Lecture Notes in Control and Information Sciences

393

Bijoy Ghosh
Clyde F. Martin
Yishao Zhou (Eds.)

Emergent Problems in Nonlinear Systems and Control

Springer

Lecture Notes
in Control and Information Sciences 393

Editors: M. Thoma, F. Allgöwer, M. Morari

Bijoy K. Ghosh, Clyde F. Martin, and
Yishao Zhou

Emergent Problems in Nonlinear Systems and Control

Springer

Authors

Bijoy K. Ghosh

Dept. of Mathematics and Statistics
Center for BioCybernetics and
Intelligent Systems
Texas Tech University
Broadway and Boston
Lubbock, TX 79409-1042
USA
E-mail: bijoy.ghosh@ttu.edu

Yishao Zhou

Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden
E-mail: yishao@math.su.se

Clyde F. Martin

Dept. of Mathematics and Statistics
Texas Tech University
Broadway and Boston
Lubbock, TX 79409-1042
USA
E-mail: clyde.f.martin@ttu.edu

ISBN 978-3-642-03626-2 e-ISBN 978-3-642-03627-9

DOI 10.1007/978-3-642-03627-9

Lecture Notes in Control and Information Sciences ISSN 0170-8643

Library of Congress Control Number: 2009933692

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

5 4 3 2 1 0

springer.com

Dedicated to our Friend and Collaborator
W. P. Dayawansa
on the Occasion of his Fiftieth Birthday

Preface

Papers in this collection partly represent the set of talks that were presented at Texas Tech University on the occasion of Daya's memorial workshop in the year 2007. Daya had a varied interest in the field of Dynamics and Control Theory and the papers bring out the essence of his involvement in these activities. He also had a large number of collaborators and this collection represent a good fraction of them. The papers included here cover his interest in control theory. Also included are papers from application areas that we believe are of strong interest to him.

As editors of this collection, we would like to thank all those individuals who participated in the 2007 workshop and subsequently took the time to contribute to the success of this publication. We would also like to thank others who were enthusiastic about this publication but were unable to submit a paper at this time. Our special thanks go to Mervyn Parakrama Ekanayake and Indika Wijayasinghe for their tireless efforts to put all the papers together.

We would hope that this publication mirrors to a large extent Daya's vigorous enthusiasm towards research, and his love for science and mathematics.

Bijoy K. Ghosh
Texas Tech University

Clyde F. Martin
Texas Tech University

Yishao Zhao
Stockholm University
2009

W.P. Dayawansa (1956 – 2006)

W. P. Dayawansa, "Daya", was born in Sri Lanka on March 27, 1956. After graduating from the University of Peradeniya in Sri Lanka, Daya came to the United States in 1979 to pursue graduate studies, first at Clarkson University, New York and then at Washington University in St Louis, Missouri, where he earned a Doctor of Science degree. Daya died on October 18, 2006 after a two-month battle with an untreatable cancer, soft tissue sarcoma. He was 50 years old at the time of his death.

During his life Daya won numerous awards for research and teaching, including at Texas Tech and the University of Maryland. In 2003 he was awarded the rank of a P. W. Horn Professorship at Texas Tech University, and in 2006 he was made Fellow of the IEEE. In the same year he won the Scientist of the Year award established by the local chapter of Achievement Rewards for College Scientists.

Daya received his Bachelor's degree in Sri Lanka and then came to Clarkson University and received a Master's degree in Electrical Engineering. He then moved to Washington University in St. Louis, where he earned a Ph. D. in the Department of System Science and Mathematics. In all three universities he was recognized as having extraordinary talent. I met Daya for the first time at a Decision and Control Conference, the fall before his graduation. He approached me about coming to Texas Tech so that he and I could work together. I will never forget the talk that he gave at that conference. A distinguished member of the audience brought up two obscure papers that he claimed had precedence over Daya's results. Daya thoughtfully explained what was in the two papers and how they related to his own work. I doubt that anyone else in the audience even knew that the papers existed. Even as a graduate student, Daya had an encyclopedic knowledge of the literature.

Daya arrived in Lubbock the next fall and immediately set to work. In the next year he wrote several papers, setting a high standard for the department. During the next three years he began work on solving a major problem involving the stabilization of low-dimensional systems. He established a new set of tools for approaching the problem, and the publication of his paper essentially solved an entire set of open problems. He went on to show that the techniques could be

used to solve many other problems. This paper established unequivocally that Daya was among the foremost control theorists and scientists in the world.

Shortly after the publication of his paper Daya moved to the Department of Electrical Engineering in the University of Maryland at College Park and began another important part of his career, becoming involved in serious interdisciplinary research. During this time he wrote, in conjunction with faculty members in mathematics and physics, the paper that is among the most quoted papers in control theory. He was tenured at the University of Maryland with support from all factions in the Electrical Engineering Department.

Daya returned to Texas Tech in the mid 1990s as a full professor. He continued to write world-class papers while becoming firmly involved with interdisciplinary research. He became interested in the ecology of the playa ecosystem of the area, which led to an interest in climatology. He was fascinated by the concept of learning, and in particular how babies learn to control posture. He worked with Jordan Berg in mechanical engineering (TTU) on the control of micro-mirrors, as well as with John Miller (USC) on problems of circadian rhythm in mammals. In addition, he worked with Sandy Dasgupta (Chemistry TTU) on applications of Kalman filtering to noise reduction, and with Bijoy K. Ghosh (Washington University and TTU) on eye motion.

Daya was only 49 when he was nominated for the Scientist of the Year, but his accomplishments reached far beyond his age. He had about 160 papers, was awarded $ 6 million in grants, and supervised many graduate students at the Master's and Ph. D. levels. He served on the editorial board of almost every major control journal. Numbers do not adequately describe his accomplishments. To understand his success you have to know how many papers he influenced but didn't coauthor, the graduate students with whom he worked but was not their advisor, and the undergraduates who learned to understand and love mathematics and its applications under his tutelage.

When Daya died the world lost a brilliant scientist. I lost a friend and a collaborator. He is missed by many. Daya is survived by his wife Samanmalee, two daughters, Samantha and Tammy, and two sons, Nuwan and Seth. He is also survived by his brother Jayantha and countless friends from around the world.

Clyde F. Matrin
Texas Tech University
2009

Contents

1

Type II Diabetes and Obesity: A Control Theoretic Model

Sam Al-Hashmi, Mervyn P.B. Ekanayake, and C.F. Martin

Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA

Summary. Diabetes mellitus can be further classified as Type I diabetes (TID) or Type II diabetes (TIID). TID is recognized as a complete failure in the pancreas β-cell islet. TIID is a failure in the body's ability to regulate glucose plasma concentration within the blood. Two aspects of TIID are well recognized: impaired insulin secretion and insulin resistance. Insulin resistance inhibits the organs' to glucose uptake and the response to insulin secretion. Insulin resistance can be linked to increased levels of Free Fatty Acid (FFA) concentration within exhibited in obese individuals and diabetes patients' offspring. Offspring of diabetes patients exhibit an extreme reduction in mitochondrial function; whereas obese individuals demonstrate high FFA concentration due to an excess of adipose tissue. FFA is linked to insulin resistance with TIID patients. Simulating glucose regulation in the body is considered to be an interesting control theory problem with significant impact. The suggested model in the paper is an attempt to address glucose regulation from a control theoretic viewpoint using the pharmacokinetic approach. The model identifies the contribution of glucose, insulin, incretins, glucagon, and FFA on glucose regulation within the body. The model consists of sub-models addressing the factors production rate (source) and clearance rate (sinks) based upon physiological responses. The model emphasizes the effect of FFA on glucose regulation, therefore creating a base for a mathematical model to simulate the behavior of early TIID diabetes glucose regulation under the effect of insulin resistance with a decreased rate in insulin secretion. The model would remain unable to simulate other behaviors of insulin impairment response exhibited within the TIID. Finally; the model creates a better understanding of TIID, further insight into prevention methods, new disease managements options in the form of diet, a better understanding of the impact of obesity on diabetes, and it can be used to investigate a possible link to cardiovascular diseases.

1.1 Introduction

The world faces a pandemic of Type II diabetes. Worldwide TIID is one of the major causes of morbidity and mortality and as a consequence it has serious economical side effects. In the United States, diabetes is the leading cause of blindness among working age adults, nontraumatic loss of limb, and is the fifth-leading cause of death. The direct medical care cost in the United States alone is $92 billion dollars. Once considered a disease of wealthy nations, Type II diabetes

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 1–19.
springerlink.com

now constitutes a truly global affliction. The International Diabetes Federation (IDF) anticipates that the worldwide incidence of diabetes among those aged 20 to 79 years will increase by around 70% in the next 20 years. Billions of people are suffering from Type II diabetes throughout the world. Hence it is a subject of interest to scientists and care providers.

Extensive research has been and is being conducted in the hope of understanding the physiological aspects of the disease [14]. Diabetes is classified as TID or TIID. TID diabetes occurs due to a complete or severe failure in the β-cells of the pancreas islet, which translates into an inability to produce insulin in response to metabolic demands and a complete failure in the glucose regulation mechanism. TIID diabetes is recognized as a failure in the regulation mechanism to maintain plasma glucose in an appropriate concentration level in response to endogenous glucose production or diet input. TIID diabetes patients typically exhibit symptoms of insulin resistance and impaired insulin secretion. Initially the body is able to adapt to maintain glucose concentration with high levels. However if Type II diabetes is not properly maintained it can evolve to Type I.

In the normal human the insulin-glucose reaction is carefully regulated. In the TID patient there is a complete failure of one component of the regulatory system and the system can be controlled by the adminstration of insulin. In the TIID patient the situation is much more complex. In this case, there seems to be multiple degradation of various components of the system and the control reduces to maintaining the input of glucose at as low a level as possible, albeit in conjunction with the administration of drugs that effect various parts of the regulatory system. The goal of the line of research described in this paper is to understand how the degradation of various components interact among each other. Seldom if ever does TIID reverse itself. How is this reflected in the mathematical model? An immediate goal of the research is to understand the genetic aspects of the disease. We suspect that there are several different avenues of genetic interaction and we want to understand this interaction.

1.2 Biological Background

Glucose is a vital substrate for our existence. Hence endogenous glucose production and uptake in the form of diet is heavily regulated throughout the body. Endogenous glucose production occurs in the liver and the kidneys, with 85% derived from hepatic production and 15% due to the kidneys, [8]. Gluconeogenesis represents the metabolic pathway for hepatic glucose production and glycogen degradation in the muscles. Also FFA can be metabolized by the adipose tissue into producing glucose if the brain demands outweigh the rest of the body. Regardless of endogenous glucose production or carbohydrate meal intake, the body's objective is to maintain an acceptable level of glucose plasma concentration. In the basal state glucose is maintained within 70-100 mg/dl, [12]. Although this level may fluctuate due to glucose input or metabolic demands, the body remains successful in maintaining it within 60-150 mg/dl [7]. Body metabolism of glucose and concentration are different during postabsorptive

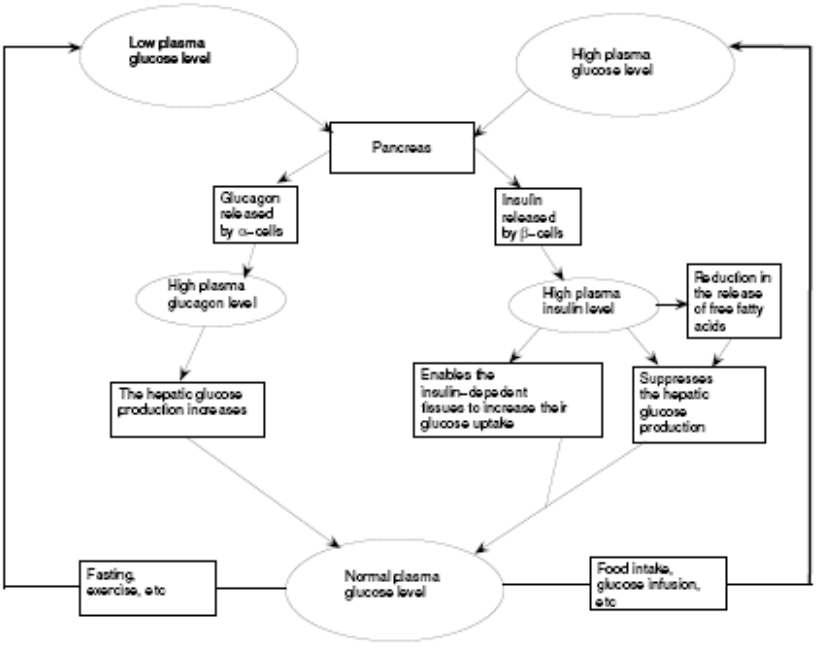


Fig. 1.1. A schematic representation of the glucose regulatory system. Based on [6].

state (10-12 hours fasting). During the postabsorptive state glucose concentration is saturated around 40 mg/dl, with the majority of the glucose consumption in insulin independent tissues. Approximately 50% of glucose concentration is utilized by the brain, 25% utilized in splanchnic area, and 25% in insulin dependent tissues primarily skeletal muscles. However during glucose intake, periphery area is responsible for (80%-85%) of glucose uptake with (4%-5%) metabolized by adipocytes [8].

The body strives to sustain an acceptable level of glucose concentration. This mechanism is successfully achieved due to several hormones with insulin and glucagon as the primary regulating hormones. A schematic detailing the regulating mechanism based upon [6] is given in Figure 1.1.

Glucagon is produced within the α-cells of the pancreas islets. It reacts as an inducer to insulin release in the β-cells of the pancreas islets and increases hepatic glucose production. Insulin is an inhibitor to glucagon release and glucose production. It also increases hepatic and periphery glucose uptake. Physical demands in the form of exercise and endurance performance has an aspect on the body fuel preference. An intake of carbohydrate meal causes the release of incretins which act as an inducer to insulin release in anticipation to the meal causing hyperinsulinemia. Hence glucagon secretion is inhibited and hepatic glucose production is suppressed, therefore maintaining normal postprandial glucose tolerance.

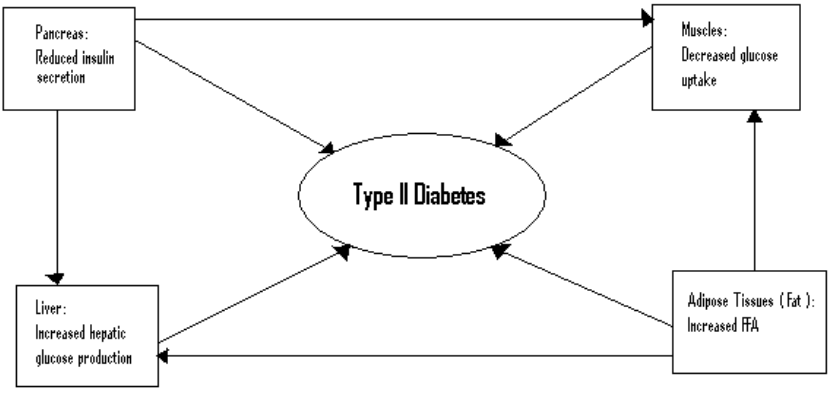


Fig. 1.2. The four main feedback system failures leading to Type II diabetes, Based on [8].

TIID is a genetic disease with failures in the regulating mechanism. To be able to target a specific control factor an overall understanding of the main categories of the physiological failures should help us identify the causes and effects.

Only rare forms of TIID exhibit monogenic defects, such as rare forms of TIID exhibited in Maturity Onset Diabetes of the Young (MODY). MODY types 1-5 are caused by mutations in genes encoding hepatic nuclear factor-4α (HNF-4α), glucokinase, HNF-1α, insulin promoter factor 1 and HNF-1β, respectively. Although the majority of Type II diabetes exhibits a polygenic defects, environment without a doubt plays as an aspect, such as increase in weight with those receptive to becoming diabetic, [3].

As in any control system, the body initially adapts to disturbances with glucose regulation such as: insulin resistance, impaired insulin secretion, and obesity. However, these disturbances have excessive demands on the body or limits it response, resulting in developing TIID.

One of the main conditions exhibited in TIID is insulin resistance. Although the causes may be diverse due to the genetics aspects, it is commonly exhibited throughout diverse ethnic backgrounds. It is also affected by the environment in the form of diet and exercise; hence it plays a key role in TIID. Indeed any successful model of TIID, should take into consideration the effect of insulin resistance. To be able to understand the failures in cell signaling mechanism and glucose uptake due to insulin resistance, a brief introduction into the cell signaling mechanism and glucose uptake in a healthy body is needed.

Glucose uptake has common aspects regardless of the tissue, such as the dependence on glucose transporter (GLUT); which is a member of the membrane family proteins for plasma glucose absorption through the myocyte. However, depending on the type of the GLUT, we can identify the tissue as an insulin dependent tissue or an insulin independent tissue. For an example; we identify

Table 1.1. Classification of glucose transport and HK activity according to their tissue distribution and functional regulation, from [9]

Organ	Glucose transporter	Hexokinase computer	Classification
Brain	GLUT1	HK-1	Glucose dependent
Erythrocyte	GLUT1	HK-1	Glucose dependent
Adipocyte	GLUT4	HK-II	Insulin dependent
Muscle	GLUT4	HK-II	Insulin dependent
Liver	GLUT2	HK-IVL	Glucose sensor
Glucokinase beta cell	GLUT2	HK-IVB (glucokinase)	Glucose sensor
Gut	GLUT3-symporter		Sodium dependent
Kidney	GLUT3-symporter		Sodium dependent

the brain as an insulin independent tissue due the GLUT1; which is associated with the brain.

Insulin dependent tissues are unable to absorb plasma glucose, without insulin activation to the GLUT associated with the tissue, such as: skeletal muscle, liver, and adipose tissue. Although GLUT2 in the liver exhibits a unique reaction by increasing in numbers in the cell membrane in anticipation due to increase plasma glucose concentration, it remains dependent on insulin activation as exhibited in GLUT4 in skeletal muscle. Cell signaling mechanism in activating GLUT4 in skeletal muscle is sufficient to clarify the signaling cycle.

In skeletal muscle, plasma glucose is brought through the myocyte by GLUT4 and then phosphorylated to glucose-6-phosphate (G6P) by hexokinase; hence trapping glucose within the cell. After isomrization, G6P is converted to G1P; which is activated to become uridine 5‘-diphospate (UDP-glucose). Finally, UDP-glucose becomes glycogen by polymerization through glycogen synthase, [14].

Insulin is a hormone that is unable to penetrate the cell membrane; hence insulin receptors are needed for insulin to interact with cells. Insulin receptors are a part of a large family of growth factor receptors withintrinsic tyrosine kinase activity. Insulin binds with insulin receptors, causing it to undergo autophosphorylation on multiple tyrosine residues. This initiates phosphorylation to insulin receptor substrate (IRS) proteins; IRS-1 in skeletal muscle and IRS-2 in the liver.

