a first phase, a *high-gain* backstepping design is used in order to obtain a system having relative degree one with respect to an output that is a linear combination of the regulation error and its first r time derivatives (with r the relative degree) and whose zero dynamics fulfill a minimum-phase assumption of the form presented in Assumption 6.1. To this system the same procedure proposed in the previous section can be thus applied by obtaining in this way a regulator solving the problem at hand except the fact that it processes not only the error but also its first r time derivatives. In the second phase, then, a *high-gain* dirty-derivative observer [18] is typically adopted in order to replace the error time derivatives with appropriate estimates by thus obtaining a pure error-feedback regulator. The high-gain tools that are typically adopted both in the backstepping and dirty-derivative observer phase, however are characterized by the fact that high-gain parameter is powered up to the value of the relative degree, by thus making the design hard to be implemented in practice in case of systems with high values of r. For this reason, in the following, we show how the idea of "low-power" high-gain adopted for the design of the internal model can be successfully employed also at this stage by thus obtaining a dynamic regulator in which the high-gain parameters characterizing the control structure are powered just to the order 2 regardless the value of d in (6.8) and of r.

We assume that the regulated plant is a relative degree r system described in the normal form

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w, z, e_1) \\ \dot{e}_i &= e_{i+1}\,, \qquad i = 1, \ldots, r-1 \\ \dot{e}_r &= q(w, z, e) + b(w, z, e)u \end{aligned} \tag{6.20}$$

in which $f(\cdot, \cdot, \cdot)$, $q(\cdot, \cdot, \cdot)$ and $b(\cdot, \cdot, \cdot)$ are smooth functions, with the high-frequency gain $b(\cdot, \cdot, \cdot)$ fulfilling (6.3), the regulation error to be steered to zero is e_1, and the zero dynamics (6.6) fulfill Assumption 1 in Sect. 6.2. The compact sets of initial conditions $w(0)$, $z(0)$ and $e_i(0)$ are $W \subset \mathbb{R}^\rho$, $Z \subset \mathbb{R}^n$ and $E_i \subset \mathbb{R}$, $i = 1, \ldots, r$, with W that is invariant for (6.2). We let $E := E_1 \times \cdots \times E_r$. As said before the design methodology presented in Sects. 6.2 and 6.3 for the relative degree-one case can be still used provided that a backstepping and dirty-derivative design steps, developed in the next two subsections, are adopted.

6.4.1 Low-Power Dynamic Backstepping Tools

The idea typically followed in the literature to deal with the case $r > 1$ is to start considering the system

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w, z, e_1) \\ \dot{e}_1 &= e_2 \\ &\vdots \\ \dot{e}_{r-1} &= e_r \end{aligned} \tag{6.21}$$

regarded as a system with input e_r and, under the minimum-phase Assumption 6.1, look for a (virtual) control law for e_r processing the error and its first $r - 1$ time derivatives that make the set $\mathscr{A} \times \{0\}$ asymptotically stable for this system with a domain of attraction containing the compact set of initial conditions. This is usually done with a *static* control law $e_r = e_r^\star$ of the form

$$e_r^\star = -(g^{r-1} b_1 e_1 + g^{r-2} b_2 e_2 + \cdots + g b_{r-1} e_{r-1})$$

in which g is a high-gain parameter and the b_i's are coefficients of an Hurwitz polynomial. As a matter of fact, after rescaling the variables e_i as $\zeta_i := g^{-(i-1)} e_i$, $i = 1, \ldots r - 1$, system (6.21) with $e_r = e_r^\star$ reads as

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w, z, \zeta_1) \\ \dot{\zeta} &= g H \zeta \end{aligned}$$

in which $\zeta = \begin{pmatrix} \zeta_1 & \ldots & \zeta_{r-1} \end{pmatrix}^T$ and H is an Hurwitz matrix. Standard high-gain tools, then, can be adopted to show that a large value of g makes the set $\mathscr{A} \times \{0\}$ asymptotically and locally exponentially stable with a domain of attraction containing the set $W \times Z \times E_1 \times \cdots \times E_{r-1}$ set of initial conditions. Motivated by the fact that if r is large the previous control law can be hard to implement due to the term g^{r-1}, in the following we propose a different construction for $e^\star$ in which *dynamic*, rather than static, high-gain stabilizers are developed with the feature that the high-gain parameter is powered just to the order 2 regardless the value of r. We assume that $r > 3$, otherwise the usual static control law presented before can be used.

The proposed dynamic controller has state $\vartheta = (\vartheta_1, \ldots, \vartheta_{r-3})^T \in \mathbb{R}^{r-3}$ described by the dynamics

$$\begin{aligned}
\dot{\vartheta}_1 &= e_3 + g^2 b_{1,1} e_1 + g b_{1,2}(e_2 - \vartheta_1) \\
\dot{\vartheta}_i &= e_{i+2} + g^2 b_{i1} \vartheta_{i-1} + g b_{i,2}(e_{i+1} - \vartheta_i) \qquad i = 2, \ldots, r-3 \\
e_r^\star &= -(g^2 b_{r-2,1} \vartheta_{r-3} + g b_{r-2,2} e_{r-1})
\end{aligned} \tag{6.22}$$

in which g is the high-gain parameter, and the $(b_{i,1}, b_{i,2})$, $i = 1, \ldots, r-2$, are coefficients to be appropriately designed.The latter, in turn, must be designed so that the block tri-diagonal matrix H defined as

$$H := \begin{pmatrix}
G_1 & S & 0 & & \ldots & \ldots & 0 \\
R_1 & G_2 & S & \ddots & & & \vdots \\
0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\
\vdots & \ddots & R_{i-1} & G_i & S & \ddots & \vdots \\
\vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\
\vdots & & & \ddots & R_{r-4} & G_{r-3} & S \\
0 & \ldots & \ldots & \ldots & 0 & R_{r-3} & G_{r-2}
\end{pmatrix}, \tag{6.23}$$

with

$$G_i := \begin{pmatrix} 0 & 1 \\ -b_{i,1} & -b_{i,2} \end{pmatrix}, \qquad R_i := \begin{pmatrix} b_{i,1} & b_{i,2} \\ 0 & 0 \end{pmatrix}, \qquad S := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

is Hurwitz. This can be always obtained using the constructive design procedure presented in Appendix 6.7.1 and noting that $M = (THT)^T$, with M defined in (6.19) taking $d = r - 1$ and the a_i's equal to the b_i's, and $T = T^T$ defined as the anti-diagonal identity matrix of dimension $r - 2$. The main result is stated in the following proposition whose proof is deferred in Appendix 6.7.3:

Proposition 6.3 *Consider system (6.21)–(6.22) with $e_r = e_r^\star$ with initial conditions taken in $W \times Z \times (E_1 \times \cdots \times E_{r-1}) \times \Theta$, where Θ is a compact set of $\mathbb{R}^{r-3}$. Let the coefficients $(b_{i,1}, b_{i,2})$, $i = 1, \ldots, r-2$, be designed so that the matrix H in (6.23) is Hurwitz. Then, there exists a $g^\star > 0$ such that for all $g \geq g^\star$ the set $\mathscr{A} \times \{0\} \times \{0\}$ is asymptotically and locally exponentially stable with a domain of attraction containing the set $W \times Z \times (E_1 \times \cdots \times E_{r-1}) \times \Theta$.*

The previous result is then instrumental to show that the problem of output regulation for systems with higher relative degree can be cast into the relative degree-one framework of Sects. 6.2 and 6.3. As a matter of fact, by changing the variable e_r as

$$e_r \;\mapsto\; e_r' := e_r + (g^2 b_{r-2,1} \vartheta_{r-3} + g b_{r-2,2} e_{r-1}) \tag{6.24}$$

it is easy to realize that the resulting system reads as

$$
\begin{aligned}
\dot{w} &= s(w)\\
\dot{z}' &= f'(w,\, z',\, e_r')\\
\dot{e}_r' &= \tilde{q}'(w,\, z',\, e_r') + b(w,\, z',\, e_r')u
\end{aligned}
$$

in which $z' = \mathrm{col}(z, e_1, \ldots, e_{r-1}, \vartheta)$ and $f'(\cdot,\cdot,\cdot)$, $q'(\cdot,\cdot,\cdot)$, $b'(\cdot,\cdot,\cdot)$ are properly defined functions with the former such that the system $\dot{z}' = f'(w, z', 0)$ coincides with (6.21)–(6.22) with $e_r = e_r^\star$. This system has relative degree one between the input u and the output e_r' and, in view of Proposition 6.3, it fulfills Assumption 6.1 with the set $\mathscr{A}$ of the assumption replaced by $\mathscr{A} \times \{0\} \times \{0\}$. Hence, the theory of Sect. 6.3 can be adopted off-the-shelf to obtain a regulator able to solve the problem for this system. By bearing in mind the theory above such a regulator takes the form

$$
\begin{aligned}
\dot{\xi} &= \phi(\xi) - \Psi\kappa e_r'\\
\dot{\vartheta}_1 &= e_3 + g^2 b_{1,1} e_1 + g b_{1,2}(e_2 - \vartheta_1)\\
\dot{\vartheta}_i &= e_{i+2} + g^2 b_{i1}\vartheta_{i-1} + g b_{i,2}(e_{i+1} - \vartheta_i) \qquad i = 2, \ldots, r-3\\
u &= \gamma(\xi) - \kappa e_r'
\end{aligned}
\tag{6.25}
$$

with e_r' defined in (6.24) and with the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ having the form (6.15). This controller is characterized by three high-gain parameters, to be fixed in order: g, introduced to deal with the high relative-degree case, ℓ, playing the role in the design of the internal model, and κ characterizing the static stabilizer. Remarkably, the power of these parameters does not exceed two, regardless the value of the relative degree (r) and the dimension of the internal model (d). Such a controller guarantees that if (6.8) is fulfilled with $\nu(\cdot) = 0$ then the set $\mathscr{B} \times \{0\} \times \{0\}$ is asymptotically stable with an appropriate domain of attraction (with the set $\mathscr{B}$ introduced in Theorem 6.1). This, by the definition of e_r', implies that the regulation error e_1 converges to zero asymptotically. On the other hand, if (6.8) is fulfilled with $\nu(\cdot) \neq 0$, then only practical regulation is achieved.

The regulator (6.25) thus solves the problem at hand in the general case with the drawback that it requires the knowledge of e_1, e_2, …, e_r, namely of the first r time derivatives of the regulation error. A pure error-feedback regulator can be obtained by replacing the error time derivatives with appropriate estimates provided by a (possibly low-power) dirty derivatives observer as detailed in the next section.

6.4.2 Low-Power Dirty-Derivative Observers

A standard high-gain observer able to provide a (dirty) estimate of the error and its first r time derivatives takes the form (see [13, 18])

$$
\begin{aligned}
\dot{\hat{e}}_i &= \hat{e}_{i+1} + c_i\, k^i\, (e_1 - \hat{e}_1)\,, \qquad i = 1, \ldots, r-1\,,\\
\dot{\hat{e}}_r &= c_r\, k^r\, (e_1 - \hat{e}_1)
\end{aligned}
$$

in which c_i, $i = 1, \ldots, r$, are coefficients of an Hurwitz polynomial and k is a high-gain parameter. The general result that is possible to prove is that, if the $r+1$ time derivative of e_1 is bounded, then the previous observer yields approximate estimates of the first r time derivatives of e_1, with an estimation dynamics that can be rendered arbitrarily fast by increasing k and with an asymptotic estimation error that can be arbitrarily decreased by also increasing k. The fact that such a observer can be successfully used to replace the e_i's in (6.25) comes from classical arguments of (nonlinear) output feedback that do not need to be repeated here see for instance [17, 18]. Similarly to the classical high-gain tools presented before, however, the main criticisms that can be raised to the previous structure is that the high-gain parameter k is powered up to the order r, which makes the practical implementation of the observer very hard in case of systems with high relative degree. For this reason, by following what proposed in [1], we propose a low-power dirty derivatives observer that takes the form

$$\begin{aligned}
\dot{\eta}_{1,1} &= \eta_{1,2} + c_{1,1}\,k\,(e_1 - \eta_{1,1})\\
\dot{\eta}_{1,2} &= \eta_{i+1,2} + c_{1,2}\,k^2\,(e_1 - \eta_{1,1})\\
\dot{\eta}_{i,1} &= \eta_{i2} + c_{i,1}\,k\,(\eta_{i-1,2} - \eta_{i,1})\\
\dot{\eta}_{i,2} &= \eta_{i+1,2} + c_{i,2}\,k^2\,(\eta_{i-1,2} - \eta_{i,1}) \qquad i = 2, \ldots, r-2\\
\dot{\eta}_{r-1,1} &= \eta_{i2} + c_{r-1,1}\,k\,(\eta_{r-2,2} - \eta_{r-1,1})\\
\dot{\eta}_{r-1,2} &= c_{r-1,2}\,k^2\,(\eta_{r-2,2} - \eta_{r-1,1})
\end{aligned} \tag{6.26}$$

with state $\eta = \mathrm{col}(\eta_1, \ldots, \eta_{r-1}) \in \mathbb{R}^{2r-2}$, $\eta_i = (\eta_{i,1},\ \eta_{i,2})^T \in \mathbb{R}^2$, coefficients $(c_{i,1}, c_{i,2})$ to be properly designed, and estimated variables $\hat{e} = \mathrm{col}(\hat{e}_1, \ldots, \hat{e}_r) \in \mathbb{R}^r$, given by

$$\hat{e} = \Gamma \eta$$

with Γ defined in (6.18). As in the classical observer, it can be shown (see [1]) that if the $r + 1$ time derivative of e_1 is bounded, the estimation dynamics of the previous observer can be rendered arbitrarily fast by increasing k and the variables $\hat{e}$ provide a practical estimation of $(e_1, \ldots, e_r)$ with an asymptotic estimation error that can be arbitrarily decreased by also increasing k. To this end the coefficients $(c_{i,1}, c_{i,2})$ must be fixed so that the matrix J defined as the M in (6.19) with d replaced by r and with the coefficients $(a_{i,1}, a_{i,2})$ in the definitions of L_i and Q_i replaced by $(c_{i,1}, c_{i,2})$ is Hurwitz (the procedure in Appendix 6.7.1 can be used to this purpose). If compared with the classical high-gain dirty derivatives observer, the structure (6.26) has the remarkable feature of having the high-gain parameter k powered just up to the order 2, regardless the value of r, at the price of extending the dimension of the observer to $2r - 2$.

It turns out that the estimate $\hat{e}$ provided by (6.26) and properly saturated can be used to replace the e_i's in (6.25) to obtain a pure error-feedback regulator. In particular, the latter assumes the form

$$\begin{aligned}
\dot{\xi} &= \phi(\xi) + \Psi v \\
\dot{\vartheta}_1 &= -g a_{1,2}\vartheta_1 + \hat{e}^{\mathrm{s}}_3 + g^2 a_{1,1}\hat{e}^{\mathrm{s}}_1 + g a_{1,2}\hat{e}^{\mathrm{s}}_2 \\
\dot{\vartheta}_i &= -g a_{i,2}\vartheta_i + g^2 a_{i1}\vartheta_{i-1} + \hat{e}^{\mathrm{s}}_{i+2} + g a_{i,2}\hat{e}^{\mathrm{s}}_{i+1} \qquad i = 2, \dots, r-3 \\
u &= \gamma(\xi) + v \\
v &= -\kappa \hat{e}^{\mathrm{s}}_r + g^2 a_{r-2,1}\vartheta_{r-3} + g a_{r-2,2}\hat{e}^{\mathrm{s}}_{r-1}
\end{aligned} \tag{6.27}$$

with $\hat{e}^{\mathrm{s}} = \mathrm{col}(\hat{e}^{\mathrm{s}}_1, \dots, \hat{e}^{\mathrm{s}}_r) \in \mathbb{R}^r$ defined as

$$\hat{e}^{\mathrm{s}} = \bar{\sigma}_L(\Gamma\eta) \tag{6.28}$$

in which $\bar{\sigma}_L(\cdot) : \mathbb{R}^r \to \mathbb{R}^r$ is a piecewise linear saturation mapping defined as $\bar{\sigma}_L(s) = \mathrm{col}(\sigma_1(s_1), \dots, \sigma_r(s_r))$ with $\sigma_i : \mathbb{R} \to \mathbb{R}$ defined as $\sigma_i(s_i) = s_i$ if $|s_i| \leq L$ and $\sigma_i(s_i) = L\,\mathrm{sign}(s_i)$ otherwise, where L is a positive constant to be fixed. It turns out that the saturation level L and the high-gain parameter k can be tuned to have the regulation objective fulfilled as detailed in the next final proposition whose proof is deferred to Appendix 6.7.4.

Proposition 6.4 *Consider the closed-loop system given by (6.20) and (6.26)–(6.27) with the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ taken as in (6.15)–(6.16), where the function $\varphi(\cdot)$ in the definition of $\phi(\cdot)$ is assumed to fulfill (6.8) with $\nu = 0$. Let the initial conditions of the system be taken in the compact set $W \times Z \times (E_1 \times \cdots \times E_r) \times \Theta \times \Sigma$ with Σ a compact set of $\mathbb{R}^{2r-2}$. Let the coefficients $(a_{i,1}, a_{i,2})$, $i = 1, \dots, d-1$, $(b_{i,1}, b_{i,2})$, $i = 1, \dots, r-2$, and $(c_{i,1}, c_{i,2})$, $i = 1, \dots, r-1$ be fixed so that the matrices M, H and J are Hurwitz. Let g, and accordingly ℓ and κ, be fixed according to Propositions 6.2, 6.3 and Theorem 6.1 so that the set $\mathscr{A} \times \{0\} \times \{0\}$ is asymptotically and locally exponentially stable for (6.20) and (6.25) with domain of attraction containing $W \times Z \times (E_1 \times \cdots \times E_{r-1}) \times \Theta$. Then, there exists a $L^\star$ and, for all $L \geq L^\star$, a $k^\star$ such that for all $k \geq k^\star$ the set $\mathscr{A} \times \{0\} \times \{0\} \times \{0\}$ is asymptotically and locally exponentially stable for (6.20) and (6.27) with domain of attraction containing $W \times Z \times (E_1 \times \cdots \times E_r) \times \Theta \times \Sigma$.*

For sake of simplicity, the previous proposition has been given in the case of asymptotic regulation, namely in case the function $\varphi(\cdot)$ embedded in the internal model makes (6.8) satisfied with $\nu(w) = 0$. It is not difficult to show that, in case $\nu(w) \neq 0$, the same error-feedback controller achieves practical regulation with a bound on the asymptotic error of the form (6.14). It is worth also noting that the overall regulator has dimension $2d + 3r - 7$ (being $2d - 2$ the dimension of ξ, $r - 3$ the dimension of ϑ, and $2r - 2$ the dimension of η) with the high-gain parameters, which are ℓ, g, k and κ, that are powered at most up to the order 2 regardless the value of d and r.

6.5 Simulation Results

We consider the problem of rejecting a disturbance d acting on the input of the linear system

$$\begin{aligned}\dot{z} &= A\,z + B\,(u - \mathrm{d})\\ e &= C\,z\end{aligned}$$

with (A, B, C) a controllable and observable triplet and the disturbance d generated by the nonlinear Duffing oscillator

$$\begin{aligned}\ddot{w}' &= \;-\alpha w'^3 - \beta w'\\ \mathrm{d} &= w'\end{aligned} \tag{6.29}$$

with α and β uncertain parameters. In this specific case the friend is $c(w') = w'$.The triplet (A, B, C) is taken so that the relative degree is 4, namely $CB = 0$, $CAB = 0$, $CA^2B = 0$ and $CA^3B \neq 0$. In the simulations, we considered

$$A = \begin{pmatrix} 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0.2 & -0.6 & -0.5 & 1\end{pmatrix}, \qquad B = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}, \qquad C = \begin{pmatrix}1 \; 0 \; 0 \; 0\end{pmatrix}.$$

Using the arguments in [5], system (6.29) can be shown to be immersed into a system of the form

$$\dot{w} \;=\; Fw + G\varphi(w)\;, \qquad \mathrm{d} = \; Hw \tag{6.30}$$

with $w \in \mathbb{R}^6$, (F, G, H) triplet of dimension 6 in prime form, and $\varphi(\cdot)$ defined as

$$\varphi(w) \;=\; -\hat{\alpha}(w)\left(36w_2^2w_3 + 18w_1w_3^2 + 24w_1w_2w_4 + 3w_1^2w_5\right) - \hat{\beta}(w)w_5 \tag{6.31}$$

with $(\hat{\alpha}(w), \hat{\beta}(w))$ defined as

$$(\hat{\alpha}(w), \hat{\beta}(w))^T = (\Upsilon(w)^T\Upsilon(w))^{-1}\Upsilon(w)^T w_{[3,6]}$$

with

$$\Upsilon(w) = \begin{pmatrix} -w_1^3 & -w_1\\ -3w_1^2w_2 & -w_2\\ -3w_3w_1^2 - 6w_1w_2^2 & -w_3\\ -3w_4w_1^2 - 18w_1w_2w_3 - 6w_2^3 & -w_4\end{pmatrix}, \qquad w_{[3,6]} = \begin{pmatrix}w_3\\w_4\\w_5\\w_6\end{pmatrix}.$$

In fact, it is easy to check that any behavior $\mathrm{d}(t)$ generated by (6.29) with initial condition $(w'(0), \dot{w}'(0))$ can be also generated by (6.30) with initial conditions taken as

$$\begin{aligned}&w_1(0) = w'(0)\,, \quad w_2(0) = \dot{w}'(0)\,, \quad w_3(0) = -\alpha w'(0)^3 - \beta w'(0)\\ &w_4(0) = -3\alpha w'(0)^2\dot{w}'(0) - \beta\dot{w}'(0)\\ &w_5(0) = -6\alpha w'(0)\dot{w}'(0)^2 - 3\alpha w'(0)^2w_3(0) - \beta w_3(0)\\ &w_6(0) = -6\alpha\dot{w}'(0)^2 - 18\alpha w'(0)\dot{w}'(0)w_3(0) - 3\alpha w'(0)^2w_4(0) - \beta w_4(0)\end{aligned}$$

Table 6.1 Asymptotic norm of the error in the various scenarios normalized with respect to the value of the asymptotic error when no internal model is present

d	0	2	3	4	5	6
$\|e\|_a$	1	10^{-3}	10^{-4}	$0.5 \cdot 10^{-4}$	10^{-5}	0

Furthermore, it turns out that $\alpha = \hat{\alpha}(w)$ and $\beta = \hat{\beta}(w)$ for all $w \in \mathbb{R}^6$. The stabilizer with state (ϑ, η) has dimension 7 (with θ having dimension 1 and η having dimension 6) with the coefficients b_i and c_i that, according to the procedure in Appendix 6.7.1, have been chosen as $b_1 = (0.24,\ 0.5)$, $b_2 = (1,\ 0.5)$, $c_1 = (4,\ 10.6)$, $c_2 = (4,\ 3.5)$, $c_3 = (4,\ 1.16)$. Different designs of internal model at increasing dimension have been then simulated. In particular, we implemented the internal model (6.15)–(6.17) in six scenarios in which the dimension d has been taken as $d = 0$ (namely no internal model is present), $d = 2$, $d = 3$, $d = 4$ and $d = 5$ with in all cases the $\varphi_s(\cdot)$ in the expression of $\phi(\cdot)$ taken as $\varphi_s(\cdot) = 0$. Practical regulation is expected in all these cases. We also simulated the case in which $d = 6$ and the $\varphi(\xi)$ in the expression of $\phi(\cdot)$ is the one in (6.31) saturated at the value $L = 50$. In this case exact regulation is expected. In all the scenarios, the coefficients a_i of the low-power internal model have been taken as $a_1 = (3.8, 16.4)$, $a_2 = (3.8, 6.5)$, $a_3 = (3.8, 3.2)$, $a_4 = (3.8, 1.6)$, $a_5 = (3.8, 0, 6)$. The high-gain parameters in the simulation have been taken as $g = 3$, $\ell = 7$, $\kappa = 40$, $k = 10^3$. The initial conditions have been taken as $(w'(0), \dot{w}'(0)) = (1, 0)$, $z(0) = (1, 1, 1, 1)$ and the origin for θ, η and ξ. The result of the asymptotic error are shown in Table 6.1 where we denoted with $|\cdot|_a$ the asymptotic norm of a signal, i.e. $|x|_a := \limsup_{t\to\infty} |x(t)|$. The values of the errors are normalized with respect to the value of the output e when the internal model is not present, namely with $d = 0$. We can see that by augmenting the dimension of the internal model the asymptotic norm of the error decreases. In the last scenario asymptotic regulation is achieved and $|e|_a = 0$.

6.6 Conclusions

The problem of output regulation for the class of nonlinear systems that have a well-defined relative degree and are minimum-phase has been investigated. The paradigm of this work follows the main idea of [3] where it has been shown that the theory of high-gain observers can be used for the design of internal models. In the recent work [1], it has been shown that it is possible to design, by means of dynamic extension, a high-gain observer with a high-gain parameter that is raised up to the order two regardless the dimension of the observed system. Here, by following the idea of [1], we presented a new design methodology for internal model-based regulators based on “low-power” high-gain tools. The methodology of [1] has been successfully extended in two directions. First, we showed that the internal model can be designed by following the high-gain observer paradigm of [3] with the new low-power approach of [1]. Second, we showed that this methodology can be extended also to the design of the stabilizer. In particular, for systems having a high relative degree, we showed that a static state-feedback, which involves a high-gain parameter that is raised up to the order of the relative degree, can be transformed into a dynamic state-feedback which involves a high-gain parameter that is raised up only to order two. A simple example has been added to show the effectiveness of the new design. We considered, for the sake of simplicity, the class of single-input single-output nonlinear systems. However, it is worth noting that the same tools can be easily applied to multi-input multi-output nonlinear systems which have the same number of inputs and outputs satisfying a "positivity" condition on the high-frequency gain matrix (see [2]).

6.7 Appendix

6.7.1 Choice of Parameters

With the definitions of L_i, Q_i and N given in Sect. 6.3, let the matrices $M_i \in \mathbb{R}^{2i \times 2i}$ be defined as $M_1 = L_1$ and

$$M_i \ := \begin{pmatrix} L_1 & N & 0 & & \dots & \dots & 0 \\ Q_2 & L_2 & N & \ddots & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & Q_j & L_j & N & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & Q_{i-1} & L_{i-1} & N \\ 0 & \dots & \dots & \dots & 0 & Q_i & L_i \end{pmatrix} \tag{6.32}$$

for $i = 2, \dots d - 1$. Note that $M_{d-1} = M$ with M defined in (6.19). Furthermore, let $G_i(s)$, $i = 1, \dots, d - 1$, be the transfer functions defined as

$$G_i(s) \ := \ B_{2(i-1)}^{\top}(sI_{2(i-1)} - M_{i-1})^{-1} B_{2(i-1)} \ ,$$

in which $B_{2(i-1)}$ is the column vector of dimension $2(i - 1)$ whose elements are all zero except the last element that is 1, and and let γ_i be defined as

$$\gamma_i \ := \ \max_{\omega \in \mathbb{R}} |G_i(j\omega)| \, . \tag{6.33}$$

Then we have the following lemma.

Lemma 6.1 *Let the coefficients $(a_{i,1}, a_{i,2})$ be chosen according to the following recursive algorithm*

- *Let $a_{11} > 0$ and $a_{12} > 0$ be any positive real numbers;*
- *let $a_{i1} = a_{(i-1)1}$ and let $a_{i2} > 0$ be chosen such that $a_{i2} \ < \ \dfrac{a_{i1}}{\gamma_{i-1}}$.*

Then the matrices M_i, $i = 1, \dots, d - 1$, are Hurwitz.

Proof The proof relies on small gain arguments. At the generic step i-th, the matrix M_i is the state matrix of the system resulting from the feedback interconnection, obtained by $u_i = y_{i-1}$, and $u_{i-1} = y_i$, of the following systems

$$\begin{cases} \dot{x}_{i-1} &= M_{i-1}x_{i-1} + B_{2(i-1)}u_{i-1} \\ y_{i-1} &= B_{2(i-1)}^{\top}x_{i-1} \end{cases}$$
$$\begin{cases} \dot{x}_i &= L_i x_i + K_i u_i \\ y_i &= B^{\top}x_i \end{cases}$$

in which $B = \begin{pmatrix} 0 & 1 \end{pmatrix}^T$ and $K_i = \begin{pmatrix} a_{i,1} & a_{i,2} \end{pmatrix}^T$. With $H_i(s)$ the transfer function of the second subsystem, an elementary computation shows that

$$\max_{\omega \in \mathbb{R}} |H_i(j\omega)| = \frac{a_{i,2}}{a_{i,1}}$$

from which the result follows by small gain arguments. □

6.7.2 Proof of Proposition 6.2

Proof By the indicated choices of the triplet $(\phi(\cdot), \Psi, \gamma(\cdot))$ in (6.15)–(6.17), it turns out that the χ subsystem in (6.11) reads as

$$\begin{aligned} \dot{\chi}_1 &= S\,\chi_1 + N\,\chi_2 + D_2(\ell)\,a_1 \left(\frac{q(w, z, 0)}{b(w, z, 0)} - C\,\chi_1 \right) \\ \dot{\chi}_i &= S\,\chi_i + N\,\chi_{i+1} + D_2(\ell)\,a_i\,(B^T\,\chi_{i-1} - C\,\chi_i)\,, \quad i = 2, \ldots, d-2 \\ \dot{\chi}_{d-1} &= S\,\chi_{d-1} + B\,\varphi_s(\Gamma\chi) + D_2(\ell)\,a_{d-1}\,(B^T\,\chi_{d-2} - C\,\chi_{d-1}) \end{aligned} \tag{6.34}$$

where (S, B, C) is a triplet in prime form of dimension 2, $a_i = \begin{pmatrix} a_{i,1} & a_{i,2} \end{pmatrix}^T$, $D_2(\ell) = \text{diag}(\ell,\ \ell^2)$, and $N = \text{diag}(0,\ 1)$. By changing variables as

$$\chi_i \;\mapsto\; \tilde{\chi}_i := \ell^{2-i}\,D_2(\ell)^{-1}\left(\chi_i - \tau_{e,i}(w)\right) \qquad i = 1, \ldots, d-1\,,$$

an easy calculation shows that system (6.34) transforms as

$$\dot{\tilde{\chi}} = \ell M\tilde{\chi} + \frac{1}{\ell^{d-1}}B_{2d-2}\Delta_\ell(\tilde{\chi}, w) + \frac{1}{\ell^{d-1}}\nu(w) + \ell L_{2d-2}\delta(w, z) \tag{6.35}$$

in which $\tilde{\chi} = \text{col}\,(\tilde{\chi}_1\,,\ \ldots\,,\ \tilde{\chi}_{d-1})$,

$$\Delta_\ell(\tilde{\chi}, w) := \varphi_s(D(\ell)\tilde{\chi} + \tau(w)) - \varphi(\tau(w))$$

with $D(\ell) = \text{diag}(\frac{1}{\ell},\ 1,\ \ell, \ldots,\ \ell^{d-3}) \oplus D_2(\ell)$, $L_{2d-2} = \text{col}\begin{pmatrix} a_1 & 0 \ldots 0 \end{pmatrix}$ and

$$\delta(w,z) := \frac{q(w,z,0)}{b(w,z,0)} - c(w)\,.$$

Note that, using the fact that $\varphi(\cdot)$ is uniformly Lipschitz on $\tau(W)$ and that $\varphi_s(\cdot)$ is bounded, it follows that there exists a constant $r > 0$ such that

$$\frac{1}{\ell^{d-1}} \|B_{2d-2}\Delta_\ell(\tilde{\chi}, w)\| \leq r\|\tilde{\chi}\|$$

for all $\tilde{\chi} \in \mathbb{R}^{2d-2}$, $w \in W$ and $\ell \geq 1$. From this and the fact that M is Hurwitz, standard Lyapunov arguments can be used to prove that the system (6.35) is Input-to-State Stable with respect to the inputs $\nu(\cdot)$ and $\delta(\cdot,\cdot)$ without restrictions on the initial state and on the input and with a linear asymptotic gains that depend on $1/\ell^d$ for the input $\nu(\cdot)$, and not dependent on ℓ for the input $\delta(\cdot,\cdot)$. The claim of Proposition 6.2 then follows by Assumption 6.1 and by the definition of $\tilde{\chi}$. □

6.7.3 *Proof of Proposition 6.3*

Proof We consider the change of variables

$$\mathrm{col}(e_1,\ldots,e_{r-1},\theta) \;\mapsto\; \tilde{e} = \mathrm{col}\left(\tilde{e}_1 \,\ldots\, \tilde{e}_{r-2}\right) \in \mathbb{R}^{2r-4}$$

in which $\tilde{e}_i \in \mathbb{R}^2$ are defined as

$$\tilde{e}_1 := \begin{pmatrix} e_1 \\ e_2 - \vartheta_1 \end{pmatrix}, \quad \tilde{e}_i := \begin{pmatrix} \vartheta_{i-1} \\ e_{i+1} - \vartheta_i \end{pmatrix}, \; i = 2,\ldots,r-3\,, \quad \tilde{e}_{r-2} := \begin{pmatrix} \vartheta_{r-3} \\ e_{r-1} \end{pmatrix}.$$

By rescaling $\tilde{e}$ into $\tilde{e}' = \mathrm{col}\left(\tilde{e}'_1 \,\ldots\, \tilde{e}'_{r-2}\right)$ with $\tilde{e}'_i \in \mathbb{R}^2$ defined as

$$\tilde{e}'_i := \begin{pmatrix} g^{1-i} & 0 \\ 0 & g^{-i} \end{pmatrix} \tilde{e}_i$$

it turns out that the closed-loop system (6.21)–(6.22) reads as

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w,\, z,\, \tilde{e}_{11}) \\ \dot{\tilde{e}}' &= g\, H\, \tilde{e}' \end{aligned}$$

by which, using Assumption 6.1, the proof of the result follows by using classical Lyapunov arguments. □

6.7.4 Proof of Proposition 6.4

Proof The proof follows standard paradigms in the field of output feedback stabilization for nonlinear systems and it is thus just sketched. Consider system (6.26) and the change of variables

$$\begin{pmatrix} \eta_{i,1} \\ \eta_{i,2} \end{pmatrix} \mapsto \tilde{\eta}_i = \begin{pmatrix} \tilde{\eta}_{i,1} \\ \tilde{\eta}_{i,2} \end{pmatrix} := \begin{pmatrix} \eta_{i,1} - e_i \\ \eta_{i,2} - e_{i+1} \end{pmatrix} \quad i = 1, \dots, r-1.$$

Furthermore, consider system (6.27) and add and subtract the term $\hat{e}_3 + g^2 a_{1,1}\hat{e}_1 + g a_{1,2}\hat{e}_2$ from the $\dot{\theta}_1$ dynamics, the term $\hat{e}_{i+2} + g a_{i,2}\hat{e}_{i+1}$ from the $\dot{\theta}_i$ dynamics, $i = 2, \dots, r-3$, and the term $\hat{e}_r + g a_{r-2,2}\hat{e}_{r-1}$ from the expression of v. By rescaling the variables $\tilde{\eta}_i$

$$\tilde{\eta}_i \;\mapsto\; \zeta_i := k^{2-i} D_2(k)^{-1}\tilde{\eta}_i \qquad i = 1, \dots, r-1$$

with $D_2(k) = \mathrm{diag}(k,\ k^2)$, it turns out that the closed-loop system reads as

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{x} &= F(w, x) + \Delta_k(x, \zeta) \\ \dot{\zeta} &= k\, J\, \zeta + \frac{1}{k^{r-1}}\, B\, \delta_k(w, x, \zeta) \end{aligned}$$

in which $x := \mathrm{col}(z, \xi, (e_1, \dots, e_r), \vartheta) \in \mathbb{R}^{n+2r+d-3}$, $\zeta := \mathrm{col}(\zeta_1 \dots, \zeta_r - 1)$, $\Delta_k(\cdot)$ and $\delta_k(\cdot)$ are appropriately defined functions (dependent on k) and $F(\cdot,\cdot)$ is such that, by construction, the set $\mathscr{B} \times \{0\} \times \{0\}$ is asymptotically and locally exponentially stable with a domain of attraction containing the set $W \times \Xi \times E_1 \times \cdots \times E_r \times \Theta$. As far as the functions $\Delta_k(\cdot)$ and $\delta_k(\cdot)$ are concerned, it is easy to see from their definition that for any compact set $X \subset \mathbb{R}^{n+2r+d-3}$ and $Z \in \mathbb{R}^{2r-2}$ there exist positive constants d_1 and d_2 (*not dependent on* k) and a value of $L^\star > 0$ such that for all $L \geq L^\star$

$$\begin{array}{c} \Delta_k(x, 0) = 0\,, \quad \|\Delta_k(x, \zeta)\| \leq d_1 \\ \|\dfrac{1}{k^{r-1}} B\delta_k(, x, \zeta)\| \leq d_2\|\zeta\| \end{array} \qquad \forall x \in X\,,\ \zeta \in Z\,,\ k > 0\,.$$

From this the result of the proposition follows by standard Lyapunov arguments.

References

1. Astolfi, D., Marconi, L.: A high-gain nonlinear observer with limited gain power. IEEE Trans. Autom. Control **60**(11), 3059–3064 (2015)
2. Astolfi, D., Isidori, A., Marconi, L., Praly, L.: Nonlinear output regulation by post-processing internal model for multi-input multi-output systems. Nonlinear Control Syst. Symp. **9**, 295–300 (2013)

3. Byrnes, C.I., Isidori, A.: Nonlinear internal models for output regulation. IEEE Trans. Autom. Control **49**(12), 2244–2247 (2004)
4. Delli Priscoli, F., Marconi, L., Isidori, A.: Adaptive observers as nonlinear internal models. Syst. Control Lett. **55**, 640–649 (2006)
5. Forte, F., Isidori, A., Marconi, L.: Robust design of internal models by nonlinear regression. In: 8th IFAC Nolcos, Tolouse, France (2013)
6. Gauthier, J.P., Kupka, I.: Deterministic Observation Theory and Applications. Cambridge University Press (2001)
7. Huang, J., Lin, C.F.: On a robust nonlinear multivariable servomechanism problem. IEEE Trans. Autom. Control **AC-39**, 1510–1513 (1994)
8. Isidori, A.: A tool for semiglobal stabilization of uncertain non-minimum-phase nonlinear systems via output feedback. IEEE Trans. Autom. Control **45**(10), 1817–1827 (2000)
9. Isidori, A.: Nonlinear output regulation: a unified design philosophy. In: Perspectives in Mathematical System Theory, Control, and Signal Processing, pp. 83–94. Springer, Berlin (2010)
10. Isidori, A., Byrnes, C.I.: Output regulation of nonlinear systems. IEEE Trans. Autom. Control **AC-25**, 131–140 (1990)
11. Isidori, A., Marconi, L., Praly, L.: Robust design of nonlinear internal models without adaptation. Automatica **48**(10), 2409–2419 (2012)
12. Marconi, L., Praly, L., Isidori, A.: Robust stabilization via nonlinear Luenberger observer. SIAM J. Control Optim. **45**(6), 2277–22298 (2007)
13. Khalil, H.K., Praly, L.: High-gain observers in nonlinear feedback control. Int. J. Robust Nonlinear Control **24**(6), 993–1015 (2014)
14. Li, R., Khalil, H.K.: Conditional integrator for non-minimum phase nonlinear systems. In: 51nd IEEE Conference on Decision and Control, pp. 4883–4887 (2012)
15. Marconi, L., Praly, L.: Uniform practical output regulation. IEEE Trans. Autom. Control **AC-53**, 1184–1202 (2008)
16. Serrani, A., Isidori, A., Marconi, L.: Semiglobal nonlinear output regulation with adaptive internal model. IEEE Trans. Autom. Control **AC-46**, 1178–1194 (2001)
17. Teel, A.R., Praly, L.: Global stabilizability and observability imply semi-global stabilizability by output feedback. Syst. Control Lett. **22**, 313–325 (1994)
18. Teel, A.R., Praly, L.: Tools for semiglobal stabilization by partial state and output feedback. SIAM J. Control Optim. **33**, 1443–1485 (1995)
19. Wang, L., Astolfi, D., Su, H., Marconi, L.: High-gain observers with limited gain power for systems with observability canonical form. Automatica (2015)

Chapter 7
Passivity-Based Control of Mechanical Systems

Romeo Ortega, Alejandro Donaire and Jose Guadalupe Romero

Abstract Stabilization of mechanical systems by shaping their energy function is a well-established technique whose roots date back to the work of Lagrange and Dirichlet. Ortega and Spong in 1989 proved that passivity is the key property underlying the stabilization mechanism of energy shaping designs and the, now widely popular, term of passivity-based control (PBC) was coined. In this chapter, we briefly recall the history of PBC of mechanical systems and summarize its main recent developments. The latter includes: (i) an explicit formula for one of the free tuning gains that simplifies the computations, (ii) addition of PID controllers to robustify and make constructive the PBC design and to track ramp references, (iii) use of PBC to solve the position feedback global tracking problem, and (iv) design of robust and adaptive speed observers.

R. Ortega (✉)
Laboratoire des Signaux et Systèmes, CNRS–SUPELEC, Plateau du Moulon,
91192 Gif–sur–Yvette, France
e-mail: ortega@lss.supelec.fr

A. Donaire
PRISMA Lab, Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione,
Università di Napoli Federico II, Via Claudio 21, 80125 Naples, Italy
e-mail: alejandro.donaire@newcastle.edu.au

A. Donaire
School of Engineering, The University of Newcastle, Callaghan, Australia

J.G. Romero
Departamento Académico de Sistemas Digitales, Instituto Tecnológico Autónomo de
México-ITAM, Rio Hondo No. 1, 01080 Distrito Federal, Mexico
e-mail: jose.romerovelazquez@itam.mx

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_7

7.1 Background on Passivity-Based Control

7.1.1 General Systems

Passive systems are a class of dynamical systems in which the energy exchanged with the environment plays a central role. In passive systems, the rate at which the energy flows into the system is not less than the increase in storage. In other words, a passive system cannot store more energy than is supplied to it from the outside, with the difference being the dissipated energy—a feature that is captured by the energy balance equation of the system. It is clear then that passivity is intimately related with the stability properties of the system. A far-reaching interpretation of the action of a controller is to view it as a process of energy exchange between two interconnected systems [53, 69]. If the overall energy balance is positive, in the sense that the energy generated by one subsystem is dissipated by the other one, the interconnection will be stable. This property explains the interest of passivity as a basic building block for control of dynamical systems. See [45] for an early account of the applications in control of input–output, and in particular passivity, theory.

The first attempts to use passivity in control theory are due to Fradkov [24] who gave an answer to the question of feedback passivation of linear time-invariant (LTI) systems. To the best of our knowledge, the use of *feedback passivation for stabilization* of nonlinear systems was first reported in [50, 55], where the work of [41] and the nonlinear Kalman–Yakubovich–Popov lemma of [28] are used as design tools for adaptive stabilization of non-feedback linearizable, but passifiable, nonlinear systems. It should be pointed out that [41] is the first paper where the fundamental concepts of stabilization, existence of Lyapunov functions and optimality are shown to be closely connected via passivity. Stabilization of cascaded systems via feedback passivation was first proposed in [44] and later generalized[1] in the groundbreaking paper [10] where the nonlinear version of Fradkov's result was reported.

7.1.2 Fully Actuated Mechanical Systems: Potential Energy Shaping

Analyzing the stability of *mechanical* systems using its total energy function dates back to Lagrange, Dirichlet, and Lord Kelvin—see [35] for a fascinating review of this circle of ideas. In the control context this approach was first used by Takegaki and Arimoto in the seminal paper [65] who proposed to shape the potential energy and to add damping to solve the point-to-point *positioning* task for a fully actuated robot manipulator. This result had a great impact in the robotics community because the controller resulting from this technique is a simple PD law, which ensures global asymptotic stability (GAS) of the desired robot position in spite of its highly com-

[1] See Remark 5.6 of [10].

plicated nonlinear dynamics. Interestingly Jonckeere [29], working independently of Takegaki and Arimoto, suggested also the use of PD-like energy shaping and damping injection controllers for stabilization of a class of Euler–Lagrange systems, which includes mechanical, electrical and electromechanical systems.

Not surprisingly, though unknown to the previous authors, the key property underlying the success of such a simple scheme is the passivity of the system dynamics. As proved in Proposition 2.2.5 of [51] a broad class of Euler–Lagrange systems define passive maps from the external forces to the derivative of the generalized coordinates, which in the case of mechanical systems are the coordinate velocities. Invoking this property the derivative action of the aforementioned PD is assimilated to a constant feedback around the passive output, while the proportional one adds a term to the systems potential energy to assign a minimum at the desired equilibrium, making the total energy function a suitable Lyapunov function. It should be mentioned that this energy-shaping plus damping injection construction proposed 34 years ago is still the basis of most developments in passivity-based control (PBC). This term, which now enjoys a wide popularity, was coined in [49] to describe a controller design procedure where the control objective is achieved via passivation.

Although the basic passivity property mentioned above suffices to explain the action of the PD controller of Takegaki and Arimoto, it is necessary to invoke another property to analyze from the passivity viewpoint the *tracking* controller of [62]—namely the now well known "*skew-symmetric*" property, which was first reported by Koditschek in [34]. Using this property it was first established in [31] (see also [32]) that robot manipulators—without potential energy—define passive maps from external forces to the filtered tracking error cleverly introduced in [62]. Since gradient parameter estimators also define passive maps [36] it was then possible to analyze, using an input–output framework, the *adaptive* control scheme of [62]. The skew-symmetric property was defined using the Christoffel symbols in [49], where the proof of passivity of the (modified) robot dynamics, with the potential energy term, appeared first.

The skew-symmetric property is the fundamental component of the recent developments in PBC of network and vision-based robotics [27, 43] as well as the so-called "Standard PBC" that is elaborated in [51] for a wide range of applications, including electromechanical systems, power electronic systems and, more recently, windmill generation systems [14, 40].

7.1.3 Underactuated Mechanical Systems: Total Energy Shaping

While fully actuated mechanical systems admit an arbitrary shaping of the potential energy by means of feedback, and therefore stabilization to any desired equilibrium, this is in general not possible for underactuated systems. In certain cases this problem can be overcome by also modifying the kinetic energy of the system. This idea of *total energy shaping* was proposed in [2] where the first solution to the problem of position feedback stabilization of robots with flexible joints was solved modifying both the kinetic and potential energies of the manipulator and adding damping through the controller. See [46] for an interpretation of this controller as an interconnection of passive dynamical systems—an approach further elaborated in [23, 53]. See also [64] for similar developments and Sect. 3.2 of [51] for the connection with the approximate differentiation scheme of [33].

Total energy shaping is achieved in [2, 23, 46, 53] viewing the controller as another *dynamical* system, with its *own energy function*, interconnected with the system to be controlled. If the interconnection is power preserving the energy and dissipation functions of plant and controller add up—achieving the desired energy shaping plus damping injection. It is also possible to modify the total energy and add damping via *static* state feedback, which is the approach adopted in the method of controlled Lagrangians (CL) [8] and interconnection and damping assignment (IDA) PBC [52], see also the closely related work [25]. In both cases stabilization (of a desired equilibrium) is achieved identifying the class of systems—Lagrangian for CL and Hamiltonian for IDA-PBC—that can possibly be obtained via feedback. The conditions under which such a feedback law exists are called *matching conditions*, and consist of a set of nonlinear partial differential equations (PDEs). In case these PDEs can be solved the original control system and the target dynamic system are said to *match*.

7.1.4 IDA-PBC and the Controlled Lagrangian Methods

Given several erroneous accounts of the history of the CL and the IDA-PBC methods reported in the literature in this subsection we give precise references to place them in their right perspective, for further technical details see [7]. In the original formulation of CL reported in [8] the (mathematically motivated) concern of preserving the symmetry of the system gives rise to two serious problems. First, in terms of energy shaping, symmetry preservation translates into the modification of the kinetic energy only, leading to designs where the closed-loop inertia matrix is *negative* definite. Leaving aside the fact that this is a rather unnatural situation for a method that claims the "preservation of the physical structure", the unavoidable presence of friction that pushes the state towards a *minimum* of the energy, renders the design practically useless, see [70]. A second problem is that it stabilizs *relative equilibria* only—for the cart–pendulum system this means that only the pendulum position is stabilized. The overcome these problems potential energy shaping was also included in [6] and later adopted under the name "symmetry-breaking potential" in [9]. It should, however, be mentioned that if the design is carried out using the so-called simplified matching equations—which were introduced in [8] to avoid the need to solve PDEs—the first problem is still present, a fact that was recognized in [5, 6] where the need to solve the PDEs is stressed.

The class of mechanical Hamiltonian systems considered in IDA-PBC strictly contains the Lagrangian systems proposed in the CL method of [8, 9]. Hence, it is not surprising that the latter is a special case of the more general IDA-PBC method. Indeed, the Hamiltonian formulation of IDA-PBC allows the inclusion of *gyroscopic forces* in the target dynamics, which translates into the presence of a free skew-symmetric matrix in the matching equation—making simpler their solution. In Proposition 7 of [7] it shown that the Lagrangian systems considered in the CL method correspond to a special selection of the aforementioned matrix. In [13], see also [7], the CL method is extended via the inclusion of external forces into the closed-loop Lagrangian system, rendering the CL method *equivalent* to IDA-PBC, both methods requiring the solution of the same PDEs. Since both methods are equivalent, in the sequel we will restrict our attention to IDA-PBC.

Caveat emptor: Because of space constraints all proofs of the claims are omitted. The interested reader is referred to the papers where the proofs are given.

Notation: Unless indicated otherwise, all vectors in the paper are *column* vectors. For $x \in \mathbb{R}^n$, $S \in \mathbb{R}^{n\times n}$, $S = S^\top > 0$, we denote $|x|^2 := x^\top x$ and $\|x\|_S^2 := x^\top S x$. Given $n, m \in \mathbb{N}$, we let I_n denote the $n \times n$ identity matrix, $0_{n\times m}$ the $n \times m$ matrix of zeros and $e_i \in \mathbb{R}^n$ the i–th Euclidean basis vector of $\mathbb{R}^n$. Given $A \in \mathbb{R}^{n\times m}$, we let $(A)_{ij}$, $(A)_j$ and $(A)^i$ denote the ijth element, jth column, and ith row of A, respectively. To simplify the expressions, the arguments of all mappings—that are assumed smooth—will be explicitly written only the first time that the mapping is defined. For a scalar function $H\colon \mathbb{R}^n \to \mathbb{R}$, we define $\nabla_{x_i} H := \frac{\partial H}{\partial x_i}$ and $\nabla_x H := \left(\frac{\partial H}{\partial x}\right)^\top$—when clear from the context the subindex in ∇ will be omitted.

7.2 Basic IDA-PBC

7.2.1 Design Procedure

As indicated in the previous section IDA-PBC was introduced in [52] to control underactuated mechanical systems described in port-Hamiltonian (pH) form by

$$\Sigma : \begin{bmatrix} \dot q \\ \dot p \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & I_n \\ -I_n & 0_{n\times n} \end{bmatrix} \nabla H(q, p) + \begin{bmatrix} 0_{n\times m} \\ G(q) \end{bmatrix} u, \tag{7.1}$$

where $q, p \in \mathbb{R}^n$ are the generalized position and momenta, respectively, $u \in \mathbb{R}^m$ is the control, $G\colon \mathbb{R}^n \to \mathbb{R}^{n\times m}$ with $\mathrm{rank}(G) = m < n$, the function $H\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$H(q, p) := \frac{1}{2} p^\top M^{-1}(q) p + V(q) \tag{7.2}$$

is the total energy with $M\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$, the positive definite inertia matrix and $V\colon \mathbb{R}^n \to \mathbb{R}$ the potential energy. The control objective is to generate a state feedback control that assigns to the closed-loop the stable equilibrium $(q,p) = (q^\star, 0)$, $q^\star \in \mathbb{R}^n$. This is achieved in IDA-PBC via a two-step procedure. The first one, called energy shaping, determines a state feedback to match the pH target dynamics

$$\Sigma_d : \begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & M^{-1}(q)\, M_d(q) \\ -M_d(q)\, M^{-1}(q) & J_2(q,p) \end{bmatrix} \nabla H_d(q,p) \tag{7.3}$$

with the new total energy function $H_d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$,

$$H_d(q,p) := \frac{1}{2} p^\top M_d^{-1}(q)\, p + V_d(q), \tag{7.4}$$

where $M_d\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is positive definite, $V_d\colon \mathbb{R}^n \to \mathbb{R}$ verifies

$$q_\star = \arg\min V_d(q), \tag{7.5}$$

and $J_2\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ fulfills the skew-symmetric condition

$$J_2(q,p) = -J_2^\top(q,p).$$

The forces associated with this matrix are the gyroscopic forces mentioned in Sect. 7.1.4, which distinguish IDA-PBC from the original CL method proposed in [8].

It is easy to see that $(q^\star, 0)$ is a stable equilibrium point of (7.3) with Lyapunov function H_d. To determine the energy-shaping term of the control we equate the right-hand sides of (7.1) and (7.3) to obtain the so-called matching equations

$$\nabla_q H - G\, u = M_d\, M^{-1}\, \nabla_q H_d - J_2\, M_d^{-1}\, p.$$

As shown in [52] these equations are equivalent to the solution of the kinetic energy (KE) PDE

$$G^\perp \left\{ M_d\, M^{-1}\, \nabla_q (p^\top M_d^{-1} p) - 2\, J_2\, M_d^{-1}\, p \right\} = -G^\perp \nabla_q (p^\top M^{-1} p), \tag{7.6}$$

the potential (PE) PDE

$$G^\perp \{ \nabla V - M_d\, M^{-1}\, \nabla V_d \} = 0_s, \tag{7.7}$$

and the (univocally defined) control

$$u_{\mathtt{ES}} = (G^\top G)^{-1}\, G^\top \left[\nabla_q H - M_d\, M^{-1}\, \nabla_q H_d + J_2\, M_d^{-1}\, p \right],$$

where $G^{\perp}\colon \mathbb{R}^n \to \mathbb{R}^{s\times n},\ s := n - m$ is a full rank left annihilator of G, i.e., $G^{\perp}G = 0_{s\times m}$ and $\operatorname{rank}(G^{\perp}) = s$. It is interesting to note that *all* matrices $G^{\perp}$ can be generated as $G^{\perp} = UU_2$, where $U \in \mathbb{R}^{s\times s}$ is an *arbitrary* full rank matrix and $U_2 \in \mathbb{R}^{n\times s}$ is determined by the singular value decomposition of G as

$$G = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma & 0_{m\times s} \\ 0_{s\times m} & 0_{s\times s} \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{\top}.$$

The second step, called damping injection, is aimed at achieving asymptotic stability. This step is carried out feeding back the natural passive output, i.e., adding to the energy shaping control a term of the form

$$u_{\mathtt{DI}} := -K_P G^{\top} M_d^{-1} p,$$

with $K_P \in \mathbb{R}^{m\times m}$ positive definite. With this new term we get

$$\dot{H}_d = -\|G^{\top} M_d^{-1} p\|_{K_P}^2 \leq 0.$$

Asymptotic stability follows if the output $G^{\top} M_d^{-1} p$ is detectable [66]. The overall control signal, then, is given as $u = u_{\mathtt{ES}} + u_{\mathtt{DI}}$.

7.2.2 A Formula for the Gyroscopic Forces and the Number of KE–PDEs

The success of IDA-PBC relies on the possibility of solving the PDEs (7.6) and (7.7). In this subsection we concentrate our attention on the KE-PDE that, as discussed in Sect. 7.1.4, is simplified with the inclusion of gyroscopic forces, i.e., the free matrix J_2. In [15] a compact representation of the KE–PDE and an explicit expression for J_2 are obtained as follows. First, note that to be consistent with (7.6), whose remaining terms are quadratic in p, the free matrix J_2 *must be linear* in p. Hence, without loss of generality we can take J_2 of the form

$$J_2(q,p) = \sum_{i=1}^{n} e_i^{\top} M_d^{-1} p\, S_i(q), \tag{7.8}$$

where $S_i\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ verify $S_i(q) = -S_i^{\top}(q)$. To streamline the presentation of the result of [15] we denote the columns of $G^{\perp}$ as

$$G^{\perp}(q) =: \begin{bmatrix} v_1^{\top}(q) \\ \vdots \\ v_s^{\top}(q) \end{bmatrix},$$

where $v_k : \mathbb{R}^n \to \mathbb{R}^n,\ k \in \bar{s} := \{1, \dots, s\}$ is given by $v_k := \mathrm{col}(v_{ki})$. Also, we introduce the mappings

$$A_k : \mathbb{R}^n \to \mathbb{R}^{n\times n},\ B_k : \mathbb{R}^n \to \mathbb{R}^{n\times n},\ \Gamma_{kj} : \mathbb{R}^n \to \mathbb{R},\ W_k : \mathbb{R}^n \to \mathbb{R}^{n\times n}.$$

as

$$A_k := M_d \left(\sum_{i=1}^{n} v_{ki} \nabla_{q_i} M^{-1} \right) M_d,\ B_k := M_d \left(\sum_{i=1}^{n} \Gamma_{ki} \nabla_{q_i} M_d^{-1} \right) M_d,\ k \in \bar{s}$$
$$\Gamma_{kj} := \sum_{i=1}^{n} v_{ki} (M_d M^{-1})_{ij},\ W_k := \begin{bmatrix} v_k^\top S_1 \\ \vdots \\ v_k^\top S_n \end{bmatrix} + \begin{bmatrix} v_k^\top S_1 \\ \vdots \\ v_k^\top S_n \end{bmatrix}^\top,\ k \in \bar{s}, j \in \bar{n} := \{1, \dots, n\}.$$

The proof of the Proposition below is given in [15].

Proposition 7.1 *The KE–PDE (7.6) is equivalent to the PDEs*

$$B_k(q) - A_k(q) = W_k(q),\ k \in \bar{s}. \tag{7.9}$$

□□□

Note that the left-hand side of (7.9) is a function of the unknown matrix M_d (and partial derivatives of its components), while the right-hand side of (7.9) is *independent* of the unknown matrix M_d (and partial derivatives of its components). Hence the number of free elements on the right-hand side of (7.9) entirely determines the number of KE-PDE's to be solved. It is shown in [15] that this number equals

$$\frac{1}{6} s (s+1)(s+2), \tag{7.10}$$

which coincides with the number reported in [11], see also [12], where it is proposed to consider other forces—besides the gyroscopic ones captured by J_2. It is also important to underscore that (7.9) gives an *explicit formula* for J_2, that was presented in a different form in [1]. It should be mentioned that, even though the number of PDEs to be solved remains unaffected, the inclusion of more general type of forces proposed in [11] effectively extends the realm of application of IDA-PBC. This issue has been elaborated in [12, 21], see also [26] where a far more general method is proposed.

7.2.3 Solving the Matching Equations

A lot of research effort has been devoted to the solution of the matching Eqs. (7.6) and (7.7). In [8] the authors give a series of conditions on the system and the assignable inertia matrices such that the PDEs can be solved. However, as pointed out in the previous subsection these "solutions" lead to negative definite inertia matrices. Analytical techniques to solve the PDEs have been reported in [6, 7] and some geometric aspects of the equations are investigated in [37].

The case of underactuation degree one systems, i.e., $s = 1$, has been studied in detail in [1, 5]. In [1] it was proved that, if the inertia matrix and the force induced by the potential energy (on the unactuated coordinate) are independent of the unactuated coordinate, then the PDEs can be *explicitly solved*. In [39] explicit solutions are also given for a class of two degrees-of-freedom systems, that includes the interesting Acrobot example. It is important to underscore that the IDA-PBC reported in [39] ensures asymptotic stability of the upward Acrobot position with a domain of attraction including a region in the lower half plane. That is, the IDA-PBC can swing up the Acrobot *without switching*. To the best of the authors' knowledge this is the first such result for *any* pendular system.

Particularly troublesome is the PDE associated to the kinetic energy which is nonlinear and *nonhomogeneous* and the solution, that defines the desired inertia matrix, must be positive definite. In [68] it is shown that we can eliminate or simplify the forcing term $G^{\perp}\nabla_q(p^{\top}M^{-1}p)$ in this PDE modifying the target dynamics and introducing a change of coordinates in the original system. In the paper the examples of pendulum on a cart and Furuta's pendulum are used to illustrate the results. Furthermore, it is shown that, in the particular case of transformation to the Lagrangian coordinates, it is possible to simplify the PDEs *if and only if* the Coriolis and centrifugal forces of the system enter into the kernel of the input matrix—see Sect. 7.3.3 where this assumption is invoked to design a robust IDA-PBC for underactuated systems.

7.3 Disturbance Rejection of IDA-PBC via Nonlinear PID

It is widely recognized that IDA-PBC designs are robust against parameter uncertainties and unmodelled dynamics, e.g., passive effects like friction. However, the (unavoidable) presence of external disturbances degrades its performance, shifting the equilibrium of the closed-loop and, possibly, inducing instability. For this reason the problem of robustification of IDA-PBC *vis-à-vis* external disturbances is of primary importance. In this section we recall some results that have been reported to address this problem. Not surprisingly the proposed answer is the addition of an outer-loop integral action. However, the nonlinear nature of the problem makes nonobvious the choice of the integral action.

7.3.1 Integral Action Around the Passive Output

To broach the subject let us start by recalling a well-known result of disturbance rejection for general pH systems reported in [47].

Proposition 7.2 *Consider the perturbed pH system*

$$\begin{aligned} \dot{x} &= F(x)\nabla H(x) + g(x)(u + d_2) \\ y &= g^\top(x)\nabla H(x) \end{aligned} \tag{7.11}$$

where $x \in \mathbb{R}^{n_x}, u \in \mathbb{R}^m$, $g : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x \times m}$ is the full rank input matrix, $d_2 \in \mathbb{R}^m$ is a constant disturbance the matrix F is such that $F(x) + F^\top(x) \leq 0$, and $H : \mathbb{R}^{n_x} \to \mathbb{R}$ is the energy function verifying

$$x^\star = \arg\min H(x).$$

Introduce an integral control around the passive output y as

$$\begin{aligned} \dot{\eta} &= K_I y \\ u &= -\eta, \end{aligned} \tag{7.12}$$

where $K_I > 0$ is an arbitrary tuning gain.

(i) The equilibrium $(x^\star, d_2)$ is stable.
(ii) There exists a (closed) ball, centered in $(x^\star, d_2)$ such that for all initial states $(x(0), \eta(0)) \in \mathbb{R}^n \times \mathbb{R}^m$ inside the ball the trajectories are bounded and $\lim_{t\to\infty} y(t) = 0$.
(iii) If, moreover, y is a detectable output for the closed-loop system (7.11), (7.12), the equilibrium is asymptotically stable.

The properties (i)–(iii) are global if $H(x)$ is positive definite and radially unbounded. □□□

The following remarks are in order.

- The disturbance is *matched*, i.e., it enters in the image of the input matrix g.
- The integral control only ensures that $y(t) \to 0$ and an additional detectability requirement is needed to ensure $x(t) \to x^\star$.

Surprisingly, the construction above fails for mechanical systems—even for fully actuated ones, i.e., when $m = n$ and $G = I_n$, contradicting the claim of [16]. Indeed, in the case of full actuation, there is no need in IDA-PBC to shape the kinetic energy and we can take $M_d = M$ and $J_2 = 0$. Consequently, applying the IDA-PBC controller to (7.1) with an additional input[2] yields the closed-loop system

$$\begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & I_n \\ -I_n & -K_P \end{bmatrix} \nabla H_d(q,p) + \begin{bmatrix} 0_{n\times m} \\ I_n \end{bmatrix} u + \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \tag{7.13}$$

where we have added d_1 and $d_2 \in \mathbb{R}^n$, which are the matched and unmatched unmeasurable disturbances—possibly time–varying, but bounded. Note that $(q^\star, 0)$ is an

[2] To avoid cluttering the notation we call this additional signal also u.

asymptotically stable equilibrium of (7.13) when $d_1 = 0$ and $d_2 = 0$. As explained in Remark 1 of [57] these external signals may represent external forces or an input measurement bias.

The addition of an integral action on the passive output, i.e., the velocities $\dot{q} = M^{-1}(q)p$, via

$$\begin{aligned} u &= -\eta \\ \dot{\eta} &= K_I M^{-1}(q)p \end{aligned}$$

sufers from the following drawbacks. If d_1 is a nonzero constant the system admits no constant equilibrium, and if $d_1 = 0$ and d_2 is constant there is an equilibrium set given by

$$\mathscr{E} = \Big\{(q, p, \eta) \,|\, p = 0,\ \nabla V(q) + \eta = d_2\Big\}.$$

Moreover, it is easy to see that, if $d_1 = 0$, the foliation

$$\mathscr{M}_\kappa = \Big\{(q, p, \eta) \,|\, K_I q - \eta = \kappa,\ \kappa \in \mathbb{R}^n\Big\},$$

is *invariant* with respect to the flow of the closed-loop system. Consequently, convergence to the desired equilibrium $(q^\star, 0, d_2)$ is attained only for a zero measure set of initial conditions.

Of course, from Proposition 7.2 we have boundedness of trajectories and stability of the equilibrium, however, the detectability requirement (iii) fails.

7.3.2 *Nonlinear PI and PID for* Fully Actuated *Mechanical Systems*

From the discussion above, it is clear that a more sophisticated approach is required to reject the disturbances in mechanical systems. This problem was addressed in [57] where we mimic the construction of the pioneering work [19], extended for general pH systems in [48].

Interestingly, the resulting controllers are, in general, nonlinear PIDs of the form

$$\begin{aligned} u &= -\mathscr{K}_{P1}(q)\nabla V_d - \mathscr{K}_{P2}(q)p - \eta_1 - \eta_2 - \mathscr{K}_D(q)\frac{d\nabla V_d}{dt} \\ \dot{\eta}_1 &= \mathscr{K}_{I1}(q)\nabla V_d \\ \dot{\eta}_2 &= \mathscr{K}_{I2}(q)p, \end{aligned} \tag{7.14}$$

with some suitably defined nonlinear gains $\mathscr{K}_{Pi}, \mathscr{K}_{Ii}, \mathscr{K}_D : \mathbb{R}^n \to \mathbb{R}^{n\times n},\ i = 1, 2$. It should be underscored that one of the proportional, integral and derivative terms are created feeding back the gravity forces ∇V_d, while the other ones are done feeding back momenta. Notice that in the simplest case when V_d is quadratic, i.e.,

$$V_d(q) = \|q - q^\star\|_K^2,$$

with $K > 0$, then u is a standard PID around the *position error* $q - q^\star$. This is the case if the system (7.1) is linear, whence the gains $\mathscr{K}_{Pi}, \mathscr{K}_D$ and $\mathscr{K}_{Ii}$ are constant, while the use of ∇V_d and nonlinear gains is necessary in the nonlinear case.

An important feature of the controllers proposed in [57] is that, similarly, to the simple addition of integral action on the passive output discussed in Proposition 7.2, the pH structure is preserved in closed-loop—in some suitably defined coordinates.

The simplest scenario considered in [57] is for *constant inertia matrix M* and constant disturbances, when a *LTI* PI (around ∇V_d) does the job as indicated below.

Proposition 7.3 *Consider the system (7.13) with* constant *inertia matrix M and* constant *disturbances* (d_1, d_2) *in closed-loop with the PI control*

$$\begin{aligned} u &= -K_P z_3 - MK_I \nabla V_d \\ \dot{z}_3 &= K_I \nabla V_d, \end{aligned}$$

with $K_P > 0$ *the damping injection gain and* $K_I > 0$.

(i) The closed-loop dynamics expressed in the coordinates $z = col(z_1, z_2, z_3)$ *with*

$$\begin{aligned} z_1 &= q \\ z_2 &= p + M(z_3 - K_P^{-1} d_2), \end{aligned} \tag{7.15}$$

takes the pH form

$$\dot{z} = \begin{bmatrix} 0_{n\times n} & I_n & -K_I \\ -I_n & -K_P & 0_{n\times n} \\ K_I & 0_{n\times n} & 0_{n\times n} \end{bmatrix} \nabla H_z(z),$$

with energy function

$$H_z(z) := H(z_1, z_2) + \frac{1}{2}\|z_3 - z_3^*\|_{K_I^{-1}}, \tag{7.16}$$

where $z_3^* := d_1 + K_P^{-1} d_2$.

(ii) The desired equilibrium point $z^\star := (q^\star, 0, z_3^*)$, *is* asymptotically stable. *The stability is almost global if* $V_d(z_1)$ *is proper and has a unique minimum.* □□□

One of the main, and rather intriguing, ideas of [19] is the way the closed-loop energy function is constructed. Indeed, the first right-hand side term of (7.16) is given by

$$H(z_1, z_2) = \frac{1}{2} z_2^\top M^{-1}(z_1) z_2 + V(z_1),$$

that is the *evaluation* of the function H given in (7.2) with the replacement $(q,p) \leftarrow (z_1,z_2)$. That is, of course, different from the composition of the function H with the change of coordinates defined in (7.15).

For time-varying disturbances we can establish an *input-to-state stability* (ISS) property. Towards this end, the inclusion of a derivative term in ∇V_d is needed—in this case, with a constant derivative gain.[3]

Proposition 7.4 *Consider the system (7.13) with* constant *mass matrix M and* time–varying *disturbances $d(t) := col(d_1(t), d_2(t))$, in closed-loop with the PID control law*

$$
\begin{aligned}
u &= -\Big(K_3R_3\Big)p - K_4\nabla V - K_5 z_3 - k_D\nabla^2 V M^{-1} p\\
\dot{z}_3 &= (M^{-1} + k_D R_3)\nabla V + R_3 p,
\end{aligned}
$$

where k_D is a positive constant, $K_3 > 0$ and $R_3 > 0$ and

$$
\begin{aligned}
K_4 &:= k_D K_P M^{-1} + k_D K_3 R_3 + K_3 M^{-1}\\
K_5 &:= \Big(K_P M^{-1} + M R_3\Big)K_3.
\end{aligned}
$$

(i) The closed-loop dynamics expressed in the coordinates $z = col(z_1, z_2, z_3)$ with

$$
\begin{aligned}
z_1 &= q\\
z_2 &= p + k_1\nabla V(q) + K_3 z_3,
\end{aligned}
$$

takes the perturbed pH form

$$
\dot{z} = \begin{bmatrix} -k_1M^{-1} & I_n & -M^{-1}\\ -I_n & -K_P & -MR_3\\ M^{-1} & R_3M & -R_3 \end{bmatrix}\nabla H_z + \begin{bmatrix} I_n & 0_{n\times n}\\ k_1\nabla^2 V(z_1) & I_n\\ 0_{n\times n} & 0_{n\times n} \end{bmatrix} d(t), \qquad (7.17)
$$

with new Hamiltonian[4] $H_z(z) = H(z_1, z_2) + \frac{1}{2}\|z_3\|_{K_3}$.

(ii) If the potential energy function V is strictly convex with bounded Hessian, then (7.17) is ISS *with respect to the time varying input disturbances $(d_1(t), d_2(t))$ with ISS Lyapunov function $H_z(z)$.*

(iii) If $d_1 = 0$ and d_2 is constant, then the desired equilibrium $z^\star := (q^*, 0, K_5^{-1}d_2)$ is asymptotically stable. □□□

When M is *not constant* it is still possible to robustly the IDA-PBC with nonlinear PIDs, but the expressions for the controller gains become quite involved, as shown below.

[3] Recall that $\frac{d\nabla V_d}{dt} = \nabla^2 V_d M^{-1} p$.

[4] To avoid cluttering we use the same symbol to denote the energy function in all cases.

Proposition 7.5 *Consider the system (7.13) under the action of unmatched and matched time-varying disturbances $d_1(t)$ and $d_2(t)$, in closed-loop with the control law*

$$\begin{aligned} u &= -k_1 K_P M^{-1}\nabla V - k_1 \nabla^2 V M^{-1} p - K_3\Big[(M^{-1} + k_1 R_3)\nabla V + R_3 p\Big] \\ &-\Big[\frac{1}{2}\sum_{i=1}^{n} e_i p^\top \nabla_{q_i} M^{-1} + K_P M^{-1} + F_{23}^\top\Big] K_3 z_3 - \frac{1}{k_1}\Big[I_n + F_{12}^\top\Big] M F_{12} M^{-1} K_3 z_3 + v(q,p) \\ \dot{z}_3 &= \Big[M^{-1} + k_1 R_3\Big]\nabla V + R_3 p \end{aligned}$$

where k_1 is a positive constant, $K_3 > 0$, the mappings $F_{12}, F_{23} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ given by

$$\begin{aligned} F_{12}(q,p) &:= -\frac{k_1}{2} M^{-1} \sum_{i=1}^{n} e_i \Big[p + k_1 \nabla V + K_3 z_3\Big]^\top M^{-1} \nabla_{q_i} M - I_n \\ F_{23}(q,p) &:= -\frac{1}{k_1} F_{12} + R_3 M. \end{aligned}$$

and the mapping $v : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ given as

$$\begin{aligned} v(q,p) := \frac{k_1}{2} &\sum_{i=1}^{n} e_i p^\top M^{-1} \nabla_{qi} M M^{-1} \nabla V - \Big(M F_{12} M^{-1} + F_{12}^\top\Big)\nabla V \\ &- \frac{1}{k_1} F_{12}^\top M F_{12} M^{-1}\Big[p + k_1 \nabla V\Big], \end{aligned}$$

(i) The closed-loop dynamics expressed in the coordinates $z = col(z_1, z_2, z_3)$ with

$$\begin{aligned} z_1 &= q \\ z_2 &= p + k_1 \nabla V(q) + K_3 z_3, \end{aligned}$$

takes the perturbed pH form

$$\dot{z} = \begin{bmatrix} -k_1 M^{-1} & F_{12} & -M^{-1} \\ -F_{12}^\top & -K_P & -F_{23}^\top \\ M^{-\top} & F_{23} & -R_3 \end{bmatrix} \nabla H_z + \begin{bmatrix} I_n & 0_{n\times n} \\ k_1 \nabla^2 V(z_1) & I_n \\ 0_{n\times n} & 0_{n\times n} \end{bmatrix} \begin{bmatrix} d_1(t) \\ d_2(t) \end{bmatrix} \tag{7.18}$$

with $H_z(z) = H(z_1, z_2) + \frac{1}{2}\|z_3\|_{K_3}$.

(ii) The closed-loop system is ISS with respect to the disturbances $(d_1(t), d_2(t))$, provided that the Hessian of the potential energy satisfies condition (ii) in Proposition 7.4.

(iii) The unperturbed *system (7.18) has an asymptotically stable equilibrium at the desired state $z^\star = (q^*, 0, 0)$.*

□□□

In [57] other variations of these PID controllers, that yield simpler expressions for some particular cases are presented. The interested reader is referred to this paper for further details.

7.3.3 Nonlinear PI and PID for Underactuated *Mechanical Systems*

Extending the previous robustification results to the case of underactuated systems is a very challenging problem. Besides the intrinsic difficulty introduced by the underactuation on the outer-loop control action, we should take into account that in this case it is not possible anymore to do an IDA-PBC design with potential energy shaping only. Since the kinetic energy also needs to be changed we cannot take $M_d = M$ and $J_2 = 0$ as in the fully–actuated case. Consequently, the perturbed system that results from the application of IDA-PBC is now of the form

$$\begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & M^{-1}M_d \\ -M_d M^{-1} & J_2 - GK_P G^\top \end{bmatrix} \nabla H_d + \begin{bmatrix} 0_{n\times m} \\ G \end{bmatrix} (u+d), \tag{7.19}$$

with H_d as in (7.4) and $d \in \mathbb{R}^m$. Notice that, in contrast to the fully actuated case, only matched disturbances are considered that we, furthermore, assume are *constant*.

To the best of the authors' knowledge, the first attempt to solve the constant disturbance rejection problem for *underactuated* mechanical systems was published in [60]. The authors consider the simplest case of 2DOF mechanical systems with constant mass matrix and underactuation degree one. Although the main idea is interesting, it is shown in [22] that there are several unfortunate errors that invalidate the claims.