IRS binds with other intracellular proteins, transmitting the signal down stream. The modification of proteins in a specific amino acid causes the transduction of the signal in a specific pathway. Insulin receptors modify IRS and allow it to bind with src homology 2 (SH2) domain; which allows IRS to interact with phosphatidlinositol (PI) 3-kinase. IRS and PI3-kinase activate GLUT to absorb plasma glucose. However, neither of them is able to activate GLUT by itself.

Insulin resistance is exhibited in the body due to several causes, such as genetic defects or environment. Some of the genetic defects that cause insulin resistance are:

- Inaccurate phosphorylation in insulin receptors; which result in failure in phosphorylation of IRS, hence failure in activation of GLUT.

- Desensitization of insulin receptors; which minimize the number of available receptors, hence reducing the rate of GLUT reaction to the presence of insulin and plasma glucose.
- Inaccurate phosphorylation in IRS; which causes a failure in interacting with PI3-kinase.
- Lipodystropy; a rare genetic disease that causes malfunction with fat storage in the adipose tissue, hence affecting the insulin sensitivity in other tissues.

Another factor to insulin resistance is high level of FFA. This is exhibited within obese individuals and even the healthy offspring's of TIID patients. Obese individuals exhibit insulin resistance due to the excess amount of fat in the body; forcing the body to store fat within the liver and skeletal muscle. Offspring's of TIID patients' exhibits insulin resistance, due to a significant reduction in the mitochondrial function; which is necessary for lipid metabolism. It also emphasizes the genetic influence in TIID. There is a strong correlation in developing insulin resistance between obese individuals and the offspring's of TIID patients. This effect could be environmental due to eating habits of the family. Also, the main effects on skeletal muscles and the liver due to insulin resistance are one of the primary interests of the research. Hence we will address the effect of high level of FFA concentration on the skeletal muscles; which can easily be related to the liver.

Insulin resistance due to factors correlated with FFA concentrations is caused by the failure in insulin signaling mechanism. Initially, measuring precursors within the proposed pathological model identified a low intracellular concentration of G6P; which can be an indication to the failure of GLUT4 activation or hexokinase phosphrylation in the glycogen synthase. However, further research has determined that it was a failure in GLUT4 activation; which can be due to a failure in hexokinase or PI-3-Kinase. Finally, further testing verified that the primary factor in the failure of GLUT4 activation is PI-3-Kinase. Studies have shown that raising the plasma FFA levels abolishes insulin IRS1 response associated with PI3-kinase. Also rats studies have verified that lipid metabolism which activate protein kinase C (PKC)-θ blunts IRS-1 phosphorylation, [14].

Mechanism for fatty acid-induced insulin resistance in human skeletal muscle. An increase in delivery of fatty acids to muscle or a decrease in intracellular metabolism of fatty acids leads to an increase in intracellular fatty acid metabolites such as diacylglycerol, fatty acyl CoA and ceramides. These metabolites activate a serine/threonine kinase cascade (possibly initiated by proteinkinase C?) leading to phosphorylation of serine/threonine sites on insulin receptor substrates (IRS-1 and IRS-2), which in turn reduces the ability of the insulin receptor substrates to activate PI 3-kinase. As a consequence, glucose transport activity and other events downstream of insulin receptor signaling are diminished [14].

1.3 The Model

The model is based on models developed in [15] and [2], under the pharmacokinetic approach to govern glucose regulation within a healthy body. In contrast to Sorensen, [15], where his research revolved around TID, our main research interest revolves around TIID. Therefore later, Alvehag, [2], began making modifications by including the pancreas as a separate compartment and including incretin effect on insulin under meal consumption. We include the adipose tissue as a separate compartment and we introduced alternative sinks for Sorensen's nonlinear sinks within the glucose sub-model. This is an attempt to adapt the model to include insulin resistance effect on glucose regulation under the FFA effect, either due to obesity or genetic dysfunction. Further modifications are necessary to address impaired insulin secretion, such as the lack of first pulse insulin response under IVGTT, to fully adapt the model to address TIID.

Glucose and insulin sub models are the primary models following the pharmacokinetic approach, where the body organs are identified as compartments.

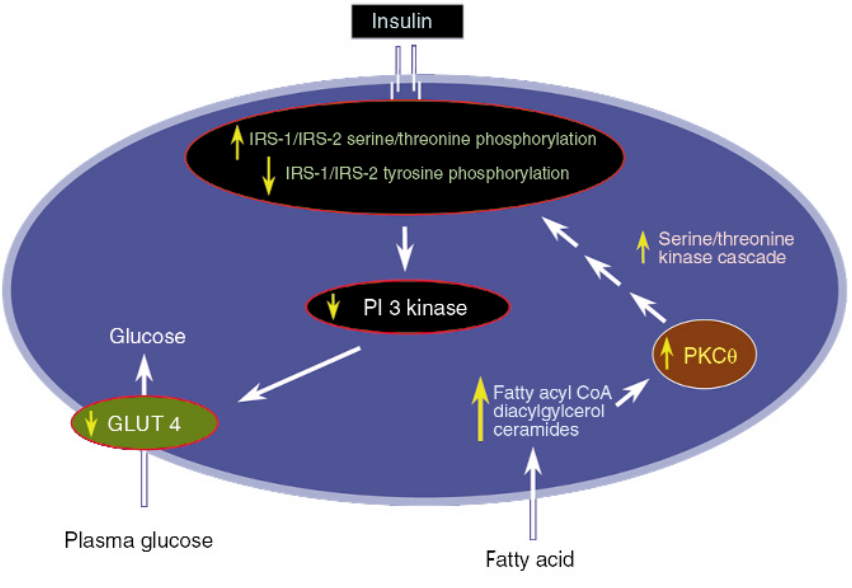


Fig. 1.3. Mechanism for fatty acid-included insulin resistance in human skeletal muscle. An increase in delivery of fatty acids to muscle or a decrease in intracellular metabolism of fatty acids leads to an increase in intracellular fatty acid metabolites such as diacylglycerol, fatty acyl CoA and ceramides. These metabolites activate a serine/threonine kinase cascade (possibly initiated by proteinkinase C) leading to phosphorylation of serine/threonine sites on insulin receptor substrates (IRS-1 and IRS-2), which in turn reduces the ability of the insulin receptor substrates to activate PI 3-kinase. As a consequence, glucose transport activity and other events downstream of insulin receptor signaling are diminished [14].

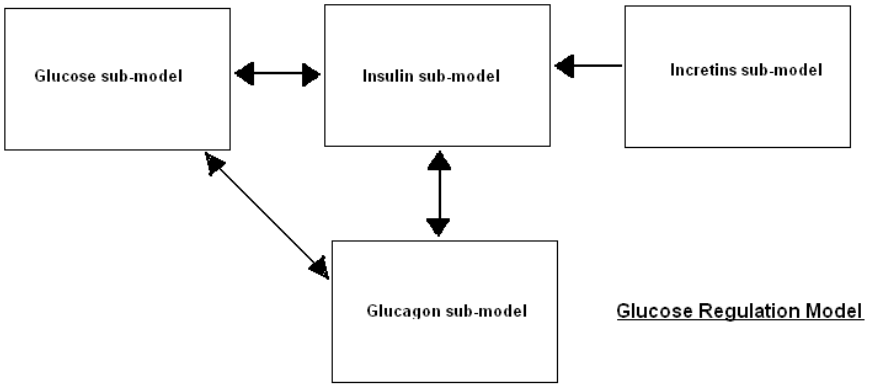


Fig. 1.4. Glucose regulation model

The sinks and sources within the glucose and insulin sub-models are based upon the body physiological response to glucose regulation. Incretins and glucagon sub-models are based upon the minimal body approach, where the body introduced as a single compartment. The glucose and insulin sub-models have been revised to include the adipose tissue compartment. Initially, skeletal muscles and adipose tissue were included in one compartment as the periphery tissue. Therefore parameters have been revised to adapt to the inclusion of the adipose tissue compartment [1].

Mass balance equations are derived, creating a system of differential equations governing the model.

Capillary blood space:

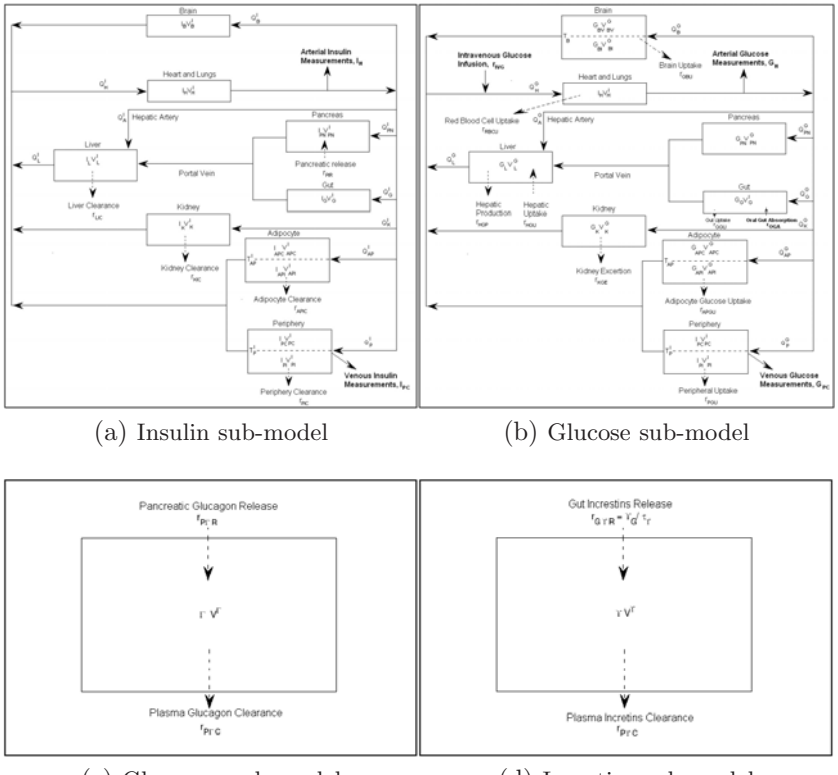
$$V_{CC}^{K}\frac{dk_{CC}}{dt} = Q_{C}^{K}(k_H - k_{CC}) + PA(k_{CI} - k_{CC}) + (r_{source} - r_{sink}) \quad (1.1)$$
$$accumulation \;=\; convection \;+\; diffusion + (metabolic\ source\ or\ sink)$$

Interstitial blood space:

$$V_{CI}^{K}\frac{dk_{CL}}{dt} = PA(k_{CC} - k_{CI}) + (r_{source} - r_{sink}) \quad (1.2)$$
$$accumulation \;=\; convection \;+\; (metabolic\ source\ or\ sink,)$$

where:

K = {the considered substance} = glucose, glucagon, insulin, or FFA.
C = considered compartment.
k = concentration of the considered substance [mass/volume].
V_C^K = volume of compartment C of substance K [volume].
Q_C^K = blood flow to/from compartment C for substance K [volume/time]



(a) Insulin sub-model (b) Glucose sub-model

(c) Glucagon sub-model (d) Incretins sub-model

Fig. 1.5. Sub-models

r_{source} = a metabolic source in the compartment for substance C [mass/time].
r_{sink} = a metabolic sink in the compartment for substance C [mass/time].
PA = permeability-are product between the capillary blood space and the interstitial blood space through the capillary blood walls.

Sinks and sources were derived in a multiplicative manner to capture the nonlinear behavior of the sink and sources based upon the experimental data. For example:

$$r = M^{\Gamma} M^{I} M^{G} M^{F} r_{basal}, \tag{1.3}$$

where:

r = metabolic rate of mass addition "source" or removal "sink" (mass/time).
M^{Γ} = multiplicative effect of glucagon (dimensionless).
M^{I} = multiplicative effect of insulin (dimensionless).

M^G = multiplicative effect of glucose (dimensionless).
M^F = multiplicative effect of FFA (dimensionless).
r_{basal} = basal metabolic rate (mass/time).

The adaptation of the multiplicative approach is commonly utilized in biological modelling [5]. Sorensen adapted the hyperbolic tangent function in defining the multipliers due to its suitability in representing the sigmoid nonlinearity observed in biological data.

$$M^K = A + B \tanh C(k^N - D), \tag{1.4}$$

where:

K = {the considered substance} = glucose, insulin, glucagon, or FFA.
k^N = normalized concentration of the considered substance [dimensionless].

The coefficients A, B, C, D are adjusted to fit experimental data based upon numerical methods. The alternative sinks, were based upon the physiological aspect of the system. The sinks are introduced within the system of equations governing the glucose sub-model.

1.4 Nomenclature

Variables:

G = glucose concentration [mg/dl].
I = insulin concentration [mU/l].
Γ = glucagon concentration [pg/ml].
γ = incretins concentration above normal levels [pmol/l].
γ_G = quantity of incretins in the gut above normal [pmol].
t = time [min].
r = metabolic sink or source [mg/min, mU/min]
"glucose and insulin" [pg/min, pmol/min] "glucagon and incretins"
M = multiplier of basal metabolic rate [dimensionless].
[IR] = multiplier approximating active insulin receptors [dimensionless].
[R] = multiplier approximating inactive insulin receptors [dimensionless].
[G] = multiplier approximating glucose concentration effect [dimensionless].

Parameters:

V = volume [dl, l] "glucose and insulin', volume [ml, l] "glucagon and incretins".
Q = blood flows [dl/min, l/min].
T = transcapillary diffusion time constant [min].
τ = time constant [min].
F = fractional clearance [dimensionless].
ζ = rate constant for incretins production [pmol/mg].

Subscript:

A = hepatic artery.
B = brain.
G = gut.
H = heart and lungs.
L = liver.
AP = adipose tissue.
P = periphery.
PN = pancreas.

The second subscripts for the sub-compartments in brain and periphery are:

C = capillary blood space.
I = interstitial fluid space.

Subscript for metabolic sinks and sources:

BGU = brain glucose uptake.
GGU = gut glucose uptake.
$APGU$ = adipose glucose uptake.
HGP = hepatic glucose production.
HGU = hepatic glucose uptake.
PGU = peripheral glucose uptake.
$RBCU$ = red blood cell glucose uptake.
KIC = kidney insulin clearance.
LIC = Liver insulin clearance.
PIC = peripheral insulin clearance.
PIC = pancreatic insulin release.
$APIC$ = adipocyte insulin clearance.
$P\Gamma C$ = plasma glucagon clearance.
$M\gamma C$ = metabolic glucagon clearance.
$P\Gamma R$ = pancreatic glucagon release.
$P\gamma C$ = plasma incretins clearance.
$M\gamma C$ = metabolic incretins clearance.
$G\gamma R$ = gut incretins release.

Superscript

G = glucose sub-model.
I = insulin sub-model.
Γ = glucagon sub-model.
γ = Incretins sub-model.
B = basal value.
N = normalized value (divided by basal value)
$r^N = \dfrac{r}{r_{basal}} = M^G M^I M^\Gamma M^{FA}$.
∞ = asymptotic or final steady state value.

1.5 Insulin Sub-model System of Equations

$$Brain: \; V_B^I \frac{dI_B}{dt} = Q_B^I (I_H - I_B) \tag{1.5}$$

$$Heart\ and\ lungs: \; V_H^I \frac{dI_H}{dt} = Q_B^I I_B + Q_L^I I_L + Q_K^I I_K + Q_P^I I_{PC} + Q_{AP}^I I_{APC} - Q_H^I I_H \tag{1.6}$$

$$Gut: \; V_G^I \frac{dI_G}{dt} = Q_G^I (I_H - I_G) \tag{1.7}$$

$$Liver: \; V_L^I \frac{dI_L}{dt} = Q_A^I I_H + Q_G^I I_G + Q_{PN}^I I_{PN} - Q_L^I I_L - r_{LIC} \tag{1.8}$$

$$Kidney: \; V_K^I \frac{dI_K}{dt} = Q_K^I (I_H - I_K) - r_{KIC} \tag{1.9}$$

$$Periphery: \; V_{PC}^I \frac{dI_{PC}}{dt} = Q_P^I (I_H - I_{PC}) - \frac{V_{PI}^I}{T_P^I} (I_{PC} - I_{PI}) \tag{1.10}$$

$$V_{PI}^I \frac{dI_{PI}}{dt} = \frac{V_{PI}^I}{T_P^I} (I_{PC} - I_{PI}) - r_{PIC} \tag{1.11}$$

$$Adipocyte: \; V_{APC}^I \frac{dI_{APC}}{dt} = Q_{AP}^I (I_H - I_{APC}) - \frac{V_{APC}^I}{T_{AP}^I} (I_{APC} - I_{API}) \tag{1.12}$$

$$V_{API}^I \frac{dI_{API}}{dt} = \frac{V_{PI}^I}{T_{AP}^I} (I_{APC} - I_{API}) - r_{APIC} \tag{1.13}$$

$$Pancreas: \; V_{PN}^I \frac{dI_{PN}}{dt} = Q_{PN}^I (I_H - I_{PN}) + r_{PIR}. \tag{1.14}$$

Insulin release is initially proposed by Gold, et al, [10] and modified for human by Sorensen, [15]. Sorensen introduced the variables $X(G)$ which affects the early

insulin response, and $Y(G)$ which affects the late insulin response. However he neglected the effect of incretins on insulin secretion. Karin Alvehag, [2], addressed this issue enabling the system to simulate glucose oral consumption. Finally, this gives us the following equations:

$$Insulin\ secretion: \ S = (M_1 Y + M_2(X-I)^+ + \varphi_2\gamma)\Lambda \quad (1.15)$$

$$Pancreatic\ insulin\ release: \ rPIR = \frac{S(G_H,\gamma)}{S(G_H^B,\gamma^B)} r_{PIR}^B. \quad (1.16)$$

The notation $(X-I)^+$ means the value of $(X-I)$ is assumed if positive and zero otherwise. Finally, the insulin sub-model assumes that the liver, kidneys, skeletal muscles, and adipose tissue respectfully remove 40, 30, 14.8, and 0.2 percent of the insulin.

$$r_{LIC} = F_{LIC}(Q_A^I I_H + Q_G^I I_G + Q_{PN}^I I_{PN}) \quad (1.17)$$

$$r_{KIC} = F_{KIC} Q_K^I I_H \quad (1.18)$$

$$r_{PIC} = F_{PIC} Q_P^I I_H \quad (1.19)$$

$$r_{APIC} = F_{APIC} Q_{AP}^I I_H. \quad (1.20)$$

1.6 Glucose Sub-model System of Equations

$$Brain: \ V_{BC}^G \frac{dG_B}{dt} = Q_B^G(G_H - G_{BC}) - \frac{V_{BI}^G}{T_B^G}(G_{BC} - G_{BI}) \quad (1.21)$$

$$V_{BI}^G \frac{dG_{BI}}{dt} = \frac{V_{BI}^G}{T_B^G}(G_{BC} - G_{BI}) - r_{BGU} \quad (1.22)$$

$$Heart\ and\ lungs: \ V_H^G \frac{dG_H}{dt} = Q_B^G G_{BC} + Q_L^G G_L + Q_K^G G_K + Q_P^G G_{PC} + Q_{AP}^G G_{APC} - Q_H^G G_H - r_{RBCU} \quad (1.23)$$

$$Gut: \ V_G^G \frac{dG_G}{dt} = Q_G^G(G_H - G_G) - r_{GGU} \quad (1.24)$$

$$Liver: \ V_L^G \frac{dG_L}{dt} = Q_A^G G_H + Q_G^G G_G + Q_{PN}^G G_{PN} - Q_L^G G_L - r_{HGU} + r_{HGP} \quad (1.25)$$

$$Kidney: \ V_K^G \frac{dG_K}{dt} = Q_K^G(G_H - G_K) - r_{KGE} \quad (1.26)$$

$$Periphery: \ V_{PC}^G \frac{dG_{PC}}{dt} = Q_P^G(G_H - G_{PC}) - \frac{V_{PI}^G}{T_P^G}(G_{PC} - G_{PI}) \quad (1.27)$$

$$V_{PI}^G \frac{dG_{PI}}{dt} = \frac{V_{PI}^G}{T_P^G}(G_{PC} - G_{PI}) - r_{PGU} \quad (1.28)$$

$$Adipocyte: \ V_{APC}^G \frac{dG_{APC}}{dt} = Q_{AP}^G(G_H - G_{APC}) - \frac{V_{API}^G}{T_{AP}^G}(G_{APC} - G_{API}) \quad (1.29)$$

$$V_{API}^{G}\frac{dG_{API}}{dt} = \frac{V_{PI}^{G}}{T_{AP}^{G}}(G_{APC} - G_{API}) - r_{APGU} \tag{1.30}$$

$$Pancreas: \; V_{PN}^{G}\frac{dG_{PN}}{dt} = Q_{PN}^{G}(G_H - G_{PN}). \tag{1.31}$$

Oral glucose input is taken as a source in the gut. Intravenous glucose injection is taken as a source in the heart and lungs compartment. The alternative sinks are derived from the cell mechanism. The sinks replace the nonlinear sinks within the glucose sub-model for the hepatic glucose uptake and the periphery glucose uptake, which are the skeletal muscles in this model. The FA constant is introduced to take into consideration the FFA effect on glucose uptake. Initially the FFA sub-model was derived consistent with physiological sinks and sources based on [11]. Due to the lack of biological data of for FFA concentrations correlated with OGTT, MGTT, and IVGTT we were unable to simulate the model. Another factor to take into consideration is the rate of FFA permeability within the body. FA exists in several forms and it is an issue to be addressed [1]. An alternative to this method is to derive a multiplier for the FA constant based upon the numerical data for a set of subjects. The weights of subjects and genetic background should be taken into consideration.

$$r_{HGU} = \exp\left(\frac{[G]([IR]-[R])}{175[FA]}\right) r_{HGU}^{B} \tag{1.32}$$

$$r_{PGU} = \exp\left(\frac{[G]([IR]-[R])}{12[FA]}\right) r_{PGU}^{B}. \tag{1.33}$$

The kidney excretion is modelled in a unique way based upon the glucose plasma concentration.

$$r_{KGE} = \begin{cases} 71 + 71\tanh(0.01)(G_K - 460) & 0 < G_K < 460mg/dl \\ -330 + 0.872G_K & G_K > 460mg/dl. \end{cases} \tag{1.34}$$

The equations governing this are based on [4]. Once the meal or glucose solution is orally taken the glucose concentration is filtered through the intestine into the blood circulation. The following differential equations govern the process.

$$Gastric\ emptying: \; \frac{dG_s}{dt} = OGC_S - \frac{1}{\tau_{GE}}G_S \tag{1.35}$$

$$Absorption\ into\ blood: \; \frac{dr_{OGA}}{dt} = \frac{1}{T_A\tau_{GE}}G_S - \frac{1}{T_A}r_{OGA}, \tag{1.36}$$

where

OCG_S = rate of hydrolyzed carbohydrates or glucose entering the stomach [mg/min]
G_S = total amount of glucose in the stomach [mg]
T_A = time constant [min]
τ_{GE} = time constant [min].

The rate of glucose absorbed through the intestine due glucose solution consumption is faster as expected than the hydrolyzed carbohydrates consumption, due to a faster gastric emptying rate. This ratio difference is addressed by adjusting the time constant. The rate of glucose entering the stomach, OCG_S), is dependent upon the oral intake of glucose. The equation governing this relationship is adapted from [13]:

$$\begin{aligned} OGC_O = & \frac{1}{\tau_S} OGC_O(t-t_0)u(t-t_0) - \frac{1}{\tau_S} OGC_O(t-t_0-1)u(t-t_0-1) - \\ & \frac{1}{\tau_S} OGC_O(t-t_0-4)u(t-t_0-4) + \\ & \frac{1}{\tau_S} OGC_O(t-t_0-5)u(t-t_0-5), \end{aligned} \tag{1.37}$$

where

$$\begin{aligned} OGC_O &= \text{the amount of oral glucose or carbohydrate intake [mg]} \\ t_0 &= \text{time of oral glucose or carbohydrate intake [min]} \\ u(t-t_k) &= \text{step function representing the entry of} \\ & \text{glucose/ carbohydrates into the gut at the time } t_k \text{ [dimensionless]} \end{aligned}$$

1.7 Glucagon Sub-model System of Equations

$$\textit{The whole body}: \; V^{\Gamma} \frac{d\Gamma}{dt} = r_{P\Gamma R} - r_{P\Gamma C} \tag{1.38}$$

The glucose source exhibits a nonlinear behavior:

$$\textit{Glucagon source}: \; r_{P\Gamma C} = M^G_{P\Gamma R} M^I_{P\Gamma R} r^B_{P\Gamma R}, \tag{1.39}$$

where:

$$\textit{Glucose multiplier}: \; M^G_{P\Gamma R} = 2.93 - 2.10 \tanh[4.18(G^N_H - .061)] \tag{1.40}$$

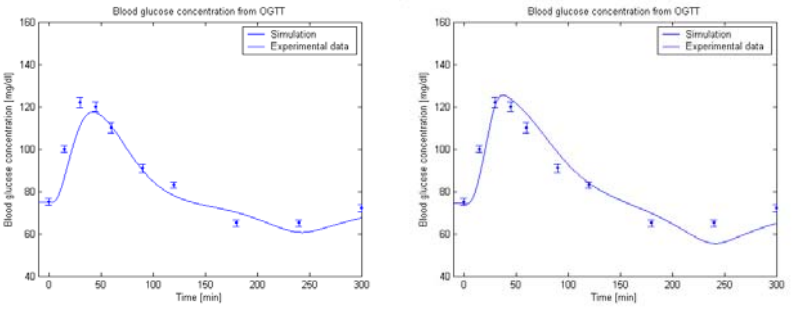
$$\textit{Insulin multiplier}: \; M^I_{P\Gamma R} = 1.31 - 0.61 \tanh[1.06(I^N_H - 0.47)] \tag{1.41}$$

On the other hand; the glucagon sink remains to exhibit a linear behavior by approximating it as a constant:

$\textit{Glucagon sink}: \; r_{P\Gamma C} = r_{M\Gamma C}\Gamma$

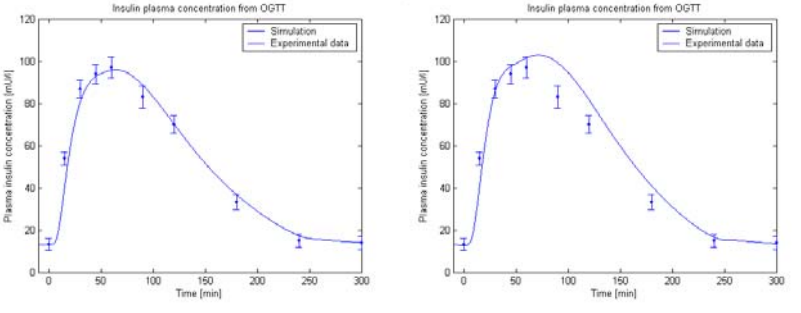
1.8 Incretins Sub-model System of Equations

$$V^{\gamma} \frac{d\gamma}{dt} = r_{G\gamma R} - r_{P\gamma C} \tag{1.42}$$



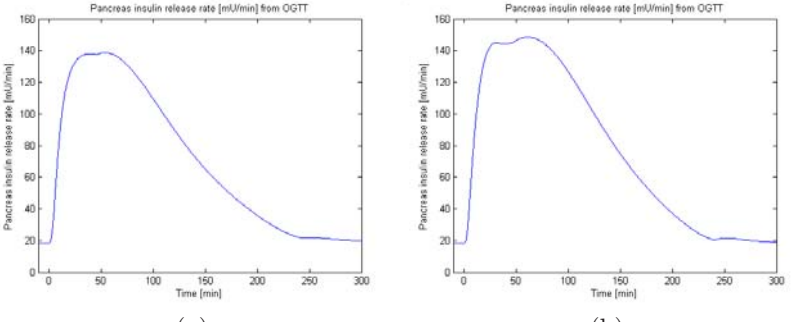
(a) (b)

Fig. 1.6. OGTT glucose blood concentration simulation



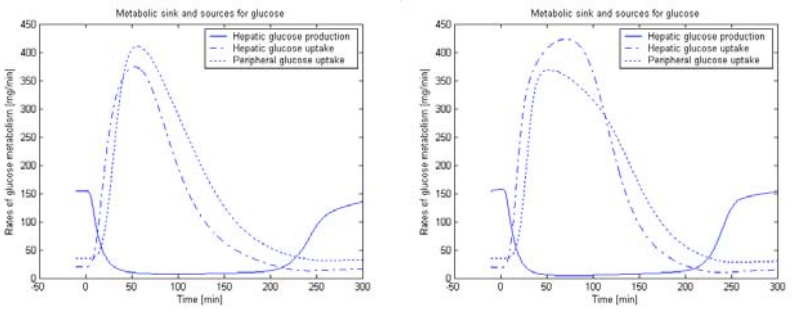
(a) (b)

Fig. 1.7. OGTT pancreas insulin release simulation



(a) (b)

Fig. 1.8. OGTT insulin plasma concentration simulation



(a) (b)

Fig. 1.9. OGTT hepatic glucose production, glucose uptake, and peripheral glucose uptake simulation

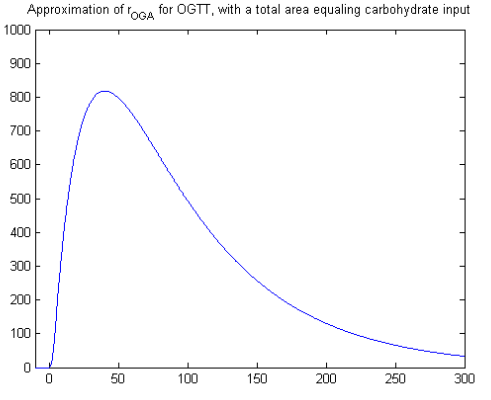


Fig. 1.10. Approximation of glucose oral input (rOGA) due to OGTT simulation

The incretins production rate in the gut is governed by the following differential equation.

$$\frac{d\gamma_G}{dt} = \zeta OGC_S - r_{G\gamma R}, \tag{1.43}$$

where:

OGC_S = rate of hydrolyzed carbohydrates or glucose entering the stomach [mg/min].

The rate of incretins production is depended upon the rate of glucose and carbohydrates entering the stomach;.

The source of incretins is given by the following equation:

$$r_{G\gamma R} = \frac{\gamma_G}{\tau_\gamma}, \tag{1.44}$$

where:

τ_γ = time delay constant describing the process of incretins absorption from the gut into the blood circulation.

ζ = a constant transforming into pmol concentration of incretins production.

The incretins sink is depended upon the incretins concentration.

$$r_{P\gamma C} = r_{M\gamma C}\gamma \tag{1.45}$$

1.9 Simulation

During all simulations we shall present the model simulation with the original sinks to the left, and then the model simulation with the alternative sinks to the right. The simulation parameters are available in [1] appendices. The OGGTT is based upon an oral consumption of a 100g of glucose and the data provided within Alvehag's code, [2].

References

1. Al-Hashmi, S.: Model and Analysis: Diabetes and Obesity, Master's Thesis, Texas Tech University (2008)
2. Alvehag, K.: Glucose regulation: A mathematical model, M. S. Thesis, The Royal Institute of Technology, Stockholm, Sweden (2006)
3. Beale, E.G., Hammer, R.E., Bénédicte, A., Claude, F.: Disregulated glyceroneogenesis: PCK1 as a candidate diabetes and obesity gene. Trends in endocrinology and metabolism 15(3), 129–135 (2004)
4. Bierman, E., Mehnert, H.: Diablog: A simulation program of insulin-glucose dynamic for education of patients. Comp. Meth. and Prog. in Biomed. 32(3-4), 311–318 (1990)
5. Carson, E.R., Cramp, D.G.: A systems model of blood glucose control. Int. J. Biomedical computing, 21–34 (1976)
6. Champe, P.C., Harvey, R.A.: Biochemistry, 2nd edn. Lippincott, Philadelphia (1994)
7. Cryer, P.E.: Hypoglycemia. In: Kasper, D.L., Fauci, A.A., Longo, D.L., Braunwald, E., Hauser, S.L., Jameson, J.L. (eds.) Harrison's principles of internal medicine, 16th edn., pp. 2180–2185. McGraw-Hill, New York (2005)
8. DeFronzo, R.A.: Pathogenesis of Type 2 diabetes mellitus. Med. Clinc, N Am. 88, 787–835 (2004)
9. DeFronzo, R.A.: Pathogenesis of Type 2 diabetes mellitus: Metablic and molecular implications for identifying diabetes genes. Diabetes 5, 177–269 (1997)
10. Gold, G., Landahl, H.D., Gishizky, M.L., Grodsky, G.M.: Heterogeneity and compartmental properties of insulin storage and secretion in rat islets. J. Clin. Invest. 69(3), 554–563 (1982)
11. Kim, J., Saidel, G.M., Carbera, M.E.: Multi-scale computational model of fuel homeostasis during exercise: Effect of hormonal control. Annals of biomedical engineering 35(1), 69–90 (2007)

12. Parker, R.S., Doyle III, F.J., Peppas, N.A.: A model-based algorithm for blood glucose control in Type I diabetic patients. IEEE Trans. Biomed. 46(2), 148–157 (1999)
13. Puckett, W.R.: Dynamic Modelling of Diabetes Mellitus, Ph. D. Dissertation, Chemical Eng. Department, University of Wisconsion, Madison (1992)
14. Shulman, G.I.: Cellular mechanisms of insulin resistance. J. Clinc. Invest. 106(2), 171–176 (2000)
15. Sorensen, J.T.: A physiological model of glucose metabolism in man and its use to design and assess improved insulin therapies for diabetes, Ph. D. Thesis, Massachusetts Institute of Technology (1985)

2

Dynamic Network Modeling of Diurnal Genes in Cyanobacteria

Thanura Elvitigala[1], Himadri B. Pakrasi[2], and Bijoy K. Ghosh[3]

[1] Electrical and Systems Engineering, Washington University, St. Louis, MO 63130, USA
[2] Department of Biology, Washington University, St. Louis, MO 63130, USA
[3] Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA

Summary. Many important processes in a living cell are controlled by DNA, RNA, proteins and dynamic interactions between them. Responding to various external conditions, such as the amount of incident light, presence or absence of nutrients and regulators (a protein complex), the genetic information stored in the DNA is decoded. This process is known as 'transcription', and it produces RNA molecules. The DNA subsequences that have the capability to generate different RNA molecules are called 'genes'. The resulting RNA molecules in turn induce the production of corresponding proteins via a process called 'translation'. Proteins perform many of the biological functions in the cell, which also include the function of 'cell regulation'. As a result, the entire process of transcription and translation can be viewed as 'gene interactions', where the product of one gene controls the activity levels of some others. Identifying the dynamic interaction between various genes would immensely help us understand the associated processes that these genes control. The problem, however, is that these interactions take place at various time scales; the chemical reactions have an inherent randomness and the dynamic equations are often nonlinear. In this paper, we would argue that simple dynamic models can capture dynamic interactions between genes locally in time resulting in a dynamic gene regulatory network. We present a feed forward dynamical systems model, to identify interactions between diurnal genes in *Cyanothece* sp. ATCC 51142, a unicellular cyanobacterium, under regular day/night cycles and altered light patterns. With the selection of appropriate parameters in the model, we can explain various gene expressions observed in the data. We construct a gene regulatory network and show how the network interconnections change under different light conditions. The model is shown to be sufficient to capture many biologically meaningful interactions between genes including, co-regulation of genes that are located in close proximity in the genome, time delays involved between regulator-target activities, increased level of interactions under transient input conditions, etc. The resultant network consists of various known network motifs. We validate some of the predicted links by showing the presence of common sequences in their corresponding conserved regulatory regions.

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 21–41.
springerlink.com

2.1 Introduction

A living cell is a very complex dynamical system, showing a remarkable ability of adapting to different environmental conditions for their survival. The introduction of various advanced technologies and involvement of experts from diverse disciplines have contributed to a significant progress in our understanding about these complex systems. However we are still far away from having a complete understanding on exact mechanisms involved in different biological processes of a living cell.

Traditionally the behavior of cells is explained using the 'Central Dogma of Molecular Biology' which consists of three main components. The DNA encodes genetic information required for the controlling of the biological processes and it is transferred from one generation to the next, through replication. The information in the DNA is decoded by a process known as transcription, which generates RNA molecules. The DNA sequences that correspond to different RNA molecules are identified as genes. Some of these RNA molecules are later translated into Proteins. The proteins perform most of the biological functions in the cells, which include the regulation of the transcription of DNA. As a result entire process can be viewed as interactions between different genes, some acting as regulator to control the activities of their targets. In Figure 2.1 we show the main elements in the central dogma of molecular biology.

Understanding the dependence between different genes in transcription/translation process is very important to study the behavior of cells. However, identifying the transcription regulatory links between genes has always been a challenging task. This is primarily due to the limited availability of gene

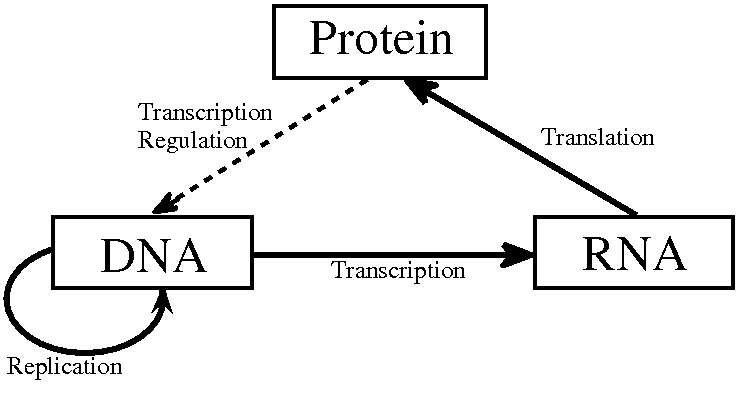


Fig. 2.1. 'Central Dogma in Molecular Biology' explains the information flow between different molecules in a living cell. Genetic information contained in DNA is transferred from one generation to the next through replication. Depending on requirements of the cells, different RNA molecules are produced in transcription process and later translated into corresponding proteins. Some of the proteins acts as regulators to control the transcription process.

expression data, compared to the large number of variables (genes) involved in the system. Despite these limitations, numerous methods have been developed to identify possible relationships between genes. Approaches such as clustering and correlation based networks [7, 15], try to identify gene groups having similar expressions across several measurements and infer the biological significance of them. Several methods based on entropy and mutual information is also available. One of the limitations of these methods is their inability to detect causal relationships between genes; namely separating regulators from the targets. In [9], the authors overcome this hurdle by limiting the regulators to the already known transcription factors. In [11], conditional mutual information was used to establish causal relationships. Various methods based on Boolean networks, Probabilistic Boolean networks, and Bayesian networks have been applied successfully, to model relatively small number of genes.