In this subsection we briefly recall some of the recent results of [22] where a class of mechanical systems for which the problem is solvable has been identified via the Assumption 7.1. Interestingly, though not surprisingly, the resulting controllers are again nonlinear PIDs of the form (7.14).

Assumption 7.1 The input matrix G and the desired mass matrix M_d are *constant* and the mass matrix $M(q)$ is independent of the non–actuated coordinates. Consequently,

$$G^\perp \nabla_q (p^\top M^{-1} p) = 0_{s\times 1}.$$

The term $G^\perp \nabla_q (p^\top M^{-1} p)$ appears in the KE-PDE (7.7) as a forcing term that makes it nonhomogeneous and introduces a quadratic term in the unknown M_d rendering very difficult its solution. As explained in Sect. 2.3 in [1] it is also assumed to be zero to provide an explicit solution of the PDE, while in [68] changes of coordinates are introduced to eliminate, or simplify, this term.

Proposition 7.6 *Consider the system (7.19), verifying Assumption 7.1, in closed-loop with the PID controller*

$$\begin{aligned} u = &-\Big[K_p G^\top M_d^{-1} G K_1 G^\top M^{-1} + K_1 G^\top \frac{dM^{-1}}{dt} + K_2 K_I\Big(K_2^\top + K_3^\top G^\top M_d^{-1} G K_1\Big) \\ &\times G^\top M^{-1}\Big]\nabla V_d - \Big[K_1 G^\top M^{-1}\nabla^2 V_d M^{-1} + (G^\top G)^{-1} G^\top J_2 M_d^{-1} \\ &+ K_2 K_I K_3^\top G^\top M_d^{-1}\Big]p - \Big(K_P G^\top M_d^{-1} G K_2 + K_3\Big) K_I z_3 \\ \dot z_3 = &\Big(K_2^\top G^\top M^{-1} + K_3^\top G^\top M_d^{-1} G K_1 G^\top M^{-1}\Big)\nabla V_d + K_3^\top G^\top M_d^{-1} p, \end{aligned}$$

where $K_1 > 0$, $K_I > 0$, $K_3 > 0$ *and*

$$K_2 := (G^\top M_d^{-1} G)^{-1}.$$

(i) The closed-loop dynamics in the coordinates $z = col(z_1, z_2, z_3)$ *with*

$$\begin{aligned} z_1 &= q \\ z_2 &= p + G K_1 G^\top M^{-1}\nabla V_d(q) + G K_2 K_I (z_3 - z_3^*), \end{aligned}$$

with $z_3^* := K_I^{-1}(K_P + K_3)^{-1} d$, *can be written in pH form as follows*

$$\begin{bmatrix} \dot z_1 \\ \dot z_2 \\ \dot z_3 \end{bmatrix} = \begin{bmatrix} -R_{11} & M^{-1} M_d & -F_{13} \\ -M_d M^{-1} & -G K_P G^\top & -G K_3 \\ F_{13}^\top & K_3^\top G^\top & -K_3 \end{bmatrix} \nabla H_z \tag{7.20}$$

with Hamiltonian

$$H_z(z) = \frac{1}{2} z_2^\top M_d^{-1} z_2 + V_d(z_1) + \frac{1}{2}\|z_3 - z_3^*\|_{K_I}^2,$$

and the mappings $R_{11} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ *and* $F_{13} : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ *given by*

$$R_{11}(q) := M^{-1} G K_1 G^\top M^{-1}, \; F_{13}(q) := M^{-1} G K_2.$$

(ii) The equilibrium $(q, p, z_3) = (q^\star, 0, z_3^*)$ *is* stable.

(iii) If the output

$$y_{D3} = \begin{bmatrix} G^\top M^{-1}\nabla V_d \\ G^\top M_d^{-1} z_2 \\ K_I(z_3 - z_3^*) \end{bmatrix}$$

is a detectable output of the dynamics (7.20), then $(q^\star, 0, z_3^*)$ *is an* asymptotically *stable equilibrium of the closed-loop.* □□□

In [22] two additional controllers, which are simplified versions of the one given in Proposition 7.6, are presented. These two controllers are obtained setting $(K_1, K_3) = (0_{m\times m}, 0_{m\times m})$ and $(K_2, K_3) = (0_{m\times m}, I_m)$. As seen in (7.20) these modifications still preserve the pH structure but eliminate some damping terms. Consequently, their corresponding detectability condition is strictly stronger than (iii) above, reducing the class of systems for which asymptotic stability is guaranteed. See [22] for further details.

7.4 Global Position Feedback Tracking

The IDA-PBC presented above, as well as the CL technique, assume that the full state is available for the controller design. As is well known, while the measurement of position is practically feasible, the one of velocity is complicated and sensitive to noise. Consequently, the design of speed observers and position feedback controllers is a problem of great practical importance that has attracted the attention of researchers for over 25 years—the reader is referred to [3, 58] for a recent list of references. The position feedback *regulation* problem was solved in [30] for fully actuated rigid manipulators and later extend to flexible joint ones in [2, 33]. However, the construction of a (smooth) controller that ensures, without velocity measurements, global *tracking* of position and velocity for all desired reference trajectories remained an open problem for a long time.

It should be mentioned that many *semi–global* results to the aforementioned position feedback tracking (PFT) problem have been reported. Semiglobal schemes intrinsically rely on high-gain injection to enlarge the domain of attraction, hence the interest in global controllers. A major contribution towards the solution of the PFT problem is due to [4], where invoking the Immersion and Invariance (I&I) techniques developed in [3], the first *globally exponentially stable* (GES) speed observer is reported—the result being applicable even for systems with nonholonomic constraints. While this contribution essentially solves the speed observation problem, the lack of a certainty equivalence principle in nonlinear systems, renders far from obvious the solution of the PFT problem. In [58] we provide the first solution to it. The design of [58] consists of the redesign of the speed observer of [4] and a new version of IDA-PBC, which combined in certainty equivalent form yields the desired result. The various components of this controller and the final result are described below.

7.4.1 A New Full State Feedback IDA-PBC

Since we are interested in the presence of Coulomb friction terms, we add a positive semidefinite matrix $\mathfrak{R} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ to the system (7.1) to get

$$\begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & I_n \\ -I_n & -\mathfrak{R}(q) \end{bmatrix} \nabla H(q,p) + \begin{bmatrix} 0_{n\times m} \\ I_n \end{bmatrix} u. \tag{7.21}$$

Notice that we are considering full actuation. Moreover, in contrast with most existing results, we *do not* assume that the inertia matrix is bounded from above.

The design of the new IDA-PBC proceeds in two steps. First, the change of coordinates in momenta proposed in [67] for observer design is used to assign a constant inertia matrix in the energy function. Second, the change of coordinates used in [57] to add integral actions to mechanical systems is combined with a new state feedback PBC to assign a pH structure with a desired energy function.

First, introduce the univocally defined, Cholesky factorization of the inverse inertia matrix

$$M^{-1}(q) = T(q)T^\top(q), \tag{7.22}$$

where $T : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is a lower triangular positive definite matrix. As shown in [67], defining the new momenta vector $\mathbf{p} := T^\top(q)p$, transforms (7.1) into

$$\begin{bmatrix} \dot{q} \\ \dot{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} 0 & T(q) \\ -T^\top(q) & S(q,\mathbf{p}) - R(q) \end{bmatrix} \nabla W(q,\mathbf{p}) + \begin{bmatrix} 0 \\ I_n \end{bmatrix} v, \tag{7.23}$$

with

$$v := T^\top(q)u, \; R(q) := T^\top(q)\mathfrak{R}(q)T(q) \tag{7.24}$$

the new control signal and dissipation matrix, respectively, $W : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

$$W(q,\mathbf{p}) = \frac{1}{2}|\mathbf{p}|^2 + V(q)$$

the new Hamiltonian function, and the the jk element of the skew-symmetric matrix $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ given by

$$S_{jk}(q,\mathbf{p}) = -\mathbf{p}^\top[(T)_j,(T)_k], \tag{7.25}$$

with $[\cdot,\cdot]$ the standard Lie bracket. See [4, 67] for its relationship with the Coriolis and centrifugal forces matrix of the Euler–Lagrange model.

Proposition 7.7 *Consider the pH system (7.23). Define the mapping* $v^\star : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n$

$$\begin{aligned} v^\star(q,\mathbf{p},t) = R(q)\mathbf{p} - \frac{d}{dt}(T^{-1}(q))R_1\tilde{q}(t) + \dot{\mathbf{p}}_d - S(q,\mathbf{p})\mathbf{p}_d(t) - T^\top(q) \\ \times\Big[\tilde{q}(t) - \nabla V(q)\Big] + \Big[S(q,\mathbf{p}) - R_2\Big]T^{-1}(q)R_1\tilde{q}(t) \\ -[T^{-1}(q)R_1T(q) + R_2]\tilde{\mathbf{p}}(t)) \end{aligned}$$

where we defined $\mathbf{p}_d := T^{-1}(q)\dot{q}_d$, *with* $q_d(t)$ *the reference trajectory,* $\tilde{q} := q - q_d$, $\tilde{p} := \mathbf{p} - \mathbf{p}_d$, *and* $R_1 > 0, R_2 > 0$ *free damping injection gains.*

(i) The closed-loop dynamics obtained setting $v = v^\star(q, \mathbf{p}, t)$ *expressed in the coordinates*

$$\begin{aligned} w_1 &= \tilde{q} \\ w_2 &= T^{-1}(q)R_1\tilde{q} + \tilde{p}, \end{aligned}$$

takes the pH form

$$\dot{w} = \begin{bmatrix} -R_1 & T(q) \\ -T^\top(q) & S(q, \mathbf{p}) - R_2 \end{bmatrix} \nabla H_w \tag{7.26}$$

with Hamiltonian function $H_w(w) = \frac{1}{2}|w|^2$.

(ii) The zero equilibrium point of (7.26) is UGES with Lyapunov function $H_w(w)$. *Consequently,* $(\tilde{q}(t), \tilde{p}(t)) \to 0$ *exponentially fast.*

□□□

Of course, there are many full-state feedback controllers ensuring exponential tracking [51]. The interest of the IDA-PBC presented above relies on the preservation of the pH structure and the addition of the *positive definite* damping matrices R_1, R_2—properties that are instrumental for the development of its position feedback version.

7.4.2 A New Exponentially Convergent I&I Momenta Observer

In [58] the exponentially convergent speed I&I observer reported in [4] is modified to estimate directly the (new) momenta **p**. An additional modification is introduced to take into account the presence of friction. Also, motivated by the developments in [61], we consider an alternative Lyapunov function for the stability analysis and add some degrees of freedom to robustify the observer design. The latter feature is essential for the proof of our main result. For the sake of brevity we do not repeat here all the observer equations but only state its existence. The interested reader is referred to [58] for further details.

Proposition 7.8 *Consider the mechanical system with friction (7.21), and assume u is such that trajectories exist for all* $t \geq 0$. *There exist smooth mappings.*

$$\begin{aligned} \mathbf{A} &: \mathbb{R}^{3n} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{3n+1} \\ \mathbf{B} &: \mathbb{R}^{3n} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \to \mathbb{R}^n \end{aligned}$$

such that the interconnection of (7.21) with

$$
\begin{aligned}
\dot{\mathrm{X}} &= \mathbf{A}(\mathrm{X}, q, v)\\
\hat{p} &= \mathbf{B}(\mathrm{X}, q),
\end{aligned}
\tag{7.27}
$$

ensures the existence of $\lambda > 0$ *such that*

$$
\lim_{t\to\infty} e^{\lambda t}[p(t) - \hat{p}(t)] = 0,
$$

for all initial conditions $(q(0), p(0), \mathrm{X}(0)) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{3n} \times \mathbb{R}_{\geq 0}$. □□□

7.4.3 A Uniformly GES Solution to the PFT Problem

In [58] it is shown that combining the IDA-PBC of Proposition 7.7 with the observer of Proposition 7.8 yields a solution to the PFT problem with the following properties:

- The closed-loop is uniformly GES that, via total stability arguments, ensures strong robustness properties.
- Only a lower bound on the inertia matrix is assumed. Hence, the result is applicable to a large class of mechanical systems, including robots with prismatic joints.
- The strong assumption of existence, exact knowledge and pervasiveness of friction is conspicuous by its absence. Instead, if friction is present, we assume it is known, treat it as a disturbance and compensate for it.[5]
- The stabilization mechanism does not rely on the use of (approximate) differentiators nor the injection of high gain into the loop. Indeed, although the proposed observer includes a dynamic scaling factor, that might take large values during the transients, it is shown to actually converge to one—hence, high-gain injection is not present in steady–state.

To the best of our knowledge, this is the first result enjoying these features reported in the literature. Again, for the sake of brevity we do not repeat here all the controller equations but only state its existence. The interested reader is referred to [58] for further details.

Proposition 7.9 *Consider the mechanical system with friction (7.21). Given any twice differentiable, bounded, reference trajectories* $q_d : \mathbb{R}_+ \to \mathbb{R}^n$*. There exist two (smooth) mappings*

$$
\begin{aligned}
\mathbf{F} &: \mathbb{R}^{3n} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^{3n+1}\\
\mathbf{H} &: \mathbb{R}^{3n} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n
\end{aligned}
$$

such that, for all initial conditions $(q(t_0), p(t_0), \varpi(t_0)) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{3n} \times \mathbb{R}_{\geq 0}$ *the system (7.1) in closed-loop with*

[5] See Sect. 7.5.1 for a robust adaptive version of the I&I momenta observer.

$$
\begin{aligned}
\dot{\varpi} &= \mathbf{F}(\varpi, q, t)\\
u &= \mathbf{H}(\varpi, q, t)
\end{aligned}
$$

verifies, for all $t \geq t_0 \geq 0$,

$$
\left\| \begin{bmatrix} q(t) - q_d(t) \\ p(t) - p_d(t) \\ \varpi(t) \end{bmatrix} \right\| \leq m \exp^{-\lambda(t-t_0)} \left\| \begin{bmatrix} q(t_0) - q_d(t_0) \\ p(t_0) - p_d(t_0) \\ \varpi(t_0) \end{bmatrix} \right\|,
$$

for some constants $m, \lambda > 0$ *(independent of* t_0*) with* $p_d := M(q)\dot{q}_d$. □□□

A transformation, similar to the one used in this section, has been presented in [17, 18], where a change of coordinates that removes the quadratic terms in velocity is found to solve the PFT problem for surface ships and mobile robots. After the publication of [58]—whose results were announced in its conference version in 2013—we became aware of [38] where a UGAS solution to the general PFT problem is reported. The controller is constructed adding an approximate differentiator to the classical full state feedback PD+ controller of [54]. An upper bound on the norm of the reference trajectories and their first and second order derivatives is imposed as a *lower bound* to the controller gains. Consequently, to track large and—or fast—varying references, high-gain and a "pure" differentiator are needed. Besides establishing only *asymptotic*, as opposed to *exponential*, stability, no friction is assumed to be present in the system and an upper bound in the inertia matrix is required.

7.5 Two Robust Adaptive Velocity Observers

The speed observer reported in Proposition 7.8 relies on the assumptions of known (or no) Coulomb friction and no disturbances. In [56] two extensions of this result were reported. First, a new globally convergent adaptive speed observer that, besides rejecting the disturbances, estimates some unknown friction coefficients for a class of mechanical systems that contains several practical examples. Second, the observer of Proposition 7.8 is robustified *vis-à-vis* constant disturbances. These two new results, which rely on the addition of (nonlinear) integral action similar to the one used in Sect. 7.3, are summarized in this section.

7.5.1 An Observer for a Class Systems with Unknown Friction and Disturbances

We consider the mechanical system

$$\begin{bmatrix} \dot{q} \\ \dot{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & I_n \\ -I_n & -\mathfrak{R} \end{bmatrix} \nabla H(q,\mathbf{p}) + \begin{bmatrix} 0_{n\times m} \\ G(q) \end{bmatrix} u + \begin{bmatrix} 0_{n\times m} \\ d \end{bmatrix}. \tag{7.28}$$

The system is subject to two different perturbations.

- *Unknown* constant disturbances $d \in \mathbb{R}^n$.
- Coulomb friction $\mathfrak{R} = \mathrm{diag}\{r_i\} \in \mathbb{R}^{n\times n}$, with *unknown constant* $r_i \geq 0,\ i \in \bar{n}$.

As customary in the observer literature, it is assumed that $u(t)$ is such that trajectories exist for all $t \geq 0$. The problem is to design a globally convergent *robust adaptive* observer for the transformed momenta $p := T^\top(q)\mathbf{p}$.[6] It should be noted that the presence of unknown friction coefficients generates products of *unmeasurable states* and *unknown parameters*, a situation for which very few results are available in the observer design literature—even for the case of linear systems.

Instrumental for the development of the adaptive observer is the change of coordinates introduced in Sect. 7.4.1. The following key assumption is made regarding the factor T.

Assumption 7.2 M^{-1} admits a factorization (7.22) with a factor T whose columns verify

$$\left[\Big(T(q)\Big)_i, \Big(T(q)\Big)_j\right] = 0,\ i,j \in \bar{n}. \tag{7.29}$$

It is well known that Assumption 7.2 is equivalent to the fact that the Riemann symbols of M are all zero.[7] It is clear from (7.23) and (7.25) that this assumption implies that the matrix of gyroscopic forces S is equal to zero. Another important observation is that the factorisation need not be equal to the Cholesky factorization. As shown in [56] the choice of T is an additional degree of freedom for the solution of the problem.

To design the robust adaptive observer, besides Assumption 7.2, a restriction on the friction coefficients is imposed. Namely, we decompose the friction matrix $\mathfrak{R}$ as $\mathfrak{R} = \mathfrak{R}_\mathbf{k} + \mathfrak{R}_\mathbf{u}$, where $\mathfrak{R}_\mathbf{k}, \mathfrak{R}_\mathbf{u}$ are $n \times n$ diagonal matrices containing the *known* and the *unknown* friction coefficients, respectively. Similarly, with an obvious definition, we decompose the transformed friction matrix (7.24) into $R(q) = R_k(q) + R_u(q)$.

To streamline the presentation all friction coefficients are grouped in a vector $r := \mathrm{col}(r_i) \in \mathbb{R}^n$ with the unknown and known coefficients in vectors $r_u \in \mathbb{R}^\ell$ and $r_k \in \mathbb{R}^{n-\ell}$, respectively. We also define a set of integers $\mathscr{N} \subset \bar{n}$ that contains the indices of the unknown coefficients of r. Finally, we define a matrix $C \in \mathbb{R}^{n\times\ell}$ such that $C^\top r = r_u$, where the matrix C verifies.

- rank $\{C\} = \ell$.
- For $j \in \mathscr{N}$, $(C)_j = e_{\ell_j}$, with ℓ_j the j-th element of $\mathscr{N}$.

Assumption 7.3 The i–th row of factor $T(q)$ is *independent of* q for $i \in \mathscr{N}$.

[6] The notation for the momenta is different from the one used in Sect. 7.4, but is consistent with the one used in [56].

[7] See Eqs. (6) and (7) of [67] for the definition of these symbols and the proof of this fact.

A consequence of Assumption 7.3 is the existence of *constant* matrices $Y_j \in \mathbb{R}^{n\times\ell}$, $j \in \bar{n}$, such that, for all vectors $a \in \mathbb{R}^n$ we have

$$R_u(q)a = (\sum_{i=1}^{n} Y_i a_i) r_u.$$

In Lemma 4.2 of [56] it is shown that

$$Y_j = \sum_{i=1}^{n} L_i e_j e_i^\top C,\ j \in \bar{n}, \tag{7.30}$$

with the *constant* matrices $L_i \in \mathbb{R}^{n\times n}$ defined as

$$L_i := T^\top(q) e_i e_i^\top T(q)\ ,\ i \in \bar{n}. \tag{7.31}$$

Proposition 7.10 *Consider the system (7.28) where the inertia matrix $M(q)$ and the friction matrix $\mathfrak{R}$ verify Assumptions 7.2 and 7.3. The $2n+\ell$ dimensional I&I adaptive momenta observer*

$$\begin{aligned}
\dot{p}_I &= -T^\top(q)[\nabla V - G(q)u - \hat{d}] - (\sum_{i=1}^{n} Y_i \hat{p}_i)\hat{r}_u - [\lambda Q(q) + R_k(q)]\hat{p}\\
\dot{r}_{u_I} &= (\sum_{i=1}^{n} Y_i^\top \hat{p}_i)(\dot{p}_I + \lambda \hat{p})\\
\dot{d}_I &= T(q)\hat{p}\\
\hat{p} &= p_I + \lambda Q(q),\ \hat{r}_u = r_{u_I} + \frac{1}{2\lambda}(\sum_{i=1}^{s} \hat{p}^\top L_i \hat{p}) e_i,\ \hat{d} = d_I + q
\end{aligned}$$

with the constant $n \times n$ matrices L_i given by (7.31), $Q(q)$ given by

$$\nabla Q(q) = T^{-1}(q),$$

$Y_i \in \mathbb{R}^{n\times\ell}$ given in (7.30) and $\lambda > 0$ a free parameter, ensures boundedness of all signals and

$$\lim_{t\to\infty}\{T^{-\top}(q(t))[\hat{p}(t) - p(t)]\} = 0$$

for all initial conditions $(q(0), p(0)) \in \mathbb{R}^n \times \mathbb{R}^n$. □□□

In [56] it is shown that the planar redundant manipulator with one elastic degree of freedom and the 2D-spider crane gantry cart satisfy the conditions of Proposition 7.10. Consequently, robust adaptive speed observation is possible for them. We note that for systems with *constant inertia matrix* Assumption 7.2 is trivially satisfied, because the factor T can be taken to be constant and Assumption 7.3 is also satisfied with $\mathscr{N} = \bar{n}$ and $C = I_n$, hence all friction coefficients can be identified.

7.5.2 A Robust Observer for Perturbed Systems with Known Friction

In [56] the I&I speed observer of Proposition 7.8 is redesigned to ensure its global convergence in spite of the presence of the *unknown* disturbances d and *known* friction forces in all coordinates.

Proposition 7.11 *Consider the system (7.28) with* known *friction matrix* $\mathfrak{R}$ *and* unknown disturbances d. *There exist smooth mappings*

$$\mathbf{C} : \mathbb{R}^{4n} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{4n+1}, \quad \mathbf{D} : \mathbb{R}^{4n} \times \mathbb{R}_{\geq 0} \times \mathbb{R}^n \to \mathbb{R}^n$$

such that the interconnection of (7.28) with

$$\dot{\mathrm{X}} = \mathbf{C}(\mathrm{X}, q, u), \quad \hat{p} = \mathbf{D}(\mathrm{X}, q),$$

ensures $\lim_{t\to\infty}[\hat{p}(t) - p(t)] = 0$, *for all initial conditions* $(q(0), p(0), \mathrm{X}(0)) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{4n} \times \mathbb{R}_{\geq 0}$. □□□

The construction of the observer above follows very closely the one of Proposition 7.8—first reported in [58]—with the only difference being the inclusion of an adaptation law for the unknown disturbance parameters d.

7.6 Constructive IDA-PBC: Shaping the Energy with a PID

As indicated in Sect. 7.2 to make the IDA-PBC method really constructive it is necessary to give explicit solutions to the PE-PDE (7.7) and the KE-PDE (7.9), which may be difficult to solve in applications. In this section we review some recent extensions of IDA-PBC where this step is obviated. Interested readers are referred to the interesting work of [42] where a dynamic version of IDA-PBC that does not require the solution of PDEs is proposed.

7.6.1 PID Control of [20]

A key feature of IDA-PBC and the CL methods is that the mechanical structure of the system is preserved in closed-loop, a condition that gives rise to the matching PDEs, which characterize the assignable Hamiltonian or Lagrangian functions, respectively. Recently in [20] it was proposed to *relax* this constraint, and concentrate our attention on the energy shaping objective only. That is, we look for a static state feedback that stabilizes the desired equilibrium assigning to the closed-loop a Lyapunov function of the same form as the energy function of the open-loop system but with new, desired inertia matrix and potential energy function. However, we *do not require* that the closed-loop system is a mechanical system with this Lyapunov function qualifying as its energy function. In this way, the need to solve the matching equations is avoided.

The controller design of [20] is carried out proceeding from a *Lagrangian representation* of the system and consists of four steps. First, the application of a (collocated) partial feedback linearization stage, *à la* [63]. Second, as done in [61], the identification of conditions on the inertia matrix and the potential energy function that ensure the Lagrangian structure is preserved. As a corollary of the Lagrangian structure preservation two new passive outputs are easily identified. Third, a PID controller around a suitable combination of these passive outputs is applied. Now, PID controllers define output strictly passive mappings with storage function the sum of the square of the PIDs input and the square of the *integrator state*—stemming from the integral action. Thus, the passivity theorem allows to immediately conclude output strict passivity—hence, $\mathscr{L}_2$–stability—of the closed-loop system with storage function the sum of the storage functions of the passive output and the PID. To achieve the aforementioned equilibrium stabilization objective a fourth step is required. Namely, to impose some integrability assumptions on the systems inertia matrix to ensure that the integral of the passive output, i.e., the integrator state, can be expressed as a function of the systems generalized coordinates and, consequently, can be added to the systems storage function to generate a *bona fide* Lyapunov function by ensuring it has a minimum at the desired position.

7.6.2 *Avoiding the Feedback Linearization Step*

As explained above the first step in the design procedure of [20] is the use of a partial linearizing state feedback that transforms the system into Spong's Normal Form [63]. It is widely recognized that feedback linearization, which involves the exact cancelation of nonlinear terms, is intrinsically non–robust. Interestingly, it has recently been shown in [59] that, for a class of systems strictly larger than the one considered in [20], it is possible to identify two new passive outputs *without* the feedback linearization step. The key modification is the introduction of a *change of coordinates* that, for systems verifying the assumption below, directly reveals the new cyclo-passive outputs around which the PID controller is added.

As done in Sect. 7.4.1 the first step is to introduce the change of coordinates $(q, \mathbf{p}) \mapsto (q, T^\top(q)p)$, where $T : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is a full rank factorization of the inverse inertia matrix, that is,

$$M^{-1}(q) = T(q)T^\top(q), \tag{7.32}$$

which transforms (7.1) into

$$\begin{bmatrix} \dot{q} \\ \dot{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} 0_{n\times n} & T(q) \\ -T^\top(q) & S(q,\mathbf{p}) \end{bmatrix} \nabla W(q,\mathbf{p}) + \begin{bmatrix} 0 \\ T^\top(q)G \end{bmatrix} u. \tag{7.33}$$

Notice that, in contrast with (7.23), the system above is underactuated and lossless.

Now, we introduce the key assumption.

Assumption 7.4 (i) The input matrix G is of the form

$$G = \begin{bmatrix} 0_{s\times m} \\ I_m \end{bmatrix}. \tag{7.34}$$

(ii) The potential energy can be written as

$$V(q) = V_a(q_a) + V_u(q_u),$$

where $q = \mathrm{col}(q_u, q_a)$, with $q_a \in \mathbb{R}^m$ and $q_u \in \mathbb{R}^s$, where $s := n - m$.

(iii) The inertia matrix depends only on the unactuated variables q_u.

(iv) The $(2,2)$ sub-block of the inertia matrix is constant.

The key observation is that Assumption 7.4 ensures the existence of a factorization (7.32) of the form

$$T(q_u) = \begin{bmatrix} T_1(q_u) & 0_{s\times m} \\ T_2(q_u) & T_3 \end{bmatrix}, \tag{7.35}$$

where $T_1 : \mathbb{R}^s \to \mathbb{R}^{s\times s}$, $T_2 : \mathbb{R}^s \to \mathbb{R}^{m\times s}$ and $T_3 \in \mathbb{R}^{m\times m}$ is *constant*.

To streamline the statement of the proposition below we introduce the partition $\mathbf{p} = \mathrm{col}(\mathbf{p}_u, \mathbf{p}_a)$ with $\mathbf{p}_u \in \mathbb{R}^s$ and $\mathbf{p}_a \in \mathbb{R}^m$.

Proposition 7.12 *Consider the underactuated mechanical system (7.33) satisfying Assumption 7.4 together with the inner-loop control*

$$u = \nabla V_a(q_a) + v. \tag{7.36}$$

Define the output signals

$$y_u := T_2(q_u)\mathbf{p}_u, \quad y_a := T_3\mathbf{p}_a.$$

The operators $v \mapsto y_u$ and $v \mapsto y_a$ are cyclo–passive *with storage functions*

$$H_u(q_u,\mathbf{p}_u) = \frac{1}{2}|\mathbf{p}_u|^2 + V_u(q_u),\ H_a(\mathbf{p}_a) = \frac{1}{2}|\mathbf{p}_a|^2, \tag{7.37}$$

respectively. More precisely,

$$\dot{H}_a = v^\top y_a\,; \quad \dot{H}_u = v^\top y_u, \tag{7.38}$$

7.6.3 PID Controller Design

Similarly to [20] the controller design is completed in [57] adding the PID

$$k_e v = -\left(K_P y_d + K_I \int_0^t y_d(s) ds + K_D \dot{y}_d \right), \tag{7.39}$$

with

$$y_d := k_a y_a + k_u y_u,$$

and a suitably chosen initial condition for the integral term. Some simple calculations, using (7.38) and (7.39), show that the function $L : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+ \mapsto \mathbb{R}$

$$L(q, \mathbf{p}, t) := k_e[k_a H_a(\mathbf{p}_a) + k_u H_u(q_u, \mathbf{p}_u)] + \frac{1}{2} \| \int_0^t y_d(s) ds \|_{K_I}^2 + \frac{1}{2} \|y_d\|_{K_D}^2,$$

verifies

$$\dot{L} \leq -\|y_d\|_{K_P}^2.$$

Therefore, it only remains to impose the assumptions required to ensure, on the one hand, the implementation (without singularities nor differentiation) of the control (7.39) and, on the other hand, guarantee the assignment of a suitable Lyapunov function. Towards this end, we impose the following assumption.

Assumption 7.5 Given Assumption 7.4, partition the inertia matrix as

$$M(q_u) = \begin{bmatrix} m_{uu}(q_u) & m_{au}^\top(q_u) \\ m_{au}(q_u) & m_{aa} \end{bmatrix},$$

where $m_{aa} \in \mathbb{R}^{m \times m}$, $m_{au} : \mathbb{R}^s \to \mathbb{R}^{s \times m}$, $m_{uu} : \mathbb{R}^n \to \mathbb{R}^{s \times s}$.

(i) The rows of m_{au} satisfy

$$\nabla (m_{au})^i = [\nabla (m_{au})^i]^\top, \ \forall i \in \bar{m}.$$

Equivalently, there exists a function $V_N : \mathbb{R}^s \to \mathbb{R}^m$ such that

$$\dot{V}_N = -m_{au}(q_u) \dot{q}_u.$$

(ii) There exist constants $k_e, k_a, k_u \in \mathbb{R}, \ K_D, K_I \in \mathbb{R}^{m \times m}, \ K_D, K_I \geq 0$, such that the following holds.

(a) $\det[K(q_u)] \neq 0, \ \forall q_u \in \mathbb{R}^s$, where $K : \mathbb{R}^s \to \mathbb{R}^{m \times m}$ is defined as

$$K(q_u) := k_e I_m + k_a K_D T_3 T_3^\top + k_u K_D T_2(q_u) T_2^\top(q_u). \tag{7.40}$$

(b) The matrix

$$M_d^{-1}(q_u)=\begin{bmatrix} k_u^2T_2^\top(q_u)K_DT_2(q_u)+k_ek_uI_s & k_ak_uT_2^\top(q_u)K_DT_3 \\ k_ak_uT_3^\top K_DT_2(q_u) & k_ek_aI_m+k_a^2T_3^\top K_DT_3 \end{bmatrix} \tag{7.41}$$

is *positive definite*.

(c) The function

$$V_d(q) := k_ek_uV_u(q_u)+\frac{1}{2}||k_aq_a+(k_u-k_a)V_N(q_u)||_{K_I}^2 \tag{7.42}$$

is proper and has an *isolated minimum* at q_*.

We are in position to present the main result of [57].

Proposition 7.13 *Consider the underactuated mechanical system (7.33) verifying Assumptions 7.4 and 7.5, together with the inner-loop controller (7.36) and the PID (7.39). The closed-loop system has a* globally stable *equilibrium at the desired point* $(q,\mathbf{p})=(q_\star,0)$ *with Lyapunov function*

$$H_d(q,\mathbf{p})=\frac{1}{2}\mathbf{p}^\top M_d^{-1}(q_u)\mathbf{p}+V_d(q). \tag{7.43}$$

with M_d *and* V_d *defined in (7.41) and (7.42), respectively. The equilibrium is GAS if the signal* y_d *is a* detectable *output for the closed-loop system.*

The following remarks are in order.

- The role of the tuning gains k_e, k_a, k_u and K_P, K_I in the energy shaping stage is clear from the expressions of M_d and V_d given in (7.41) and (7.42), respectively. It is important to highlight that there is no sign constraint on the scalar quantities, which gives a large flexibility to shape the energy functions.
- Condition (i) in Assumption 7.5 is imposed to be able to add the new term $\frac{1}{2}||k_aq_a+(k_u-k_a)V_N(q_u)||_{K_I}^2$ in the desired potential energy function. This motivates the name $V_N(q_u)$.
- Notice that the systems potential energy V_u appears now multiplied by k_ek_u, whose sign can be used to "flip" this function, as done for the cart-pendulum example in [57].

7.6.4 Tracking Constant Speed Trajectories

In [57] it is shown that the controller methodology presented in Proposition 7.13 can be directly extended to track constant speed trajectories in the actuated coordinates with constant positions in the underactuated ones. To formulate this problem we define the generalized coordinates errors as

$$\tilde{q}(t) = \begin{bmatrix} \tilde{q}_u(t) \\ \tilde{q}_a(t) \end{bmatrix} := \begin{bmatrix} q_u(t) - q_u^* \\ q_a(t) - rt \end{bmatrix}, \tag{7.44}$$

with $q_u^* \in \mathbb{R}^s$ and $r \in \mathbb{R}^m$ a *constant* vector. Consistent with the desired trajectories we define the errors in momenta as

$$\tilde{\mathbf{p}} = \begin{bmatrix} \tilde{\mathbf{p}}_u \\ \tilde{\mathbf{p}}_a \end{bmatrix} := \begin{bmatrix} \mathbf{p}_u \\ \mathbf{p}_a - T_3^{-1} r \end{bmatrix}. \tag{7.45}$$

The tracking objective is to ensure

$$\lim_{t\to\infty} \begin{bmatrix} \tilde{q}(t) \\ \tilde{\mathbf{p}}(t) \end{bmatrix} = 0, \tag{7.46}$$

and the main result is as follows.

Proposition 7.14 *Consider the underactuated mechanical system (7.33) satisfying Assumptions 7.4 and 7.5 together with*

$$u = \nabla V_a(q_a) + \mathbf{v}$$

and the PID control

$$k_e \mathbf{v} = -\left(K_P \mathbf{y}_d + K_I \int_0^t \mathbf{y}_d(s)ds + K_D \dot{\mathbf{y}}_d \right), \tag{7.47}$$

with

$$\mathbf{y}_d := k_a T_3 \tilde{\mathbf{p}}_a + k_u T_2(\tilde{q}_u + q_u^*)\tilde{\mathbf{p}}_u.$$

All trajectories of the closed-loop system are bounded, the zero equilibrium of the error system is stable and (7.46) is satisfied if $\mathbf{y}_d$ *is a* detectable *output for the closed-loop system.*

Acknowledgements R. Ortega is supported by Government of Russian Federation (grant 074-U01, GOSZADANIE 2014/190 (project 2118)), the Ministry of Education and Science of Russian Federation (project 14.Z50.31.0031). The work of J.G. Romero is supported by a public grant overseen by the French National Research Agency (ANR) as part of the Investissement d'Avenir program, through the iCODE Institute, research project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

References

1. Acosta, J.A., Ortega, R., Astolfi, A., Mahindrakar, A.M.: Interconnection and damping assignment passivity-based control of mechanical systems with underactuation degree one. IEEE Trans. Autom. Control **50**(12), 1936–1955 (2005)

2. Ailon, A., Ortega, R.: An observer-based controller for robot manipulators with flexible joints. Syst. Control Lett. **21**(4), 329–335 (1993)
3. Astolfi, A., Karagiannis, D., Ortega, R.: Nonlinear and Adaptive Control with Applications. Springer, Berlin (2007)
4. Astolfi, A., Ortega, R., Venkataraman, A.: A globally exponentially convergent immersion and invariance speed observer for mechanical systems with non-holonomic constraints. Automatica **46**(1), 182–189 (2010)
5. Auckly, D., Kapitanski, L.: On the λ-equations for matching control laws. SIAM J. Control Optim. **41**(5), 1372–1388 (2002)
6. Auckly, D., Kapitanski, L., White, W.: Control of nonlinear underactuated systems. Commun. Pure Appl. Math. **53**(3), 354–369 (2000)
7. Blankenstein, G., Ortega, R., van der Schaft, A.J.: The matching conditions of controlled Lagrangians and interconnection and damping assignment passivity based control. Int. J. Control **75**(9), 645–665 (2002)
8. Bloch, A., Leonard, N., Marsden, J.: Controlled Lagrangians and the stabilization of mechanical systems I: the first matching theorem. IEEE Trans. Autom. Control **45**(12), 2253–2270 (2000)
9. Bloch, A.M., Chang, D.E., Leonard, N., Marsden, J.E.: Controlled Lagrangians and the stabilization of mechanical systems II: potential shaping. IEEE Trans. Autom. Control **46**(10), 1556–1571 (2001)
10. Byrnes, C., Isisdori, A., Willems, J.C.: Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems. IEEE Trans. Autom. Control **36**(11), 1228–1240 (1991)
11. Chang, D.E.: Generalization of the IDA–PBC method for stabilization of mechanical systems. In: The 18th Mediterranean Conference on Control and Automation, pp. 226–230. Marrakech, Morocco (2010)
12. Chang, D.E.: On the method of interconnection and damping assignment passivity-based control for the stabilization of mechanical systems. Regular Chaotic Dyn. **19**(5), 556–575 (2014)
13. Chang, D.E., Bloch, A.M., Leonard, N.E., Marsden, J.E., Woolsey, C.A.: The equivalence of controlled Lagrangian and controlled Hamiltonian systems for simple mechanical systems. ESAIM: Control Optim. Calc. Var. **8**, 393–422 (2002)
14. Cisneros, R., Mancilla-David, F., Ortega, R.: Passivity-based control of a grid-connected small-scale windmill with limited control authority. IEEE J. Emerg. Sel. Top. Power Electr. **1**(4), 2168–6777 (2013)
15. Crasta, N., Ortega, R., Pillai, H.: On the matching equations of energy shaping controllers for mechanical systems. Int. J. Control **88**(9), 1757–1765 (2015)
16. Dirksz, D., Scherpen, J.: Power-based control: Canonical coordinate transformations, integral and adaptive control. Automatica **48**(6), 1045–1056 (2012)
17. Do, K.D., Pan, J.: Underactuated ships follow smooth paths with integral actions and without velocity measurements for feedback: theory and experiments. IEEE Trans. Control Syst. Technol. **14**(2), 308–322 (2006)
18. Do, K.D., Jiang, Z.P., Pan, J.: A global output-feedback controller for simultaneous tracking and stabilization of unicycle-type mobile robots. IEEE Trans. Robot. Autom. **20**(3), 589–594 (2004)
19. Donaire, A., Junco, S.: On the addition of integral action to port-controlled Hamiltonian systems. Automatica **45**(8), 1910–1916 (2009)
20. Donaire, A., Mehra, R., Ortega, R., Satpute, S., Romero, J.G., Kazi, F., Singh, N.M.: Shaping the energy of mechanical systems without solving partial differential equations. IEEE Trans. Autom. Control **61**(4), 1051–1056 (2016)
21. Donaire, A., Ortega, R., Romero, J.G.: Simultaneous interconnection and damping assignment passivity–based control of mechanical systems using dissipative forces. Syst. Control Lett. **94**, 118–126 (2016)

22. Donaire, A., Romero, J.G., Ortega, R., Siciliano, B., Crespo, M.: Robust IDA–PBC for underactuated mechanical systems subject to matched disturbances. Int. J.Robust Nonlinear Control (to appear) (2017)
23. Duindam, V., Macchelli, A., Stramigioli, S., Bruyninckx, H. (eds.): Modeling and Control of Complex Physical System: The Port-Hamiltonian Approach. Springer, Berlin (2009)
24. Fradkov, A.L.: Synthesis of an adaptive system for linear plant stabilization. Autom. Remote Control **35**(12), 1960–1966 (1974)
25. Fujimoto, K., Sugie, T.: Canonical transformations and stabilization of generalized Hamiltonian systems. Syst. Control Lett. **42**(3), 217–227 (2001)
26. Grillo, S., Marsden, J.E., Nair, S.: Lyapunov constraints and global asymptotic stabilization. J. Geom. Mech. **3**(2), 145–196 (2011)
27. Hatanaka, T., Chopra, N., Fujita, M., Spong, M.W.: Passivity-Based Control and Estimation in Networked Robotics. Springer International, Cham, Switzerland (2015)
28. Hill, D., Moylan, P.: The stability of nonlinear dissipative systems. IEEE Trans. Autom. Control **25**(5), 708–711 (1976)
29. Jonckeere, E.: Lagrangian theory of large scale systems. In: European Conference on Circuit Theory and Design, The Hague, The Netherlands, pp. 626–629 (1981)
30. Kelly, R.: A simple set–point robot controller by using only position measurement. In: IFAC World Congress, Sydney, Australia, pp. 173–176 (1994)
31. Kelly, R., Ortega, R.: Adaptive control of robot manipulators: an input–output approach. In: IEEE International Conference on Robotics and Automation, PA, USA, pp. 699–703 (1988)
32. Kelly, R., Carelli, R., Ortega, R.: Adaptive motion control design of robot manipulators: an input-output approach. Int. J. Control **49**(12), 2563–2581 (1989)
33. Kelly, R., Ortega, R., Ailon, A., Loria, A.: Global regulation of flexible joints robots using approximate differentiation. IEEE Trans. Autom. Control **39**(6), 1222–1224 (1994)
34. Koditschek, D.E.: Natural motion of robot arms. In: IEEE Conference on Decision and Control, Las Vegas, USA, pp. 733–735 (1984)
35. Koditschek, D.E.: Robot planning and control via potential functions. In: Khatib, O., Craig, J.J., Lozano-Pérez, T. (eds.) The Robotics Review 1, pp. 349–367. The MIT Press, Cambridge (1989)
36. Landau, I.D.: Adaptive Control: The Model Reference Approach. Marcel Dekker, New York (1979)
37. Lewis, A.: Notes on energy shaping. In: IEEE Conference on Decision and Control, Paradise Island, Bahamas, pp. 4818–4823 (2004)
38. Loria, A.: Observers are unnecessary for output-feedback control of Lagrangian systems. IEEE Trans. Autom. Control **61**(4), 905–920 (2016)
39. Mahindrakar, A.D., Astolfi, A., Ortega, R., Viola, G.: Further constructive results on IDA-PBC of mechanical systems: the Acrobot example. Int. J. Robust Nonlinear Control **16**, 671–685 (2006)
40. Monroy, A., Alvarez-Icaza, L., Espinosa-Perez, G.: Passivity-based control for variable speed constant frequency operation of a DFIG wind turbine. Int. J. Control **81**(9), 1399–1407 (2008)
41. Moylan, P., Anderson, B.D.O.: Nonlinear regulator theory and an inverse optimal control problem. IEEE Trans. Autom. Control **18**(5), 460–465 (1973)
42. Nunna, K., Sassano, M., Astolfi, A.: Constructive interconnection and damping assignment for port-controlled Hamiltonian systems. IEEE Trans. Autom. Control **60**(9), 2350–2361 (2015)
43. Nuño, E., Ortega, R., Basañez, L.: An adaptive controller for nonlinear teleoperators. Automatica **46**(2), 155–159 (2010)
44. Ortega, R.: Passivity properties for stabilization of cascaded nonlinear systems. Automatica **27**(2), 423–424 (1991)
45. Ortega, R.: Applications of input–output techniques to control problems. In: European Control Conference, Grenoble, France, pp. 1307–1313 (1991)

46. Ortega, R., Borja, P.: New results on control by interconnection and energy–balancing passivity–based control of port–Hamiltonian systems. In: IEEE Conference on Decision and Control, Los Angeles, USA, pp. 2346–2351 (2014)
47. Ortega, R., García-Canseco, E.: Interconnection and damping assignment passivity-based control: a survey. Eur. J. Control **10**, 432–450 (2004)
48. Ortega, R., Romero, J.G.: Robust integral control of port-Hamiltonian systems: the case of non-passive outputs with unmatched disturbances. Syst. Control Lett. **61**(1), 11–17 (2012)
49. Ortega, R., Spong, M.W.: Adaptive motion control of rigid robots: a tutorial. Automatica **25**(6), 877–888 (1989)
50. Ortega, R., Rodriguez, A., Espinosa, G.: Adaptive stabilization of non-linearizable systems under a matching condition. In: American Control Conference, San Diego, USA, pp. 67–72 (1990)
51. Ortega, R., Loria, A., Nicklasson, P., Sira-Ramirez, H.: Passivity-Based Control of Euler-Lagrange Systems: Mechanical, Electrical, and Electromechanical Applications. Springer, London (1998)
52. Ortega, R., Spong, M.W., Gomez, F., Blankenstein, G.: Stabilization of underactuated mechanical systems via interconnection and damping assignment. IEEE Trans. Autom. Control **47**(8), 1218–1233 (2002)
53. Ortega, R., van der Schaft, A., Castaños, F., Astolfi, A.: Control by interconnection and standard passivity-based control of port-Hamiltonian systems. IEEE Trans. Autom. Control **53**(11), 2527–2542 (2008)
54. Paden, B., Panja, R.: Globally asymptotically stable PD+ controller for robot manipulators. Int. J. Control **4**, 1697–1712 (1988)
55. Rodriguez, A., Ortega, R.: Adaptive stabilization of nonlinear systems: the non–feedback–linearizable case. In: IFAC World Congress, Tallinn, USSR, pp. 121–124 (1990)
56. Romero, J.G., Ortega, R.: Two globally convergent adaptive speed observers for mechanical systems. Automatica **60**, 7–11 (2015)
57. Romero, J.G., Donaire, A., Ortega, R.: Robust energy shaping control of mechanical systems. Syst. Control Lett. **62**(9), 770–780 (2013)
58. Romero, J.G., Ortega, R., Sarras, I.: A globally exponentially stable tracking controller for mechanical systems using position feedback. IEEE Trans. Autom. Control **60**(3), 818–823 (2015)
59. Romero, J.G., Ortega, R., Donaire, A.: Energy shaping of mechanical systems via PID control and extension to constant speed tracking. IEEE Trans. Autom. Control **61**(11), 3551–3556 (2016)
60. Ryalat, M., Laila, D., Torbati, M.: Integral IDA–PBC and PID–like control for port–controlled Hamiltonian systems. In: American Control Conference, Chicago, USA, pp. 5365–5370 (2015)
61. Sarras, I., Acosta, J.A., Ortega, R., Mahindrakar, A.: Constructive immersion and invariance stabilization for a class of underactuated mechanical systems. Automatica **49**(5), 1442–1448 (2013)
62. Slotine, J., Li, W.: Adaptive manipulator control: a case study. IEEE Trans. Autom. Control **33**(11), 995–1003 (1988)
63. Spong, M.W.: Partial feedback linearization of underactuated mechanical systems. In: The IEEE/RSJ International Conference on Intelligent Robots and Systems, Munich, Germany, pp. 314–321 (1994)
64. Stramigioli, S., Maschke, B., van der Schaft, A.: Passive output feedback and port interconnection. In: IFAC Symposium on Nonlinear Control Systems, Enschede, The Netherlands, pp. 613–618 (1998)
65. Takegaki, M., Arimoto, S.: A new feedback for dynamic control of manipulators. Trans. ASME: J. Dyn. Syst. Meas. Control **12**, 119–125 (1981)
66. van der Schaft, A.: L_2-Gain and Passivity Techniques in Nonlinear Control. Springer, Berlin (2000)
67. Venkatraman, A., Ortega, R., Sarras, I., van der Schaft, A.: Speed observation and position feedback stabilization of partially linearizable mechanical systems. IEEE Trans. Autom. Control **55**(5), 1059–1074 (2010)

68. Viola, G., Ortega, R., Banavar, R., Acosta, J.A., Astolfi, A.: Total energy shaping control of mechanical systems: simplifying the matching equations via coordinate changes. IEEE Trans. Autom. Control **52**(6), 1093–1099 (2007)
69. Willems, J.C.: The behavioral approach to open and interconnected systems. IEEE Control Syst. Mag. **27**(6), 46–99 (2007)
70. Woolsey, C., Bloch, A., Leonard, N., Marsden, J.: Physical dissipation and the method of controlled Lagrangians. In: European Control Conference, Porto, Portugal, pp. 2570–2575 (2001)

Chapter 8
Asymptotic Stabilization of Some Finite and Infinite Dimensional Systems by Means of Dynamic Event-Triggered Output Feedbacks

Christophe Prieur and Aneel Tanwani

Abstract The problem of designing dynamic sampling routines for output feedback stabilization of controlled plants is considered. Instead of the more conventional periodic sampling, our approach is based on using event-triggered conditions for sampling, which potentially allow for reduced rate of communication between the plant and the controller. Several classes of control systems, from finite dimensional to infinite dimensional, are considered in this chapter, each within its own problem setup. Within the setup of finite dimensional systems, we consider plants comprising linear and nonlinear ordinary differential equations, and controlled via dynamic output feedback controllers. For such systems, we provide (different) event-based dynamic algorithms to determine sampling times for outputs and control inputs. In the linear case, it is further shown that the proposed algorithms are robust with respect to communication errors due to quantization, and if the parameters of the quantizers are updated appropriately, then the state of the closed-loop system converges asymptotically to the equilibrium. For the plants modeled as hyperbolic system of conservation laws, an event-triggered sampling algorithm of the boundary control results in state converging to the origin. Together with the asymptotic stabilization of the closed-loop system, it is also shown that there exists a minimum inter-sampling time and thus Zeno solutions are avoided in the closed-loop system despite the state-dependent occurrence of discrete dynamics. For all the considered control problems, Lyapunov functions are instrumental to define the sampling sequences, the desired robustness properties of the controller are formalized using input-to-state stability notion, and the tools from stability of cascaded systems and certainty equivalence principle are essential for analysis carried out in our work.

C. Prieur (✉)
GIPSA-lab, CNRS, University Grenoble Alpes, 38000 Grenoble, France
e-mail: Christophe.Prieur@gipsa-lab.fr

A. Tanwani
Laboratoire d'Analyse et Architecture des Systèmes (LAAS), CNRS,
7 Ave. Colonel Roche, 31400 Toulouse, France
e-mail: aneel.tanwani@laas.fr

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_8

8.1 Introduction

Modern day control systems often involve the interface between a physical plant and a digital computer through a communication channel. Appropriate exchange of information between the plant and controller is essential for obtaining desired performance from such control systems. Due to limited capacity of the communication channel, control practitioners have to design algorithms that determine how *frequently* and *accurately*, the plant and the control need to transmit data for an acceptable outcome. For finite dimensional systems, such problems have been studied under the framework of sampled-data control, quantized control, or more broadly, within the context of *network control systems*, see [19, 25] for a survey, and list of references, on these topics.

The problems studied in this chapter relate to asymptotic stabilization of certain finite and infinite dimensional dynamical systems using output feedback controllers, when the information cannot be transmitted continuously and hence the signals need to be sampled. In addition, we are also interested in dealing with the uncertainties, or errors in the communication that are typically introduced in digital communication. The layout of this problem setup is sketched in Fig. 8.1.

The dynamical plants that we consider include finite dimensional linear and nonlinear ordinary differential equations (ODEs), and linear hyperbolic partial differential equations (PDEs). We are mainly interested in algorithms to determine sampling times for inputs and outputs using the recently revived framework of *event-based strategies* [6], where the basic idea is to sample a signal based on its current value instead of using pre-calculated sampling times as done conventionally. The class of controllers in the ODE setup comprises dynamic output feedback controllers, whereas for the hyperbolic PDEs, we restrict ourselves to the static output feedback boundary controls. For the case of linear ODEs, we also consider added communication errors due to quantization, and the design of dynamic quantizers which result in asymptotic stability.

When analyzing control systems in the presence of sampling errors, the controllers are required to be robust with respect to output measurement errors, typically formalized using *input-to-state stability* (ISS) notion. The basic idea behind the event-triggered sampling is to implement such controllers and keep the sampling error relative to the current value of state sufficiently small to ensure asymptotic sta-

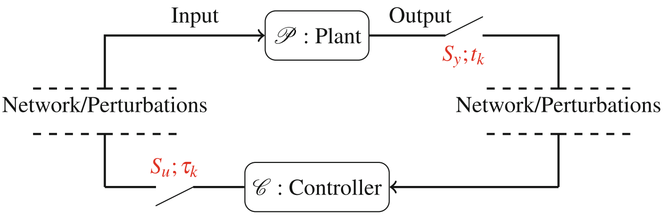


Fig. 8.1 Output feedback control of dynamical systems with perturbations or network

bility [18, 39]. Various variants of this idea have now appeared in literature [4, 24, 26, 34]. However, the implementation of such schemes in the presence of dynamic output feedback controllers has remained a challenge. The current literature mostly addresses linear ODEs and even in that case, the current results either ensure practical stability of the system, or they build on periodic sampling results [1, 17].

Designing feasible sampling algorithms (based on event-based strategies) for the plants controlled via dynamic output feedback, and analyzing the closed-loop system via Lyapunov methods, are the focal points of this chapter. Depending on the plant under consideration, we study the sampling problem under different setups. Basically, as we generalize the system class, the related problem context becomes more specific. Starting with the very specific case of linear ODEs, we consider the most general problem setup, where in addition to determining sampling times for implementation of output feedback controllers, we also consider errors in communication of outputs and inputs that arise due to quantization. In contrast to existing approaches, the sampling times for outputs and control inputs are not necessarily synchronized. In addition, these algorithms are shown to be robust with respect to uncertainties in communication of outputs and inputs. These uncertainties are modeled as quantization errors, and algorithms are also provided to design dynamic quantizers where the quantization error converges to zero as the state of plant gets closer to the origin. As a result, the overall closed loop is shown to be asymptotically stable for our choice of sampling and quantization algorithms.

We then move onto plants modeled as nonlinear ODEs, and by now, there exists a large literature dealing with output feedback stabilization, in particular, those based on designing the observer and static state feedback control laws separately (see e.g., [5, 42]). Building on the ideas developed for the linear case, we now design auxiliary dynamical systems which determine the sampling times. The introduction of discrete dynamics, in addition to the already present continuous dynamics, calls for the framework of hybrid systems as introduced in [12, 28]; see also the textbook [13] on this subject. The tools from the theory of stability of cascaded nonlinear systems, and the hybrid systems, are thus used to analyze the stability of the closed-loop system (as done in e.g. [32]).

Finally, we move on to the infinite dimensional systems, and in this chapter, we look at the plants modeled as linear hyperbolic PDEs with boundary control. At this moment, our result for this system class only comprises (synchronous) sampling algorithm for static output feedback boundary control laws. However, the sampling algorithms still use additional dynamics for event-triggered sampling but with certain differences from finite dimensional setup.

The development of this chapter aims at highlighting the techniques based on analysis with Lyapunov functions, which were the central ingredient for design and analysis of the aforementioned problems. The basic idea is to address the problem of sampled-data control for very general system classes under a unifying framework. The results presented here are based on authors' recent work, and for detailed reading and proofs, we refer the reader to the following papers:

- For linear ODEs, paper [40] studies the problem of designing the sampling sequences, and quantization schemes for both outputs and control inputs separately. These results are summarized in Sect. 8.2.
- For nonlinear ODEs controlled via dynamic output feedback, paper [41] suggests dynamic sampling algorithms for both inputs and outputs; see Sect. 8.3.
- For linear hyperbolic PDEs, paper [10] suggests an event-based sampling routine for the boundary control; see Sect. 8.4.

Let us emphasize that for each of these results, not only the asymptotic stability of the closed-loop system is obtained, but it is also proven analytically that there is no accumulation of sampling times over a finite interval, which is important for implementing event-based algorithms.

8.2 Design of Quantized and Event-Triggered Controllers for Linear Systems

In this section, we consider linear time-invariant plants described as:

$$\mathscr{P} : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \,, \\ y(t) = Cx(t), \end{cases} \tag{8.1}$$

where $x(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^p$ is the measured output, and $u(t) \in \mathbb{R}^m$ is the control input. We are interested in feedback stabilization of the control system (8.1) by realizing the control architecture proposed in Fig. 8.2. That is, the outputs and inputs are subjected to zero-order sample and hold and the sampling instants need to be computed separately for both signals. Furthermore, the sampled outputs and control inputs cannot be transmitted exactly to the controller and the plant, respectively, and must be encoded using finitely many alphabets.

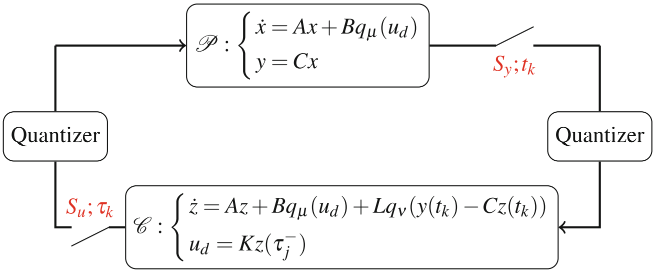

$\mathscr{P} : \begin{cases} \dot{x} = Ax + Bq_\mu(u_d) \\ y = Cx \end{cases}$
$S_y; t_k$
Quantizer
Quantizer
$S_u; \tau_k$
$\mathscr{C} : \begin{cases} \dot{z} = Az + Bq_\mu(u_d) + Lq_\nu(y(t_k) - Cz(t_k)) \\ u_d = Kz(\tau_j^-) \end{cases}$

Fig. 8.2 Feedback loop with time-sampled and quantized inputs and outputs

The proposed dynamic controller $\mathscr{C}$ has the following form[1]:

$$\mathscr{C} : \begin{cases} \dot{z}(t) = Az(t) + Bu(t) + Lq_\nu(y(t_k) - Cz(t_k)), & t \in [t_k, t_{k+1}), \\ u(t) = q_\mu(Kz(\tau_j)), \quad t \in [\tau_j, \tau_{j+1}), & \end{cases} \tag{8.2}$$

where $y(t_k)$, and $u(\tau_j)$, $k, j \in \mathbb{N}$, denote the sampled values of output and input, respectively. The output quantizer $q_\nu : \mathbb{R}^p \to \mathscr{Q}_y$, and input quantizer $q_\mu : \mathbb{R}^m \to \mathscr{Q}_u$ for some finite sets $\mathscr{Q}_y$ and $\mathscr{Q}_u$, include the design parameters $\nu : [t_0, \infty) \to \mathbb{R}_+$, and $\mu : [\tau_0, \infty) \to \mathbb{R}_+$, respectively, which are piecewise constant and are only updated at t_k and τ_j, respectively. For notational convenience, we will often denote $\nu(t_k)$ by ν_k, and $\mu(\tau_j)$ by μ_j. In writing Eq. (8.2), it must be noted that the discrete measurements received by the controller have been passed through a sample-and-hold device, and that the state $z(\cdot)$ evolves continuously. This approach is essentially different from some of the existing techniques adopted in for example, [2, 23], where the state of the observer/controller is updated in discrete manner whenever the new measurements are available (periodically). We remark that the stability of dynamical systems with quantized measurements has been studied extensively over the past decade, see the survey [25] and references therein. Stability using quantized and periodically sampled output measurements (without asymptotic observer) was considered in [23], and using asymptotic observers (without sampling) was considered in [22]. To the best of our knowledge, stability where both the inputs and outputs are quantized and aperiodically sampled has not been treated. In doing so, we find that the quantization parameter for control input depends upon the parameter chosen for output quantization, in order to capture the growth of the estimated state, and that there is a trade-off between how fast we sample and how precisely we quantize due to the choice of respective parameters.

The proposed controller (8.2) is based on the principle of *certainty equivalence*. In the absence of quantization or sampling errors, such controllers drive the state estimation error $(x - z)$ to zero, and the control input replicates the full state static feedback law to drive the plant state to the origin. Thus, the following basic assumptions are necessary:

(L-1) The pair (A, B) is stabilizable, and hence for every symmetric positive definite matrix Q_c (denoted as $Q_c > 0$) there exists a matrix $P_c > 0$ such that

$$(A + BK)^\top P_c + P_c(A + BK) \leq -Q_c. \tag{8.3}$$

(L-2) The pair (A, C) is observable, so that for every $Q_o > 0$, there exists a matrix $P_o > 0$ such that

$$(A - LC)^\top P_o + P_o(A - LC) \leq -Q_o. \tag{8.4}$$

[1] Implementing the controller (8.2) requires that the variable z is also known for computing the output sampling times and choosing the appropriate encoding symbol. Thus, a copy of (8.2) is also implemented next to the plant output sensors to compute the value of z.

Now, using Fig. 8.2 as the template for the remainder of the section, we provide the algorithms to compute the output sampling times t_k, and input sampling times τ_j, $k, j \in \mathbb{N}$ in Sects. 8.2.1, and 8.2.2, respectively, along with the encoding strategies for the respective quantizers.

8.2.1 Output Processing Unit

Our first objective is to provide an algorithm that gives sampling times at which the plant output must be sent to the controller, followed by an update rule for output quantization parameters which provide an encoding scheme for the sampled output measurements. As stated earlier, the controller needs the output information in order to estimate the state of the plant, and hence the output samples must be sent often enough so that the state estimation error is always decaying with respect to some chosen metric. It is thus useful to introduce the estimation error $\tilde{x} := x - z$. Equations (8.1) and (8.2) then result in the following equations for error dynamics:

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) - Lq_{v_k}(\tilde{y}(t_k)) \\ \tilde{y}(t) &= C\tilde{x}(t). \end{aligned} \tag{8.5}$$

8.2.1.1 Event-Triggered Sampling of the Output

To provide the basic idea behind calculation of the sampling times, let us momentarily ignore the error due to quantization of the output, and replace the term $q_{v_k}(\tilde{y}(t_k))$ in (8.5) by $\tilde{y}(t_k) = \tilde{y}(t) + \tilde{y}(t_k) - \tilde{y}(t)$. If the sampling times t_k are chosen such that

$$|\tilde{y}(t) - \tilde{y}(t_k)| \le \sigma_o |\tilde{x}(t)|$$

for $\sigma_o > 0$ sufficiently small, then $|\tilde{x}(t)|$ converges to zero.

However, in addition to the quantization errors, the term $\tilde{x}$ remains unknown. But, using this intuition, we first aim to find an invertible map from $\tilde{x}(t)$ to some past-sampled output values of $\tilde{y}$ measured over the interval $[t_0, t)$ using Eq. (8.5), so that $\tilde{x}(t)$ can be directly expressed in terms of a certain number of past output samples. To do so, it is seen that in (8.5), $q_{v_k}(\tilde{y}(t_k))$ acts as a known term, and using $\tilde{y}$ as the output, it is possible under the observability assumption to reconstruct $\tilde{x}$ as a function of (sufficiently many) sampled values of $\tilde{y}$ over any compact interval. Toward this end, let

$$\psi(s_1, s_2, s_3) := Ce^{As_1} \int_{s_2}^{s_3} e^{-As} L \, ds, \quad s_1 \le s_2 \le s_3, \tag{8.6}$$

so that ψ takes values in $\mathbb{R}^p$. Now, for $t > t_k > t_{k-1} > \cdots > t_{k-\mathrm{n_s}-1} \ge t_0$, define the lower triangular matrix $\Psi_{k,\mathrm{n_s}}(t)$ as

$$\Psi_{k,\mathtt{n}_\mathtt{s}}(t) := \begin{bmatrix} \psi(t_k, t_k, t) & 0 & 0 \cdots \\ \psi(t_{k-1}, t_k, t) & \psi(t_{k-1}, t_{k-1}, t_k) & 0 \cdots \\ \psi(t_{k-2}, t_k, t) & \psi(t_{k-2}, t_{k-1}, t_k) & \psi(t_{k-2}, t_{k-2}, t_{k-1}) \cdots \\ \vdots & \vdots & \vdots, \end{bmatrix}$$

where $\mathtt{n}_\mathtt{s} \in \mathbb{N}$ is some strictly positive integer, and to keep the notation short, it is noted that $\Psi_{k,\mathtt{n}_\mathtt{s}}(t)$ depends on $(t, t_k, t_{k-1}, \dots, t_{k-\mathtt{n}_\mathtt{s}-1})$. Next, introduce the following notation:

$$\begin{aligned} N_{k,\mathtt{n}_\mathtt{s}}(t) &:= \operatorname{col}\left(Ce^{-A(t-t_k)}, Ce^{-A(t-t_{k-1})}, \cdots, Ce^{-A(t-t_{k-\mathtt{n}_\mathtt{s}+1})}\right), \\ \widetilde{Y}_{k,\mathtt{n}_\mathtt{s}} &:= \operatorname{col}\left(\tilde{y}(t_k), \tilde{y}(t_{k-1}), \cdots, \tilde{y}(t_{k-\mathtt{n}_\mathtt{s}+1})\right), \\ q_{\nu_{k,\mathtt{n}_\mathtt{s}}}(\widetilde{Y}_{k,\mathtt{n}_\mathtt{s}}) &:= \operatorname{col}\left(q_{\nu_k}(\tilde{y}(t_k)), q_{\nu_{k-1}}(\tilde{y}(t_{k-1})), \cdots, q_{\nu_{k-\mathtt{n}_\mathtt{s}+1}}(\tilde{y}(t_{k-\mathtt{n}_\mathtt{s}+1}))\right). \end{aligned} \tag{8.7}$$

Applying the variation of constants formula to system (8.5), it is seen that, for any $t > t_k > t_{k-1} > \cdots > t_{k-\mathtt{n}_\mathtt{s}+1}$, we have

$$N_{k,\mathtt{n}_\mathtt{s}}(t)\tilde{x}(t) = \widetilde{Y}_{k,\mathtt{n}_\mathtt{s}} - \Psi_{k,\mathtt{n}_\mathtt{s}}(t) q_{\nu_{k,\mathtt{n}_\mathtt{s}}}(\widetilde{Y}_{k,\mathtt{n}_\mathtt{s}}). \tag{8.8}$$

We will use this important relation (8.8) to define the sampling times for output measurements, but before proceeding to that, let us recall a few well-known results. An important requirement in our analysis of minimum inter-sampling time is the invertibility of the matrix $N_{k,\mathtt{n}_\mathtt{s}}(t)$, for each $t > t_{\mathtt{n}_\mathtt{s}-1}$, and is achieved due to the following result from [43]:

Lemma 8.1 *Let* $\operatorname{Im}(\lambda(A))$ *denote the imaginary part of the eigenvalue* λ *of the matrix* A*, and let* $\omega := \max_{1\le i,j\le n}\{\operatorname{Im}(\lambda_i(A) - \lambda_j(A))\}$. *If*

$$\mathtt{n}_\mathtt{s} > 2(n-1) + \frac{T_s}{2\pi}\omega \tag{8.9}$$

then the matrix $\operatorname{col}(Ce^{As_1}, Ce^{As_2}, \cdots, Ce^{As_{\mathtt{n}_\mathtt{s}}})$ *is left invertible for all* $s_1, s_2, \dots, s_{\mathtt{n}_\mathtt{s}} \in [0, T_s]$.

In the context of Lemma 8.1, $\mathtt{n}_\mathtt{s}$ could be interpreted as the number of samples required for observability of the discretized system (8.5). It is well known from linear systems theory that for an observable matrix pair (A, C) where A has only real eigenvalues, it would suffice to take $\mathtt{n}_\mathtt{s}$ to be the observability index for invertibility of $N_{k,\mathtt{n}_\mathtt{s}}(t)$ for each $t \ge 0$. However, in the presence of complex eigenvalues, there are some isolated points on real line where $N_{k,\mathtt{n}_\mathtt{s}}(t)$ may loose rank. It is proved in [43, Theorem 1] that the number of roots of the determinant of $N_{k,\mathtt{n}_\mathtt{s}}(t)$ are upper bounded on any bounded interval, and if one takes $\mathtt{n}_\mathtt{s}$ to satisfy the bound (8.9), then injectivity of $N_{k,\mathtt{n}_\mathtt{s}}(t)$ holds for all $t \in [0, T_s]$.

To use this lemma for our problem setup, we first fix some integer $\mathtt{n}_\mathtt{s}^* > 2(n-1)$. Choose $t_0 < t_1 < \cdots < t_{\mathtt{n}_\mathtt{s}^*-1}$ arbitrarily, and let

$$f(t, \widetilde{Y}_{k,\mathtt{n}_\mathrm{s}}) = \widetilde{Y}_{k,\mathtt{n}_\mathrm{s}} - \Psi_{k,\mathtt{n}_\mathrm{s}}(t) q_{v_{k,\mathtt{n}_\mathrm{s}}}(\widetilde{Y}_{k,\mathtt{n}_\mathrm{s}}), \qquad k \geq \mathtt{n}_\mathrm{s} - 1.$$

The sampling times t_{k+1}, for $k \geq \mathtt{n}_\mathrm{s}^*$, are then defined recursively as follows:

$$t_{k+1}^{\mathrm{event}} := \inf \left\{ t > t_k \mid \|N_{k,\mathtt{n}_\mathrm{s}}(t)\| \cdot |\tilde{y}(t) - \tilde{y}(t_k)| \geq \sigma_o \cdot |f(t, \widetilde{Y}_{k,\mathtt{n}_\mathrm{s}})| \right\} \tag{8.10a}$$

$$t_{k+1}^{\mathrm{sample}} := \inf \left\{ t > t_k \mid t - t_{k-\mathtt{n}_\mathrm{s}^*+1} > \min \left\{ \frac{2\pi}{\omega} \left(\mathtt{n}_\mathrm{s}^* - 2(n-1) \right), \mathtt{n}_\mathrm{s}^* T \right\} \right\} \tag{8.10b}$$

$$t_{k+1} := \min\{ t_{k+1}^{\mathrm{event}}, t_{k+1}^{\mathrm{sample}} \}, \tag{8.10c}$$

where $\sigma_o := \varepsilon_o \dfrac{\lambda_{\min}(Q_o)}{2\,\|P_o L\|}$, for some $\varepsilon_o \in (0, 1)$ and $T > 0$ is some prespecified constant, which can be arbitrarily large, but finite. The sampling rule (8.10) guarantees that the output measurements are transmitted persistently to the controller.

8.2.1.2 Output Quantization

We now define an encoding strategy that is used to transmit $\tilde{y}(t_k)$ at each time instant t_k using a string of finite length. The quantization model we use is adopted from [22], which is a dynamic one. For output measurements, we assume that the quantizer has a scalable parameter v and has the form:

$$q_v(y) = v q^y \left(\frac{y}{v} \right), \tag{8.11}$$

where $q^y(\cdot)$ denotes a finite-level quantizer with sensitivity parameterized by Δ_y and range R_y, that is, **if** $|y| \leq R_y$, **then** $|q^y(y) - y| \leq \Delta_y$. This way, the range of the quantizer $q_v(\cdot)$ is $R_y v$ and the sensitivity is $\Delta_y v$. Increasing v would mean that we are increasing the range of the quantizer with large quantization errors and decreasing v corresponds to finer quantization with smaller range. It will be assumed that the quantizer is centered around the origin, that is, $q^y(y) = 0$ if $|y| < \Delta_y$.

We now specify an update rule for the parameter v, so that the state estimation error $\tilde{x}$ converges to zero. First, we pick $v_0, \cdots, v_{\mathtt{n}_\mathrm{s}-1}$ to be arbitrary. It is assumed that $v_{\mathtt{n}_\mathrm{s}}$ is chosen such that[2] $\tilde{x}(t_{\mathtt{n}_\mathrm{s}})$ is contained in an ellipsoid:

$$V_o(\tilde{x}(t_{\mathtt{n}_\mathrm{s}})) \leq \frac{\lambda_{\min}(P_o) R_y^2}{\|C\|^2} {v_{\mathtt{n}_\mathrm{s}}}^2, \tag{8.12}$$

[2] If $\tilde{x}(t_0)$ is known to belong to a known bounded set, then $v_{\mathtt{n}_\mathrm{s}}$ satisfying (8.12) is computed from calculating an upper bound on $|\tilde{x}(t_{\mathtt{n}_\mathrm{s}})|$ using the differential equation (8.5). One can also use the relation (8.8) to obtain an upper bound on $\tilde{x}$ at certain time, or use the strategy proposed in [22, 36] to get a bound on state estimation error. To keep the notation simple, we have used the same index for sampling times and quantization parameter.

then $|\tilde{x}(t_{\mathtt{n_s}})| \leq \frac{R_y}{\|C\|} v_0$ and $|C\tilde{x}(t_{\mathtt{n_s}})| \leq R_y v_0$. Suppose that we have chosen v_k such that (8.12) holds for $\tilde{x}(t_k)$, for some $k \geq \mathbb{N}$. We now specify v_{k+1} such that (8.12) holds for $\tilde{x}(t_{k+1})$, for all $k \in \mathbb{N}$, and at the same time $\lim_{k\to\infty} v_k = 0$.

Since the controller receives the quantized measurements only, the observer takes the following form over the interval $t \in [t_k, t_{k+1})$:

$$\dot{z}(t) = Az(t) + Bu(t) + Lq_{v_k}(y(t_k) - Cz(t_k)). \tag{8.13}$$

The dynamics of the state estimation error for the interval $[t_k, t_{k+1})$ are:

$$\begin{aligned}
\dot{\tilde{x}}(t) &= A\tilde{x}(t) - Lq_{v_k}(\tilde{y}(t_k)) && \text{(8.14a)}\\
&= A\tilde{x}(t) - L\tilde{y}(t_k) - Lq_{v_k}(\tilde{y}(t_k)) + L\tilde{y}(t_k) && \text{(8.14b)}\\
&= (A - LC)\tilde{x}(t) + L(\tilde{y}(t) - \tilde{y}(t_k)) - v_k L\left(q^y\left(\frac{\tilde{y}(t_k)}{v_k}\right) - \frac{\tilde{y}(t_k)}{v_k}\right). && \text{(8.14c)}
\end{aligned}$$

Pick $V_o(\tilde{x}) = \tilde{x}^\top P_o \tilde{x}$ as the Lyapunov function, and we see that the measurement update rule (8.10) leads to the following bound for $t \in [t_k, t_{k+1})$, $k \geq \mathtt{n_s}$:

$$\dot{V}_o(\tilde{x}(t)) \leq -(1 - \varepsilon_o)\lambda_{\min}(Q_o)|\tilde{x}(t)|^2 + 2v_k\, \Delta_y\, \|P_o L\|\, |\tilde{x}(t)|. \tag{8.15}$$

Thus, within two measurement updates, the error converges to a ball parameterized by v_k. In particular, for some $0 < \alpha_o < \frac{(1-\varepsilon_0)\lambda_{\min}(Q_o)}{\lambda_{\max}(P_o)}$, if we let

$$\chi_o := \frac{2\,\|P_o L\|}{(1 - \varepsilon_o)\lambda_{\min}(Q_o) - \alpha_o \lambda_{\max}(P_o)} \tag{8.16}$$

then $|\tilde{x}(t)| \geq \chi_o \Delta_y v_k$ implies that

$$\dot{V}_o(\tilde{x}(t)) \leq -\alpha_o V_o(\tilde{x}(t)).$$

Thus, it follows that, for $t \in [t_k, t_{k+1})$, $k \geq \mathtt{n_s}$:

$$V_o(\tilde{x}(t)) \leq \max\{\lambda_{\max}(P_o)\chi_o^2 \Delta_y^2 v_k^2, e^{-\alpha_o(t-t_k)} V_o(\tilde{x}(t_k))\}.$$

For each $k \geq 0$, letting

$$\Theta_{k+1}^y := \max\left\{\lambda_{\max}(P_o)\chi_o^2\Delta_y^2 v_k^2, e^{-\alpha_o(t_{k+1}-t_k)} \frac{\lambda_{\min}(P_o)R_y^2}{\|C\|^2} v_k^2\right\},$$

it follows that $V_o(\tilde{x}(t_{k+1})) \leq \Theta^y_{k+1}$, for $k \geq \mathtt{n_s}$. If we now pick $v_{k+1}, k \geq \mathtt{n_s}$, as follows:

$$v_{k+1} := \frac{\|C\|}{R_y} \sqrt{\frac{\Theta^y_{k+1}}{\lambda_{\min}(P_o)}} \tag{8.17}$$

then it is guaranteed that $|\tilde{x}(t_{k+1})| \leq \frac{R_y}{\|C\|} v_{k+1}$ and $|C\tilde{x}(t_{k+1})| \leq R_y v_{k+1}$. To ensure v_k converges to zero as k gets large, we impose a bound on the number of quantization levels determined by the ratio $\frac{\Delta_y}{R_y}$, see (8.22) in the statement of Theorem 8.1.