Whether 'gene interaction' should be modeled as a deterministic or a stochastic process has been debated for a long time. Arguments in favor of the stochastic modeling, is based on the randomness observed during the molecular interactions. However stochastic modeling requires considerably large number of data points and is therefore difficult to use. Though there is inherent randomness in interactions at a molecular level, in order to understand overall response of genes, it is usually sufficient to study the average behavior of gene products. Furthermore some of the environmental changes, such as day/night cycle, take place at a much slower time scale compared to molecular interactions. As discussed by [17], gene behaviors under such input conditions can be considered as purely deterministic.

Many deterministic systems can successfully be modeled using differential equations. Many such models have been proposed based on the interaction patterns observed in the actual system. In [19], Feed Forward Loop (FFL) has been identified as a dominant motif in gene interaction networks. Coherent FFL based models are used in [4] to study the dynamic interactions between genes and three FFLs are successfully identified in yeast.

In this paper, we first introduce different gene classification methods to identify oscillatory and transient behaviors in the gene expressions under normal day/night cycles and altered light patterns. Existence of different oscillatory patterns is discussed in details. Genes are categorized into two main groups. The 'Circadian Controlled Genes', which do not show a significant change from their regular oscillatory behavior under transient light inputs where as the 'Light Responding Genes' show a significant shift. We also find that in addition to the natural 24h oscillations, a number of genes exist with a 12h oscillation. Genes are put into sub groups based on their main oscillatory frequencies.

We propose the use of dynamical systems based on FFLs, to model interactions between different diurnal genes; both 'Circadian Controlled' and 'Light Responding'. We describe the specific simplification made to the general model, to suit the data set being analyzed. We discuss how to select appropriate parameters and function formats based on the type of gene being modeled. We show that the model is sufficient to explain the interaction between diurnal genes

under regular light/dark cycles and transient light conditions. We discuss how we can obtain a global gene interaction network based on the proposed model and how it can be improved by utilizing the already existing biological insight. Various features in the resultant network are discussed in details. We study the changes in the network, under different light conditions and gene groups. The resultant network is shown to be rich, with various interaction patterns already identified in other biological systems. Within the target gene groups picked by the model, we identify many regulatory region motifs that are highly significant, which suggest that many interactions predicted by the model are likely to be actually present.

2.2 Preliminary Data Processing

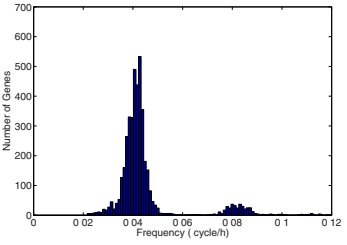
Data analyzed in this paper has came from two time course microarray experiments conducted with *Cyanothece* sp. ATCC 51142 (Cyanothece hereafter), a unicellular cyanobacterium, using Agilent (www.agilent.com) custom-made two channel Microarrays. In [16], cells have been grown under alternating 12h light and dark conditions. We refer to this experiment as DLDL hereafter and the samples are extracted every 4h over a period of 48h. In [18], cells are kept in alternating 12h light-dark cycles before being transferred to constant light conditions. This experiment is referred to as LDLL hereafter. The samples are obtained under both light-dark (LD) and constant light conditions (LL). In both experiments, Cyanothece cells are grown under nitrogen fixing conditions. The microarray raw data are normalized using LOWESS normalization algorithm to avoid systematic intensity based bias which is commonly observed in two channel microarray chips, [14]. The statistical analysis is based on student T-test and indicates a good consistency in the data. The preliminary data processing steps have been described in [16] and [18].

2.2.1 Identification of Cyclic Genes

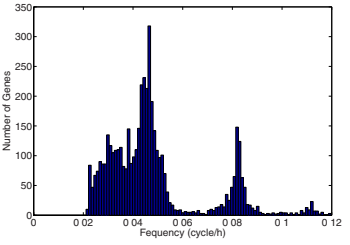
Several approaches for identification of cyclic behaviors have been discussed in the literature. In [12], a Fourier score based method with random permutation is introduced for the detection of cyclic behaviors. For any time course signal x_i, Fourier score at a given frequency f can be defined as

$$F_i = \sqrt{\left(\sum_t \sin \omega t \cdot x_i(t)\right)^2 + \left(\sum_t \cos \omega t \cdot x_i(t)\right)^2}, \tag{2.1}$$

where $\omega = 2\pi f$. In order to identify the main frequency components of the gene expressions, Fast Fourier Transform (FFT) [6], is performed on the mean deducted data. In Figure 2.2a we have shown that the histogram for the distribution of the main frequencies in DLDL. A clear sharp peak observed at .0415/h corresponds to oscillations with 24h period. We also notice that a considerable number of genes have a main frequency close to .083/h which corresponds to 12h



(a) Distribution of main frequencies in DLDL.



(b) Distribution of main frequencies in LDLL.

Fig. 2.2. Main frequencies present in the gene expressions are found using FFT. Distribution of frequencies in two experiments shows clear differences, suggesting significant influence by the incident light pattern.

period. In a similar analysis of data from LDLL, the histogram is observed to have a spread over a wider range of frequencies, and crowded towards the lower end as observed in Figure 2.2b. The frequencies corresponding to extended periods (40h-48h) are mainly assigned to those genes which failed to continue their cyclic behavior once the light has been switched to LL. However even in this situation, a significant number of genes maintain a main period of around 24h and 12h. A clear difference between the two histograms, resulting from changes in the transcription levels of a large number of genes, confirms the significance of light control as opposed to an internal circadian clock control. Both 24h and 12h periods are subsequently used for the calculations of the Fourier score.

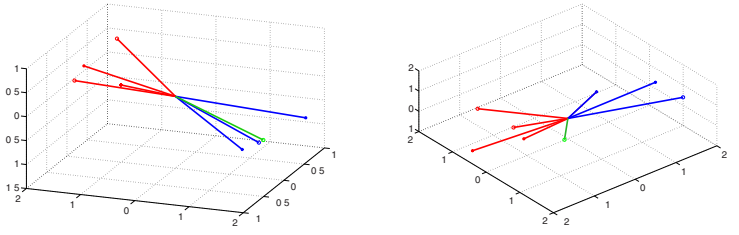
If a given signal consists of a dominant cyclic component of a specific frequency, we expect it to have a large Fourier score. In order to quantify the significance of the Fourier score, we compare the value for the original expression with the Fourier scores obtained for a large collection of artificially generated signals. These signals are constructed by randomly permuting the original gene expressions at various time points. We used the False Discovery Rate (FDR) [2] based approach to estimate the expected rate of false classifications. An empirical FDR for a chosen threshold t for the Fourier score can be define as,

$$FDR(t) = \frac{\sum_{j=1}^{M}\sum_{k=1}^{N} I(F_{j,k} \geq t)/M}{\sum_{k=1}^{N} I(F_k^o \geq t)}, \tag{2.2}$$

where $M, N, F_{j,k}, F_j^o$ are respectively the number of permutations used for the null hypothesis, the total number of genes, the Fourier score for the random signal j obtained from gene k and Fourier score for the original expression of gene k. $I(x)$ is an indicator function taking values,

$$I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

For all our calculations we picked the value of M to be $10,000$. Original signals are scaled to have a unit standard deviation, so that Fourier scores of different



(a) Hydrogenase gene *cce_2318* (b) Hydrogenase gene *cce_1063*

Fig. 2.3. Distribution of vectors corresponding to different light regimes for two Hydrogenase genes. A gene which does not change its behavior significantly under subjective dark is shown in Figure 2.3a while Figure 2.3b shows a gene which changes its behavior significantly. Red-under light, Blue-under dark and Green-under last 12h in LDLL

genes are comparable. A Fourier score cutoff of 6.5 has been selected for both experiments. This correspond to a FDR of 2% for the DLDL and 3% for the LDLL. We have allowed slightly relaxed threshold for the LDLL, based on initial observations made in [18], that some genes showed oscillations with lower amplitude under constant light conditions. Using this criteria we have identified a total of 2138 and 1056 genes to have significant cyclic behavior in DLDL and LDLL experiment respectively. As shown earlier, most of these oscillatory genes have periods around $24h$ while some have oscillations with a period around $12h$.

2.2.2 Identification of Transient Behaviors Using Angular Distance

The method described above has primarily been used to separate cyclic behaviors from non-cyclic ones. In this subsection, we propose a method based on angular distance, in order to correctly classify the transient behaviors under constant light conditions. We separate the time course data into 12h regions based on the input conditions, light and dark. This gives us four 3-dimensional vectors for each gene, for each of the two experiments. We calculate the pair wise angular distance between different pair of vectors using,

$$d_{1,2} = \left(1 - x_1 {x_2}^T / (x_1 {x_1}^T)^{1/2} (x_2 {x_2}^T)^{1/2}\right), \tag{2.3}$$

where x_1 and x_2 are two vectors that correspond to two input regions. The angular distance $d_{1,2}$ can have values in the range from 0 to 2 where 0 represent vectors having same direction while 2 represent vectors in opposite directions.

The idea of using angular distance for characterizing gene behavior under different light regimes is graphically shown in Figure 2.3. It shows the distribution of vectors corresponding to different light regimes for two selected genes. First gene shows oscillations under both conditions while second gene ceases to oscillate under constant light conditions. Clearly for the gene which show change in its behavior under constant light conditions, the vector corresponds to last

12h in LDLL is located away from vectors corresponds to regular light and dark regimes.

A combination of two methods improves the correct classification of gene behaviors significantly. Especially we are able to identify subtle changes in the behavior patterns under constant light conditions. Table 2.1 lists the final classifications of genes based on their behaviors. We identify 448 and 5 genes with 24h and 12h oscillations respectively, under both alternating and constant light inputs. These identified genes are strongly circadian controlled. We also identify 722 and 45 additional genes having corresponding frequencies. However the expressions of these genes are altered significantly under constant light conditions. We refer to them as light responding genes. Additionally there are some genes showing oscillations with different frequencies under two conditions.

Table 2.1. Classification of genes based on their behavior in each of the two experiments. The periods $24h$ and $12h$ correspond to the periods of the primary oscillations. N.C: Not Cyclic

	Experiment-DLDL	
Experiment-LDLL	24h	12h
24h	448	3
12h	49	5
N.C.	722	45

In Figure 2.4 we present the various gene groups using a graphical representation. Two main gene groups; 'Circadian Controlled' and 'Light Responding' genes with 24h oscillations are shown as two rings. Genes are separated into different clusters based on the peak times of their activities. Number of genes having ultradian oscillations (i.e. oscillations with less than 24h periods) is also shown. Genes are colored based on their activity levels; red representing high activity level while blue indicating lower activity level.

2.3 Dynamical System Model Based on Feed Forward Loops

In order to explain the existence of different behavioral patterns and to study possible interactions between genes, we propose a dynamical systems model based on 'Feed Forward Loop, introduced in [4]. This model can be presented as,

$$\dot{Y}(t) = -\alpha_y Y(t) + \beta_y f(X(t), K_{xy}), \tag{2.4}$$

$$\dot{Z}(t) = -\alpha_z Z(t) + \beta_z g(X(t), Y(t), K_{xz}, K_{yz}), \tag{2.5}$$

where $X(t)$, $Y(t)$ and $Z(t)$ represent expression levels of genes X, Y and Z respectively. The activation function $f(X(t), K) = (X(t)/K)^H/(1 + (X(t)/K)^H)$ has two parameters H and K. The parameter H controls the steepness of $f(u, K)$. Its value is kept at $H = 2$ in [4]. However, as discussed in 2.3.1 and

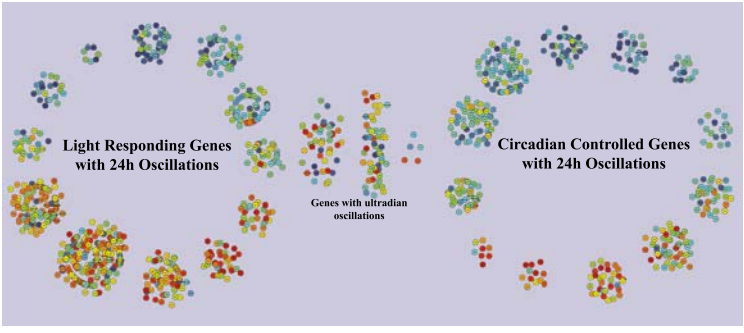


Fig. 2.4. Main gene categories identified using gene classification methods. 'Circadian Controlled' and 'Light Responding' genes with 24h oscillations are clustered in small groups based on their activity levels; red representing high and blue representing low, at a given point of time. We also identified several genes with ultradian oscillations

2.3.1, we select both $H = 1$ and $H = 2$ depending on the gene group we model. The parameter K defines the expression of Gene X required to significantly activate the expression of the other genes. We assume that the regulators operate away from the saturated regions and pick $K >> X(t)$. In [4], the genes X and Y of 2.5, are assumed to be acting independently and $g(t)$ was selected to have the form $g(t) = f_x(X(t), K_{xz})f_y(Y(t), K_{yz})$. In this paper, we consider an additional form, $g(t) = f_x(X(t), K_{xz}) + f_y(Y(t), K_{yz})$, where X and Y acts additively.

The models 2.4 and 2.5 are Linear Time Invariant (LTI) dynamical system with $f(t)$ and $g(t)$ being inputs. These models can be solved analytically and the solutions are given by,

$$Y(t) = e^{-\alpha_y t}Y(0) + \beta_y \int_0^t e^{-\alpha_y(t-\sigma)}u(\sigma)d\sigma. \tag{2.6}$$

with $u(t)$, the input to the system.

Since the system is asymptotically stable, for large values of time t the first term can be ignored. Moreover when $u(t)$ is a periodic function, the expression of the target gene $Y(t)$ would also be oscillating with the same frequency but possibly with some phase shift.

2.3.1 Explaining Different Gene Groups Using the Model

Based on the model, oscillations of the target genes are determined by the oscillations of their regulators. Different types of regulatory relationships give rise to different patterns of behaviors. We assume that some of the higher level regulators get input from two global factors, namely circadian oscillator and/or external light input and subsequently propagate those signals to the target genes.

Genes with a main period of 24h

We select $H = 1$ and assume that the most of genes are regulated by a single regulator, which also has a main period of 24h. These regulatory relationships are first modeled using 2.4. For those genes which could not be explained using a single regulator, we assume the regulation relation to be of 2.5, where two regulators act additively. In this case we try to fit the data using the $g(t) = f_x(X(t), K_{xz}) + f_y(Y(t), K_{yz})$.

Based on the model 2.6, target gene would also be oscillating with a period of 24h. If a gene is under circadian control directly or indirectly then it continues to show the same behavior when the light pattern changes to LDLL as well. However if it has a significant direct influence from the incident light pattern, then it ceases to oscillate under the LL condition. This explains the possible mechanism to observe two different groups of genes, first having 24h oscillations under both experiments and second having 24h oscillations only under DLDL.

Genes with a main period of 12h

Similar to the explanation given for the 24h genes, if the regulator itself has a 12h oscillation, then the target would also have the same period. This is just one of the possible scenarios. However it is still not clear how the 12h oscillations are originated at the first place, since the natural oscillations are of 24h period irrespective of whether they are coming from the circadian clock or the oscillatory diurnal cycle of the light input.

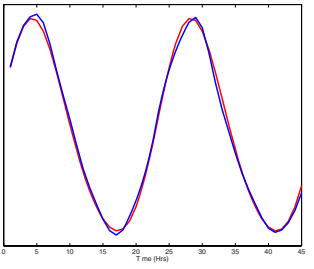
We propose two possible scenarios where a regulator with 24h oscillations can give rise to 12h oscillations in the target. First, it can be according to the model 2.4 with $H = 2$. In this case there is a single regulator gene. Second, it might be based on 2.5 with two independent regulators targeting a single target. In this case $g(t)$ takes the form $g(t) = f_x(X(t), K_{xz}) f_y(Y(t), K_{yz})$. Both these models can generate 12h oscillations with an input having 24h period.

Genes having oscillations with different periods in the two experiments

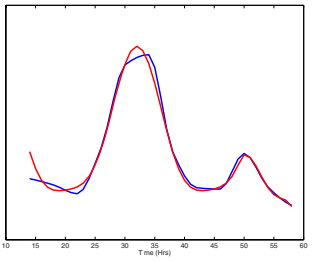
These type of behaviors can easily be modeled by using 2.5 with two regulators working in parallel, thus $g(t)$ taking the form $g(t) = f_x(X(t), K_{xz}) + f_y(Y(t), K_{yz})$. Two regulators oscillate with two different frequencies and depending on the external conditions, their influence on the target would be different, giving rise to different frequencies in the target under two conditions.

2.3.2 Approximation of Gene Expressions

In [4], the authors discuss in detail, how the process of finding numerical solutions to ordinary differential equations, can be simplified. They propose to expand the original expression data using differentiable basis functions so that



(a) Expressions under DLDL



(b) Expressions under LDLL

Fig. 2.5. A good approximation of a gene expression under both the experimental conditions. With 2.7, we are able to get a good approximation for more than 75% of genes

the derivatives can be directly computed. The approximation problem can be efficiently solved using least square techniques.

Since the main behavioral pattern of the data we analyzed is oscillations, we use sinusoidal functions as the basis functions for this problem. Original expressions are approximated as a linear combination of sinusoidal functions along with a linear trend, as given by,

$$X(t) = a + bt + \sum_{j=1}^{N} \alpha_j \sin(j\omega t + \phi_j), \tag{2.7}$$

where $\omega = 2\pi/24$ is the angular frequency corresponding to 24h and ϕ_i is the phase angles of the approximated signals. Parameters a, b, α_j and ϕ_j are estimated using least square optimization method. Since the original data is sampled at 4h, N is limited to 2.

In order to better capture the transient behavior of 'light responding' genes under constant light conditions, this approximation is done separately for oscillatory region and the transient region. With the selected parameters, model captures at least 75% of total energy in the original signal for more than 99% and 80% of genes in DLDL and LDLL respectively. In Figure 2.5, we show approximation of an expression of a light responding gene using 2.7 under two experiments. Genes that are not approximated accurately with this model are excluded from further analysis.