8.2.2 Input Processing Unit

We can tailor the aforementioned ideas to derive a sampling algorithm and a quantization strategy for the control input.

8.2.2.1 Sampling Algorithms for Inputs

Let $\tau_0 = t_{\mathtt{n_s}}$, and choose the control input u, so that $u(t) = 0$, for $t \in [t_0, \tau_0)$, and $u(\tau_0) = Kz(\tau_0) = Kz(t_{\mathtt{n_s}})$, and the next update is performed at τ_{j+1}, which, for $j \geq 0$, is defined as follows:

$$\tau^{\text{event}}_{j+1} := \inf \left\{ t > \tau_j \,\middle|\, |Kz(t) - Kz(\tau_j)| \geq \left(\sigma_c |z(t)| + \gamma_c \frac{R_y}{\|C\|} v(t) \right) \right\}, \tag{8.18a}$$

$$\tau^{\text{pers}}_{j+1} := \inf\{t > \tau_j \mid t - \tau_j \geq T\}, \tag{8.18b}$$

$$\tau_{j+1} := \min\{\tau^{\text{event}}_j, \tau^{\text{pers}}_j\}, \tag{8.18c}$$

where $\sigma_c := \varepsilon_c \dfrac{\lambda_{\min}(Q_c)}{2\|P_c B\|}$, and $\gamma_c := \tilde{\beta}\sigma_c \|C\|$ for some $\varepsilon_c \in (0, 1)$, and $\tilde{\beta} > 0$.

Note that the term $v(\cdot)$ is only piecewise constant and does not vary continuously with time. In case there is a time $t_k > \tau_j$ such that $|Kz(t_k) - Kz(\tau_j)| < \sigma_c |z(t_k)| + \gamma_c R_y v(t_k^-))$, and due to sudden change in the value of v at time t_k, it happens that $|Kz(t_k) - Kz(\tau_j)| \geq \sigma_c |z(t_k)| + \gamma_c R_y v(t_k^+)$, then in that case we assume that $\tau_{j+1} = t_k$, and hence the control input is updated instantaneously without any delay.

8.2.2.2 Input Quantization

In our setup, the control input cannot be transmitted to the plant with exact precision and only $q_{\mu_j}(Kz(\tau_j)), j \in \mathbb{N}$ is transmitted to the plant. The quantization model used for control inputs is similar to the one adopted for outputs, that is,

$$q_\mu(u) = \mu\, q^u\left(\frac{u}{\mu}\right),$$

where μ denotes the scaling parameter, and q^u is a finite-level quantizer whose range is denoted by R_u, and the sensitivity by Δ_u. We specify an update rule for the parameter μ_j associated with the input quantizer such that the resulting closed-loop system is still globally asymptotically stable. In order to do that, we choose $z(\tau_0)$ such that

$$V_c(z(\tau_0)) \le \frac{\lambda_{\min}(P_c)R_u^2}{\|K\|^2}\mu_0^2.$$

With quantized inputs and outputs, the dynamical system (8.2) is thus written as:

$$\begin{aligned}\dot z(t) &= (A+BK)z(t) + B(u(t)-Kz(t)) + L(q_{\nu_k}(y(t_k)-Cz(t_k))), \quad t, t_k \in [\tau_j, \tau_{j+1})\\ &= (A+BK)z(t) + BK(z(\tau_j)-z(t)) + \mu_j B\left(q\left(\frac{Kz(\tau_j)}{\mu_j}\right) - \frac{Kz(\tau_j)}{\mu_j}\right)\\ &\quad + L(q_{\nu_k}(y(t_k)-Cz(t_k))).\end{aligned}$$

With $V_c(z) = z^\top P_c z$ as the Lyapunov function, and the control update rule (8.18), we observe that

$$\begin{aligned}\dot V_c \le -(1-\varepsilon_c)\lambda_{\min}(Q_c)|z(t)|^2 + |z(t)|(\tilde\beta\varepsilon_c\lambda_{\min}(Q_c)R_y\nu_{k^*(t)} + 2\|P_cB\|\Delta_u\mu_j)\\ +2\,|z(t)|\,\|P_cL\|(|\tilde y(t_{k^*(t)})| + \nu_{k^*(t)}\Delta_y),\end{aligned} \tag{8.19}$$

where

$$k^*(t) := \max\{k \in \mathbb{N} : t_k \le t\}.$$

From our output quantization scheme, we have that $|\tilde y(t_k)| = |C\tilde x(t_k)| \le R_y\nu_k$, for all $k \in \mathbb{N}$, and $\nu_{k^*(\tau_j)} \ge \nu_{k^*(t)}$, for $t \ge \tau_j$. For a fixed $0 < \alpha_c < \frac{(1-\varepsilon_c)\lambda_{\min}(Q_c)}{\lambda_{\max}(P_c)}$, we introduce the constants

$$\chi_c := \frac{2\|P_cB\|}{(1-\varepsilon_c)\lambda_{\min}(Q_c) - \alpha_c\lambda_{\max}(P_c)} \tag{8.20}$$

and

$$\xi_1 := \frac{\tilde\beta\varepsilon_c\lambda_{\min}(Q_c) + 2\|P_cL\|}{(1-\varepsilon_c)\lambda_{\min}(Q_c) - \alpha_c\lambda_{\max}(P_c)}, \quad \xi_2 := \frac{2\|P_cL\|}{(1-\varepsilon_c)\lambda_{\min}(Q_c) - \alpha_c\lambda_{\max}(P_c)}.$$

It is note that if $|z(t)| \ge \overline{\chi}_j := \chi_c\Delta_u\mu_j + (\xi_1R_y + \xi_2\Delta_y)\nu_{k,j}^*$, then

$$\dot V_c(z(t)) \le -\alpha_c V_c(z(t)).$$

Assuming that $z(\tau_j)$ is contained in an ellipsoid defined as:

$$V_c(z(\tau_j)) \leq \frac{\lambda_{\min}(P_c)R_u^2}{\|K\|^2}\mu_j^2,$$

we let

$$\Theta_{j+1}^u := \max\left\{\lambda_{\max}(P_c)\overline{\chi}_j^2, e^{-\alpha_c(\tau_{j+1}-\tau_j)}\frac{\lambda_{\min}(P_c)R_u^2}{\|K\|^2}\mu_j^2\right\}.$$

Choose μ_{j+1} such that

$$\mu_{j+1}^2 = \frac{\|K\|^2\Theta_{j+1}^u}{R_u^2\lambda_{\min}(P_c)} \tag{8.21}$$

then it is guaranteed that $|z(\tau_{j+1})| \leq \frac{R_u}{\|K\|}\mu_{j+1}$, and $|Kz(\tau_{j+1})| \leq R_u\mu_{j+1}$. The convergence of μ_j to zero will again follow from the bound in (8.22) in Theorem 8.1.

Remark 8.1 In order to implement the quantization algorithm for the control inputs, it must be noted that the parameter μ actually depends on the parameter ν used for the quantization of $\tilde{y}$. This is done because the evolution of the controller state z actually depends upon the quantized values of $\tilde{y}$, and to determine the region that contains the state z at current time instant, we use the knowledge of how large $\tilde{y}$ is, which is indeed captured by the most recent value of ν.

8.2.3 Convergence Result

Based on the sampling strategies and quantization algorithms developed in Sects. 8.2.1 and 8.2.2, we now state our first main result which relates to the asymptotic stability of the origin in closed-loop (8.1), (8.2).

Theorem 8.1 *Assume that the information transmitted between the plant $\mathscr{P}$ to the controller $\mathscr{C}$, given by $q_{\nu_k}(\tilde{y}(t_k))$, $k \geq 1$, and $q_{\mu_j}(Kz(\tau_j))$, $j \geq 1$, are such that*

- *The output sampling instants t_k, $k \geq \mathtt{n_s}$, are determined by the relation (8.10); and the input sampling instants τ_j, $j \geq 1$ are determined by (8.18).*
- *For the output dynamic quantizer (8.11), the parameter $\nu_{\mathtt{n_s}}$ is chosen to satisfy (8.12) and ν_k, $k > \mathtt{n_s}$, is updated according to (8.17). The parameter μ_j for the dynamic quantization of the input is updated according to (8.21).*
- *The number of output quantization levels determined by R_y and Δ_y, and the input quantization levels determined by R_u and Δ_u are such that*

$$\frac{\Delta_y}{R_y} = \sqrt{\frac{\lambda_{\min}(P_o)}{\lambda_{\max}(P_o)}}\cdot\frac{\overline{\rho}_y}{\chi_o\,\|C\|}, \quad \textit{and} \quad \frac{\Delta_u}{R_u} = \sqrt{\frac{\lambda_{\min}(P_c)}{\lambda_{\max}(P_c)}}\frac{\overline{\rho}_u}{\chi_c\|K\|} \tag{8.22}$$

for some $\overline{\rho}_y, \overline{\rho}_u \in (0, 1)$.

Then the following statements hold:

- *There is a minimum dwell time between two consecutive sampling times of the input and the output, that is, there exists $t_D, \tau_D > 0$ such that $t_{k+1} - t_k \geq t_D$, and $\tau_{j+1} - \tau_j \geq \tau_d$ for each $k \geq \mathtt{n}_\mathtt{s}$, $j \geq 1$.*
- *The origins of the error dynamics (8.5) and of the plant dynamics (8.1) are asymptotically stable.*

Remark 8.2 (*Trade-off between sampling and quantization*) In order to maximize the inter-sampling time for outputs, one may choose σ_o in the expression (8.10a) by selecting a large value of ε_o. However, the value ε_o closer to 1 results in the larger value of χ_o introduced in (8.16). From expression (8.22), it now follows that larger values of χ_o mean that we require a large number of quantization levels to guarantee asymptotic convergence. Hence, slower sampling allows for lesser number of quantization levels and leads to faster convergence of the parameter ν, and vice versa. In order to minimize both the sampling rate and the quantization levels, that is, increase σ_o without increasing χ, one way is to maximize the ratio $\frac{\lambda_{\min}(Q_o)}{\|P_o L\|}$ by selecting L and Q_o appropriately. Similar observation can be made for input sampling and quantization.

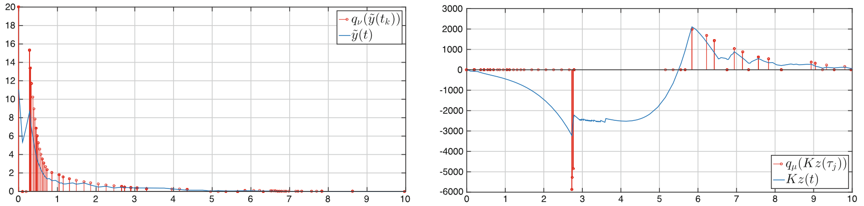
8.2.4 Illustrative Example

Consider the plant (8.1) with $A = \begin{bmatrix} 1 & 1 \\ 0 & 0.5 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \end{bmatrix}$. Since the matrix A does not have any complex eigenvalues, it suffices to take $\mathtt{n}_\mathtt{s} = 2$. For the state estimation part, we choose the output injection gain $L = [4\ 3]^\top$, $Q_o = \left[\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right]$, and $\varepsilon_o = 0.75$ which results in $P_o := \left[\begin{smallmatrix} 1.63 & -1.47 \\ -1.47 & 1.93 \end{smallmatrix}\right]$ and $\sigma_o = 0.09$. For quantization of the sampled output, we pick a quantizer q^y which rounds off the real-valued output to the nearest integer, so that $\Delta_y = 1$. The value of parameter $\alpha_o = 0.1 \frac{(1-\varepsilon_0)\lambda_{\min}(Q_o)}{\lambda_{\max}(P_o)} = 0.0038$ results in $\chi_o = 37.94$. Finally by selecting $\overline{\rho}_y = 0.975$, it is seen that the number of quantization levels required for convergence of state estimation error is

$$\frac{R_y}{\Delta_y} = 127,$$

that is, we need $\lceil \log_2(127) \rceil = 7$ bit quantizer for the output. It must be recalled that no optimality criterion was placed in obtaining the required number of bits for convergence and it could be reduced for other choices of matrices L and Q_o.

For the control input, the feedback gain matrix $K = -[6\ 4.5]$ is chosen with $Q_c = Q_o$, and $\varepsilon_c = 0.85$. This results in $P_c = \left[\begin{smallmatrix} 2.5 & 0.25 \\ 0.25 & 0.5 \end{smallmatrix}\right]$, and $\sigma_c = 0.38$. For the quantization, we again pick q^u such that its input is rounded of to the nearest integer, so that $\Delta_u = 1$. The value of parameter $\alpha_c = 0.1 \frac{(1-\varepsilon_c)\lambda_{\min}(Q_c)}{\lambda_{\max}(P_c)} = 0.0048$ results in $\chi_c = 9.94$. Finally by selecting $\overline{\rho}_u = 0.85$, it is seen that the desired value of



(a) The sampled output which is transmitted to the controller.

(b) Sampled control inputs that are transmitted to the system.

Fig. 8.3 Simulation results for event-triggered sampling and quantization in linear plants

$R_u = 318$, so we need $\lceil \log_2(318) \rceil = 9$ bit quantizer for the control input. The results of the simulation are given in Fig. 8.3. As expected, the states of the system converge to zero under the proposed algorithm, and the plots in Fig. 8.3a, b show the sampled and real values of the output and input, respectively.

8.3 Dynamic Sampling for Nonlinear Systems

We next consider nonlinear dynamical systems of the form

$$\mathscr{P} : \begin{cases} \dot{x} = f(x,u) \ , \\ y = h(x) \end{cases} \tag{8.23}$$

where x, u, y denote the state trajectory, the input, and the output respectively. For stabilization of system (8.23), we choose to work with the following class of controllers:

$$\mathscr{C} : \begin{cases} \dot{z} = g(z,u,y) \ , \\ u = k(z) \ . \end{cases} \tag{8.24}$$

Going by the approach adopted for controller design in the previous section, the dynamical system given by the first line of (8.24) plays the role of state estimator, and the control input u is some function of the estimated state variable z. The problem of stabilization of nonlinear systems with dynamic output feedback is well studied in the literature, see [3] for a survey, or [42] for various tools developed for solving this problem.

In the previous section, the algorithm that was designed to determine the output sampling times resembled a discrete-time system. Inspired by this development, the sampling algorithms that we propose in this section for nonlinear systems are based on designing auxiliary dynamical systems which determine when the next output, or input sample must be transmitted. In this regard, Fig. 8.4 lays out the sketch of the control loop that we wish to implement in this section. Before addressing this problem of designing sampling algorithms, we first describe some basic hypotheses on the system and controller data (8.24) that will be used later.

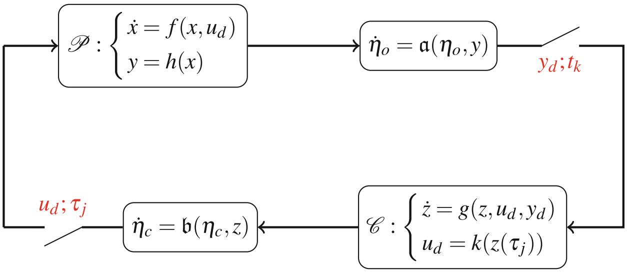

$\mathscr{P}: \begin{cases} \dot{x} = f(x, u_d) \\ y = h(x) \end{cases}$
$\dot{\eta}_o = \mathfrak{a}(\eta_o, y)$
$y_d; t_k$
$u_d; \tau_j$
$\dot{\eta}_c = \mathfrak{b}(\eta_c, z)$
$\mathscr{C}: \begin{cases} \dot{z} = g(z, u_d, y_d) \\ u_d = k(z(\tau_j)) \end{cases}$

Fig. 8.4 Feedback loop where the inputs and outputs are time-sampled and the sampling instants are determined by the dynamic filters η_c and η_o, respectively

8.3.1 Nominal Output Feedback

Basic assumptions on system (8.23) and the controller (8.24) which relate to robust (with respect to measurement errors) asymptotic stabilization of the closed-loop system are now listed.

(NL-1) The vector fields $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^n$ are continuous in each of their arguments. The function $h : \mathbb{R}^n \to \mathbb{R}^p$ is continuous and there exists a class $\mathscr{K}$ function α_h such that

$$|h(x)| \leq \alpha_h(|x|).$$

(NL-2) **An ISS state estimator**: There exist a continuously differentiable function $V_o : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, and functions $\alpha_o, \underline{\alpha}_o, \overline{\alpha}_o, \gamma_o$ of class $\mathscr{K}_\infty$, which satisfy the following inequalities:

$$\underline{\alpha}_o(|\tilde{x}|) \leq V_o(\tilde{x}) \leq \overline{\alpha}_o(|\tilde{x}|) \tag{8.25a}$$

$$\langle \nabla V_o, f(x, u) - g(z, u, y + d_y) \rangle \leq -\alpha_o(V_o(\tilde{x})) + \gamma_o(|d_y|), \tag{8.25b}$$

where $\tilde{x} := x - z$ denotes the state estimation error.

(NL-3) **An ISS control law**: There exist a continuously differentiable function $V_c : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, functions $\alpha_c, \underline{\alpha}_c, \overline{\alpha}_c, \gamma_c$ of class $\mathscr{K}_\infty$, and a state feedback control law $k : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\underline{\alpha}_c(|x|) \leq V_c(x) \leq \overline{\alpha}_c(|x|) \tag{8.26a}$$

$$\langle \nabla V_c, f(x, k(x + d_c)) \rangle \leq -\alpha_c(V_c(x)) + \gamma_c\left(\frac{|d_c|}{2}\right). \tag{8.26b}$$

(NL-4) As $r \to 0^+$, we have $(\gamma_c \circ \underline{\alpha}_o^{-1})(r) = O(\alpha_o(r))$, that is, there exists a constant $M > 0$ such that

$$\lim_{r\to 0^+} \frac{(\gamma_c \circ \underline{\alpha}_o^{-1})(r)}{\alpha_o(r)} \leq M. \tag{8.27}$$

Hypotheses **(NL-2)** and **(NL-3)** allow us to decompose the problem of dynamic output feedback into two components: first is to design a state estimator, and then apply a static control law which is robust with respect to measurement errors. Designing control laws which are ISS with respect to measurements of state variable has remained a topic of major interest in the control community and several techniques now exist depending on the system class. The state estimators, that one typically designs for a nonlinear system (using high-gain, or passivity approach), are robust with respect to output measurement error but the estimate of the form (8.25b) is typically not stated in such works. We refer the reader to a recent paper [35] which deals with designing estimators of this form. The assumption **(NL-4)** is imposed to construct a particular Lyapunov function for the closed-loop system. This construction is inspired by [37] and it allows us to invoke arguments related to the stability of the cascaded nonlinear systems. If the functions $\underline{\alpha}_c$ and γ_c are quadratic, and α_o is linear, as one would usually obtain in the linear case with quadratic V_o and V_c, then **(NL-4)** is satisfied.

8.3.2 Sampling Algorithms

To design sampling algorithms, we introduce the following auxiliary dynamical system:

$$\dot{\eta}_o := -\beta_o(\eta_o) + \rho_o(|y(t)|) + \gamma_o(|y(t) - y(t_k)|) \tag{8.28a}$$

and on the controller side

$$\dot{\eta}_c := -\beta_c(\eta_c) + \rho_c\left(\frac{|z(t)|}{2}\right) + \gamma_c(|z(t) - z(\tau_j)|) \tag{8.28b}$$

with initial conditions $\eta_o(0) > 0$, and $\eta_c(0) > 0$. In the above equations, β_o, ρ_o, ρ_c, and β_c are all functions of class $\mathcal{K}$, which would be specified later. One may notice that, if $\beta_o(r) = \alpha_o(r)$ and $\beta_c(r) = \alpha_c(r)$ are linear, then in the light of **(NL-2)** and **(NL-3)**, the dynamic filters in (8.28a) and (8.28b) play the role of norm estimators [38] for error dynamics $\tilde{x} = x - z$, and the closed-loop dynamics for the state x, respectively.

We use the sample-and-hold strategy for sampling, that is, the outputs and inputs are updated at certain discrete times, and in between updates, they are held constant. The algorithms that determine the sampling instants for inputs and outputs can now be defined as a function of η_o, η_c given by (8.28).

Output Sampling Rule:
It is assumed that the output sent to the controller is updated at time instants t_k, $k \in \mathbb{N}$, which are defined inductively as:

$$t_{k+1} := \inf\{t > t_k \mid |y(t) - y(t_k)| \geq \sigma_o(\eta_o(t))\}, \tag{8.29}$$

where $\sigma_o : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is some positive definite, nondecreasing function to be specified later.

Input Sampling Rule:
The control input $u(\cdot)$ is updated at time instants $\tau_j, j \in \mathbb{N}$, according to the following rule:

$$\tau_{j+1} := \inf\{t > \tau_j \mid |z(t) - z(\tau_j)| \geq \sigma_c(\eta_c(t))\}, \tag{8.30}$$

where $\sigma_c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a positive definite and nondecreasing function which will be specified later.

8.3.3 Stability Analysis

Using the sampling algorithms from the previous section, the dynamics of the closed-loop system are now written in the framework of hybrid systems [13], where we specify the continuous and discrete dynamics, along with their respective domains. We then invoke tools from the literature related to the stability of such systems to show that for certain choice of the design parameters in (8.28) and appropriately chosen functions σ_o, σ_c in (8.29), (8.30), the origin of the closed-loop system is asymptotically stable.

8.3.3.1 Hybrid Model of the System

Using y_d and z_d to denote the sampled output and sampled controller state, respectively, we can let $\overline{x} := (x, z, \eta_c, \eta_o, y_d, z_d) \in \mathbb{R}^{\overline{n}}$, where $\overline{n} = 3n + p + 2$, is the augmented state variable for the closed-loop system. The flow set $\mathfrak{C}$ for the state variables (where they all satisfy a certain ordinary differential equation) is defined as

$$\mathfrak{C} := \mathfrak{C}_o \cap \mathfrak{C}_c \cap \mathfrak{C}_\eta,$$

where we define

$$\mathfrak{C}_o := \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid |h(x) - y_d| \leq \sigma_o(\eta_o)\}, \tag{8.31a}$$

$$\mathfrak{C}_c := \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid |z - z_d| \leq \sigma_c(\eta_c)\}, \tag{8.31b}$$

$$\mathfrak{C}_\eta := \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid \eta_o \geq 0 \wedge \eta_c \geq 0\}. \tag{8.31c}$$

The jump set $\mathfrak{D}$ where the state variables may get reset is given by:

$$\mathfrak{D} := \mathfrak{D}_c \cup \mathfrak{D}_o,$$

where

$$\mathfrak{D}_o := \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid |h(x) - y_d| \geq \sigma_o(\eta_o)\} \tag{8.32a}$$

$$\mathfrak{D}_c := \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid |z - z_d| \geq \sigma_c(\eta_c)\}. \tag{8.32b}$$

Clearly, the sets $\mathfrak{C}$ and $\mathfrak{D}$ are closed. By construction, the jump set $\mathfrak{D}$ for the closed-loop hybrid system also allows for two jumps simultaneously, that is, y_d and z_d may get updated at the same time instant. The corresponding sets of differential and difference equation on these sets are:

$$\overline{x} \in \mathfrak{C} : \begin{cases} \dot{x} &= f(x, k(z_d)) \\ \dot{z} &= g(z, k(z_d), y_d) \\ \dot{z}_d &= 0 \\ \dot{y}_d &= 0 \\ \dot{\eta}_o &= -\beta_o(\eta_o) + \rho_o(|h(x)|) + \gamma_o(|h(x) - y_d|) \\ \dot{\eta}_c &= -\beta_c(\eta_c) + \rho_c\left(\frac{|z|}{2}\right) + \gamma_c(|z - z_d|) \end{cases} \tag{8.33a}$$

$$\overline{x} \in \mathfrak{D}_c : \left\{ z_d^+ = z \right. \tag{8.33b}$$

$$\overline{x} \in \mathfrak{D}_o : \left\{ y_d^+ = h(x). \right. \tag{8.33c}$$

The closed-loop system (8.33) satisfies the basic assumptions listed in [13, Assumption 6.5], and is, hence, nominally well posed.

8.3.3.2 Design of Sampling Functions

The choice of functions σ_o, σ_c depends on the construction of a function q which will also be used to define the Lyapunov function of system (8.33). Under assumption **(NL-4)**, it is possible to introduce a continuous nondecreasing function $q : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that satisfies [37, Lemma 2]:

$$q(s) \geq \frac{4(\gamma_c \circ \underline{\alpha}_o^{-1})(s)}{\alpha_o(s)}.$$

The functions β_o, β_c, σ_o, σ_c, ρ_o and ρ_c under consideration should satisfy the following design criteria for the stability result to follow:

(D1) Let β_o and β_c be two smooth functions of class $\mathcal{K}$;
(D2) Let θ_o be a function of class $\mathcal{K}_\infty$ defined as

$$\theta_o(r) := \alpha_o^{-1}(2\gamma_o(r)).$$

Choose the functions σ_o and σ_c in (8.31) and (8.32) such that for some $\varepsilon \in (0,1)$, and for each $s \geq 0$:

$$(\gamma_o \circ \sigma_o)(s) \cdot \left[1 + (q \circ \theta_o \circ \sigma_o)(s)\right] \leq (1-\varepsilon)\beta_o(s)$$
$$2(\gamma_c \circ \sigma_c)(s) \leq (1-\varepsilon)\beta_c(s).$$

(D3) The functions ρ_o and ρ_c in (8.33) are chosen such that for each $s \geq 0$:

$$0 \leq (\rho_o \circ \alpha_h \circ \underline{\alpha}_c^{-1})(s) \leq (1-2\varepsilon)\alpha_c(s),$$
$$0 \leq \rho_c(s) \leq \min\left\{(1-\varepsilon)\gamma_c(s), \varepsilon\alpha_c(\underline{\alpha}_c(s))\right\}.$$

The basic idea behind the aforementioned design criteria is to work with the following candidate Lyapunov function:

$$V(\overline{x}) := l(V_o(\tilde{x})) + V_c(x) + \eta_o + \eta_c, \tag{8.34}$$

where $l : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is defined as

$$l(s) := \int_0^s q(r)dr.$$

Since q is a continuous nondecreasing function, it follows that $l(\cdot)$ is a continuously differentiable function of class $\mathcal{K}_\infty$. The foregoing bounds are introduced to obtain $\dot{V} \leq 0$ during the flows, whereas, by construction, $V^+ = V^-$ when a jump occurs. Arguments based on LaSalle's invariance principle can then be invoked to show that the origin is asymptotically stable.

Theorem 8.2 *Consider the closed-loop system (8.33) under the hypotheses **(NL-1)**, **(NL-2)**, **(NL-3)**, and **(NL-4)**. If the functions $\beta_o, \beta_c, \rho_o, \rho_c$ in (8.33a) and the functions σ_o, σ_c in (8.31c) are chosen to meet the design criteria **(D1)**, **(D2)**, and **(D3)**, then the origin $\{0\} \in \mathbb{R}^{\overline{n}}$ is globally asymptotically stable (GAS) for the closed-loop system (8.33).*

8.3.4 Dwell Time Between Sampling Instants

We next want to show that the proposed sampling algorithms given in Sect. 8.3.2 do not lead to the accumulation of jump times over a finite time interval, and over each compact interval, there exists a lower bound on inter-sampling times. Our strategy for showing the existence of minimal inter-sampling time is primarily based on the approach adopted in [39]. However, unlike [39], we do not get an autonomous differential inequality that gives uniform lower bounds; instead, we obtain time-varying differential inequalities, and hence the inter-sampling times depend upon the interval under consideration.

For the proposed sampling routines, the minimum time between two output (respectively, control input) updates is the time taken by the term $\frac{|y(t)-y_d(t)|}{\sigma_o(\eta_o(t))}$ (resp. $\frac{|z(t)-z_d(t)|}{\sigma_c(\eta_c(t))}$) to go from 0 to 1, after each time y_d has been reset to current value of y (resp. z_d has been reset to z). In order to derive lower bounds on minimal time between updates, we will first introduce certain assumptions on the gain functions given in Sect. 8.3.1 and the ones used to define sampling instants in Sect. 8.3.2.

(A1) The system dynamics defined by f, h, and the controller functions g, k are bounded by a linear growth rate, which allow us to write $|f(x, k(x + d_c))| \le L_{fk}(|x| + |d_c|)$ and $|g(z, k(z + d_z), h(x) + d_y)| \le L_{gk}(|z| + |d_z|) + L_{gh}(|x| + d_y)$. Also, $\|\partial h/\partial x\|$ is bounded by a constant.

(A2) The functions α_o and α_c are linear, and $\alpha_o(r) < \alpha_c(r)$.

(A3) The function $\gamma_c \circ \underline{\alpha}_o^{-1}$ is bounded by a linear growth rate: There exists $L_{co} > 0$ such that $\gamma_c(\underline{\alpha}_o^{-1}(r)) \le L_{co}\, r$.

(A4) The functions σ_o, σ_c are same up to multiplication by a constant $C > 0$, that is, $\sigma_o = C\sigma_c$. Furthermore, let $\sigma := \min\{\sigma_o, \sigma_c\}$, and assume there are constants $C_{\sigma,1}, C_{\sigma,2} > 0$, that satisfy

$$\sigma(r) \ge C_{\sigma,1} \max\{\underline{\alpha}_c^{-1}(r), \underline{\alpha}_o^{-1}(r)\} \tag{8.35a}$$

$$\sigma'(r) \cdot r \le C_{\sigma,2}\, \sigma(r). \tag{8.35b}$$

In addition, there exists a continuous locally integrable function $\chi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ such that for every $r, s \ge 0$

$$\frac{\sigma(r)}{\sigma(s)} \le \chi\left(\frac{r}{s}\right). \tag{8.35c}$$

Remark 8.3 (*How restrictive are **(A1)–(A4)**?*) One typically requires f, g to be locally Lipschitz for existence of solutions, which would ensure that the linear growth rate condition holds on every compact set. A global linear bound (which is satisfied for globally Lipschitz functions) has been introduced to avoid the semi-global arguments in this chapter. For **(A2)**, note that it is always possible to modify the Lyapunov functions V_o, V_c so that the dissipation functions α_o and α_c are linear, see [27, Lemma 12]. However, this would also modify the gain function γ_c and one must be careful in verifying hypothesis **(NL-4)**. The most restrictive aspect of our approach is to verify **(A4)**: As one would usually observe in the linear case, if $\underline{\alpha}_c(r) = \underline{\alpha}_o(r) = r^2$, then one can choose $\sigma(r) = \sqrt{r}$ (modulo multiplication with certain constants). In general, having $\sigma(r) = r^\alpha$, for $\alpha > 0$, would satisfy (8.35b) and (8.35c). To find a constructive proof for existence of such σ is a topic of ongoing work.

In the light of these assumptions, we now impose the following additional criteria on the design functions introduced in (8.28a), (8.28b), (8.29) and (8.30):

(D4) The functions β_o, β_c are linear and for each $r \geq 0$

$$\beta_o(r) \leq \alpha_o(r) < \beta_c(r) \leq \alpha_c(r).$$

(D5) The functions ρ_o and ρ_c are chosen such that

$$\begin{aligned}\max\{\rho_o \circ \underline{\alpha}_c^{-1}(r), \rho_o \circ \underline{\alpha}_o^{-1}(r)\} &\leq C_\rho\, r\\ \max\{\rho_c \circ \underline{\alpha}_c^{-1}(r), \rho_c \circ \underline{\alpha}_o^{-1}(r)\} &\leq C_\rho\, r\end{aligned}$$

for some constants $C_\rho > 0$.

The main result on Zeno-freeness now follows.

Theorem 8.3 *If, in addition to the hypotheses of Theorem 8.2, assumptions **(A1)**–**(A4)** hold and the functions $\beta_o, \beta_c, \rho_o, \rho_c$ satisfy **(D4)** and **(D5)**, then there is no accumulation point of the sampling times for outputs and inputs over a compact interval.*

8.3.5 Example

In order to demonstrate our design, and observe practical feasibility of our algorithms, we take a nonlinear system with globally Lipschitz vector field. The calculations carried out in this example would carry over to linear systems with very slight modification. Consider the system

$$\begin{aligned}\dot{x}_1 &= x_2 + 0.25\,|x_1|\\ \dot{x}_2 &= \mathrm{sat}(x_1) + u\end{aligned}$$

with $y = x_1$. The notation $\mathrm{sat}(x_1)$ denotes the saturation function, that is, $\mathrm{sat}(x_1) = \min\{1, \max\{-1, x_1\}\}$. The nominal output feedback controller is:

$$\begin{aligned}\dot{z}_1 &= z_2 + 0.25\,|y| + l_1(y - z_1)\\ \dot{z}_2 &= \mathrm{sat}(y) + u + l_2(y - z_1)\\ u &= k(z) = \mathrm{sat}(z_1) - k_1 z_1 - k_2 z_2,\end{aligned}$$

where we pick $L^\top := [l_1\ l_2] = [2\ 2]$, and $K := [k_1\ k_2] = L^\top$. By choosing, $V_o(\tilde{x}) = \tilde{x}^\top P_o \tilde{x}$ and $V_c(x) = x^\top P_c x$, with $P_o = \left[\begin{smallmatrix} 2 & -1\\ -1 & 1\end{smallmatrix}\right]$ and $P_c = \left[\begin{smallmatrix} 1 & 0.5\\ 0.5 & 0.5\end{smallmatrix}\right]$, hypotheses **(NL-2)** and **(NL-3)** hold. Indeed, if the controller is driven by the sampled output y_d, then

$$\dot{V}_o \leq -\alpha_o V_o + \overline{\gamma}_o |y_d - y|^2,$$

where $\overline{\gamma}_o := \frac{\|P_o L\|^2}{p_o \lambda_{\min}(P_o)}$, and $\alpha_o + p_o = 2$. Similarly, with $u_d = k(z_d)$, the derivative of $V_c(x) = x^\top P_c x$ satisfies

$$\dot{V}_c \leq -\alpha_c V_c + \overline{\gamma}_c |z_d - z|^2 + \overline{\gamma}_c |\tilde{x}|^2,$$

where $\overline{\gamma}_c := \frac{2\|P_c BK\|^2}{p_c \lambda_{\min}(P_c)}$, and $\alpha_c + p_c = 2 - \|P\|$.

For sampling algorithms, we can let $q(s) = \overline{q} := \frac{4\overline{\gamma}_c}{\alpha_o \lambda_{\min}(P_o)}$ to be the constant function and consider the following dynamic filters which satisfy the design criteria **(D1)–(D3)**:

$$\begin{aligned}
\dot{\eta}_o &= -\alpha_o \eta_o + \overline{\rho}_o |y(t)|^2 + \overline{\gamma}_o |y - y_d|^2 \\
\dot{\eta}_c &= -\alpha_c \eta_c + \overline{\rho}_c \frac{|z|^2}{4} + \overline{\gamma}_c |z - z_d|^2,
\end{aligned}$$

where in the notation of (8.28), we have chosen $\beta_o(r) = \alpha_o\, r$ and $\beta_c(r) = \alpha_c\, r$ for simplicity. Also, we let $\overline{\rho}_o := \frac{(1-2\varepsilon)\lambda_{\min}(P_o)}{\|C\|^2}$ and $\overline{\rho}_c := \min\{(1-\varepsilon)\overline{\gamma}_c, \varepsilon\, \alpha_c \lambda_{\min}(P_c)\}$. The jump sets for the closed-loop system which determine the sampling times are now defined as follows:

$$\begin{aligned}
\mathfrak{D}_o &= \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid |h(x) - y_d| \geq \overline{\sigma}_o \sqrt{\eta_o}\,\} \\
\mathfrak{D}_c &= \{\overline{x} \in \mathbb{R}^{\overline{n}} \mid |z - z_d| \geq \overline{\sigma}_c \sqrt{\eta_c}\,\}
\end{aligned}$$

where $\overline{\sigma}_o := \frac{(1-\varepsilon)\alpha_o}{(1+\overline{q})\overline{\gamma}_o}$, and $\overline{\sigma}_c = \frac{(1-\varepsilon)\alpha_c}{2\overline{\gamma}_c}$. From Theorem 8.2, asymptotic stability of the closed-loop system with sampled outputs and inputs now follows. It can also be verified that **(A1)–(A4)** hold by construction.

The results of the simulation appear in Fig. 8.5. We observed that, even though the constants $\overline{\sigma}_o$ and $\overline{\sigma}_c$ are relatively small in magnitude, it was possible to slow down the sampling rate by increasing the initial values of η_o, and η_c. Also, the Lyapunov functions $V_o(\tilde{x})$ and $V_c(x)$ are not always decaying but the function V in (8.34) associated with the closed loop is indeed decaying with time.

8.4 Event-Triggered Sampling in Hyperbolic Systems

Moving from the class of finite dimensional to infinite dimensional systems, we now consider the problem of designing sampling algorithm for control of linear hyperbolic plants described by the following equations:

$$\mathscr{P} : \begin{cases} \partial_t x(w,t) + \Lambda \partial_w x(w,t) = 0\ , \ w \in [0,1],\ \ t \geq 0 \\ x(0,t) = Hx(1,t) + Bu(t)\ , \quad t \geq 0 \\ y(t) = x(1,t)\ , \qquad\qquad\qquad t \geq 0. \end{cases} \tag{8.36}$$

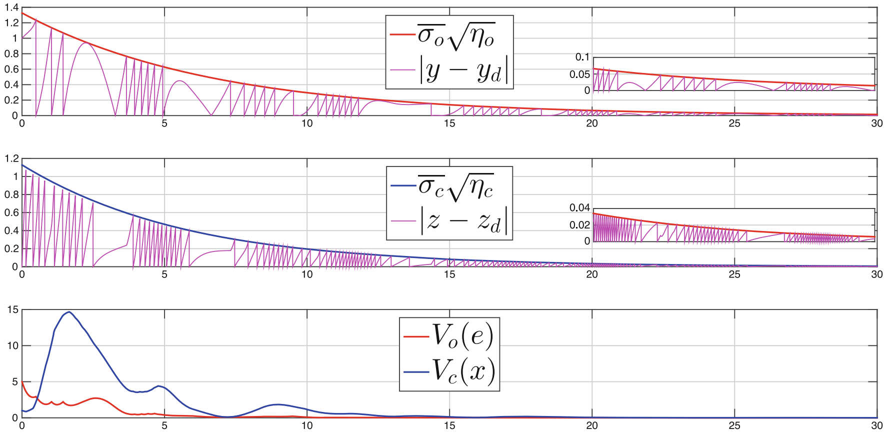


Fig. 8.5 Simulation results: In the *top plot*, whenever $|y(t) - y_d(t)|$ reaches the sampling threshold $\overline{\sigma}_o\sqrt{\eta_o(t)}$, y_d is updated. The *middle plot* shows $\overline{\sigma}_c\sqrt{\eta_c(t)}$ and $|z(t) - z_d(t)|$. The *bottom plot* shows time evaluation of Lyapunov functions $V_o(e)$ and $V_c(x)$ which are not always decaying

In (8.36), $x(w, t)$ in $\mathbb{R}^n$ is the state depending on the space variable $w \in [0, 1]$ and on the time variable $t \geq 0$, $y(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ denote, respectively, the measured output and boundary control input at time t. The matrices $H \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ define the boundary condition, and $\Lambda \in \mathscr{D}_{n,+}$, where $\mathscr{D}_{n,+}$ denotes the set of diagonal positive definite matrices.

8.4.1 Nominal Output Feedback Law

The presence of the control input in (8.36) allows us to modify the boundary condition $x(0, t)$, which is then transported according to the rate determined by the entries of the matrix Λ. Within the class of linear output feedback controllers, if we define

$$u(t) = Ky(t) \tag{8.37}$$

for some suitably chosen matrix $K \in \mathbb{R}^{n\times m}$, then the boundary condition of the closed-loop system is described by

$$x(0, t) = Gx(1, t)\ , \quad t \geq 0, \tag{8.38}$$

where $G := H + BK$. Such feedback laws for boundary control have been found to be useful in various applications, most notably the traffic flow control [11, 15], and the open-channel regulation [8, 29, 30] among others.

Several techniques are available in the literature to design the feedback gain matrix K for asymptotic stability of the closed-loop system, see the books [16, 21].

For our purposes, the design methodology based on the appropriate choice of Lyapunov functions [9, 14] is more relevant. Since the event-triggered sampling is built on working with Lyapunov function for closed-loop system, we choose to work with the Lyapunov function proposed in [7]. Based on that, a sufficient condition, which guarantees asymptotic stability of the origin (w.r.t. $L^2(0,1)$ norm), is to choose K such that

$$\rho_2(G) < 1, \tag{8.39}$$

where $\rho_2(G) = \inf\{\|\varDelta G \varDelta^{-1}\|,\ \varDelta \in \mathscr{D}_{n,+}\}$. As a particular case of [7, Theorem 2], this condition is obtained from working with the following Lyapunov function candidate V defined for all x in $L^2(0,1)$ by

$$V(x) = \int_0^1 e^{-\mu w} x^\top P x \, dw \tag{8.40}$$

where P is a suitable matrix in $\mathscr{D}_{n,+}$ and $\mu > 0$ is sufficiently small. To see the role of condition (8.39) in ensuring stability, we compute the time derivative of V along the solutions of (8.36) with the boundary condition (8.38):

$$\begin{aligned} \dot V &= -2 \int_0^1 x(w,t)^\top P \varLambda \partial_w x(w,t) e^{-\mu w} dw \\ &= -\left[x(w,t)^\top P \varLambda x(w,t) e^{-\mu w}\right]_0^1 - \mu \int_0^1 x(w,t)^\top P \varLambda x(w,t) e^{-\mu w} dw, \end{aligned}$$

where the integration by parts has been used together with the property $P\varLambda = \varLambda P$ (which is implied by P and $\varLambda$ diagonal). Now since $P\varLambda$ and P are positive definite, there exists $\alpha_c > 0$ such that $-\mu P \varLambda \leq -\alpha_c P$. Using the boundary condition (8.38), and the fact that P, $\varLambda$ are diagonal, we obtain

$$\begin{aligned} \dot V &\leq x(1,t)^\top \left[G^\top P \varLambda G - P \varLambda e^{-\mu}\right] x(1,t) - \alpha_c V \\ &\leq x(1,t)^\top (P\varLambda)^{1/2} \left[(P\varLambda)^{-1/2} G^\top P \varLambda G (P\varLambda)^{-1/2} - e^{-\mu} I\right] (P\varLambda)^{1/2} x(1,t) - \alpha_c V, \end{aligned} \tag{8.41}$$

where $I \in \mathbb{R}^{n \times n}$ denotes the identity matrix. Due to the condition $\rho_2(G) < 1$, there exists $\varDelta$ in $\mathscr{D}_{n,+}$ such that $\|\varDelta G \varDelta^{-1}\| < 1$ which implies that the matrix $\varDelta^{-1} G^\top \varDelta \varDelta G \varDelta^{-1} - I$ is negative definite. Hence, there exists $\mu > 0$ such that

$$\varDelta^{-1} G^\top \varDelta \varDelta G \varDelta^{-1} - e^{-\mu} I \leq 0\,. \tag{8.42}$$

Now define P as $P = \varDelta^2 \varLambda^{-1}$, it follows from (8.41) and (8.42) that

$$\dot V \leq -\alpha_c V$$

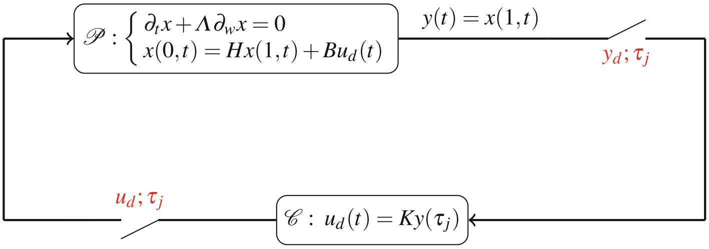

$\mathscr{P}: \begin{cases} \partial_t x + \Lambda \partial_w x = 0 \\ x(0,t) = Hx(1,t) + Bu_d(t) \end{cases}$
$y(t) = x(1,t)$
$y_d; \tau_j$
$u_d; \tau_j$
$\mathscr{C}: u_d(t) = Ky(\tau_j)$

Fig. 8.6 Feedback loop with time-sampled boundary control. The output and input sampling instants are synchronized due to static control

and thus the origin is exponentially stable in L^2 norm for the system (8.36) with the boundary condition (8.38), if (8.39) holds.

8.4.2 Sampling Algorithm

In contrast to the dynamic controllers for finite dimensional systems, the nominal control law chosen for (8.36) is actually a static one. For this reason, it makes sense to choose the sampling times for the output y and the control input u to be the same. We are thus interested in computing a sequence of sampling times $\{\tau_j\}_{j\in\mathbb{N}}$ such that the piecewise constant control input u_d, defined as,

$$\mathscr{C}: \begin{cases} u_d(t) = 0, & \forall\, t \in [\tau_0, \tau_1), \\ u_d(t) = Ky(\tau_j), & \forall\, t \in [\tau_j, \tau_{j+1}), \quad j \geq 1, \end{cases} \tag{8.43}$$

renders the plant (8.36) asymptotically stable in closed loop.[3] This control architecture is graphically illustrated in Fig. 8.6.

To compute the sampling times, we first rewrite the boundary condition $x(0,t)$ in the presence of sampling error as follows:

$$\begin{aligned} x(0,t) &= Hx(1,t) + BKx(1,\tau_j) \\ &= (H + BK)x(1,t) + BK(-x(1,t) + x(1,\tau_j)) \\ &= Gx(1,t) + d(t,\tau_j), \end{aligned} \tag{8.44}$$

[3] Instead of letting $u_d(t) = 0$ for all t in $[0, \tau_1)$ in the first line of (8.43), other choices may be possible, as $u_d(t) = Ky(0)$ for all t in $[0, \tau_1)$. However, we were not able to find a choice such that the estimation (8.47) in Theorem 8.4 below holds for all t in $[0, \tau_1)$, but only for all $t \geq \tau_1$. This estimation of the time derivative of the Lyapunov function V along the solutions to (8.36) in closed-loop with (8.43) is indeed a key step in our proof.

where $d(t, \tau_j) := BK(-x(1,t) + x(1, \tau_j))$ is seen as a perturbation due to sampling. A natural question is how to derive an ISS property with respect to d, from the Lyapunov function V defined for the nominal system. ISS properties for hyperbolic systems have been studied for specific cases in [31, 33]. To obtain the desired ISS estimate in the current context, we work with the Lyapunov function V introduced in (8.40), where the parameters P and μ are fixed as obtained in the nominal case. The time derivative of V along the solutions of (8.36) with perturbed boundary condition (8.44) yields the following modification of (8.41):

$$\begin{aligned}\dot{V} \leq x(1,t)^\top (P\Lambda)^{1/2} \left[(P\Lambda)^{-1/2} G^\top P\Lambda G (P\Lambda)^{-1/2} - e^{-\mu} I\right] (P\Lambda)^{1/2} x(1,t) \\ + d(t,\tau_j)^\top P\Lambda d(t,\tau_j) - \alpha_c V.\end{aligned}$$

Therefore, under the condition $\rho_2(G) < 1$, with the same definition for α_c as in the nominal case, it holds that

$$\dot{V} \leq -\alpha_c V(t) + \gamma_c \|d(t,\tau_j)\|^2,$$

where $\gamma_c > 0$ is the largest eigenvalue of $P\Lambda$. We rewrite the foregoing inequality as

$$\dot{V} \leq -\frac{\alpha_c}{2} V + \left(-\frac{\alpha_c}{2} V + \gamma_c \|d(t,\tau_j)\|^2\right),$$

which leads us to define the increasing sequence of sampling times $\{\tau_j\}_{j\in\mathbb{N}}$. We let $\tau_0 = 0$, $\tau_1 = \frac{1}{\underline{\lambda}}$ (where $\underline{\lambda}$ denotes the smallest term on the diagonal of Λ), and τ_{j+1} for $j \geq 1$ is defined iteratively as

$$\tau_{j+1} = \inf\left\{t > \tau_j \mid \gamma_c \|d(t,\tau_j)\|^2 \geq \frac{\alpha_c}{2} V + \eta_o e^{-\beta_o t}\right\}. \tag{8.45}$$

In (8.45), the parameters η_o and β_o, which make the sampling rule dynamic, are to be chosen appropriately (see Theorem 8.4 below).

The definition of the sequence $\{\tau_j\}_{j\in\mathbb{N}}$ depends on the values of the Lyapunov function $V(t)$ at time t and the state of an external filter $\dot{\eta}_o = -\beta_o \eta_o$. As far as $V(x(.,t))$ is concerned, using (8.36) and the boundary condition (8.44), it holds that (with a slight abuse of notation)

$$V(x(.,t)) = \sum_{i=1}^{n} p_i \int_0^1 \left(H_i y(t - \tfrac{w}{\lambda_i}) + B_i u_d(t - \tfrac{w}{\lambda_i})\right)^2 e^{-\mu w} dw, \tag{8.46}$$

where $P = \operatorname{diag}(p_i)$, and H_i and B_i denote the i-th row of the matrices H and B, respectively. Therefore the definition (8.45) of the sequence $\{\tau_j\}_{j\in\mathbb{N}}$ is a function of the current output, the boundary control, and of the auxiliary variable η_o only.

We can now state the main result of this section. See [10] for a complete proof which is quite tedious, in particular, the proof of the well posedness in appropriate state space preventing accumulation of discontinuities (and thus preventing Zeno solutions).[4]

Theorem 8.4 *Consider the plant $\mathscr{P}$ defined by (8.36) in closed-loop with the controller $\mathscr{C}$ defined by (8.43). Under the assumption $\rho_2(G) < 1$, and for suitable choices of the matrix $P \in \mathscr{D}_{n,+}$, and positive scalars μ, β_o and $\eta_o(0)$, the closed-loop system has a unique solution $x(.,t)$ in $L^2(0,1)$ for all $t \geq 0$, and for all initial condition $x(0,t) \in \mathscr{C}_{lpw}(0,1)$. Moreover, the origin is globally exponentially convergent, that is, there exist $\overline{\alpha}_c > 0$, $C_1 > 0$ and $C_2 > 0$ such that for every $x^0 \in \mathscr{C}_{lpw}(0,1)$, the solution of (8.36) in closed loop with (8.43) satisfies, for all $t \geq 0$,*

$$\|x(\cdot,t)\|_{L^2(0,1)} \leq \left(C_1 \|x(0,t)\|_{L^2(0,1)} + C_2\right) e^{-\overline{\alpha}_c t} \,.$$

Finally, for the solution of the closed-loop system (8.36), (8.43), it holds that, $\forall\, t \geq \frac{1}{\underline{\lambda}}$,

$$\dot{V} \leq -(1-\sigma)\alpha_c V(t) + \overline{\eta}_o e^{-\beta_o t} \tag{8.47}$$

for a suitable σ in $(0,1)$, and $\overline{\eta}_o > 0$.

Note that, Theorem 8.4 does not state any stability property due to the presence of the term $e^{-\beta_o t}$. Modifying the sampling algorithm to get the global asymptotic stability of the origin is a topic of actual research.

8.4.3 Numerical Simulations

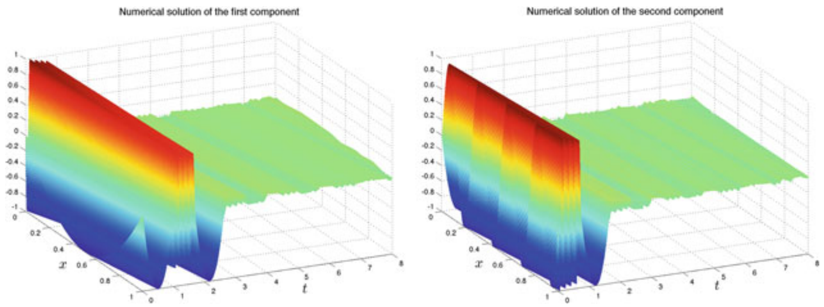
Let us consider a 2 × 2 linear hyperbolic system of conservation laws

$$\partial_t x(w,t) + \Lambda \partial_w x(w,t) = 0 \qquad w \in [0,1], \quad t \geq 0$$

where $x(w,t) \in \mathbb{R}^2$, $\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{2} \end{bmatrix}$, $H = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top$, $K = \begin{bmatrix} -1 & 0 \end{bmatrix}$. We may check the condition $\rho_2(G) < 1$, which implies the exponential convergence of the origin of (8.36) in closed-loop with the nominal controller (8.37), and also the global convergence of the origin of (8.36) in closed loop with time-sampled controller (8.43) is obtained for suitable choices of the parameters considered in Theorem 8.4.

For simulation, let the initial condition be defined by $x(w,0) = \begin{bmatrix} 4w(w-1) & \sin(8\pi w) \end{bmatrix}^\top$ for $w \in [0,1]$. The solution of the closed-loop system (8.36) and (8.43) is numerically computed using a Weighted Essentially Non-Oscillatory scheme (as done in [20]). The convergence of the state x is observed in Fig. 8.7, where the first and the second components of the solution are given.

[4] The well-posed property for solutions has to be understood in the set $\mathscr{C}_{lpw}(0,1)$ of piecewise left-continuous functions. In particular there are only a finite number of discontinuities in any bounded time interval.



(a) The first state component $x_1(w,t)$. (b) The second state component $x_2(w,t)$.

Fig. 8.7 Simulation results for the solution x of (8.36) in closed loop with (8.43)

For some additional discussions on the triggering condition and the numerical computation of the positive inter-execution time (between two updates of the input of plant $\mathscr{P}$), please see [10].

8.5 Conclusion

Novel techniques within the paradigm of event-triggered algorithms have been suggested for sampled-data control of the following classes of plants: linear control systems; nonlinear control systems; and linear hyperbolic systems. In contrast to existing methods, the use of dynamic filters was proposed to overcome certain limitations seen while using static inequalities to determine sampling times. The analysis and design were built on appropriately chosen ISS Lyapunov functions for the closed-loop system. To show that such schemes are robust to communication errors in case of finite dimensional linear systems, the transmitted inputs and outputs are also dynamically quantized to ensure asymptotic stability of the closed-loop system. For all the suggested control algorithms, it is analytically proven that the Zeno solutions do not occur, and hence the sampling times do not possess an accumulation point over any finite time interval.

Acknowledgements The presentation given in MinesParisTech, France in July 2015 by the first author and this companion chapter are dedicated to Laurent Praly on the occasion of his sixtieth birthday. The authors would like to thank Laurent for many inspiring discussions. This chapter has collected some results contained in [10, 40, 41] obtained in collaboration with N. Espitia, M. Fiacchini, A. Girard, N. Marchand, and A. Teel.

References

1. Abdelrahim, M., Postoyan, R., Daafouz, J.: Event-triggered control of nonlinear singularly perturbed systems based only on the slow dynamics. Automatica **52**, 15–22 (2015)
2. Andrieu, V., Nadri, M., Serres, U., Vivalda, J.-C.: Self-triggered continuous discrete observer with updated sampling period. Automatica **62**, 106–113 (2015)
3. Andrieu, V., Praly, L.: A unifying point of view on output feedback designs for global asymptotic stabilization. Automatica **45**(8), 1789–1798 (2009)
4. Anta, A., Tabuada, P.: To sample or not to sample: self-triggered control for nonlinear systems. IEEE Trans. Autom. Control **55**(9), 2030–2042 (2010)
5. Astolfi, D., Praly, L.: Output feedback stabilization with an observer in the original coordinates for nonlinear systems. In: Proceedings of the 52nd IEEE Conference on Decision and Control, pp. 5927–5932, Firenze, Italy (2013)
6. Astrom, K.: Event based control. In: Analysis and Design of Nonlinear Control Systems, pp. 127–148. Springer (2008)
7. Bastin, G., Coron, J.-M., d'Andréa Novel, B.: Using hyperbolic systems of balance laws for modeling, control and stability analysis of physical networks. In: 17th IFAC World Congress Lecture notes for the Pre-Congress Workshop on Complex Embedded and Networked Control Systems. Seoul, Korea (2008)
8. Bastin, G., Coron, J.-M., d'Andréa Novel, B.: On Lyapunov stability of linearised Saint-Venant equations for a sloping channel. Netw. Heterog. Media **4**(2), 177–187 (2009)
9. Coron, J.-M., d'Andréa Novel, B., Bastin, G.: A strict Lyapunov function for boundary control of hyperbolic systems of conservation laws. IEEE Trans. Autom. Control **52**(1), 2–11 (2007)
10. Espitia, N., Girard, A., Marchand, N., Prieur, C.: Event-based control of linear hyperbolic systems of conservation laws. Automatica **70**, 275–287 (2016)
11. Garavello, M., Piccoli, B.: Traffic Flow on Networks. American Institute of Mathematical Sciences, Springfield (2006)
12. Goebel, R., Hespanha, J., Teel, A.R., Cai, C., Sanfelice, R.: Hybrid systems: generalized solutions and robust stability. In: IFAC Symposium on Nonlinear Control Systems (NOLCOS), pp. 1–12, Stuttgart, Germany (2004)
13. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid Dynamical Systems: Modeling, Stability, and Robustness. Princeton University Press (2012)
14. Gugat, M., Herty, M.: Existence of classical solutions and feedback stabilization for the flow in gas networks. ESAIM: Control Optim. Calc. Var. **17**(1), 28–51 (2011)
15. Gugat, M., Herty, M., Klar, A., Leugering, G.: Optimal control for traffic flow networks. J. Optim. Theory Appl. **126**(3), 589–616 (2005)
16. Hale, J.K., Lunel, S.M.V.: Introduction to functional differential equations, vol. 99. Springer Science & Business Media (2013)
17. Heemels, W.P.M.H., Donkers, M.C.F., Teel, A.R.: Periodic event-triggered control for linear systems. IEEE Trans. Autom. Control **58**, 847–861 (2013)
18. Heemels, W.P.M.H., Johansson, K.H., Tabuada, P.: An introduction to event-triggered and self-triggered control. In: Proceedings 51st IEEE Conference Decision and Control, pp. 3270–3285, Maui, HI (2012)
19. Hespanha, J., Naghshtabrizi, P., Xu, Y.: A survey of recent results in networked control systems. Proc. IEEE **95**(1), 138–162 (2007)
20. Lamare, P.-O., Girard, A., Prieur, C.: Switching rules for stabilization of linear systems of conservation laws. SIAM J. Control Optim. **53**(3), 1599–1624 (2015)
21. Li, T.-T.: Global classical solutions for quasilinear hyperbolic systems. RAM: Research in Applied Mathematics, vol. 32. John Wiley & Sons, Paris (1994)
22. Liberzon, D.: Hybrid feedback stabilization of of systems with quantized signals. Automatica **39**(9), 1543–1554 (2003)
23. Liberzon, D.: On stabilization of linear systems with limited information. IEEE Trans. Autom. Control **48**(2), 304–307 (2003)

24. Liu, T., Jiang, Z.-P.: A small-gain approach to robust event-triggered control of nonlinear systems. IEEE Trans. Autom. Control **60**(8), 2072–2085 (2015)
25. Nair, G., Fagnani, F., Zampieri, S., Evans, R.J.: Feedback control under data rate constraints: an overview. Proc. IEEE **95**(1), 108–137 (2007)
26. Postoyan, R., Tabuada, P., Nesic, D., Anta, A.: A framework for the event-triggered stabilization of nonlinear systems. IEEE Trans. Autom. Control **60**(4), 982–996 (2015)
27. Praly, L., Wang, Y.: Stabilization in spite of matched unmodeled dynamics and an equivalent definition of input-to-state stability. Math. Control Signals Syst. **9**, 1–33 (1996)
28. Prieur, C.: Uniting local and global controllers with robustness to vanishing noise. Math. Control Signals Syst. **14**(2), 143–172 (2001)
29. Prieur, C.: Control of systems of conservation laws with boundary errors. Netw. Heterog. Media **4**(2), 393–407 (2009)
30. Prieur, C., de Halleux, J.: Stabilization of a 1-D tank containing a fluid modeled by the shallow water equations. Syst. Control Lett. **52**(3–4), 167–178 (2004)
31. Prieur, C., Mazenc, F.: ISS-Lyapunov functions for time-varying hyperbolic systems of balance laws. Math. Control Signals Syst. **24**(1), 111–134 (2012)
32. Prieur, C., Tarbouriech, S., Zaccarian, L.: Lyapunov-based hybrid loops for stability and performance of continuous-time control systems. Automatica **49**(2), 577–584 (2013)
33. Prieur, C., Winkin, J., Bastin, G.: Robust boundary control of systems of conservation laws. Math. Control Signals Syst. **20**(2), 173–197 (2008)
34. Seuret, A., Prieur, C., Marchand, N.: Stability of non-linear systems by means of event-triggered sampling algorithms. IMA J. Math. Control Inf. **31**(3), 415–433 (2014)
35. Shim, H., Liberzon, D.: Nonlinear observers robust to measurement disturbances in an ISS sense. IEEE Trans. Autom. Control **61**, 48–61 (2016)
36. Shim, H., Tanwani, A., Ping, Z.: Back-and-forth operation of state observers and norm estimation of estimation error. In: Proceedings 51st IEEE Conference Decision and Control, pp. 3221–3226, Maui, HI (2012)
37. Sontag, E.D., Teel, A.: Changing supply functions in input/state stable systems. IEEE Trans. Autom. Control **40**(8), 1476–1478 (1995)
38. Sontag, E.D., Wang, Y.: Output-to-state stability and detectability of nonlinear systems. Syst. Control Lett. **29**(5), 279–290 (1997)
39. Tabuada, P.: Event-triggered real-time scheduling of stabilizing control tasks. IEEE Trans. Autom. Control **52**(9), 1680–1685 (2007)
40. Tanwani, A., Prieur, C., Fiacchini, M.: Observer-based feedback stabilization of linear systems with event-triggered sampling and dynamic quantization. Syst. Control Lett. **94**, 46–56 (2016)
41. Tanwani, A., Teel, A.R., Prieur, C.: On using norm estimators for event-triggered control with dynamic output feedback. In: Proceedings 54th IEEE Conference on Decision and Control, pp. 5500–5505, Osaka, Japan (2015)
42. Teel, A.R., Praly, L.: Tools for semiglobal stabilization by partial state and output feedback. SIAM J. Control Optim. **33**(5), 1443–1488 (1995)
43. Wang, L.Y., Li, C., Yin, G.G., Guo, L., Xu, C.-Z.: State observability and observers of linear-time-invariant systems under irregular sampling and sensor limitations. IEEE Trans. Autom. Control **56**(11), 2639–2654 (2011)

Chapter 9
Incremental Graphical Asymptotic Stability for Hybrid Dynamical Systems

Yuchun Li and Ricardo G. Sanfelice

Abstract This chapter introduces an incremental asymptotic stability notion for sets of hybrid trajectories $\mathcal{S}$. The elements in $\mathcal{S}$ are functions defined on hybrid time domains, which are subsets of $\mathbb{R}_{\geq 0} \times \mathbb{N}$ with a specific structure. For this abstract system, incremental asymptotic stability is defined as the property of the graphical distance between every pair of solutions to the system having stable behavior (incremental graphical stability) and approaching zero asymptotically (incremental graphical attractivity). Necessary conditions for $\mathcal{S}$ to have such properties are presented. When $\mathcal{S}$ is generated by hybrid systems given in terms of hybrid inclusions, that is, differential equations and difference equations with state constraints, further necessary conditions on the data are highlighted. In addition, sufficient conditions for incremental graphical asymptotic stability involving the data of the hybrid inclusion are presented. Throughout the chapter, examples illustrate the notions and results.

9.1 Introduction

9.1.1 Motivation

In contrast to asymptotic stability, which can be interpreted as a property of each system solution relative to a set, incremental stability consists of a property for every pair of solutions to the system. More precisely, for a continuous-time system of the form $\dot{x} = f(x)$, the uniform version of such a property requires every pair of solutions $t \mapsto \phi_1(t)$ and $t \mapsto \phi_2(t)$ to $\dot{x} = f(x)$ to satisfy

$$|\phi_1(t) - \phi_2(t)| \leq \beta(|\phi_1(0) - \phi_2(0)|, t) \tag{9.1}$$

Y. Li · R.G. Sanfelice (✉)
University of California, Santa Cruz 95064, USA
e-mail: ricardo@ucsc.edu

Y. Li
e-mail: yuchunli@ucsc.edu

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_9

for each t in the domain of definition of ϕ_1 and ϕ_2, where β is a class-$\mathscr{KL}$ function; see, e.g., [1–3]. The bound (9.1) implies that the Euclidean distance between two solutions is upper bounded by a function of the difference between their initial conditions and also decreases as t gets arbitrarily large (when the domain of definition of the solutions is unbounded to the right).

Unfortunately, the incremental stability notions available in the literature (most of which are for continuous-time systems) cannot be applied directly to systems with variables that can change continuously and, at times, jump. These systems, known as *hybrid systems*, are capable of modeling a wide range of complex dynamical systems, including robotic, automotive, and power systems as well as natural processes. Hybrid systems are dynamical systems that exhibit characteristics typical of both continuous-time and discrete-time behaviors. As a set stability theory in terms of Lyapunov functions is available (see [4, 5]), the availability of an incremental stability notion for this class of systems would enable the study of similar properties for them as the current notion for continuous-time systems allows. However, as we make clear in Sect. 9.2, mismatch of jump times and length of domains of pairs of solutions starting nearby makes characterizing and guaranteeing incremental stability properties in hybrid systems difficult.

9.1.2 Results in This Chapter

In this chapter, we introduce a notion of graphical incremental asymptotic stability for a set of hybrid trajectories, which we denote $\mathscr{S}$ and contains all trajectories that cannot be further extended (namely, they are maximal). A set of hybrid trajectories can be considered an abstract system on itself, or can be generated using hybrid inclusions. For such class of systems, we establish necessary and sufficient conditions for graphical incremental asymptotic stability. More precisely, we establish the following results:

1. The set $\mathscr{S}$ is neither graphically incrementally stable nor graphically incrementally attractive if there exists two elements in $\mathscr{S}$ with nearby initial conditions such that the amount of flow or jump is not the same, as in Propositions 9.1, 9.2 and 9.3.
2. The set $\mathscr{S}$ is not incrementally graphically stable if there exists one element in $\mathscr{S}$ that is not unique, as in Proposition 9.4.
3. When elements in $\mathscr{S}$ are generated by all maximal solutions to a hybrid system given in terms of a hybrid inclusion with a nonempty jump set D, under mild assumptions, Theorem 9.1 reveals that it is necessary to have a finite-time convergence like property from points that are nearby the jump set D. Proposition 9.6 provides a sufficient condition to guarantee such a property.
4. In Theorem 9.2, sufficient conditions for a set $\mathscr{S}$ consisting of all maximal solutions to a hybrid inclusion to be incrementally graphically asymptotically stable are given. A special case of this result (with the jump set D being discrete) is

established in Corollary 9.1. Both results require the flow map to induce a contraction during flows.

5. An extension of the result in Theorem 9.2 is presented in Theorem 9.3, where the jump map is required to be a weak contraction mapping.

To the best of our knowledge, the notion of incremental stability and its properties for hybrid systems have not been thoroughly studied before, only discussed briefly in [6] for a class of transition systems in the context of bisimulations, and in [7] for a particular class of hybrid systems prioritizing ordinary time t; see also related definitions in [8].

9.1.3 *Organization of the Chapter*

The remainder of this chapter is organized as follows. Section 9.2 briefly discusses notions of incremental stability for continuous-time (discrete) systems and introduces a notion of graphical incremental stability for sets of hybrid trajectories. Section 9.3 establishes several sufficient and necessary conditions for the proposed notion. Examples are discussed throughout the chapter to illustrate the results.

Notation: The set $\mathbb{B}$ denotes a closed unit ball in Euclidean space with appropriate dimension. Given a set $S \subset \mathbb{R}^n$, the closure of S is the intersection of all closed sets containing S, denoted by $\overline{S}$; S is said to be discrete if nonempty and there exists $\delta > 0$ such that for each $x \in S$, $(x + \delta\mathbb{B}) \cap S = \{x\}$; $\overline{\mathrm{con}}S$ is the closure of the convex hull of the set S. $\mathbb{R}_{\geq 0} := [0, \infty)$ and $\mathbb{N} := \{0, 1, 2, \dots\}$. Given vectors $v \in \mathbb{R}^n$, $w \in \mathbb{R}^m$, $|v|$ defines the Euclidean vector norm $|v| = \sqrt{v^\top v}$, and $[v^\top \; w^\top]^\top$ is equivalent to (v, w); given a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$, i.e., $P = P^\top > 0$, the weighted norm $|v|_P = \sqrt{v^\top P v}$. Given a function $f : \mathbb{R}^m \to \mathbb{R}^n$, its domain of definition is denoted by dom$f$, i.e., dom$f := \{x \in \mathbb{R}^m : f(x) \text{ is defined}\}$. The range of f is denoted by rge f, i.e., rge $f := \{f(x) : x \in \mathrm{dom} f\}$. The right limit of the function f is defined as $f^+(x) := \lim_{v\to 0^+} f(x + v)$ if it exists. Given a point $y \in \mathbb{R}^n$ and a closed set $\mathscr{A} \subset \mathbb{R}^n$, $|y|_{\mathscr{A}} := \inf_{x\in\mathscr{A}} |x - y|$. A function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a class-$\mathscr{K}_\infty$ function, also written $\alpha \in \mathscr{K}_\infty$, if α is zero at zero, continuous, strictly increasing, and unbounded; α is positive definite, also written $\alpha \in \mathscr{P}\mathscr{D}$, if $\alpha(s) > 0$ for all $s > 0$ and $\alpha(0) = 0$. A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a class-$\mathscr{K}\mathscr{L}$ function, also written $\beta \in \mathscr{K}\mathscr{L}$, if it is nondecreasing in its first argument, nonincreasing in its second argument, $\lim_{r\to 0^+} \beta(r, s) = 0$ for each $s \in \mathbb{R}_{\geq 0}$, and $\lim_{s\to\infty} \beta(r, s) = 0$ for each $r \in \mathbb{R}_{\geq 0}$. Given a function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^r$, $\nabla_x f(x, y) := \frac{\partial f}{\partial x}(x, y)$. Given a matrix $A \in \mathbb{R}^{n\times n}$, eig$(A)$ is the set of eigenvalues of A; $\overline{\lambda}(A) = \max\{\mathrm{Re}(\lambda) : \lambda \in \mathrm{eig}(A)\}$; $\underline{\lambda}(A) = \min\{\mathrm{Re}(\lambda) : \lambda \in \mathrm{eig}(A)\}$; $|A| := \max\{|\lambda|^{\frac{1}{2}} : \lambda \in \mathrm{eig}(A^\top A)\}$. Given a real number $x \in \mathbb{R}$, f loor(x) is the closest integer to x from below. A function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a Lyapunov function with respect to a set $\mathscr{A}$ if V is continuously differentiable and such that $c_1(|x|_{\mathscr{A}}) \leq V(x) \leq c_2(|x|_{\mathscr{A}})$ for all $x \in \mathbb{R}^n$ and some functions $c_1, c_2 \in \mathscr{K}_\infty$. Given a set $\mathscr{A} \subset \mathbb{R}^n$, a point $x \in \mathbb{R}^n$ and a metric d on $\mathbb{R}^n$, the distance $|x|^d_{\mathscr{A}} := \sup_{z\in\mathscr{A}} d(x, z)$.

9.2 Definition of Incremental Stability for Hybrid Systems

Informally, incremental stability is typically defined as the property of every pair of trajectories staying close when they start close (stability) and, as time gets large, converging to each other (attractivity). To formally state this notion, let the set of trajectories to a system with state in $\mathbb{R}^n$ be denoted by $\mathscr{S}$ and the time variable parameterizing such trajectories be denoted by s. The variable s parameterizes the trajectories in forward time from $s_\circ = 0$. This parameter takes values from $\mathbb{R}_{\geq 0}$ when the system is a continuous-time system, in which case $\mathscr{S}$ is a set of continuous-time trajectories and every element $\phi \in \mathscr{S}$ has a domain $\operatorname{dom} \phi$ that is a subset of $\mathbb{R}_{\geq 0}$. The parameter takes values from $\mathbb{N}$ when the system is a discrete-time system, in which case $\mathscr{S}$ is a set of discrete-time trajectories and elements in $\mathscr{S}$ have a domain that is a subset of $\mathbb{N}$. Let the function d denote a metric on $\mathbb{R}^n \times \mathbb{R}^n$ measuring the distance between pairs of elements in $\mathscr{S}$. An element $\phi \in \mathscr{S}$ is said to be maximal if there is no $\phi' \in \mathscr{S}$ such that ϕ is a proper truncation of ϕ' and complete if $\operatorname{dom} \phi$ is unbounded. Since we are interested in the behavior of maximal elements in $\mathscr{S}$, without loss of generality, from now on, it is assumed that $\mathscr{S}$ is a set of maximal hybrid trajectories.

The set of trajectories $\mathscr{S}$ is incrementally asymptotically stable with respect to a metric d if it is incrementally stable, in the sense that for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\begin{aligned}&\phi_1, \phi_2 \in \mathscr{S}, \quad d(\phi_1(s_\circ), \phi_2(s_\circ)) \leq \delta \\ &\qquad \Rightarrow \quad \operatorname{dom} \phi_1 = \operatorname{dom} \phi_2, \quad d(\phi_1(s), \phi_2(s)) \leq \varepsilon \quad \forall s \in \operatorname{dom} \phi_1 (= \operatorname{dom} \phi_2)\end{aligned} \tag{9.2}$$

and incrementally attractive, in the sense that there exists $\mu > 0$ such that

$$\begin{aligned}&\phi_1, \phi_2 \in \mathscr{S}, \quad d(\phi_1(s_\circ), \phi_2(s_\circ)) \leq \mu \\ \Rightarrow \quad &\operatorname{dom} \phi_1 = \operatorname{dom} \phi_2 \text{unbounded}, \qquad \lim_{s \to \infty} d(\phi_1(s), \phi_2(s)) = 0.\end{aligned} \tag{9.3}$$

When incremental attractivity holds for any $\mu > 0$, we say that the set of trajectories $\mathscr{S}$ is globally incrementally stable.

The notion defined above captures the nominal version of [2, Definition 2.1] for continuous-time systems when the elements in $\mathscr{S}$ are generated by a nonlinear continuous-time system of the form $\dot{x} = f(x)$. It also captures the notion for discrete-time systems of the form $x^+ = g(x)$, see, e.g., [9, 10]. To assess this notion for the hybrid case, we define hybrid trajectories as functions on hybrid time domains.

Definition 9.1 (*hybrid time domain*) A subset $E \subset \mathbb{R}_{\geq 0} \times \mathbb{N}$ is a *compact hybrid time domain* if

$$E = \bigcup_{j=0}^{J-1} \left([t_j, t_{j+1}], j \right)$$

for some finite sequence of times $0 = t_0 \leq t_1 \leq t_2 \leq \ldots \leq t_J$. It is a *hybrid time domain* if for all $(T, J) \in E$, $E \cap ([0, T] \times \{0, 1, \ldots, J\})$ is a compact hybrid time domain.

Given a hybrid time domain E, we define

$$\sup_t E := \sup_{(t,j)\in E} t, \qquad \sup_j E := \sup_{(t,j)\in E} j.$$

Definition 9.2 (*hybrid trajectory*) A function $\phi : \operatorname{dom}\phi \to \mathbb{R}^n$ is a *hybrid trajectory* (or hybrid arc) if $\operatorname{dom}\phi$ is a hybrid time domain and if for each $j \in \mathbb{N}$, the function $t \mapsto \phi(t,j)$ is locally absolutely continuous on the interval $I_j : = \{t : (t,j) \in \operatorname{dom}\phi\}$.

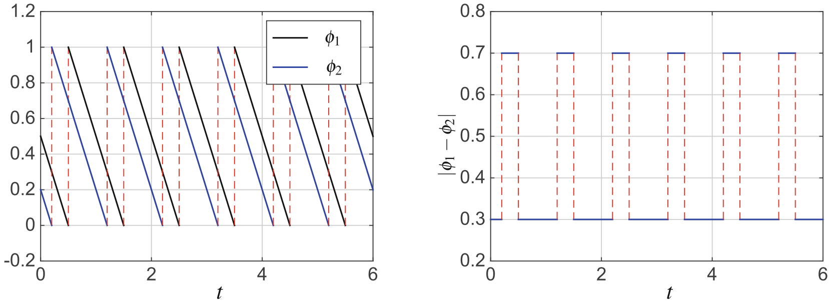
Remark 9.1 When every $\phi \in \mathscr{S}$ is such that $\operatorname{dom}\phi \subset \mathbb{R}_{\geq 0} \times \{0\}$ and $\operatorname{dom}\phi$ has more than one point, $\mathscr{S}$ is a set of continuous-time trajectories, while when every $\phi \in \mathscr{S}$ is such that $\operatorname{dom}\phi \subset \{0\} \times \mathbb{N}$ and $\operatorname{dom}\phi$ has more than one point, $\mathscr{S}$ is a set of discrete-time trajectories. Finally, when every $\phi \in \mathscr{S}$ is such that ϕ is either bounded or $\operatorname{dom}\phi$ is unbounded, $\mathscr{S}$ is said to be pre-forward complete.

For the case when the elements in $\mathscr{S}$ are hybrid trajectories, it is natural to consider an extension of the notion above when s takes values from $\mathbb{R}_{\geq 0} \times \mathbb{N}$ and is written as $s = (t, j)$, and $s_\circ = (0, 0)$. Unfortunately, there are several subtleties that make such extension of the notion above limiting for hybrid systems, some of which we illustrate next in simple examples. The first example illustrates issues measuring the distance between a pair of trajectories for a system that one would expect to be incrementally stable (but not incrementally attractive). The second example illustrates issues in measuring such distance for pairs of trajectories with dramatically different hybrid time domains.

Example 9.1 (mismatch of event times) Let $\mathscr{S}$ be the set of hybrid trajectories with (maximal and complete) elements ϕ defined as

$$\begin{aligned} \phi(t,j) =& \phi(0,0) - (t - j) \\ & \forall (t,j) \ : \ t \in \left[\max\{j-1, 0\} + \operatorname{ceil}\left(\frac{j}{j+1}\right)\phi(0,0), j + \phi(0,0)\right], j \in \mathbb{N} \end{aligned}$$

with $\phi(0,0) \geq 0$. (This set of trajectories can be generated using the hybrid inclusion given in Example 9.7.) Each trajectory in $\mathscr{S}$ reaches zero in finite flow time, at which event is reset to one instantaneously and from where it periodically reaches zero and gets reset to one. Figure 9.1a shows two trajectories with initial values within $\delta = 0.3$. This figure appears to suggest that trajectories from $\mathscr{S}$ starting close stay close. However, condition (9.2) does not hold unless $\phi_1(0,0) = \phi_2(0,0)$. In fact, consider two such trajectories, ϕ_1 and ϕ_2, with initial values satisfying $|\phi_1(0,0) - \phi_2(0,0)| \leq \delta$ and $\phi_1(0,0) \neq \phi_2(0,0)$. First, $\operatorname{dom}\phi_1 \neq \operatorname{dom}\phi_2$ since $(\phi_1(0,0), 1) \in \operatorname{dom}\phi_1$ and $(\phi_2(0,0), 1) \in \operatorname{dom}\phi_2$ but $\phi_1(0,0) \neq \phi_2(0,0)$. Without loss of generality, assume



(a) The projections of two hybrid trajectories from $\phi_1(0,0) = 0.5$ and $\phi_2(0,0) = 0.3$ on the t direction.

(b) Euclidean distance between ϕ_1 and ϕ_2.

Fig. 9.1 Two elements ϕ_1 and ϕ_2 from the set $\mathscr{S}$ given in Example 9.1. The Euclidean distance, which, precisely, is given by $|\phi_1(t,j_1(t)) - \phi_2(t,j_2(t))|$ for all $(t,j_i(t)) \in \operatorname{dom}\phi_i$, $j_i(t) = \min_{(t,j_i')\in\operatorname{dom}\phi_i} j_i'$, assumes the value 0.7 for 0.3 s periodically. On the other hand, the "graphical distance" from ϕ_1 to ϕ_2 is zero for $\varepsilon = 0.3$, while the "graphical distance" from ϕ_2 to ϕ_1 converges to zero in 0.3 s

$0 < \phi_1(0,0) < \phi_2(0,0)$. Then, even when the condition of equal domains is omitted, we have

$$|\phi_1(t_1,1) - \phi_2(t_1,0)| = |1 - \phi_2(t_1,0)| = |1 - \phi_2(0,0) + \phi_1(0,0)|,$$

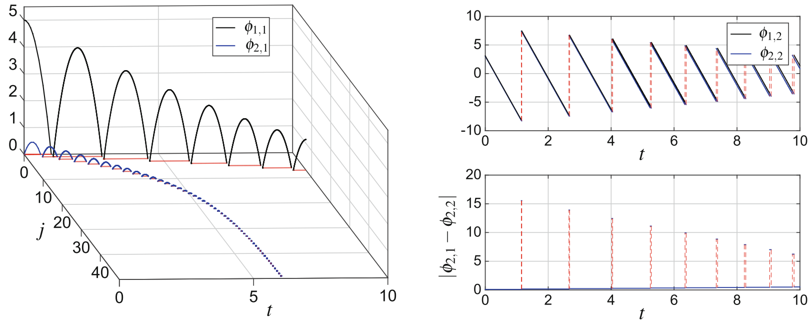
where we used the fact that $t_1 = \phi_1(0,0)$. No matter how small $\delta \in (0,1)$ is chosen, $|\phi_1(t_1,1) - \phi_2(t_1,0)| \geq 1 - \delta$. This property makes it impossible for the Euclidean distance between ϕ_1 and ϕ_2 to satisfy the ε-δ criterion in (9.2). In such case, the Euclidean distance (or any other metric d) may not be a good candidate of a distance function for the study of incremental properties. △

Example 9.1 suggests that a notion of incremental stability for hybrid systems has to allow for a mismatch of the jump times of two hybrid trajectories. This example also highlights that the pointwise (in $s = (t,j)$) distance is not appropriate for the purposes of defining incremental stability for sets of hybrid trajectories.

Example 9.2 (mismatch of length of domains) Let $\mathscr{S}$ be the set of hybrid trajectories with elements ϕ defined as

$$\phi(t,j) = \begin{bmatrix} -\frac{\gamma}{2}(t-t_j)^2 + \phi_2(t_j,j)(t-t_j) + \phi_1(t_j,j) \\ -\gamma(t-t_j) + \phi_2(t_j,j) \end{bmatrix}$$
$$\forall (t,j) \in \bigcup_{i\in\mathbb{N}} \left([t_i,t_{i+1}] \times \{i\}\right)$$

with $\phi(0,0) \in \mathbb{R}_{\geq 0} \times \mathbb{R}$, where $t_0 = 0$, $t_1 = \frac{\phi_2(0,0)+\sqrt{\phi_2(0,0)^2+2\gamma\phi_1(0,0)}}{\gamma}$,



(a) The first component (height $\phi_{i,1}$) of hybrid trajectories starting at $\phi_1(0,0) = (5,0)$ and $\phi_2(0,0) = (0,3)$.

(b) The projection of the second component (velocity $\phi_{i,2}$) of hybrid trajectories from $\phi_1(0,0) = (3,3)$ and $\phi_2(0,0) = (3,3.1)$ on the t direction.

Fig. 9.2 Hybrid trajectories in Example 9.2. The Euclidean distance, which is $|\phi_{1,2}(t, j_1(t)) - \phi_{2,2}(t, j_2(t))|$ for all $(t, j_i(t)) \in \operatorname{dom}\phi_i$, has repetitive large peaks, where $j_i(t) = \min_{(t,j_i)\in\operatorname{dom}\phi_i} j_i$

$$t_j = t_1 + \frac{2(\gamma t_1 - \phi_2(0,0))}{\gamma} \sum_{i=1}^{j-1} \lambda^i \qquad \forall j \in \mathbb{N} \setminus \{0,1\}$$

$$\phi_2(t_{j+1}, j+1) = -\lambda\phi_2(t_{j+1}, j) \qquad \forall j \in \mathbb{N}$$

$\gamma > 0$, and $\lambda \in (0,1)$. These trajectories capture the evolution of the height (ϕ_1) and vertical velocity (ϕ_2) of a ball bouncing on a ground at zero height, where γ represents the gravity constant and λ the restitution coefficient. A hybrid inclusion generating this set of hybrid trajectories is given in [4, Examples 1.1 and 2.12]. Each element $\phi \in \mathscr{S}$ is such that

$$\sup_t \operatorname{dom}\phi = \frac{\phi_2(0,0)}{\gamma} + \frac{1+\lambda}{\gamma(1-\lambda)}\sqrt{\phi_2(0,0)^2 + 2\gamma\phi_1(0,0)} \tag{9.4}$$

Figure 9.2a shows the position (first) component of two hybrid trajectories ($\phi_i = (\phi_{i,1}, \phi_{i,2})$ for $i \in \{1,2\}$) from initial conditions $\phi_1(0,0) = (5,0)$ (ball starting at a positive height with zero velocity) and $\phi_2(0,0) = (0,3)$ (ball starting at the ground with a positive velocity). As Fig. 9.2a shows, the jumps in ϕ_2 accumulate at about $t = 6$ s while ϕ_1 is still describing the motion of the ball bouncing.

Given two elements $\phi_1, \phi_2 \in \mathscr{S}$ with $\phi_1(0,0) \neq \phi_2(0,0)$, according to (9.4), $\sup_t \operatorname{dom}\phi_1 \neq \sup_t \operatorname{dom}\phi_2$. Without loss of generality, assuming that $\sup_t \operatorname{dom}\phi_2 < \sup_t \operatorname{dom}\phi_1$, then we have that ϕ_2 is not defined at points $(t', j') \in \operatorname{dom}\phi_1$ with $t' + j' \geq \sup_t \operatorname{dom}\phi_2$. Hence, at such points, it is not possible to measure the distance between ϕ_1 and ϕ_2. Note that for such points (t', j') we have that $(t', j) \notin \operatorname{dom}\phi_2$ for any $j \in \mathbb{N}$, which indicates that it is not possible to relax the incremental stability notion by instead requiring that the distance between the trajectories be small for each common t and potentially different values of the jump parameter j. Even when we omit such points, for points $(t,j) \in \operatorname{dom}\phi_2$ with t close to $\sup_t \operatorname{dom}\phi_2$ and points $(t,j'') \in \operatorname{dom}\phi_1$, we have that j is much larger than j'' since j grows unbounded as t approaches $\sup_t \operatorname{dom}\phi_2$. This fact makes comparing trajectories using the graphical distance in this particular set of hybrid solutions very difficult. A similar situation is encountered if, instead, the pointwise distance is used. As shown in Fig. 9.2b, the pointwise distance between velocity (second) components of two solutions ($\phi_{i,2}$ for $i \in \{1,2\}$) has repetitive large peaks, even though they are initialized very close to each other. △

While Example 9.1 already has elements in $\mathscr{S}$ with different domains, Example 9.2 pinpoints a key difficulty in measuring the distance between solutions with jump times that accumulate, namely, Zeno solutions. In fact, when accumulation of events occurs in finite time t, determining the appropriate distance function to certify incremental stability is rather difficult since, when the accumulation time depends on the initial condition as in Example 9.2, the distance between the trajectories may not be quantifiable over an unbounded set. On the other hand, a notion of incremental stability for a set of continuous-time trajectories or for a set of discrete-time trajectories with elements having different time domains can be formulated by only requiring the stability condition to hold over the intersection of the domains of definition of every pair of trajectories starting nearby.

Motivated by the issues mentioned above, we propose a notion of incremental asymptotic stability that employs the graphical distance between the graphs defined by the hybrid trajectories.

Definition 9.3 ([4, Definition 5.20]) The graph of a hybrid trajectory $\phi : \operatorname{dom}\phi \to \mathbb{R}^n$ is a set in $\mathbb{R}^{n+2}$ given by

$$\operatorname{gph}\phi = \{(t,j,x) : (t,j) \in \operatorname{dom}\phi,\ x = \phi(t,j)\}. \tag{9.5}$$

To measure the distance between the graphs of two hybrid trajectories, given a metric d, we use the following graphical distance notion for hybrid trajectories.

Definition 9.4 ([4, Definition 4.11]) Given $\varepsilon > 0$, two hybrid trajectories ϕ_1 and ϕ_2 are *graphically ε-close with respect to d* if

(a) for each $(t,j) \in \operatorname{dom}\phi_1$ there exists s such that $(s,j) \in \operatorname{dom}\phi_2$, $|t-s| \le \varepsilon$, and

$$d(\phi_1(t,j), \phi_2(s,j)) \le \varepsilon,$$

(b) for each $(t,j) \in \operatorname{dom}\phi_2$ there exists s such that $(s,j) \in \operatorname{dom}\phi_1$, $|t-s| \le \varepsilon$, and

$$d(\phi_2(t,j), \phi_1(s,j)) \le \varepsilon.$$

To characterize the distance between the graphs of two hybrid arcs over a finite horizon, we use the following graphical (τ, ε)-closeness notion for hybrid trajectories.

Definition 9.5 ([4, Definition 5.23]) Given $\tau, \varepsilon > 0$, two hybrid trajectories ϕ_1 and ϕ_2 are *graphically* (τ, ε)*-close with respect to* d if

(a) for each $(t, j) \in \operatorname{dom} \phi_1$ with $t + j \leq \tau$ there exists s such that $(s, j) \in \operatorname{dom} \phi_2$, $|t - s| \leq \varepsilon$, and

$$d(\phi_1(t, j), \phi_2(s, j)) \leq \varepsilon,$$

(b) for each $(t, j) \in \operatorname{dom} \phi_2$ with $t + j \leq \tau$ there exists s such that $(s, j) \in \operatorname{dom} \phi_1$, $|t - s| \leq \varepsilon$, and

$$d(\phi_2(t, j), \phi_1(s, j)) \leq \varepsilon.$$

To characterize the property of hybrid trajectories graphically converging to each other, we introduce the following notion.

Definition 9.6 Given $\varepsilon > 0$, two hybrid trajectories ϕ_1 and ϕ_2 are *eventually graphically* ε*-close with respect to* d if

(a) there exists $T > 0$ such that for each $(t, j) \in \operatorname{dom} \phi_1$ and $t + j \geq T$, there exists $(s, j) \in \operatorname{dom} \phi_2$ satisfying $|t - s| \leq \varepsilon$ and

$$d(\phi_1(t, j), \phi_2(s, j)) \leq \varepsilon, \tag{9.6}$$

(b) there exists $T > 0$ such that for each $(t, j) \in \operatorname{dom} \phi_2$ and $t + j \geq T$, there exists $(s, j) \in \operatorname{dom} \phi_1$ satisfying $|t - s| \leq \varepsilon$ and

$$d(\phi_2(t, j), \phi_1(s, j)) \leq \varepsilon. \tag{9.7}$$

Remark 9.2 If two hybrid trajectories ϕ_1 and ϕ_2 are not complete, then the property in Definition 9.6 holds for free. In particular, the property would hold vacuously for $T > \max\{T_1 + J_1, T_2 + J_2\}$, where $T_1 = \sup_t \operatorname{dom} \phi_1$, $J_1 = \sup_j \operatorname{dom} \phi_1$, $T_2 = \sup_t \operatorname{dom} \phi_2$, and $J_2 = \sup_j \operatorname{dom} \phi_2$.

Now, we are ready to define incremental asymptotic stability for sets of hybrid trajectories.

Definition 9.7 (*incremental graphical asymptotic stability*) The set of hybrid trajectories $\mathscr{S}$ is said to be

1. *incrementally graphically stable* (δS) *with respect to* d if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\begin{aligned} &\phi_1, \phi_2 \in \mathscr{S}, \quad d(\phi_1(0,0), \phi_2(0,0)) \leq \delta \\ &\qquad \Rightarrow \quad \phi_1 \text{ and } \phi_2 \text{ are graphically } \varepsilon\text{-close with respect to d} \end{aligned} \tag{9.8}$$

2. *incrementally graphically locally attractive* (δLA) *with respect to* d if there exists $\mu > 0$ such that for every $\varepsilon > 0$

$$\begin{aligned}&\phi_1, \phi_2 \in \mathscr{S}, \quad d(\phi_1(0,0), \phi_2(0,0)) \le \mu \\ &\qquad \Rightarrow \quad \phi_1 \text{ and } \phi_2 \text{ are eventually graphically } \varepsilon\text{-close with respect to d}\end{aligned} \tag{9.9}$$

3. *incrementally graphically locally asymptotically stable* (δLAS) *with respect to d* if it is both δS and δLA.

When δLA holds for every $\mu > 0$, we say that the set of hybrid trajectories $\mathscr{S}$ is incrementally graphically globally attractive (δGA).

Remark 9.3 The notion in Definition 9.7 covers the special cases of $\mathscr{S}$ being a set of continuous-time trajectories or a set of discrete-time trajectories. In particular, when $\mathscr{S}$ is a set of complete discrete-time trajectories, condition (9.8) reduces to

$$\begin{aligned}&\phi_1, \phi_2 \in \mathscr{S}, \quad d(\phi_1(0,0), \phi_2(0,0)) \le \delta \\ &\qquad \Rightarrow \quad d(\phi_1(0,j), \phi_2(0,j)) \le \varepsilon \quad \forall j \in \mathbb{N}.\end{aligned} \tag{9.10}$$

Due to requiring a property for every possible pair of trajectories, incremental graphical global attractivity only holds when $\mathscr{S}$ is either a set of continuous-time trajectories or of discrete-time trajectories (see [8]). As a difference to those in [8, Definition 3], both the δS and δLA notions in Definition 9.7 exploit the graphically ε-closeness notion in [4, Definition 4.11], which in [4] is shown to be a structural property of solutions to well-posed hybrid systems.

Note that unboundedness of the domain of the elements in a generic set $\mathscr{S}$ is not required, but when there are elements with dramatically different domains, incremental stability may not hold—in particular, the set of solutions in Example 9.2 would not be δLA. The following result formalizes this fact.

Proposition 9.1 *Let $\mathscr{S}$ be a set of hybrid trajectories. Suppose that no matter how small $\delta' > 0$ is, there exist complete $\phi_1, \phi_2 \in \mathscr{S}$ with $|\phi_1(0,0) - \phi_2(0,0)| \le \delta'$ such that* $\sup_t \operatorname{dom} \phi_2 < \sup_t \operatorname{dom} \phi_1 < \infty$. *Then, $\mathscr{S}$ is neither δS nor δLA with respect to any metric d.*

Proof We proceed by contradiction. Let d be any metric, $t_1^z = \sup_t \operatorname{dom} \phi_1$, and $t_2^z = \sup_t \operatorname{dom} \phi_2$. Since $\operatorname{dom} \phi_1$ and $\operatorname{dom} \phi_2$ are unbounded and $\sup_t \operatorname{dom} \phi_2 < \sup_t \operatorname{dom} \phi_1$, there exists $T \in (t_2^z, t_1^z)$. Pick $\varepsilon \in (0, \min\{T - t_2^z, t_1^z - T\})$. By continuity of d and the fact that $d(x,x) = 0$ for all $x \in \mathbb{R}^n$,

$$\begin{aligned}&\text{for each } \rho > 0 \text{ there exists } \delta'' > 0 \text{ such that} \\ &\qquad d(x', y') \le \rho \text{ for all } x', y' \text{ such that } |x' - y'| \le \delta''.\end{aligned} \tag{9.11}$$

Now, suppose that $\mathscr{S}$ is δS with respect to d. With ε as above, let δ be such that (9.8) holds. Pick $\rho \le \delta$ and let δ'' be generated by the continuity property of d in (9.11). Using δ' such that $\delta' \le \delta''$ in the assumption of the claim, in which ϕ_1 and ϕ_2 start within δ' in terms of the Euclidean distance, in particular, we have that

$d(\phi_1(0,0),\phi_2(0,0)) \leq \delta$ and ϕ_1 and ϕ_2 are graphically ε-close with respect to d. However, since $\sup_t \operatorname{dom}\phi_2 < T$, there exists $(t,j) \in \operatorname{dom}\phi_1$ with $t > T$ such that $(t',j') \notin \operatorname{dom}\phi_2$ for each t' satisfying $|t-t'| < \varepsilon$ and for some $j' \in \mathbb{N}$. This fact contradicts graphical ε-closeness with respect to d guaranteed by (9.8). The case when $\mathscr{S}$ is δLA follows similarly. □

Next, we revisit Example 9.1 and show that the set of hybrid trajectories therein is δS. More examples illustrating the proposed notions will be given in Sect. 9.3, in which sets of solutions $\mathscr{S}$ are generated by hybrid inclusions.

Example 9.3 We show that $\mathscr{S}$ given in Example 9.1 is δS. For a given $\varepsilon > 0$, let $0 < \delta < \varepsilon$ and assume $|\phi_1(0,0) - \phi_2(0,0)| < \delta$ and pick corresponding trajectories $\phi_1, \phi_2 \in \mathscr{S}$. Without loss of generality, we further suppose $0 \leq \phi_1(0,0) \leq \phi_2(0,0)$ and pick corresponding trajectories $\phi_1, \phi_2 \in \mathscr{S}$. Then, the hybrid trajectory ϕ_1 jumps before ϕ_2. For each $j \in \mathbb{N} \setminus \{0\}$, let $\bar{t}_j = \max_{(t,j-1)\in\operatorname{dom}\phi_1\cap\operatorname{dom}\phi_2} t$ and $\bar{t}'_j = \min_{(t,j)\in\operatorname{dom}\phi_1\cap\operatorname{dom}\phi_2} t$. Then, we have that for each $t \in [0,\bar{t}_1]$, there exists $(s,0) \in \operatorname{dom}\phi_2$ such that $s = t$ and

$$|\phi_1(t,0) - \phi_2(t,0)| = |\phi_1(0,0) - t - \phi_2(0,0) + t| \leq \delta < \varepsilon. \tag{9.12}$$

For each $t \in [\bar{t}_1, \bar{t}'_1]$,

$$\begin{aligned} |\phi_1(\bar{t}_1,0) - \phi_2(t,0)| &= |\phi_2(0,0) - t| \\ &\leq |\phi_2(0,0) - \bar{t}_1| = |\phi_2(0,0) - \phi_1(0,0)| \leq \delta < \varepsilon, \end{aligned} \tag{9.13}$$

where we used the fact that $\phi_1(\bar{t}_1,0) = \phi_1(0,0) - \bar{t}_1 = 0$. Moreover, $\phi_2(\bar{t}'_1,0) = \phi_2(0,0) - \bar{t}'_1 = 0$. Then, $|\bar{t}'_1 - \bar{t}_1| = |\phi_2(0,0) - \phi_1(0,0)| \leq \delta < \varepsilon$. Therefore, for each $t \in [\bar{t}_1, \bar{t}'_1]$,

$$|\phi_1(t,1) - \phi_2(\bar{t}'_1,1)| = |1 - (t - \bar{t}_1) - 1| \leq \delta < \varepsilon. \tag{9.14}$$

Proceeding similarly and using (9.14), for each $t \in [\bar{t}'_{i-1}, \bar{t}_i]$, where $i \in \mathbb{N} \setminus \{0,1\}$,

$$|\phi_1(t,i-1) - \phi_2(t,i-1)| = |\phi_1(\bar{t}'_{i-1},i-1) - \phi_2(\bar{t}'_{i-1},i-1)| \leq \delta < \varepsilon.$$

Moreover, since $\phi_1(\bar{t}_i,i-1) = 0$, for each $t \in [\bar{t}_i, \bar{t}'_i]$, where $i \in \mathbb{N} \setminus \{0,1\}$,

$$\begin{aligned} |\phi_1(\bar{t}_i,i-1) - \phi_2(t,i-1)| &= |\phi_1(\bar{t}_i,i-1) - \phi_2(\bar{t}_i,i-1) + (t - \bar{t}_i)| \\ &\leq |\phi_2(\bar{t}_i,i-1) - \phi_1(\bar{t}_i,i-1)| \leq \delta < \varepsilon, \end{aligned}$$

and $|\phi_1(t,i) - \phi_2(\bar{t}'_i,i)| = |1-(t-\bar{t}_i)-1| \leq \delta < \varepsilon$. Therefore, the set $\mathscr{S}$ is δS.[1] On the other hand, since the distance between ϕ_1 and ϕ_2 does not converge to zero, ϕ_1, ϕ_2 are not eventually ε-close and thus the set $\mathscr{S}$ is neither δLA nor δGA. △

To further illustrate the notion in Definition 9.7, the following example shows that a set $\mathscr{S}$ is δLAS.

Example 9.4 Let $\mathscr{S}$ be the set of hybrid trajectories with elements ϕ

$$\phi(t,j) = \left(\phi(t_j,j) - \text{ceil}\left(\frac{j}{j+1}\right)\right)\exp(-t+t_j) \tag{9.15}$$

for all $(t,j) \in \bigcup_{i\in\mathbb{N}, i<J}\left([t_i,t_{i+1}]\times\{i\}\right)\bigcup([t_J,\infty)\times\{J\})$ with $\phi(0,0) \subset \bigcup_{i\in\{2k:k\in\mathbb{N}\}}[i,i+1]$, where $J = \frac{1}{2}\,\text{floor}(\phi(0,0))$, $t_0 = 0$, and, for $J > 0$, $t_J = t_{J-1}$ and

$$\begin{aligned} t_j &= \ln(\phi(0,0)) - \ln(\text{floor}(\phi(0,0))) \\ &\quad + \sum_{k=1}^{j-1}(\ln(\text{floor}(\phi(0,0)) - k) - \ln(\text{floor}(\phi(0,0)) - k - 1)) \quad \forall j \in \mathbb{N}\setminus\{0\}, j \leq J. \end{aligned}$$

(This set of trajectories can be generated using the hybrid inclusion given in Example 9.6.) Given $\varepsilon > 0$, consider two elements $\phi_1, \phi_2 \in \mathscr{S}$ such that $|\phi_1(0,0) - \phi_2(0,0)| \leq \delta$, where $0 \leq \delta < \min\{1,\varepsilon\}$. Then, it is guaranteed that $\bar{J} := \sup_j \text{dom}\,\phi = \sup_j \text{dom}\,\phi_2 < \infty$ since $\text{floor}(\phi_1(0,0)) = \text{floor}(\phi_2(0,0))$. For each $j \in \mathbb{N}\setminus\{0\}$, let $\bar{t}_j = \max_{(t,j-1)\in\text{dom}\,\phi_1\cap\text{dom}\,\phi_2} t$ and $\bar{t}'_j = \min_{(t,j)\in\text{dom}\,\phi_1\cap\text{dom}\,\phi_2} t$. Without loss of generality, assume $\phi_2(0,0) > \phi_1(0,0) \geq 2$, then ϕ_1 jumps first. Then, we have that for each $t \in [0,\bar{t}_1]$, there exists $(s,0) \in \text{dom}\,\phi_2$ such that $s = t$ and

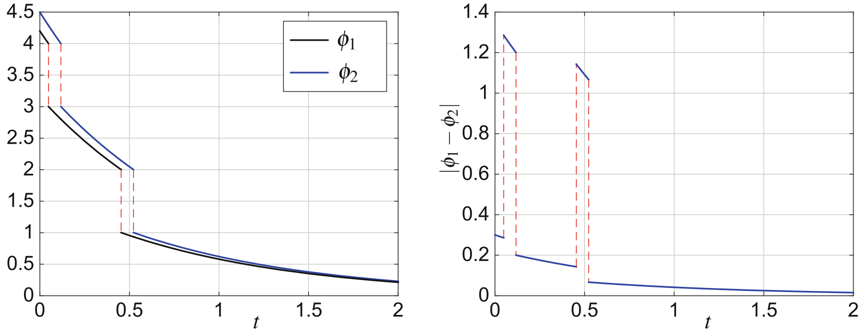
$$|\phi_1(t,0) - \phi_2(t,0)| = |\phi_1(0,0)\exp(-t) - \phi_2(0,0)\exp(-t)| \leq \delta < \varepsilon. \tag{9.16}$$

For each $t \in [\bar{t}_1, \bar{t}'_1]$,

$$\begin{aligned} |\phi_1(\bar{t}_1,0) - \phi_2(t,0)| &= |\exp(-\bar{t}_1)\phi_1(0,0) - \exp(-t)\phi_2(0,0)| \\ &\leq |\exp(-\bar{t}_1)\phi_1(0,0) - \exp(-\bar{t}_1)\phi_2(0,0)| \leq \delta < \varepsilon, \end{aligned} \tag{9.17}$$

where we used the property $\exp(-\bar{t}_1)\phi_1(0,0) = \text{floor}(\phi_1(0,0)) = \text{floor}(\phi_2(0,0)) = \exp(-\bar{t}'_1)\phi_2(0,0)$. Note that $\bar{t}_1 = \ln(\phi_1(0,0)) - \ln(\text{floor}(\phi_1(0,0)))$ and $\bar{t}'_1 = \ln(\phi_2(0,0)) - \ln(\text{floor}(\phi_2(0,0)))$. Therefore, $\bar{t}'_1 - \bar{t}_1 = \ln(\phi_2(0,0)) - \ln(\phi_1(0,0))$. Furthermore, by the mean value theorem, there exists $\phi_0^\star \in [\phi_1(0,0), \phi_2(0,0)]$ such that $|\bar{t}'_1 - \bar{t}_1| = \frac{1}{\phi_0^\star}|\phi_1(0,0) - \phi_2(0,0)| \leq |\phi_1(0,0) - \phi_2(0,0)| \leq \delta < \varepsilon$. Similarly, for each $t \in [\bar{t}_1, \bar{t}'_1]$,

[1] Using the ideas in [11], it may be possible to construct an alternative distance function that is decreasing along trajectories.



(a) The projections of two elements from $\phi_1(0,0) = 4.2$ and $\phi_2(0,0) = 4.5$ on the t direction.

(b) Comparison between graphical distance and Euclidean distance between ϕ_1 and ϕ_2.

Fig. 9.3 Two maximal elements ϕ_1 and ϕ_2. Unlike the Euclidean distance, which is $|\phi_1(t, j_1(t)) - \phi_2(t, j_2(t))|$ for all $(t, j_i(t)) \in \operatorname{dom}\phi_i$ and $j_i(t) = \min_{(t,j_i)\in\operatorname{dom}\phi_i} j_i$, which does not decrease along hybrid trajectories, the "graphical distance" from ϕ_1 to ϕ_2 is zero for $\varepsilon = 0.3$ and the "graphical distance" from ϕ_2 to ϕ_1 converges to zero

$$
\begin{aligned}
|\phi_1(t,1) - \phi_2(\bar{t}_1', 1)| &= |\exp(-t + \bar{t}_1)\phi_1(\bar{t}_1, 1) - \phi_2(\bar{t}_1', 1)| \\
&= \phi_2(\bar{t}_1', 1) - \exp(-t + \bar{t}_1)\phi_1(\bar{t}_1, 1) \\
&\le \phi_2(\bar{t}_1', 1) - \exp(-\bar{t}_1' + \bar{t}_1)\phi_1(\bar{t}_1, 1) \\
&\le \phi_2(\bar{t}_1', 1) - \exp(-\ln(\phi_2(0,0)) + \ln(\phi_1(0,0)))\phi_1(\bar{t}_1, 1) \\
&\le \operatorname{floor}(\phi_2(0,0)) \left(1 - \exp\left(\ln \frac{\phi_1(0,0)}{\phi_2(0,0)}\right)\right) \\
&\le \frac{\operatorname{floor}(\phi_2(0,0))}{\phi_2(0,0)} (\phi_2(0,0) - \phi_1(0,0)) \\
&\le \phi_2(0,0) - \phi_1(0,0) \le \delta.
\end{aligned}
\tag{9.18}
$$

Note that the derivation in (9.18) can be repeated for $\bar{J}$ times.

If $\phi_1(0,0), \phi_2(0,0) \in [0,1]$, we have that $|\phi_1(t,0) - \phi_2(t,0)| \le \exp(-t)|\phi_1(0,0) - \phi_2(0,0)| \le \delta$ for all $(t,0) \in \operatorname{dom}\phi_1 = \operatorname{dom}\phi_2$. In fact, $\lim_{t\to\infty,(t,0)\in\operatorname{dom}\phi_1} |\phi_1(t,0) - \phi_2(t,0)| = 0$. Therefore, the set $\mathscr{S}$ is δLAS.

As shown in Fig. 9.3a, the domains of two elements in the set $\mathscr{S}$ may be different from each other. The Euclidean distance between ϕ_1 and ϕ_2 has peaks during the mismatch part of the hybrid time domain, i.e., the time instances (t) when two solutions have different values of j, as shown in Fig. 9.3b. △

While the notion introduced in Definition 9.7 appears to be suitable for the study of incremental stability properties of sets of hybrid trajectories, in particular, for those generated using hybrid inclusions, conditions guaranteeing it are not obvious due to the noncausality nature of the notion. Necessary and sufficient conditions for this notion are proposed in the next section, both for sets of hybrid trajectories as well as hybrid inclusions.

9.3 Necessary and Sufficient Conditions for Incremental Graphical Stability Notions

In this section, we explore several necessary and sufficient conditions of incremental graphical stability properties for hybrid systems that satisfy certain assumptions. In particular, Proposition 9.2 implies a basic necessary condition for two hybrid arcs to be ε-close and eventually ε-close, respectively. Proposition 9.4 shows that maximal elements in a set $\mathscr{S}$ are unique if $\mathscr{S}$ is δS. In Theorem 9.2, a sufficient condition for $\mathscr{H}$ to be δLAS is presented for a hybrid system with generic jump sets. When D is a discrete set, Corollary 9.1 provides sufficient conditions for $\mathscr{H}$ to be δLAS. Moreover, Proposition 9.6 establishes Lyapunov-like sufficient conditions for item (2) of Corollary 9.1. Then, a finite-time stability property is shown to be necessary for $\mathscr{H}$ to be δS or δLA in Theorem 9.1. Furthermore, Theorem 9.3 studies conditions for which $\mathscr{H}$ is δLAS when the jump map is Lipschitz.

For them to be constructive, some of the necessary and sufficient conditions are stated for sets of hybrid trajectories generated by hybrid system given by hybrid inclusions. A hybrid system $\mathscr{H}$ has data (C, f, D, g) and is defined by

$$\begin{aligned} \dot{z} &= f(z) \quad z \in C, \\ z^+ &= g(z) \quad z \in D, \end{aligned} \tag{9.19}$$

where $z \in \mathbb{R}^n$ is the state, f defines the flow map capturing the continuous dynamics and C defines the flow set on which f is effective. The map g defines the jump map and models the discrete behavior, while D defines the jump set, which is the set of points from where jumps are allowed. A solution ϕ to $\mathscr{H}$ is hybrid trajectory that satisfies the dynamics of (9.19). A solution is Zeno if it is complete and its domain is bounded in the t direction. A solution is precompact if it is complete and bounded. The set of hybrid trajectories $\mathscr{S}_{\mathscr{H}}$ contains all maximal solutions to $\mathscr{H}$, and the set $\mathscr{S}_{\mathscr{H}}(\xi)$ contains all maximal solutions to $\mathscr{H}$ from ξ. Note the use of single-valued maps f and g in (9.19) is necessary when studying incremental stability; see Proposition 9.4.

Definition 9.8 A hybrid system $\mathscr{H} = (C, f, D, g)$ is said to satisfy the hybrid basic conditions if

(a) the sets C and D are closed;
(b) the functions $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ are continuous.

We refer the reader to [4] and [5] for more details on these notions and the hybrid systems' framework.

9.3.1 Necessary Conditions

The following result highlights a necessary property of the hybrid time domains of two hybrid arcs that are graphically close. In particular, it holds for every pair of elements in a set $\mathscr{S}$ that is δS, δLA, or δGA.

Proposition 9.2 *Given $\varepsilon > 0$ and two elements $\phi_1, \phi_2 \in \mathscr{S}$, the following holds:*

1. *if ϕ_1 and ϕ_2 are graphically ε-close, or*
2. *if ϕ_1 and ϕ_2 are complete and graphically eventually ε-close,*

then

$$\sup_j \operatorname{dom} \phi_1 = \sup_j \operatorname{dom} \phi_2. \tag{9.20}$$

Proof We proceed by contradiction. Given $\varepsilon > 0$, consider two hybrid arcs ϕ_1, ϕ_2 that are graphically ε-close. Suppose that $J_1 = \sup_j \operatorname{dom} \phi_1$, $J_2 = \sup_j \operatorname{dom} \phi_2$ and $J_1 \neq J_2$. Moreover, without loss of generality, assume that J_1 and J_2 are both finite and $J_1 > J_2$. Then, $J_1 > 0$. Let $(t_{J_1}, J_1) \in \operatorname{dom} \phi_1$ be such that $(t_{J_1}, J_1 - 1) \in \operatorname{dom} \phi_1$. Then, $(t, J_1) \notin \operatorname{dom} \phi_2$ for any $t \in \mathbb{R}_{\geq 0}$, which implies that there does not exist $(t, J_1) \in \operatorname{dom} \phi_2$ such that $|t - t_{J_1}| \leq \varepsilon$ and $d(\phi_1(t_{J_1}, J_1), \phi_2(t, J_1)) \leq \varepsilon$. This contradicts the fact that ϕ_1 and ϕ_2 are graphically ε-close. The situation where either J_1 or J_2 is ∞ follows similarly.

When ϕ_1 and ϕ_2 are complete and eventually graphically ε-close, given $\varepsilon > 0$, there exists $T > 0$ such that ϕ_1 and ϕ_2 satisfy (9.6) and (9.7) for all $(t_1, j_1) \in \operatorname{dom} \phi_1$ and $(t_2, j_2) \in \operatorname{dom} \phi_2$ such that $t_1 + j_1 > T$ and $t_2 + j_2 > T$. Proceeding by contradiction, suppose that $J_1 = \sup_j \operatorname{dom} \phi_1$, $J_2 = \sup_j \operatorname{dom} \phi_2$ and $J_1 \neq J_2$. Moreover, without loss of generality, assume that J_1 and J_2 are both finite and $J_1 > J_2$. Then, $J_1 > 0$. Let $(t_{J_1}, J_1) \in \operatorname{dom} \phi_1$ be such that $(t_{J_1}, J_1 - 1) \in \operatorname{dom} \phi_1$. Pick $(t, J_1) \in \operatorname{dom} \phi_1$ and $t + J_1 > T$, which is always possible since ϕ_1 is complete. Then, $(t, J_1) \notin \operatorname{dom} \phi_2$ which implies that there does not exists $(t, J_1) \in \operatorname{dom} \phi_2$ such that $|t - t_{J_1}| \leq \varepsilon$ and $d(\phi_1(t_{J_1}, J_1), \phi_2(t, J_1)) \leq \varepsilon$. This contradicts the fact that ϕ_1 and ϕ_2 are eventually graphically ε-close. The situation where either J_1 or J_2 is ∞ follows similarly. □

Example 9.5 Consider the set $\mathscr{S}$ given in Example 9.4, and two elements ϕ_1 and ϕ_2 with $\phi_1(0,0) = 4.5$ and $\phi_2(0,0) = 1$, respectively. The hybrid trajectory ϕ_1 jumps twice while the hybrid trajectory ϕ_2 never jumps. Therefore, ϕ_1 and ϕ_2 are not graphically eventually ε-close according to Proposition 9.2. This property prevents the set $\mathscr{S}$ from being δGA while Example 9.4 shows that this set is δLAS. △

Proposition 9.3 *Let $\mathscr{S}$ be a set of hybrid trajectories.*

1. *If $\mathscr{S}$ is δS or δLA with respect to a metric d, there exists $\delta > 0$ such that*

$$\phi_1, \phi_2 \in \mathscr{S}, \quad d(\phi_1(0,0), \phi_2(0,0)) \leq \delta \quad \Rightarrow \sup_t \operatorname{dom} \phi_1 = \sup_t \operatorname{dom} \phi_2. \tag{9.21}$$

2. *If $\mathscr{S}$ is δGA with respect to a metric d,*

$$\sup_t \operatorname{dom} \phi_1 = \sup_t \operatorname{dom} \phi_2. \tag{9.22}$$

Proof Proceeding by contradiction, suppose $\mathscr{S}$ is δS and, no matter how small $\delta > 0$ is chosen, there exist $\phi_1, \phi_2 \in \mathscr{S}$ such that $d(\phi_1(0,0), \phi_2(0,0)) < \delta$ and $\sup_t \operatorname{dom} \phi_1 \neq \sup_t \operatorname{dom} \phi_2$. Then, by Proposition 9.1, $\mathscr{S}$ is neither δS nor δLA. The argument follows similarly when $\mathscr{S}$ is δGA. □

The following result establishes that uniqueness is a necessary condition for δS. In turn, according to Proposition 9.4, it justifies the choice of using single-valued flow and jump maps in the definition of $\mathscr{H}$ in (9.19).

Proposition 9.4 (uniqueness of elements in $\mathscr{S}$) *Let $\mathscr{S}$ be a set of hybrid trajectories. Suppose $\mathscr{S}$ is δS with respect to a metric d. Then, every element of $\mathscr{S}$ is unique.*

Proof We proceed by contradiction. Assume that there exist two elements $\phi_1, \phi_2 \in \mathscr{S}$ such that $\phi_1(0,0) = \phi_2(0,0)$ but $\phi_1 \not\equiv \phi_2$. We have the following cases:

1. $\operatorname{dom} \phi_1 \neq \operatorname{dom} \phi_2$. If $\sup_j \operatorname{dom} \phi_1 \neq \sup_j \operatorname{dom} \phi_2$, by Proposition 9.2, ϕ_1 and ϕ_2 cannot be graphically ε-close, which contradicts that $\mathscr{S}$ is δS. While if

$$\sup_t \operatorname{dom} \phi_1 \neq \sup_t \operatorname{dom} \phi_2,$$

according to Proposition 9.3, ϕ_1 and ϕ_2 cannot be graphically ε-close, which contradicts that $\mathscr{S}$ is δS. If $\sup_j \operatorname{dom} \phi_1 = \sup_j \operatorname{dom} \phi_2$ and $\sup_t \operatorname{dom} \phi_1 = \sup_t \operatorname{dom} \phi_2$, since $\operatorname{dom} \phi_1 \neq \operatorname{dom} \phi_2$, there exists $(t^\star, j^\star) \in \operatorname{dom} \phi_1$ such that $(t^\star, j^\star) \notin \operatorname{dom} \phi_2$. Without loss of generality, assume the ϕ_1 and ϕ_2 have their domains of definition unbounded in the t direction. It must be one of the following cases:

a. $(t^\star, \bar{j}) \in \operatorname{dom} \phi_2$ for some $j^\star \neq \bar{j} \in \mathbb{N}$. Then,

i. if $\bar{j} < j^\star$, it follows that there exists $\bar{t} > t^\star$ such that $(\bar{t}, j^\star) \in \operatorname{dom} \phi_2$. Moreover, $(t, j^\star) \notin \operatorname{dom} \phi_2$ for all $t \in [t^\star - \frac{1}{2}(\bar{t} - t^\star), t^\star + \frac{1}{2}(\bar{t} - t^\star)]$. Then, for $\varepsilon = \frac{1}{2}(\bar{t} - t^\star)$, there does not exists $(t, j^\star) \in \operatorname{dom} \phi_2$ such that $|t - t^\star| \leq \varepsilon$ and $d(\phi_1(t^\star, j^\star), \phi_2(t, j^\star)) \leq \varepsilon$. This contradicts the fact that ϕ_1 and ϕ_2 are graphically ε-close due to the set $\mathscr{S}$ being δS.

ii. the case when $\bar{j} > j^\star$ follows similarly.

b. $(\bar{t}, j^\star) \in \operatorname{dom} \phi_2$ for some $\bar{t} \neq t^\star$ and $\bar{t} \in \mathbb{R}_{\geq 0}$. Then,

i. if $\bar{t} < t^\star$, let $\bar{t}' = \max\{t : (t, j^\star) \in \operatorname{dom} \phi_2, t \leq t^\star\}$. Then, $\bar{t}' < t^\star$. Furthermore, either $j^\star = \sup_j \operatorname{dom} \phi_2$ or $(\bar{t}', j^\star + 1) \in \operatorname{dom} \phi_2$. In either case, pick $\varepsilon = \frac{1}{2}(t^\star - \bar{t}')$, and note it is not possible to find $(t, j^\star) \in \operatorname{dom} \phi_2$ such that $|t - t^\star| \leq \varepsilon$ and $d(\phi_2(t, j^\star), \phi_1(t^\star, j^\star)) \leq \varepsilon$. This contradicts the fact that ϕ_1 and ϕ_2 are graphically ε-close.

ii. the case when $\bar{t} > t^\star$ follows similarly.

2. $\operatorname{dom}\phi_1 = \operatorname{dom}\phi_2$ but there exists $(t^\star, j^\star) \in \operatorname{dom}\phi_1$ such that $\phi_1(t^\star, j^\star) \neq \phi_2(t^\star, j^\star)$. Suppose $(t^\star, j^\star)$ is not an "end point," i.e., $(t^\star, j^\star - 1) \notin \operatorname{dom}\phi_1$ and $(t^\star, j^\star + 1) \notin \operatorname{dom}\phi_1$. Denote $\bar{\varepsilon} = d(\phi_1(t^\star, j^\star), \phi_2(t^\star, j^\star)) > 0$. Since $t \mapsto \phi_1(t, j^\star)$ and $t \mapsto \phi_2(t, j^\star)$ are locally absolutely continuous for all t such that $(t, j^\star) \in \operatorname{dom}\phi_1$ and $(t, j^\star) \in \operatorname{dom}\phi_2$ according to Definition 9.2, there exists $\delta > 0$ such that $|t - t^\star| \leq \delta$ implies that $d(\phi_1(t, j^\star), \phi_1(t^\star, j^\star)) \leq \frac{1}{2}\bar{\varepsilon}$ and $d(\phi_2(t, j^\star), \phi_2(t^\star, j^\star)) \leq \frac{1}{2}\bar{\varepsilon}$. Therefore, by triangle inequality,

$$\begin{aligned} d(\phi_1(t, j^\star), \phi_2(t^\star, j^\star)) &\geq d(\phi_1(t^\star, j^\star), \phi_2(t^\star, j^\star)) - d(\phi_1(t, j^\star), \phi_1(t^\star, j^\star)) \\ &\geq \bar{\varepsilon} - \frac{1}{2}\bar{\varepsilon} = \frac{1}{2}\bar{\varepsilon}. \end{aligned}$$

Thus, for $\varepsilon = \frac{1}{4}\bar{\varepsilon}$, no matter how small δ is chosen, we have that

$$d(\phi_1(0,0), \phi_2(0,0)) = 0 < \delta$$

and ϕ_1 and ϕ_2 are not graphically ε-close which contradicts the assumption that the set $\mathscr{S}$ is δS with respect to d. The situation where $(t^\star, j^\star)$ is an "end point" can be proved similarly. □

When the set $\mathscr{S}$ is generated by solutions to a hybrid system $\mathscr{H} = (C, f, D, g)$, a sufficient condition for guaranteeing uniqueness of maximal solutions requires f to be locally Lipschitz and no flow from $C \cap D$—a rigorous statement is given in [4, Proposition 2.11]. According to Proposition 9.4, assuming uniqueness of solutions to $\mathscr{H}$ is not at all restrictive, in fact, when studying incremental graphical stability, it is necessary. Hence, in the following results we impose the following uniqueness of solutions assumption.

Assumption 9.1 The hybrid system $\mathscr{H} = (C, f, D, g)$ is such that each maximal solution ϕ to $\mathscr{H}$ is unique.

Next, we show that, to have δS or δLA, a finite-time convergence property within a neighborhood of the jump set D is a necessary condition for a set of hybrid trajectories generated by hybrid system $\mathscr{H}$. Indeed, without the finite-time convergence property nearby D and $g(D)$, the graphs of the solutions would not be graphically close.

Theorem 9.1 (necessary condition for δS and δLA) *Consider a hybrid system $\mathscr{H} = (C, f, D, g)$ with state $z \in \mathbb{R}^n$ satisfying Assumption 9.1 and the hybrid basic conditions. Suppose $D \neq \emptyset$ and $g(D) \subset C \cup D$. If $\mathscr{S}_{\mathscr{H}}$ is δS or δLA with respect to a metric d, then there exists $\delta_0 > 0$ such that each maximal solution ϕ to $\mathscr{H}$ from $\phi(0,0)$ satisfying $|\phi(0,0)|_D^d \leq \delta_0$ and $\phi(0,0) \in C$ converges to D within finite time, i.e., there exists $s > 0$ such that $|\phi(s,0)|_D^d = 0$.*

Proof Let $\varepsilon > 0$ be given. Proceeding by contradiction, for all $\delta_0 > 0$, there exists $\phi \in \mathscr{S}_{\mathscr{H}}$ satisfying

$$\phi(0,0) \in C, \quad |\phi(0,0)|_D^d \le \delta_0 \tag{9.23}$$

and $|\phi(t,0)|_D^d > 0$ for all $(t,0) \in \operatorname{dom}\phi$. Let $z^\star \in D$ be such that $|\phi(0,0)|_D^d = d(\phi(0,0), z^\star) = \delta_0$. Consider a solution $\phi_1 \in \mathscr{S}_{\mathscr{H}}$ from $z^\star$. Then, we have that $d(\phi_1(0,0), \phi(0,0)) \le \delta_0$ which implies that ϕ_1 and ϕ are graphically ε-close due to $\mathscr{S}$ being δS with respect to d (using $\delta = \delta_0$ in the definition). Since each maximal solution to $\mathscr{H}$ is unique under Assumption 9.1, $(0,1) \in \operatorname{dom}\phi_1$. Then, since $\phi(t,0) \notin D$ for all $(t,0) \in \operatorname{dom}\phi$, there does not exist $(s,1) \in \operatorname{dom}\phi$ such that $d(\phi_1(0,1), \phi(s,1)) < \varepsilon$ with $|s| \le \varepsilon$. This contradicts the assumption that ϕ and ϕ_1 are graphically ε-close. Now suppose $\mathscr{S}$ is δLA. For any $T > 0$, $t + j \ge T$ and $(t,j) \in \operatorname{dom}\phi_1$, there does not exist $(s,j) \in \operatorname{dom}\phi$ with $|s - t| \le \varepsilon$ such that $d(\phi_1(t,j), \phi(s,j)) \le \varepsilon$. This contradicts the fact that ϕ_1 and ϕ are graphically eventually ε-close. □

The δS property leads to the following necessary condition pertaining to dependence of solutions with respect to initial conditions.

Proposition 9.5 (necessary condition for δS) *Consider a hybrid system $\mathscr{H} = (C, f, D, g)$ with state $z \in \mathbb{R}^n$ satisfying Assumption 9.1. Suppose $\mathscr{S}_{\mathscr{H}}$ is δS with respect to a metric d. Then, $\mathscr{S}_{\mathscr{H}}$ satisfies the following property: for every $\phi \in \mathscr{S}_{\mathscr{H}}$, and for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every solution $\bar{\phi} \in \mathscr{S}_{\mathscr{H}}(\phi(0,0) + \delta\mathbb{B})$, $\bar{\phi}$ and ϕ are graphically ε-close with respect to d.*

Proof Since the set $\mathscr{S}_{\mathscr{H}}$ is δS, for a given $\varepsilon > 0$, there exists $\bar{\delta} > 0$ such that for $\phi_1, \phi_2 \in \mathscr{S}_{\mathscr{H}}$,

$$d(\phi_1(0,0), \phi_2(0,0)) \le \bar{\delta} \Longrightarrow \phi_1, \phi_2 \text{ are graphically } \varepsilon\text{-close.}$$

Let $\delta > 0$ be small enough such that $|\phi_1(0,0) - \phi_2(0,0)| \le \delta$ implies that

$$d(\phi_1(0,0), \phi_2(0,0)) \le \bar{\delta}.$$

Then, for any ϕ and $\bar{\phi}$ picked as in the theorem, $|\phi(0,0) - \bar{\phi}(0,0)| \le \delta$ implies that $d(\bar{\phi}(0,0), \phi(0,0)) \le \bar{\delta}$. Therefore, using the δS property of the set $\mathscr{S}_{\mathscr{H}}$, $\bar{\phi}$ and ϕ are graphically ε-close. □

9.3.2 Sufficient Conditions

To establish sufficient conditions for δLAS, we impose the following assumptions. The first assumption is that each maximal solution to $\mathscr{H}$ has its domain of definition unbounded in the t direction. The second assumption enables each maximal solution to $\mathscr{H}$ to flow for sufficient amount of time in between jumps. A sufficient condition for Assumption 9.3 can be found in [12, Lemma 2.7].

Assumption 9.2 The hybrid system $\mathscr{H} = (C, f, D, g)$ is such that every $\phi \in \mathscr{S}_{\mathscr{H}}$ satisfies $\sup_t \operatorname{dom}\phi = \infty$.

Assumption 9.3 The hybrid system $\mathscr{H} = (C, f, D, g)$ is such that there exists $\gamma > 0$ such that for each $\phi \in \mathscr{S}_{\mathscr{H}}$, the flow time between two consecutive jumps is lower bounded by γ.

Moreover, we will use the following forward invariance notion.

Definition 9.9 (*forward invariance from away of* D) A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be forward invariant for $\mathscr{H}$ from away of D if for each solution ϕ to $\mathscr{H}$ from $\phi(0,0) \in \mathscr{A} \setminus D$, $\phi(t,0) \in \mathscr{A}$ for all $(t,0) \in \operatorname{dom}\phi$.

Remark 9.4 Note that the standard forward invariance notion for a set captures the property that every solution from the set stays within the set for all time, see, e.g., [4, Definition 6.25].

Now, we are ready to present the sufficient condition.

Theorem 9.2 (δLAS through flow for generic D) *Consider a hybrid system* $\mathscr{H} = (C, f, D, g)$ *with state* $z \in \mathbb{R}^n$. *Suppose* $\mathscr{H}$ *satisfies Assumptions 9.1, 9.2 and 9.3, and the hybrid basic conditions. Let* γ *be generated from Assumption 9.3. If there exist* $P = P^\top > 0$, $\beta > 0$, *and* $\delta_0 > 0$ *such that* $\mathscr{H}$ *satisfies*

(1) $\nabla f^\top(z)P + P\nabla f(z) \leq -2\beta P$ *for all* $z \in \overline{\operatorname{con}}C$*;*

(2) for each $\delta \in [0, \delta_0]$, *each* $\phi \in \mathscr{S}_{\mathscr{H}}$ *from* $\phi(0,0)$ *satisfying*

$$\phi(0,0) \in C, \quad |\phi(0,0)|_D = \delta \tag{9.24}$$

is such that there exists $s \in [0, \delta]$ *for which we have*

$$|\phi(s,0)|_D = 0 \tag{9.25}$$

and the set $\phi(s,0) + \delta\mathbb{B}$ *is forward invariant from away of D, and each* $\bar{\phi} \in \mathscr{S}_{\mathscr{H}}(g(\phi(s,0)) + \delta\mathbb{B})$ *satisfies*

$$\bar{\phi}(t,0) \in g(\phi(s,0)) + \delta\mathbb{B} \tag{9.26}$$

for all $t \in [0, s]$*;*

(3) the jump map g is locally Lipschitz on D with Lipschitz constant $L_1 \in [0,1]$,[2] *i.e.,* $|g(z_1) - g(z_2)| \leq L_1|z_1 - z_2|$ *for all* $z_1, z_2 \in D$ *such that* $|z_1 - z_2| \leq \delta_0$*; and*

(4) $c < \exp(\beta\gamma)$, *where* $c = \sqrt{\frac{\bar{\lambda}(P)}{\underline{\lambda}(P)}}$*;*

then, the set $\mathscr{S}_{\mathscr{H}}$ *is* δ*LAS with d being the Euclidean distance.*

Proof Given $\varepsilon > 0$, and using δ_0, γ as in the assumption, consider $\phi_1, \phi_2 \in \mathscr{S}_{\mathscr{H}}$ such that $|\phi_1(0,0) - \phi_2(0,0)| < \delta$, where δ is chosen such that

[2] Such g is also known as a weak contraction map.

$$0 < \delta \le \min\left\{\frac{\varepsilon}{c}, \frac{\delta_0}{c}, \gamma - \frac{1}{\beta}\ln c\right\}.$$

First, we show that $\mathcal{S}_{\mathcal{H}}$ is δS for the case when $\phi_1(0,0), \phi_2(0,0) \in C$ and $\sup_j \operatorname{dom}\phi_1 = \sup_j \operatorname{dom}\phi_2 = 0$, i.e., no jump occurs to either ϕ_1 or ϕ_2. By the generalized mean value theorem (for vector-valued functions), for almost all $(t,0) \in \operatorname{dom}\phi_1 (= \operatorname{dom}\phi_2 = [0,\infty) \times \{0\})$, we have that

$$\begin{aligned}\dot{\phi}_1(t,0) - \dot{\phi}_2(t,0) &= f(\phi_1(t,0)) - f(\phi_2(t,0))\\ &= \int_0^1 \nabla f(\eta(t,s))ds\,(\phi_1(t,0) - \phi_2(t,0)),\end{aligned}$$

where $\eta(t,s) = \phi_1(t,0) + s(\phi_2(t,0) - \phi_1(t,0))$. Then, using item (1), for almost all $t \in [0,\infty)$, we have

$$\begin{aligned}&\frac{d}{dt}|\phi_1(t,0) - \phi_2(t,0)|_P^2\\ &= (\phi_1(t,0) - \phi_2(t,0))^\top \left(\int_0^1 \left(\nabla f^\top(\eta(t,s))P + P\nabla f(\eta(t,s))\right) ds\right)(\phi_1(t,0) - \phi_2(t,0))\\ &\le -\int_0^1 2\beta(\phi_1(t,0) - \phi_2(t,0))^\top P(\phi_1(t,0) - \phi_2(t,0))ds\\ &\le -2\beta|\phi_1(t,0) - \phi_2(t,0)|_P^2, \qquad (9.27)\end{aligned}$$

where we used the property that $\eta(t,s) \in \overline{\operatorname{con}} C$ for all $t \in [0,\infty)$ and $s \in [0,1]$. Therefore, by the comparison lemma, we have, for all $t \in [0,\infty)$,

$$|\phi_1(t,0) - \phi_2(t,0)|_P \le \exp(-\beta t)|\phi_1(0,0) - \phi_2(0,0)|_P. \qquad (9.28)$$

Then, using the property

$$\underline{\lambda}(P)|z|^2 \le |z|_P^2 = z^\top P z \le \overline{\lambda}(P)|z|^2 \quad \forall z \in \mathbb{R}^n \qquad (9.29)$$

and the choice of δ, we obtain

$$\begin{aligned}|\phi_1(t,0) - \phi_2(t,0)| &\le \frac{1}{\sqrt{\underline{\lambda}(P)}}|\phi_1(t,0) - \phi_2(t,0)|_P\\ &\le c\exp(-\beta t)|\phi_1(0,0) - \phi_2(0,0)| \le \varepsilon. \qquad (9.30)\end{aligned}$$

Next, we show $\mathcal{S}_{\mathcal{H}}$ is δS for the case when either ϕ_1 or ϕ_2 jump. By the choice of δ and item (2), $\sup_j \operatorname{dom}\phi_1 = \sup_j \operatorname{dom}\phi_2 =: J$. Without loss of generality, assume ϕ_1 jumps first and $J = \infty$. Furthermore, for each $j \in \mathbb{N} \setminus \{0\}$, let $\overline{t}_j = \max_{(t,j-1)\in \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2} t$ and $\overline{t}'_j = \min_{(t,j)\in \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2} t$, and $\overline{t}'_0 = 0$, where $\overline{t}_j$ denotes the minimum time when one of the two solutions ϕ_1, ϕ_2 jumps j times, while $\overline{t}'_j$ denotes the minimum time when both solutions have jumped j times. In fact,

$[\bar{t}'_j, \bar{t}_{j+1}] \times \{j\} \subset \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2$ for all $j \in \mathbb{N}$. For simplicity, assume that the time when j-th jump occurs to ϕ_1 is always smaller than or equal to that of ϕ_2 for $j \in \mathbb{N}$.

(I) If $\phi_1(0,0), \phi_2(0,0) \in C$. Similarly as in (9.30), for all $t \in [0, \bar{t}_1]$, we have that

$$|\phi_1(t,0) - \phi_2(t,0)| \leq c\exp(-\beta t)|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon. \tag{9.31}$$

When $t = \bar{t}_1$, since ϕ_1 jumps first, $\phi_1(\bar{t}_1, 0) \in D$ and $\phi_1(\bar{t}_1, 1) = g(\phi_1(\bar{t}_1, 0))$. Note that under item (3) of Assumption 9.3, $g(D) \cap D = \emptyset$. Then,

a. if $\phi_2(\bar{t}_1, 0) \in D$, i.e., $\bar{t}_1 = \bar{t}'_1$, by (9.31), $|\phi_1(\bar{t}_1, 0) - \phi_2(\bar{t}_1, 0)| \leq \delta$ and

$$\phi_1(\bar{t}_1, 0), \phi_2(\bar{t}_1, 0) \in D.$$

Then, we can apply the argument in item (II);

b. If $\phi_2(\bar{t}_1, 0) \notin D$, i.e., $\bar{t}_1 < \bar{t}'_1$, by (9.31), it follows that $\phi_2(\bar{t}_1, 0) \in (D + \delta\mathbb{B}) \setminus D$. For each $t \in [\bar{t}_1, \bar{t}'_1]$, since, $|\phi_2(\bar{t}_1, 0)|_D \leq |\phi_2(\bar{t}_1, 0) - \phi_1(\bar{t}_1, 0)| = \bar{\delta}_1$ for some $\bar{\delta}_1 \in [0, \delta]$, by (9.24) and (9.25) in item (2), it follows that $\bar{t}'_1 - \bar{t}_1 \leq \delta$. Since the set $\phi(\bar{t}_1, 0) + \bar{\delta}_1\mathbb{B}$ is forward invariant from away of D according to item (2), we obtain, for each $t \in [\bar{t}_1, \bar{t}'_1]$,

$$|\phi_2(t,0) - \phi_1(\bar{t}_1, 0)| \leq |\phi_2(\bar{t}_1, 0) - \phi_1(\bar{t}_1, 0)|. \tag{9.32}$$

Furthermore, since $\phi_1(\bar{t}_1, 0), \phi_2(\bar{t}'_1, 0) \in D$, by item (3),

$$\begin{aligned}|\phi_2(\bar{t}'_1, 1) - \phi_1(\bar{t}_1, 1)| &\leq |\phi_2(\bar{t}'_1, 0) - g(\phi_1(\bar{t}_1, 0))|\\ &\leq |\phi_2(\bar{t}'_1, 0) - \phi_1(\bar{t}_1, 0)|.\end{aligned}$$

Then, since $\phi_1(\bar{t}_1, 1) \in \phi_2(\bar{t}'_1, 1) + \bar{\delta}_1\mathbb{B}$ according to (9.32), by item (2), for each $t \in [\bar{t}_1, \bar{t}'_1]$,

$$\begin{aligned}|\phi_1(t,1) - \phi_2(\bar{t}'_1, 1)| &\leq |\phi_1(\bar{t}_1, 1) - \phi_2(\bar{t}'_1, 1)|\\ &\leq |\phi_1(\bar{t}_1, 0) - \phi_2(\bar{t}'_1, 0)|.\end{aligned} \tag{9.33}$$

In general, for each $j \in \mathbb{N}$ and $t \in [\bar{t}'_j, \bar{t}_{j+1}]$, since $\phi_1(\bar{t}'_j, j), \phi_2(\bar{t}'_j, j) \in C$, similarly as for (9.31), we have

$$|\phi_1(t,j) - \phi_2(t,j)| \leq c\exp(-\beta(t - \bar{t}'_j))|\phi_1(\bar{t}'_j, j) - \phi_2(\bar{t}'_j, j)|. \tag{9.34}$$

While for $j \in \mathbb{N} \setminus \{0\}$ and $t \in [\bar{t}_j, \bar{t}'_j]$, we have $[\bar{t}_j, \bar{t}'_j] \times \{j\} \subset \operatorname{dom}\phi_1$, $[\bar{t}_j, \bar{t}'_j] \times \{j-1\} \subset \operatorname{dom}\phi_2$ and $|\bar{t}'_j - \bar{t}_j| \leq \delta$. Then, similarly as for (9.32) and (9.33), we obtain

i. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}_j, \bar{t}'_j]$:

$$|\phi_2(t,j-1) - \phi_1(\bar{t}_j, j-1)| \leq |\phi_2(\bar{t}_j, j-1) - \phi_1(\bar{t}_j, j-1)|, \tag{9.35}$$

ii. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}_j, \bar{t}'_j]$:

$$|\phi_1(t,j) - \phi_2(\bar{t}'_j, j)| \leq |\phi_1(\bar{t}_j, j-1) - \phi_2(\bar{t}_j, j-1)|. \tag{9.36}$$

Therefore, using (9.34), (9.35), (9.36) and $|\phi_1(0,0) - \phi_2(0,0)| \leq \delta$, it follows that

i. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}'_j, \bar{t}_{j+1}]$:

$$\begin{aligned}
|\phi_1(t,j) - \phi_2(t,j)| &\leq c \exp(-\beta(t - \bar{t}'_j))|\phi_1(\bar{t}'_j, j) - \phi_2(\bar{t}'_j, j)| \\
&\leq c \exp(-\beta(t - \bar{t}'_j))|\phi_1(\bar{t}_j, j-1) - \phi_2(\bar{t}_j, j-1)| \\
&\leq c^2 \exp(-\beta(t - \bar{t}'_j)) \exp(-\beta(\bar{t}_j - \bar{t}'_{j-1})) \\
&\quad \times |\phi_1(\bar{t}'_{j-1}, j-1) - \phi_2(\bar{t}'_{j-1}, j-1)| \\
&\quad \vdots \\
&\leq c^{j+1} \exp(-\beta(t - \bar{t}'_{j-1} + \Delta_j))|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon,
\end{aligned} \tag{9.37}$$

where $\Delta_j := \sum_{k=1}^{j} (\bar{t}_k - \bar{t}'_{k-1})$. In particular, the first inequality in (9.37) uses (9.34) with $t \in [\bar{t}'_j, \bar{t}_{j+1}]$, the second inequality in (9.37) uses (9.36) with $t = \bar{t}'_j$, and the third inequality in (9.37) uses (9.34) with $t = \bar{t}_j$.

ii. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}_j, \bar{t}'_j]$:

$$\begin{aligned}
|\phi_2(t,j-1) - \phi_1(\bar{t}_j, j-1)| &\leq |\phi_2(\bar{t}_j, j-1) - \phi_1(\bar{t}_j, j-1)| \\
&\leq c \exp(-\beta(\bar{t}_j - \bar{t}'_{j-1}))|\phi_1(\bar{t}'_{j-1}, j-1) - \phi_2(\bar{t}'_{j-1}, j-1)| \\
&\quad \vdots \\
&\leq c^j \exp(-\beta \Delta_j)|\phi_1(0,0) - \phi_2(0,0)| \\
&\leq \exp(-(\beta(\gamma - \delta) - \ln c)j)|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon,
\end{aligned} \tag{9.38}$$

where the first inequality follows from (9.35) with $t \in [\bar{t}_j, \bar{t}'_j]$, and the second inequality follows from (9.34) with $t = \bar{t}_j$.

iii. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}_j, \bar{t}'_j]$:

$$\begin{aligned} |\phi_1(t,j) - \phi_2(\bar{t}'_j,j)| &\leq |\phi_1(\bar{t}_j,j-1) - \phi_2(\bar{t}_j,j-1)| \\ &\leq c\exp(-\beta(\bar{t}_j - \bar{t}'_{j-1}))|\phi_1(\bar{t}'_{j-1},j-1) - \phi_2(\bar{t}'_{j-1},j-1)| \\ &\;\vdots \\ &\leq c^j\exp(-\beta\Delta_j)|\phi_1(0,0) - \phi_2(0,0)| \\ &\leq \exp(-(\beta(\gamma-\delta) - \ln c)j)|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon. \end{aligned} \tag{9.39}$$

In particular, the first inequality in (9.39) uses (9.36) with $t \in [\bar{t}_j, \bar{t}'_j]$, and the second inequality in (9.39) uses (9.34) with $t = \bar{t}_j$.

Therefore, ϕ_1 and ϕ_2 are ε-close.

(II) If $\phi_1(0,0), \phi_2(0,0) \in D$, by condition (3) that the jump map is locally Lipschitz on D with Lipschitz constant $L_1 \leq 1$, we obtain

$$|\phi_1(1,0) - \phi_2(1,0)| = |g(\phi_1(0,0)) - g(\phi_2(0,0))| \leq |\phi_1(0,0) - \phi_2(0,0)|. \tag{9.40}$$

Note that after the jump, $\phi_1(1,0), \phi_2(1,0) \in C$, we can apply the arguments in item (I).

(III) If $\phi_1(0,0) \in C, \phi_2(0,0) \in D$, the arguments follow similarly as in item (I).

Therefore, by combining arguments in items (I), (II), (III), it is proved that ϕ_1 and ϕ_2 are ε-close. Note that the case when $J < \infty$ follows similarly. Therefore, $\mathscr{S}_{\mathscr{H}}$ is δS with respect to Euclidean distance.

Now, we show that $\mathscr{S}_{\mathscr{H}}$ is δLA. Consider the case in item (I) (the other cases follow similarly). Note that $\bigcup_{j=1}^{\infty}[\bar{t}'_j, \bar{t}_{j+1}] = \infty$ if $J = \infty$ (or $\bigcup_{j=1}^{J-1}[\bar{t}'_j, \bar{t}_{j+1}] \bigcup [\bar{t}'_J, \infty) = \infty$ if $J < \infty$ respectively). Moreover, since $[\bar{t}'_j, \bar{t}_{j+1}] \times \{j\} \subset \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2$ for all $j \in \mathbb{N}$. Then, on each interval $[\bar{t}'_j, \bar{t}_{j+1}]$, we have that $|\phi_1(t,j+1) - \phi_2(t,j+1)| \leq \exp(-\beta(t - \bar{t}'_j))|\phi_1(\bar{t}'_j,j) - \phi_2(\bar{t}'_j,j)|$ for all $t \in [\bar{t}'_j, \bar{t}_{j+1}]$. In particular, pick

$$\mu = \delta < \min\left\{\frac{\delta_0}{c}, \gamma - \frac{1}{\beta}\ln c\right\},$$

for a given $\varepsilon' > 0$, pick

$$T = -\frac{1}{\beta(\gamma-\delta)}\ln\left(\min\left\{1, \frac{\varepsilon'}{c\delta}\right\}\right).$$

Then, using (9.37), (9.38), and (9.39), we obtain

1. for (t, j) such that $j \geq T$ and $t \in [\bar{t}'_j, \bar{t}_{j+1}]$:

$$\begin{aligned}|\phi_1(t,j) - \phi_2(t,j)| &\leq c\exp(-(\beta(\gamma-\delta) - \ln c)j)|\phi_1(0,0) - \phi_2(0,0)|\\ &\leq c\exp(-(\beta(\gamma-\delta) - \ln c)T)|\phi_1(0,0) - \phi_2(0,0)|\\ &\leq \min\left\{1, \frac{\varepsilon'}{c\delta}\right\} c|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon',\end{aligned}$$

2. for (t, j) such that $j \geq T$ and $t \in [\bar{t}_j, \bar{t}'_j]$:

$$|\phi_2(t,j-1) - \phi_1(\bar{t}_j, j-1)| \leq \exp(-(\beta(\gamma-\delta) - \ln c)j)|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon',$$

3. for (t, j) such that $j \geq T$ and $t \in [\bar{t}_j, \bar{t}'_j]$:

$$|\phi_1(t,j) - \phi_2(\bar{t}'_j, j)| \leq \exp(-(\beta(\gamma-\delta) - \ln c)j)|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon'.$$

Therefore, for ϕ_1, ϕ_2 such that $|\phi_1(0,0) - \phi_2(0,0)| \leq \mu$, ϕ_1, ϕ_2 are eventually ε-close and $\mathcal{S}_{\mathcal{H}}$ is δLA. □

When the jump set D is discrete, the conditions in Theorem 9.2 simplify and we obtain the following result.

Corollary 9.1 (δLAS through flow with D being a discrete set) *Consider a hybrid system $\mathcal{H} = (C, f, D, g)$ with state $z \in \mathbb{R}^n$ and D being a discrete set. Suppose $\mathcal{H}$ satisfies Assumptions 9.1, 9.2, 9.3, and the hybrid basic conditions. Let γ be generated from Assumption 9.3. If there exist $P = P^\top > 0$, $\beta > 0$, and $\delta_0 > 0$ such that $\mathcal{H}$ satisfies*

(1) $\nabla f^\top(z)P + P\nabla f(z) \leq -2\beta P$ *for all* $z \in \overline{\text{con}}C$*;*

(2) for each $\delta \in [0, \delta_0]$*, each* $\phi \in \mathcal{S}_{\mathcal{H}}$ *from* $\phi(0,0)$ *satisfying*

$$\phi(0,0) \in C, \quad |\phi(0,0)|_D = \delta \tag{9.41}$$

is such that there exists $s \in [0, \delta]$ *for which we have*

$$|\phi(s,0)|_D = 0, \quad |\phi(t,0)|_D \leq \delta \quad \forall t \in [0, s], \tag{9.42}$$

and

$$|\bar{\phi}(t,0) - g(\phi(s,0))| \leq \delta \quad \forall t \in [0, s], \tag{9.43}$$

where $\bar{\phi} \in \mathcal{S}_{\mathcal{H}}(g(\phi(s,0)))$*; and*

(3) $c < \exp(\beta\gamma)$*, where* $c := \sqrt{\frac{\bar{\lambda}(P)}{\underline{\lambda}(P)}}$*;*

then, the set $\mathcal{S}_{\mathcal{H}}$ is δLAS with d being the Euclidean distance.

Proof Given $\varepsilon > 0$, and using δ_0, γ as in the assumption, consider $\phi_1, \phi_2 \in \mathcal{S}_{\mathcal{H}}$ such that $|\phi_1(0,0) - \phi_2(0,0)| < \delta$, where δ is chosen such that

$$0 < \delta \leq \min\left\{\frac{\varepsilon}{c}, \frac{\delta_0}{c}, \gamma - \frac{1}{\beta}\ln c\right\}$$

and, for each $z \in D$, $(z + \delta\mathbb{B}) \cap D = \{z\}$; namely $z + \delta\mathbb{B}$ is a small neighborhood around z that does not intersect D. Note that from condition (3), $\gamma - \frac{1}{\beta}\ln c > 0$.