Once the original gene expression is approximated, its derivative can be calculated easily as follows

$$\dot{X}(t) = b + \sum_{j=1}^{N} j\alpha_j\omega_t \cos(j\omega t + \phi_j). \tag{2.8}$$

2.3.3 Model Fitting

Model fitting is done in several steps. For all possible gene pairs, approximated expressions and their derivatives are fitted using model 2.4. Optimal parameter values α and β are obtained using nonlinear least square method, minimizing

$$F(\alpha_y, \beta_y) = \| \dot{Y}(t) + \alpha_y Y(t) - \beta_y f(X(t)) \| . \tag{2.9}$$

Using the optimal parameter values, the normalized error is calculated using

$$\text{Normalized Error} = \mathrm{F}(\mathrm{k_x}, \mathrm{k_u})^2_{\mathrm{opt}} / \| \dot{\mathrm{x}}(\mathrm{t}) \|^2. \tag{2.10}$$

Gene pairs giving rise to a normalized error $<= 10\%$ are considered as possible regulator-target pairs.

If a gene cannot be approximated using a single regulator, we try to fit the data using 2.5. If a particular target is approximated well using a single regulator in the other experimental condition, that regulator is picked as one of the candidates. This is based on the assumption that most regulatory relationships are preserved under changing conditions but additional regulators can be recruited, specific to the different conditions. If the selected gene does not produce a good model fitting in conjunction with any another gene, acting as the second regulator, we try the possibility of other gene pairs as regulators, starting with those that gives rise to smaller errors.

2.3.4 Robustness of the Regulatory Links

Robustness is an essential feature in gene regulations. Biological systems are required to be able to maintain the proper target-regulator relationship in the presence of various disturbances arising from external and internal causes. In order to evaluate the robustness of the regulatory links identified using the model, we changed the parameters α and β by $\pm 5\%$ from the optimal values and error is calculated using 2.10 with the modified parameters.

2.3.5 Link Filtering

One of the challenges in deriving Gene Regulatory Networks (GRN) is identifying valid links between genes, from many possible candidates. Since the number of different time points is significantly less than the variables in the system, problems are mostly under determined. As a result identification of most likely relationships needs to be performed using known biological insight about the system. Following are some of the assumptions generally made about the gene interactions in bacteria.

1. Genes having same phase are likely to regulated by a single regulator.
2. Biological networks tend to follow power law; few hubs with many genes and many hubs with few genes.

3. Regulatory links between genes are likely to be preserved under changing conditions. Level of influence of regulators might change under different treatment/condition and may become visible only under a specific condition.
4. Genes located in close proximity in the genome may belong to a single operon and are regulated by a single regulator.
5. Regulatory relationships between genes are resilient to external noise.

On one hand the assumptions described above can be used to filter out some of the possible links between genes. On the other hand a realistic model should be capable of preserving some of the above basic assumptions. So we can use these as a criteria to measure the acceptability of the model for the purpose of explaining the observed data.

2.4 Network Structure

A total of 1251 genes identified as light responding or circadian controlled are used in the analysis. Using 2.7, a total of 1012 genes are well approximated (approximation captured 75% of the energy of the original signal) for both the experiments and network design has been limited to them. We have found that expressions of 968 (95%) genes in DLDL could be explained using single regulator-target model given by 2.4. The remaining genes require at least two regulators and are fitted with the model 2.5. In the case of LDLL only 476 (47%) genes are approximated using 2.4, which consist of 334 circadian controlled genes and 137 light responding genes. Furthermore 24 circadian controlled genes, 307 light responding genes and 44 intermediary genes are approximated using 2.5 while behavior of 166 genes are not captured using either of the models. This clearly shows the existence of a more complex level of gene interactions under transient light patterns.

It is observed that the majority of the possible regulator-target links are resilient to parameter variations. With 5% deviation from the optimal values, more than 75% of those links remained valid (having a model fit with $< 10\%$ error). Final GRN was derived while preserving the properties described in 2.3.5. Only those links which were resilient to parameter fluctuations were considered in the network. The network for DLDL consisted of 167 unique regulator genes while the network for LDLL consisted of 250 unique regulators. This represent about 3.5-5% of the total genome. It should be noted that in other well studied bacterial systems such as E.*coli*, percentage of transcription factors is around 3.7%.

Number of targets for a given regulator varied from 1-65, with a power-law distribution that has an exponent of -1.9. Using a robust least squares fit we note that the correlation coefficient is 97% for the log-log plot between the distribution of the number of targets and their frequencies.

2.4.1 Direct Regulation vs. Indirect Regulation

Cyanothece genome consists of 194 annotated regulatory-function genes representing about 4% of total genes. Out of them 28 genes were included in the network, representing 2.7% of the total gene in the network. We found that 304 genes in the network can be associated with these 28 genes using either 2.4 or 2.5. We identify these links as likely direct regulation between genes. Other links might represent either indirect regulations or unclassified regulatory functions.

2.4.2 Core Network and Extended Network

Network for DLDL consists of 607 regulatory links while LDLL network consist of 822 links. We see that some of the interactions have more influence in one condition compared to the other. As observed in [13], it suggests the existence of superimposed circadian signaling and diurnal signaling, where one type becomes significant under specific conditions.

There are only 130 conserved links under both conditions. This number represents close to 10% of the entire network. We identify that these genes belong to a core gene network. The remaining links are condition specific, indicating that they have a significant influence only during one experimental condition. These genes give rise to an extended network. In the core network, only at 5% of the times, a gene with a known regulator-function is present as a regulator. In contrast, among extended network, this percentage rises to 26%. We believe that this is an indication of the dynamic role of the regulatory genes required in order for the genome to adapt to changing environmental conditions.

Furthermore in the core-network, 70% of the target genes belong to the 'Circadian Controlled' group. Here 80% of the regulators came from the same group. However in the extended network, Circadian Controlled genes represent only 35% of the targets and regulators. The remaining genes are from 'Light Responding' group. Figure 2.6 show the distribution of genes in the core network and the extended network.

We observe clear correspondence between the number of links in the network and the gene categories identified in the previous work [8]. In the final network, 33% and 61% of the 'Circadian Controlled' genes had just 1 and 2 regulators respectively. In contrast only 4% of 'Light Responding' genes had a single regulator. Further 28% and 67% of 'Light Responding' genes contained 2 and 3 regulators respectively. Among genes identified as 'Intermediary' namely having two dominant frequencies in the two conditions, 92% had 3 regulators in the final network.

2.4.3 Regulation of Possible Operons

Genes belonging to a single operon consist of a single regulatory region and are transcribed as a group. However depending on the respective positions in the operon their transcription levels show differences. A transcription control model should be flexible enough to assign genes belonging to possible operons to a

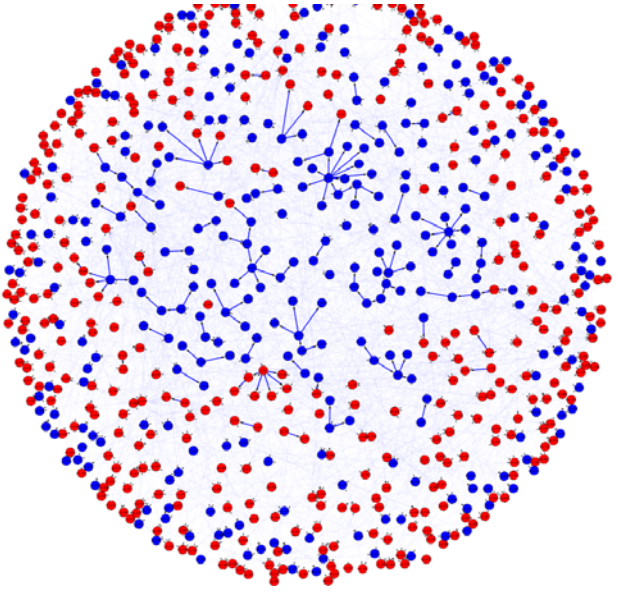


Fig. 2.6. Distribution of genes in the final network. Genes belonging to the core network are located in the center of network and are rich in 'Circadian Controlled' genes. Blue: 'Circadian Controlled', Red: 'Light Responding'

single regulator, despite the changes in the transcription levels. As explained in 2.4.7, we treated those genes, located in the same DNA strand and have a separation of less than 100 base pairs between their Open Reading Frames (ORFs), as members of an operon. Among the genes in the network there were 275 such genes giving rise to 110 operons. We observe that genes in 43 operons can be associated with the same regulator. Expressions of genes from different groups are significantly different so that they are not associated with the same regulators.

2.4.4 Regulators of Different Biological Processes

Some of the regulators in the network are associated with specific biological processes. The significance of the dependence between the regulator and the biological process is measured using 'Fisher's Exact Test',[1]. In Figure 2.7, distributions of target genes for top regulators are shown. It is clear that many regulators are associated with only a few pathways. Except for the first 10 regulators, others are associated only with $<= 5$ different pathways. Similarly most of the important biological processes are associated with only a few regulators.

2.4.5 Phase Difference between Regulator-Target Pairs

One of the important features of the transcription control model, proposed in this paper, is the ability to associate genes with possible phase differences. Moreover,

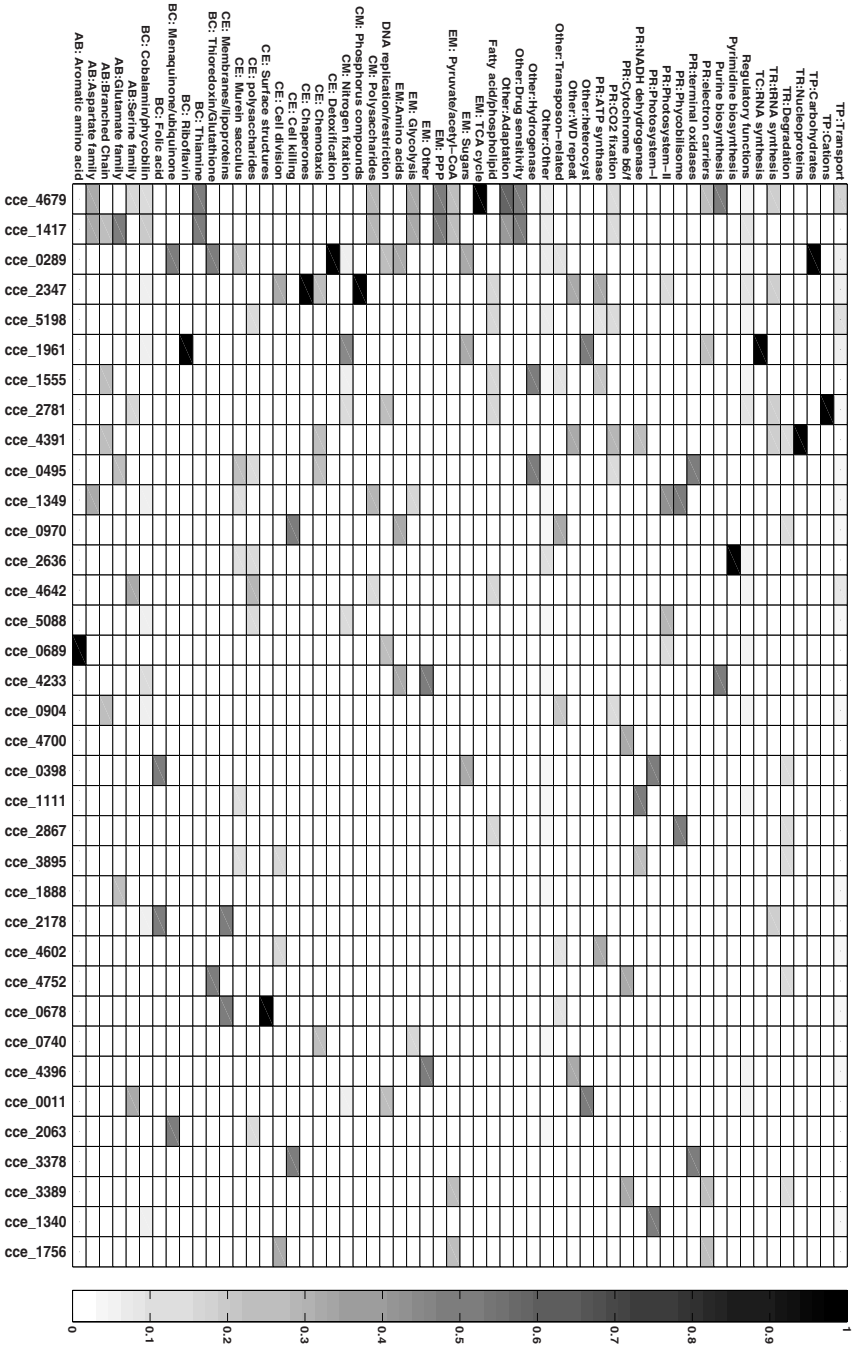
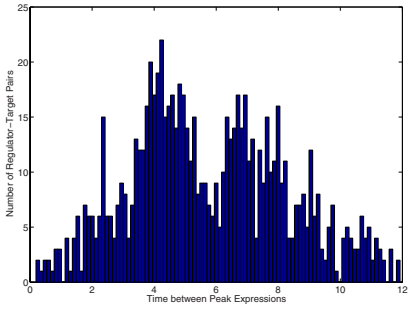
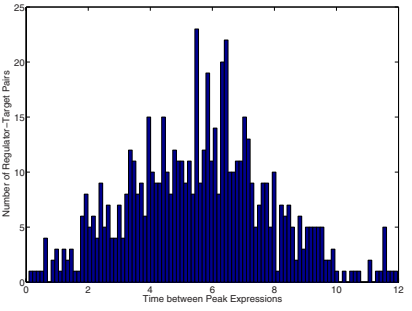


Fig. 2.7. Distributions of different gene functions among different regulators



(a) For Activator type Regulations



(b) For Repressor type Regulations

Fig. 2.8. Phase Differences between Regulator-Targets pairs. We observe that for majority of regulations there is 4-6 hour time difference between peak times of regulator and target

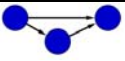
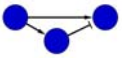
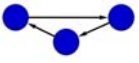
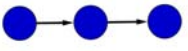
using the phase difference between regulator and target, it is possible to identify if the particular interaction is positive (inductive) or negative (repressive). Based on the final GRN (Gene Regulatory Network), majority of the phase differences between regulator-target pairs are observed to be between 4–5h. The gene pairs showing a phase difference close to $12h$, are identified as having a negative relationship. We observed that close to 45% of genes show negative regulation. This suggests that in a bacterial system, both inductive as well as repressive regulation takes place with similar proportions. This is observed in E.*coli*, where activator and repressor percentages are 48% and 52% respectively. Figure 2.8 present the histogram for distribution of phase differences between regulator-target pairs.

A close examination of expressions of genes classified as 'Light Responding' shows that majority of them alter their regular oscillatory behavior, once switched to constant light conditions, only after some delay. This fact supports the time delay observed, in the model, between the regulator and target genes.

2.4.6 Network Motifs

Even though the individual interactions are determined using FFL, resultant GRN showed a very rich structure, containing many network motifs reported in other systems [19, 3]. There are about 70 genes which does not seem to have any regulatory role and have only a single regulator. The majority of the genes are modeled by 2.5 and has more than two regulators. The obtained network consists of both coherent (regulators acting as inducers) and incoherent (at least one regulator acting as a repressor) type FFLs. It also contains several genes with auto-regulatory and cyclic type relationships. There are many regulator chains observed in the network as well. In several of these chains, we see one known regulatory gene acting as a regulator for another known regulatory gene,

Table 2.2. Some of the already known network motifs found within the gene interactions

Type of Motif	Structure	Number of Occurrences
Auto Regulation		4
Coherent FFL		10
Incoherent FFL		9
Cyclic		1
Single Input		70
Multiple Input		> 300
Chain		70

giving rise to a hierarchy of regulators. In Table 2.2, we list some of the common network motifs and their number of occurrences.

2.4.7 Regulatory Motifs

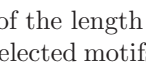
If the target genes associated with a given regulator are truly interacting, we expect them to share a common regulatory region motif. This idea can be used as a method of measuring the accuracy of the GRN. If we can find over-represented motif among the possible targets of a given regulator, it increases the chances of those regulatory relationships to be actual and direct.

In order to identify conserved regions in the upstream regions of the genes we use multiple sequence alignment program 'CONSENSUS', [10]. In bacterial systems, the upstream regions of genes are not well characterized. As a result we used following criteria to extract the relevant regions.

1. If two genes in the same strand are separated by less than 100 nucleotides, we consider them to be a part of an operon. Then we move forward in the strand until we have a wider separation between genes and consider the upstream regions corresponding to relevant gene, so obtained. We make sure that the upstream region of an operon is included only once in the calculation.
2. Criteria for minimum separation is applied only for genes in the same strand. If consecutive genes are on opposite strand we do not treat them as co-regulated.
3. Upstream region is limited to 500 base pairs forward or sequence up to end of the gene located ahead, which ever the shorter.

Table 2.3. Some of the regulator genes and over-represented upstream region motifs of their targets.

$$Ratio = \frac{\%\ of\ times\ motif\ was\ present\ in\ target\ genes}{\%\ of\ times\ motif\ was\ present\ in\ rest\ of\ the\ genes}$$

Gene	Function	Motifs	P-Value	Ratio
cce_1349	Other categories		6.10E-09	96.8
cce_3378	Regulatory		5.70E-15	54.3
cce_2124	Branched chain		4.09E-16	51.1
cce_2540	Regulatory		1.23E-15	51.1
cce_0206	Other categories		3.54E-13	41.4
cce_0398	not available		5.50E-23	40.3
cce_3206	not available		4.40E-17	35.9
cce_0970	Regulatory		2.57E-19	29.1
cce_1083	not available		5.31E-18	27.8
cce_4602	not available		1.86E-23	26.4
cce_1555	not available		8.04E-23	23.0
cce_1978	not available		8.57E-13	20.7

We search for the consensus sequence of the length 8 in the upstream regions of the relevant genes. Significance of the selected motifs are evaluated by comparing the proportion of genes containing the given motif among the possible targets and among the rest of the genes in the network.

One of the ways to measure the reliability of a GRN is to check whether the target genes corresponding to a particular regulator, contain over-represented motifs in their upstream regions. Analysis of the upstream regions using 'CONSENSUS' is able to predict several conserved regions. The significance of the obtained motifs to be non-random is calculated by the algorithm and results in very small p-values. Additionally, we calculate the ratio of observing the motifs among the target genes and compare that to all the remaining genes. There are many motifs for which this ratio exceeded 20. Using these two criteria we are able to identify several highly specific probable binding site motifs. Table 2.3 lists some of the highest ranked motifs. Figures of conserved motifs are generated using 'WebLogo3' [5].