First, we show that $\mathcal{S}_{\mathcal{H}}$ is δS for the case when $\phi_1(0,0), \phi_2(0,0) \in C$ and $\sup_j \operatorname{dom}\phi_1 = \sup_j \operatorname{dom}\phi_2 = 0$, i.e., no jump occurs to either ϕ_1 or ϕ_2. Similarly as derived in (9.27), using condition (1) and the comparison lemma, we obtain for all $(t,0) \in \operatorname{dom}\phi_1 (= \operatorname{dom}\phi_2 = [0,\infty) \times \{0\})$,

$$|\phi_1(t,0) - \phi_2(t,0)|_P \leq c\exp(-\beta t)|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon. \tag{9.44}$$

Next, we show $\mathcal{S}_{\mathcal{H}}$ is δS for the case when either ϕ_1 or ϕ_2 jump. By the choice of δ and item (2), $\sup_j \operatorname{dom}\phi_1 = \sup_j \operatorname{dom}\phi_2 =: J$. This can be verified as follows. When ϕ_1 flows to the jump set D, ϕ_2 is within the δ neighborhood of ϕ_1, then, by item (2), ϕ_2 flows into the jump set D within δ time. Furthermore, since $\gamma \geq \delta$, therefore, ϕ_1 will not jump again before ϕ_2 jumps. Without loss of generality, assume ϕ_1 jumps first and $J = \infty$ (Alternatively, we could pick J large enough, but ∞ suffices). Furthermore, for each $j \in \mathbb{N} \setminus \{0\}$, let $\bar{t}_j = \max_{(t,j-1)\in\operatorname{dom}\phi_1\cap\operatorname{dom}\phi_2} t$ and $\bar{t}'_j = \min_{(t,j)\in\operatorname{dom}\phi_1\cap\operatorname{dom}\phi_2} t$, and $\bar{t}'_0 = 0$, where $\bar{t}_j$ denotes the minimum time in $\operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2$ when at least one of the two solutions ϕ_1, ϕ_2 has jumped j times (note that $\bar{t}_j$ and $\bar{t}'_j$ are not necessarily jump times of both solutions), while $\bar{t}'_j$ denotes the minimum time when both solutions have jumped j times. In fact, $[\bar{t}'_j, \bar{t}_{j+1}] \times \{j\} \subset \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2$ for all $j \in \mathbb{N}$. Note that with Assumption 9.3 and the choice of δ, $\phi_1(\bar{t}_1, 0) = \phi_2(\bar{t}'_1, 0)$ and $\bar{t}'_1 > \bar{t}_1$. By the uniqueness of solutions, $\phi_1(\bar{t}_1, 1) = \phi_2(\bar{t}'_1, 1)$ and ϕ_2 is "following" the trajectory of ϕ_1 after that, which implies that ϕ_1 and ϕ_2 jumps one after another. In particular, after the j-th jump occurs to ϕ_1, the j-th jump occurs to ϕ_2 before the $(j+1)$-th jump occurs to ϕ_1. The derivation follows the steps as in Theorem 9.2. The main difference is that in the derivation of (9.32) and (9.33), instead of using the condition (2) in Theorem 9.2, we use condition (2) of Corollary 9.1.

The proof for δLA follows similarly as that in Theorem 9.2 with $\mu > 0$ and $\mu < \min\left\{\frac{\delta_0}{c}, \gamma - \frac{1}{\beta}\ln c\right\}$.

Remark 9.5 Item (1) in Corollary 9.1 guarantees strict decrease of the distance between every pair of maximal solutions to $\mathcal{H}$ on the intersections of their hybrid time domains. In fact, these conditions guarantee a contraction property of the nonlinear system with right-hand side given by f; see, e.g., [9]. The second item in Corollary 9.1 implies that, over the mismatch parts of their hybrid time domains, the graphical distance between them does not grow. The third item in Corollary 9.1 ensures that every pair of maximal solutions can flow for enough time to overcome the possible overshoot on the distance between them. When $P = I$, the third condition is satisfied for free.

The necessity of item (2) of Corollary 9.1 is justified in Theorem 9.1. This condition is guaranteed by the following sufficient condition.

Proposition 9.6 *Consider a hybrid system $\mathscr{H} = (C, f, D, g)$ with state $z \in \mathbb{R}^n$ and D being a discrete set. Suppose $\mathscr{H}$ satisfies the hybrid basic conditions. Then, item (2) of Corollary 9.1 holds if there exists $\delta_0 > 0$ such that, for any $z^\star \in D$, the following holds: there exist $c_1, c_2 > 0$, $c_2 \in (0, c_1]$, and $\alpha \in (0, 1)$ such that*

(1) the function $V_1(z) := |z - z^\star|^2$ satisfies $\langle \nabla V_1(z), f(z)\rangle + c_1 V_1^\alpha(z) \leq 0$ and $|z - z^\star|^{1-2\alpha} \leq c_1(1-\alpha)$ for all $z \in C \bigcap ((z^\star + \delta_0\mathbb{B}) \setminus D)$,
(2) the function $V_2(z) := |z - g(z^\star)|^2$ satisfies $\langle \nabla V_2(z), f(z)\rangle - c_2 V_2^\alpha(z) \leq 0$ for all $z \in C \bigcap (g(z^\star) + \delta_0\mathbb{B})$.

Proof Let δ be such that $0 < \delta \leq \delta_0$ and for each $z \in D$, $(z + \delta\mathbb{B}) \cap D = \{z\}$. Given $z^\star \in D$, consider $\phi \in \mathscr{S}_{\mathscr{H}}(C \bigcap \big((z^\star + \delta\mathbb{B}) \setminus D\big))$. By item (1) in Proposition 9.6 and by integrating $t \mapsto \frac{dV_1^{1-\alpha}}{dt}(\phi(t, 0))$ over $[0, t_1] \times \{0\} \subset \operatorname{dom}\phi$, it follows that

$$V_1(\phi(t,0))^{1-\alpha} \leq -c_1(1-\alpha)t + V_1(\phi(0,0))^{1-\alpha} \quad \forall (t,0) \in \operatorname{dom}\phi. \tag{9.45}$$

Note that, since V_1 is a positive definite function with respect to $z^\star$, using the property that $|z - z^\star|^{1-2\alpha} \leq c_1(1-\alpha)$ for all $z \in C \bigcap ((z^\star + \delta_0\mathbb{B}) \setminus D)$, ϕ converges to $z^\star$ within $t^\star$ seconds, where

$$t^\star = \frac{V_1(\phi(0,0))^{1-\alpha}}{c_1(1-\alpha)} = \frac{|\phi(0,0) - z^\star|^{2-2\alpha}}{c_1(1-\alpha)} \tag{9.46}$$

$$\implies \quad t^\star \leq \frac{c_1(1-\alpha)|\phi(0,0) - z^\star|}{c_1(1-\alpha)} = |\phi(0,0) - z^\star|. \tag{9.47}$$

Moreover, by (9.45) and the fact that $V_1(\phi(t,0)) = |\phi(t,0) - z^\star|^2$,

$$|\phi(t,0) - z^\star| = \sqrt{V_1(\phi(t,0))} \leq \sqrt{V_1(\phi(0,0))} = |\phi(0,0) - z^\star| \quad \forall (t,0) \in \operatorname{dom}\phi. \tag{9.48}$$

It is implied from (9.47) that there exists $s \in [0, |\phi(0,0) - z^\star|]$ such that $\phi(s,0) = z^\star$ and, from (9.48), $|\phi(t,0)|_D \leq \delta$ for all $t \in [0, s]$. Now using item (2) of the assumptions and proceeding similarly to arrive to (9.45), the maximal solution $\bar{\phi} \in \mathscr{S}_{\mathscr{H}}(g(\phi(s,0)))$ satisfies

$$V_2(\bar{\phi}(t,0))^{1-\alpha} \leq c_2(1-\alpha)t + V_2(\bar{\phi}(0,0))^{1-\alpha} \quad \forall (t,0) \in \operatorname{dom}\bar{\phi}. \tag{9.49}$$

Since $V_2(\bar{\phi}(0,0)) = |g(\phi(s,0)) - g(z^\star)|^2 = 0$, using (9.46), (9.48), and (9.49), we obtain that for all $t \in [0, s]$,

$$
\begin{aligned}
|\bar{\phi}(t,0) - g(\phi(s,0))| = |\bar{\phi}(t,0) - g(z^\star)| = \sqrt{V_2(\bar{\phi}(t,0))} &\leq \sqrt{(c_2(1-\alpha)t)^{\frac{1}{1-\alpha}}} \\
&\leq \sqrt{\left(c_2(1-\alpha)\frac{V_1(\phi(0,0))^{1-\alpha}}{c_1(1-\alpha)}\right)^{\frac{1}{1-\alpha}}} \\
&\leq \sqrt{\left(\frac{c_2}{c_1}\right)^{\frac{1}{1-\alpha}}}|\phi(0,0) - z^\star| \leq |\phi(0,0) - z^\star| \leq \delta,
\end{aligned}
$$

where we used the property $0 < c_2 \leq c_1$. □

Remark 9.6 In item (1) of Proposition 9.6, if c_1 and α can be chosen as $c_1 \geq 2$ and $\alpha = \frac{1}{2}$, then, for any $z^\star \in D$, the condition $|z - z^\star|^{1-2\alpha} \leq c_1(1-\alpha)$ is true for any $z \in \mathbb{R}^n$ since $|z - z^\star|^{1-2\alpha} = |z - z^\star|^0 = 1 \leq \frac{1}{2}c_1$.

The following example illustrates the sufficient condition in Corollary 9.1, for which Proposition 9.6 is used to guarantee that item (2) in Corollary 9.1 holds.

Example 9.6 Consider the following hybrid system $\mathscr{H} = (C, f, D, g)$ with state $z \in \mathbb{R}$ and data given by

$$
\begin{aligned}
f(z) &= -z \qquad \forall\, z \in \mathbb{R} \\
C &:= \bigcup_{i \in \{2k : k \in \mathbb{N}\}} [i, i+1] \\
g(z) &= z - 1 \qquad \forall\, z \in D := \{2k : k \in \mathbb{N} \setminus \{0\}\},
\end{aligned}
$$

where $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{N} \to \mathbb{N}$. The conditions in Corollary 9.1 can be verified as follows. Each $\phi \in \mathscr{S}_{\mathscr{H}}$ is complete and its domain is unbounded in the t direction. Moreover, the flow map is continuously differentiable on $\overline{\text{con}}C$. Furthermore, for any $\phi \in \mathscr{S}_{\mathscr{H}}$ from $\phi(0,0) \in (C \cup D)$, denote $\rho^\star := \max\{x : x \in C, x \leq \phi(0,0)\}$. If $\rho^\star \leq 1$, then ϕ never jumps and the jump time between two consecutive jumps is bounded below by ∞. If $\rho^\star \geq 2$, the flow time between two consecutive jumps of ϕ is bounded below by $\bar{\rho} := \ln \frac{\rho^\star}{\rho^\star - 1}$. For all $z \in \overline{\text{con}}C$, $\nabla f(z) + \nabla f(z)^\top = -2$, so item (1) in Corollary 9.1 is satisfied with $\beta = 1$ and $P = I$. Moreover, given $z^\star \in D$, the function $V_1(z) = |z - z^\star|^2$ satisfies $\langle \nabla V_1(z), f(z)\rangle = 2(z - z^\star)(-z) \leq -2z^\star(z - z^\star) = -2z^\star V_1^{\frac{1}{2}}(z)$ for $z \in C \bigcap ((z^\star + \bar{\rho}\mathbb{B}) \setminus D)$, where we used the property that $z \geq z^\star$ for all $z \in C \bigcap ((z^\star + \bar{\rho}\mathbb{B}) \setminus D)$. Furthermore, the function $V_2(z) = |z - g(z^\star)|^2$ satisfies $\langle \nabla V_2(z), f(z)\rangle = 2(z - g(z^\star))(-z) \leq 2z^\star(g(z^\star) - z) = 2g(z^\star)V_2^{\frac{1}{2}}(z)$ for $z \in (g(z^\star) + \bar{\rho}\mathbb{B}) \bigcap C$, where we used the property that $z \leq g(z^\star)$ for all $z \in (g(z^\star) + \bar{\rho}\mathbb{B}) \bigcap C$ and $g(z^\star) = z^\star - 1 < z^\star$. Then, Proposition 9.6 is satisfied with $c_1 = 2z^\star$, $\alpha = 1/2$, and $c_2 = 2(z^\star - 1) \in (0, c_1]$. Thus, the condition (2) in Corollary 9.1 is verified. Note that the condition (3) in Corollary 9.1 holds for free since $\beta = 1$, $c = 1$ and $\gamma = \bar{\rho} > 0$. Then, by Corollary 9.1, we have that $\mathscr{H}$ is δLAS. △

The following result establishes a sufficient condition for a set $\mathcal{S}_{\mathcal{H}}$ to be δLAS "through jumps." In particular, under such conditions, the graphical distance between any two maximal solutions to a hybrid system $\mathcal{H}$ strictly decreases during jumps. Due to such requirement, we need to impose the following assumption to guarantee that every maximal solution to $\mathcal{H}$ jumps infinitely many times.

Assumption 9.4 The hybrid system $\mathcal{H} = (C, f, D, g)$ is such that every $\phi \in \mathcal{S}_{\mathcal{H}}$ satisfies $\sup_j \operatorname{dom} \phi = \infty$.

Theorem 9.3 (δLAS through jump for generic D) *Consider a hybrid system $\mathcal{H} = (C, f, D, g)$ with state $z \in \mathbb{R}^n$. Suppose $\mathcal{H}$ satisfies Assumptions 9.1, 9.3, and the hybrid basic conditions. If there exist $\delta_0, L_1, L_2 > 0, P = P^\top > 0$ such that*

(1) $\nabla f(z)^\top P + P\nabla f(z) \leq 0$ *for all* $z \in \overline{\text{con}}C$*;*
(2) for each $\delta \in [0, \delta_0]$*, each maximal solution* ϕ *to* $\mathcal{H}$ *from* $\phi(0,0)$ *satisfying*

$$\phi(0,0) \in C, \quad |\phi(0,0)|_D = \delta$$

satisfies $|\phi(s,0)|_D = 0$ *for some* $s \in [0, \delta]$*;*
(3) for each $z \in D$ *and each* $\delta \in [0, \delta_0]$*, the set* $z + \delta\mathbb{B}$ *is forward invariant for* $\mathcal{H}$ *from away of D;*
(4) the jump map g is locally Lipschitz on D with Lipschitz constant L_1*, i.e.,* $|g(z_1) - g(z_2)| \leq L_1|z_1 - z_2|$ *for all* $z_1, z_2 \in D$ *such that* $|z_1 - z_2| \leq \delta_0$*;*
(5) f is bounded on $\overline{\text{con}}C$ *with bound* L_2*, i.e.,* $|f(z)| \leq L_2$ *for all* $z \in \overline{\text{con}}C$*;*
(6) $c(L_1 + L_2) \leq 1$ *where* $c = \sqrt{\frac{\overline{\lambda}(P)}{\underline{\lambda}(P)}}$*;*

then, the set $\mathcal{S}_{\mathcal{H}}$ is δS with d being the Euclidean distance. Furthermore, if L_1 and L_2 can be chosen such that $c(L_1 + L_2) < 1$ and $\mathcal{H}$ satisfies Assumption 9.4, then, $\mathcal{S}_{\mathcal{H}}$ is δLAS with d being the Euclidean distance.

Proof Given $\varepsilon > 0$ and using δ_0 as in the item (2)–(5) of assumption and γ as in Assumption 9.3, consider $\phi_1, \phi_2 \in \mathcal{S}_{\mathcal{H}}$ such that $|\phi_1(0,0) - \phi_2(0,0)| < \delta$, where δ is chosen such that

$$0 < \delta \leq \min\left\{\frac{\varepsilon}{c}, \frac{\delta_0}{c}, \gamma\right\}.$$

First, we show that $\mathcal{S}_{\mathcal{H}}$ is δS for the case when $\phi_1(0,0), \phi_2(0,0) \in C$ and $\sup_j \phi_1 = \sup_j \operatorname{dom} \phi_2 = 0$. Similarly as derived in (9.27), using item (1) and the comparison lemma, we have, for all $t \in [0, \infty)$,

$$|\phi_1(t,0) - \phi_2(t,0)| \leq c|\phi_1(0,0) - \phi_2(0,0)| \leq \varepsilon.$$

Next, we show $\mathcal{S}_{\mathcal{H}}$ is δS for the case when either ϕ_1 or ϕ_2 jump. By the choice of δ and item (2), $\sup_j \operatorname{dom} \phi_1 = \sup_j \operatorname{dom} \phi_2 =: J$. Without loss of generality, assume ϕ_1 jumps first and $J = \infty$. Furthermore, for each $j \in \mathbb{N} \setminus \{0\}$, let

$\bar{t}_j = \max_{(t,j-1)\in \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2} t$ and $\bar{t}'_j = \min_{(t,j)\in \operatorname{dom}\phi_1 \cap \operatorname{dom}\phi_2} t$, and $\bar{t}'_0 = 0$. For simplicity, assume that the time when j-th jump occurs to ϕ_1 is always smaller than or equal to that of ϕ_2 for $j \in \mathbb{N}$.

(I) If $\phi_1(0,0), \phi_2(0,0) \in C$, using item (1) and similar derivations in Theorem 9.2, we obtain for all $t \in [0, \bar{t}_1]$,

$$|\phi_1(t,0) - \phi_2(t,0)| \le c|\phi_1(0,0) - \phi_2(0,0)| \le \varepsilon. \tag{9.50}$$

When $t = \bar{t}_1$, since ϕ_1 jumps first, $\phi_1(\bar{t}_1, 0) \in D$ and $\phi_1(\bar{t}_1, 1) = g(\phi_1(\bar{t}_1, 0))$. Note that under Assumption 9.3, $g(D) \cap D = \emptyset$. Then,

a. if $\phi_2(\bar{t}_1, 0) \in D$, i.e., $\bar{t}_1 = \bar{t}'_1$, by (9.50) and the choice of δ, $|\phi_1(\bar{t}_1, 0) - \phi_2(\bar{t}_1, 0)| \le \delta$ and $\phi_1(\bar{t}_1, 0), \phi_2(\bar{t}_1, 0) \in D$. By condition (4) and (9.50),

$$|\phi_1(\bar{t}_1, 1) - \phi_2(\bar{t}_1, 1)| \le L_1 |\phi_1(\bar{t}_1, 0) - \phi_2(\bar{t}_1, 0)| \le \varepsilon. \tag{9.51}$$

Since $\phi_1(\bar{t}_1, 1), \phi_2(\bar{t}_1, 1) \in C$, we can recursively apply the arguments in (I)

b. If $\phi_2(\bar{t}_1, 0) \notin D$, i.e., $\bar{t}_1 < \bar{t}'_1$, by (9.50), it follows that $\phi_2(\bar{t}_1, 0) \in (D + \delta\mathbb{B}) \setminus D$. For each $t \in [\bar{t}_1, \bar{t}'_1]$, since, $\phi_1(\bar{t}_1, 0) \in D$ and $|\phi_2(\bar{t}_1, 0) - \phi_1(\bar{t}_1, 0)| = \bar{\delta}_1$ for some $\bar{\delta}_1 \in [0, \delta]$, by item (2) and item (3), we obtain

i. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}'_j, \bar{t}_{j+1}]$:

$$|\phi_1(t,j) - \phi_2(t,j)| \le (L_1 + L_2)^j c^{j+1} |\phi_1(0,0) - \phi_2(0,0)| \le \varepsilon. \tag{9.52}$$

ii. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}_j, \bar{t}'_j]$:

$$|\phi_2(t,j-1) - \phi_1(\bar{t}_j, j-1)| \le (L_1 + L_2)^{j-1} c^j |\phi_1(0,0) - \phi_2(0,0)| \le \varepsilon. \tag{9.53}$$

iii. for each $j \in \mathbb{N} \setminus \{0\}$ and each $t \in [\bar{t}_j, \bar{t}'_j]$:

$$|\phi_1(t,j) - \phi_2(\bar{t}'_j, j)| \le (L_1 + L_2)^{j-1} c^j |\phi_1(0,0) - \phi_2(0,0)| \le \varepsilon. \tag{9.54}$$

Therefore, ϕ_1 and ϕ_2 are ε-close.

The other cases follow similarly. Therefore, $\mathcal{S}_{\mathcal{H}}$ is δS with respect to Euclidean distance.

When the domain of each $\phi \in \mathcal{S}_{\mathcal{H}}$ is unbounded in the j direction and $c(L_1 + L_2) < 1$, the δLA property can be established by picking $0 < \mu \le \min\left\{\frac{\delta_0}{c}, \gamma\right\}$, for a given $\varepsilon' > 0$, pick $T = \max\left\{1, \log_{c(L_1+L_2)} \frac{\varepsilon'}{c\mu}\right\} + 1$. □

The following example illustrates the conditions in Theorem 9.3.

Example 9.7 Consider a timer system $\mathscr{H}$ with state $z \in \mathbb{R}$ and data given by

$$\begin{aligned} \dot{z} &= -1 \qquad && z \geq 0, \\ z^+ &= 1 \qquad && z = 0. \end{aligned}$$

Each maximal solution ϕ to it has a domain that is unbounded in the t and j directions. Moreover, the flow time between two consecutive jumps of ϕ is lower bounded by 1. The condition in item (1) of Theorem 9.3 can be verified with $P = I$ as $\nabla f(z) + \nabla f(z)^\top = 0$ for all $z \in \overline{\mathrm{con}} C$. The condition in item (2) can be verified according to Proposition 9.6. Consider $\delta_0 \in (0, 1)$ and the function $V(z) = |z|_D^2 = z^2$. For each $z \in (D + \delta_0 \mathbb{B}) \cap C \setminus D$, we have $\langle \nabla V(z), f(z) \rangle = -2z = -2V^{\frac{1}{2}}(z)$, where we used the property that $z \in [0, 1]$. Item (3) of Theorem 9.3 follows from the fact $D = \{0\}$ is a singleton and $\langle \nabla V(z), f(z) \rangle = -2z < 0$ for all $z \in (D + \delta_0 \mathbb{B}) \cap C \setminus D$. Item (4) of Theorem 9.3 is satisfied with $L_1 = 0$, and item (5) of Theorem 9.3 is satisfied with $L_2 = 1$. Item (6) of Theorem 9.3 holds for free since $c = 1$. Therefore, the set $\mathscr{S}_{\mathscr{H}}$ is δS with d being the Euclidean distance. △

9.4 Final Remarks

In this chapter, we introduced and studied several notions of graphical incremental stability for hybrid systems. When compared to the pointwise distance, the proposed graphical notion can be applied to systems with "peaking phenomenon," which is a typical behavior in tracking and observer design for hybrid systems. Graphical incremental stability involves a convergence property where solutions converge to each other. Several sufficient and necessary conditions for a hybrid system to be graphically incrementally stable and graphically incrementally attractive were provided and illustrated in examples.

An alternative approach to use the graphical distance is to prioritize ordinary time. When one prioritizes ordinary time t, i.e., studying the incremental property of solutions' projection to the t direction, it leads to the result as in [7]. Note that the notion defined therein imposes the incremental stability property in some of the state components. This is due to the fact that when studying the incremental stability for certain hybrid systems, such as mechanical systems and dynamical systems that are dominated by continuous-time behavior, one may not be interested in having state components pertaining to variables such as timers, logic variables, and memory states to have the incremental stability property.

The results in [7] cover results for continuous-time system as in [2]. In [2], several sufficient and necessary conditions for continuous-time systems to be incrementally stable are provided. For continuous-time systems, incremental stability has also been studied in more general spaces and using general distance notions, such as the Riemannian distance in the context of contraction theory; see, e.g., the study of contracting and nonexpansive flows in [13, 14], the local arguments in [9], and the regional results in [15] in the context of observer design. Due to often being misinterpreted as a property of convergent systems [16], the authors in [17] provide a rigorous comparison between incremental stability and the property of convergent systems, and conclude that neither implies the other.

Following the ideas in [7, 18], one could alternatively define a notion that prioritizes jumps and mimics the case of purely discrete-time systems. Unfortunately, such a notion would only apply to a narrow class of hybrid system due to the general aforementioned difficulty. For instance, for the rather elementary set of hybrid trajectories in Example 9.1, the pointwise distance between every pair of trajectories with different initial conditions is clearly nondecreasing as a function of t, while the graphical distance between them is small and, as shown in Example 9.3, the system is graphically incrementally stable. As argued in this chapter, for hybrid systems that exhibit a "peaking phenomenon," see, e.g., [19, 20], approaches that prioritize ordinary time t or jump time j in the incremental stability notion do not have broad applicability in the analysis of incremental stability for hybrid dynamical systems.

Acknowledgements This research has been partially supported by the National Science Foundation under CAREER Grant no. ECCS-1450484 and Grant no. CNS-1544396, and by the Air Force Office of Scientific Research under YIP Grant no. FA9550-12-1-0366 and Grant no. FA9550-16-1-0015.

References

1. Feng, C., Wang, P.: The existence of almost periodic solutions of some delay differential equations. Comput. Math. Appl. **47**(8–9), 1225–1231 (2004)
2. Angeli, D.: A Lyapunov approach to incremental stability properties. IEEE Trans. Autom. Control **47**(3), 410–421 (2002)
3. Zamani, M., Tabuada, P.: Backstepping design for incremental stability. IEEE Trans. Autom. Control **56**(9), 2184–2189 (2011)
4. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid Dynamical Systems: Modeling, Stability, and Robustness. Princeton University Press, New Jersey (2012)
5. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid dynamical systems. IEEE Control Syst. Mag. **29**(2), 28–93 (2009)
6. Prabhakar, P., Liu, J., Murray, R.M.: Pre-orders for reasoning about stability properties with respect to input of hybrid systems. In: In Proceedings of the International Conference on Embedded Software (EMSOFT), pp. 1–10, Sept 2013
7. Li, Y., Phillips, S., Sanfelice, R.G.: Results on incremental stability for a class of hybrid systems. In: Proceedings of IEEE 53rd Annual Conference on Decision and Control, pp. 3089–3094, Dec 2014
8. Postoyan, R., Biemond, J.J.B., Heemels, W.P.M.H., van De Wouw, N.: Definitions of incremental stability for hybrid systems. In: Proceedings of IEEE 54th Annual Conference on Decision and Control, Dec 2015
9. Lohmiller, W., Slotine, J.-J.: On contraction analysis for nonlinear systems. Automatica **34**(6), 671–682 (1998)
10. Pavlov, A., van de Wouw, N.: Steady-state analysis and regulation of discrete-time nonlinear systems. IEEE Trans. Autom. Control **57**(7), 1793–1798 (2012)

11. Biemond, J.J.B., van de Wouw, N., Heemels, W.P.M.H., Nijmeijer, H.: Tracking control for hybrid systems with state-triggered jumps. IEEE Trans. Autom. Control **58**(4), 876–890 (2013)
12. Sanfelice, R.G., Goebel, R., Teel, A.R.: Invariance principles for hybrid systems with connections to detectability and asymptotic stability. IEEE Trans. Autom. Control **52**(12), 2282–2297 (2007)
13. Lewis, D.C.: Metric properties of differential equations. Am. J. Math. **71**, 294–312 (1949)
14. Isac, G., Németh, S.Z.: Scalar and Asymptotic Scalar Derivatives: Theory and Applications, 4th edn. Springer (2008)
15. Sanfelice, R.G., Praly, L.: Convergence of nonlinear observers on $\mathbb{R}^n$ with a Riemannian metric (Part I). IEEE Trans. Autom. Control **57**(7), 1709–1722 (2012)
16. Demidovich, B.P.: Dissipativity of a nonlinear system of differential equations. Ser. Mat.; Mekh. Part I.6 (1961); Part II.1, 3–8 (1962) (in Russian), pp. 19–27
17. Ruffer, B.S., van de Wouw, N., Mueller, M.: Convergent systems vs. incremental stability. Syst. Control Lett. **62**(3), 277–285 (2013)
18. Li, Y., Phillips, S., Sanfelice, R.G.: Basic properties and characterizations of incremental stability prioritizing flow time for a class of hybrid systems. Syst. Control Lett. **90**, 7–15 (2016)
19. Sanfelice, R.G., Biemond, J.J.B., van de Wouw, N., Maurice, W.P., Heemels, H.: An embedding approach for the design of state-feedback tracking controllers for references with jumps. Int. J. Robust Nonlinear Control **24**, 1585–1608 (2014)
20. Galeani, S., Menini, L., Potini, A.: Robust trajectory tracking for a class of hybrid systems: an internal model principle approach. IEEE Trans. Autom. Control **57**(2), 344–359 (2012)

Part III
New Perspectives

Chapter 10
Exponential Stability of Semi-linear One-Dimensional Balance Laws

Georges Bastin and Jean-Michel Coron

Abstract Raman amplifiers and plug flow chemical reactors are typical examples of engineering systems that are conveniently represented by *semi-linear one-dimensional systems of balance laws*. The main goal of this chapter is to explain how a quadratic Lyapunov function can be used to prove the exponential stability of the steady state for this class of hyperbolic systems.

10.1 Introduction

The Lyapunov method is a well-established tool in stability analysis of dynamical systems. The principal merit of the method is that the actual solution (whether analytical or numerical) of the concerned system is not required. Meanwhile, the main drawback is that no systematic procedure exists for deriving Lyapunov functions and Laurent Praly is definitely one of the scientists who made the greatest contributions to their construction (see e.g., [3, 9–11, 14]). In this chapter, we bring a modest additional stone to this building. The main goal is to explain how a quadratic Lyapunov function can be used to prove the exponential stability of the steady state of *semi-linear one-dimensional hyperbolic systems of balance laws*. As a motivation, in the next section, we present some interesting physical examples of such systems.

G. Bastin
Université catholique de Louvain, ICTEAM, Avenue G. Lemaitre 4,
1348 Louvain-la-Neuve, Belgium
e-mail: georges.bastin@uclouvain.be

J.-M. Coron (✉)
Laboratoire Jacques-Louis Lions, Sorbonne Université, UPMC Univ Paris 06,
UMR 7598, Place Jussieu 4, 75252 Paris, France
e-mail: jean-michel.coron@upmc.fr

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_10

10.1.1 Raman Amplifiers

Raman amplifiers are electro-optical devices that are used for compensating the natural power attenuation of laser signals transmitted along optical fibers in long distance communications. Their operation is based on the *Raman effect* which was discovered by [12]. The simplest implementation of Raman amplification in optical telecommunications is depicted in Fig. 10.1. The transmitted information is encoded by amplitude modulation of a laser signal with wavelength ω_s. The signal is provided by an optical source at the channel input and received by a photo-detector at the output. A pump laser beam with wavelength ω_p is injected backward in the optical fiber. If the wavelengths are appropriately selected, the energy of the pump is transferred to the signal and produces an amplification that counteracts the natural attenuation. The dynamics of the signal and pump powers along the fiber are represented by the following system of two balance laws [4]:

$$\begin{aligned} &\partial_t S + \lambda_s \Big(\partial_x S + \alpha_s S - \beta_s SP \Big) = 0, \\ & && t \in [0, +\infty), \quad x \in [0, L], \qquad (10.1) \\ &\partial_t P - \lambda_p \Big(\partial_x P - \alpha_p P - \beta_p PS \Big) = 0, \end{aligned}$$

where $S(t, x)$ is the power of the transmitted signal, $P(t, x)$ is the power of the pump laser beam, λ_s and λ_p are the propagation group velocities of the signal and pump waves respectively, α_s and α_p are the attenuation coefficients per unit length, β_s and β_p are the amplification gains per unit length. All these positive constant parameters α_s and α_p, β_s and β_p, λ_s and λ_p are characteristic of the fiber material and dependent of the wavelengths ω_s and ω_p.

As the input signal power and the launch pump power can be exogenously imposed, the boundary conditions are

$$S(t, 0) = U_0, P(t, L) = U_L, \qquad (10.2)$$

with constant inputs U_0 and U_L.

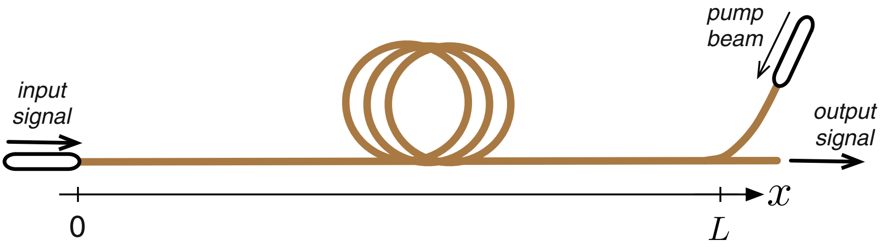


Fig. 10.1 Optical communication with Raman amplification

10.1.2 Plug Flow Chemical Reactors

A plug flow chemical reactor (PFR) is a tubular reactor where a liquid reaction mixture circulates. The reaction proceeds as the reactants travel through the reactor. Here, we consider the case of an horizontal PFR where a simple monomolecular reaction takes place

$$A \rightleftarrows B.$$

A is the reactant species and B is the desired product. The reaction is supposed to be exothermic and a jacket is used to cool the reactor. The cooling fluid flows around the wall of the tubular reactor. The dynamics of the PFR are described by the following system of balance laws:

$$\begin{aligned} &\partial_t T_c - V_c \partial_x T_c - k_o(T_c - T_r) = 0, \\ &\partial_t T_r + V_r \partial_x T_r + k_o(T_c - T_r) - k_1 r(T_r, C_A, C_B) = 0, \\ &\partial_t C_A + V_r \partial_x C_A + r(T_r, C_A, C_B) = 0, \\ &\partial_t C_B + V_r \partial_x C_B - r(T_r, C_A, C_B) = 0, \end{aligned} \tag{10.3}$$

where $t \in [0, +\infty)$, $x \in [0, L]$, $T_c(t, x)$ is the coolant temperature, $T_r(t, x)$ is the reactor temperature. The variables $C_A(t, x)$ and $C_B(t, x)$ denote the concentrations of the chemicals in the reaction medium. V_c is the constant coolant velocity in the jacket, V_r is the constant reactive fluid velocity in the reactor. The function $r(T_r, C_A, C_B)$ represents the reaction rate. A typical form of this function is

$$r(T_r, C_A, C_B) = (aC_A - bC_B) \exp\Big(-\frac{E}{RT_r}\Big),$$

where a and b are rate constants, E is the activation energy and R is the Boltzmann constant.

The system is subject to the following constant boundary conditions:

$$T_r(t, 0) = T_r^{\text{in}}, \quad C_A(t, 0) = C_A^{\text{in}}, \quad C_B(t, 0) = 0, \quad T_c(t, 0) = T_c^{\text{in}}. \tag{10.4}$$

10.1.3 Chemotaxis

Chemotaxis refers to the motion of certain living microorganisms (bacteria, slime molds, leukocytes ...) in response to the concentrations of chemicals. A simple model for one-dimensional chemotaxis, known as the Kac-Goldstein model, has been proposed in [5] in order to explain the spatial pattern formations in chemosensitive populations. Revisited in [6], this model, in its simplest form, is a system of two balance laws of the form

$$\begin{aligned} &\partial_t \varrho^+ + \gamma \partial_x \varrho^+ + \phi(\varrho^+, \varrho^-)(\varrho^- - \varrho^+) = 0, \\ &\partial_t \varrho^- - \gamma \partial_x \varrho^- + \phi(\varrho^+, \varrho^-)(\varrho^+ - \varrho^-) = 0, \end{aligned} \quad t \in [0, +\infty), \quad x \in [0, L], \tag{10.5}$$

where ϱ^+ denotes the density of right-moving cells and ϱ^- the density of left-moving cells. The function $\phi(\varrho^+, \varrho^-)$ is called the "turning function". The constant parameter γ is the velocity of the cell motion. With the change of coordinates $\varrho \triangleq \varrho^+ + \varrho^-$, $q \triangleq \gamma(\varrho^+ - \varrho^-)$, we have the following alternative equivalent model:

$$\begin{aligned} &\partial_t \varrho + \partial_x q = 0, \\ &\partial_t q + \gamma^2 \partial_x \varrho - 2\phi\big(\frac{\varrho}{2} + \frac{q}{2\gamma}, \frac{\varrho}{2} - \frac{q}{2\gamma}\big) q = 0, \end{aligned}$$

where ϱ is the total density and q is a flux proportional to the difference of densities of right and left-moving cells. Remark that we have $q = \varrho V$ where

$$V \triangleq \gamma \frac{\varrho^+ - \varrho^-}{\varrho^+ + \varrho^-}$$

can be interpreted as the average group velocity of the moving cells.

Various possible turning functions are reviewed in [8]. A typical example is

$$\phi(\varrho^+, \varrho^-) = \alpha \varrho^+ \varrho^- - \mu,$$

where α and μ are positive constants.

A special case of interest (see, e.g., [7]) is when the cells are confined in the domain $[0, L]$. This situation may be represented by "no-flow boundary conditions" of the form

$$\begin{aligned} q(t, 0) = \gamma\big(\varrho^+(t, 0) - \varrho^-(t, 0)\big) = 0, \\ q(t, L) = \gamma\big(\varrho^+(t, L) - \varrho^-(t, L)\big) = 0. \end{aligned} \tag{10.6}$$

10.2 Exponential Stability of Semi-linear Hyperbolic Systems of Balance Laws

The examples given above are special cases of the general semi-linear hyperbolic system

$$\mathbf{Y}_t + \Lambda \mathbf{Y}_x + G(\mathbf{Y}) = \mathbf{0}, \quad t \in [0, +\infty), \quad x \in [0, L], \tag{10.7}$$

$$\mathcal{B}\big(\mathbf{Y}(t, 0), \mathbf{Y}(t, L)\big) = \mathbf{0}, \quad t \in [0, +\infty), \tag{10.8}$$

where

- t and x are the two independent variables: a time variable $t \in [0, +\infty)$ and a space variable $x \in [0, L]$ over a finite interval;

- $\mathbf{Y} : [0, +\infty) \times [0, L] \to \mathcal{Y}$ is the vector of state variables, with $\mathcal{Y}$ a nonempty connected open subset of $\mathbb{R}^n$;

- $\varLambda \in \mathcal{M}_{n,n}(\mathbb{R})$ is the diagonal matrix defined as

$$\varLambda \triangleq \begin{pmatrix} \varLambda^+ & 0 \\ 0 & -\varLambda^- \end{pmatrix} \quad \text{with} \quad \begin{cases} \varLambda^+ = \operatorname{diag}\left\{\lambda_1, \dots, \lambda_m\right\}, \\ \varLambda^- = \operatorname{diag}\left\{\lambda_{m+1}, \dots, \lambda_n\right\}, \end{cases} \tag{10.9}$$

  where $m \in [0, n]$ and $\lambda_i > 0 \ \forall i$;

- $G \in C^2(\mathcal{Y}, \mathbb{R}^n)$ is the vector of *source* terms;

- $\mathcal{B} \in C^2(\mathcal{Y} \times \mathcal{Y}, \mathbb{R}^n)$ is the vector of boundary conditions.

A steady state $\mathbf{Y}^*(x)$ is a solution of the ordinary differential equation $\varLambda \mathbf{Y}^*_x(x) + G(\mathbf{Y}^*(x)) = \mathbf{0}$ satisfying the boundary condition $\mathcal{B}\big(\mathbf{Y}^*(0), \mathbf{Y}^*(L)\big) = \mathbf{0}$.

We define the following change of coordinates:

$$\mathbf{Z}(t, x) \triangleq \mathbf{Y}(t, x) - \mathbf{Y}^*(x), \qquad \mathbf{Z} = (Z_1, \dots, Z_n)^{\mathsf{T}}.$$

In the $\mathbf{Z}$ coordinates, the system (10.7), (10.8) is rewritten

$$\mathbf{Z}_t + \varLambda \mathbf{Z}_x + B(\mathbf{Z}, x) = \mathbf{0}, \tag{10.10}$$

$$\mathcal{B}\big(\mathbf{Z}(t, 0) + \mathbf{Y}^*(0), \mathbf{Z}(t, L) + \mathbf{Y}^*(L)\big) = \mathbf{0}, \tag{10.11}$$

where

$$B(\mathbf{Z}, x) \triangleq \Big[G(\mathbf{Z} + \mathbf{Y}^*(x)) - G(\mathbf{Y}^*(x))\Big].$$

Since $B(\mathbf{0}, x) = \mathbf{0}$ by definition of the steady state, it follows that there exists a matrix $M(\mathbf{Z}, x) \in \mathcal{M}_{n \times n}(\mathbb{R})$ such that (10.10) may be rewritten as

$$\mathbf{Z}_t + \varLambda \mathbf{Z}_x + M(\mathbf{Z}, x)\mathbf{Z} = \mathbf{0}, \tag{10.12}$$

with

$$M(\mathbf{0}, x) = \frac{\partial B}{\partial \mathbf{Z}}(\mathbf{0}, x).$$

In order to have a well-posed Cauchy problem, a basic requirement is that "at each boundary point the incoming information $\mathbf{Z}_{\text{in}}$ is determined by the outgoing information $\mathbf{Z}_{\text{out}}$" [13, Sect. 3], with the definitions

$$\mathbf{Z}_{\text{in}}(t) \triangleq \begin{pmatrix} \mathbf{Z}^{+}(t,0) \\ \mathbf{Z}^{-}(t,L) \end{pmatrix} \quad \text{and} \quad \mathbf{Z}_{\text{out}}(t) \triangleq \begin{pmatrix} \mathbf{Z}^{+}(t,L) \\ \mathbf{Z}^{-}(t,0) \end{pmatrix}, \tag{10.13}$$

where $\mathbf{Z}^{+}$ and $\mathbf{Z}^{-}$ are defined as follows:

$$\mathbf{Z}^{+} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}, \qquad \mathbf{Z}^{-} = \begin{pmatrix} Z_{m+1} \\ \vdots \\ Z_n \end{pmatrix}.$$

This means that the system (10.12) is subject to boundary conditions having the form

$$\mathbf{Z}_{\text{in}}(t) = \mathcal{H}\big(\mathbf{Z}_{\text{out}}(t)\big), \tag{10.14}$$

where the map $\mathcal{H} \in C^1(\mathbb{R}^n; \mathbb{R}^n)$.

Our concern is to analyze the exponential stability of the steady state $\mathbf{Z}(t,x) \equiv \mathbf{0}$ of the system (10.12) under the boundary condition (10.14) and under an initial condition

$$\mathbf{Z}(0,x) = \mathbf{Z}_{\text{o}}(x), \quad x \in [0,L]. \tag{10.15}$$

which satisfies the compatibility condition

$$\begin{pmatrix} \mathbf{Z}_{\text{o}}^{+}(0) \\ \mathbf{Z}_{\text{o}}^{-}(L) \end{pmatrix} = \mathcal{H}\begin{pmatrix} \mathbf{Z}_{\text{o}}^{+}(L) \\ \mathbf{Z}_{\text{o}}^{-}(0) \end{pmatrix}. \tag{10.16}$$

Let us first recall the following theorem on the well-posedness of the Cauchy problem (10.12), (10.14), (10.15).

Theorem 10.1 *There exists $\delta_0 > 0$ such that, for every $\mathbf{Z}_o \in H^1((0,L); \mathbb{R}^n)$ satisfying*

$$\|\mathbf{Z}_o\|_{H^1((0,L);\mathbb{R}^n)} \leqslant \delta_0$$

and the compatibility condition (10.16), the Cauchy problem (10.12), (10.14), (10.15) has a unique maximal classical solution

$$\mathbf{Z} \in C^0([0,T), H^1((0,L); \mathbb{R}^n)) \tag{10.17}$$

with $T \in (0, +\infty]$.

Moreover, if

$$\|\mathbf{Z}(t,\cdot)\|_{H^1((0,L);\mathbb{R}^n)} \leqslant \delta_0, \ \forall t \in [0,T),$$

then $T = +\infty$.

A proof of this theorem is easily adapted from [1, Appendix B] by considering the special case of a constant matrix Λ which allows to replace $H^2((0,L); \mathbb{R}^n)$ by $H^1((0,L); \mathbb{R}^n)$.

The definition of the exponential stability is as follows.

Definition 10.1 The steady state $\mathbf{Z}(t,x) \equiv 0$ of the system (10.12), (10.14) is exponentially stable for the H^1-norm if there exist $\delta > 0$, $\nu > 0$ and $C > 0$ such that, for every $\mathbf{Z}_{\text{o}} \in H^1((0,L);\mathbb{R}^n)$ satisfying $\|\mathbf{Z}_{\text{o}}\|_{H^1((0,L);\mathbb{R}^n)} \leqslant \delta$ and the compatibility conditions (10.16), the solution $\mathbf{Z}$ of the Cauchy problem (10.12), (10.14), (10.15) is defined on $[0,+\infty) \times [0,L]$ and satisfies

$$\|\mathbf{Z}(t,.)\|_{H^1((0,L);\mathbb{R}^n)} \leq Ce^{-\nu t}\|\mathbf{Z}_{\text{o}}\|_{H^1((0,L);\mathbb{R}^n)}, \quad \forall t \in [0,+\infty). \tag{10.18}$$

Let us now define the matrix $\mathbf{K}$ as the linearization of the map $\mathcal{H}$ at the steady state

$$\mathbf{K} \triangleq \mathcal{H}'(\mathbf{0}).$$

We then have the following stability theorem.

Theorem 10.2 *The steady state $\mathbf{Z}(t,x) \equiv \mathbf{0}$ of the system (10.12), (10.14) is exponentially stable for the H^1-norm if there exists a map Q satisfying*

$$Q(x) \triangleq \mathbf{diag}\{Q^+(x), Q^-(x)\},$$
$$Q^+(x) \triangleq \mathbf{diag}\{q_1(x), \ldots, q_m(x)\}, \quad Q^-(x) \triangleq \mathbf{diag}\{q_{m+1}(x), \ldots, q_n(x)\},$$
$$q_i \in C^1([0,L];\mathbb{R}_+) \ \forall i.$$

such that the following Matrix Inequalities hold:

(i) *the matrix*

$$\begin{pmatrix} Q^+(L)\Lambda^+ & 0 \\ 0 & Q^-(0)\Lambda^- \end{pmatrix} - \mathbf{K}^{\mathsf{T}} \begin{pmatrix} Q^+(0)\Lambda^+ & 0 \\ 0 & Q^-(L)\Lambda^- \end{pmatrix} \mathbf{K} \tag{10.19}$$

is positive semi-definite;

(ii) *the matrix*

$$-Q'(x)\Lambda + Q(x)M(\mathbf{0},x) + M^{\mathsf{T}}(\mathbf{0},x)Q(x)$$

is positive definite $\forall x \in [0,L]$.

10.3 Proof in the Case Where $m = n$

For the clarity of the demonstration, we shall first prove the theorem in the special case where $m = n$, which means that the matrix Λ is the positive diagonal matrix $\text{diag}\{\lambda_1, \ldots, \lambda_n\}$ with $\lambda_i > 0 \ \ \forall i = 1, \ldots, n$. In that case, the boundary condition (10.14) and the compatibility conditions (10.16) are simply rewritten

$$\mathbf{Z}(t,0) = \mathcal{H}\Big(\mathbf{Z}(t,L)\Big), \tag{10.20}$$

$$\mathbf{Z}_{\mathrm{o}}(0) = \mathcal{H}\Big(\mathbf{Z}_{\mathrm{o}}(L)\Big). \tag{10.21}$$

Moreover, condition (i) of Theorem 10.2 is restated as

(i-bis) *the matrix* $Q(L)\Lambda - \mathbf{K}^{\mathsf{T}}Q(0)\Lambda\mathbf{K}$ *is positive semi-definite.*

For the stability analysis, we adopt the H^1 Lyapunov function candidate

$$\mathbf{V} \triangleq \mathbf{V}_1 + \mathbf{V}_2 \tag{10.22}$$

such that

$$\mathbf{V}_1 = \int_0^L \mathbf{Z}^{\mathsf{T}}Q(x)\mathbf{Z}\,dx, \tag{10.23}$$

$$\mathbf{V}_2 = \int_0^L \mathbf{Z}_t^{\mathsf{T}}Q(x)\mathbf{Z}_t\,dx, \tag{10.24}$$

where, by definition, the notation $\mathbf{Z}_t$ must be understood as

$$\mathbf{Z}_t \triangleq -\Lambda\mathbf{Z}_x - B(\mathbf{Z},x).$$

Let us remark that by (10.17) $\mathbf{V}$ is a continuous function of t. In order to prove Theorem 10.2, we temporarily assume that $\mathbf{Z}$ is of class C^2 on $[0,T]\times[0,L]$ and therefore that $\mathbf{V}$ is of class C^1 in $[0,T]$. Under this assumption (that will be relaxed later on) the first step of the proof is to compute the following estimates of $d\mathbf{V}_1/dt$ and $d\mathbf{V}_2/dt$.

Estimate of $d\mathbf{V}_1/dt$

The time derivative of $\mathbf{V}_1$ along the solutions of (10.12), (10.20) is[1]

$$\begin{aligned}\frac{d\mathbf{V}_1}{dt} &= \int_0^L 2\mathbf{Z}^{\mathsf{T}}Q(x)\mathbf{Z}_t dx\\ &= \int_0^L 2\mathbf{Z}^{\mathsf{T}}Q(x)\Big(-\Lambda\mathbf{Z}_x - B(\mathbf{Z},x)\Big)dx.\end{aligned}$$

Then, using integrations by parts, we get

$$\frac{d\mathbf{V}_1}{dt} = \mathcal{T}_{11} + \mathcal{T}_{12}, \tag{10.25}$$

with

[1] The notation M^{T} denotes the transpose of the matrix M.

$$\mathcal{T}_{11} \triangleq \Big[- \mathbf{Z}^\mathsf{T} Q(x) \Lambda \mathbf{Z} \Big]_0^L, \tag{10.26}$$

$$\mathcal{T}_{12} \triangleq \int_0^L -\mathbf{Z}^\mathsf{T} Q'(x) \Lambda \mathbf{Z} - 2\mathbf{Z}^\mathsf{T} Q(x) B(\mathbf{Z}, x) dx. \tag{10.27}$$

From (10.26), we have

$$\mathcal{T}_{11} = -\mathbf{Z}^\mathsf{T}(t, L) Q(L) \Lambda \mathbf{Z}(t, L) + \mathbf{Z}^\mathsf{T}(t, 0) Q(0) \Lambda \mathbf{Z}(t, 0). \tag{10.28}$$

Let us introduce a notation in order to deal with estimates on “higher order terms”. We denote by $\mathcal{O}(X; Y)$, with $X \geqslant 0$ and $Y \geqslant 0$, quantities for which there exist $C > 0$ and $\varepsilon > 0$, independent of $\mathbf{Z}$ and $\mathbf{Z}_t$, such that

$$(Y \leqslant \varepsilon) \Rightarrow (|\mathcal{O}(X; Y)| \leqslant CX).$$

Then from (10.28), using the boundary condition (10.20), we have

$$\mathcal{T}_{11} = -\mathbf{Z}^\mathsf{T}(t, L) \Big[Q(L) \Lambda - \mathbf{K}^\mathsf{T} Q(0) \Lambda \mathbf{K} \Big] \mathbf{Z}(t, L) + \mathcal{O}(|\mathbf{Z}(t, L)|^3; |\mathbf{Z}(t, L)|), \tag{10.29}$$

and from (10.27) we have

$$\begin{aligned} \mathcal{T}_{12} = &- \int_0^L \mathbf{Z}^\mathsf{T} \Big[- Q'(x) \Lambda + M^\mathsf{T}(\mathbf{0}, x) Q(x) + Q(x) M(\mathbf{0}, x) \Big] \mathbf{Z}\, dx \\ &+ \mathcal{O}\Big(\int_0^L |\mathbf{Z}|^3 dx; |\mathbf{Z}(t, .)|_0 \Big), \end{aligned} \tag{10.30}$$

where, for $f \in C^0([0, L]; \mathbb{R}^n)$, we denote $|f|_0 = \max\{|f(x)|;\ x \in [0, L]\}$.

Estimate of $d\mathbf{V}_2/dt$

By time differentiation of the system equations (10.12), (10.20), $\mathbf{Z}_t$ can be shown to satisfy the following hyperbolic dynamics:

$$\mathbf{Z}_{tt} + \Lambda \mathbf{Z}_{tx} + \frac{\partial B}{\partial \mathbf{Z}}(\mathbf{Z}, x) \mathbf{Z}_t = \mathbf{0}, \tag{10.31}$$

$$\mathbf{Z}_t(t, 0) = \mathcal{H}'(\mathbf{Z}(t, L)) \mathbf{Z}_t(t, L). \tag{10.32}$$

The time derivative of $\mathbf{V}_2$ along the solutions of (10.12), (10.20), (10.31), (10.32) is

$$\begin{aligned} \frac{d\mathbf{V}_2}{dt} &= \int_0^L 2\mathbf{Z}_t^\mathsf{T} Q(x) (\mathbf{Z}_t)_t dx \\ &= \int_0^L 2\mathbf{Z}_t^\mathsf{T} Q(x) \Big(- \Lambda \mathbf{Z}_{tx} - \frac{\partial B}{\partial \mathbf{Z}}(\mathbf{Z}, x) \mathbf{Z}_t \Big) dx. \end{aligned}$$

Then, using integrations by parts, we get

$$\frac{d\mathbf{V}_2}{dt} = \mathcal{T}_{21} + \mathcal{T}_{22}, \tag{10.33}$$

with

$$\mathcal{T}_{21} \triangleq \Big[- \mathbf{Z}_t^\mathsf{T} Q(x) \Lambda \mathbf{Z}_t \Big]_0^L, \tag{10.34}$$

$$\mathcal{T}_{22} \triangleq \int_0^L \mathbf{Z}_t^\mathsf{T} Q'(x) \Lambda \mathbf{Z}_t + 2\mathbf{Z}_t^\mathsf{T} Q(x) \Big(\frac{\partial B}{\partial \mathbf{Z}}(\mathbf{Z}, x) \mathbf{Z}_t \Big) dx. \tag{10.35}$$

From (10.34), we have

$$\mathcal{T}_{21} = -\mathbf{Z}_t^\mathsf{T}(t,L) Q(L) \Lambda \mathbf{Z}_t(t,L) + \mathbf{Z}_t^\mathsf{T}(t,0) Q(0) \Lambda \mathbf{Z}_t(t,0). \tag{10.36}$$

Then, using the boundary condition (10.32), we get

$$\begin{aligned} \mathcal{T}_{21} = &-\mathbf{Z}_t^\mathsf{T}(t,L) \Big[Q(L)\Lambda - \mathbf{K}^\mathsf{T} Q(0) \Lambda \mathbf{K} \Big] \mathbf{Z}_t(t,L) \\ &+ \mathcal{O}(|\mathbf{Z}_t(t,L)|^2 |\mathbf{Z}(t,L)|; |\mathbf{Z}(t,L)|). \end{aligned} \tag{10.37}$$

Moreover $\mathcal{T}_{22}$ is written

$$\begin{aligned} \mathcal{T}_{22} = &- \int_0^L \mathbf{Z}_t^\mathsf{T} \Big[- Q'(x)\Lambda + M^\mathsf{T}(\mathbf{0}, x) Q(x) + Q(x) M(\mathbf{0}, x) \Big] \mathbf{Z}_t \, dx \\ &+ \mathcal{O}\Big(\int_0^L |\mathbf{Z}_t|^2 |\mathbf{Z}| dx; |\mathbf{Z}(t,.)|_0 \Big). \end{aligned} \tag{10.38}$$

In the next lemma, we shall now use these estimates to show that the Lyapunov function exponentially decreases along the system trajectories.

Lemma 10.1 *There exist positive real constants α, β and δ such that, for every $\mathbf{Z}$ such that $|\mathbf{Z}|_0 \leq \delta$, we have*

$$\frac{1}{\beta} \int_0^L (|\mathbf{Z}|^2 + |\mathbf{Z}_x|^2) dx \leqslant \mathbf{V} \leqslant \beta \int_0^L (|\mathbf{Z}|^2 + |\mathbf{Z}_x|^2) dx, \tag{10.39}$$

$$\frac{d\mathbf{V}}{dt} \leq -\alpha \mathbf{V}. \tag{10.40}$$

Proof Inequalities (10.39) follow directly from the definition of $\mathbf{V}$ and straightforward estimations.

Let us introduce the following compact matrix notations:

$$\mathcal{K} \triangleq Q(L)\Lambda - \mathbf{K}^\mathsf{T} Q(0) \Lambda \mathbf{K}, \tag{10.41}$$

$$\mathcal{L}(x) \triangleq -Q'(x)\Lambda + M^\mathsf{T}(\mathbf{0}, x) Q(x) + Q(x) M(\mathbf{0}, x). \tag{10.42}$$

Then it follows from (10.28), (10.30), (10.37), (10.38) that

$$\begin{aligned}\frac{d\mathbf{V}}{dt} = -\mathbf{Z}^{\mathsf{T}}(t,L)\mathcal{K}\,\mathbf{Z}(t,L) - \mathbf{Z}_t^{\mathsf{T}}(t,L)\mathcal{K}\,\mathbf{Z}_t(t,L) \\ + \mathcal{O}(|\mathbf{Z}(t,L)|(|\mathbf{Z}(t,L)|^2 + |\mathbf{Z}_t(t,L)|^2); |\mathbf{Z}(t,L)|) \\ - \int_0^L \Big(\mathbf{Z}^{\mathsf{T}}\mathcal{L}(x)\,\mathbf{Z} + \mathbf{Z}_t^{\mathsf{T}}\mathcal{L}(x)\,\mathbf{Z}_t\Big)dx \\ + \mathcal{O}(\int_0^L \Big((|\mathbf{Z}|^2| + |\mathbf{Z}_t|^2)|\mathbf{Z}|\Big)dx; |\mathbf{Z}(t,.)|_0\big).\end{aligned} \tag{10.43}$$

Then, by assumption **(i-bis)** of Theorem 10.2 and from (10.41), there exists $\delta_1 > 0$ such that if $|\mathbf{Z}(t,L)| < \delta_1$ then

$$\begin{aligned}-\mathbf{Z}^{\mathsf{T}}(t,L)\mathcal{K}\,\mathbf{Z}(t,L) - \mathbf{Z}_t^{\mathsf{T}}(t,L)\mathcal{K}\,\mathbf{Z}_t(t,L) \\ + \mathcal{O}(|\mathbf{Z}(t,L)|(|\mathbf{Z}(t,L)|^2 + |\mathbf{Z}_t(t,L)|^2); |\mathbf{Z}(t,L)|) \leqslant 0.\end{aligned} \tag{10.44}$$

Let us recall the following Sobolev inequality, see, e.g., [2]: for a function $\varphi \in C^1([0,L];\mathbb{R}^n)$, there exists $C_1 > 0$ such that

$$|\varphi|_0 \leqslant C_1 \int_0^L (|\varphi(x)|^2 + |\varphi'(x)|^2)dx. \tag{10.45}$$

Moreover, from (10.10) and (10.31), we know also that there exist $\delta_2 > 0$ and $C_2 > 0$ such that, if $|\mathbf{Z}(t,x)| + |\mathbf{Z}_t(t,x)| < \delta_2$, then

$$|\mathbf{Z}_t(t,x)| \leqslant C_2\big(|\mathbf{Z}(t,x)| + |\mathbf{Z}_x(t,x)|\big), \tag{10.46}$$

$$|\mathbf{Z}_x(t,x)| \leqslant C_2\big(|\mathbf{Z}(t,x)| + |\mathbf{Z}_t(t,x)|\big). \tag{10.47}$$

Using repeatedly, inequalities (10.45) to (10.47), it follows that there exists $\delta_3 > 0$ and $C_3 > 0$ such that, if $|\mathbf{Z}(t,.)|_0 < \delta_3$, then

$$\mathcal{O}(\int_0^L \Big((|\mathbf{Z}|^2| + |\mathbf{Z}_t|^2)|\mathbf{Z}|\Big)dx; |\mathbf{Z}(t,.)|_0\big) \leqslant C_3|\mathbf{Z}(t,.)|_0\mathbf{V}. \tag{10.48}$$

Using assumption (ii) of Theorem 10.2, there exists $\gamma > 0$ such that

$$-\int_0^L \Big(\mathbf{Z}^{\mathsf{T}}\mathcal{L}(x)\,\mathbf{Z} + \mathbf{Z}_t^{\mathsf{T}}\mathcal{L}(x)\,\mathbf{Z}_t\Big)dx \leqslant -2\gamma\mathbf{V}. \tag{10.49}$$

Finally it follows from (10.43), (10.44), (10.48) and (10.49) that, if $\delta < \min(\delta_1, \delta_3)$ is taken sufficiently small, then $\alpha > 0$ can be selected such that

$$\frac{d\mathbf{V}}{dt} = (-2\gamma + C_3|\mathbf{Z}(t,.)|_0)\mathbf{V} \leqslant -\alpha\mathbf{V},$$

for every $\mathbf{Z}(t,.)$ such that $|\mathbf{Z}(t,.)|_0 \leq \delta$. This concludes the proof of Lemma 10.1.

In this lemma, the estimates (10.39) and (10.40) were obtained under the assumption that $\mathbf{Z}$ is of class C^2 on $[0,T] \times [0,L]$. But the selection of α and β does not depend on the C^2-norm of $\mathbf{Z}$: they depend only on the $C^0([0,T];H^1((0,L);\mathbb{R}^n))$-norm of $\mathbf{Z}$. Hence, using a classical density argument (see, e.g., [1, Comment 4.6]), the estimates (10.39) and (10.40) remain valid in the distribution sense if $\mathbf{Z}(.,.)$ is only of class C^1.

Let us now introduce

$$\varepsilon \triangleq \min\left\{\frac{\delta}{2C_1\beta}, \frac{\delta_0}{\beta}\right\}. \tag{10.50}$$

Note that $\beta \geqslant 1$ and therefore that $\delta \leqslant \delta_0$. Using Lemma 10.1, (10.45) and (10.50), for every $t \in [0,T]$

$$\left(\|\mathbf{Z}(t,.)\|_{H^1((0,L);\mathbb{R}^n)} \leqslant \varepsilon\right) \Longrightarrow \left(|\mathbf{Z}(t,.)|_0 \leq \frac{\delta}{2} \text{ and } \mathbf{V}(t) \leqslant \beta\varepsilon^2\right), \tag{10.51}$$

$$\left(|\mathbf{Z}(t,.)|_0 \leq \delta \text{ and } \mathbf{V} \leqslant \beta\varepsilon^2\right) \Longrightarrow \left(|\mathbf{Z}(t,.)|_0 \leq \frac{\delta}{2} \text{ and } \|\mathbf{Z}(t,.)\|_{H^1((0,L);\mathbb{R}^n)} \leqslant \delta_0\right), \tag{10.52}$$

$$\left(|\mathbf{Z}(t,.)|_0 \leq \delta\right) \Longrightarrow \left(\frac{d\mathbf{V}}{dt} \leqslant 0\right) \quad \text{in the distribution sense.} \tag{10.53}$$

Let $\mathbf{Z}_o \in H^1((0,L);\mathbb{R}^n)$ satisfy the compatibility condition (10.21) and

$$\|\mathbf{Z}_o\|_{H^1((0,L);\mathbb{R}^n)} < \varepsilon.$$

Let $\mathbf{Z} \in C^0([0,T^*), H^1((0,L);\mathbb{R}^n))$ be the maximal classical solution the Cauchy problem (10.12), (10.14), (10.15). Using implications (10.51) to (10.53) for $T \in [0,T^*)$, we get that

$$|\mathbf{Z}(t,\cdot)|_{H^1((0,L);\mathbb{R}^n)} \leqslant \delta_0, \; \forall t \in [0,T^*), \tag{10.54}$$

$$|\mathbf{Z}(t,\cdot)|_0 + |\mathbf{Z}_t(t,\cdot)|_0 \leqslant \delta, \; \forall t \in [0,T^*). \tag{10.55}$$

Using (10.54) and Theorem 10.1, we have that $T = +\infty$. Using Lemma 10.1 and (10.55), we finally obtain that

$$\|\mathbf{Z}(t,\cdot)\|^2_{H^1((0,L);\mathbb{R}^n)} \leqslant \beta\mathbf{V}(t) \leqslant \beta\mathbf{V}(0)e^{-\alpha t} \leqslant \beta^2\|\mathbf{Z}_{\mathrm{o}}\|^2_{H^1((0,L);\mathbb{R}^n)}e^{-\alpha t}.$$

This concludes the proof of Theorem 10.2.

10.4 Proof in the Case Where 0 < *m* < *n*

In this section, we explain the modifications of the proof that must be used to deal with the case $0 < m < n$. (Of course,the case $m = 0$ is equivalent to the case $m = n$ by considering $\mathbf{Z}(t, L - x)$ instead of $\mathbf{Z}(t, x)$.)

The major difference lies in functions $\mathcal{T}_{11}$ and $\mathcal{T}_{21}$ which are now written as follows:

$$\begin{aligned}\mathcal{T}_{11} = &-\begin{pmatrix}\mathbf{Z}^+(t,L)\\ \mathbf{Z}^-(t,0)\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}Q^+(L)\Lambda^+ & 0\\ 0 & Q^-(0)\Lambda^-\end{pmatrix}\begin{pmatrix}\mathbf{Z}^+(t,L)\\ \mathbf{Z}^-(t,0)\end{pmatrix}\\ &+\begin{pmatrix}\mathbf{Z}^+(t,0)\\ \mathbf{Z}^-(t,L)\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}Q^+(0)\Lambda^+ & 0\\ 0 & Q^-(L)\Lambda^-\end{pmatrix}\begin{pmatrix}\mathbf{Z}^+(t,0)\\ \mathbf{Z}^-(t,L)\end{pmatrix},\end{aligned}$$

$$\begin{aligned}\mathcal{T}_{21} = &-\begin{pmatrix}\mathbf{Z}_t^+(t,L)\\ \mathbf{Z}_t^-(t,0)\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}Q^+(L)\Lambda^+ & 0\\ 0 & Q^-(0)\Lambda^-\end{pmatrix}\begin{pmatrix}\mathbf{Z}_t^+(t,L)\\ \mathbf{Z}_t^-(t,0)\end{pmatrix}\\ &+\begin{pmatrix}\mathbf{Z}_t^+(t,0)\\ \mathbf{Z}_t^-(t,L)\end{pmatrix}^{\mathsf{T}}\begin{pmatrix}Q^+(0)\Lambda^+ & 0\\ 0 & Q^-(L)\Lambda^-\end{pmatrix}\begin{pmatrix}\mathbf{Z}_t^+(t,0)\\ \mathbf{Z}_t^-(t,L)\end{pmatrix}.\end{aligned}$$

Using the boundary condition (10.14) and assumption **(i)** in these equations, it is then a straightforward exercise to verify that Theorem 10.2 can be established for the case $0 < m < n$ in a manner completely parallel to the one we have followed in the case $m = n$.

10.5 Conclusion

The main goal of this chapter was to explain how a quadratic Lyapunov function can be used to prove the exponential stability of the steady state of *semi-linear one-dimensional hyperbolic systems of balance laws*. Further stability results for hyperbolic systems of balance laws can be found in the textbook [1].

Acknowledgements This work was supported by the ERC advanced grant 266907 (CPDENL) of the 7th Research Framework Programme (FP7) and by the Belgian Programme on Interuniversity Attraction Poles (IAP VII/19).

References

1. Bastin, G., Coron, J.M.: Stability and boundary stabilization of 1-D hyperbolic systems. Progress in Nonlinear Differential Equations and their Applications, **88**. Subseries in Control. Birkhäuser/Springer, [Cham], xiv+307 (2016). ISBN: 978-3-319-32060-1; 978-3-319-32062-5
2. Brezis, H.: Analyse fonctionnelle, théorie et applications. Collection Mathématiques appliquées pour la maîtrise. Masson, Paris (1983)
3. Coron, J.M., Praly, L., Teel, A.: Feedback stabilization of nonlinear systems: sufficient conditions and Lyapunov and input-output techniques. In: Trends in control (Rome, 1995), pp. 293–348. Springer, Berlin (1995)
4. Dower, P., Farrel, P.: On linear control of backward pumped Raman amplifiers. In: Proceedings IFAC Symposium on System Identification, pp. 547–552. Newcastle, Australia (2006)
5. Goldstein, S.: On diffusion by discontinuous movements, and the telegraph equation. Q. J. Mech. Appl. Math. **4**, 129–156 (1951)
6. Kac, M.: A stochastic model related to the telegrapher's equation. Rocky Mt. J. Math. **4**, 497–509 (1956)
7. Lutscher, F.: Modeling alignment and movement of animals and cells. J. Math. Biol. **45**, 234–260 (2002)
8. Lutscher, F., Stevens, A.: Emerging patterns in a hyperbolic model for locally interacting cell systems. J. Nonlinear Sci. **12**, 619–640 (2002)
9. Mazenc, F., Praly, L.: Adding integrations, saturated controls, and stabilization for feedforward systems. IEEE Trans. Autom. Control **41**(11), 1559–1578 (1996). doi:10.1109/9.543995
10. Praly, L.: Une introduction à l'utilisation de fonctions de Lyapunov pour la stabilisation et l'atténuation de perturbations., chap. 2 in Commandes nonlinéaires. Lavoisier (2003)
11. Praly, L., Carnevale, D., Astolfi, A.: Dynamic versus static weighting of Lyapunov functions. IEEE Trans. Autom. Control **58**(6), 1557–1561 (2013). doi:10.1109/TAC.2012.2229813
12. Raman, C., Krishnan, K.: A new type of secondary radiation. Nature **121**, 501–502 (1928)
13. Russell, D.: Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions. SIAM Rev. **20**(4), 639–739 (1978)
14. Teel, A.R., Praly, L.: A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions. ESAIM Control Optim. Calc. Var. **5**, 313–367 (electronic) (2000). doi:10.1051/cocv:2000113

Chapter 11
Checkable Conditions for Contraction After Small Transients in Time and Amplitude

Michael Margaliot, Tamir Tuller and Eduardo D. Sontag

Abstract Contraction theory is a powerful tool for proving asymptotic properties of nonlinear dynamical systems including convergence to an attractor and entrainment to a periodic excitation. We consider generalizations of contraction with respect to a norm that allow contraction to take place after small transients in time and/or amplitude. These generalized contractive systems (GCSs) are useful for several reasons. First, we show that there exist simple and checkable conditions guaranteeing that a system is a GCS, and demonstrate their usefulness using several models from systems biology. Second, allowing small transients does not destroy the important asymptotic properties of contractive systems like convergence to a unique equilibrium point, if it exists, and entrainment to a periodic excitation. Third, in some cases as we change the parameters in a contractive system it becomes a GCS just before it looses contractivity with respect to a norm. In this respect, generalized contractivity is the analogue of marginal stability in Lyapunov stability theory.

11.1 Introduction

Differential analysis studies nonlinear dynamical systems based on the time evolution of the distance between trajectories emanating from different initial conditions. A dynamical system is called *contractive* if any two trajectories converge to one

M. Margaliot (✉)
School of Electrical Engineering-Systems and the Sagol School of Neuroscience,
Tel Aviv University, 69978 Tel Aviv, Israel
e-mail: michaelm@eng.tau.ac.il

T. Tuller
Department of Biomedical Engineering and the Sagol School of Neuroscience,
Tel-Aviv University, 69978 Tel-aviv, Israel
e-mail: tamirtul@post.tau.ac.il

E.D. Sontag
Department of Mathematics and the Center for Quantitative Biology,
Rutgers University, Piscataway, NJ 08854, USA
e-mail: eduardo.sontag@gmail.com

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_11

other at an exponential rate. A contractive system has many important properties including convergence to a unique attractor (if it exists), and entrainment to periodic excitations [2, 21, 34]. These properties can be proven even when the equilibrium point or attractor are not known explicitly. Contraction theory found applications in control theory [22], observer design [10], synchronization of coupled oscillators [3, 44], and more. It has also been extended in many directions including the notion of partial contraction [38], analysis of networks of interacting agents using contraction theory [6, 35], and Lyapunov and Lyapunov-Finsler characterizations of incremental stability [4] and contraction [18]. The latter also leads to a LaSalle-type principle for contractive systems [18]. There is also a growing interest in design techniques for controllers that render control systems contractive or incrementally stable; see, e.g. [45] and the references therein, and also the incremental ISS condition in [15].

A contractive system with added diffusion terms or random noise still satisfies certain asymptotic properties [1, 28]. Also, there exist explicit bounds on the deviations between trajectories of the system and those of its discretization [15]. In this respect, contraction is a *robust* property.

Contraction can in general be defined with respect to a norm that depends on time and/or space [21]. However, establishing that a given dynamical systems is contractive with respect to such a norm may be difficult (see, e.g. [8]). There are, however, *easy to check* conditions for establishing contraction with respect to a fixed norm that are based on the corresponding matrix measure.

Since contraction is usually used to prove asymptotic properties, i.e. properties that hold as time goes to infinity, it is natural to consider systems that are *eventually contractive*, i.e. that become contractive after some time $T > 0$. However, finding *checkable* conditions that guarantee this property seems difficult.

In this chapter, we consider three forms of generalized contractive systems (GCSs). These are motivated by requiring contraction, with respect to a fixed norm, to take place after arbitrarily small transients in time and/or amplitude. We give easy to check sufficient conditions for GSC that are based on matrix measures. In some cases as we change the parameters in a contractive system it becomes a GCS just before it looses contractivity. In this respect, a GCS is the analogue of marginal stability in Lyapunov stability theory. We demonstrate the usefulness of these generalizations using examples of systems that are *not* contractive with respect to any norm, yet are GCSs.

The remainder of this chapter is organized as follows. The next section provides a brief review of some ideas from contraction theory. Section 11.3 presents three generalizations of contraction with respect to a fixed norm. Section 11.4 details sufficient conditions for their existence and describes their implications. The proofs of all the results are placed in Sect. 11.5. The GSCs reviewed here were introduced in [42] (see also [24]). Due to space constraints, [24, 42] did not include the proofs of the main results. These are included here, as well as several new results and examples.

11.2 Preliminaries

We begin with a brief review of some ideas from contraction theory. For more details, including the historic development of contraction theory, and the relation to other notions, see e.g. [20, 33, 40].

Consider the time-varying dynamical system

$$\dot{x} = f(t, x), \tag{11.1}$$

with the state x evolving on a positively invariant convex set $\Omega \subseteq \mathbb{R}^n$. We assume that $f(t, x)$ is differentiable with respect to x, and that both $f(t, x)$ and $J(t, x) := \frac{\partial f}{\partial x}(t, x)$ are continuous in (t, x). Let $x(t, t_0, x_0)$ denote the solution of (11.1) at time $t \geq t_0$ with $x(t_0) = x_0$. For the sake of simplicity, we assume from here on that $x(t, t_0, x_0)$ exists and is unique for all $t \geq t_0 \geq 0$ and all $x_0 \in \Omega$.

We say that (11.1) is *contractive* on Ω with respect to a norm $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ if there exists $c > 0$ such that

$$|x(t_2, t_1, a) - x(t_2, t_1, b)| \leq \exp(-(t_2 - t_1)c)|a - b| \tag{11.2}$$

for all $t_2 \geq t_1 \geq 0$ and all $a, b \in \Omega$. This means that any two trajectories contract to one another at an exponential rate. This implies in particular that the initial condition is "quickly forgotten."

Note that Ref. [21] provides a more general definition, where contraction is with respect to a time- and state-dependent metric $M(t, x)$ rather than to a fixed norm (see also [37] for a general treatment of contraction on a Riemannian manifold). Some of the results below may be stated using this more general framework. But, for a given dynamical system finding such a metric may be difficult. Another extension of contraction is incremental stability [4].

Our approach is based on the fact that there exists a simple sufficient condition guaranteeing (11.2), so generalizing (11.2) appropriately leads to *checkable* sufficient conditions for a system to be a GCS. Another advantage of our approach is that a GCS retains the important property of entrainment to periodic signals.

Recall that a vector norm $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ induces a matrix measure $\mu : \mathbb{R}^{n\times n} \to \mathbb{R}$ defined by $\mu(A) := \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon}(||I + \epsilon A|| - 1)$, where $||\cdot|| : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ is the matrix norm induced by $|\cdot|$. It is well known (see, e.g. [34]) that if there exist a vector norm $|\cdot|$ and $c > 0$ such that the induced matrix measure $\mu : \mathbb{R}^{n\times n} \to \mathbb{R}$ satisfies

$$\mu(J(t, x)) \leq -c, \tag{11.3}$$

for all $t \geq 0$ and all $x \in \Omega$ then (11.2) holds. This is in fact a particular case of using a Lyapunov-Finsler function to prove contraction [18].

We list here the matrix measures corresponding to some vector norms (see, e.g. [43, Chap. 3]). The matrix measure induced by the L_1 vector norm is

$$\mu_1(A) = \max\{c_1(A), \dots, c_n(A)\}, \tag{11.4}$$

where

$$c_j(A) := A_{jj} + \sum_{\substack{1 \le i \le n \\ i \ne j}} |A_{ij}|, \tag{11.5}$$

i.e., the sum of the entries in column j of A, with non diagonal elements replaced by their absolute values. The matrix measure induced by the L_∞ norm is

$$\mu_\infty(A) = \max\{d_1(A), \dots, d_n(A)\}, \tag{11.6}$$

where

$$d_j(A) := A_{jj} + \sum_{\substack{1 \le i \le n \\ i \ne j}} |A_{ji}|, \tag{11.7}$$

i.e., the sum of the entries in row j of A, with non diagonal elements replaced by their absolute values.

Often it is useful to work with scaled norms. Let $|\cdot|_*$ be some vector norm, and let $\mu_* : \mathbb{R}^{n\times n} \to \mathbb{R}$ denote its induced matrix measure. If $P \in \mathbb{R}^{n\times n}$ is an invertible matrix, and $|\cdot|_{*,P} : \mathbb{R}^n \to \mathbb{R}_+$ is the vector norm defined by $|z|_{*,P} := |Pz|_*$ then the induced matrix measure is $\mu_{*,P}(A) = \mu_*(PAP^{-1})$.

One important implication of contraction is *entrainment* to a periodic excitation. Recall that $f : \mathbb{R}_+ \times \Omega \to \mathbb{R}^n$ is called *T-periodic* if

$$f(t,x) = f(t+T,x)$$

for all $t \ge 0$ and all $x \in \Omega$. Note that for the system $\dot{x}(t) = f(u(t), x(t))$, with u an input (or excitation) function, f will be T periodic if u is a T-periodic function. It is well known [21, 34] that if (11.1) is contractive and f is T-periodic then for any $t_1 \ge 0$ there exists a unique periodic solution $\alpha : [t_1, \infty) \to \Omega$ of (11.1), of period T, and every trajectory converges to α. Entrainment is important in various applications ranging from biological systems [23, 34] to the stability of a power grid [17]. Note that for the particular case where f is time-invariant, this implies that if Ω contains an equilibrium point e then it is unique and all trajectories converge to e.

11.3 Definitions of Contraction After Small Transients

We begin by defining three generalizations of (11.2).

Definition 11.1 The time-varying system (11.1) is said to be *contractive after a small overshoot and short transient* (SOST) on Ω w.r.t. a norm $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ if for each $\varepsilon > 0$ and each $\tau > 0$ there exists $\ell = \ell(\tau, \varepsilon) > 0$ such that

$$|x(t_2+\tau,t_1,a)-x(t_2+\tau,t_1,b)| \le (1+\varepsilon)\exp(-(t_2-t_1)\ell)|a-b| \tag{11.8}$$

for all $t_2 \ge t_1 \ge 0$ and all $a, b \in \Omega$.

This definition is motivated by requiring contraction at an exponential rate, but only after an (arbitrarily small) time τ, and with an (arbitrarily small) overshoot $(1+\varepsilon)$. However, as we will see below when the convergence rate ℓ may depend on ε a somewhat richer behavior may occur.

Definition 11.2 The time-varying system (11.1) is said to be *contractive after a small overshoot* (SO) on Ω w.r.t. a norm $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ if for each $\varepsilon > 0$ there exists $\ell = \ell(\varepsilon) > 0$ such that

$$|x(t_2,t_1,a)-x(t_2,t_1,b)| \le (1+\varepsilon)\exp(-(t_2-t_1)\ell)|a-b| \tag{11.9}$$

for all $t_2 \ge t_1 \ge 0$ and all $a, b \in \Omega$.

The definition of SO is thus similar to that of SOST, yet now the convergence rate ℓ depends only on ε, and there is no time transient τ (i.e., $\tau = 0$). In other words, SO is a uniform (in τ) version of SOST.

Definition 11.3 The time-varying system (11.1) is said to be *contractive after a short transient* (ST) on Ω w.r.t. a norm $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ if for each $\tau > 0$ there exists $\ell = \ell(\tau) > 0$ such that

$$|x(t_2+\tau,t_1,a)-x(t_2+\tau,t_1,b)| \le \exp(-(t_2-t_1)\ell)|a-b| \tag{11.10}$$

for all $t_2 \ge t_1 \ge 0$ and all $a, b \in \Omega$.

This definition allows the contraction to "kick in" only after a time transient of length τ.

It is clear that every contractive system is SOST, SO, and ST. Thus, all these notions are generalizations of contraction. Also, both SO and ST imply SOST and, as we will see below, under a mild technical condition on (11.1) SO and SOST are equivalent. Figure 11.2 on p. 16 summarizes the relations between these GCSs (as well as other notions defined below).

The next simple example demonstrates a system that does not satisfy (11.2), but is a GCS.

Example 11.1 Consider the *scalar* time-varying system

$$\dot{x}(t) = -\alpha(t)x(t), \tag{11.11}$$

where the state x evolves on $\Omega := [-1, 1]$, and $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is a class K function (i.e. α is continuous and strictly increasing, with $\alpha(0) = 0$). It is straightforward to show that this system does not satisfy (11.2) w.r.t. *any* norm (consider the trajectories emanating from $x(0) = 0$ and from $x(0) = \varepsilon$, with $\varepsilon > 0$ sufficiently small), yet it is ST, with $\ell(\tau) = \alpha(\tau) > 0$, for any given $\tau > 0$.

The next section analyzes the properties of the three forms of GCSs introduced above, with an emphasis on checkable conditions for establishing that a system is a GCS based on matrix measures. We assume from here on that the state space $\Omega \subset \mathbb{R}^n$ is compact and convex. The proofs of all the results are placed in Sect. 11.5.

11.4 Properties of GCSs

The next three subsections study the three forms of GCSs defined above.

11.4.1 Contraction After a Small Overshoot and Short Transient (SOST)

Just like contraction, SOST implies entrainment to a periodic excitation. To demonstrate this, assume for example that the vector field f in (11.1) is T periodic. Pick $t_0 \geq 0$. Define $m : \Omega \to \Omega$ by $m(a) := x(T + t_0, t_0, a)$. In other words, m maps a to the solution of (11.1) at time $T + t_0$ for the initial condition $x(t_0) = a$. Then m is continuous and maps the convex and compact set Ω to itself, so by the Brouwer fixed point theorem (see, e.g. [11, Chap. 6]) there exists $\zeta \in \Omega$ such that $m(\zeta) = \zeta$, i.e., $x(T + t_0, t_0, \zeta) = \zeta$. This implies that (11.1) admits a periodic solution $\gamma : [t_0, \infty) \to \Omega$ with period T. Assuming that the system is also SOST, pick $\tau, \varepsilon > 0$. Then there exists $\ell = \ell(\tau, \varepsilon) > 0$ such that

$$|x(t - t_0 + \tau, t_0, a) - x(t - t_0 + \tau, t_0, \zeta)| \leq (1 + \varepsilon) \exp(-(t - t_0)\ell)|a - \zeta|,$$

for all $a \in \Omega$ and all $t \geq t_0$. Taking $t \to \infty$ implies that every solution converges to γ. In particular, there cannot be two distinct periodic solutions. Thus, we proved the following.

Proposition 11.1 *Suppose that the time-varying system (11.1), with state x evolving on a compact and convex state-space $\Omega \subset \mathbb{R}^n$, is SOST, and that the vector field f is T-periodic. Then for any $t_0 \geq 0$ it admits a unique periodic solution $\gamma : [t_0, \infty) \to \Omega$ with period T, and $x(t, t_0, a)$ converges to γ for any $a \in \Omega$.*

Since both SO and ST imply SOST, Proposition 11.1 holds for all three forms of GCSs.

Our next goal is to derive a sufficient condition for SOST. One may naturally expect that if (11.1) is contractive w.r.t. a set of norms $|\cdot|_\zeta$, with, say $\zeta \in (0, p], p > 0$, and that $\lim_{\zeta \to 0} |\cdot|_\zeta = |\cdot|$ then (11.1) is a GCS w.r.t. the norm $|\cdot|$. In fact, this can be further generalized by requiring (11.1) to be contractive w.r.t. $|\cdot|_\zeta$ only on suitable subsets Ω_ζ of the state-space. This leads to the following definition.

Definition 11.4 System (11.1) is said to be *nested contractive* (NC) on Ω with respect to a norm $|\cdot|$ if there exist convex sets $\Omega_\zeta \subseteq \Omega$, and norms $|\cdot|_\zeta : \mathbb{R}^n \to \mathbb{R}_+$, where $\zeta \in (0, 1/2]$, such that the following conditions hold.

(a) $\cup_{\zeta\in(0,1/2]}\Omega_\zeta = \Omega$, and

$$\Omega_{\zeta_1} \subseteq \Omega_{\zeta_2}, \quad \text{for all } \zeta_1 \geq \zeta_2. \tag{11.12}$$

(b) For every $\tau > 0$ there exists $\zeta = \zeta(\tau) \in (0, 1/2]$, with $\zeta(\tau) \to 0$ as $\tau \to 0$, such that for every $a \in \Omega$ and every $t_1 \geq 0$

$$x(t, t_1, a) \in \Omega_\zeta, \quad \text{for all } t \geq t_1 + \tau, \tag{11.13}$$

and (11.1) is contractive on Ω_ζ with respect to $|\cdot|_\zeta$.

(c) The norms $|\cdot|_\zeta$ converge to $|\cdot|$ as $\zeta \to 0$, i.e., for every $\zeta > 0$ there exists $s = s(\zeta) > 0$, with $s(\zeta) \to 0$ as $\zeta \to 0$, such that

$$(1 - s)|y| \leq |y|_\zeta \leq (1 + s)|y|, \quad \text{for all } y \in \Omega.$$

Equation (11.13) means that after an arbitrarily short time ζ every trajectory enters and remains in a subset Ω_ζ of the state space on which we have contraction with respect to $|\cdot|_\zeta$. We can now state the main result in this subsection. Recall that the proofs of all the results are placed in Sect. 11.5.

Theorem 11.1 *If the system (11.1) is NC w.r.t. the norm $|\cdot|$ then it is SOST w.r.t. the norm $|\cdot|$.*

The next result is an application of Theorem 11.1 to systems with a cyclic structure (see, e.g. [6, 7] and the references therein). It also shows that as we change the parameters in a contractive system, it may become a GCS when it hits the "verge" of contraction (as defined in 11.2). This is reminiscent of an asymptotically stable system that becomes marginally stable as it looses stability.

Proposition 11.2 *Consider the system*

$$\begin{aligned}
\dot{x}_1 &= -f_1(x_1) + g(x_n),\\
\dot{x}_2 &= -f_2(x_2) + k_1 x_1,\\
\dot{x}_3 &= -f_3(x_3) + k_2 x_2,\\
&\vdots\\
\dot{x}_n &= -f_n(x_n) + k_{n-1} x_{n-1}.
\end{aligned} \tag{11.14}$$

Suppose that the following properties hold for all i: $k_i > 0$, $f_i(0) = 0$, $f_i'(s)$ is a non-decreasing function of s with $f_i'(0) > 0$, $g(0) > 0$, and $g'(s)$ is a strictly decreasing function of s with $g'(s) > 0$ for all $s \geq 0$. (Note that these properties imply in particular that $\mathbb{R}^n_+$ is an invariant set of the dynamics. We further assume that there *exists a compact and convex set $\Omega \subset \mathbb{R}^n_+$ that is an invariant set of the dynamics.) Let $k := \prod_{i=1}^{n-1} k_i$. For $\varepsilon > 0$, let*

$$D_\varepsilon := \operatorname{diag}\left(1, \frac{f_1'(0)-\varepsilon}{k_1}, \frac{(f_1'(0)-\varepsilon)(f_2'(0)-\varepsilon)}{k_1 k_2}, \dots, \prod_{i=1}^{n-1} \frac{f_i'(0)-\varepsilon}{k_i}\right).$$

If

$$\prod_{i=1}^{n} f_i'(0) > kg'(0) \tag{11.15}$$

then (11.14) is contractive on Ω w.r.t. the scaled norm $|\cdot|_{1,D_\varepsilon}$ for all $\varepsilon > 0$ sufficiently small. If $\prod_{i=1}^{n} f_i'(0) = kg'(0)$ then (11.14) does not satisfy (11.2), w.r.t. any (fixed) norm on Ω, yet it is SOST on Ω w.r.t. the norm $|\cdot|_{1,D_0}$.

Example 11.2 Consider the cyclic system

$$\begin{aligned} \dot{x}_1 &= -\alpha_1 x_1 + g(x_n), \\ \dot{x}_2 &= -\alpha_2 x_2 + x_1, \\ \dot{x}_3 &= -\alpha_3 x_3 + x_2, \\ &\vdots \\ \dot{x}_n &= -\alpha_n x_n + x_{n-1}, \end{aligned} \tag{11.16}$$

where $\alpha_i > 0$, and

$$g(u) := \frac{1+u}{c+u}, \quad \text{with } c > 1.$$

As explained in [39, Chap. 4] this is a model for a simple biochemical feedback control circuit for protein synthesis in the cell. The x_i's represent concentrations of various macromolecules in the cell and are therefore non-negative. It is straightforward to see that this system satisfies all the properties in Proposition 11.2 with $f_i(s) = \alpha_i s$, and $k_i = 1$. Using the fact that $g(u) < 1$ for all $u \geq 0$ it is straightforward to show that the set $\Omega_r := r([0, \alpha_1^{-1}] \times [0, (\alpha_1\alpha_2)^{-1}] \times \cdots \times [0, \alpha^{-1}])$ is an invariant set of the dynamics for all $r \geq 1$.

Let $\alpha := \prod_{i=1}^{n} \alpha_i$. We conclude that if

$$c - 1 < c^2 \alpha$$

then (11.16) is contractive on Ω_r w.r.t. the scaled norm $|\cdot|_{1,D_\varepsilon}$ for all $\varepsilon > 0$ sufficiently small, where $D_\varepsilon := \operatorname{diag}\left(1, \alpha_1 - \varepsilon, (\alpha_1 - \varepsilon)(\alpha_2 - \varepsilon), \dots, \prod_{i=1}^{n-1}(\alpha_i - \varepsilon)\right)$. On the other hand, if $c - 1 = c^2\alpha$ then (11.16) does not satisfy (11.2), w.r.t. any (fixed) norm on Ω_r, yet it is SOST on Ω_r w.r.t. the norm $|\cdot|_{1,D_0}$. Intuitively speaking, this means that the system is contractive when the "total dissipation" α is strictly larger than the maximal value of the feedback's derivative $g'(0)$, and looses contractivity to become SOST when these two values are equal.

Thus, (11.16), with $c - 1 \leq c^2\alpha$, admits a unique equilibrium point $e \in \Omega_1$ and

$$\lim_{t\to\infty} x(t,a) = e, \quad \text{for all } a \in \mathbb{R}^n_+.$$

This property also follows from a more general result [39, Prop. 4.2.1] that is proved using the theory of irreducible cooperative dynamical systems. Yet the contraction approach leads to new insights. For example, it implies that the distance between trajectories can only decrease, and can also be used to prove entrainment to suitable generalizations of (11.16) that include periodically varying inputs.

We now describe another application of Theorem 11.1 to a model from systems biology. Cells often respond to stimulus by modification of proteins. One mechanism for doing this is *phosphorelay* (also called phosphotransfer) in which a phosphate group is transferred through a serial chain of proteins from an initial histidine kinase (HK) down to a final response regulator (RR). The next example uses Theorem 11.1 to analyze a model for phosphorelay studied in [13].

Example 11.3 Consider the system

$$\begin{aligned}
\dot{x}_1 &= (p_1 - x_1)c - \eta_1 x_1 (p_2 - x_2),\\
\dot{x}_2 &= \eta_1 x_1 (p_2 - x_2) - \eta_2 x_2 (p_3 - x_3),\\
&\vdots\\
\dot{x}_{n-1} &= \eta_{n-2} x_{n-2} (p_{n-1} - x_{n-1}) - \eta_{n-1} x_{n-1} (p_n - x_n),\\
\dot{x}_n &= \eta_{n-1} x_{n-1} (p_n - x_n) - \eta_n x_n,
\end{aligned} \tag{11.17}$$

where $\eta_i, p_i > 0$, and $c : [t_1, \infty) \to \mathbb{R}_+$. In the context of phosphorelay [13], $c(t)$ is the strength at time t of the stimulus activating the HK, $x_i(t)$ is the concentration of the phosphorylated form of the protein at the ith layer at time t, and p_i denotes the total protein concentration at that layer. Note that $\eta_n x_n$ is the flow of the phosphate group to an external receptor molecule.

In the particular case where $p_i = 1$ for all i (11.17) becomes the *ribosome flow model* (RFM) [32]. This is the mean-field approximation of an important model from nonequilibrium statistical physics called the *totally asymmetric simple exclusion process* (TASEP) [9]. In the RFM, $x_i \in [0, 1]$ is the normalized occupancy at site i, where $x_i = 0$ [$x_i = 1$] means that site i is completely free [full], and η_i is the capacity of the link that connects site i to site $i + 1$. This has been used to model mRNA translation, where every site corresponds to a group of codons on the mRNA strand, $x_i(t)$ is the normalized occupancy of ribosomes at site i at time t, $c(t)$ is the initiation rate at time t, and η_i is the elongation rate from site i to site $i + 1$.

Our original motivation for introducing GCSs was to prove entrainment in the RFM [23]. For more on the analysis of the RFM, and networks of interconnected RFMs, using tools from systems and control theory, see [25–27, 29–31, 46].

Assume that there exists $\eta_0 > 0$ such that $c(t) \geq \eta_0$ for all $t \geq t_1$. Let $\Omega := [0, p_1] \times \cdots \times [0, p_n]$ denote the state-space of (11.17). Then, as shown in Sect. 11.5, (11.17) does not satisfy (11.2), w.r.t. any norm, on Ω, yet it is SOST on Ω w.r.t. the L_1 norm.

Systems in which every state variable describes the amount of "material" in a compartment, and the dynamics describes the flow between the compartments and the environment are called compartmental systems [19]. Both (11.16) and (11.17) are thus compartmental systems. Analysis of contraction in such systems using the matrix measure corresponding to the scaled L_1 norm goes back at least to the work of Sandberg [36].

Considering Theorem 11.1 in the special case where all the sets Ω_ζ in Definition 11.4 are equal to Ω yields the following result.

Corollary 11.1 *Suppose that (11.1) is contractive on Ω w.r.t. a set of norms $|\cdot|_\zeta$, $\zeta \in (0, 1/2]$, and that condition (c) in Definition 11.4 holds. Then (11.1) is SOST on Ω w.r.t. $|\cdot|$.*

Corollary 11.1 may be useful in cases where some matrix measure of the Jacobian J of (11.1) turns out to be non positive on Ω, but not strictly negative, suggesting that the system is "on the verge" of satisfying (11.2). The next result demonstrates this for the time-invariant system

$$\dot{x} = f(x), \tag{11.18}$$

and the particular case of the matrix measure $\mu_1 : \mathbb{R}^{n \times n} \to \mathbb{R}$ induced by the L_1 norm. Recall that this is given by (11.4) with the c_js defined in (11.5).

Proposition 11.3 *Consider the Jacobian $J(\cdot) : \Omega \to \mathbb{R}^{n \times n}$ of the time-invariant system (11.18). Suppose that Ω is compact and convex, and that the set $\{1, \ldots, n\}$ can be divided into two nonempty disjoint sets S_0 and S_- such that the following properties hold for all $x \in \Omega$:*

1. *for any $k \in S_0$, $c_k(J(x)) \leq 0$;*
2. *for any $j \in S_-$, $c_j(J(x)) < 0$;*
3. *for any $i \in S_0$ there exists an index $z = z(i) \in S_-$ such that $J_{zi}(x) > 0$.*

Then (11.18) is SOST on Ω w.r.t. the L_1 norm.

The proof of Proposition 11.3 is based on the following idea. By compactness of Ω, there exists $\delta > 0$ such that

$$c_j(J(x)) < -\delta, \quad \text{for all } j \in S_- \text{ and all } x \in \Omega. \tag{11.19}$$

The conditions stated in the proposition imply that there exists a diagonal matrix P such that $c_k(PJP^{-1}) < 0$ for all $k \in S_0$. Furthermore, there exists such a P with diagonal entries *arbitrarily close* to 1, so $c_j(PJP^{-1}) < -\delta/2$ for all $j \in S_-$. Thus, $\mu_1(PJP^{-1}) < 0$. Now Corollary 11.1 implies SOST. Note that this implies that the compactness assumption may be dropped if for example it is known that (11.19) holds.

Example 11.4 Consider the system:

$$\begin{aligned} \dot{x} &= -\delta x + k_1 y - k_2(e_T - y)x, \\ \dot{y} &= -k_1 y + k_2(e_T - y)x, \end{aligned} \tag{11.20}$$

where $\delta, k_1, k_2, e_T > 0$, and $\Omega := [0, \infty) \times [0, e_T]$. This is a basic model for a transcriptional module that is ubiquitous in both biology and synthetic biology (see, e.g. [14, 34]). Here $x(t)$ is the concentration at time t of a transcriptional factor X that regulates a downstream transcriptional module by binding to a promoter with concentration $e(t)$ yielding a protein-promoter complex Y with concentration $y(t)$. The binding reaction is reversible with binding and dissociation rates k_2 and k_1, respectively. The linear degradation rate of X is δ, and as the promoter is not subject to decay, its total concentration, e_T, is conserved, so $e(t) = e_T - y(t)$. The Jacobian of (11.20) is $J = \begin{bmatrix} -\delta - k_2(e_T - y) & k_1 + k_2 x \\ k_2(e_T - y) & -k_1 - k_2 x \end{bmatrix}$, and all the properties in Proposition 11.3 hold with $S_- = \{1\}$ and $S_0 = \{2\}$. Indeed, $J_{12} = k_1 + k_2 x > k_1 > 0$ for all $\begin{bmatrix} x & y \end{bmatrix}^T \in \Omega$. Thus, (11.20) is SOST on Ω w.r.t. the L_1 norm. Note that Ref. [34] showed that (11.20) is contractive w.r.t. a certain *weighted* L_1 norm. Here we showed SOST w.r.t. the (unweighted) L_1 norm.

Example 11.5 A more general example studied in [34] is where the transcription factor regulates several independent downstream transcriptional modules. This leads to the following model:

$$\begin{aligned} \dot{x} &= -\delta x + k_{11}y_1 - k_{21}(e_{T,1} - y_1)x + k_{12}y_2 - k_{22}(e_{T,2} - y_2)x \\ &\quad + \cdots + k_{1n}y_n - k_{2n}(e_{T,n} - y_n)x, \\ \dot{y}_1 &= -k_{11}y_1 + k_{21}(e_{T,1} - y_1)x, \\ &\vdots \\ \dot{y}_n &= -k_{1n}y_n + k_{2n}(e_{T,n} - y_n)x, \end{aligned} \tag{11.21}$$

where n is the number of regulated modules. The state-space is $\Omega = [0, \infty) \times [0, e_{T,1}] \times \cdots \times [0, e_{T,n}]$. The Jacobian of (11.21) is

$$J = \begin{bmatrix} -\delta - \sum_{i=1}^{n} k_{2i}(e_{T,i} - y_i) & k_{11} + k_{21}x & k_{12} + k_{22}x & \ldots & k_{1n-1} + k_{2n-1}x & k_{1n} + k_{2n}x \\ k_{21}(e_{T,1} - y_1) & -k_{11} - k_{21}x & 0 & \ldots & 0 & 0 \\ k_{22}(e_{T,2} - y_2) & 0 & -k_{12} - k_{22}x & 0 & \ldots & 0 \\ & \vdots & & & & \\ k_{2n}(e_{T,n} - y_n) & 0 & 0 & \ldots & 0 & -k_{1n} - k_{2n}x \end{bmatrix},$$

and all the properties in Proposition 11.3 hold with $S_- = \{1\}$ and $S_0 = \{2, 3, \ldots, n\}$. Thus, this system is SOST on Ω w.r.t. the L_1 norm.

Arguing as in the proof of Proposition 11.3 for the matrix measure μ_∞ induced by the L_∞ norm (see 11.7) yields the following result.

Proposition 11.4 *Consider the Jacobian $J(\cdot) : \Omega \to \mathbb{R}^{n\times n}$ of the time-invariant system (11.18). Suppose that Ω is compact and that the set $\{1,\dots,n\}$ can be divided into two nonempty disjoint sets S_0 and S_- such that the following properties hold for all $x \in \Omega$:*

1. *$d_j(J(x)) \le 0$ for all $j \in S_0$;*
2. *$d_k(J(x)) < 0$ for all $k \in S_-$;*
3. *for any $j \in S_0$ there exists an index $z = z(j) \in S_-$ such that $J_{jz}(x) \neq 0$.*

Then (11.18) is SOST on Ω w.r.t. the L_∞ norm.

11.4.2 Contraction After a Small Overshoot (SO)

A natural question is under what conditions SO and SOST are equivalent. To address this issue, we introduce the following definition.

Definition 11.5 We say that (11.1) is *weakly expansive* (WE) if for each $\delta > 0$ there exists $\tau_0 > 0$ such that for all $a, b \in \Omega$ and all $t_0 \ge 0$

$$|x(t,t_0,a) - x(t,t_0,b)| \le (1+\delta)|a-b|, \quad \text{for all } t \in [t_0, t_0+\tau_0]. \tag{11.22}$$

Proposition 11.5 *Suppose that (11.1) is WE. Then (11.1) is SOST if and only if it is SO.*

Remark 11.1 Suppose that f in (11.1) is Lipschitz globally in Ω uniformly in t, i.e., there exists $L > 0$ such that

$$|f(t,x) - f(t,y)| \le L|x-y|, \quad \text{for all } x, y \in \Omega,\ t \ge 0.$$

Then by Gronwall's Lemma (see, e.g. [41, Appendix C])

$$|x(t,t_0,a) - x(t,t_0,b)| \le \exp\big(L(t-t_0)\big)\,|a-b|,$$

for all $t \ge t_0 \ge 0$, and this implies that (11.22) holds for $\tau_0 := \frac{1}{L}\ln(1+\delta) > 0$. In particular, if Ω is compact and f is periodic in t then WE holds under rather weak continuity arguments on f.

To explore the connection of SO with other notions related to contraction, we require the following definitions.

Definition 11.6 We say that (11.1) is *non expansive* (NE) w.r.t. a norm $|\cdot|$ if for all $a, b \in \Omega$ and all $s_2 > s_1 \ge 0$

$$|x(s_2,s_1,a) - x(s_2,s_1,b)| \le |a-b|. \tag{11.23}$$

We say that (11.1) is *weakly contractive* (WC) if (11.23) holds with $\le$ replaced by $<$.

One may perhaps expect that any of the three generalizations of contraction also implies WC. However, the next example shows that SO does not imply WC.

Example 11.6 Consider the scalar system

$$\dot{x} = \begin{cases} -2x, & 0 \le |x| < 1/2, \\ -\frac{x}{|x|}, & \frac{1}{2} \le |x| \le 1, \end{cases} \tag{11.24}$$

with x evolving on $\Omega := [-1, 1]$. Clearly, this system is not WC. However, it is not difficult to show that it satisfies the definition of SO with $\ell = \ell(\varepsilon) := \min\{\ln(1 + \varepsilon), 1\}$.

11.4.3 Contraction After a Short Transient (ST)

For *time-invariant* systems whose state evolves on a convex and compact set Ω it is possible to give a simple sufficient condition for ST. Let $\text{Int}(S)$ $[\partial S]$ denote the interior [boundary] of a set S.

Definition 11.7 The time-invariant system (11.18) with the state x evolving on a compact and convex set $\Omega \subset \mathbb{R}^n$, is said to be *interior contractive* (IC) w.r.t. a norm $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ if the following properties hold:

(a) for every $x_0 \in \partial\Omega$,
$$x(t, x_0) \notin \partial\Omega, \quad \text{for all } t > 0; \tag{11.25}$$

(b) for every $x \in \text{Int}(\Omega)$,
$$\mu(J(x)) < 0, \tag{11.26}$$

where $\mu : \mathbb{R}^{n\times n} \to \mathbb{R}$ is the matrix measure induced by $|\cdot|$.

In other words, the matrix measure is negative in the interior of Ω, and the boundary of Ω is "repelling". Note that these conditions do not necessarily imply that the system satisfies (11.2) on Ω, as it is possible that $\mu(J(x)) = 0$ for some $x \in \partial\Omega$. Yet, (11.26) does imply that (11.18) is NE on Ω. We can now state the main result in this subsection.

Theorem 11.2 *If the system (11.18) is IC w.r.t. a norm $|\cdot|$ then it is ST w.r.t. $|\cdot|$.*

The proof of this result is based on showing that IC implies that for each $\tau > 0$ there exists $d = d(\tau) > 0$ such that

$$\text{dist}(x(t, x_0), \partial\Omega) \ge d, \quad \text{for all } x_0 \in \Omega \text{ and all } t \ge \tau,$$

and then using this to conclude that for any $t \ge \tau$ all the trajectories of the system are contained in a convex and compact set $D \subset \text{Int}(\Omega)$. In this set the system is contractive with rate $c := \max_{x \in D} \mu(J(x)) < 0$. The next example, that is a variation of a system studied in [34], demonstrates this reasoning.

Example 11.7 Consider a transcriptional factor X that regulates a downstream transcriptional module by irreversibly binding, at a rate $k_2 > 0$, to a promoter E yielding a protein-promoter complex Y. The promoter is not subject to decay, so its total concentration, denoted by $e_T > 0$, is conserved. Assume also that X is obtained from an inactive form X_0, for example through a phosphorylation reaction that is catalyzed by a kinase with abundance $u(t)$ satisfying $u(t) \geq u_0 > 0$ for all $t \geq 0$. The sum of the concentrations of X_0, X, and Y is constant, denoted by z_T, with $z_T > e_T$. Letting $x_1(t), x_2(t)$ denote the concentrations of X, Y at time t yields the model

$$\begin{aligned} \dot{x}_1 =& (z_T - x_1 - x_2)u - \delta x_1 - k_2(e_T - x_2)x_1, \\ \dot{x}_2 =& k_2(e_T - x_2)x_1, \end{aligned} \tag{11.27}$$

with the state evolving on $\Omega := [0, z_T] \times [0, e_T]$. Here $\delta \geq 0$ is the dephosphorylation rate $X \to X_0$. Let $P := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and consider the matrix measure $\mu_{\infty,P}$. A calculation yields

$$\begin{aligned} \tilde{J} &:= PJP^{-1} \\ &= \begin{bmatrix} -u - \delta & \delta \\ k_2(e_T - x_2) & k_2(x_2 - x_1 - e_T) \end{bmatrix}, \end{aligned}$$

so $d_1(\tilde{J}) = -u - \delta + |\delta| \leq -u_0 < 0$, and

$$\begin{aligned} d_2(\tilde{J}) &= k_2(x_2 - x_1 - e_T) + |k_2(e_T - x_2)| \\ &= -k_2 x_1. \end{aligned}$$

Letting $S := \{0\} \times [0, e_T]$, we conclude that $\mu_{\infty,P}(x) < 0$ for all $x \in (\Omega \setminus S)$. For any $x \in S$, $\dot{x}_1 = (z_T - x_2)u \geq (z_T - e_T)u_0 > 0$, and arguing as in the proof of Theorem 11.2 (see Sect. 11.5), we conclude that for any $\tau > 0$ there exists $d = d(\tau) > 0$ such that

$$x_1(t, a) \geq d, \quad \text{for all } a \in \Omega \text{ and all } t \geq \tau.$$

In other words, after time τ all the trajectories are contained in the closed and convex set $D = D(\tau) := [d, z_T] \times [0, e_T]$. Letting $c := c(\tau) = \max_{x \in D} \mu_{\infty,P}(J(x))$ yields $c < 0$ and

$$|x(t + \tau, a) - x(t + \tau, b)|_{\infty,P} \leq \exp(ct)|a - b|_{\infty,P}, \quad \text{for all } a, b \in \Omega \text{ and all } t > 0,$$

so (11.27) is ST w.r.t. $|\cdot|_{\infty,P}$.

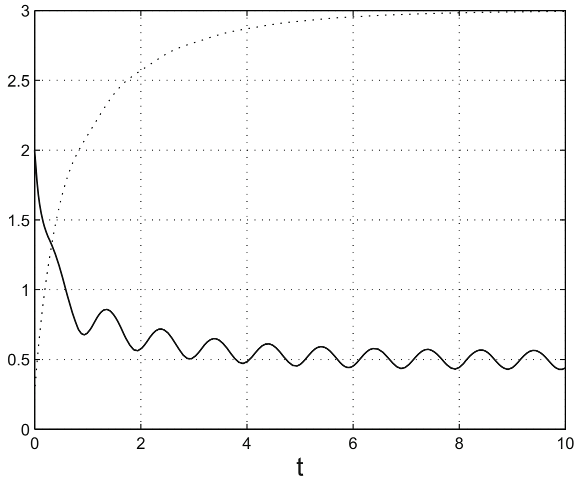


Fig. 11.1 Solution $x_1(t)$ (*solid line*) and $x_2(t)$ (*dashed line*) of the system in Example 11.8 as a function of t

As noted above, one motivation for GCSs is that contraction is used to deduce asymptotic results, so allowing initial transients should increase the class of systems that can be analyzed. The next result demonstrates this.

Corollary 11.2 *If (11.18) is IC with respect to some norm then it admits a unique equilibrium point* $e \in \text{Int}(\Omega)$, *and* $\lim_{t\to\infty} x(t,a) = e$ *for all* $a \in \Omega$.

Remark 11.2 The proof of Corollary 11.2, given in the Appendix, is based on Theorem 11.2. It is possible to give another proof using the *variational system* (see, e.g. [18]) associated with (11.18):

$$\begin{aligned} \dot{x} &= f(x), \\ \dot{\delta x} &= J(x)\delta x. \end{aligned} \tag{11.28}$$

The function $V(x, \delta x) := |\delta x|$, where $|\cdot| : \mathbb{R}^n \to \mathbb{R}_+$ is the vector norm corresponding to the matrix measure μ in (11.26), is a Lyapunov-Finsler function of (11.28), and Corollary 11.2 follows from the LaSalle invariance principle described in [18].

Since IC implies ST and this implies SOST, it follows from Proposition 11.1 that IC implies entrainment to T-periodic vector fields.[1] The next example demonstrates this.

Example 11.8 Consider again the system in Example 11.7, and assume that the kinase abundance $u(t)$ is a strictly positive and periodic function of time with period T. Since we already showed that this system is ST, it admits a unique periodic solution γ, of period T, and any trajectory of the system converges to γ. Figure 11.1

[1] Note that the proof that IC implies ST used a result for time-invariant systems, but an analogous argument holds for the time-varying case as well.

depicts the solution of (11.27) for $\delta = 2, k_2 = 1, z_T = 4, e_T = 3, u(t) = 2 + \sin(2\pi t)$, and initial condition $x_1(0) = 2, x_2(0) = 1/4$. It may be seen that both state variables converge to a periodic solution with period $T = 1$. (In particular, x_2 converges to the constant function $x_2(t) \equiv e_T$ that is of course periodic with period T.)

Contraction can be characterized using a Lyapunov-Finsler function [18]. The next result describes a similar characterization for ST. For simplicity, we state this for the time-invariant system (11.18).

Proposition 11.6 *The following two conditions are equivalent.*

(a) The time-invariant system (11.18) is ST w.r.t. a norm $|\cdot|$.
(b) For any $\tau > 0$ there exists $\ell = \ell(\tau) > 0$ such that for any $a, b \in \Omega$ and any c on the line connecting a and b the solution of (11.28) with $x(0) = c$ and $\delta x(0) = b - a$ satisfies

$$|\delta x(t+\tau)| \le \exp(-\ell t)|\delta x(0)|, \quad \textit{for all } t \ge 0. \tag{11.29}$$

Note that (11.29) implies that the function $V(x, \delta x) := |\delta x|$ is a generalized *Lyapunov-Finsler function in the following sense. For any $\tau > 0$ there exists $\ell = \ell(\tau) > 0$ such that along solutions of the variational system:*

$$V\left(x(t+\tau, x(0)), \delta x(t+\tau, \delta x(0), x(0))\right) \le \exp(-\ell t) V(x(0), \delta x(0)),$$

for all $t \ge 0$.

Figure 11.2 summarizes the relations between the various contraction notions.

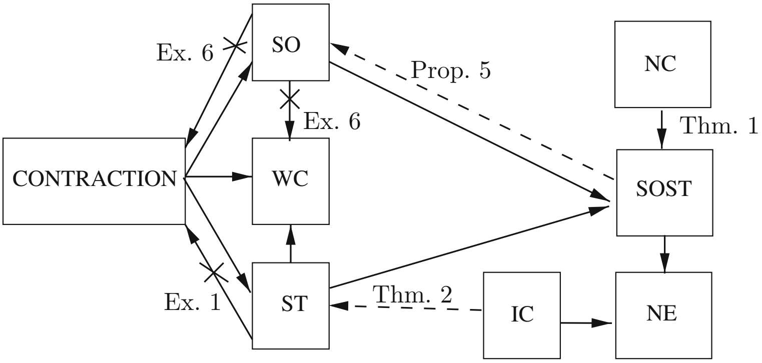


Fig. 11.2 Relations between various contraction notions. A *solid arrow* denotes implication; a *crossed out arrow* denotes that the implication is in general false; and a *dashed arrow* denotes an implication that holds under some additional conditions. Some of the relations are immediate. Others follow from the results marked near the *arrows*

11.5 Proofs

Proof of Theorem 11.1 Fix arbitrary $t_1 \geq 0$. The function $\zeta = \zeta(\tau) \in (0, 1/2]$ is as in the statement of the Theorem. For each $\tau > 0$, let $c_\zeta > 0$ be a contraction constant on Ω_ζ, where we write $\zeta = \zeta(\tau)$ here and in what follows. Pick $a, b \in \Omega$ and $\tau > 0$. By (11.13), $x(t, t_1, a), x(t, t_1, b) \in \Omega_\zeta$ for all $t \geq t_1 + \tau$, so

$$\begin{aligned}|x(t, t_1, a)&-x(t, t_1, b)|_\zeta \\ &\leq \exp(-c_\zeta(t - t_1 - \tau))|x(t_1 + \tau, t_1, a) - x(t_1 + \tau, t_1, b)|_\zeta,\end{aligned}$$

for all $t \geq t_1 + \tau$. In particular,

$$|x(t, t_1, a) - x(t, t_1, b)|_\zeta < |x(t_1 + \tau, t_1, a) - x(t_1 + \tau, t_1, b)|_\zeta, \tag{11.30}$$

for all $t > t_1 + \tau$. From the convergence property of norms in the Theorem statement, there exist $v_\zeta, w_\zeta > 0$ such that

$$|y| \leq v_\zeta |y|_\zeta \leq w_\zeta v_\zeta |y|, \quad \text{for all } y \in \Omega, \tag{11.31}$$

and $v_\zeta \to 1$, $w_\zeta \to 1$ as $\tau \to 0$. Combining this with (11.30) yields

$$|x(t, t_1, a) - x(t, t_1, b)| < v_\zeta w_\zeta |x(t_1 + \tau, t_1, a) - x(t_1 + \tau, t_1, b)|,$$

for all $t > t_1 + \tau$. Note that taking $\tau \to 0$ yields

$$|x(t, t_1, a) - x(t, t_1, b)| \leq |a - b|, \quad \text{for all } t > t_1. \tag{11.32}$$

Now for $t \geq t_1 + \tau$ let $p := t - t_1 - \tau$. Then

$$\begin{aligned}|x(t, t_1, a) - x(t, t_1, b)| &\leq v_\zeta |x(t, t_1, a) - x(t, t_1, b)|_\zeta \\ &\leq v_\zeta \exp(-c_\zeta p)|x(t_1 + \tau, t_1, a) - x(t_1 + \tau, t_1, b)|_\zeta \\ &\leq v_\zeta w_\zeta \exp(-c_\zeta p)|x(t_1 + \tau, t_1, a) - x(t_1 + \tau, t_1, b)| \\ &\leq v_\zeta w_\zeta \exp(-c_\zeta p)|a - b|,\end{aligned}$$

where the last inequality follows from (11.32). Now pick $\varepsilon > 0$. Since $v_\zeta \to 1$, $w_\zeta \to 1$ as $\tau \to 0$, $v_\zeta w_\zeta \leq 1 + \varepsilon$ for $\tau > 0$ small enough. We conclude that there exists $\tau_m > 0$ sufficiently small such that for all $\bar{\tau} \in [0, \tau_m]$

$$|x(t + \bar{\tau}, t_1, a) - x(t + \bar{\tau}, t_1, b)| \leq (1 + \varepsilon) \exp(-c_\zeta(t - t_1))|a - b|, \tag{11.33}$$

for all $a, b \in \Omega$ and all $t \geq t_1$. Now pick $\bar{\tau} > \tau_m$. For any $t \geq t_1$, let $s := t + \bar{\tau} - \tau_m$. Then

$$\begin{aligned}|x(t+\bar{\tau},t_1,a)-x(t+\bar{\tau},t_1,b)| &= |x(s+\tau_m,t_1,a)-x(s+\tau_m,t_1,b)|\\ &\leq (1+\varepsilon)\exp(-c_\zeta(s-t_1))|a-b|\\ &\leq (1+\varepsilon)\exp(-c_\zeta(t-t_1))|a-b|,\end{aligned}$$

and this completes the proof of Theorem 11.1. □

Proof of Proposition 11.2 The Jacobian of (11.14) is

$$J(x)=\begin{bmatrix}-f_1'(x_1) & 0 & 0 & \dots & 0 & g'(x_n)\\ k_1 & -f_2'(x_2) & 0 & \dots & 0 & 0\\ 0 & k_2 & -f_3'(x_3) & \dots & 0 & 0\\ & & \vdots & & & \\ 0 & 0 & 0 & \dots & k_{n-1} & -f_n'(x_n)\end{bmatrix}, \tag{11.34}$$

so

$$D_\varepsilon J(x)D_\varepsilon^{-1}=\begin{bmatrix}-f_1'(x_1) & 0 & 0\dots & 0 & \frac{g'(x_n)}{\prod_{i=1}^{n-1}\frac{f_i'(0)-\varepsilon}{k_i}}\\ f_1'(0)-\varepsilon & -f_2'(x_2) & 0\dots & 0 & 0\\ 0 & f_2'(0)-\varepsilon & 0\dots & 0 & 0\\ & & \vdots & & \\ 0 & 0 & 0\dots & f_{n-1}'(0)-\varepsilon & -f_n'(x_n)\end{bmatrix}.$$

Thus, for any sufficiently small $\varepsilon>0$, $\mu_{1,D_\varepsilon}(J(x))$ is the maximum of the n values:

$$v_1 := f_1'(0)-f_1'(x_1)-\varepsilon,\dots,v_{n-1} := f_{n-1}'(0)-f_{n-1}'(x_{n-1})-\varepsilon,$$

and

$$v_n := \frac{kg'(x_n)-f_n'(x_n)\prod_{i=1}^{n-1}(f_i'(0)-\varepsilon)}{\prod_{i=1}^{n-1}(f_i'(0)-\varepsilon)}.$$

Since f_i' is nondecreasing, $v_i\leq-\varepsilon$ for all $i=1,\dots,n-1$. Suppose that $\prod_{i=1}^n f_i'(0) > kg'(0)$. Then since $f_i'(x_n)\geq f_i'(0)$ and $g'(x_n)\leq g'(0)$, there exists a sufficiently small $\varepsilon>0$ such that $v_n\leq-\varepsilon/2$, so $\mu_{1,D_\varepsilon}(J(x))\leq-\varepsilon/2$ for all $x\in\mathbb{R}_+^n$, and thus the system is contractive on $\mathbb{R}_+^n$ w.r.t. $|\cdot|_{1,D_\varepsilon}$.

Now assume that

$$\prod_{i=1}^n f_i'(0)=kg'(0). \tag{11.35}$$

By (11.34),

$$\det(J(x))=(-1)^n\left(\prod_{i=1}^n f_i'(x_i)-kg'(x_n)\right),$$

so (11.35) implies that $\det(J(0)) = 0$, and thus the system does not satisfy (11.2) w.r.t. any (fixed) norm on $\mathbb{R}^n_+$.

We now use Theorem 11.1 to prove that (11.14) is SOST on $\mathbb{R}^n_+$. For $\zeta \in (0, 1/2]$, let

$$\Omega_\zeta := \{x \in \mathbb{R}^n_+ : x \geq \zeta\}.$$

It is straightforward to verify that (11.14) satisfies condition (BR) in [23, Lemma 1], and this implies that for every $\tau > 0$ there exists $\varepsilon(\tau) > 0$ such that $x(t) \in \Omega_\varepsilon$ for all $t \geq \tau$. Then $g'(x_n) < g'(0)$, and $f_n'(x_n) \geq f_n'(0)$ so for any sufficiently small $\varepsilon > 0$,

$$kg'(x_n) - f_n'(x_n) \prod_{i=1}^{n-1} (f_i'(0) - \varepsilon) < kg'(0) - f_n'(0) \prod_{i=1}^{n-1} f_i'(0) = 0.$$

We already showed that this implies that there exists a $\zeta > 0$ and a norm $|\cdot|_{1,D_\zeta}$ such that (11.14) is contractive on Ω_ε w.r.t. this norm. Summarizing, all the conditions in Theorem 11.1 hold, and we conclude that (11.14) is SOST on $\mathbb{R}^n_+$ w.r.t. $|\cdot|_{1,D_0}$. □

Analysis of the system in Example 11.3. For $a \in \Omega$, let $x(t, t_1, a)$ denote the solution of (11.17) at time $t \geq t_1$ for the initial condition $x(t_1) = a$. Pick $\tau > 0$. Equation (11.17) satisfies condition (BR) in [23, Lemma 1], and this implies that there exists $\varepsilon = \varepsilon(\tau) > 0$ such that for all $a \in \Omega$, all $i = 1, \dots n$, and all $t \geq t_1 + \tau$

$$x_i(t, t_1, a) \geq \varepsilon.$$

Furthermore, if we define $y_i(t) := p_{n-i+1} - x_{n-i+1}(t)$, $i = 1, \dots, n$, then the y system also satisfies condition (BR) in [23, Lemma 1], and this implies that there exists $\varepsilon_1 = \varepsilon_1(\tau) > 0$ such that for all $a \in \Omega$, all $i = 1, \dots n$, and all $t \geq t_1 + \tau$

$$y_i(t, t_1, a) \geq \varepsilon_1.$$

We conclude that after an arbitrarily short time $\tau > 0$ every state variable $x_i(t)$, $t \geq \tau + t_1$, is separated from 0 and from p_i. This means the following. For $\zeta \in [0, 1/2]$, let

$$\Omega_\zeta := \{x \in \Omega : \zeta p_i \leq x_i \leq (1 - \zeta) p_i, \ i = 1, \dots, n\}.$$

Note that $\Omega_0 = \Omega$, and that Ω_ζ is a strict subcube of Ω for all $\zeta \in (0, 1/2]$. Then for any $t_1 \geq 0$, and any $\tau > 0$ there exists $\zeta = \zeta(\tau) \in (0, 1/2)$, with $\zeta(\tau) \to 0$ as $\tau \to 0$, such that

$$x(t, t_1, a) \in \Omega_\zeta, \quad \text{for all } t \geq t_1 + \tau \text{ and all } a \in \Omega. \tag{11.36}$$

The Jacobian of (11.17) satisfies $J(t, x) = L(x) - \operatorname{diag}(c(t), 0, \dots, 0, \eta_n)$, where

$$L(x)=\begin{bmatrix} -\eta_1(p_2-x_2) & \eta_1 x_1 & 0 & 0\\ \eta_1(p_2-x_2) & -\eta_1 x_1-\eta_2(p_3-x_3) & \dots & 0\\ 0 & \eta_2(p_3-x_3) & \dots & 0\\ & & \ddots & \\ 0 & \dots & -\eta_{n-2}x_{n-2}-\eta_{n-1}(p_n-x_n) & \eta_{n-1}x_{n-1}\\ 0 & \dots & \eta_{n-1}(p_n-x_n) & -\eta_{n-1}x_{n-1} \end{bmatrix}.$$

Note that $L(x)$ is Metzler, tridiagonal, and has zero sum columns for all $x \in \Omega$. Note also that for any $x \in \Omega_\zeta$ every entry L_{ij} on the sub- and superdiagonal of L satisfies $\zeta s_1 \le L_{ij} \le (1-\zeta)s_2$, with $s_2 := \max_i\{\eta_i p_i\} > s_1 := \min_i\{\eta_i p_i\} > 0$.

Note also that there exist $x \in \partial\Omega$ such that $J(x)$ is singular (e.g., when $x_1 = 0$ and $x_3 = p_3$ the second column of J is all zeros), and this implies that the system does not satisfy (11.2) on Ω w.r.t. any norm.

By [23, Theorem 4], for any $\zeta \in (0, 1/2]$ there exists $\varepsilon = \varepsilon(\zeta) > 0$, and a diagonal matrix $D = \text{diag}(1, q_1, q_1q_2, \dots, q_1q_2 \dots q_{n-1})$, with $q_i = q_i(\varepsilon) > 0$, such that (11.17) is contractive on Ω_ζ w.r.t. the scaled L_1 norm defined by $|z|_{1,D} := |Dz|_1$. Furthermore, we can choose ε such that $\varepsilon(\zeta) \to 0$ as $\zeta \to 0$, and $D(\varepsilon) \to I$ as $\varepsilon \to 0$. Summarizing, all the conditions in Definition 11.4 hold, so (11.17) is NC on Ω and applying Theorem 11.1 concludes the analysis. □

Proof of Proposition 11.3 Without loss of generality, assume that $S_0 = \{1, \dots, k\}$, with $1 \le k < n-1$, so that $S_- = \{k+1, \dots, n\}$. Fix $\varepsilon \in (0, 1)$. Let $D = \text{diag}(d_1, \dots, d_n)$ with the d_is defined as follows. For every $i \in S_0$, $d_i = 1$ and $d_{z(i)} = 1 - \varepsilon$. All the other d_is are one. Let $\tilde{J} := DJD^{-1}$. Then $\tilde{J}_{ij} = \frac{d_i}{d_j} J_{ij}$. We now calculate $\mu_1(\tilde{J})$. Fix $j \in S_0$. Then $d_j = 1$, so

$$\begin{aligned} c_j(\tilde{J}) &= \tilde{J}_{jj} + \sum_{\substack{1\le i\le n\\ i\ne j}} |\tilde{J}_{ij}| \\ &= J_{jj} + \sum_{\substack{i\in S_0\\ i\ne j}} d_i|J_{ij}| + \sum_{\substack{k\in S_-\\ k\ne j}} d_k|J_{kj}| \\ &= J_{jj} + \sum_{\substack{i\in S_0\\ i\ne j}} |J_{ij}| + \sum_{\substack{k\in S_-\\ k\ne j}} d_k|J_{kj}| \\ &< c_j(J), \end{aligned}$$

where the inequality follows from the fact that $d_k \le 1$ for all k, and for the specific value $k = z(j) \in S_-$ we have $d_k = 1 - \varepsilon$ and $|J_{kj}| > 0$. We conclude that for every $j \in S_0$, $c_j(\tilde{J}) < c_j(J) = 0$. It follows from property 11.3) in the statement of Proposition 11.3 and the compactness of Ω that there exists $\delta > 0$ such that $c_j(J(x)) < -\delta$ for all $j \in S_-$ and all $x \in \Omega$, so for $\varepsilon > 0$ sufficiently small we have $c_j(\tilde{J}(x)) < -\delta/2$ for all $j \in S_-$ and all $x \in \Omega$. We conclude that for all $\varepsilon > 0$ sufficiently small, $\mu_1(DJD^{-1}) = \max_j c_j(\tilde{J}) < 0$, i.e., the system is contractive w.r.t. $|\cdot|_{1,D}$. Clearly, $|\cdot|_{1,D} \to |\cdot|_1$ as $\varepsilon \to 0$, and applying Corollary 11.1 completes the proof. □

Proof of Proposition 11.5 Suppose that (11.1) is WE, and also SOST w.r.t. some norm $|\cdot|_v$. Pick $\varepsilon > 0$. Since the system is WE, there exists $\tau_0 = \tau_0(\varepsilon) > 0$ such that

$$|x(t,t_0,a)-x(t,t_0,b)|_v \le \left(1+\frac{\varepsilon}{2}\right)|a-b|_v,$$

for all $t \in [t_0, t_0+\tau_0]$. Letting $\ell_2 := \frac{1}{\tau_0}\ln(\frac{1+\varepsilon}{1+(\varepsilon/2)})$ yields

$$|x(t,t_0,a)-x(t,t_0,b)|_v \le (1+\varepsilon)\exp(-(t-t_0)\ell_2)|a-b|_v, \tag{11.37}$$

for all $t \in [t_0, t_0+\tau_0]$. It is not difficult to show that SOST implies that there exists $\ell_1 = \ell_1(\tau_0,\varepsilon) > 0$ such that

$$|x(t,t_0,a)-x(t,t_0,b)|_v \le (1+\varepsilon)\exp(-(t-t_0)\ell_1)|a-b|_v,$$

for all $t \ge t_0+\tau_0$. Combining this with (11.37) yields

$$|x(t,t_0,a)-x(t,t_0,b)|_v \le (1+\varepsilon)\exp(-(t-t_0)\ell)|a-b|_v,$$

for all $t \ge t_0$, where $\ell := \min\{\ell_1,\ell_2\} > 0$. This proves SO. □

Proof of Theorem 11.2 We require the following result.

Lemma 11.1 *If system (11.18) is IC then for each* $\tau > 0$ *there exists* $d = d(\tau) > 0$ *such that*

$$\operatorname{dist}(x(t,x_0),\partial\Omega) \ge d, \quad \textit{for all } x_0 \in \Omega \textit{ and all } t \ge \tau.$$

Proof of Lemma 11.1 Pick $\tau > 0$ and $x_0 \in \Omega$. Since Ω is an invariant set, $\operatorname{Int}(\Omega)$ is also an invariant set (see, e.g. [5, Lemma III.6]), so (11.25) implies that $x(t,x_0) \notin \partial\Omega$ for all $t > 0$. Since $\partial\Omega$ is compact, $e_{x_0} := \operatorname{dist}(x(\tau,x_0),\partial\Omega) > 0$. Thus, there exists a neighborhood U_{x_0} of x_0, such that $\operatorname{dist}(x(\tau,y),\partial\Omega) \ge e_{x_0}/2$ for all $y \in U_{x_0}$. Cover Ω by such U_{x_0} sets. By compactness of Ω, we can pick a finite subcover. Pick smallest e in this subcover, and denote this by d. Then $d > 0$ and we have that $\operatorname{dist}(x(\tau,x_0),\partial\Omega) \ge d$ for all $x_0 \in \Omega$. Now, pick $t \ge \tau$. Let $x_1 := x(t-\tau,x_0)$. Then

$$\begin{aligned}\operatorname{dist}(x(t,x_0),\partial\Omega) &= \operatorname{dist}(x(\tau,x_1),\partial\Omega)\\ &\ge d,\end{aligned}$$

and this completes the proof of Lemma 11.1. □

We can now prove Theorem 11.2. We recall some definitions from the theory of convex sets. Let $B(x,r)$ denote the closed ball of radius r around x (in the Euclidean norm). Let K be a compact and convex set with $0 \in \operatorname{Int}(K)$. Let $s(K)$ denote the *inradius* of k, i.e., the radius of the largest ball contained in K. For $\lambda \in [0, s(K)]$ the *inner parallel set of* K *at distance* λ is

$$K_{-\lambda} := \{x \in K : B(x,\lambda) \subseteq K\}.$$

Note that $K_{-\lambda}$ is a compact and convex set; in fact, $K_{-\lambda}$ is the intersection of all the translated support hyperplanes of K, with each hyperplane translated "inwards" through a distance λ (see [12, Section 17]). Assume, without loss of generality, that $0 \in \text{Int}(\Omega)$. Pick $\tau > 0$. Let $M = M(\tau) := \{x(t, x_0) : t \geq \tau,\ x_0 \in \Omega\}$. By Lemma 11.1, $M \subset \Omega$ and $\text{dist}(y, \partial\Omega) \geq d > 0$ for all $y \in M$. Let $\lambda = \lambda(\tau) := \frac{1}{2} \min \{d, s(\Omega)\}$. Then $\lambda > 0$. Pick $z \in M$. We claim that $B(z, \lambda) \subseteq \Omega$. To show this, assume that there exists $v \in B(z, \lambda)$ such that $v \notin \Omega$. Then there is a point q on the line connecting v and z such that $q \in \partial\Omega$. Therefore,

$$\begin{aligned} \text{dist}(z, \partial\Omega) &\leq |z - q| \\ &\leq |z - v| \\ &\leq \lambda \\ &\leq d/2, \end{aligned}$$

and this is a contradiction as $z \in M$. We conclude that $M \subseteq K_{-\lambda}$. Let $c = c(\tau) := \max_{x \in K_{-\lambda}} \mu(J(x))$. Then (11.26) implies that $c < 0$. Thus, the system is contractive on $K_{-\lambda}$, and for all $a, b \in \Omega$ and all $t \geq 0$

$$|x(t + \tau, a) - x(t + \tau, b)| \leq \exp(ct)|a - b|,$$

where $|\cdot|$ is the vector norm corresponding to the matrix measure μ. This establishes ST, and thus completes the proof of Theorem 11.2. □

Proof of Corollary 11.2. Since Ω is convex, compact, and invariant, it includes an equilibrium point e of (11.18). Clearly, $e \in \text{Int}(\Omega)$. By Theorem 11.2, the system is ST. Pick $a \in \Omega$ and $\tau > 0$, and let $\ell = \ell(\tau) > 0$. Applying (11.10) with $b = e$ yields

$$|x(t + \tau, a) - e| \leq \exp(-\ell t)|a - e|,$$

for all $t \geq 0$. Taking $t \to \infty$ completes the proof. □

Remark 11.3 Another possible proof of Corollary 11.2 is based on defining $V : \Omega \to \mathbb{R}_+$ by $V(x) := |x - e|$. Then for any $a \in \Omega$, $V(x(t, a))$ is nondecreasing, and the LaSalle invariance principle tells us that $x(t, a)$ converges to an invariant subset of the set $\{y \in \Omega : |y - e| = r\}$, for some $r \geq 0$. If $r = 0$ then we are done. Otherwise, pick y in the omega limit set of the trajectory. Then $y \notin \partial\Omega$, so (11.26) implies that V is strictly decreasing. This contradiction completes the proof.

Proof of Proposition 11.6. Pick $a, b \in \Omega$. Let $\gamma : [0, 1] \to \Omega$ be the line $\gamma(r) := (1 - r)a + rb$. Note that since Ω is convex, $\gamma(r) \in \Omega$ for all $r \in [0, 1]$. Let

$$w(t, r) := \frac{d}{dr} x(t, \gamma(r)).$$

This measures the sensitivity of the solution at time t to a change in the initial condition along the line γ. Note that $w(0, r) = \frac{d}{dr}\gamma(r) = b - a$, and

$$\dot{w}(t, r) = J(x(t, \gamma(r)))w(t, r).$$

Comparing this to (11.28) implies that $w(t,r)$ is equal to the second component, $\delta x(t)$, of the solution of the variational system (11.28) with initial condition

$$\begin{aligned} x(0) &= (1-r)a + rb, \\ \delta x(0) &= b - a. \end{aligned} \tag{11.38}$$

Suppose that the time-invariant system (11.18) is ST. Pick $\tau > 0$. Let $\ell = \ell(\tau) > 0$. Then for any $r \in [0,1)$ and any $\varepsilon \in [0, 1-r]$,

$$|x(t+\tau, \gamma(r+\varepsilon)) - x(t+\tau, \gamma(r))| \le \exp(-t\ell)|\gamma(r+\varepsilon) - \gamma(r)|.$$

Dividing both sides of this inequality by ε and taking $\varepsilon \downarrow 0$ implies that

$$|w(t+\tau, r)| \le \exp(-t\ell)|b-a|, \tag{11.39}$$

so

$$|\delta x(t+\tau)| \le \exp(-t\ell)|\delta x(0)|.$$

This proves the implication (a) → (b). To prove the converse implication, assume that (11.29) holds. Then (11.39) holds and thus

$$\begin{aligned} |x(t+\tau, b) - x(t+\tau, a)| &= \left| \int_0^1 \frac{d}{dr} x(t+\tau, \gamma(r)) \mathrm{d}r \right| \\ &\le \int_0^1 |w(t+\tau, r)| \, \mathrm{d}r \\ &\le \int_0^1 \exp(-\ell t)|b-a| \mathrm{d}r \\ &= \exp(-\ell t)|b-a|, \end{aligned}$$

so the system is ST. □

Above, we have used several times the fact that singularity of the Jacobian implies that the system $\dot{x} = f(x)$ cannot be contractive (as defined in 11.2) w.r.t. any (fixed) norm. For the sake of completeness, we now show this.

Pick any point $a \in \text{Int}(\Omega)$ and any fixed $\varepsilon > 0$ such that the sphere B of radius ε around a is contained in Ω. Pick any $b = a + q$, $q \in B$, and let $\gamma : [0,1] \to \Omega$ be the line $\gamma(r) := (1-r)a + rb = a + rq$. Since Ω is convex, this line is contained in Ω. Let $w(t,r) := \frac{d}{dr} x(t, \gamma(r))$. Since $\dot{w}(t,r) = J(x(t,\gamma(r)))w(t,r)$, we have that for any vector norm and for any $\tau > 0$,

$$\begin{aligned} |w(\tau, 0)| - |w(0,0)| &= |(I + \tau J(x(0, \gamma(0))) + o(\tau))w(0,0)| - |w(0,0)| \\ &= |(I + \tau J(a))q| - |q| + o(\tau)\,. \end{aligned}$$

Pick $r \in [0, 1)$, and $\varepsilon > 0$ sufficiently small. If the system is contractive then there exist a vector norm $|\cdot|$ and $\eta > 0$ such that for all $t \geq 0$,

$$|x(t, \gamma(r+\varepsilon)) - x(t, \gamma(r))| \leq \exp(-\eta t)|\gamma(r+\varepsilon) - \gamma(r)|.$$

Dividing both sides by ε and taking limits as $\varepsilon \to 0$ yields $|w(t, r)| \leq \exp(-\eta t)|q|$, for all $t \geq 0$, and all $r \in [0, 1)$. In particular,

$$|w(\tau, 0)| - |w(0, 0)| \leq (\exp(-\eta\tau) - 1)|q|.$$

Combining all this information, we have that

$$|(I + \tau J(a))q| - |q| + o(\tau) \leq (\exp(-\eta\tau) - 1)|q|$$

and therefore, dividing by $|q|$ and $\tau > 0$,

$$\frac{\frac{|(I+\tau J(a))q|}{|q|} - 1}{\tau} \leq -\eta + \frac{o(\tau)}{\tau}.$$

For each fixed τ, pick a $q = q(\tau)$ so that $\|I + \tau J(a)\| = \frac{|(I+\tau J(a))q|}{|q|}$, so the inequality gives

$$\frac{\|I + \tau J(a)\| - 1}{\tau} \leq -\eta + \frac{o(\tau)}{\tau}.$$

Taking the limit as $\tau \searrow 0$ gives that $\mu(J(a)) \leq -\eta$, where μ is the matrix measure associated to the given norm. It follows that the real part of every eigenvalue of $J(a)$ is also less than $-\eta$ [16, p. 35], so $J(a)$ is nonsingular. There remains the case when a is not in the interior of Ω. However, continuity of eigenvalues implies that the conclusion that the real part of every eigenvalue of $J(a)$ is $\leq -\eta$ is true as well.

11.6 Conclusions

Contraction theory is a powerful tool for studying nonlinear dynamical systems. Contraction implies several desirable asymptotic properties such as convergence to a unique attractor (if it exists), and entrainment to periodic excitation. This holds even if the equilibrium point or periodic attractor are not known in explicit form. However, proving contraction is in many cases nontrivial.

We considered three generalizations of contraction. These are motivated by allowing contraction to take place after an arbitrarily small transient in time and/or amplitude. In particular, this means that they have the same asymptotic properties as contractive systems. We provided checkable conditions guaranteeing that a dynamical system is a GCS, and demonstrated their usefulness by using them to analyze a number of models from systems biology. Some of these models do not satisfy (11.2), w.r.t. any (fixed) norm, yet are a GCS.

Acknowledgements We thank Zvi Artstein and George Weiss for helpful comments. EDS's work is supported in part by grants NIH 1R01GM100473, AFOSR FA9550-14-1-0060, and ONR N00014-13-1-0074. The research of MM and TT is partly supported by a research grant from the Israeli Ministry of Science, Technology and Space. The research of EDS, MM and TT is also supported by a research grant from the US-Israel Binational Science Foundation. We thank an anonymous reviewer for carefully reading this chapter and providing us several useful comments.

References

1. Aminzare, Z., Sontag, E.D.: Logarithmic lipschitz norms and diffusion-induced instability. Nonlinear Anal. Theory Methods Appl. **83**, 31–49 (2013)
2. Aminzare, Z., Sontag, E.D.: Contraction methods for nonlinear systems: a brief introduction and some open problems. In: Proceedings of 53rd IEEE Conference on Decision and Control, Los Angeles, CA, pp. 3835–3847 (2014)
3. Andrieu, V., Jayawardhana, B., Praly, L.: On transverse exponential stability and its use in incremental stability, observer and synchronization. In: Proceedings of 52nd IEEE Conference on Decision and Control, Florence, Italy, pp. 5915–5920 (2013)
4. Angeli, D.: A Lyapunov approach to incremental stability properties. IEEE Trans. Autom. Control **47**, 410–421 (2002)
5. Angeli, D., Sontag, E.D.: Monotone control systems. IEEE Trans. Autom. Control **48**, 1684–1698 (2003)
6. Arcak, M.: Certifying spatially uniform behavior in reaction-diffusion PDE and compartmental ODE systems. Automatica **47**(6), 1219–1229 (2011)
7. Arcak, M., Sontag, E.D.: Diagonal stability of a class of cyclic systems and its connection with the secant criterion. Automatica **42**(9), 1531–1537 (2006)
8. Aylward, E.M., Parrilo, P.A., Slotine, J.J.E.: Stability and robustness analysis of nonlinear systems via contraction metrics and SOS programming. Automatica **44**(8), 2163–2170 (2008)
9. Blythe, R.A., Evans, M.R.: Nonequilibrium steady states of matrix-product form: a solver's guide. J. Phys. A Math. Theor. **40**(46), R333–R441 (2007)
10. Bonnabel, S., Astolfi, A., Sepulchre, R.: Contraction and observer design on cones. In: Proceedings of 50th IEEE Conf. on Decision and Control and European Control Conference, Orlando, FL, pp. 7147–7151 (2011)
11. Border, K.C.: Fixed Point Theorems with Applications to Economics and Game Theory. Cambridge University Press (1985)
12. Chakerian, G.D., Sangwine-Yager, J.R.: Synopsis and exercises for the theory of convex sets (2009). https://www.math.ucdavis.edu/deloera/TEACHING/MATH114/
13. Csikasz-Nagy, A., Cardelli, L., Soyer, O.S.: Response dynamics of phosphorelays suggest their potential utility in cell signaling. J. Roy. Soc. Interface **8**, 480–488 (2011)
14. Del Vecchio, D., Ninfa, A.J., Sontag, E.D.: Modular cell biology: retroactivity and insulation. Mol. Syst. Biol. **4**(1), 161 (2008)
15. Desoer, C., Haneda, H.: The measure of a matrix as a tool to analyze computer algorithms for circuit analysis. IEEE Trans. Circuit Theory **19**, 480–486 (1972)
16. Desoer, C., Vidyasagar, M.: Feedback Synthesis: Input-Output Properties. SIAM, Philadelphia, PA (2009)
17. Dorfler, F., Bullo, F.: Synchronization and transient stability in power networks and nonuniform Kuramoto oscillators. SIAM J. Control Optim. **50**, 1616–1642 (2012)

18. Forni, F., Sepulchre, R.: A differential Lyapunov framework for contraction analysis. IEEE Trans. Autom. Control **59**(3), 614–628 (2014)
19. Jacquez, J.A., Simon, C.P.: Qualitative theory of compartmental systems. SIAM Rev. **35**(1), 43–79 (1993)
20. Jouffroy, J.: Some ancestors of contraction analysis. In: Proceedings of 44th IEEE Conference on Decision and Control, Seville, Spain, pp. 5450–5455 (2005)
21. Lohmiller, W., Slotine, J.J.E.: On contraction analysis for non-linear systems. Automatica **34**, 683–696 (1998)
22. Lohmiller, W., Slotine, J.J.E.: Control system design for mechanical systems using contraction theory. IEEE Trans. Autom. Control **45**, 984–989 (2000)
23. Margaliot, M., Sontag, E.D., Tuller, T.: Entrainment to periodic initiation and transition rates in a computational model for gene translation. PLoS ONE **9**(5), e96,039 (2014)
24. Margaliot, M., Sontag, E.D., Tuller, T.: Contraction after small transients. Automatica **67**, 178–184 (2016)
25. Margaliot, M., Tuller, T.: On the steady-state distribution in the homogeneous ribosome flow model. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**, 1724–1736 (2012)
26. Margaliot, M., Tuller, T.: Stability analysis of the ribosome flow model. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**, 1545–1552 (2012)
27. Margaliot, M., Tuller, T.: Ribosome flow model with positive feedback. J. Roy. Soc. Interface **10**, 20130, 267 (2013)
28. Pham, Q.C., Tabareau, N., Slotine, J.J.: A contraction theory approach to stochastic incremental stability. IEEE Trans. Autom. Control **54**, 816–820 (2009)
29. Poker, G., Zarai, Y., Margaliot, M., Tuller, T.: Maximizing protein translation rate in the non-homogeneous ribosome flow model: a convex optimization approach. J. Roy. Soc. Interface **11**(100), 20140, 713 (2014)
30. Raveh, A., Margaliot, M., Sontag, E.D., Tuller, T.: A model for competition for ribosomes in the cell. J. Roy. Soc. Interface **13**(116) (2016)
31. Raveh, A., Zarai, Y., Margaliot, M., Tuller, T.: Ribosome flow model on a ring. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**(6), 1429–1439 (2015)
32. Reuveni, S., Meilijson, I., Kupiec, M., Ruppin, E., Tuller, T.: Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput. Biol. **7**, e1002, 127 (2011)
33. Rüffer, B.S., van de Wouw, N., Mueller, M.: Convergent systems versus incremental stability. Syst. Control Lett. **62**, 277–285 (2013)
34. Russo, G., Di Bernardo, M., Sontag, E.D.: Global entrainment of transcriptional systems to periodic inputs. PLoS Comput. Biol. **6**, e1000, 739 (2010)
35. Russo, G., di Bernardo, M., Sontag, E.D.: A contraction approach to the hierarchical analysis and design of networked systems. IEEE Trans. Autom. Control **58**, 1328–1331 (2013)
36. Sandberg, I.W.: On the mathematical foundations of compartmental analysis in biology, medicine, and ecology. IEEE Trans. Circuits Syst. **25**(5), 273–279 (1978)
37. Simpson-Porco, J.W., Bullo, F.: Contraction theory on Riemannian manifolds. Syst. Control Lett. **65**, 74–80 (2014)
38. Slotine, J.J.E.: Modular stability tools for distributed computation and control. Int. J. Adapt. Control Signal Process. **17**, 397–416 (2003)
39. Smith, H.L.: Monotone Dynamical Systems: an Introduction to the Theory of Competitive and Cooperative Systems, Mathematical Surveys and Monographs, vol. 41. American Mathematical Society, Providence, RI (1995)
40. Soderlind, G.: The logarithmic norm. history and modern theory. BIT Numer. Math. **46**, 631–652 (2006)
41. Sontag, E.D.: Mathematical control theory: deterministic finite-dimensional systems. In: Texts in Applied Mathematics, vol. 6, 2 edn. Springer, New York (1998)
42. Sontag, E.D., Margaliot, M., Tuller, T.: On three generalizations of contraction. In: Proceedings of 53rd IEEE Conference on Decision and Control, pp. 1539–1544. Los Angeles, CA (2014)
43. Vidyasagar, M.: Nonlinear Systems Analysis. Prentice Hall, Englewood Cliffs, NJ (1978)

44. Wang, W., Slotine, J.J.: On partial contraction analysis for coupled nonlinear oscillators. Biol. Cybern. **92**, 38–53 (2005)
45. Zamani, M., van de Wouw, N., Majumdar, R.: Backstepping controller synthesis and characterizations of incremental stability. Syst. Control Lett. **62**(10), 949–962 (2013)
46. Zarai, Y., Margaliot, M., Tuller, T.: Explicit expression for the steady state translation rate in the infinite-dimensional homogeneous ribosome flow model. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**, 1322–1328 (2013)

Chapter 12
Asymptotic Expansions of Laplace Integrals for Quantum State Tomography

Pierre Six and Pierre Rouchon

Abstract Bayesian estimation of a mixed quantum state can be approximated via maximum likelihood (MaxLike) estimation when the likelihood function is sharp around its maximum. Such approximations rely on asymptotic expansions of multi-dimensional Laplace integrals. When this maximum is on the boundary of the integration domain, as is the case when the MaxLike quantum state is not full rank, such expansions are not standard. We provide here such expansions, even when this maximum does not lie on the smooth part of the boundary, as in the case when the rank deficiency exceeds two. Aside from the MaxLike estimate of the quantum state, these expansions provide confidence intervals for any observable. They confirm the formula proposed and used without precise mathematical justifications by the authors in an article recently published in Physical Review A.

12.1 Introduction

When the probability laws of observed data Y with respect to a continuous parameter p to be estimated are given by an analytic model, the Maximum Likelihood (MaxLike) reconstruction method is widely used (see, e.g., [1]). It chooses an estimate of p, denoted by p_{ML}, the value of p that maximizes the conditional probability $\mathbb{P}(Y \mid p)$ of the data Y. Indeed, when the data Y consists of a large amount of independent measurements, the function $p \mapsto \mathbb{P}(Y \mid p)$ becomes extremely sharp around its maximal value, and the MaxLike estimate p_{ML} is a good approximation of the Bayesian mean estimate denoted by p_{BM}:

P. Six · P. Rouchon (✉)
Centre Automatique et Systèmes, Mines-ParisTech, PSL Research University,
60 Bd Saint-Michel, 75006 Paris, France
e-mail: pierre.rouchon@mines-paristech.fr

P. Six
e-mail: pierre.six@mines-paristech.fr

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_12

$$p_{BM} = \int_{\mathscr{D}} p\, \mathbb{P}\,(p \mid Y)\; \mathrm{d}p = \frac{\int_{\mathscr{D}} p\, \mathbb{P}\,(Y \mid p)\; \mathbb{P}_0(p)\, \mathrm{d}p}{\int_{\mathscr{D}} \mathbb{P}\,(Y \mid p)\; \mathbb{P}_0(p)\, \mathrm{d}p},$$

where $\mathscr{D} \subset \mathbb{R}^{\dim p}$ is a set of physically acceptable values for p; $\mathbb{P}\,(p \mid Y)$ is the probability density of p knowing Y; and $\mathbb{P}_0(p)$ is any a priori probability density for p.