As observed in many experimentally verified transcription factor binding sites, we see some conserved nucleotides in the vicinity of the predicted motifs. This increases the chances of these motifs being true binding sites for transcription factors.

2.5 Conclusion

In this paper, we have analyzed two microarray data sets with the objective of identifying oscillatory behaviors in genes in *Cyanothece* under diurnal conditions. Using two different approaches, we have characterized sustained and altered oscillations under changing input conditions. Based on this we have identified two main groups of genes, viz. 'Circadian Controlled' and 'Light Responding' genes. Subgroups are created based on the main frequencies of their oscillations.

We have proposed a biologically realistic dynamical systems model based on FFL for identification of interactions between diurnal genes. The proposed model has several desirable features. With selection of appropriate parameters and suitable input functions, the model is able to capture the main dynamics of the measurements. It has successfully reproduced oscillations with multiple frequencies, altered and damped oscillations under changing input conditions, etc. The model is clearly stable, which is an essential feature of any biological system. Interactions identified using the model are directional; i.e. the target and regulator are clearly defined. Since the model allows a phase shift between input and output, it can accommodate the delay between the transcription of a regulator and the action of corresponding protein, controlling its target after translation and post translational modifications. These types of relationships are not modeled by traditional correlation based methods.

Majority of the transcriptional relationships inferred by the model are shown to be consistent under parameter modifications. It implies that the relationships we detect are resilient to small variations in the signals and/or parameters. This is an important feature, any realistic biological model should posses.

The model is able to infer interactions for more than 80% of the genes considered to be diurnal and used for the analysis. We have shown that the network for DLDL experiment, where cells are under regular dark/light cycle, has considerably less number of interactions compared to that for the LDLL experiment. This is due to various alterations of gene expressions under constant light conditions. We infer that this added complexity of the network indicate additional regulatory relationships, which become visible under altered environmental conditions. We have identified consistent links between two conditions and found that majority of the genes involved in those links are categorized as 'Circadian Controlled'.

Using the model we are able to associate about 30% of the genes that are already known to be regulatory. We have also identified about 100 possible operons based on the gene locations in the genome and we showed that genes in 43 of them could be associated with the same regulators. Model also suggested that many of the important biological processes are primarily controlled by relatively small number of regulators. We see that there is about $4--5$h time lag between regulator and target genes. This is in good agreement with the delay observed in gene expression data from one behavior to an altered behavior once the incident light is switched from oscillatory to constant condition. Many of the above features are characteristics observed in other bacterial systems as well and the model proposed here is able to capture them to a great extent.

The final network is rich with many known network motifs. In addition to the FFL, which the proposed model is based on, a variety of other network structures are observed, such as auto-regulations, cyclic regulations, single and multi input genes and chain type of regulations. We have identified many hierarchical regulatory relationships as well.

From the upstream regions of the target gene groups we are able to detect many conserved binding site motifs. We have shown that many of these motifs are very specific to selected gene groups. Also we are able to detect several conserved nucleotides in the vicinity of the identified motifs. These observations increase the possibility that these regions are indeed transcription factor binding sites. It also increases the acceptability of the proposed network model.

Finally we would like to acknowledge that the proposed network is not complete. In this paper our focus has been limited to only the diurnal genes. We show that the proposed feed-forward loop based model is sufficient to capture many interaction patterns between these genes. In this paper, we identify the network under regular day/night conditions as the core network. We also identify additional interactions under modified light conditions. It is quite possible that more interactions would become visible if system is perturbed by other conditions. With the availability of such data, the model might need to be refined.

Acknowledgement. This work is part of a Membrane Biology EMSL Scientific Grand Challenge project at the W. R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U. S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle. The project is also supported in part by National Science Foundation FIBR program under grant number 0425749.

We would like to thank Jana Stockel and Joerg Toeple for sharing their unpublished data, manuscripts etc. We would also like to thank Wenxue Wang, Abhay Singh and Eric Welsh for number of fruitful discussions and suggestions and help provided in drafting the manuscript.

References

1. Agresti, A.: Categorical Data Analysis, 2nd edn. John Wiley & Sons, Inc, Chichester (2002)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300 (1995)
3. Blais, A., Dynlacht, B.D.: Constructing transcriptional regulatory networks. Genes. Dev. 19, 1499–1511 (2005)
4. Cao, J., Zhao, H.: Estimating dynamic models for gene regulation networks. Bioinformatics 24, 1619–1624 (2008)
5. Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E.: WebLogo: A Sequence Logo Generator. Genome Res. 14, 1188–1190 (2004)
6. Duhamel, P., Vetterli, M.: Fast Fourier Transforms: A Tutorial Review and a State of the Art. Signal Processing 19, 259–299 (1990)

7. D'Haeseleer, P., Liang, S., Somogyi, R.: Genetic network inference: From co-expression clustering to reverse engineering. Bioinformatics 16, 707–726 (2000)
8. Elvitigala, T.R., Pakrasi, H.B., Ghosh, B.K.: Controlling diurnal rhythms by light. In: Proc.of the 10^{th} International Conference on Control Automation Robotics & Vision, pp. 1367–1372 (2008)
9. Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., Gardner, T.S.: Large-scale mapping and validation of Escherichia Coli transcriptional regulation from a compendium of expression profiles. PLoS Biology 5(1) (2007)
10. Hertz, G.Z., Stormo, G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563–577 (1999)
11. Liang, K., Wang, X.: Gene regulatory network reconstruction using conditional mutual information. EURASIP Journal on Bioinformatics and Systems Biology (2008)
12. de Lichtenberg, U., Jensen, L.J., Fausboll, A., Jensen, T.S., Bork, P., Brunak, S.: Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics 21, 1164–1171 (2005)
13. Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M., Givan, S.A., Yanovsky, M., Hong, F., Kay, S.A., Chory, J.: Network discovery pipeline elucidates conserved time-of-day specific cis-regulatory modules. PLoS Genetics 4, e14 (2008)
14. Quackenbush, J.: Microarray data normalization and transformation. J. Nature Genetics 32, 496–501 (2002)
15. Stuart, J.M., Segal, E., Koller, D., Kim, S.K.: A Gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255 (2003)
16. Stockel, J., Welsh, E.A., Liberton, M., Kunnavakkan, R., Aurora, R., Pakrasi, H.B.: Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes. Proceedings of the National Academy of Science 105, 6156–6161 (2008)
17. Tagkopoulos, I., Liu, Y.C., Tavazoie, S.: Predictive behavior within microbial genetic networks. Science 320, 1313–1317 (2008)
18. Toepel, J., Welsh, E., Summerfield, T.C., Pakrasi, H.B., Sherman, L.A.: Differential transcriptional analysis of the cyanobacterium Cyanothece sp. Strain ATCC 51142 during light-dark and continuous-light growth. J. Bacteriol. 190, 3904–3913 (2008)
19. Uri, A.: An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/CRC, Boca Raton (2006)

3

On Stability of Limit Cycles of a Prototype Problem of Piecewise Linear Systems

O. Eriksson, J. Tègner, and Y. Zhou

[1] Dept. of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden and The Computational Medicine group, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Solna, SE-171 76 Stockholm, Sweden
[2] The Computational Medicine group, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Karolinska University Hospital, Solna, SE-171 76 Stockholm, Sweden and Division of Computational Biology, Department of Physics, Chemistry and Biology, The Institute of Technology, Linköping University, SE-581 83 Linköping, Sweden
[3] Department of Mathematics, Stockholm University, SE-10691, Stockholm, Sweden

Summary. The purpose of this paper is to develop a machinery to analyze existence and stability of limit cycle of a prototype of piecewise linear systems, possibly with delays in switching rules. The study of this type of problems is motivated by modelling cell cycle regulation. The results are applied to a cell cycle model of fission yeast. It is shown that the cell cycle model has a limit cycle and it is stable and criterion of the stability regions are also given.

3.1 Introduction

Consider the following piecewise linear system of prototype

$$\begin{aligned} \dot{x}(t) &= A_\alpha x(t) + B, \quad \text{for } t > t_0, \\ x(s) &= \varphi(s), \quad \text{for } t_0 - \tau \le s \le t_0 \end{aligned} \tag{3.1}$$

with a switching rule

$$\alpha(x) \in \{1, 2, ...\} \tag{3.2}$$

where $x \in \mathbb{R}^n$ is the state, $\varphi(s)$ is given in $\mathbb{R}^n$, A_α is an invertible $n \times n$ matrix and B is an $n \times 1$ matrix.

In order to distinguish the time at which we inspect the state from the variable passing through the interval $[t_0 - \tau, t_0]$ we shall, as usual in the theory of delay equations (see Hale [13]), write throughout the paper $x_t(s) := x(t+s)$ for $t \ge t_0$ and $t_0 - \tau \le s \le t_0$. With this notation, x_t is the state at time t. Clearly the solution to (3.1) is

$$x(t) = \begin{cases} e^{A_\alpha(t-t_0)}(\varphi(t) + A_\alpha^{-1}B) - A_\alpha^{-1}B, & \text{for } t_0 - \tau \le t \le t_0 \\ e^{A_\alpha(t-t_0)}(\varphi(t_0) + A_\alpha^{-1}B) - A_\alpha^{-1}B, & \text{for } t > t_0 \end{cases} \tag{3.3}$$

for a fixed α.

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 43–55.
springerlink.com

The motivation of studying this class of piecewise linear systems is highly inspired by a desire for understanding the complexity of cell cycle regulation and for making mathematical analysis accessible to these complex systems, as comprehensive as possible. For a detailed study on how a highly nonlinear complex cell cycle model [1] can be reduced to the piecewise linear system described above we refer to [2, 3]. For references on stability analysis of piecewise linear systems without unstructural delay we refer the reader to e.g. [4, 5].

In this paper we shall give a general analysis of the systems defined by (3.1)-(3.2), and prove that there is a limit cycle in the cell cycle system discussed in [3] and that it is locally stable. The stability regions of the limit cycle will also be discussed.

Without loosing insight in detailed analysis, we assume that $\alpha(x) \in \{i, j\}$ and i, j correspond to the following rule

$$\begin{aligned} i &= \begin{cases} 1 & \text{if } Cx_t \leq \theta_1 \\ 2 & \text{if } Cx_t > \theta_1 \end{cases}, \\ j &= \begin{cases} 1 & \text{if } Cx \leq \theta_2 \\ 2 & \text{if } Cx > \theta_2 \end{cases}, \end{aligned} \tag{3.4}$$

with C being a $1 \times n$ matrix. Denote the hyperplanes $Cx - \theta_1 = 0$ and $Cx - \theta_2 = 0$ by S^I and S^D. Note that the index j in (3.4) indicates a delay of τ for (x, α) passing through the hyperplane S^D. Thus we sometimes say the immediate respectively delayed switching plane. For simplicity let S^I lie to the left of S^D and we only consider, throughout the paper, systems where, if $x(t'') \in S^D$ then $x(t) \notin S^D \cup S^I$ for $t \in]t'', t'' + \tau]$. See motivation in Section 3.5.

The paper is organized as follows. In Section 3.2, we study the existence of limit cycles. Then we turn to analysis on stability, especially Section 3.3 deals with local stability and Section 3.4 stability regions. In Section 3.5, we apply the results in Section 3.3 and Section 3.4 to a reduced cell cycle model proposed in [3]. Finally the paper is concluded by some further comments in Section 3.6.

3.2 Existence of Limit Cycle

Assume that a limit cycle generated by (3.3)-(3.4) passes the switching planes in a clockwise consecutive order according to Figure 3.1, and that the delayed switch of this limit cycle after passing S^D from left to right takes place on the right side of S^D, while the delayed switch after passing S^D from right to left takes place on the left side of S^I. Then a limit cycle solution can be constructed by integrate the subsystems according to the switching rules, (3.3) and (3.4), respectively.

Let $\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3, \widetilde{x}_4$ be the intersection points between the trajectory generated by (3.3)-(3.4) and the hyperplanes S^I and S^D as indicated in Figure 3.1.

Let the time taken from $\widetilde{x}_i$ to $\widetilde{x}_{i+1}$ be $t_{i,i+1}$ where $i+4 = i, i = 1, 2, 3, 4$. Then $t_{34} < \tau < \min(t_{23}, t_{34} + t_{41})$ according to the assumptions on delay described earlier.

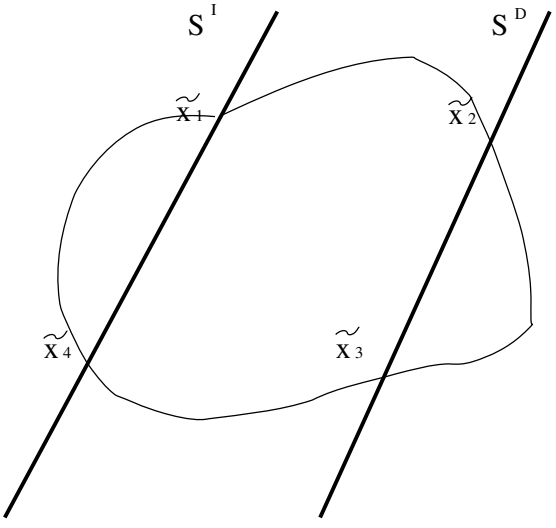


Fig. 3.1. Switching planes and trajectory of limit cycle.

Note that the matrices A_{11}, A_{12}, A_{21} and A_{22} are assumed to be invertible. Thus the solution (3.3), together with the switch rules (3.4), can be written explicitly as follows:

$$\begin{aligned}
\widetilde{x}_2 &= e^{A_{21}t_{12}}(\widetilde{x}_1 + A_{21}^{-1}B) - A_{21}^{-1}B,\\
\widetilde{x}_3 &= e^{A_{22}(t_{23}-\tau)}(e^{A_{21}\tau}(\widetilde{x}_2 + A_{21}^{-1}B) - A_{21}^{-1}B) + A_{22}^{-1}B) - A_{22}^{-1}B,\\
\widetilde{x}_4 &= e^{A_{22}t_{34}}(\widetilde{x}_3 + A_{22}^{-1}B) - A_{22}^{-1}B,\\
\widetilde{x}_1 &= e^{A_{11}(t_{41}-(\tau-t_{34}))}((e^{A_{12}(\tau-t_{34})}(\widetilde{x}_4 + A_{12}^{-1}B) - A_{12}^{-1}B) + A_{11}^{-1}B) - A_{11}^{-1}B.
\end{aligned}$$

Now, by this construction we can in principle formulate the conditions for the existence of limit cycle. Obviously, it takes t_i^* time for actual switch of the system, where $t_1^* = t_{12} + \tau$, $t_2^* = t_{34} + t_{23} - \tau$, $t_3^* = \tau - t_{34}$ and $t_4^* = t_{41} + t_{34} - \tau$.

By successive elimination of $\widetilde{x}_2, \widetilde{x}_3$ in the expression of $\widetilde{x}_1$ we have

$$\begin{aligned}
\widetilde{x}_1 = (I - E_4E_3E_2E_1)^{-1}\,[&E_4E_3E_2(E_1 - I)z_1 + E_4E_3(E_2 - I)z_2\\
&+E_4(E_3 - I)z_3 + (E_4 - I)z_4] \quad (3.5)
\end{aligned}$$

where $E_i = e^{A_i t_i^*}$ and $z_i = A_i^{-1}B$ with $A_1 = A_{21}, A_2 = A_{22}, A_3 = A_{12}, A_4 = A_{11}$. In the same way we obtain

$$\begin{aligned}
\widetilde{x}_4 = (I - E_3E_2E_1E_4)^{-1}\,[&E_1E_2E_3(E_4 - I)z_4 + E_1E_2(E_3 - I)z_3\\
&+E_1(E_2 - I)z_2 + (E_1 - I)z_1]\,, \quad (3.6)
\end{aligned}$$

$$\begin{aligned}
\widetilde{x}_2 = (I - E_{12}E_4E_3E_2E_\tau)^{-1}\,[&(E_{12}E_4E_3E_2(E_\tau - I) + E_{12} - I)z_1\\
&+E_{12}E_4E_3(E_2 - I)z_2 + E_{12}E_4(E_3 - I)z_3 + E_{12}(E_4 - I)z_4]\,, \quad (3.7)
\end{aligned}$$

$$\widetilde{x}_3 = (I - E_{23}E_1E_4E_3E_{34})^{-1}\left[(E_{23}E_1E_4E_3(E_{34} - I) + E_{23} - I)z_2 \right.$$
$$\left. + E_{23}E_1E_4(E_3 - I)z_3 + E_{23}E_1(E_4 - I)z_4 + E_{23}(E_1 - I)z_1\right], \quad (3.8)$$

where $E_\tau = e^{A_1\tau}$, $E_{12} = e^{A_1 t_{12}}$, $E_{23} = e^{A_{21}(t_{23}-\tau)}$ and $E_{34} = e^{A_{21}t_{23}}$. Obviously, $E_{23}E_{34} = E_2$ and $E_\tau E_{12} = E_1$. Since $\widetilde{x}_2$ and $\widetilde{x}_3$ lie on S^D and $\widetilde{x}_1$ and $\widetilde{x}_4$ lie on S^I, respectively, they satisfy $C\tilde{x}_1 - \theta_1 = 0$, $C\tilde{x}_2 - \theta_2 = 0$, $C\tilde{x}_3 - \theta_2 = 0$, $C\tilde{x}_4 - \theta_1 = 0$. Thus we have the following result on the existence of a limit cycle.

Proposition 3.2.1. *Assume that there exists a periodic solution with four switches per cycle and period* $t^* = t_1^* + t_2^* + t_3^* + t_4^* > 0$. *Assume further that* $\tilde{x}_j$*:s are defined by* (3.5)-(3.8) *and* $g_1(t_1^*, t_2^*, t_3^*, t_4^*) = C\tilde{x}_1 - \theta_1$, $g_2(t_1^*, t_2^*, t_3^*, t_4^*) = C\tilde{x}_2 - \theta_2$, $g_3(t_1^*, t_2^*, t_3^*, t_4^*) = C\tilde{x}_3 - \theta_2$, $g_4(t_1^*, t_2^*, t_3^*, t_4^*) = C\tilde{x}_4 - \theta_1$. *Then the following conditions hold*

$$\begin{cases} g_1(t_1^*, t_2^*, t_3^*, t_4^*) = 0 \\ g_2(t_1^*, t_2^*, t_3^*, t_4^*) = 0 \\ g_3(t_1^*, t_2^*, t_3^*, t_4^*) = 0 \\ g_4(t_1^*, t_2^*, t_3^*, t_4^*) = 0 \end{cases}$$

and the period solution is governed by system with A_{21} *on* $[0, t_1^*)$, A_{22} *on* $[t_1^*, t_1^* + t_2^*)$, A_{12} *on* $[t_1^* + t_2^*, t_1^* + t_2^* + t_3^*)$, *and* A_{11} *on* $[t_1^* + t_2^* + t_3^*, t^*)$. *Furthermore, the periodic solution is obtained with initial condition* $\widetilde{x}_i$ *for* $i = 1, 2, 3, 4$.

Note that if the initial condition does not belong to any switching surface the existence of a limit cycle still holds, for the trajectory will cross one switching surface after a finite time.

3.3 Local Stability of Limit Cycles

The idea is to analyze the effect of a small perturbation of the initial condition $\widetilde{x}_1$ on S^I that generates a limit cycle (or other points as defined in the previous section) to the first return map. Let the return map be T from some point in a small neighbourhood of $\widetilde{x}_1 \in S^I$, to the point where the trajectory returns to S^I. It is well-known that the limit cycle is locally stable if all eigenvalues of the Jacobian of T are inside the unit circle.

To this end we have to find the Jacobian of the return map. Starting at $x(t_0) = \widetilde{x}_1 \in S^I$, $x(t) = e^{A_{21}(t-t_0)}(x(t_0) + A_{21}^{-1}B) - A_{21}^{-1}B$, if $t < t_{12} + \tau$, thus $\widetilde{x}_2 = e^{A_{21}t_{12}}(\widetilde{x}_1 + A_{21}^{-1}B) - A_{21}^{-1}B$. Now let $x(t_0) = \widetilde{x}_1 + \widetilde{\delta x_1}$ where $\widetilde{\delta x_1}$ is arbitrary and the norm of which is small, but $x(t_0)$ is on the switching plane, i.e. it is such that $C(\widetilde{x}_1 + \widetilde{\delta x_1}) - \theta_1 = 0$. The solution with this initial condition is

$$x(t) = e^{A_{21}t}(\widetilde{x}_1 + \widetilde{\delta x_1} + A_{21}^{-1}B) - A_{21}^{-1}B.$$

Assuming the solution reaches the switching plane S^D at time $t_{12} + \delta_1 t_{12}$ we have

$$x(t_{12}+\delta_1 t_{12}) = e^{A_{21}(t_{12}+\delta_1 t_{12})}(\widetilde{x}_1 + \widetilde{\delta x_1} + A_{21}^{-1}B) - A_{21}^{-1}B,$$

Taylor expanding the term $e^{A_{21}\delta_1 t_{12}}$, together with the fact that $e^{A_{21}t_{12}}(A_{21}\widetilde{x}_1 + B) = A_{21}x(t_{12}) + B$, gives

$$\begin{aligned} x(t_{12}+\delta_1 t_{12}) &= \widetilde{x}_2 + e^{A_{21}t_{12}}\widetilde{\delta x_1} + e^{A_{21}t_{12}}(A_{21}\widetilde{x}_1 + A_{21}^{-1}B)\delta_1 t_{12} + O(\delta_1^2) \\ &= \widetilde{x}_2 + e^{A_{21}t_{12}}\widetilde{\delta x_1} + (A_{21}\widetilde{x}_2 + B)\delta_1 t_{12} + O(\delta_1^2) \end{aligned}$$

Since the trajectory passes S^D at $t_{12}+\delta_1 t_{12}$, $Cx(t_{12}+\delta_1 t_{12}) = \theta_2$. Then neglecting the higher order terms and using $\theta_2 = Cx(t_{12})$ we have

$$Ce^{A_{21}t_{12}}\widetilde{\delta x_1} + Cv_1\delta_1 t_{12} = 0.$$

where $v_1 = A_{21}\widetilde{x}_2 + B$. If $Cv_1 \neq 0$ (that is, the solution is transversal to S^D), then

$$\delta_1 t_{12} = -\frac{Ce^{A_{21}t_{12}}}{Cv_1}\widetilde{\delta x_1}.$$

Now we have

$$x(t_{12}+\delta_1 t_{12}) = \widetilde{x}_2 + W_1\widetilde{\delta x_1}$$

i.e. $\widetilde{\delta x_2} = W_1\widetilde{\delta x_1}$ where $W_1 = \left(I - \frac{v_1 C}{Cv_1}\right)e^{A_{21}t_{12}}$.

Next, let $x(t_0) = \tilde{x}_2 + \widetilde{\delta x_2}$, $\tilde{x}_2 \in S^D$, $\widetilde{\delta x_2}$ is arbitrary and the norm of which is small but $x(t_0) \in S^D$. Compute now solution with this initial condition and assume it reaches the switching plane S^I at time $t_{23}+\delta_2 t_{23}$. By a straightforward calculation as before:

$$\begin{aligned} x(t_{23}+\delta_2 t_{23}) &= \tilde{x}_3 + e^{A_{22}(t_{23}-t_\tau)}e^{A_{21}t_\tau}\widetilde{\delta x_2} \\ &+ A_{22}e^{A_{22}(t_{23}-\tau)}((e^{A_{21}t_\tau}(\tilde{x}_2 + A_{21}^{-1}B) - A_{21}^{-1}B) + A_{22}^{-1}B)\delta_2 t_{23} + o(\delta_2^2) \\ &= \tilde{x}_3(A_{22}\tilde{x}_3 + M)\delta_2 t_{23} + e^{A_{22}(t_{23}-t_\tau)}e^{A_{21}t_\tau}\widetilde{\delta x_2} + o\delta_2^2) \end{aligned}$$

if $t_{34}+\varepsilon_1 < \tau < \min(t_{23}, t_{34}+t_{41}) - \varepsilon_2$ for some $\varepsilon_1, \varepsilon_2 > 0$. Neglecting the higher order terms and use the same argument as that in computing W_1 yields

$$\delta_2 t_{23} = -\frac{Ce^{A_{22}(t_{23}-t_\tau)}e^{A_{21}t_\tau}}{C(A_{22}\tilde{x}_3 + B)}\widetilde{\delta x_2}.$$

Hence

$$x(t_{23}+\delta_2 t_{23}) - \tilde{x}_3 \approx \left(I - \frac{(A_{22}\tilde{x}_3 + B)C}{C(A_{22}\tilde{x}_3 + B)}\right)e^{A_{22}(t_{23}-\tau)}e^{A_{21}\tau}\widetilde{\delta x_2}.$$

Set $W_2 := \left(I - \frac{v_2 C}{Cv_2}\right)$ where $v_2 = A_{22}\widetilde{x}_3 + B$.

Using similar calculations and neglecting the higher order terms, we obtain

$$\widetilde{\delta x_4} = W_3\widetilde{\delta x_3} = W_3W_2W_1\widetilde{\delta x_1}.$$

$$x(t_{41}+\delta_4 t_{41}) - \widetilde{x}_1 = W_4\widetilde{\delta x_4} = W_4W_3W_2W_1\widetilde{\delta x_1}.$$

where

$$W_3 = \left(I - \frac{v_2 C}{C v_2}\right) e^{A_{22} t_{34}},$$
$$W_4 = \left(I - \frac{v_4 C}{C v_4}\right) e^{A_{11} t_4^*} e^{A_{12} t_3^*}$$

and $v_2 = A_{22}\widetilde{x}_3 + B$, $v_4 = A_{11}\widetilde{x}_1 + B$, and $\tilde{x}_2, \tilde{x}_3 \in S^D$, $\tilde{x}_1, \tilde{x}_4 \in S^I$. Note that to derive W_2 and W_4 we needed a technical assumption that there exist small numbers $\varepsilon_1 > 0, \varepsilon_2 > 0$ such that $t_{34} + \varepsilon_1 < \tau < \min(t_{23}, t_{34} + t_{41}) - \varepsilon_2$.

Now the Jacobian of the return map is $W = W_4 W_3 W_2 W_1$. If the eigenvalues of W are inside the unit circle, then the limit cycle under consideration is locally stable. Therefore we have proved the following theorem.

Theorem 3.3.1. *Consider the piecewise linear system* (3.3) *and* (3.4)*. Assume there exists a limit cycle with period t^* as described in* Proposition 3.2.1*, and that there exist small numbers $\varepsilon_1 > 0, \varepsilon_2 > 0$ such that $t_{34} + \varepsilon_1 < \tau < \min(t_{23}, t_{34} + t_{41}) - \varepsilon_2$. Assume further that the limit cycle is transversal to the switching planes S^D, S^I at $\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3, \widetilde{x}_4$, respectively. Then the Jacobian of the return map T is given by $W = W_4 W_3 W_2 W_1$. Furthermore, the limit cycle (if existing) is locally stable if all eigenvalues of W lie inside the unit circle.*

3.4 On Stability Regions

In this section we discuss the question arising from the global analysis of limit cycles. However, the analysis given below applies to both cycles and fixed points. Our analysis leads to a description of a stability region, that is, all points in this region will generate solutions that will converge to either an asymptotically stable fixed point or an asymptotically stable limit cycle.

To find the stability regions we study the maps from a subset of one switching plane to a subset of another switching plane. We will give conditions to ensure that the maps we have found are contractive, which in turn provide the condition for asymptotical stability of fixed points or limit cycles. To find these maps for the delay piecewise linear system (3.3)-(3.4) we have to set up some necessary notations and definitions.

Let S_1, S_2 be two switching planes. Let $x(0) = \widetilde{x}_1 + \Delta_1$. Define t_{Δ_1} as the set of all times $t \geq 0$ such that the trajectory $x(t)$ with initial condition $x(0) \in S_1$ and $x(t)$ in the closure of the solution set on $[0, t]$. Note that we have taken the initial time $t_0 = 0$, which is not restricted. Define also the set of *expected switching times* of the map, called impact map, from Δ_1 in a subset of S_1, $S_1^d - \widetilde{x}_1$ called departure set, that generates the trajectory to Δ_2 in a subset of S_2, $S_2^a - \widetilde{x}_2$ called arrival set, to which the trajectory arrives, as $\mathcal{T} = \{t \mid t \in t_{\Delta_1}, \Delta_1 \in S_1^d - \widetilde{x}_1\}$. We denote $x(t, \widetilde{x})$ the trajectory generated by the initial condition $\widetilde{x}$.

Now we turn to finding such maps, called impact maps, for the system (3.3) and (3.4). Remember that we have four switching possibilities: $S^I \to S^I$, $S^I \to S^D$, $S^D \to S^D$, $S^D \to S^I$, that make four maps: (i) S_d^I to S_a^D from left to right,

(ii) S_d^D to S_a^D from right side of S^D, (iii) S_d^D to S_a^I from right to left, and finally (iv) S_d^I to S_a^I from left side of S^I. Denote also the expected switching time sets as $\mathcal{T}_i$, where i is in accordance with the above four cases.

Let $x(0) = \tilde{x}_1 \in S_d^I$. Then

$$x(t, \tilde{x}_1) = e^{A_{21}t}\tilde{x}_1 + \int_0^t e^{A_{21}(t-s)}Bds. \tag{3.9}$$

We require that $x(t, \tilde{x}_1) \in S_a^D$. Hence t is a switching time. Note that switching time may not be unique.

Let $x_1 = \tilde{x}_1 + \Delta_1 \in S_d^I$, $x_2 = \tilde{x}_2 + \Delta_2 \in S_a^D$, and $\tilde{x}_1 \in S_d^I$, $\tilde{x}_2 \in S_a^D$. Then $C\Delta_1 = C\Delta_2 = 0$. From the expression of $x(t, \tilde{x}_1)$ above, we have

$$\Delta_2 = e^{A_{21}t}\Delta_1 + e^{A_{21}t}\tilde{x}_1 + \int_0^t e^{A_{21}(t-s)}Bds - \tilde{x}_2 = e^{A_{21}t}\Delta_1 + x(t, \tilde{x}_1) - \tilde{x}_2.$$

Since $C\Delta_2 = 0$ and $C\tilde{x}_2 = \theta_2$,

$$Ce^{A_{21}t}\Delta_1 = \theta_2 - Cx(t, \tilde{x}_1).$$

Assume that $Cx(t, \tilde{x}_1) \neq \theta_2$. Then

$$\frac{Ce^{A_{21}t}\Delta_1}{\theta_2 - Cx(t, \tilde{x}_1)} = 1$$

showing that

$$\Delta_2 = e^{A_{21}t}\Delta_1 + (x(t, \tilde{x}_1) - \tilde{x}_2) \cdot 1 = \left(I + \frac{(x(t, \tilde{x}_1) - \tilde{x}_2)C}{\theta_2 - Cx(t, \tilde{x}_1)}\right) e^{A_{21}t}\Delta_1.$$

Therefore

$$H_1(t, \tau) := \left(I + \frac{(x(t, \tilde{x}_1) - \tilde{x}_2)C}{\theta_2 - Cx(t, \tilde{x}_1)}\right) e^{A_{21}t}, \quad \tilde{x}_1 \in S_d^I, \tilde{x}_2 \in S_a^D, t \in \mathcal{T}_1 \tag{3.10}$$

is the desired map from $\Delta_1 \in S_d^I - \tilde{x}_1$ to $\Delta_2 \in S_a^D - \tilde{x}_2$ for all $t \in \mathcal{T}_1$.

In the same manner, we can derive the other three maps, denoted by $H_2(t, \tau)$ from $\Delta_2 \in S_d^D - \tilde{x}_2$ to $\Delta_3 \in S_a^D - \tilde{x}_3$, $H_3(t, \tau)$ from $\Delta_3 \in S_d^D - \tilde{x}_3$ to $\Delta_4 \in S_a^I - \tilde{x}_4$, and $H_4(t, \tau)$ from $\Delta_4 \in S_d^I - \tilde{x}_4$ to $\Delta_5 \in S_a^I - \tilde{x}_5$.

$$H_2(t, \tau) = \left(I + \frac{(x(t, \tilde{x}_2) - \tilde{x}_3)C}{\theta_2 - Cx(t, \tilde{x}_2)}\right) e^{A_{22}(t-\tau)}e^{A_{21}\tau}, \ \tilde{x}_2 \in S_d^D, \tilde{x}_3 \in S_a^D, \forall t \in \mathcal{T}_2 \text{ and } t > \tau \tag{3.11}$$

$$H_3(t, \tau) = \left(I + \frac{(x(t, \tilde{x}_3) - \tilde{x}_4)C}{\theta_1 - Cx(t, \tilde{x}_3)}\right) e^{A_{22}t}, \quad \tilde{x}_3 \in S_d^D, \tilde{x}_4 \in S_a^I, \forall t \in \mathcal{T}_3 \tag{3.12}$$

$$H_4(t, \tau) = \left(I + \frac{(x(t, \tilde{x}_4) - \tilde{x}_5)C}{\theta_1 - Cx(t, \tilde{x}_4)}\right) e^{A_{11}(t-\tilde{t})}e^{A_{12}\tilde{t}},$$
$$\tilde{x}_4 \in S_d^I, \tilde{x}_5 \in S_d^I, \forall t \in \mathcal{T}_4, t - \tilde{t} > 0, \text{ for some } 0 < \tilde{t} < \tau \tag{3.13}$$

where $Cx(t, \tilde{x}_i) \neq \theta_1$, $(i = 3, 4)$ and $Cx(t, \tilde{x}_2) \neq \theta_2$.

We summarize this as a theorem:

Theorem 3.4.1. *Assume that* $Cx(t,\tilde{x}_i) \neq \theta_1$, $(i = 3,4)$ *and* $Cx(t,\tilde{x}_i) \neq \theta_2$, $(i = 1,2)$ *for* $\widetilde{x}_1 \in S_d^I$, $\widetilde{x}_2 \in S_d^D$, $\widetilde{x}_3 \in S_d^D$ *and* $\widetilde{x}_4 \in S_d^I$. *Define* H_i *as above,* (3.10)-(3.13). *Then, for any* $\Delta_i \in S_d^i - \tilde{x}_i$ *there exists a* $t \in \mathcal{T}_i$ *such that*

$$\Delta_{i+1} = H_i(t,\tau)\Delta_i,$$

Such $t \in t_{\Delta_i}$ *is the switching time associated with* Δ_{i+1}, *for* $i = 1,2,3,4$, *where* $S^1 = S^4 = S^I, S^2 = S^3 = S^D$.

Furthermore, if the initial states are chosen so that these maps are contractive, then the limit cycle, or fixed point, is stable.

Note also that $C\Delta_i = 0$ in the derivation of H_i. This indicates that the maps H_i, in fact, takes place in $\mathbb{R}^{n-1}$. To see this, let $C^\perp$ be an $n \times (n-1)$ matrix with columns orthonormal to C'. Then $\Delta_{i+1} = H_i(t)\Delta_i$ is equivalent to $C^\perp\tilde{\Delta}_{i+1} = H_i(t,\tau)C^\perp\tilde{\Delta}_i$, where $\tilde{\Delta}_i, \tilde{\Delta}_{i+1} \in \mathbb{R}^{n-1}$. Thus,

$$\tilde{\Delta}_{i+1} = (C^\perp)'H_i(t,\tau)C^\perp\tilde{\Delta}_i.$$

Thus

$$\tilde{\Delta}_{i+1} = \tilde{H}_i(t,\tau)\tilde{\Delta}_i.$$

where $\tilde{H}_i(t,\tau) := (C^\perp)'H_i(t,\tau)C^\perp$.

It is in general not easy to check contraction of these maps. However, if the state space is two-dimensional, then the difficulty is reduced significantly. Note that $H_i(t,\tau)$ becomes scalar. To prove that $H_i(t)$ is contractive is equivalent to proving that $|\tilde{H}_i(t,\tau)| < 1$, for each i, since $\tilde{H}_i$ is a scalar, i.e.

$$|(C^\perp)'H_i(t)C^\perp| < 1. \tag{3.14}$$

Next step is to find the largest interval in S^I and S^D around $\tilde{x}_i$ where the impact map from some $U_i \subset S^I(S^D)$ to the next switch on the switching plane is continuous, and a set of initial conditions of interval in S^I or S^D such that every point in this set has switching time in $\mathcal{T}_i$. Define $C_1(t) = Ce^{A_{21}t}C^\perp$, $C_2(t) = Ce^{A_{22}t}e^{A_{21}\tau}C^\perp$, $C_3(t) = Ce^{A_{12}t}C^\perp$, $C_4(t) = Ce^{A_{11}(t-\tilde{t})}e^{A_{12}\tilde{t}}C^\perp$, $d_1(t) = \theta_2 - Cx(t,\tilde{x}_1)$, $d_2(t) = \theta_2 - Cx(t,\tilde{x}_2)$, $d_3(t) = \theta_1 - Cx(t,\tilde{x}_3)$, and $d_4(t) = \theta_1 - Cx(t,\tilde{x}_4)$. We have

Theorem 3.4.2. *Assume that* (3.14) *holds for all* $t \in \mathcal{T}_i := [t_{i-}, t_{i+}]$. *Define*

$$R_i^C = \min_{t\in\mathcal{T}_i} |\dot{d}_i(t)|/|\dot{C}_i(t)|, \quad \bar{R}_i = \inf_{t\notin\mathcal{T}_i} |d_i(t)|/|C_i(t)|.$$

Then the impact map in the domain $\{\tilde{x}_i + C^\perp\tilde{\Delta}_i : |\tilde{\Delta}_i| < \min\{R_i^C, \bar{R}_i\}\}$ *is a contraction.*

The proof is similar to the ones in [4].