Reliance only on MaxLike estimation has the advantage of providing easy-to-compute algorithms. The first and second derivatives of $\mathbb{P}\,(Y \mid p)$ versus p can be derived using the finite difference method, and gradient-like optimization methods can be used. The Cramér–Rao bound can also be extracted from the Hessian of the log-likelihood function to define a lower bound of the mean estimation error when this Hessian matrix is not degenerate. Nevertheless, some technical problems can arise, in particular for quantum state tomography [2], where the estimated value of parameter p corresponds to a quantum state ρ, an element of the compact convex domain $\mathscr{D}$ formed by the set of non-negative Hermitian matrices of trace one. In practice, MaxLike estimates of ρ_{ML} may be of low rank, for instance on the boundary of $\mathscr{D}$ as noted in [3] and observed in [4].

All these reasons have led us to consider Bayesian Mean Estimations (BME) in the general setting when the parameter p exists in a finite-dimensional and compact domain $\mathscr{D}$ with piecewise smooth boundary. As the magnitude of $\mathbb{P}\,(Y \mid p)$ grows (or decreases) at an exponentially high rate compared to the number of independent measurements N generating the measurement set Y, we consider the scaled log-likelihood function $f(p) = \frac{1}{N} \log\left(\mathbb{P}\,(Y \mid p)\right)$. We then address the problem of computing the asymptotic development when N tends toward infinity, for any smooth scalar functions f and g and under various conditions of the Laplace's integral:

$$\mathscr{I}_g(N) = \int_{\mathscr{D}} g(p) \exp\left(Nf(p)\right)\, \mathrm{d}p. \tag{12.1}$$

Such asymptotic expansions, which have long been investigated, involve integration by parts, Watson's lemma, Laplace's method, stationary phase, steepest descents, and Hironaka's resolution of singularities: see [5] for $\dim p = 1$; and the regular case when $\dim p \geq 1$; also see [6] for the singular case in arbitrary dimensions and its much more elaborate analysis. In the analytic case and around the maximum of f at p_{ML} inside domain $\mathscr{D}$, these expansions rely on terms such as $e^{Nf(p_{ML})} \frac{(\log N)^k}{N^\alpha}$, where k is a non-negative integer less than $\dim p - 1$ and where α is rational and strictly positive [6, p. 231]. Fundamental connections between algebraic geometry and statistical learning theory in the singular case stem from such series expansions, for example, when the Hessian of f at p_{ML} is not negative definite. This is the object of singular learning theory developed in [7, 8].

It is interesting to note that, as far as we know, very few results can be found when p_{ML} lies on the boundary of $\mathscr{D}$, except in the case when p_{ML} is on a smooth part of the boundary. In [5, Sect. 8.3], the derivation of the leading term is explained when p_{ML} is on the smooth part of the boundary and when the Hessian of the restriction of f to this smooth part is negative definite; Sect. 8.3.4 of [6] provides precise indications showing that, when the Hessian of the restriction of f is degenerate, an asymptotic expansion exists which is similar to that obtained for p_{ML} in the interior of $\mathscr{D}$.

For quantum state estimation, this ensures the existence of asymptotic expansion in any case when ρ_{ML} has either a full rank (interior of $\mathscr{D}$) or rank deficiency of one (smooth part of the boundary of $\mathscr{D}$). For rank deficiency strictly exceeding one, ρ_{ML} does not lie on the smooth part of the boundary. As far as we know, the derivations of asymptotic expansions in these singular cases, when the rank deficiency of ρ_{ML} exceeds two, have not been precisely addressed before now. This paper is a first attempt at deriving such asymptotic expansion of the Bayesian mean and variance when the log-likelihood function reaches it maximum on the boundary of $\mathscr{D}$, that is, when ρ_{ML} is of low rank.

The goal of this paper is twofold. First, we provide the leading terms of specific asymptotic expansions when p_{ML} lies in a half-space. This is the object of Sect. 12.2, where we assume that restricting f to the boundary admits a non-degenerate maximum at p_{ML} (see Theorem 12.2). Second, we consider quantum state estimation and reformulate these leading terms intrinsically in terms of operators and trace. This is the object of Sect. 12.3, where we recall the precise structures of f and g in this case and exploit convexity and unitary invariance. We provide in this section precise mathematical justifications of the necessary and sufficient optimality conditions given without details in [4, Eq. (8)] (see Lemma 12.2 below) and of the Bayesian variance approximation corresponding to Eq. (10) in [4] (see Theorem 12.3).

12.2 Asymptotic Expansion of Laplace's Integral

Here, we assume that p is of dimension n and that $\mathscr{D} = (-1, 1)^n$. Set $p = z$ with $z \in \mathbb{R}^n$. Then (12.1) may be written

$$\mathscr{I}_g(N) = \int_{z\in(-1,1)^n} g(z)\exp\left(Nf(z)\right)\,\mathrm{d}z. \tag{12.2}$$

Theorem 12.1 *Consider (12.2) where f and g are analytic functions of z on a compact neighbourhood of $\overline{\mathscr{D}}$, the closure of $\mathscr{D}$. Assume that f admits a unique maximum on $\overline{\mathscr{D}}$ at $z = 0$ with $\left.\frac{\partial^2 f}{\partial z^2}\right|_0$ negative definite.*

If $g(0) \neq 0$, we have the following dominant term in the asymptotic expansion of $\mathscr{I}_g(N)$ for large N

$$\mathscr{I}_g(N) = \left(\frac{g(0)\,(2\pi)^{n/2}\,e^{Nf(0)}N^{-n/2}}{\sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0\right)\right|}}\right) + O\Big(e^{Nf(0)}N^{-n/2-1}\Big). \tag{12.3}$$

If $g(0) = 0$, *with* $\left.\frac{\partial g}{\partial z}\right|_0 = 0$, *then we have*

$$\mathscr{I}_g(N) = \left(\frac{Tr\left(- \left.\frac{\partial^2 g}{\partial z^2}\right|_0 \left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0 \right)^{-1} \right) (2\pi)^{n/2}}{2\sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0 \right)\right|}} \right) e^{Nf(0)} N^{-n/2-1} + O\left(e^{Nf(0)} N^{-n/2-2} \right). \quad (12.4)$$

Proof Since f is analytic, $f(z) = f(0) - h(z)$ where h is an analytic function of z only with $h(0) = 0$, $\left.\frac{\partial h}{\partial z}\right|_0 = 0$ and $\left.\frac{\partial^2 h}{\partial z^2}\right|_0 = - \left.\frac{\partial^2 f}{\partial z^2}\right|_0$ positive definite.

Via the Morse lemma (see, e.g., [9]), there exists a local diffeomorphism on z around 0, written $\tilde{z} = \psi(z)$, such that $\psi(0) = 0$ and $h(z) = \frac{1}{2}\sum_{k=1}^n (\psi_k(z))^2$. Moreover, we can choose ψ such that $\left.\frac{\partial \psi}{\partial z}\right|_0 = \sqrt{- \left.\frac{\partial^2 f}{\partial z^2}\right|_0}$ is a positive definite symmetric matrix.

For $\eta \in (0, 1)$ small, there exists a $c < f(0)$ such that, $\forall z \in (-1, 1)^n/(-\eta, \eta)^n$, $f(z) \leq c$. Since

$$\begin{aligned}\mathscr{I}_g(N) &= \int_{z\in(-\eta,\eta)^n} g(z)e^{Nf(z)}\,\mathrm{d}z + \int_{z\in(-1,1)^n/(-\eta,\eta)^n} g(z)e^{Nf(z)}\,\mathrm{d}z \\ &= e^{Nf(0)} \left(\int_{z\in(-\eta,\eta)^n} g(z)e^{N(f(z)-f(0))}\,\mathrm{d}z + O\left(e^{-N(f(0)-c)}\right) \right),\end{aligned}$$

we keep only

$$I_\eta(N) = \int_{z\in(-\eta,\eta)^n} g(z)e^{N(f(z)-f(0))}\,\mathrm{d}z.$$

Since η is small, we can consider the change of variable $\tilde{z} = \psi(z)$ that yields:

$$I_\eta(N) = \int_{\tilde{z}\in\psi((-\eta,\eta)^n)} \tilde{g}(\tilde{z})e^{-\frac{N}{2}\sum_{k=1}^n \tilde{z}_k^2}\,\mathrm{d}\tilde{z},$$

where

$$\tilde{g}(\tilde{z}) = \frac{g(\psi^{-1}(\tilde{z}))}{\sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0 \right)\right|}} (1 + \tilde{d}(\tilde{z})) \quad (12.5)$$

and $\tilde{d}$ is an analytic function with $\tilde{d}(0) = 0$. There exists $\tilde{\eta} > 0$ such that $(-\tilde{\eta}, \tilde{\eta})^n \subset \psi((-\eta, \eta)^n)$. Thus, as in the passage from $\mathscr{I}_g(N)$ to $I_\eta(N)$, then up to exponentially small terms versus N, we consider only the asymptotic expansion of

$$\tilde{I}_{\tilde{\eta}} = \int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \tilde{g}(\tilde{z})e^{-\frac{N}{2}\sum_{k=1}^n \tilde{z}_k^2}\,\mathrm{d}\tilde{z}. \quad (12.6)$$

When $g(0) \neq 0$, we have $\tilde{g}(0) \neq 0$. Setting $\tilde{g}(\tilde{z}) = \tilde{g}(0) + \sum_{k=1}^{n} \tilde{z}_k \tilde{h}_k(\tilde{z})$ with $\tilde{h}_k$ bounded analytic functions on $(-\tilde{\eta}, \tilde{\eta})^n$, we obtain

$$\tilde{I}_{\tilde{\eta}} = \tilde{g}(0) \int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} e^{-\frac{N}{2}\sum_{k=1}^{n}\tilde{z}_k^2} \, \mathrm{d}\tilde{z} + \int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \left(\sum_{k=1}^{n} \tilde{z}_k \tilde{h}_k(\tilde{z}) \right) e^{-\frac{N}{2}\sum_{k=1}^{n}\tilde{z}_k^2} \, \mathrm{d}\tilde{z}.$$

Up to exponentially small terms versus N, the first integral on the right-hand side can be replaced by

$$\int_{\tilde{z}\in(-\infty,+\infty)^n} e^{-\frac{N}{2}\sum_{k=1}^{n}\tilde{z}_k^2} \, \mathrm{d}\tilde{z} = \left(\frac{2\pi}{N} \right)^{n/2}.$$

A single integration by parts versus z_k yields

$$\int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \tilde{z}_k \tilde{h}_k(\tilde{z}) e^{-\frac{N}{2}\sum_{k=1}^{n}\tilde{z}_k^2} \, \mathrm{d}\tilde{z}$$
$$= \frac{1}{N} \int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \frac{\partial \tilde{h}_k}{\partial \tilde{z}_k}(\tilde{z}) e^{-\frac{N}{2}\sum_{k=1}^{n}\tilde{z}_k^2} \, \mathrm{d}\tilde{z} + O(e^{-\tilde{\eta}^2 N/2}/N).$$

This implies (12.3), via $\tilde{I}_{\tilde{\eta}} = \tilde{g}(0) \left(\frac{2\pi}{N} \right)^{n/2} (1 + O(1/N))$ and $\tilde{g}(0) = \frac{g(0)}{\sqrt{\left| \det\left(\frac{\partial^2 f}{\partial z^2}\Big|_0 \right) \right|}}$.

Assuming now that $g(0) = 0$ and $\frac{\partial g}{\partial z}\Big|_0 = 0$, and considering then the function $\tilde{g}$ in (12.5), we have $\tilde{g}(0) = 0$ and $\frac{\partial \tilde{g}}{\partial \tilde{z}}\Big|_0 = 0$. Moreover, writing

$$\kappa_0 = \sqrt{\left| \det\left(\frac{\partial^2 f}{\partial z^2}\Big|_0 \right) \right|},$$

we have

$$\kappa_0 \tilde{g}(\psi(z)) = g(z)\,(1 + e(z))$$

where e is an analytic function such that $e(0) = 0$. Thus, for any $i, j \in \{1, \dots, n\}$,

$$\frac{\partial^2 g}{\partial z_i \partial z_j}\Big|_0 = \kappa_0 \sum_{k,k'=1}^{n} \frac{\partial^2 \tilde{g}}{\partial \tilde{z}_k \partial \tilde{z}_{k'}}\Big|_0 \frac{\partial \psi_k}{\partial z_i}\Big|_0 \frac{\partial \psi_{k'}}{\partial z_i}\Big|_0.$$

Since $\frac{\partial \psi}{\partial z}\Big|_0 = \sqrt{-\frac{\partial^2 f}{\partial z^2}\Big|_0}$, we have

$$\kappa_0 \left.\frac{\partial^2 \tilde{g}}{\partial \tilde{z}^2}\right|_0 = \left(\sqrt{-\left.\frac{\partial^2 f}{\partial z^2}\right|_0}\right)^{-1} \left.\frac{\partial^2 g}{\partial z^2}\right|_0 \left(\sqrt{-\left.\frac{\partial^2 f}{\partial z^2}\right|_0}\right)^{-1},$$

and thus

$$\mathrm{Tr}\left(\left.\frac{\partial^2 \tilde{g}}{\partial \tilde{z}^2}\right|_0\right) = \frac{\mathrm{Tr}\left(-\left.\frac{\partial^2 g}{\partial z^2}\right|_0 \left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0\right)^{-1}\right)}{\sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0\right)\right|}} \tag{12.7}$$

Since $\tilde{g}$ and its first partial derivatives with respect to $\tilde{z}_k$ vanish, we have

$$\tilde{g}(\tilde{z}) = \sum_{k,k'=1}^{n} \tilde{z}_k \tilde{z}_{k'} \tilde{b}_{k,k'}(\tilde{z}),$$

where the function $\tilde{b}_{k,k'}$ is analytic. To evaluate the integral in (12.6), we have to consider the dominant terms of the following integrals:

$$B_{k,k'} = \int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \tilde{z}_k \tilde{z}_{k'} \tilde{b}_{k,k'}(\tilde{z}) e^{-\frac{N}{2}\sum_{l=1}^{n} \tilde{z}_l^2}\, \mathrm{d}\tilde{z}.$$

For $k \neq k'$, one integration by parts versus $\tilde{z}_k$ followed by another versus $\tilde{z}_{k'}$ yield $B_{k,k'} = O\left(N^{-n/2-2}\right)$. For $k = k'$, we can perform a single integration by parts versus $\tilde{z}_k$:

$$\begin{aligned}
&\int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \tilde{z}_k^2 \tilde{b}_{k,k}(\tilde{z}) e^{-\frac{N}{2}\sum_{l=1}^{n} \tilde{z}_l^2}\, \mathrm{d}\tilde{z} \\
&\qquad = \frac{1}{N}\int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \left(\tilde{b}_{k,k}(\tilde{z}) + \tilde{z}_k \frac{\partial \tilde{b}_{k,k}}{\partial \tilde{z}_k}(\tilde{z})\right) e^{-\frac{N}{2}\sum_{l=1}^{n} \tilde{z}_l^2}\, \mathrm{d}\tilde{z} + O(e^{-N\tilde{\eta}^2/2}) \\
&\qquad\qquad = \frac{\tilde{b}_{k,k}(0)}{N}\left(\frac{2\pi}{N}\right)^{n/2} + O\left(N^{-n/2-2}\right).
\end{aligned}$$

The sum $\sum_{k,k'} B_{k,k'}$ corresponds to the integral $\tilde{I}_{\tilde{\eta}}$ and becomes

$$\tilde{I}_{\tilde{\eta}}(N) = \frac{\sum_{k=1}^{n} \tilde{b}_{k,k}(0)}{N}\left(\frac{2\pi}{N}\right)^{n/2} + O\left(N^{-n/2-2}\right).$$

Since $\tilde{I}_{\tilde{\eta}}$ and $e^{-Nf(0)}\mathscr{I}_g(N)$ coincide up to exponentially small terms, we obtain (12.4) using (12.7) since $\sum_{k=1}^{n} \tilde{b}_{k,k}(0) = \frac{1}{2}\mathrm{Tr}\left(\left.\frac{\partial^2 \tilde{g}}{\partial \tilde{z}^2}\right|_0\right)$. ■

We assume now that $p \in \mathbb{R}^{n+1}$, $n+1$ being the dimension of p (n non-negative integers), and that $\mathscr{D} = (0,1)\times(-1,1)^n$. Set $p = (x,z)$ with $x \in \mathbb{R}$ and $z \in \mathbb{R}^n$. Then, when $g(x,z)$ is replaced by $x^m g(x,z)$, with m a non-negative integer, (12.1) becomes

$$\mathscr{I}_g(N) = \int_{x\in(0,1)} \int_{z\in(-1,1)^n} x^m g(x,z) \exp\left(Nf(x,z)\right) \,\mathrm{d}x\, \mathrm{d}z. \tag{12.8}$$

Theorem 12.2 *Consider (12.8), where f and g are analytic functions of* (x, z) *on a compact neighbourhood of* $\overline{\mathscr{D}}$*, the closure of* $\mathscr{D}$*. Assume that f admits a unique maximum on* $\overline{\mathscr{D}}$ *at* $(x, z) = (0, 0)$*, with* $\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}$ *negative definite and* $\frac{\partial f}{\partial x}\Big|_{(0,0)} < 0$.

If $g(0, 0) \neq 0$*, we have the following dominant term in the asymptotic expansion of* $\mathscr{I}_g(N)$ *for large N:*

$$\mathscr{I}_g(N) = \left(\frac{g(0,0)\; m!\; (2\pi)^{n/2}\; e^{Nf(0,0)} N^{-m-n/2-1}}{\sqrt{\left|\det\left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)\right|}\; \left(-\frac{\partial f}{\partial x}\Big|_{(0,0)}\right)^{m+1}} \right) + O\Big(e^{Nf(0,0)} N^{-m-n/2-2}\Big). \tag{12.9}$$

If $g(0, 0) = 0$*, with* $\frac{\partial g}{\partial x}\Big|_{(0,0)} = 0$ *and* $\frac{\partial g}{\partial z}\Big|_{(0,0)} = 0$*, we have*

$$\mathscr{I}_g(N) = \left(\frac{Tr\left(-\frac{\partial^2 g}{\partial z^2}\Big|_{(0,0)} \left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)^{-1}\right)\; m!\; (2\pi)^{n/2}}{2\sqrt{\left|\det\left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)\right|}\; \left(-\frac{\partial f}{\partial x}\Big|_{(0,0)}\right)^{m+1}} \right) e^{Nf(0,0))} N^{-m-n/2-2} + O\left(e^{Nf(0,0))} N^{-m-n/2-3}\right). \tag{12.10}$$

For clarity, we consider here only the analytic situation, despite the fact that the above asymptotics are also valid in the C^{m+3} case.

Proof We adapt here the method sketched in Sect. 8.3.4 of [6] for oscillatory integrals in a half-space. Since f is analytic, we have

$$f(x, z) = f(0, 0) - xf_1(x, z) - h(z),$$

where f_1 is analytic with $f_1(0, 0) = -\frac{\partial f}{\partial x}\Big|_{(0,0)} > 0$ and h is an analytic function of z only, with $h(0) = 0$, $\frac{\partial h}{\partial z}\Big|_0 = 0$ and $\frac{\partial^2 h}{\partial z^2}\Big|_0 = -\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}$ positive definite.

Set $\phi(x, z) = xf_1(x, z)$. Consider the following map $(x, z) \mapsto (\tilde{x} = \phi(x, z), z)$. It is a local diffeomorphism around $(0, 0)$ that preserves the sign of x, i.e., $x\phi(x, z) \geq 0$. Moreover, using the Morse lemma (see, e.g., [9]), there exists a local diffeomorphism on z around 0, $\tilde{z} = \psi(z)$, such that $\psi(0) = 0$ and $h(z) = \frac{1}{2}\sum_{k=1}^{n}(\psi_k(z))^2$ (see, e.g., [9]). Moreover, we can choose ψ such that $\frac{\partial \psi}{\partial z}\Big|_0 = \sqrt{-\frac{\partial^2 f}{\partial z^2}\Big|_0}$ is a positive definite symmetric matrix.

To summarize, there is a local analytic diffeomorphism $\Xi : V \ni (x,z) \mapsto (\tilde{x},\tilde{z}) \in \tilde{V}$ from an open connected neighbourhood V of 0 to another open connected neighbourhood of 0 such that

- for all $(x,z) \in V$, we have $\phi(x,z) > 0$ (resp. $< 0, = 0$) when $x > 0$ (resp. $< 0, = 0$).
- $\forall (x,z) \in V, f(x,z) = -\phi(x,z) - \frac{1}{2}\sum_{k=1}^{n}(\psi_k(z))^2$.
- $\det\left.\begin{pmatrix} \frac{\partial\phi}{\partial x} & \frac{\partial\phi}{\partial z} \\ \frac{\partial\psi}{\partial x} & \frac{\partial\psi}{\partial z}\end{pmatrix}\right|_{(x,z)} = \left|\left.\frac{\partial f}{\partial x}\right|_{(0,0)}\right| \sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_{(0,0)}\right)\right|}(1 + d(x,z))$ where d is analytic on V with $d(0,0) = 0$.

Since V is a neighbourhood of 0, there exists a $\eta \in (0,1)$ such that $\mathscr{C}_\eta = (0,\eta) \times (-\eta,\eta)^n \subset V$. Moreover, there exists $c < f(0,0)$ such that, $\forall (x,z) \in \mathscr{D}/\mathscr{C}_\eta, f(x,z) \leq c$. Since

$$\mathscr{I}_g(N) = \int_{(x,z)\in\mathscr{C}_\eta} x^m g(x,z) e^{Nf(x,z)}\,\mathrm{d}x\,\mathrm{d}z + \int_{(x,z)\in\mathscr{D}/\mathscr{C}_\eta} x^m g(x,z)e^{Nf(x,z)}\,\mathrm{d}x\,\mathrm{d}z$$

$$= e^{Nf(0,0)}\left(\int_{(x,z)\in\mathscr{C}_\eta} x^m g(x,z)e^{N(f(x,z)-f(0,0))}\,\mathrm{d}x\,\mathrm{d}z + e^{-N(f(0,0)-c)}\int_{(x,z)\in\mathscr{D}/\mathscr{C}_\eta} x^m g(x,z)e^{N(f(x,z)-c)}\,\mathrm{d}x\,\mathrm{d}z\right)$$

$$= e^{Nf(0,0)}\left(\int_{(x,z)\in\mathscr{C}_\eta} x^m g(x,z)e^{N(f(x,z)-f(0,0))}\,\mathrm{d}x\,\mathrm{d}z + O\left(e^{-N(f(0,0)-c)}\right)\right)$$

we have to consider only the asymptotic expansion of

$$I_\eta(N) = \int_{(x,z)\in\mathscr{C}_\eta} x^m g(x,z)e^{N(f(x,z)-f(0,0))}\,\mathrm{d}x\,\mathrm{d}z.$$

Since $\mathscr{C}_\eta \subset V$, we can consider the change of variable $(\tilde{x},\tilde{z}) = \Xi(x,z)$ that yields

$$I_\eta(N) = \int_{(\tilde{x},\tilde{z})\in\Xi(\mathscr{C}_\eta)} \tilde{x}^m \tilde{g}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z},$$

where

$$\tilde{g}(\tilde{x},\tilde{z}) = \frac{g(\Xi^{-1}(\tilde{x},\tilde{z}))}{\left(f_1(\Xi^{-1}(\tilde{x},\tilde{z}))\right)^m \left|\left.\frac{\partial f}{\partial x}\right|_{(0,0)}\right| \sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_{(0,0)}\right)\right|}}\,(1 + \tilde{d}(\tilde{x},\tilde{z}))$$

and $\tilde{d}$ is an analytic function with $\tilde{d}(0,0) = 0$. Since, for all $(\tilde{x},\tilde{z}) \in \Xi(\mathscr{C}_\eta)$ we have $\tilde{x} \geq 0$, there exists an $\tilde{\eta} > 0$ such that $\widetilde{\mathscr{C}_{\tilde{\eta}}} = (0,\tilde{\eta}) \times (-\tilde{\eta},\tilde{\eta})^n \subset \Xi(\mathscr{C}_\eta)$. Thus, as in the passage from $\mathscr{I}_g(N)$ to $I_\eta(N)$, up to exponentially small terms versus N we consider only the asymptotic expansion of

$$\tilde{I}_{\tilde{\eta}} = \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} \tilde{x}^m \tilde{g}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}. \tag{12.11}$$

When $g(0,0) \neq 0$, we have $\tilde{g}(0,0) \neq 0$. Set $\tilde{g}(\tilde{x},\tilde{z}) = \tilde{g}(0,0) + \tilde{x}\tilde{g}_1(\tilde{x},\tilde{z}) + \sum_{k=1}^{n} \tilde{z}_k \tilde{h}_k(\tilde{x},\tilde{z})$ with $\tilde{g}_1$ and $\tilde{h}_k$ bounded analytic functions on $\widetilde{\mathscr{C}_{\tilde{\eta}}}$. We obtain

$$\begin{aligned}\tilde{I}_{\tilde{\eta}} = \tilde{g}(0,0) \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} &\tilde{x}^m e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z} \\ &+ \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} \tilde{x}^{m+1}\tilde{g}_1(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z} \\ &+ \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} \tilde{x}^m \left(\sum_{k=1}^{n} \tilde{z}_k \tilde{h}_k(\tilde{x},\tilde{z})\right) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z}.\end{aligned}$$

Up to exponentially small terms versus N, the first integral on the right-hand side can be replaced by

$$\int_{(\tilde{x},\tilde{z})\in(0,+\infty)\times(-\infty,+\infty)^n} \tilde{x}^m e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z} = \frac{m!}{N^{m+1}} \left(\frac{2\pi}{N}\right)^{n/2}.$$

For the second integral, $m+1$ integrations by parts versus $\tilde{x}$ are necessary

$$\begin{aligned}&\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} \tilde{x}^{m+1}\tilde{g}_1(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z} \\ &\qquad = \int_{\tilde{z}\in(-\tilde{\eta},\tilde{\eta})^n} \left(\int_0^{\tilde{\eta}} \tilde{x}^{m+1}\tilde{g}_1(\tilde{x},\tilde{z}) e^{-N\tilde{x}}\, \mathrm{d}\tilde{x}\right) e^{-\frac{N}{2}\sum_{k=1}^{n}\tilde{z}_k^2}\, \mathrm{d}\tilde{z},\end{aligned}$$

where, $m+1$ integrations by parts give

$$\int_0^{\tilde{\eta}} \tilde{x}^{m+1}\tilde{g}_1(\tilde{x},\tilde{z}) e^{-N\tilde{x}}\, \mathrm{d}\tilde{x} = \frac{1}{N^{m+1}} \int_0^{\tilde{\eta}} \tilde{g}_{m+2}(\tilde{x},\tilde{z}) e^{-N\tilde{x}}\, \mathrm{d}\tilde{x} + O(e^{-\tilde{\eta}N}/N)$$

with $\tilde{g}_{m+2} = \frac{\partial^{m+1}}{\partial\tilde{x}^{m+1}}\left(\tilde{x}^{m+1}\tilde{g}_1(\tilde{x},\tilde{z})\right)$. We obtain

$$\begin{aligned}&\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} \tilde{x}^{m+1}\tilde{g}_1(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z} \\ &\quad = \frac{1}{N^{m+1}} \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}_{\tilde{\eta}}}} \tilde{g}_{m+2}(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\, \mathrm{d}\tilde{x}\, \mathrm{d}\tilde{z} + O(e^{-\tilde{\eta}N}/N) \\ &\quad = O\left(\frac{1}{N^{m+n/2+2}}\right),\end{aligned}$$

since $\int_0^{\tilde{\eta}} \tilde{g}_{m+2}(\tilde{x},\tilde{z}) e^{-N\tilde{x}}\, \mathrm{d}\tilde{x}$ is of order $1/N$.

Similarly, m integrations by parts versus $\tilde{x}$ give

$$\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{x}^m \tilde{z}_k \tilde{h}_k(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}$$
$$= \frac{1}{N^m}\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{z}_k \tilde{q}_{k,m}(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z} + O(e^{-\tilde{\eta}N}/N),$$

where $\tilde{q}_{k,m}(\tilde{x},\tilde{z}) = \frac{\partial^m}{\partial\tilde{x}^m}\left(\tilde{x}^m\tilde{h}_k(\tilde{x},\tilde{z})\right)$. A single integration by parts versus $\tilde{z}_k$ yields

$$\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{z}_k \tilde{q}_{k,m}(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}$$
$$= \frac{1}{N}\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \frac{\partial\tilde{q}_{k,m}}{\partial\tilde{z}_k}(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z} + O(e^{-\tilde{\eta}^2N/2}/N).$$

This implies that

$$\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{x}^m \tilde{z}_k \tilde{h}_k(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{kl=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}$$
$$= \frac{1}{N^{m+1}}\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \frac{\partial\tilde{r}_{k,m}}{\partial\tilde{z}_k}(\tilde{x},\tilde{z}) e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z} + O(e^{-\tilde{\eta}^2N/2}/N)$$
$$= O\left(\frac{1}{N^{m+n/2+2}}\right).$$

Thus, we obtain (12.9), since $\tilde{I}_{\tilde{\eta}} = \frac{\tilde{g}(0,0)m!}{N^{m+1}}\left(\frac{2\pi}{N}\right)^{n/2}(1+O(1/N))$ and $\tilde{g}(0,0) = \frac{g(0,0)}{\left(\left|\frac{\partial f}{\partial x}\Big|_{(0,0)}\right|\right)^{m+1}\sqrt{\left|\det\left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)\right|}}$.

Assuming now that $g(0,0)=0$, $\frac{\partial g}{\partial x}\Big|_{(0,0)} = 0$ and $\frac{\partial g}{\partial z}\Big|_{(0,0)} = 0$, and considering the function $\tilde{g}$ in (12.11), we have $\tilde{g}(0,0)=0$, $\frac{\partial\tilde{g}}{\partial\tilde{x}}\Big|_{(0,0)} = 0$ and $\frac{\partial\tilde{g}}{\partial\tilde{z}}\Big|_{(0,0)} = 0$. Moreover, denoting

$$\lambda_0 = \sqrt{\left|\det\left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)\right|}\left(-\frac{\partial f}{\partial x}\Big|_{(0,0)}\right)^{m+1},$$

we have

$$\lambda_0\tilde{g}(\phi(x,z),\psi(z)) = g(x,z)\,(1+e(x,z)),$$

where e is an analytic function such that $e(0,0)=0$. As in (12.7), we obtain

$$\mathrm{Tr}\left(\frac{\partial^2\tilde{g}}{\partial\tilde{z}^2}\Big|_{(0,0)}\right) = \frac{\mathrm{Tr}\left(-\frac{\partial^2 g}{\partial z^2}\Big|_{(0,0)}\left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)^{-1}\right)}{\sqrt{\left|\det\left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)}\right)\right|}\left(-\frac{\partial f}{\partial x}\Big|_{(0,0)}\right)^{m+1}}. \tag{12.12}$$

Since $\tilde{g}$ and its first partial derivatives versus $\tilde{x}$ and $\tilde{z}_k$ vanish, we have

$$\tilde{g}(\tilde{x},\tilde{z}) = \tilde{x}^2\tilde{a}(\tilde{x},\tilde{z}) + \sum_{k,k'=1}^{n} \tilde{z}_k\tilde{z}_{k'}\tilde{b}_{k,k'}(\tilde{x},\tilde{z}) + \sum_{k=1}^{n} \tilde{x}\tilde{z}_k\tilde{c}_k(\tilde{x},\tilde{z}),$$

where the functions $\tilde{a}$, $\tilde{b}_{k,k'}$ and $\tilde{c}_k$ are analytic. To evaluate the integral in (12.11), we have to consider the dominant terms of three kinds of integrals:

$$\begin{aligned}
A &= \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{x}^{m+2}\tilde{a}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z},\\
B_{k,k'} &= \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{x}^{m}\tilde{z}_k\tilde{z}_{k'}\tilde{b}_{k,k'}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z},\\
C_k &= \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{x}^{m+1}\tilde{z}_k\tilde{c}_k(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}.
\end{aligned}$$

As done previously, $m+2$ integrations by parts on $\tilde{x}$ yield $A = O\left(N^{-m-n/2-3}\right)$. Also as previously, $m+1$ integrations by parts versus $\tilde{x}$ and a single integration by parts versus $\tilde{z}_k$ provide $C_k = O\left(N^{-m-n/2-3}\right)$. For $k \neq k'$, m integrations by parts versus $\tilde{x}$, one integration by parts versus $\tilde{z}_k$ and another versus $\tilde{z}_{k'}$, yield a similar expression to $B_{k,k'} = O\left(N^{-m-n/2-3}\right)$. For $k = k'$, we start with m integrations by parts versus $\tilde{x}$

$$\begin{aligned}
B_{k,k} &= \int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{x}^{m}\tilde{z}_k^2\tilde{b}_{k,k}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}\\
&= \frac{1}{N^m}\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{z}_k^2\tilde{q}_{k,m}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{k=1}^{n}\tilde{z}_k^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z} + O(e^{-\tilde{\eta}N}/N),
\end{aligned}$$

where $\tilde{q}_{k,m}(\tilde{x},\tilde{z}) = \frac{\partial^m}{\partial\tilde{x}^m}\left(\tilde{x}^m\tilde{b}_{k,k}(\tilde{x},\tilde{z})\right)$. We notice that $\tilde{q}_{k,m}(0) = m!\tilde{b}_{k,k}(0)$. A single integration by parts versus $\tilde{z}_k$ yields

$$\begin{aligned}
&\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \tilde{z}_k^2\tilde{q}_{k,m}(\tilde{x},\tilde{z})e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}\\
&= \frac{1}{N}\int_{(\tilde{x},\tilde{z})\in\widetilde{\mathscr{C}}_{\tilde{\eta}}} \left(\tilde{q}_{k,m}(\tilde{x},\tilde{z}) + \tilde{z}_k\frac{\partial\tilde{q}_{k,m}}{\partial\tilde{z}_k}(\tilde{x},\tilde{z})\right)e^{-N\left(\tilde{x}+\frac{1}{2}\sum_{l=1}^{n}\tilde{z}_l^2\right)}\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z} + O(e^{-N\tilde{\eta}^2/2})\\
&= \tilde{q}_{k,m}(0)\frac{1}{N^2}\left(\frac{2\pi}{N}\right)^{n/2} + O\left(N^{-n/2-3}\right).
\end{aligned}$$

With $\tilde{q}_{k,m}(0) = m!\tilde{b}_{k,k}(0)$, the sum $A + \sum_k C_k + \sum_{k,k'} B_{k,k'}$ corresponding to the integral in (12.11) becomes

$$\tilde{I}_{\tilde{\eta}}(N) = \frac{\sum_{k=1}^{n} m!\tilde{b}_{k,k}(0)}{N^{m+2}}\left(\frac{2\pi}{N}\right)^{n/2} + O\left(N^{-m-n/2-3}\right).$$

Since $\tilde{I}_{\tilde{\eta}}$ and $e^{-Nf(0)}\mathscr{I}_g(N)$ coincide up to exponentially small terms, we obtain (12.10) from (12.12), since $\sum_{k=1}^{n} \tilde{b}_{k,k}(0) = \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 \tilde{g}}{\partial \tilde{z}^2}\Big|_0 \right)$. ■

The asymptotic expansions of Theorems 12.1 and 12.2 directly yield the following approximations of the Bayesian mean and variance.

Corollary 12.1 *Consider the analytic function $f(z)$ of Theorem 12.1. Then we have the following asymptotic for any analytic function $g(z)$:*

$$\mathscr{M}_g(N) \triangleq \frac{\int_{z\in(-1,1)^n} g(z) \exp\left(Nf(z)\right) \mathrm{d}z}{\int_{z\in(-1,1)^n} \exp\left(Nf(z)\right) \mathrm{d}z} = g(0) + O(N^{-1}). \tag{12.13}$$

We have also

$$\mathscr{V}_g(N) \triangleq \frac{\int_{z\in(-1,1)^n} \left(g(z) - \mathscr{M}_g(N)\right)^2 \exp\left(Nf(z)\right)\ \mathrm{d}z}{\int_{z\in(-1,1)^n} \exp\left(Nf(z)\right)\ \mathrm{d}z} = \frac{Tr\left(- \frac{\partial^2 g}{\partial z^2}\Big|_0 \left(\frac{\partial^2 f}{\partial z^2}\Big|_0 \right)^{-1} \right)}{2N} + O\left(N^{-2}\right). \tag{12.14}$$

Consider the analytic function $f(x,z)$ of Theorem 12.2. Then, we have the following asymptotic for any analytic function $g(x,z)$:

$$\mathscr{M}_g(N) \triangleq \frac{\int_{x\in(0,1)} \int_{z\in(-1,1)^n} x^m g(x,z) \exp\left(Nf(x,z)\right)\ \mathrm{d}x\ \mathrm{d}z}{\int_{x\in(0,1)} \int_{z\in(-1,1)^n} x^m \exp\left(Nf(x,z)\right)\ \mathrm{d}x\ \mathrm{d}z} = g(0,0) + O(N^{-1}). \tag{12.15}$$

We have also

$$\mathscr{V}_g(N) \triangleq \frac{\int_{x\in(0,1)} \int_{z\in(-1,1)^n} x^m \left(g(x,z) - \mathscr{M}_g(N)\right)^2 \exp\left(Nf(x,z)\right) \mathrm{d}x\ \mathrm{d}z}{\int_{x\in(0,1)} \int_{z\in(-1,1)^n} x^m \exp\left(Nf(x,z)\right) \mathrm{d}x\ \mathrm{d}z} = \frac{Tr\left(- \frac{\partial^2 g}{\partial z^2}\Big|_{(0,0)} \left(\frac{\partial^2 f}{\partial z^2}\Big|_{(0,0)} \right)^{-1} \right)}{2N} + O\left(N^{-2}\right). \tag{12.16}$$

In the proof of Theorem 12.1, we have shown during the passage from z to $\tilde{z}$ coordinates the following lemma.

Lemma 12.1 *Take two C^2 real-value functions f and g of $z \in \mathbb{R}^n$. Assume that 0 is a regular critical point of f and a critical point of g. Take any C^2 diffeomorphism ϕ defined locally around 0: $\tilde{z} = \phi(z)$. Then*

$$Tr\left(- \left.\frac{\partial^2 g}{\partial z^2}\right|_0 \left(\left.\frac{\partial^2 f}{\partial z^2}\right|_0 \right)^{-1} \right) = Tr\left(- \left.\frac{\partial^2 \tilde{g}}{\partial \tilde{z}^2}\right|_{\phi(0)} \left(\left.\frac{\partial^2 \tilde{f}}{\partial \tilde{z}^2}\right|_{\phi(0)} \right)^{-1} \right)$$

where $\tilde{f}(\phi(z)) = f(z)$ *and* $\tilde{g}(\phi(z)) = g(z)$.

This lemma simply states that the above trace formula is coordinate-free—that is, independent of the local coordinates chosen to compute the Hessian of f and g at their common critical point.

12.3 Application to Quantum State Tomography

As explained in [4], the estimated parameter p corresponds to a density operator ρ (quantum state), which is a square matrix with complex entries and belonging to the convex compact set $\mathscr{D}$ formed by Hermitian $d \times d$ non-negative matrices of trace one. Then, the log-likelihood function admits the following structure:

$$f(\rho) = \sum_{\mu \in \mathscr{M}} \log \left(\mathrm{Tr} \left(\rho Y_\mu \right) \right), \tag{12.17}$$

where the set $\mathscr{M}$ is finite and each measurement data Y_μ belongs also to $\mathscr{D}$. For any Hermitian $d \times d$ matrix A (a quantum observable), we are interested in providing an approximation of the Bayesian estimate of $\mathrm{Tr}\,(\rho A)$,

$$I_A(N) = \frac{\int_{\mathscr{D}} \mathrm{Tr}\,(\rho A)\, e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho}{\int_{\mathscr{D}} e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho}, \tag{12.18}$$

and of the Bayesian variance

$$V_A(N) = \frac{\int_{\mathscr{D}} \Big(\mathrm{Tr}\,(\rho A) - I_A(N) \Big)^2 e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho}{\int_{\mathscr{D}} e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho}. \tag{12.19}$$

Here, $\mathrm{d}\rho$ stands for the standard Euclidian volume element on $\mathscr{D}$, derived from the Frobenius product of $n \times n$ Hermitian matrices, and $\mathbb{P}_0 > 0$ is a probability density on ρ prior to the measurement data (Y_μ). Since the number of real parameters to describe ρ is large in general, it is difficult to compute these integrals even numerically using the Monte Carlo method.

The following lemma provides a unitary invariance characterization of any $\bar{\rho}$ argument of the maximum of f on $\mathscr{D}$.

Lemma 12.2 *Assume that the* $d \times d$ *Hermitian matrix* $\overline{\rho}$ *is an argument of the maximum of* $f : \mathscr{D} \ni \rho \mapsto f(\rho) \in [-\infty, 0]$ *defined in (12.17) over* $\mathscr{D}$ *(the set of density operators). Then* $\overline{\rho}$ *necessarily satisfies the following conditions:*

- *$Tr\left(\overline{\rho} Y_\mu\right) > 0$ for each $\mu \in \mathscr{M}$;*
- *$\left[\overline{\rho}\,,\ \nabla f|_{\overline{\rho}}\right] = \overline{\rho} \cdot \nabla f|_{\overline{\rho}} - \nabla f|_{\overline{\rho}} \cdot \overline{\rho} = 0$, where $\nabla f|_{\overline{\rho}} = \sum_{\mu\in\mathscr{M}} \frac{Y_\mu}{Tr(\overline{\rho} Y_\mu)}$ is the gradient of f at $\overline{\rho}$ for the Frobenius scalar product;*
- *there exists $\overline{\lambda} > 0$ such that $\overline{\lambda P} = \overline{P}\ \nabla f|_{\overline{\rho}}$ and $\nabla f|_{\overline{\rho}} \leq \overline{\lambda} I$, where $\overline{P}$ is the orthogonal projector on the range of $\overline{\rho}$ and I is the identity operator.*

These conditions are also sufficient and characterize the unique maximum when, additionally, the vector space spanned by the Y_μ's coincides with the set of Hermitian matrices.

Proof Since f is a concave function of ρ, we can use the standard optimality criterion for a convex optimization problem (see, e.g., [10, Sect. 4.2.3]): $\overline{\rho}$ maximizes f over the convex compact set $\mathscr{D}$ if, and only if, $\rho \in \mathscr{D}$, $\text{Tr}\left((\rho - \overline{\rho})\ \nabla f|_{\overline{\rho}}\right) \leq 0$.

Assume that $f(\overline{\rho})$ is maximum. Since $f(I) > -\infty$, for each μ we have $\text{Tr}\left(\overline{\rho} Y_\mu\right) > 0$. Taking $\rho = e^{-iH}\overline{\rho} e^{iH}$, where H is an arbitrary Hermitian operator, we have

$$\text{Tr}\left(e^{-iH}\overline{\rho} e^{iH}\ \nabla f|_{\overline{\rho}}\right) \leq \text{Tr}\left(\overline{\rho}\ \nabla f|_{\overline{\rho}}\right).$$

For H close to zero, we have via the Baker-Campbell-Hausdorff formula, $e^{-iH}\overline{\rho} e^{iH} = \overline{\rho} - i[H, \overline{\rho}] + O(\text{Tr}\left(H^2\right))$. The above inequality implies that for all H sufficiently small, $\text{Tr}\left([H, \overline{\rho}]\ \nabla f|_{\overline{\rho}}\right) = \text{Tr}\left(H\left[\overline{\rho}, \nabla f|_{\overline{\rho}}\right]\right) = 0$ and thus $\overline{\rho}$ and $\nabla f|_{\overline{\rho}}$ commute.

Consider the spectral decomposition $\overline{\rho} = U\overline{\Delta}U^\dagger$ where U is unitary and $\overline{\Delta}$ diagonal with entries $0 \leq \overline{\Delta}_1 \leq \overline{\Delta}_2 \leq \cdots \leq \overline{\Delta}_d \leq 1$. Since $\overline{\rho}$ and $\nabla f|_{\overline{\rho}}$ commute, we also have $\nabla f|_{\overline{\rho}} = U\overline{\Lambda}U^\dagger$ with $\overline{\Lambda}$ diagonal with entries $(\overline{\Lambda}_k)$. Since ∇f is non-negative, these entries are also non-negative. Taking $\rho = U\Delta U^\dagger$, where Δ is any diagonal matrix with non-negative entries and of trace one, we have

$$\text{Tr}\left((\rho - \overline{\rho})\ \nabla f|_{\overline{\rho}}\right) = \text{Tr}\left((\Delta - \overline{\Delta})\overline{\Lambda}\right) \leq 0.$$

This means that, for any $(\Delta_1, \ldots, \Delta_d) \in [0, 1]^d$ such that $\sum_{k=1}^d \Delta_k = 1$ we have:

$$\sum_{k=1}^{d} (\Delta_k - \overline{\Delta}_k)\overline{\Lambda}_k \leq 0.$$

Take $\epsilon > 0$, $(k_1, k_2) \in \{1, \ldots, d\}^2$ such that $\overline{\Delta}_{k_1} > 0$ and $k_2 \neq k_1$. For $k \in \{1, \ldots, d - 1\}/\{k_1, k_2\}$ set $\Delta_k = \overline{\Delta}_k$, and take $\Delta_{k_1} = \overline{\Delta}_{k_1} - \epsilon$ with $\Delta_{k_2} = \overline{\Delta}_{k_2} + \epsilon$. By construction $\text{Tr}\,(\Delta) = 1$ and, for $\epsilon > 0$ sufficiently small, $\Delta_k \geq 0$ for all $k \in \{1, \ldots, d\}$. The previous inequality implies that

$$\forall (k_1, k_2) \in \{1, \ldots, d\}^2 \text{ such that } \overline{\Delta}_{k_1} > 0 \text{ and } k_1 \neq k_2, \quad \overline{\Lambda}_{k_2} \leq \overline{\Lambda}_{k_1}.$$

Thus for all k_1, k_2 such that $\overline{\Delta}_{k_1} > 0$ and $\overline{\Delta}_{k_2} > 0$, $\overline{\Lambda}_{k_1} = \overline{\Lambda}_{k_2} = \overline{\lambda} \geq 0$. For k_1, k_2 such that $\overline{\Delta}_{k_1} > 0$ and $\overline{\Delta}_{k_2} = 0$, we also have $\overline{\Lambda}_{k_2} \leq \overline{\Lambda}_{k_1} = \overline{\lambda}$. Thus we obtain $\overline{\Lambda} \leq \overline{\lambda} I$. With $\overline{\Theta}$, the diagonal matrix of entries $\overline{\Theta}_k = 0$ (resp. $= 1$) when $\overline{\Delta}_k = 0$ (resp. > 0), we have $\overline{P} = U\overline{\Theta}U^\dagger$, and we obtain $\overline{\lambda P} = \overline{P}\ \nabla f|_{\overline{\rho}}$. Since $\nabla f|_{\overline{\rho}}$ is non-negative and cannot be zero, we have $\overline{\lambda} > 0$.

Take $\overline{\rho}$ satisfying the conditions of Lemma 12.2. Since they are unitary invariant, we can assume that $\overline{\rho}$ and $\nabla f|_{\overline{\rho}}$ are diagonal operators $\overline{\Delta}$ and $\overline{\Lambda}$. Since we are in the convex situation, it is enough to prove that $\overline{\rho}$ is a local maximum. Any local variation of ρ around $\overline{\rho}$ and remaining inside $\mathscr{D}$ is parameterized via the following mapping:

$$(H, D) \mapsto e^{-iH}(\overline{\Delta} + D)e^{iH} = \rho_{H,D},$$

where H is any Hermitian matrix and D is any diagonal matrix of zero trace such that $\overline{\Delta} + D \geq 0$. We have the following expansion for H and D around zero:

$$\rho_{H,D} = \overline{\Delta} + D - i[H, \overline{\Delta}] - i[H, D] - \tfrac{1}{2}[H, [H, \overline{\Delta}]] + O(\mathrm{Tr}\left(H^3 + D^3\right)).$$

This yields the following second-order expansion of $(H, D) \mapsto f(\rho_{H,D})$ around zero:

$$\begin{aligned} f(\rho_{H,D}) = & f(\overline{\rho}) + \mathrm{Tr}\left(\overline{\Lambda}\left(D - i[H, \overline{\Delta}] - i[H, D] - \frac{1}{2}[H, [H, \overline{\Delta}]]\right)\right) \\ & - \sum_{\mu \in \mathscr{M}} \frac{\mathrm{Tr}^2\left((\rho_{H,D} - \overline{\rho})Y_\mu\right)}{2\mathrm{Tr}^2\left(\overline{\rho}Y_\mu\right)} + O(\|\rho_{H,D} - \overline{\rho}\|^3). \end{aligned}$$

By assumption, $\overline{\Lambda}$, $\overline{\Delta}$ and D are diagonal. Thus $\mathrm{Tr}\left(\overline{\Lambda}\left(-i[H, \overline{\Delta}] - i[H, D]\right)\right) = 0$. Some elementary arguments exploiting $\overline{\lambda\Theta} \leq \overline{\Delta} \leq \overline{\lambda} I$ show that $\mathrm{Tr}\left(\overline{\Lambda} D\right) \leq 0$ since D is such that $\overline{\Delta} + D$ is non-negative and of trace one. We also have

$$-\mathrm{Tr}\left(\overline{\Lambda}\left([H, [H, \overline{\Delta}]]\right)\right) = \mathrm{Tr}\left([H, \overline{\Lambda}]\ [H, \overline{\Delta}]\right) = -2 \sum_{k_1 \in P, k_2 \in Q} \overline{\Delta}_{k_1}(\overline{\lambda} - \overline{\Lambda}_{k_2})|H_{k_1 k_2}|^2 \leq 0$$

where $P = \{k \mid \overline{\Delta}_k > 0\}$ and $Q = \{k \mid \overline{\Delta}_k = 0\}$.

Consequently,

$$f(\rho_{H,D}) \leq f(\overline{\rho}) - \sum_{\mu \in \mathscr{M}} \frac{\mathrm{Tr}^2\left((\rho_{H,D} - \overline{\rho})Y_\mu\right)}{2\mathrm{Tr}^2\left(\overline{\rho}Y_\mu\right)} + O(\|\rho_{H,D} - \overline{\rho}\|^3).$$

Since the vector space spanned by the Y_μ coincide with the set of Hermitian matrices, the quadratic form $X \mapsto \sum_{\mu \in \mathscr{M}} \frac{\mathrm{Tr}^2(XY_\mu)}{2\mathrm{Tr}^2(\overline{\rho}Y_\mu)}$ is non-degenerate (X is any Hermitian matrix) and f is strongly concave. Thus we have $f(\rho) < f(\overline{\rho})$ for $\rho \neq \overline{\rho}$ close to $\overline{\rho}$.

Consequently, $\overline{\rho}$ is a strict local maximum and this maximum is unique and global since f is concave. ■

Theorem 12.3 *Consider the log-likelihood function f defined in (12.17). Assume that the Y_μ's span the set of Hermitian matrices. Denote by $\overline{\rho}$ the unique maximum of f on $\mathscr{D}$ and define a projector $\overline{P}$ such that, in addition to the necessary and sufficient conditions of Lemma 12.2, we have* $\ker\left(\overline{\lambda} I - \nabla f|_{\overline{\rho}}\right) = \ker(I - \overline{P})$. *Then, for any Hermitian operator A, its Bayesian mean defined in (12.18) admits the following asymptotic expansion*

$$I_A(N) = Tr\left(A\overline{\rho}\right) + O(1/N)$$

and its Bayesian variance defined in (12.19) satisfies

$$V_A(N) = Tr\left(A_{\|}\ \left(\overline{\boldsymbol{F}}\right)^{-1}(A_{\|})\right)/N + O(1/N^2)$$

where

- *for any Hermitian operator B, $B_{\|}$ stands for is orthogonal projection on the tangent space at $\overline{\rho}$ to the submanifold of Hermitian matrices with a rank equal to the rank of $\overline{\rho}$ and of unit trace, written as*

$$B_{\|} = B - \frac{Tr\left(B\overline{P}\right)}{Tr\left(\overline{P}\right)}\overline{P} - (I - \overline{P})B(I - \overline{P}); \tag{12.20}$$

  when $\overline{\rho}$ is full rank, $B_{\|} = B - Tr\,(B)\,I/d$ since $\overline{P} = I$;
- *the linear super-operator $\overline{\boldsymbol{F}}$ corresponds to the Hessian at $\overline{\rho}$ of the restriction of f to the manifold of Hermitian matrices of rank equal to the rank of $\overline{\rho}$ and with trace one. For any Hermitian operator X, it is written as*

$$\overline{\boldsymbol{F}}(X) = \sum_\mu \frac{Tr\left(XY_{\mu\|}\right)}{Tr^2\left(\overline{\rho}Y_\mu\right)} Y_{\mu\|} + \left(\overline{\lambda} I - \nabla f|_{\overline{\rho}}\right) X\overline{\rho}^+ + \overline{\rho}^+ X\left(\overline{\lambda} I - \nabla f|_{\overline{\rho}}\right), \tag{12.21}$$

  with $\overline{\rho}^+$ the Moore-Penrose pseudoinverse of $\overline{\rho}$; the restriction of $X \mapsto Tr\left(X\overline{\boldsymbol{F}}(X)\right)$ to the tangent space at $\overline{\rho}$ is positive definite; thus the restriction of $\overline{\boldsymbol{F}}$ to this tangent space is invertible and can be seen as the analogue of the Fisher information; its inverse at $A_{\|}$ is denoted here above by $\left(\overline{\boldsymbol{F}}\right)^{-1}(A_{\|})$.

Proof The Hessian of f at $\rho \in \mathscr{D}$ where $f(\rho) > -\infty$ is

$$\nabla^2 f\Big|_\rho (X, Z) = -\sum_\mu \frac{\mathrm{Tr}\left(XY_\mu\right)\mathrm{Tr}\left(ZY_\mu\right)}{\mathrm{Tr}^2\left(\rho Y_\mu\right)},$$

where X and Z are any Hermitian matrices. Since it is positive definite, f is strongly concave. Consequently the argument of the maximum of f on $\mathscr{D}$ is unique, denoted $\overline{\rho}$, and satisfies the condition of Lemma 12.2. Take a small neighbourhood $\mathscr{V}$ of $\overline{\rho}$ in $\mathscr{D}$. Then there exists a $\epsilon > 0$ such that, for $\rho \in \mathscr{D}/\mathscr{V}, f(\rho) \leq f(\overline{\rho}) - \epsilon$. To investigate $\int_{\mathscr{V}} e^{N(f(\rho)-f(\overline{\rho}))}\ \mathbb{P}_0(\rho)\ \mathrm{d}\rho$, we consider the following local coordinates based on the spectral decomposition of $\overline{\rho} = U\overline{\Delta}U^\dagger$ with U unitary and $\overline{\Delta}$ diagonal with entries $0 = \overline{\delta}_1 \leq \overline{\delta}_2 \leq \cdots \leq \cdots, \overline{\delta}_d \leq 1$ with $\sum_{k=1}^{d} \overline{\delta}_k = 1$. Denote by r the rank of $\overline{\rho}$ and assume that $r < d$ (the case $r = d$ is much simpler since it relies on Theorem 12.1, and is left to the reader). We have $\overline{\delta}_k = 0$ for k between 1 and $d - r$, and $\overline{\delta}_k > 0$ for k between $d - r + 1$ and d. Since the volume element $\mathrm{d}\rho$ used in (12.18) and (12.19) is unitary invariant [11, page 42], we can assume without lost of generality that $\overline{\rho}$ is diagonal (change $\mathbb{P}_0(\bullet)$ to $\mathbb{P}_0(U \bullet U^\dagger)$ and replace each Y_μ by $U^\dagger Y_\mu U$ in the definition of f in (12.17)). Consider the following map

$$(\xi, \zeta, \omega) \mapsto Y = \exp\left(\begin{bmatrix} 0 & \omega \\ -\omega^\dagger & 0 \end{bmatrix}\right) \begin{bmatrix} \xi & 0 \\ 0 & \overline{\Delta}_r + \zeta - \frac{\mathrm{Tr}(\xi)}{r} I_r \end{bmatrix} \exp\left(\begin{bmatrix} 0 & -\omega \\ \omega^\dagger & 0 \end{bmatrix}\right)$$

where ξ is a $(d-r) \times (d-r)$ Hermitian matrix, ω is a $(d-r) \times r$ matrix with complex entries, ζ is a $r \times r$ Hermitian matrix of trace 0, I_r is the identity matrix of size r and $\overline{\Delta} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Delta}_r \end{bmatrix}$. This map is a local diffeomorphism from a neighbourhood of $(0, 0, 0)$ to a neighbourhood of $\overline{\rho}$ in the set of Hermitian matrices of trace one since its tangent map at zero, given by

$$(\delta\xi, \delta\zeta, \delta\omega) \mapsto \begin{bmatrix} \delta\xi & \delta\omega\, \overline{\Delta}_r \\ \overline{\Delta}_r\, \delta\omega^\dagger & \delta\zeta - \frac{\mathrm{Tr}(\delta\xi)}{r} I_r \end{bmatrix} = \delta\rho \tag{12.22}$$

is bijective (local inversion theorem). Thus we have

$$\int_{\mathscr{V}} e^{N(f(\rho)-f(\overline{\rho}))}\ \mathbb{P}_0(\rho)\ \mathrm{d}\rho$$
$$= \int_{Y^{-1}(\mathscr{V})} e^{N(f(\xi,\zeta,\omega)-f(0,0,0))} \mathbb{P}_0(\xi, \zeta, \omega) J(\xi, \zeta, \omega) \mathrm{d}\xi\ \mathrm{d}\zeta\ \mathrm{d}\omega$$

where $f(\xi, \zeta, \omega)$ and $\mathbb{P}_0(\xi, \zeta, \omega)$ stand for $f(Y(\xi, \zeta, \omega))$ and $\mathbb{P}_0(Y(\xi, \zeta, \omega))$, and where $J(\xi, \zeta, \omega)$ is the Jacobian of this change of coordinates.

Since the constraint $Y(\xi, \zeta, \omega) \geq 0$ may be written $\xi \geq 0$, we consider another change of variables to parameterize $\xi \geq 0$ around 0: $\Xi\ :\ \ (x, \sigma, \zeta, \omega) \mapsto (x\sigma = \xi, \zeta, \omega)$, where $x \geq 0$ and σ is a $(d-r) \times (d-r)$ density matrix. Then,

$$
\begin{aligned}
&\int_{\mathscr{V}} e^{N(f(\rho)-f(\overline{\rho}))}\ \mathbb{P}_0(\rho)\ \mathrm{d}\rho \\
&= \int_{\Xi^{-1}\left(\Upsilon^{-1}(\mathscr{V})\right)} e^{N(f(x\sigma,\zeta,\omega)-f(0,0,0))} \mathbb{P}_0(x\sigma,\zeta,\omega) J(x\sigma,\zeta,\omega) x^m\ \mathrm{d}x\ \mathrm{d}\sigma\ \mathrm{d}\zeta\ \mathrm{d}\omega
\end{aligned}
$$

with $m = (d-r+1)(d-r-1)$. This change of variables is singular, since for $x = 0$ it is not invertible; however, since the set of coordinates verifying $x = 0$ is of zero measure, this has no impact on the integral. Take $\eta > 0$ small enough and adjust the neighbourhood $\mathscr{V}$ of $\overline{\rho}$ such that $\Xi^{-1}\left(\Upsilon^{-1}(\mathscr{V})\right)$ coincides with the set where $x \in (0,\eta)$, $\sigma \in \mathscr{D}_{d-r}$ and all the real and imaginary parts of the ζ and ω entries belong to $(-\eta,\eta)$. Following the notations of Theorem 12.1, set $z = (\zeta,\omega)$. We have $z \in (-\eta,\eta)^n$ with $n = 2r(d-r) + (r+1)(r-1)$ and

$$
\begin{aligned}
&\int_{\mathscr{V}} e^{N(f(\rho)-f(\overline{\rho}))}\ \mathbb{P}_0(\rho)\ \mathrm{d}\rho \\
&\qquad = \int_{\sigma\in\mathscr{D}_{d-r}} \left(\int_{(x,z)\in(0,\eta)\times(-\eta,\eta)^n} e^{Nf(x\sigma,z)} x^m J(x\sigma,z)\mathbb{P}_0(x\sigma,z)\ \mathrm{d}x\ \mathrm{d}z \right) \mathrm{d}\sigma.
\end{aligned}
$$

For each $\sigma \in \mathscr{D}_{d-r}$, let us use (12.9), with $J(x\sigma,z)\mathbb{P}_0(x\sigma,z)$ standing for $g(x,z)$. We have $g(0,0) = J(0,0)\mathbb{P}_0(0,0) > 0$. By construction, we have

$$
f(x\sigma,z) = xf_1(x,\sigma,z) + f(0,z),
$$

where $f_1(x,\sigma,z)$ is analytic versus (x,z) and $f_1(0,\sigma,0) = \left(\mathrm{Tr}\left(\Lambda_{d-r}\sigma\right) - \overline{\lambda}\right)$. This is based on (12.22) and on the diagonal structure $\nabla f|_{\overline{\rho}} = \begin{bmatrix} \Lambda_{d-r} & 0 \\ 0 & \overline{\lambda} I_r \end{bmatrix}$. By assumption, $\Lambda_{d-r} < \overline{\lambda} I_{d-r}$. Thus, there exists $\epsilon' > 0$ such that, for all σ, $f_1(0,\sigma,0) < -\epsilon'$ and $\frac{\partial f}{\partial x} < -\epsilon'$ at $(x,z) = 0$, for any $\sigma \in \mathscr{D}_{d-r}$. Let us consider now the expansion of $z \mapsto f(0,z)$ up to order 2 versus z. Using $\delta z = (\delta\zeta,\delta\omega)$ and (12.22), completed via second-order terms derived from the Baker-Campbell-Hausdorff formula, we find

$$
\delta\rho = \begin{bmatrix} \delta\omega\ \overline{\Delta}_r\ \delta\omega^\dagger & \delta\omega\ (\overline{\Delta}_r + \delta\zeta) \\ (\delta\zeta + \overline{\Delta}_r)\ \delta\omega^\dagger & \delta\zeta - \frac{\delta\omega^\dagger\delta\omega\overline{\Delta}_r + \overline{\Delta}_r\delta\omega^\dagger\delta\omega}{2} \end{bmatrix} + 0(\|\delta z\|^3).
$$

Consequently,

$$
\begin{aligned}
f(0,\delta z) =& f(\overline{\rho}) + \mathrm{Tr}\left(\nabla f|_{\overline{\rho}}\ \delta\rho\right) + \tfrac{1}{2}\ \nabla^2 f\Big|_{\overline{\rho}}(\delta\rho,\delta\rho) + O(\|\delta\rho\|^3) \\
&= f(\overline{\rho}) - \mathrm{Tr}\left((\overline{\lambda} I_{d-r} - \Lambda_{d-r})\delta\omega\ \overline{\Delta}_r\ \delta\omega^\dagger\right) - \tfrac{1}{2}\sum_{\mu} \frac{\mathrm{Tr}^2\left(\delta\rho Y_\mu\right)}{\mathrm{Tr}^2\left(\overline{\rho} Y_\mu\right)}. \qquad (12.23)
\end{aligned}
$$

This shows that $\frac{\partial f}{\partial z}$ vanishes at $(0, z)$ and that $\frac{\partial^2 f}{\partial z^2}$ is negative definite at $(0, z)$ ($\overline{\lambda} I_{d-r} > \Lambda_{d-r}$) and independent of σ. All the assumptions necessary for (12.9) are fulfilled and we can write:

$$\int_{\mathscr{D}} e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho = \kappa_0\, e^{f(\overline{\rho})N} N^{-m-n/2-1} \int_{\sigma \in \mathscr{D}_{d-r}} \frac{\mathrm{d}\sigma}{\left(\overline{\lambda} - \mathrm{Tr}\left(\Lambda_{d-r}\sigma\right)\right)^{m+1}} + O\left(e^{f(\overline{\rho})N} N^{-m-n/2-2}\right)$$

where $\kappa_0 = \frac{\mathbb{P}_0(\overline{\rho}) J(0,0) m!\ (2\pi)^{n/2}}{\sqrt{\left|\det\left(\left.\frac{\partial^2 f}{\partial z^2}\right|_{(0,0)}\right)\right|}}$.

Similarly we have

$$\int_{\mathscr{D}} \mathrm{Tr}\,(\rho A)\, e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho = \kappa_0 \mathrm{Tr}\left(A\overline{\rho}\right)\ e^{f(\overline{\rho})N} N^{-m-n/2-1} \int_{\sigma \in \mathscr{D}_{d-r}} \frac{\mathrm{d}\sigma}{\left(\overline{\lambda} - \mathrm{Tr}\left(\Lambda_{d-r}\sigma\right)\right)^{m+1}} + O\left(e^{f(\overline{\rho})N} N^{-m-n/2-2}\right).$$

Consequently, we have proved that $I_A(N) = \mathrm{Tr}\left(\overline{\rho} A\right) + O(1/N)$.

Simple computations show that the expansion of $V_A(N)$ reduces to the expansion of the integral $\int_{\mathscr{D}} \mathrm{Tr}^2\left((\rho - \overline{\rho})A\right) e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, \mathrm{d}\rho$ based on (12.10) with $g(x, \sigma, z) = J(x\sigma, z)\mathbb{P}_0(x\sigma, z) h(x\sigma, z)$, $h(x\sigma, z) = \mathrm{Tr}^2\left((\Upsilon(x\sigma, z) - \overline{\rho})A\right)$ and $z = (\zeta, \omega)$. Since $\left.\frac{\partial^2 g}{\partial z^2}\right|_{(0,\sigma,0)} = J(0,0)\mathbb{P}_0(\overline{\rho}) \left.\frac{\partial^2 h}{\partial z^2}\right|_{(0,\sigma,0)}$ is independent of σ, we have from (12.10):

$$\int_{\mathscr{D}} \mathrm{Tr}^2\left((\rho - \overline{\rho})A\right) e^{Nf(\rho)}\, \mathbb{P}_0(\rho)\, d\rho = \kappa_0 \frac{\mathrm{Tr}\left(-\left.\frac{\partial^2 h}{\partial z^2}\right|_{(0,0)} \left(\left.\frac{\partial^2 f}{\partial z^2}\right|_{(0,0)}\right)^{-1}\right)}{2} e^{f(\overline{\rho})N} N^{-m-n/2-2} \int_{\sigma \in \mathscr{D}_{d-r}} \frac{d\sigma}{\left(\overline{\lambda} - \mathrm{Tr}\left(\Lambda_{d-r}\sigma\right)\right)^{m+1}} + O\left(e^{f(\overline{\rho})N} N^{-m-n/2-3}\right).$$

Consequently, we have $V_A(N) = \frac{\mathrm{Tr}\left(-\left.\frac{\partial^2 h}{\partial z^2}\right|_{(0,0)} \left(\left.\frac{\partial^2 f}{\partial z^2}\right|_{(0,0)}\right)^{-1}\right)}{2N} + O(N^{-2})$. The fact that the trace in the numerator coincides with $2\mathrm{Tr}\left(A_{\|}\ \left(\overline{\boldsymbol{F}}\right)^{-1}(A_{\|})\right)$ results from the following computations.

- Formula (12.20) is unitary invariant. In the frame where $\overline{\rho} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{\Delta}_r \end{bmatrix}$ is diagonal, the tangent space to the manifold of rank r Hermitian matrices at $\overline{\rho}$ is given by $\delta\rho$ satisfying (12.22) with $\delta\xi = 0$ and $(\delta\zeta, \delta\omega)$ arbitrary. One can check that (12.20) provides the following block decomposition $\begin{bmatrix} 0 & A_{0,r} \\ A_{0,r}^{\dagger} & A_r - \frac{\mathrm{Tr}(A_r)}{r} I_r \end{bmatrix}$ for $A_{\parallel}$ when $A = \begin{bmatrix} A_0 & A_{0,r} \\ A_{0,r}^{\dagger} & A_r \end{bmatrix}$. One can also check that $A_{\parallel}$ belongs to this tangent space and that $\mathrm{Tr}\,(A\delta\rho) = \mathrm{Tr}\,\left(A_{\parallel}\delta\rho\right)$ for any tangent element $\delta\rho$.

- Since $h(0,z) = \mathrm{Tr}^2\left((\Upsilon(0,z) - \overline{\rho})A\right)$, we have

$$\left.\frac{\partial^2 h}{\partial z^2}\right|_{(0,0)} (\delta z, \delta z) = 2\mathrm{Tr}^2\,(\delta\Upsilon\, A) = 2\mathrm{Tr}^2\left(\delta\Upsilon\, A_{\parallel}\right)$$

with $\delta\Upsilon = \begin{bmatrix} 0 & \delta\omega\, \overline{\varDelta}_r \\ \overline{\varDelta}_r\, \delta\omega^{\dagger} & \delta\zeta \end{bmatrix}$ and $\delta z = (\delta\zeta, \delta\omega)$. This means that $\left.\frac{\partial^2 h}{\partial z^2}\right|_{(0,0)}$ is collinear with the orthogonal projector on the direction given by $A_{\parallel}$ in the tangent space to $\overline{\rho}$. This implies that $\mathrm{Tr}\left(\left.\frac{\partial^2 h}{\partial z^2}\right|_{(0,0)} \left(\left.\frac{\partial^2 f}{\partial z^2}\right|_{(0,0)}\right)^{-1}\right)$ corresponds to twice the value at $A_{\parallel}$ of the quadratic form attached to the inverse of the Hessian at $\overline{\rho}$ of the restriction of f to the manifold of rank r Hermitian matrices of trace one (we use here Lemma 12.1).

- This Hessian is given by (12.21) since, for $X = \delta\Upsilon = \begin{bmatrix} 0 & \delta\omega\, \overline{\varDelta}_r \\ \overline{\varDelta}_r\, \delta\omega^{\dagger} & \delta\zeta \end{bmatrix}$, we have

$$\mathrm{Tr}\left(X\left(\overline{\lambda}I - \nabla f|_{\overline{\rho}}\right)X\overline{\rho}^{+} + X\overline{\rho}^{+}X\left(\overline{\lambda}I - \nabla f|_{\overline{\rho}}\right)\right)$$
$$= 2\mathrm{Tr}\left((\overline{\lambda}I_{d-r} - \varLambda_{d-r})\delta\omega\, \overline{\varDelta}_r\, \delta\omega^{\dagger}\right)$$

because $\overline{\rho}^{+} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{\varDelta}_r^{-1} \end{bmatrix}$. We recover from (12.23) that $f(0,z) = f(\overline{\rho}) - \frac{1}{2}$ $\mathrm{Tr}\left(X\,\overline{\boldsymbol{F}}(X)\right)$, i.e., that $\overline{\boldsymbol{F}}$ is indeed the Hessian at $\overline{\rho}$ of the restriction of f to rank r Hermitian matrices of trace one.

■

12.4 Concluding Remark

When maximum likelihood estimation provides a quantum state of reduced rank, we have expressed, based on asymptotic expansions of specific multidimensional Laplace integrals, an estimate of the Bayesian mean and variance for any observable. We speculate that similar asymptotic expansions could be of some interest for quantum compress sensing [12] when the dimension of the underlying Hilbert space is large and the rank is small.

References

1. Cappé, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Models. Springer Series in Statistics (2005)
2. Paris, M.G.A., Rehacek, J.: Quantum State Estimation. Springer (2004)
3. Blume-Kohout, R.: Optimal, reliable estimation of quantum states. New J. Phys. **12**(4), 043034 (2010)
4. Six, P., Campagne-Ibarcq, P., Dotsenko, I., Sarlette, A., Huard, B., Rouchon, P.: Quantum state tomography with noninstantaneous measurements, imperfections, and decoherence. Phys. Rev. A **93**, 012109 (2016)
5. Bleistein, N., Handelsman, R.A.: Asymptotic Expansions of Integrals. Dover, New York (1986)
6. Arnold, V.I., Gusein-Zade, S.M., Varchenko, A.N.: Singularities of Differentiable Maps, vol. II. Birkhäuser, Boston (1985)
7. Lin, S.: Algebraic Methods for Evaluating Integrals in Bayesian Statistics. PhD thesis, University of California, Berkeley (2011)
8. Watanabe, S.: Algebraic Geometry and Statistical Learning Theory. Cambridge University Press (2009)
9. Milnor, J.: Morse Theory. Princeton University Press (1963)
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2009)
11. Mehta, M.L.: Random Matrices, 3rd edn. Elsevier, Academic Press (2004)
12. Gross, D., Liu, Y.-K., Flammia, S.-T., Becker, S., Eisert, J.: Quantum state tomography via compressed sensing. Phys. Rev. Lett. **105**(15), 150401 (2010)

Chapter 13
Recent Developments in Stability Theory for Stochastic Hybrid Inclusions

Andrew R. Teel

Abstract Stochastic hybrid systems (SHS) combine continuous evolution, instantaneous jumps, and random inputs that affect each type of evolution. Various types of SHS have been studied for over three decades and can be used to model many interesting systems in science and engineering. The most recent developments regarding SHS focus on models that permit nonunique solutions, perhaps thereby modeling the effect of an adversarial input on the system dynamics, and robustness properties, which can again be linked to the effect of adversaries. We call such systems "stochastic hybrid inclusions" (SHI). Using, as a departure point, developments over the past ten years on modeling, sequential compactness of the solution space, and robustness of stability for non-stochastic hybrid systems, a comprehensive modeling framework for SHI is being developed. The ultimate goal is an extensive, robust stability theory for SHI. In this paper, we review recent results that have been obtained in this direction, describing a solution concept for a class of SHI, defining stability notions like recurrence and asymptotic stability in probability, stating equivalent characterizations (involving uniformity and robustness) that follow from a sequential compactness result, providing Lyapunov-based necessary and sufficient conditions for these properties, and describing relaxed sufficient conditions that are based on an invariance-like principle.

13.1 Introduction: Praly and Robustness

It is an extreme pleasure to contribute to this volume, which honors Laurent Praly and his contributions to the field of nonlinear control on the occasion of his sixtieth birthday. My joint journal papers with Laurent have been few, and not recent, but they have always been enlightening to me. While his postdoctoral student in 1992, we published tools for semi-global stabilization by output feedback [52] and applied those tools to the general output feedback stabilization problem for nonlinear sys-

A.R. Teel (✉)
ECE Department, University of California, Santa Barbara, CA 93106-9560, USA
e-mail: teel@ece.ucsb.edu

N. Petit (ed.), *Feedback Stabilization of Controlled Dynamical Systems*,
Lecture Notes in Control and Information Sciences 473,
DOI 10.1007/978-3-319-51298-3_13

tems [51]. Later, we teamed up to present general results on disturbance attenuation for nonlinear systems, focusing on techniques that permit the use of non-smooth Lyapunov functions [54]. The same year we developed novel converse Lyapunov theorems for differential inclusions [55]. This work with Laurent allowed me to hone my technical skills, and learn new analysis tricks. For this experience, I will always be grateful.

In addition to these experiences, each of the papers with Laurent heightened my interest in the role of robustness in asymptotic stability studies. Indeed, robustness turned out to be the key to converse Lyapunov theorems, as had already been noted in [17, 19], and very explicitly in [5]; see also [18]. Subsequently, I was fascinated to discover that, for a nonlinear discrete-time system with a discontinuous right-hand side, the origin can be globally asymptotically stable with no robustness margin [16]. My students and I pointed out that this phenomenon could occur in the closed loop when employing a commonly advocated model predictive control algorithm [12]. With these observations as inspiration we knew that, when we investigated stability theory for *hybrid systems*, we had to elucidate the weakest assumptions under which asymptotic stability is automatically robust. Our work in this area started in [7, 10] and culminated in the tutorial article [8] and the research monograph [9].

Now, as my collaborators and I turn our attention to stochastic systems, the same principles guide us: we look for a stability theory that applies to a very wide class of stochastic hybrid systems and that automatically entails robustness. To cut our teeth, we began by looking at stability theory for stochastic difference inclusions. Our results for such systems are contained in [11, 37, 38, 43, 47–49]. Most recently, we have turned our attention to stochastic hybrid systems, or inclusions, which are the topic of this chapter. The results on stochastic hybrid inclusions that are recalled here are adopted from [39, 42, 44–46]. The interested reader may also wish to consult [53] for a survey of other stability theory results available in the stochastic hybrid systems literature.

Most of our focus is on Lyapunov function methods for establishing stability properties, which brings me to one more important comment about Laurent Praly: I have never seen anyone so adept at finding Lyapunov functions for nonlinear systems. At his birthday celebration we joked that the best "app" available for finding Lyapunov functions is the "Ask Laurent!" app. Around the time of Laurent's birth, in the mid-1950s, our predecessors recognized the urgency of establishing the existence of smooth Lyapunov functions [2, 17, 21–23] since Laurent was born to find them. Tongue in cheek, one of the goals of my talk at that celebration and of this paper is to encourage "Ask Laurent!" 6.0 to include the functionality of finding Lyapunov functions for stochastic hybrid systems.

The rest of this chapter is organized as follows. In Sect. 13.2 we recall some main results about stability theory and robustness for constrained differential inclusions. We hint at similar observations about constrained difference inclusions in Sect. 13.3. However, since these systems provide a special case of hybrid systems, we do not go into much detail. Instead, we discuss a robust stability theory for non-stochastic hybrid inclusions in Sect. 13.4. Then, we turn our attention to our current research on stochastic hybrid inclusions. In each section, we point out the role regularity assumptions and, in the case of stochastic systems, also causality assumptions play in guaranteeing a coherent, robust stability theory. Finally, we end with some conclusions.

13.2 Constrained Differential Inclusions

In this section, we consider asymptotic stability and global recurrence for constrained differential inclusions. Asymptotic stability is a standard property considered in the systems and control literature. Recurrence is a property that is much more common to find in the stochastic systems and control literature. This is because it is possible for a stochastic system to exhibit recurrence to an open, bounded set and yet possess no compact, forward invariant set. This phenomenon does not occur in non-stochastic systems, as we explain below.

13.2.1 Motivation

In the analysis of control systems, there are several motivations for studying differential inclusions rather than differential equations.

One motivation is the convenience of using a differential inclusion to analyze a differential equation under the influence of an arbitrary time-varying disturbance, $\dot{x}(t) = f(x(t), d(t))$, $(x(t), d(t)) \in \mathbb{R}^n \times \mathbb{R}^m$, when the disturbance is expected to satisfy a state-dependent constraint $d(t) \in S(x(t))$ for all $t \geq 0$, where $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$. The double arrows here, following the notation appearing in [30], indicate that the values of S are subsets of $\mathbb{R}^m$. In this case, we may be motivated to analyze the behavior of the differential inclusion $\dot{x} \in F(x)$, where $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is defined by $F(x) := f(x, S(x))$ for all $x \in \mathbb{R}^n$, or perhaps by $\overline{\text{co}} f(x, S(x))$ for all $x \in \mathbb{R}^n$, where $\overline{\text{co}}$ indicates taking the closed convex hull. Convex hulls are appropriate in continuous-time systems because it is possible for the derivative to switch arbitrarily fast among the available values in the set-valued map, essentially replicating the effect of any value in the convex hull of the derivative set.

Another motivation for differential inclusions occurs when a continuous feedback control system $\dot{x} = f(x, u)$, $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$, employs a discontinuous feedback function $u = k(x)$ where $k : \mathbb{R}^n \to \mathbb{R}^m$. Since discontinuous differential equations may not have solutions (in a standard sense) or because the solutions of a discontinuous differential equation may not give an accurate picture of the behavior under small perturbations, we may be motivated to study instead the differential inclusion $\dot{x} \in F(x)$, where $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is defined by $F(x) := \overline{\text{co}} f(x, K(x))$ and K is the outer semicontinuous hull[1] of k; that is, K is the set-valued mapping whose graph coincides with the closure of the graph of k, which is the set $\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m : y = k(x)\}$. This differential inclusion is sometimes called the Krasovskii regularization of the

[1] For more information, see [30, pp. 154–155].

discontinuous differential equation and the solutions of the differential inclusion are sometimes called the Krasovskii solutions of the original discontinuous differential equation. For more information, see [13] or [33].

There are also reasons to consider constrained differential equations or inclusions, of the form $x \in C$, $\dot{x} \in F(x)$. For example, a particular subset of the state space may be known to be forward invariant; restricting the state to that forward invariant set may facilitate a simpler analysis. In this case, we can take C to be the forward invariant set. Similarly, the state of such a system may evolve on a manifold embedded in a Euclidean space. In this particular case, we can take C to be the manifold on which the state evolves. One situation where evolution on a manifold is helpful is when converting a time-varying, periodic system to a time-invariant one. In this case, the state may include a clock variable that rotates around the unit circle with uniform rate. In this situation, considering initial conditions that are not constrained to this circle would make little sense. Finally, it may be known that once the state of $\dot{x} \in F(x)$ reaches a set D, it behaves as desired. Then to study whether the solutions of system eventually behave well from any initial condition, we may consider studying the behavior of the solutions of $x \in \overline{\mathbb{R}^n \backslash D} =: C$, $\dot{x} \in F(x)$. For this constrained system, we may aim to prove that each solution either behaves as desired or is forced to stop because it reaches the boundary of and attempts to leave C. In this case, as a solution of the original system, it would reach D, and then start behaving as desired.

A special case of constrained differential equations corresponds to the situation where $C = \mathbb{R}^n$ and F is a function, i.e., $\dot{x} = F(x)$.

13.2.2 Model and Solution Concept

We consider a constrained differential inclusion of the form

$$x \in C, \quad \dot{x} \in F(x). \tag{13.1}$$

Throughout the discussion of this system, we impose the following assumption:

Assumption 13.1 The data of the constrained differential inclusion (13.1) are such that

1. the set $C \subset \mathbb{R}^n$ is closed, and
2. the set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is outer semicontinuous[2] and locally bounded[3] with values on C that are nonempty and convex. ■

We note that when $f : C \to \mathbb{R}^n$ is a continuous function and C is closed, the set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by $F(x) := \{f(x)\}$ for $x \in C$ and $F(x) := \varnothing$ for $x \in \mathbb{R}^n \backslash C$ is outer semicontinuous and locally bounded with nonempty convex values on C. The outer semicontinuity follows from the fact that the graph of F is

[2] See [30, Definition 5.4].

[3] See [30, Definition 5.14].

closed in this situation. Having a closed graph is an equivalent characterization of outer semicontinuity [30, Theorem 5.7(a)]. The values are convex on C by virtue of being singletons at such points.

For the constrained differential inclusion (13.1), a *solution* is any locally absolutely continuous function x defined on a set $\mathrm{dom}(x)$ of the form $[0, T)$ or $[0, T]$ with $T \geq 0$, or $[0, \infty)$, such that $x(t) \in C$ and $\dot{x}(t) \in F(x(t))$ for almost all $t \in \mathrm{dom}(x)$. Given a set $K \subset \mathbb{R}^n$, $\mathscr{S}(K)$ denotes the set of solutions starting in the set K.

13.2.3 Asymptotic Stability: Definitions and Results

We give a brief overview of stability theory for the constrained differential inclusion (13.1), where the attractor is denoted $\mathscr{A}$. All of the subsequent results use the following assumption:

Assumption 13.2 The set $\mathscr{A} \subset \mathbb{R}^n$ is compact. ■

Since the attractor is assumed to be compact, the discussion is more general than a stability discussion for equilibria. This is especially appropriate for hybrid systems (which we consider eventually) where it is quite common for some states to persistently change their values. On the other hand, it does not allow for unbounded attractors, like might be required for the analysis of time-varying, non-periodic systems when attempting to use results for time-invariant systems. As suggested by the discussion above about clock variables as examples of state constraints, it does account for stability theory for time-varying, periodic systems.

We now give a sequence of definitions, culminating in a definition of robust, uniform, global asymptotic stability.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *Lyapunov stable* for (13.1) if, for each $\varepsilon > 0$, there exists $\delta > 0$ such that $x(t) \in \mathscr{A} + \varepsilon\mathbb{B}$ for each $x \in \mathscr{S}(\mathscr{A} + \delta\mathbb{B})$ and each $t \in \mathrm{dom}(x)$.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly Lagrange stable* for (13.1) if, for each $\delta > 0$, there exists $\varepsilon > 0$ such that $x(t) \in \mathscr{A} + \varepsilon\mathbb{B}$ for each $x \in \mathscr{S}(\mathscr{A} + \delta\mathbb{B})$ and each $t \in \mathrm{dom}(x)$.

Note that Lyapunov stability of the set $\mathscr{A}$ is a characterization of how the system behaves near the set $\mathscr{A}$ while Lagrange stability of $\mathscr{A}$ is a characterization of how the system behaves far from $\mathscr{A}$.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *globally attractive* for (13.1), if there are no finite escape times (that is, each solution is bounded on each bounded subset of its domain) and every solution with an unbounded time domain satisfies $\lim_{t \to \infty} |x(t)|_{\mathscr{A}} = 0$. In this definition, $|x|_{\mathscr{A}}$ denotes the distance of a vector x to the set $\mathscr{A}$, i.e., $|x|_{\mathscr{A}} := \inf_{y \in \mathscr{A}} |x - y|$ where $| \cdot |$ denotes the standard Euclidean norm.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly globally attractive* for (13.1) if there are no finite escape times and for each $\varepsilon > 0$ and $\Delta > 0$ there exists $T \geq 0$ such that $x(t) \in \mathscr{A} + \varepsilon\mathbb{B}$ for all $x \in \mathscr{S}(\mathscr{A} + \Delta\mathbb{B})$ and all $t \in [T, \infty) \cap \mathrm{dom}(x)$. Notice that, for a given $x \in \mathscr{S}(\mathscr{A} + \Delta\mathbb{B})$, if $[T, \infty) \cap \mathrm{dom}(x) = \emptyset$ then there is nothing to check for the solution x.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *globally asymptotically stable* (GAS) for (13.1), if it is Lyapunov stable and globally attractive.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly globally asymptotically stable* (UGAS) for (13.1) if it is Lyapunov stable, uniformly Lagrange stable, and uniformly globally attractive.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *robustly uniformly globally asymptotically stable* (R-UGAS) for (13.1), if there exists a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ satisfying $\rho(x) > 0$ for all $x \in \mathbb{R}^n \backslash \mathscr{A}$, such that $\mathscr{A}$ is UGAS for the constrained differential inclusion $x \in C_\rho$, $\dot{x} \in F_\rho(x)$, where

$$C_\rho := \{x \in \mathbb{R}^n : (x + \rho(x)\mathbb{B}) \cap C \neq \emptyset\} \tag{13.2a}$$

$$F_\rho(x) := \overline{\text{co}} F((x + \rho(x)\mathbb{B}) \cap C) + \rho(x)\mathbb{B}. \tag{13.2b}$$

The conditions of Assumptions 13.1 and 13.2 provide the somewhat surprising result that R-UGAS is not actually stronger than GAS.

Theorem 13.1 *Under Assumptions 13.1 and 13.2, the set $\mathscr{A} \subset \mathbb{R}^n$ is R-UGAS for (13.1) if and only if it is GAS for (13.1).*

For the case where $C = \mathbb{R}^n$, Theorem 13.1 is a consequence of [55, Proposition 3, Theorem 3 and Propositions 2]. For the general case, the result of Theorem 13.1 is contained in [4, Theorem 7.9].

We emphasize through an example that GAS may not imply UGAS when Assumption 13.1 does not hold.

Example 13.1 Consider the case where $\mathscr{A} := \{0\}$, $C := [0, \infty)$ and $F : \mathbb{R}^n \to \mathbb{R}^n$ where

$$F(x) := \begin{cases} -x & x \in [0, 1] \\ -\sqrt{x-1} & x \in (1, \infty). \end{cases}$$

This function is not continuous at 1. In particular, it does not have a closed graph and thus is not outer semicontinuous when viewed as a set-valued mapping. The solution starting at a point $x_\circ \in [0, 1]$ is $x(t) = \exp(-t)x_\circ$. The solution starting at a point $x_\circ > 1$ is

$$x(t) := \begin{cases} 1 + \left(\sqrt{x_\circ - 1} - 0.5t\right)^2 & t \in [0, 2\sqrt{x_\circ - 1}) \\ \exp(-t + 2\sqrt{x_\circ - 1}) & t \geq 2\sqrt{x_\circ - 1}. \end{cases}$$

Therefore the origin is UGAS. However, for each continuous function $\rho : \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfying $\rho(x) > 0$ for all $x \in \mathbb{R} \backslash \{0\}$, the differential inclusion $\dot{x} = \overline{\text{co}} F((x + \rho(x)\mathbb{B}) \cap C) + \rho(x)\mathbb{B}$ will have a solution $x(t) = c$ where c is any real number greater than or equal to one satisfying $\sqrt{c-1} \leq \rho(c)$. This condition holds for $c = 1$ and for other values near one since $\rho(1) > 0$ and ρ is continuous. ■

13.2.4 Lyapunov Functions for Asymptotic Stability

One of the most convenient ways to establish GAS (equivalently, R-UGAS) of a compact set is by means of a Lyapunov function. A function $V : \mathrm{dom}(V) \to \mathbb{R}_{\geq 0}$ is a *Lyapunov function candidate* for the set $\mathscr{A} \subset \mathbb{R}^n$ and the data (C, F) of (13.1), i.e., for $(\mathscr{A}, (C, F))$, if $C \subset \mathrm{dom}(V)$, it is continuously differentiable on an open set containing C, $V(x) = 0$ for all $x \in \mathscr{A}$, $V(x) > 0$ for all $x \in C \backslash \mathscr{A}$, and if the sequence of points $\left\{x_i\right\}_{i=1}^{\infty}$, with each point belonging to C, is unbounded then the sequence of values $\left\{V(x_i)\right\}_{i=1}^{\infty}$ is unbounded. It is a *Lyapunov function* if $\langle \nabla V(x), f \rangle < 0$ for all $x \in C \backslash \mathscr{A}$ and $f \in F(x)$. It is a *Krasovskii–LaSalle function* if $\langle \nabla V(x), f \rangle \leq 0$ for all $x \in C$ and $f \in F(x)$ and there does not exist a solution x with an unbounded time domain that renders $t \mapsto V(x(t))$ constant and nonzero. The latter condition is satisfied if V is also a Lyapunov function. It is an *exponentially decreasing Lyapunov function*, if if there exists $\lambda > 0$ such that $\langle \nabla V(x), f \rangle \leq -\lambda V(x)$ for all $x \in C$ and $f \in F(x)$.

The existence of an exponentially decreasing Lyapunov function does not necessarily imply that the solutions converge exponentially to the set $\mathscr{A}$. Indeed, it turns out that the existence of an exponentially decreasing Lyapunov function is equivalent to the GAS property.

Theorem 13.2 *Under Assumptions 13.1 and 13.2, $(\mathscr{A}, (C, F))$ admits an exponentially*

decreasing Lyapunov function if and only if $\mathscr{A}$ is GAS for (13.1).

For the case where $C = \mathbb{R}^n$, Theorem 13.1 is established via the combination of [55, Propositions 3, 2, Theorem 3 and Theorem 1]. For the general case, the result of Theorem 13.1 is contained in [4, Theorem 3.13].

Example 13.1 above shows that GAS may not imply the existence of a Lyapunov function when Assumption 13.1 is omitted. Indeed, since $\lim_{x \to 1^+} F(x) = 0$ and ∇V is continuous, $\langle \nabla V(x), F(x) \rangle$ must approach zero as $x \to 1^+$; on the other hand, $\lim_{x \to 1^+} -\lambda V(x) = -\lambda V(1) < 0$.

Fortunately, we are not required to find an exponentially decreasing Lyapunov function in order to establish R-UGAS. The existence of a Krasovskii–LaSalle function is enough, as the next theorem states.

Theorem 13.3 *Under Assumptions 13.1 and 13.2, if $(\mathscr{A}, (C, F))$ admits a Krasovskii–LaSalle function then the set $\mathscr{A}$ is R-UGAS for (13.1).*

For the case $C = \mathbb{R}^n$, the result of Theorem 13.3 can be pieced together from results in [6, Chap. 3] or [1, 31] (which locate the ω-limit set of each bounded solution and help to establish attractivity and, in turn, GAS) and [4] (which converts GAS to R-UGAS). Similarly, for the general case, Theorem 13.3 follows by combining the results of [4, 32].

The conclusion of Theorem 13.3 may fail when Assumption 13.1 is omitted, as the next example illustrates.

Example 13.2 Consider the case where $\mathscr{A} := \{0\}$, $C := [0, \infty)$ and $F : \mathbb{R}^n \to \mathbb{R}^n$ where

$$F(x) := \begin{cases} -x & x \in [0, 1] \\ -(x-1) & x \in (1, \infty). \end{cases}$$

This function is discontinuous at $x = 1$ and hence does not have a closed graph. The solution starting at a point $x_\circ \in [0, 1]$ is $x(t) = \exp(-t)x_\circ$. The solution starting at a point $x_\circ > 1$ is $x(t) = 1 + \exp(-t)[x_\circ - 1]$. Hence the origin is not globally attractive. With $V(x) = x^2$, we have that $\langle \nabla V(x), F(x) \rangle < 0$ for all $x \in C$, and there is no solution that renders $t \mapsto V(x(t))$ constant and nonzero. ■

An alternative to attempting to rule out solutions that keep V equal to a nonzero constant involves employing Matrosov functions. See [20, 34, 50]. In this approach, we do not need to know anything about solutions to the constrained differential inclusion. Instead, we must work to find additional functions whose derivatives have the effect of ruling out solutions that keep V equal to a nonzero constant. We defer to the references above for more details.

13.2.5 Recurrence: Definitions and Results

In this section, we consider an attractivity-like property, called recurrence, which plays a prominent role in the study of stochastic systems. We study it here for non-stochastic systems. We use $\mathscr{O}$ to denote the recurrent set, and assume the following:

Assumption 13.3 The set $\mathscr{O} \subset \mathbb{R}^n$ is open and bounded. ■

We give a sequence of definitions, culminating in the definition of robust, uniform global recurrence.

A set $\mathscr{O} \subset \mathbb{R}^n$ is said to be *globally recurrent* (GR) for (13.1) if there are no finite escape times and for each solution x with an unbounded time domain there exists $t \in \mathrm{dom}(x)$ such that $x(t) \in \mathscr{O}$.

A set $\mathscr{O} \subset \mathbb{R}^n$ is said to be *uniformly globally recurrent* (UGR) for (13.1) if there are no finite escape times and for each compact set $K \subset \mathbb{R}^n$ there exists $T > 0$ such that, for each solution $x \in \mathscr{S}(K)$ with a time domain that contains T, there exists $t \in \mathrm{dom}(x) \cap [0, T)$ such that $x(t) \in \mathscr{O}$.

A set $\mathscr{O} \subset \mathbb{R}^n$ is said to be *robustly uniformly globally recurrent* (R-UGR) for (13.1) if there exists a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}_{>0}$ such that $\mathscr{O}$ is uniformly globally recurrent for the constrained differential inclusion $x \in C_\rho$, $\dot{x} \in F_\rho(x)$, where the pair (C_ρ, F_ρ) is defined in (13.2).

The conditions of Assumptions 13.1 and 13.3 provide the somewhat surprising result that R-UGR is not actually stronger than global recurrence.

Theorem 13.4 *Under Assumptions 13.1 and 13.3, the set $\mathscr{O} \subset \mathbb{R}^n$ is R-UGR for (13.1) if and only if it is globally recurrent for (13.1).*

The result of Theorem 13.4 is contained in [41, Theorem 4]. Global recurrence may not imply R-UGR if either Assumption 13.1 or 13.3 does not hold.

Example 13.3 Consider the case where $F : \mathbb{R} \to \mathbb{R}$ is defined by $F(x) := x^2(1-x)$ for all $x \in \mathbb{R}$ and $\mathcal{O} := [-0.1, 0] \cup [0.9, 1.1]$. Notice that $\mathcal{O}$ is bounded but not open. The set $\mathcal{O}$ is globally recurrent since every solution starting to the left of or at the origin converges to the origin (and thus reaches $[-0.1, 0]$ in finite time) while every solution that starts to the right of the origin converges to 1 (and thus reaches $[0.9, 1.1]$ in finite time). However, $\mathcal{O}$ is not uniformly globally recurrent because the time it takes to reach $\mathcal{O}$ grows unbounded as the initial condition approaches the origin from the right. ■

Example 13.4 Consider the case where $C := \mathbb{R} \times [0, 2]$ and $f : C \to \mathbb{R}^2$ is defined by

$$f(x) := \begin{bmatrix} -x_1\lambda(x_2) \\ 0 \end{bmatrix}$$

where $\lambda : [0, 2] \to \mathbb{R}_{>0}$ satisfies

$$\lambda(x_2) := \begin{cases} 1 & x_2 = 1 \\ (1-x_2)^2 & x_2 \in [0, 2] \setminus \{1\}\,. \end{cases}$$

Let $\mathcal{O} := (-0.1, 0.1) \times (-1, 3)$. The set $\mathcal{O}$ is open and bounded, but F is not continuous. The set $\mathcal{O}$ is globally recurrent (in fact, globally attractive), but it is not uniformly globally recurrent since the time it takes the x_1 component to become small from $x_1(0) = 1$ grows unbounded as the initial value of x_2 approaches 1. ■

Finally, we can make a connection between global recurrence and global asymptotic stability. This connection is made through the definition of the Ω-limit set for (13.1) from a set of initial conditions K, defined as

$$\Omega(K) := \left\{ z \in \mathbb{R}^n : z = \lim_{i\to\infty} x_i(t_i), x_i \in \mathscr{S}(K), t_i \in \operatorname{dom}(x_i), \lim_{i\to\infty} t_i = \infty \right\}.$$

Theorem 13.5 *Under Assumptions 13.1 and 13.3, if the set $\mathcal{O} \subset \mathbb{R}^n$ is globally recurrent for (13.1) and $\Omega(\overline{\mathcal{O}})$ is nonempty then the latter is a UGAS compact set for (13.1).*

For more details about this result, see [41, Sect. 5.2]. As we will explain later, there is no reason to expect an analogous result for stochastic systems.

13.2.6 Foster Functions for Recurrence

Like for asymptotic stability, a convenient tool for establishing global recurrence is a Lyapunov-like function. A function $V : \operatorname{dom}(V) \to \mathbb{R}_{\geq 0}$ is a Lyapunov–Foster function candidate for the set $\mathcal{O} \subset \mathbb{R}^n$ and the data (C, F) of (13.1), i.e., for $(\mathcal{O}, (C, F))$, if $C \subset \mathrm{dom}(V)$, it is continuously differentiable on an open set containing C, and if the sequence of points $\{x_i\}_{i=1}^{\infty}$, with each point belonging to C, is unbounded then the sequence of values $\{V(x_i)\}_{i=1}^{\infty}$ is unbounded. (There is no requirement that $V(x) = 0$ for $x \in \mathcal{O}$.) It is a Lyapunov–Foster function if $\langle \nabla V(x), f \rangle < 0$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$. It is a Krasovskii–LaSalle–Foster function if $\langle \nabla V(x), f \rangle \leq 0$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$ and there does not exist a solution x with an unbounded time domain that never intersects $\mathcal{O}$ and that renders $t \mapsto V(x(t))$ constant. The latter property is satisfied if V is also a Lyapunov–Foster function. It is a uniformly decreasing Lyapunov–Foster function if there exists $\lambda > 0$ such that $\langle \nabla V(x), f \rangle \leq -\lambda$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$.

Theorem 13.6 *Under Assumptions 13.1 and 13.3, $(\mathcal{O}, (C, F))$ admits a uniformly decreasing Lyapunov–Foster function if and only if $\mathcal{O}$ is globally recurrent for (13.1).*

Theorem 13.6 was established for more general (i.e., hybrid) systems in [41, Theorem 5].

Regarding the necessary and sufficient conditions for global recurrence in Theorem 13.6, Example 13.1 above, with $\mathcal{O}$ a small open neighborhood of the origin, provides a counterexample to the necessity when Assumption 13.1 is omitted.

We are not required to find a uniformly decreasing Lyapunov–Foster function in order to prove robust, uniform global recurrence, as the next theorem states.

Theorem 13.7 *Under Assumptions 13.1 and 13.3, if $(\mathcal{O}, (C, F))$ admits a Krasovskii–LaSalle–Foster then the set $\mathcal{O}$ is R-UGR for (13.1).*

The result of Theorem 13.7 is a combination of the results in [6, Chap. 3] or [1, 31] (about locating the ω-limit set of each bounded solution that has an unbounded time domain, to prove recurrence) and [41, Theorem 4] (on the equivalence of recurrence and R-UGR).

Example 13.2 above, with $\mathcal{O}$ a small open neighborhood of the origin, provides a counterexample to the conclusion of Theorem 13.7 when Assumption 13.1 is omitted.

The idea behind Matrosov functions can also be applied easily to rule out solutions x that never intersect $\mathcal{O}$ and render $t \mapsto V(x(t))$ constant.

13.3 Constrained Difference Inclusions

In this section, we allude to results on asymptotic stability and global recurrence for constrained difference inclusions. We do not go into detail since the available results are very similar to those available for constrained differential inclusions and because they are contained in the upcoming results for hybrid inclusions.

13.3.1 Motivation

The motivations for considering constrained difference inclusions, rather than just constrained difference equations, are analogous to the motivations for considering constrained differential inclusions. However, when constructing the set-valued mappings that prescribe the possible next value of the state, there is no reason to invoke the convex hull, as there is no analogy to arbitrarily fast switching in discrete-time systems. Like in continuous-time systems, sometimes discontinuous feedbacks are useful, or even necessary [29], or simply manifest themselves when the control is designed by solving an optimization problem like in model predictive control [27]. But the best predictor of the behavior of a discontinuous system under small perturbations is the difference inclusion that uses the outer semicontinuous hull of the discontinuous function [16, 33]. Recall that the outer semicontinuous hull of a discontinuous mapping is the unique set-valued mapping whose graph is equal to the closure of the graph of the original mapping.

Also like in continuous-time systems, there are many reasons to consider constrained difference inclusions. In addition to the reasons encountered for continuous-time systems, another natural reason is that discrete-time systems often naturally include variables that take values in a discrete set and so it makes no sense to consider solutions from all initial conditions in the underlying Euclidean space.

13.3.2 Model and Solutions

The model of a constrained difference inclusion has the form

$$x \in D, \quad x^+ \in G(x). \tag{13.3}$$

These systems are often studied under the following conditions:

Assumption 13.4 The data of the constrained difference inclusion (13.3) are such that

1. the set $D \subset \mathbb{R}^n$ is closed, and
2. the set-valued mapping $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is outer semicontinuous and locally bounded with values on D that are nonempty. ■

For a constrained difference inclusion $x \in D, x^+ \in G(x)$, a *solution* is any function x defined on a set $\mathrm{dom}(x)$ of the form $\{0, \ldots, k\}$, where k is a nonnegative integer, or $\mathbb{Z}_{\geq 0}$, such that $x(0) \in D$ and if both j and $j + 1$ belong to $\mathrm{dom}(x)$ then $x(j) \in D$ and $x(j + 1) \in G(x(j))$.

13.3.3 Asymptotic Stability: Basic Definitions and Results

The definitions of stability for constrained difference equations parallel those for constrained differential equations but with j replacing t as the time variable. Also, in the definitions of R-UGAS and R-UGR, the inflated system is $x \in D_\rho$, $x^+ \in G_\rho(x)$ where

$$D_\rho := \{x \in \mathbb{R}^n : (x + \rho(x)\mathbb{B}) \cap D \neq \varnothing\} \tag{13.4a}$$

$$G_\rho(x) := \{g \in \mathbb{R}^n : g + \rho(g)\mathbb{B}, g \in G((x + \rho(x)\mathbb{B}) \cap D)\} . \tag{13.4b}$$

For discrete-time systems, there is no particular need for a Lyapunov function to be continuously differentiable, or even continuous, as long as it is uniformly decreasing along solutions and it can be upper and lower bounded by $\mathscr{K}_\infty$ functions of the distance of the state to the attractor $\mathscr{A}$.

The theorems and counterexamples of Sect. 13.2 for constrained differential inclusions apply to constrained difference inclusions, *mutatis mutandis*. Moreover, such theorems are special cases of upcoming results for hybrid systems. Hence, those results are omitted here.

13.4 Hybrid Inclusions

Now we turn our attention to hybrid inclusions, demonstrating results pertaining to asymptotic stability and recurrence that parallel results in continuous-time systems and discrete-time systems.

13.4.1 Motivation

Hybrid systems, or perhaps more appropriately "hybrid inclusions," combine constrained differential inclusions (13.1) and constrained difference inclusions (13.3) [9]. One strong motivation for studying hybrid systems stems from the role that hybrid feedback can play in robustly stabilizing nonlinear continuous-time systems. For example, logic-based switching control has been shown to be useful for stabilizing the origin of difficult systems like the non-holonomic integrator [15]. In addition, hysteresis is a very effective mechanism for achieving robust, global stabilization of a point on a manifold without boundary [24–26]. Moreover, the hybrid systems formalism can address a wide range of systems, including mechanical systems with impacts and networked control systems, which combine continuous-time evolution and communication logic and switching; see, for example, [8].

While constrained differential inclusions and constrained difference inclusions can exhibit nonunique solutions, because the allowed derivative or the allowed next value is not unique, for hybrid systems nonuniqueness can also arise at points in the state space where both continuous evolution and instantaneous change are allowed.

13.4.2 Model and Solutions

A model of a hybrid inclusion is written formally as

$$x \in C \quad \dot{x} \in F(x) \tag{13.5a}$$

$$x \in D \quad x^+ \in G(x). \tag{13.5b}$$

The data (C, F) and (D, G) are supposed to satisfy Assumptions 13.1 and 13.4, respectively.

For a hybrid system, each solution is defined on a hybrid time domain, which combines continuous time and discrete time. A *compact hybrid time domain* is a set of the form $\cup_{i=0}^{J}\left([t_i, t_{i+1}] \times \{i\}\right)$ where $J \in \mathbb{Z}_{\geq 0}$, and $0 = t_0 \leq t_1 \leq \cdots \leq t_{J+1}$. A *hybrid time domain* is a set $E \subset \mathbb{R}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ such that, for each $(T, J) \in E$, the set $E \cap ([0, T] \times \{0, \ldots, J\})$ is a compact hybrid time domain.

A solution of the hybrid system (13.5) is a function x defined on a hybrid time domain $\mathrm{dom}(x)$ such that $t \mapsto x(t, j)$ is locally absolutely continuous for each $j \in \mathbb{Z}_{\geq 0}$, $x(0, 0) \in C \cup D$, and

1. if $(t_1, j), (t_2, j) \in \mathrm{dom}(x)$ with $t_1 < t_2$ then, for almost all $t \in [t_1, t_2]$,

$$x(t, j) \in C \tag{13.6a}$$

$$\dot{x}(t, j) \in F(x(t, j)); \tag{13.6b}$$

2. if $(t, j), (t, j + 1) \in \mathrm{dom}(x)$ then

$$x(t, j) \in D \tag{13.7a}$$

$$x(t, j + 1) \in G(x(t, j)). \tag{13.7b}$$

13.4.3 Asymptotic Stability: Basic Definitions and Results

Stability theory for an attractor $\mathscr{A}$ will be discussed under Assumption 13.2 together with Assumptions 13.1 and 13.4. The definitions of stability parallel those for continuous-time and discrete-time systems. We make those definitions explicit here to be clear.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *Lyapunov stable* for (13.5) if, for each $\varepsilon > 0$, there exists $\delta > 0$ such that $x(t, j) \in \mathscr{A} + \varepsilon\mathbb{B}$ for each $x \in \mathscr{S}(\mathscr{A} + \delta\mathbb{B})$ and each $(t, j) \in \mathrm{dom}(x)$. For future reference, it is worth noting that this condition is equivalent to asking that

$$\text{graph}(x) \subset \mathbb{R}^2 \times (\mathscr{A} + \varepsilon\mathbb{B}) \qquad \forall x \in \mathscr{S}(\mathscr{A} + \delta\mathbb{B}) \tag{13.8}$$

where

$$\text{graph}(x) := \{(t,j,z) : (t,j) \in \text{dom}(x),\ z = x(t,j)\}\,. \tag{13.9}$$

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly Lagrange stable* for (13.5) if, for each $\delta > 0$, there exists $\varepsilon > 0$ such that $x(t,j) \in \mathscr{A} + \varepsilon\mathbb{B}$ for each $x \in \mathscr{S}(\mathscr{A} + \delta\mathbb{B})$ and each $(t,j) \in \text{dom}(x)$, that is, (13.8) holds.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *globally attractive* for (13.5) if there are no finite escape times and every solution with an unbounded time domain satisfies $\lim_{t+j\to\infty} |x(t,j)|_{\mathscr{A}} = 0$.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly globally attractive* for (13.5) if there are no finite escape times and, for each $\varepsilon > 0$ and $\Delta > 0$, there exists $T \geq 0$ such that $x(t,j) \in \mathscr{A} + \varepsilon\mathbb{B}$ for all $x \in \mathscr{S}(\mathscr{A} + \Delta\mathbb{B})$ and all $(t,j) \in \text{dom}(x)$ satisfying $t + j \geq T$; in other words, defining $\mathscr{T}_{\geq T} := \left\{(s,i) \in \mathbb{R}_{\geq 0} \times \mathbb{Z}_{\geq 0} : s + i \geq T\right\}$, we have

$$\text{graph}(x) \cap (\mathscr{T}_{\geq T} \times \mathbb{R}^n) \subset \mathbb{R}^2 \times (\mathscr{A} + \varepsilon\mathbb{B}). \tag{13.10}$$

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *globally asymptotically stable* (GAS) for (13.5), if it is Lyapunov stable and globally attractive.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly globally asymptotically stable* (UGAS) for (13.5) if it is Lyapunov stable, uniformly Lagrange stable, and uniformly globally attractive.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *robustly uniformly globally asymptotically stable* (R-UGAS) for (13.5), if there exists a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ satisfying $\rho(x) > 0$ for all $x \in \mathbb{R}^n \backslash \mathscr{A}$, such that $\mathscr{A}$ is uniformly globally asymptotically stable for the hybrid system

$$x \in C_\rho \quad \dot{x} \in F_\rho(x) \tag{13.11a}$$

$$x \in D_\rho \quad x^+ \in G_\rho(x) \tag{13.11b}$$

where the pair (C_ρ, F_ρ) is defined in (13.2) and the pair (D_ρ, G_ρ) is defined in (13.4).

The main stability theory results also carry over from the continuous-time and discrete-time settings. For example, the conditions of Assumptions 13.2 with Assumptions 13.1 and 13.4 provide the somewhat surprising result that R-UGAS is not actually stronger than GAS.

Theorem 13.8 *Under Assumptions 13.1, 13.2, and 13.4, the set $\mathscr{A} \subset \mathbb{R}^n$ is R-UGAS for (13.5) if and only if it is GAS for (13.5).*

The result of Theorem 13.8 is taken from [4, Theorem 7.9].

Example 13.1 illustrates that GAS may not imply R-UGAS when Assumption 13.1 does not hold. Here, we provide an alternative example that involves continuous functions but where C is not closed. A similar example can be constructed for D not closed.

Example 13.5 Let $\mathscr{A} \subset \mathbb{R}$ be the origin, and consider the data (C, f, D, g) where $C := [0, 1), f : C \to \mathbb{R}$ is given by $f(x) = -x(1 - x)$ for all $x \in C$, $D := [1, \infty)$, and $g(x) := 0$ for all $x \in D$. Each solution starting in C never jumps and converges to the origin monotonically. All solutions starting in D jump once, to the origin, and then flow, remaining at the origin for all subsequent time. Hence, the origin is GAS. However, it is not R-UGAS, as any inflation of the data would satisfy $[0, 1] \subset C_\rho$ and $0 \in F_\rho(1)$, thereby admitting a solution that remains at 1, while flowing, for all time. ■

13.4.4 Lyapunov Functions for Asymptotic Stability

Again, a Lyapunov function is a convenient tool for establishing asymptotic stability. Moreover, converse Lyapunov theorems establish that it is reasonable to search for Lyapunov functions for hybrid inclusions. A function $V : \mathrm{dom}(V) \to \mathbb{R}_{\geq 0}$ is a *Lyapunov function candidate* for the set $\mathscr{A} \subset \mathbb{R}^n$ and the hybrid data (C, F, D, G), i.e., for $(\mathscr{A}, (C, F, D, G))$, if $C \cup D \cup G(D) \subset \mathrm{dom}(V)$, it is continuous on its domain, it is continuously differentiable on an open set containing C, $V(x) = 0$ for all $x \in \mathrm{dom}(V) \cap \mathscr{A}$, $V(x) > 0$ for all $x \in C \cup D \backslash \mathscr{A}$, and if the sequence of points $\left\{x_i\right\}_{i=1}^{\infty}$, with each point belonging to $C \cup D$, is unbounded then the sequence of values $\left\{V(x_i)\right\}_{i=1}^{\infty}$ is unbounded. It is a *Lyapunov function* if $\langle \nabla V(x), f \rangle < 0$ for all $x \in C \backslash \mathscr{A}$ and $f \in F(x)$ and $V(g) - V(x) < 0$ for all $x \in D \backslash \mathscr{A}$ and $g \in G(x)$. It is a *Krasovskii–LaSalle function* if $\langle \nabla V(x), f \rangle \leq 0$ for all $x \in C$ and $f \in F(x)$, $V(g) - V(x) \leq 0$ for all $x \in D$ and $g \in G(x)$, and there does not exist a solution x with an unbounded time domain that renders $(t, j) \mapsto V(x(t, j))$ constant and nonzero. The latter condition is satisfied if V is also a Lyapunov function. It is an *exponentially decreasing Lyapunov function* if if there exists $\lambda > 0$ such that $\langle \nabla V(x), f \rangle \leq -\lambda V(x)$ for all $x \in C$ and $f \in F(x)$, and $V(g) \leq \exp(-\lambda)V(x)$ for all $x \in D$ and $g \in G(x)$.

The following results parallel the earlier Theorems 13.2 and 13.3.

Theorem 13.9 *Under Assumptions 13.1, 13.2, and 13.4, $(\mathscr{A}, (C, F, D, G))$ admits an exponentially decreasing Lyapunov function if and only if $\mathscr{A}$ is GAS for (13.5).*

The result of Theorem 13.9 is contained in [4, Theorem 3.13].

Theorem 13.10 *Under Assumptions 13.1, 13.2, and 13.4, if $(\mathscr{A}, (C, F, D, G))$ admits a*

Krasovskii–LaSalle function then the set $\mathscr{A}$ is R-UGAS for (13.12)

For the general case, Theorem 13.10 follows by combining the results of [32] and [4].

Like for continuous-time and discrete-time systems, an alternative to attempting to rule out solutions that keep V equal to a nonzero constant involves employing Matrosov functions. See [34]. In this approach, we do not need to know anything about solutions to the hybrid system. Instead, we must work to find additional functions whose derivatives have the effect of ruling out solutions that keep V equal to a nonzero constant.

13.4.5 Recurrence: Definitions and Results

In this section, we consider recurrence for hybrid inclusions. Again, we use $\mathcal{O}$ to denote the recurrent set.

A set $\mathcal{O} \subset \mathbb{R}^n$ is said to be *globally recurrent* (GR) for (13.5) if there are no finite escape times and, for each solution x with an unbounded time domain, there exists $(t,j) \in \mathrm{dom}(x)$ such that $x(t,j) \in \mathcal{O}$.

A set $\mathcal{O} \subset \mathbb{R}^n$ is said to be *uniformly globally recurrent* (UGR) for (13.5) if there are no finite escape times and for each compact set $K \subset \mathbb{R}^n$ there exists $T > 0$ such that, for each solution $x \in \mathcal{S}(K)$, either $\mathrm{dom}(x) \subset \left\{(t,j) \in \mathbb{R}_{\geq 0} \times \mathbb{Z}_{\geq 0} : t + j < T\right\}$ $=: \mathcal{T}_{<T}$ or else there exists $(t,j) \in \mathrm{dom}(x) \cap \mathcal{T}_{<T}$ such that $x(t,j) \in \mathcal{O}$.

A set $\mathcal{O} \subset \mathbb{R}^n$ is said to be *robustly uniformly globally recurrent* (R-UGR) for (13.5) if there exists a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}_{>0}$ such that $\mathcal{O}$ is uniformly globally recurrent for (13.11) where (C_ρ, F_ρ) are defined in (13.2) and (D_ρ, G_ρ) are defined in (13.4).

The conditions of Assumptions 13.1, 13.3, and 13.4 provide the (by now expected) result that R-UGR is not actually stronger than recurrence.

Theorem 13.11 *Under Assumptions 13.1, 13.3, and 13.4, the set $\mathcal{O} \subset \mathbb{R}^n$ is R-UGR for (13.5) if and only if it is globally recurrent for (13.5).*

The result of Theorem 13.11 was established in [41, Theorem 4].

Finally, like for continuous-time and discrete-time systems, we can make a connection between recurrence and global asymptotic stability. This connection is made through the definition of the Ω-limit set for (13.5) from a set of initial conditions K, defined as

$$\Omega(K) := \left\{ z \in \mathbb{R}^n : z = \lim_{i\to\infty} x_i(t_i, j_i), x_i \in \mathcal{S}(K), (t_i, j_i) \in \mathrm{dom}(x_i), \lim_{i\to\infty} t_i + j_i = \infty \right\}.$$

Theorem 13.12 *Under Assumptions 13.1, 13.3, and 13.4, if the set $\mathcal{O} \subset \mathbb{R}^n$ is recurrent for (13.1) and $\Omega(\overline{\mathcal{O}})$ is nonempty then the latter is a UGAS compact set for (13.5).*

For more details, see [41, Sect. 5.2].

13.4.6 Foster Functions for Recurrence

A function $V : \mathrm{dom}(V) \to \mathbb{R}_{\geq 0}$ is a *Lyapunov–Foster function candidate* for the set $\mathcal{O} \subset \mathbb{R}^n$ and the data (C, F, D, G) of (13.5), i.e., for $(\mathcal{O}, (C, F, D, G))$, if $C \cup D \cup G(D) \subset \mathrm{dom}(V)$, it is continuous on its domain, continuously differentiable on an open set containing C, and if the sequence of points $\left\{x_i\right\}_{i=1}^{\infty}$, with each point belonging to $C \cup D$, is unbounded then the sequence of values $\left\{V(x_i)\right\}_{i=1}^{\infty}$ is unbounded. It is a *Lyapunov–Foster function* if $\langle \nabla V(x), f\rangle < 0$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$ and

$V(g) - V(x) < 0$ for all $x \in D\backslash\mathcal{O}$ and $g \in G(x)$. It is a *Krasovskii–LaSalle–Foster function* if $\langle \nabla V(x), f\rangle \leq 0$ for all $x \in C\backslash\mathcal{O}$ and $f \in F(x)$, $V(g) - V(x) \leq 0$ for all $x \in D\backslash\mathcal{O}$ and $g \in G(x)$, and there does not exist a solution x with an unbounded time domain that never intersects $\mathcal{O}$ and that renders $(t, j) \mapsto V(x(t, j))$ constant. The latter is satisfied if V is also a Lyapunov–Foster function. It is a *uniformly decreasing Lyapunov–Foster function* if there exists $\lambda > 0$ such that $\langle \nabla V(x), f\rangle \leq -\lambda$ for all $x \in C\backslash\mathcal{O}$ and $f \in F(x)$ and $V(g) - V(x) \leq -\lambda$ for all $x \in D\backslash\mathcal{O}$ and $g \in G(x)$.

Theorem 13.13 *Under Assumptions 13.1, 13.3, and 13.4, $(\mathcal{O}, (C, F, D, G))$ admits a uniformly decreasing Foster function if and only if $\mathcal{O}$ is globally recurrent for (13.5).*

Theorem 13.13 was established in [41, Theorem 5].

Theorem 13.14 *Under Assumptions 13.1, 13.3, and 13.4, if $(\mathcal{O}, (C, F, D, G))$ admits a*
Krasovskii–LaSalle–Foster then the set $\mathcal{O}$ is R-UGR for (13.5).

The result of Theorem 13.14 is a combination of the results in [32] (about locating the ω-limit set of each bounded solution that has an unbounded time domain) and [41, Theorem 4] (on the equivalence of recurrence and R-UGR).

Example 13.2 above, with $\mathcal{O}$ a small open neighborhood of the origin, provides a counterexample to the conclusion of Theorem 13.14 when Assumption 13.1 is omitted.

13.5 A Class of Stochastic Hybrid Inclusions

13.5.1 Motivation

Now, we get to the setting that has been developed most recently. In particular, we discuss stochastic hybrid inclusions. While some results for these systems have been developed for the case where both the flows and the jumps are affected by stochastic inputs, much more progress has been made for the case where randomness appears only in the jumps. For simplicity, we focus on that case here, but refer the reader to [46, 53] for what has been developed for the more general setting.

In addition to finding motivation in the same types of problems that motivated hybrid inclusions, the study of stochastic hybrid inclusions is motivated by problems that include analyzing the effect of random-in-time updates in networked control systems [14], random updates in sampled-data multi-agent systems to eliminate correlated actions [28], randomness in algorithms used for consensus on manifolds (see [36] and the references therein), like the circle, and sampled-data nonlinear observer problems with randomness in the measurements [3], to name just a few.

13.5.2 Model and Solutions

A model of a stochastic hybrid inclusion with randomness only in the jumps has the form

$$x \in C, \quad \dot{x} \in F(x) \tag{13.12a}$$
$$x \in D, \quad x^+ \in G(x, v^+) \qquad v \sim \mu(\cdot) \tag{13.12b}$$

where v^+ is a placeholder for a sequence of independent, identically distributed (iid) random variables, defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with distribution μ; that is, letting $\{\mathbf{v}_i\}_{i=1}^{\infty}$ denote the sequence of random variables, we have $\mu(A) = \mathbb{P}\left(\omega \in \Omega : \mathbf{v}_i(\omega) \in A\right)$ for each $i \in \mathbb{Z}_{\geq 1}$ and each $A \in \mathbf{B}(\mathbb{R}^m)$, the latter being the Borel σ-field on $\mathbb{R}^m$. We use $\mathscr{V}$ to denote the set of all possible values of $\mathbf{v}_i(\omega), \omega \in \Omega, i \in \mathbb{Z}_{\geq 1}$. A solution $\mathbf{x}$ from some point $x \in C \cup D$, denoted $\mathbf{x} \in \mathscr{S}_r(x)$, is a mapping from Ω to hybrid arcs such that, for almost every $\omega \in \Omega$, and with the definition $\phi_\omega := \mathbf{x}(\omega)$,

1. $\phi_\omega(0,0) = x$;
2. if $(t_1,j), (t_2,j) \in \mathrm{dom}(\phi_\omega)$ with $t_1 < t_2$ then, for almost all $t \in [t_1, t_2]$,

$$\phi_\omega(t,j) \in C \tag{13.13a}$$
$$\dot{\phi}_\omega(t,j) \in F(\phi_\omega(t,j)) \tag{13.13b}$$

3. If $(t,j), (t,j+1) \in \mathrm{dom}(\phi_\omega)$ then

$$\phi_\omega(t,j) \in D \tag{13.14a}$$
$$\phi_\omega(t,j+1) \in G(\phi_\omega(t,j), \mathbf{v}_{j+1}(\omega)). \tag{13.14b}$$

In addition, to qualify as a solution, $\mathbf{x}$ must have an appropriate measurability property. This property, defined below, enables measuring probabilities and expected values associated with solutions; it also enforces a causal relationship between solutions and the sequence of random variable inputs. We express measurability in terms of the graphs of the sample paths $\mathbf{x}(\omega)$ (see (13.9)). A solution is required to be such that, for each $i \in \mathbb{Z}_{\geq 0}$, the set-valued mapping

$$\omega \mapsto \mathrm{graph}(\mathbf{x}(\omega)) \cap \left(\mathbb{R}_{\geq 0} \times \{0, \ldots, i\} \times \mathbb{R}^n\right) \tag{13.15}$$

is an $\mathscr{F}_i$-measurable set-valued mapping, where $\mathscr{F}_0 := \{\emptyset, \Omega\}$, and $\{\mathscr{F}_i\}_{i=1}^{\infty}$ is the natural filtration associated to $\{\mathbf{v}_i\}_{i=1}^{\infty}$; that is, the space of events

$$\mathscr{F}_i := \left\{\left\{\omega \in \Omega : (\mathbf{v}_1(\omega), \ldots, \mathbf{v}_i(\omega)) \in A\right\}, A \in \mathbf{B}((\mathbb{R}^m)^i)\right\}.$$

This particular measurability assumption imposes causality constraints on how the solution depends on the random inputs. For more details, see [42]. See also Example 13.6 below. It also guarantees that $\omega \mapsto \mathrm{graph}(\mathbf{x}(\omega))$ is $\mathscr{F}$-measurable [30, Proposition 14.11(b)]. For more information about measurability of set-valued mappings, see [30, Chap. 14].

13.5.3 Asymptotic Stability: Basic Definitions and Results

Stability theory for an attractor $\mathscr{A}$ proceeds under Assumptions 13.1, 13.2, and the following extension of Assumption 13.4.

Assumption 13.5 The data (D, G) are such that

1. the set $D \subset \mathbb{R}^n$ is closed, and
2. the set-valued mapping $v \mapsto \mathrm{graph} G(\cdot, v) := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in G(x, v)\}$ is measurable with closed values. ■

The fact that the values of $v \mapsto \mathrm{graph} G(\cdot, v)$ are closed implies that, for each $v \in \mathbb{R}^m$, the set-valued mapping $G(\cdot, v)$ is outer semicontinuous.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *Lyapunov stable in probability* for (13.5) if, for each $\varepsilon > 0$ and $\rho > 0$ there exists $\delta > 0$ such that

$$\mathbb{P}\left(\mathrm{graph}(\mathbf{x}) \subset \mathbb{R}^2 \times (\mathscr{A} + \varepsilon\mathbb{B})\right) \geq 1 - \rho \qquad \forall \mathbf{x} \in \mathscr{S}_r(\mathscr{A} + \delta\mathbb{B}). \tag{13.16}$$

The reader may wish to compare the condition in (13.16) to the condition in (13.8).

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly Lagrange stable in probability* for (13.5) if, for each $\delta > 0$ and $\rho > 0$ there exists $\varepsilon > 0$ such that (13.16) holds.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *globally attractive almost surely* for (13.5) if, for every solution $\mathbf{x}$ and almost every $\omega \in \Omega$, $\phi_\omega := \mathbf{x}(\omega)$ is a hybrid arc without finite escape time and if its time domain is unbounded then $\lim_{t+j\to\infty} |\phi_\omega(t, j)|_{\mathscr{A}} = 0$.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly globally attractive in probability* for (13.5) if every solution is almost surely bounded and, for each $\varepsilon > 0$, $\rho > 0$ and $\Delta > 0$, there exists $T \geq 0$ such that, for all $\mathbf{x} \in \mathscr{S}_r(\mathscr{A} + \Delta\mathbb{B})$,

$$\mathbb{P}\left(\mathrm{graph}(x) \cap (\mathscr{T}_{\geq T} \times \mathbb{R}^2) \subset \mathbb{R}^2 \times (\mathscr{A} + \varepsilon\mathbb{B})\right) \geq 1 - \rho. \tag{13.17}$$

The reader may wish to compare the condition in (13.17) to the condition in (13.10).

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *globally asymptotically stable in probability* (GASp) for (13.5) if it is Lyapunov stable in probability and globally attractive almost surely.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *uniformly globally asymptotically stable in probability* (UGASp) for (13.5) if it is Lyapunov stable in probability, uniformly Lagrange stable in probability, and uniformly globally attractive in probability.

A set $\mathscr{A} \subset \mathbb{R}^n$ is said to be *robustly uniformly globally asymptotically stable in probability* (R-UGASp) for (13.5) if there exists a continuous function $\rho : \mathbb{R}^n \to$

$\mathbb{R}_{\geq 0}$ satisfying $\rho(x) > 0$ for all $x \in \mathbb{R}^n \backslash \mathscr{A}$, such that $\mathscr{A}$ is uniformly globally asymptotically stable for the hybrid system

$$x \in C_\rho \quad \dot{x} \in F_\rho(x) \tag{13.18a}$$

$$x \in D_\rho \quad x^+ \in G_\rho(x, v^+) \tag{13.18b}$$

where the pair (C_ρ, F_ρ) is defined in (13.2) and the pair (D_ρ, G_ρ) is such that D_ρ is defined in (13.4) and

$$G_\rho(x, v^+) := \left\{ g \in \mathbb{R}^n \, : \, g + \rho(g)\mathbb{B}, g \in G((x + \rho(x)\mathbb{B}, v^+) \cap D) \right\}. \tag{13.19}$$

In turns out that, for the system (13.12) under Assumptions 13.1, 13.2 and 13.4, GASp is equivalent to UGASp. On the other hand, so far we have established the equivalence between GASp and R-UGASp only for the case where $C = \varnothing$, i.e., for stochastic difference inclusions. However, a more general result does not appear to face any obstructions and should be forthcoming.

Theorem 13.15 *Under Assumptions 13.1, 13.2 and 13.4, the set $\mathscr{A} \subset \mathbb{R}^n$ is UGASp for (13.12) if and only if it is GASp for (13.12).*

The result of Theorem 13.15 is contained in [45].

Theorem 13.16 *Suppose $C = \varnothing$. Under Assumptions 13.1, 13.2 and 13.4, the set $\mathscr{A} \subset \mathbb{R}^n$ is R-UGASp for (13.12) if and only if it is GASp for (13.12).*

The result of Theorem 13.16 is contained in [49].

13.5.4 Lyapunov Functions for Asymptotic Stability

Lyapunov functions prove useful for stochastic hybrid systems as well. A function $V : \mathrm{dom}(V) \to \mathbb{R}_{\geq 0}$ is a *Lyapunov function candidate* for the set $\mathscr{A} \subset \mathbb{R}^n$ and the hybrid data (C, F, D, G, μ), i.e., for $(\mathscr{A}, (C, F, D, G, \mu))$, if $C \cup D \cup G(D, \mathscr{V}) \subset \mathrm{dom}(V)$, it is continuous on its domain, it is continuously differentiable on an open set containing C, $V(x) = 0$ for all $x \in \mathrm{dom}(V) \cap \mathscr{A}$, $V(x) > 0$ for all $x \in C \cup D \backslash \mathscr{A}$, and if the sequence of points $\left\{x_i\right\}_{i=1}^\infty$, with each point belonging to $C \cup D$, is unbounded then the sequence of values $\left\{V(x_i)\right\}_{i=1}^\infty$ is unbounded. It is a *Lyapunov function* if $\langle \nabla V(x), f \rangle < 0$ for all $x \in C \backslash \mathscr{A}$ and $f \in F(x)$ and

$$\int_{\mathbb{R}^m} \max_{g \in G(x,v)} V(g)\mu(dv) < V(x) \qquad \forall x \in D \backslash \mathscr{A}. \tag{13.20}$$

It is a *Krasovskii–LaSalle function* if $\langle \nabla V(x), f \rangle \leq 0$ for all $x \in C$ and $f \in F(x)$,

$$\int_{\mathbb{R}^m} \max_{g\in G(x,v)} V(g)\mu(dv) \le V(x) \qquad \forall x \in D, \tag{13.21}$$

and there does not exist a solution $\mathbf{x}$ with unbounded time domain almost surely that renders V constant and nonzero almost surely. The latter property holds if V is also a Lyapunov function.

The following results parallel the earlier Theorems 13.2 and 13.3.

Theorem 13.17 *Suppose $C = \emptyset$. Under Assumptions 13.1, 13.2, and 13.4, $(\mathscr{A}, (C, F, D, G, \mu))$ admits a Lyapunov function if and only if $\mathscr{A}$ is GASp for (13.5).*

The result of Theorem 13.17 is contained in [49].

Theorem 13.18 *Under Assumptions 13.1, 13.2, and 13.4, if $(\mathscr{A}, (C, F, D, G, \mu))$ admits a Krasovskii–LaSalle function then the set $\mathscr{A}$ is UGASp for (13.12) (and R-UGASp when $C = \emptyset$).*

The first part of Theorem 13.18 follows by combining the results in [39] with Theorem 13.15, which comes from [45]. The result for $C = \emptyset$ follows by combining the results in [47] with Theorem 13.16, which comes from [49].

The following example illustrates that the results of Theorems 13.17 and 13.18 would fail if the causality constraint, i.e., the $\mathscr{F}_i$-measurability of the set-valued mapping in (13.15), were not a part of the solution definition.

Example 13.6 Consider the stochastic difference inclusion

$$x_1 \in \{-0.6, 0.6\} \qquad x_1^+ \in \{-0.6, 0.6\} \tag{13.22a}$$

$$x_2 \in \mathbb{R} \qquad x_2^+ = (x_1 + v^+)x_2 \tag{13.22b}$$

where the distribution μ associated to the iid random process driving the system satisfies $\mu(\{-0.6\}) = \mu(\{0.6\}) = 0.5$. Given $x \in D := \{-0.6, 0.6\} \times \mathbb{R}$, the mapping $\omega \mapsto \mathbf{x}(\omega)$ defined for almost all $\omega \in \Omega$ by $\mathbf{x}(\omega) := \phi_\omega$, $\phi_\omega(0) := x$ and, for all $j \in \mathbb{Z}_{\ge 0}$,

$$\phi_{1,\omega}(j+1) = \mathbf{v}_{j+2}(\omega) \in \{-0.6, 0.6\} \tag{13.23a}$$

$$\phi_{2,\omega}(j+1) = \left(\phi_{1,\omega}(j) + \mathbf{v}_{j+1}(\omega)\right) \phi_{2,\omega}(j) \tag{13.23b}$$

satisfies the appropriate recursion but not the causality constraint associated with the set-valued mapping in (13.15). This particular mapping has the property that

$$\mathbb{P}\left(|\mathbf{x}_2(j)| = (1.2)^j |x_2(0)| \quad \forall j \in \mathbb{Z}_{\ge 0}\right) = 0.5. \tag{13.24}$$

In particular, the compact set $\mathscr{A} := \{-0.6, 0.6\} \times \{0\}$ is not GASp.

On the other hand, it is straightforward to show that the function $V(x) = x_2^2$ is a Lyapunov function for $(\mathscr{A}, (D, G))$. ■

Once again, an alternative to attempting to rule out solutions that keep V equal to a nonzero constant almost surely involves employing Matrosov functions. See [42]. In this approach, we do not need to know anything about solutions to the stochastic hybrid inclusion. Instead, we must work to find additional functions that have the effect of ruling out solutions that keep V equal to a nonzero constant almost surely.

13.5.5 Recurrence: Definitions and Results

In this section, we study recurrence. The recurrence property plays a more important role for stochastic systems than it does for non-stochastic systems. This is because, in the stochastic case, a system may exhibit a recurrent, open, and bounded set but not exhibit an asymptotically stable compact set. For example, consider the system with $C = \emptyset$, $D = \mathbb{R}$, $g(x, v^+) = v^+$ and such that μ does not have compact support, perhaps because it corresponds to a Gaussian distribution. In this case, there is no compact set that is almost surely forward invariant, and thus no compact set that is Lyapunov stable in probability. Compare this observation for the stochastic case with Theorem 13.12 for the non-stochastic case. More results have been established for recurrence than for asymptotic stability for the system (13.12).

We continue to use $\mathcal{O}$ to denote the recurrent set and will continue to assume that it is open and bounded.

A set $\mathcal{O} \subset \mathbb{R}^n$ is said to be *globally recurrent* (GR) for (13.12) if, for each solution $\mathbf{x}$, there are no finite escape times almost surely and, almost surely, the sample paths with unbounded time domains reach $\mathcal{O}$, i.e., for almost all $\omega \in \Omega$, either $\mathrm{dom}(\mathbf{x}(\omega))$ is bounded or there exists $(t,j) \in \mathrm{dom}(\mathbf{x}(\omega))$ such that, with $\phi_\omega := \mathbf{x}(\omega)$, $\phi_\omega(t,j) \in \mathcal{O}$.

A set $\mathcal{O} \subset \mathbb{R}^n$ is said to be *uniformly globally recurrent* (UGR) for (13.5) if there are no finite escape times and for each compact set $K \subset \mathbb{R}^n$ and $\rho > 0$ there exists $T > 0$ such that, for each solution $x \in \mathcal{S}(K)$,

$$\mathbb{P}\left(\Omega_1 \cup \Omega_2\right) \geq 1 - \rho \tag{13.25}$$

where

$$\Omega_1 := \left\{\omega \in \Omega : \mathrm{dom}(\mathbf{x}(\omega)) \subset \mathcal{T}_{<T}\right\} \tag{13.26a}$$

$$\Omega_2 := \left\{\omega \in \Omega : \mathrm{graph}(\mathbf{x}(\omega)) \cap \left(\mathcal{T}_{<T} \times \mathcal{O}\right) \neq \emptyset\right\}. \tag{13.26b}$$

A set $\mathcal{O} \subset \mathbb{R}^n$ is said to be *robustly uniformly globally recurrent* (R-UGR) for (13.5) if there exists a continuous function $\rho : \mathbb{R}^n \to \mathbb{R}_{>0}$ such that $\mathcal{O}$ is uniformly globally recurrent for (13.18) where (C_ρ, F_ρ) are defined in (13.2) and (D_ρ, G_ρ) are defined in (13.4).

The conditions of Assumptions 13.1, 13.3, and 13.4 confer the property that R-UGR and global recurrence are equivalent.

Theorem 13.19 *Under Assumptions 13.1, 13.3, and 13.4, the set $\mathcal{O} \subset \mathbb{R}^n$ is R-UGR for (13.12) if and only if it is globally recurrent for (13.12).*

Theorem 13.19 was established in [40].

13.5.6 Foster Functions for Recurrence

A function $V : \mathrm{dom}(V) \to \mathbb{R}_{\geq 0}$ is a *Lyapunov–Foster function candidate* for the set $\mathcal{O} \subset \mathbb{R}^n$ and the data (C, F, D, G, μ) of (13.12), i.e., for $(\mathcal{O}, (C, F, D, G, \mu))$, if $C \cup D \cup G(D, \mathcal{V}) \subset \mathrm{dom}(V)$, it is continuous on its domain, continuously differentiable on an open set containing C, and if the sequence of points $\left\{x_i\right\}_{i=1}^{\infty}$, with each point in $C \cup D$, is unbounded then the sequence of values $\left\{V(x_i)\right\}_{i=1}^{\infty}$ is unbounded. It is a *Lyapunov–Foster function* if $\langle \nabla V(x), f \rangle < 0$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$, and

$$\int_{\mathbb{R}^m} \max_{g \in G(x,v)} V(g) < V(x) \qquad \forall x \in D \backslash \mathcal{O}. \tag{13.27}$$

It is a *Krasovskii–LaSalle–Foster function* if $\langle \nabla V(x), f \rangle \leq 0$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$,

$$\int_{\mathbb{R}^m} \max_{g \in G(x,v)} V(g) \leq V(x) \qquad \forall x \in D \tag{13.28}$$

and there does not exist a solution $\mathbf{x}$ that is has an unbounded time domain almost surely and that almost surely never intersects $\mathcal{O}$ while keeping V equal to a nonzero constant. The latter property holds if V is a Lyapunov–Foster function. In the stochastic case, recurrence does not guarantee the type of uniformly decreasing Lyapunov–Foster function we encountered in the non-stochastic case. Instead, we consider the following definition: It is a *uniformly decreasing (on compact sets) Lyapunov–Foster function* if there exists a continuous function $\lambda : \mathbb{R}^n \to \mathbb{R}_{>0}$ such that $\langle \nabla V(x), f \rangle \leq -\lambda(x)$ for all $x \in C \backslash \mathcal{O}$ and $f \in F(x)$ and

$$\int_{\mathbb{R}^m} \max_{g \in G(x,v) \cap (\mathbb{R}^n \backslash \mathcal{O})} V(g) \mu(dv) - V(x) \leq -\lambda(x) \qquad \forall x \in D \backslash \mathcal{O}. \tag{13.29}$$

The following result is contained in [35].

Theorem 13.20 *Under Assumptions 13.1, 13.3, and 13.4, $(\mathcal{O}, (C, F, D, G, \mu))$ admits a uniformly decreasing (on compact sets) Lyapunov–Foster function if and only if $\mathcal{O}$ is recurrent for (13.5).*

The following theorem is a combination of results in [39] and Theorem 13.19, which came from [40].

Theorem 13.21 *Under Assumptions 13.1, 13.3, and 13.4, if $(\mathscr{O},(C,F,D,G,\mu))$ admits a*
Krasovskii–LaSalle–Foster then the set $\mathscr{O}$ is R-UGR for (13.5).

Example 13.2 above, with $\mathscr{O}$ a small open neighborhood of the origin, provides a counterexample to the conclusion of Theorem 13.20 when Assumption 13.1 is omitted.

13.6 Conclusions

In this paper, we have summarized the stability theory that is currently available for stochastic hybrid inclusions. Most of the pieces are in place, though some important open questions remain, especially for the case where there is randomness in the differential inclusion. This chapter is light on examples, though examples have been studied elsewhere, either in the literature on stochastic hybrid systems with unique solutions or in papers like [42]. For example, the latter studies the stochastic bouncing ball, which exhibits Zeno sample paths (typically each with different Zeno times). Both asymptotic stability and recurrence are considered for that example. We believe that many interesting additional applications are waiting to be made. Indeed, stochastic hybrid inclusions provides a very rich modeling framework for systems that combine continuous change, instantaneous change, worst-case disturbances, and random inputs.

Acknowledgements The research reported in this chapter has been supported by the US Air Force Office of Scientific Research through grants FA9550-12-1-0127 and FA9550-15-1-0155, and by the US National Science Foundation through grants ECCS-1232035 and ECCS-1508757.

References

1. Bacciotti, A., Ceragioli, F.: Stability and stabilization of discontinuous systems and nonsmooth Lyapunov functions. ESAIM-COCV **4**, 361–376 (1999)
2. Barbashin, E.A., Krasovskii, N.N.: On the existence of a function of Lyapunov in the case of asymptotic stability in the large. Prikl. Mat. Mekh. **18**, 345–350 (1954)
3. Brodtkorb, A.H., Teel, A.R., Sorensen, A.J.: Sensor-based hybrid observer for dynamic positioning. In: 54th IEEE Conference on Decision and Control (CDC), pp. 948–953, Dec 2015
4. Cai, C., Teel, A.R., Goebel, R.: Smooth Lyapunov functions for hybrid systems. Part II: (Pre-)asymptotically stable compact sets. IEEE Trans. Autom. Control **53**(3), 734–748, Apr 2008
5. Clarke, F.H., Ledyaev, Y.S., Stern, R.J.: Asymptotic stability and smooth Lyapunov functions. J. Differ. Equ. **149**, 69–114 (1998)
6. Filippov, A.F.: Differential Equations with Discontinuous Right-Hand Sides. Kluwer (1988)
7. Goebel, R., Hespanha, J., Teel, A.R., Cai, C., Sanfelice, R.: Hybrid systems: generalized solutions and robust stability. In: IFAC Symposium on Nonlinear Control Systems, pp. 1–12. Stuttgart, Germany (2004)

8. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid dynamical systems. IEEE Control Syst. Mag. **29**(2), 28–93 (2009)
9. Goebel, R., Sanfelice, R.G., Teel, A.R.: Hybrid Dynamical Systems. Princeton University Press (2012)
10. Goebel, R., Teel, A.R.: Solutions to hybrid inclusions via set and graphical convergence with stability theory applications. Automatica **42**, 573–587 (2006)
11. Grammatico, S., Subbaraman, A., Teel, A.R.: Discrete-time stochastic control systems: a continuous Lyapunov function implies robustness to strictly causal perturbations. Automatica **49**(10), 2939–2952 (2013)
12. Grimm, G., Messina, M.J., Tuna, S.E., Teel, A.R.: Examples when nonlinear model predictive control is nonrobust. Automatica **40**(10), 1729–1738 (2004)
13. Hàjek, O.: Discontinuous differential equations I. J. Differ. Equ. **32**, 149–170 (1979)
14. Hespanha, J.P.: A model for stochastic hybrid systems with application to communication networks. Nonlinear Anal. Spec. Issue Hybrid Syst. **62**, 1353–1383 (2005)
15. Hespanha, J.P., Morse, A.S.: Stabilization of nonholonomic integrators via logic-based switching. Automatica **35**(3), 385–393 (1999)
16. Kellett, C.M., Teel, A.R.: Smooth Lyapunov functions and robustness of stability for differential inclusions. Sys. Control Lett. **52**, 395–405 (2004)
17. Kurzweil, J.: On the inversion of Ljapunov's second theorem on stability of motion. Am. Math. Soc. Trans. Ser. **2**(24), 19–77. Originally appeared in. Czechoslovak Mathematical Journal **81**(1956), 217–259 (1963)
18. Ledyaev, Y.S., Sontag, E.D.: A Lyapunov characterization of robust stabilization. Nonlinear Anal. **37**, 813–840 (1999)
19. Lin, Y., Sontag, E.D., Wang, Y.: A smooth converse Lyapunov theorem for robust stability. SIAM J. Control Optim. **34**(1), 124–160 (1996)
20. Loria, A., Panteley, E., Popovic, D., Teel, A.R.: A nested Matrosov theorem and persistency of excitation for uniform convergence in stable nonautonomous systems. IEEE Trans. Autom. Control **50**(2), 183–198 (2005)
21. Malkin, I.G.: On the question of the reciprocal of Lyapunov's theorem on asymptotic stability. Prikl. Mat. Meh. **18**, 129–138 (1954)
22. Massera, J.L.: On Liapounoff's conditions of stability. Ann. Math. **50**, 705–721 (1949)
23. Massera, J.L.: Contributions to stability theory. Ann. Math. **64**, 182–206 (1956). (Erratum: Annals of Mathematics, vol. 68, (1958), 202.)
24. Mayhew, C.G., Sanfelice, R.G., Teel, A.R.: Quaternion-based hybrid control for robust global attitude tracking. IEEE Trans. Autom. Control **56**(11), 2555–2566 (2011)
25. Mayhew, C.G., Teel, A.R.: Global stabilization of spherical orientation by synergistic hybrid feedback with application to reduced-attitude tracking for rigid bodies. Automatica **49**(7), 1945–1957 (2013)
26. Mayhew, C.G., Teel, A.R.: Synergistic hybrid feedback for global rigid-body attitude tracking on *SO*(3). IEEE Trans. Autom. Control **58**(11), 2730–2742 (2013)
27. Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O.M.: Constrained model predictive control: stability and optimality. Automatica **36**, 789–814 (2000)
28. Poveda, J.I., Teel, A.R., Nesic, D.: Flexible Nash seeking using stochastic difference inclusions. In: American Control Conference, pp. 2236–2241, July 2015
29. Rawlings, J.B., Muske, K.R.: The stability of constrained receding horizon control. IEEE Trans. Autom. Control **38**, 1512–1516 (1993)
30. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer (1998)
31. Ryan, E.P.: An integral invariance principle for differential inclusions with applications in adaptive control. SIAM J. Control Optim. **36**(3), 960–980 (1998)
32. Sanfelice, R.G., Goebel, R., Teel, A.R.: Invariance principles for hybrid systems with connections to detectability and asymptotic stability. IEEE Trans. Autom. Control **52**(12), 2282–2297 (2007)
33. Sanfelice, R.G., Goebel, R., Teel, A.R.: Generalized solutions to hybrid dynamical systems. ESAIM: Control Optim. Calc. Var. **14**(4):699–724 (2008)

34. Sanfelice, R.G., Teel, A.R.: Asymptotic stability in hybrid systems via nested Matrosov functions. IEEE Trans. Autom. Control **54**(7), 1569–1574 (2009)
35. Subbaraman, A.: Robust stability theory for stochastic dynamical systems. Ph.D. dissertation, University of California, Santa Barbara (2015)
36. Subbaraman, A., Hartman, M., Teel, A.R.: A stochastic hybrid algorithm for robust global almost sure synchronization on the circle: All-to-all communication. In: 52nd IEEE Conference on Decision and Control, pp. 600–605, Dec 2013
37. Subbaraman, A., Teel, A.R.: A converse Lyapunov theorem for strong global recurrence. Automatica **49**(10), 2963–2974 (2013)
38. Subbaraman, A., Teel, A.R.: A Matrosov theorem for strong global recurrence. Automatica **49**(11), 3390–3395 (2013)
39. Subbaraman, A., Teel, A.R.: A Krasovskii-LaSalle function based recurrence principle for a class of stochastic hybrid systems. In: Proceedings of the 53rd IEEE Conference on Decision and Control, pp. 2310–2315 (2014)
40. Subbaraman, A., Teel, A.R.: Robustness of recurrence for a class of stochastic hybrid systems. In: Proceedings of the IFAC Conference on the Analysis and Design of Hybrid Systems, pp. 304–309 (2015)
41. Subbaraman, A., Teel, A.R.: On the equivalence between global recurrence and the existence of a smooth Lyapunov function for hybrid systems. Syst. Control Lett. **88**, 54–61 (2016)
42. Teel, A.R.: Lyapunov conditions certifying stability and recurrence for a class of stochastic hybrid systems. Ann. Rev. Control **37**, 1–24 (2013)
43. Teel, A.R.: A Matrosov theorem for adversarial Markov decision processes. IEEE Trans. Autom. Control **58**(8), 2142–2148 (2013)
44. Teel, A.R.: On a recurrence principle for a class of stochastic hybrid systems. In: Proceedings of the American Control Conference, pp. 4518–4523 (2014)
45. Teel, A.R.: On sequential compactness of solutions for a class of stochastic hybrid systems. In: Proceedings of the American Control Conference, pp. 4512–4517 (2014)
46. Teel, A.R.: Stochastic hybrid inclusions with diffusive flows. In: Proceedings of the 53rd IEEE Conference on Decision and Control, pp. 3071–3076 (2014)
47. Teel, A.R.: A recurrence principle for stochastic difference inclusions. IEEE Trans. Autom. Control **60**(2), 420–435 (2015)
48. Teel, A.R., Hespanha, J., Subbaraman, A.: Stochastic difference inclusions: results on recurrence and asymptotic stability in probability. In: Proceedings of the 51st IEEE Conference on Decision and Control, pp. 4051–4056 (2012)
49. Teel, A.R., Hespanha, J.P., Subbaraman, A.: A converse Lyapunov theorem and robustness for asymptotic stability in probability. IEEE Trans. Autom. Control **59**(9), 2426–2441 (2014)
50. Teel, A.R., Nesic, D., Lee, T.-C., Tan, Ying: A refinement of Matrosov's theorem for differential inclusions. Automatica **68**, 378–383 (2016)
51. Teel, A.R., Praly, L.: Global stabilizability and observability imply semi-global stabilizability by output feedback. Syst. Control Lett. **22**(5), 313–325 (1994)
52. Teel, A.R., Praly, L.: Tools for semiglobal stabilization by partial state and output feedback. SIAM J. Control Optim. **33**(5), 1443–1488 (1995)
53. Teel, A.R., Subbaraman, A., Sferlazza, A.: Stability analysis for stochastic hybrid systems: a survey. Automatica **50**(10), 2435–2456 (2014)
54. Teel, A.R., Praly, L.: On assigning the derivative of a disturbance attenuation control Lyapunov function. Math. Control Signals Syst. **13**, 95–124 (2000)
55. Teel, A.R., Praly, L.: A smooth Lyapunov function from a class-$\mathscr{K}\mathscr{L}$ estimate involving two positive semidefinite functions. ESAIM Control Optim. Calc. Var. **5**, 313–367 (2000)