If the piecewise linear system has a local limit cycle with period t^*, and the limit cycle crosses transversely 4 switching planes per cycle we can continue discussing the stability region. To find the stability region of the limit cycle we have to find the conditions for contraction of the four impact maps simultaneously. This is summarized in the following:

Theorem 3.4.3. *Let $\mathcal{T}_i$ be the largest set such that* (3.14) *holds for all $t_i \in \mathcal{T}_i$, $i = 1, 2, 3, 4$. Let $R = \min\{R_i : i = 1, ..., 4\}$. Then the solution starting inside of any of the set defined by $\{\tilde{x}_i + C^{\perp}\tilde{\Delta}_i : |\tilde{\Delta}_i| < R\} \subset S^I$ or S^D, i =1,...,4, converges asymptotically to the limit cycle.*

3.5 Application to a Reduced Cell Cycle Model

The purpose of this section is to illustrate that the study carried out in preceding sections is useful in the global analysis of dynamical behavior of the reduced cell cycle model [3]. The reduced cell cycle system is defined as follows:

$$\dot{x}_{\mathrm{Cdc13t}}(t) = -\, s_1(t-\tau)x_{\mathrm{Cdc13t}}(t) + k_1 M, \tag{3.15}$$

$$\dot{x}_{\mathrm{PreMPF}}(t) = s_2(t)x_{\mathrm{Cdc13t}}(t) - s_3(t, t-\tau)x_{\mathrm{PreMPF}}(t), \tag{3.16}$$

$$y_{\mathrm{MPF}}(t) = x_{\mathrm{Cdc13t}}(t) - x_{\mathrm{PreMPF}}(t), \tag{3.17}$$

$$s_1(t-\tau) = k_2' + s_{slp/ste}(y_{\mathrm{MPF}}(t-\tau)), \tag{3.18}$$

$$s_2(t) = s_{wee}(y_{\mathrm{MPF}}(t)), \tag{3.19}$$

$$s_3(t, t-\tau) = s_{wee}(y_{\mathrm{MPF}}(t)) + s_{25}(y_{\mathrm{MPF}}(t)) + k_2' + s_{slp/ste}(y_{\mathrm{MPF}}(t-\tau)), \tag{3.20}$$

$$s_{25}(z) = \begin{cases} l_{25} & \text{if } z \le \theta_{25/\mathrm{wee}} \\ h_{25} & \text{if } z > \theta_{25/\mathrm{wee}} \end{cases}, \tag{3.21}$$

$$s_{\mathrm{wee}}(z) = \begin{cases} h_{\mathrm{wee}} & \text{if } z \le \theta_{25/\mathrm{wee}} \\ l_{\mathrm{wee}} & \text{if } z > \theta_{25/\mathrm{wee}} \end{cases}, \tag{3.22}$$

$$s_{\mathrm{slp/ste}}(z) = \begin{cases} l_{\mathrm{slp/ste}} & \text{if } z \le \theta_{\mathrm{slp}} \\ h_{\mathrm{slp/ste}} & \text{if } z > \theta_{\mathrm{slp}} \end{cases}, \tag{3.23}$$

where the parameters are

$$\begin{gathered} \tau = 15, \quad k_1 = 0.03, \quad k_2' = 0.03, \quad l_{25} = 0.2, \quad h_{25} = 5, \quad \theta_{25/\mathrm{wee}} = 0.25, \\ h_{\mathrm{wee}} = 1.3,\ l_{\mathrm{wee}} = 0.15,\ l_{\mathrm{slp/ste}} = 0,\ h_{\mathrm{slp/ste}} = 1.3,\ \theta_{\mathrm{slp/ste}} = 0.4, \quad \mu = 0.005. \end{gathered} \tag{3.24}$$

In the original model [1], the cell mass M is a slow time dependent variable. We here treat M as a constant parameter in order to examine model behaviour for different values of M. In the following analysis $M = 1.8$.

Let $x = (x_{\mathrm{Cdc13t}} \quad x_{\mathrm{PreMPF}})'$ represent the state of the cell cycle system and $u_{ext} = M$ the external input, and let $y = y_{\mathrm{MPF}}$ be the output from the cell cycle system. Then, the DPL described in the preceding section can be put in the matrix form

$$\dot{x} = Ax + B, y = Cx, \tag{3.25}$$

where $C = (1 \quad -1)$, $A = \begin{pmatrix} -s_1(t-\tau) & 0 \\ s_2(t) & -s_3(t,t-\tau) \end{pmatrix}$, and s_1,s_2,s_3 are combinations of step functions defined by (3.18)-(3.20) and $B = (k_1 u_{ext} \quad 0)'$ and k_1 a constant parameter. The system matrix A takes four possible forms, indexed by A_{ij}, $i, j \in$

$\{1, 2\}$, where $i = i(y(t))$ and $j = j(y(t-\tau))$ (A change of index i corresponds to a change of step functions s_{25} and s_{wee} and a change of j to a change of $s_{slp/ste}$). Then

$$\dot{x}(t) = A_{ij}x(t) + B \tag{3.26}$$
$$y(t) = Cx(t), \tag{3.27}$$

where i and j correspond to the following switching rules

$$\begin{aligned} i(y(t)) &= \begin{cases} 1, & \text{if } y(t) \leq \theta_{25/wee} \\ 2, & \text{if } y(t) > \theta_{25/wee} \end{cases}, \\ j(y(t-\tau)) &= \begin{cases} 1, & \text{if } y(t-\tau) \leq \theta_{slp/ste} \\ 2, & \text{if } y(t-\tau) > \theta_{slp/ste} \end{cases}, \end{aligned} \tag{3.28}$$

and $\theta_{25/wee}$ and θ_{slp} correspond to the switching thresholds of the different step functions. The DPL-model is illustrated in Figure 3.2. The resulting A_{ij}-matrices are obtained from (3.15)-(3.23), and correspond to

$$\begin{aligned} A_{11} &= \begin{bmatrix} -(k_2' + l_{slp/ste}) & 0 \\ h_{wee} & -(h_{wee} + l_{25} + k_2' + l_{slp/ste}) \end{bmatrix} \\ A_{12} &= \begin{bmatrix} -(k_2' + h_{slp/ste}) & 0 \\ h_{wee} & -(h_{wee} + l_{25} + k_2' + h_{slp/ste}) \end{bmatrix} \\ A_{21} &= \begin{bmatrix} -(k_2' + l_{slp/ste}) & 0 \\ l_{wee} & -(l_{wee} + h_{25} + k_2' + l_{slp/ste}) \end{bmatrix} \\ A_{22} &= \begin{bmatrix} -(k_2' + h_{slp/ste}) & 0 \\ l_{wee} & -(l_{wee} + h_{25} + k_2' + h_{slp/ste}) \end{bmatrix}. \end{aligned} \tag{3.29}$$

Here $h_{slp/ste}$, h_{wee}, h_{25} and $l_{slp/ste}$, l_{wee}, l_{25} are the high and low values of the step functions and k_2' is the parameter from the original NT-model [1]. Note that the matrices A_{ij} are invertibe and have all eigenvalues real negative.

Following the discussion in Sections 3.2 and 3.3 we could find a limit cycle going from $\tilde{x}_1 = (1.646, 1.396)$ on S^I: $(1, -1)x = \theta_{25/wee}$, it reaches $\tilde{x}_2 = (1.646, 1.246)$ on S^D: $(1, -1)x = \theta_{slp/ste}$, continues to $\tilde{x}_3 = (0.412, 0.012)$ on S^D, then to $\tilde{x}_4 = (0.257, 0.007)$ on S^I and finally goes back to $\tilde{x}_1$. The period is 112.253 and the switching times $t_1^* = 0.023$, $t_2^* = 16.13$, $t_3^* = 0.40$ and $t_4^* = 95.70$. This is depicted in Figure 3.2.

Note that in order to prove the local stability of a limit cycle we have to show that the Jacobian of the return map W defined in Theorem 3.3.1 should have all eigenvalues inside the unit circle.

If all parameters are fixed as in (3.24), and cell mass $M = 1.8$, we can easily compute the eigenvalues of W to check if the eigenvalues are inside the unit circle. The eigenvalues of W are both ≈ 0.

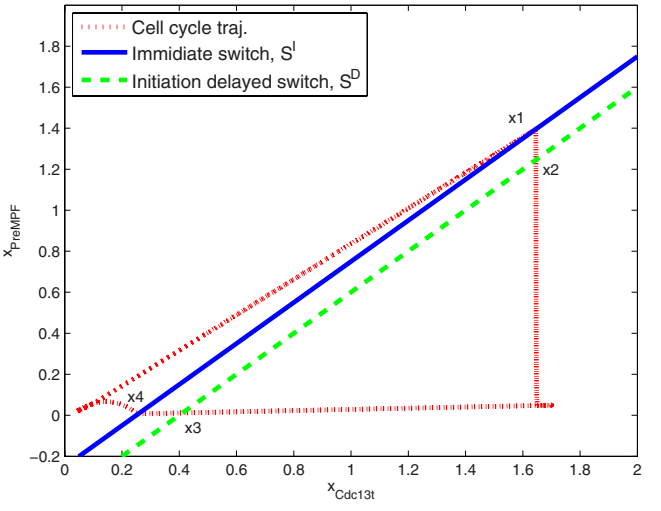


Fig. 3.2. A numerical simulation of the reduced cell cycle model (3.15)-(3.23) together with the switching lines S^I and S^D.

The following estimation will allow us to find parameters such that a limit cycle is locally stable if existing. It is well-known that $|\lambda_i(W)| \leq \|W\|$, where λ_i is denoted as the eigenvalues of W and $\|\cdot\|$ is the norm of an operator $\cdot$ and we shall take the spectral norm, i.e. $\|\cdot\| = \max_{i\in 1,2}(\lambda_i((\cdot)'(\cdot)))^{1/2}$. Now

$$\|W\| \leq \|W_4\|\|W_3\|\|W_2\|\|W_1\|.$$

Then, to guarantee $|\lambda_i(W)| < 1$, it suffices to find conditions such that $\|W_i\| < 1$. To this end we estimate the norms of the matrices W_i.

Lemma 3.5.1. *Let* $A = \begin{bmatrix} \alpha & 0 \\ \gamma & \beta \end{bmatrix}$ *with* $\alpha, \beta, \gamma \in \mathbb{R}$. *If* $\alpha \neq \beta$,

$$e^{2(\alpha+\beta)t} < 1$$
$$1 - (e^{2\alpha t} + e^{2\beta t} + \left(\frac{e^{\alpha t} - e^{\beta t}}{\alpha - \beta}\right)^2 \gamma^2) + e^{2(\alpha+\beta)t} > 0$$

or if $\alpha = \beta$,

$$e^{4\alpha t} < 1$$
$$1 - e^{2\alpha t}(1 + \gamma^2) + e^{4\alpha t} > 0,$$

then $\|e^{At}\| < 1$.

Proof. By a straightforward calculation

$$e^{A't}e^{At} = \begin{bmatrix} e^{2\alpha t} + \left(\frac{e^{\alpha t}-e^{\beta t}}{\alpha-\beta}\right)^2 \gamma^2 & e^{\beta t}\left(\frac{e^{\alpha t}-e^{\beta t}}{\alpha-\beta}\right)\gamma \\ e^{\beta t}\left(\frac{e^{\alpha t}-e^{\beta t}}{\alpha-\beta}\right)\gamma & e^{2\beta t} \end{bmatrix},$$

if $\alpha \neq \beta$. Then the eigenvalues of $e^{A't}e^{At}$ lie inside of the unit circle is equivalent to

$$e^{2(\alpha+\beta)t} < 1$$

$$1 - (e^{2\alpha t} + e^{2\beta t} + \left(\frac{e^{\alpha t}-e^{\beta t}}{\alpha-\beta}\right)^2 \gamma^2) + e^{2(\alpha+\beta)t} > 0.$$

Similarly if $\alpha = \beta$, then

$$e^{A't}e^{At} = e^{2\alpha t}\begin{bmatrix} (1+\gamma^2) & \gamma \\ \gamma & 1 \end{bmatrix},$$

The second alternative follows.

Lemma 3.5.2. *Let* $C = \begin{bmatrix} 1 & -1 \end{bmatrix}$, v_i *defined earlier be* $\begin{bmatrix} a_i \\ b_i \end{bmatrix}$. *Then* $\|W_i\| < 1$, $i = 1, 2, 3, 4$, *if*

$$\sqrt{\frac{2(a_i^2+b_i^2)}{(a_i-b_i)^2}} < \frac{1}{\|e^{A_i}\|},$$

where $A_1 = A_{21}t_{23}$, $A_2 = A_{22}t_{23} + (A_{21} - A_{22})\tau$, $A_3 = A_{12}t_{34}$, $A_4 = A_{11}t_4^* + A_{12}t_3^*$.

Proof. Since A_{21} and A_{22} commute, and A_{11} and A_{12} commute, we have $e^{A_{22}(t_{23}-\tau)}e^{A_{21}\tau} = e^{A_{22}t_{23}+(A_{21}-A_{22})\tau} = e^{A_2}$ and $e^{A_{11}t_4^*}e^{A_{12}t_3^*} = e^{A_{11}t_4^*+A_{12}t_3^*} = e^{A_4}$.

A simple calculation yields that the eigenvalues of $(I - \frac{v_iC}{Cv_i})'(I - \frac{v_iC}{Cv_i})$ are 0 and $\frac{2(a_i^2+b_i^2)}{(a_i-b_i)^2} > 0$, where $a_i \neq b_i$ according to the definition of v_i. Then

$$\|I - \frac{v_iC}{Cv_i}\| = \sqrt{\frac{2(a_i^2+b_i^2)}{(a_i-b_i)^2}}.$$

Hence

$$\|W_i\| \leq \|I - \frac{v_iC}{Cv_i}\|\|e^{A_i}\| < 1,$$

completing the proof.

3.6 Conclusions

We have investigated a class of piecewise linear systems with explicit delay in this paper. The main contribution is giving a set of conditions for local stability of the limit cycle and stability regions of such solutions. Although it is not possible to

provide a fully analytical result, our theory provides a computationally checkable tool based on a rigorous analysis. To deal with unstructural delay was new to our best knowledge.

It is worth pointing out that our analysis, with some small modifications, can be carried out for several switching surfaces and also if the delay occurs in a different way. For the essence of the analysis we have chosen the DPL-structure which we think is the most representative (also in the degree of difficulty).

The theory developed in this paper can also be applied to other models, without assuming that the subsystem matrices are invertible or Hurwitz, by a slight modification in our proofs.

Piecewise linear systems with memory delay both in states and switching rules are under investigation. This will hopefully allow us to analyze systems of delay-differential equation such as the one used in [7].

Acknowledgement. Olivia Eriksson is grateful for the finacial support from SSF (Swedish Foundation for Strategic Research).

References

1. Novak, B., Pataki, Z., Ciliberto, A., Tyson, J.J.: Mathematical model of the cell division cycle of fission yeast. Chaos 11, 277–286 (2001)
2. Eriksson, O., Zhou, Y., Tegner, J.: Modeling complex cellular networks - Robust switching in the cell cycle ensures a piecewise linear reduction of the regulatory network. In: Proc. of the IEEE Conference on Decision and Control, vol. 1, pp. 117–123 (2004)
3. Eriksson, O., Brinne, B., Zhou, Y., Björkegren, J., Tegnér, J.: Deconstructing the core dynamics from a complex time-lagged regulatory biological circuit. IET Syst. Biol. 3, 113–129 (2009)
4. Gonçalves, J.M.: Region of stability for limit cycles of piecewise linear systems. In: Proc. of the IEEE Conference on Decision and Control (2003)
5. Gonçalves, J.M., Megretski, A., Dahleh, M.A.: Global analysis of limit cycles of piecewise linear systems using impact maps and surface Lyapunov functions. IEEE Trans. Automat. Contr. 48, 2089–2106 (2003)
6. Hale, J.K.: Theory of functional differential equations. Springer, New York (1997)
7. Monk, N.A.M.: Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptal time delays. Curr. Biol. 13, 1409–1413 (2003)
8. Tyson, J.J., Hong, C.I., Thron, C.D., Novak, B.: A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. Biophys J. 77, 2411–2417 (1999)

4

On the Existence and Uniqueness of Minimum Time Optimal Trajectory for a Micro Air Vehicle under Wind Conditions

Ram V. Iyer[1], Rachelle Arizpe[2], and Phillip R. Chandler[3]

[1] Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409-1042, USA
[2] Antonian College Preparatory, 6425 West Ave San Antonio, Texas 78213, USA
[3] U.S. Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio 45433-7531, USA

Summary. An important subproblem in the area of cooperative control of multiple, autonomous, unmanned air vehicles is the determination of the minimum-time optimal paths for the agents to fly from one destination to the next. The tasks for the air vehicles are usually tightly coupled in time, and hence estimates of the times taken for each air vehicle to fly from one destination to the next is highly critical for correct assignment of tasks. In this article, we discuss the existence and uniqueness of minimum time solutions for the trajectory planning problem for a Micro Air Vehicle (MAV) under wind conditions. We show that there exists a minimum time solution for the trajectory planning problem with a minimum turn radius constraint for the air vehicle, and for a non-zero, time-varying wind vector field satisfying certain easily checked sufficient conditions. We also prove uniqueness for almost every combination of initial and final conditions in the case of a wind vector field that can vary with time but is constant in the spatial variable at each time instant.

4.1 Introduction

Cooperative Control of multiple, autonomous unmanned air vehicles (UAVs) is an active area of research that holds enormous potential for military and civilian applications [1, 2, 3]. This new paradigm for control has been implemented in the MultiUAV simulation platform by the Air Force Research Laboratory [3]. The MultiUAV platform has a hierarchical architecture, in which, at the highest level the dynamics of the controlled agents are suppressed, and a task allocation for the agents is performed using graph theory. The tasks for the agents are usually tightly coupled in time [3, 4], and hence estimates of the times taken for each agent to fly from one destination to the next is highly critical for correct assignment of tasks. This estimate of times is usually obtained by considering a kinematic model of the air vehicle, along with kinematic turn-radius constraints, to keep the computation time at a manageable level [5, 6, 7]. The key result that is used in this computation is Dubins' result on the existence of minimum time solutions for a kinematic model with minimum turn-radius constraint [8].

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 57–77.
springerlink.com

However, this result is only for valid for zero-wind, and hence all of the available cost estimation algorithms are only valid for zero-wind.

In this article, we discuss the existence and uniqueness of minimum time solutions for the trajectory planning problem for a Micro Air Vehicle (MAV) under wind conditions. Numerical results from algorithms based on this paper can be found in [10]. MAV's are powered by batteries that typically have a very short life [2, 5]. Therefore, before deployment, it is desirable to know: (a) whether the MAV can complete its mission - which requires flight from Point A with velocity V_0 to Point B with velocity V_f, , in the presence of wind; (b) if the answer to the previous item is in the affirmative, then the control inputs that achieve the mission. We model an MAV flying with a constant speed in the wind axes and at a constant altitude. The kinematic equations of motion for the MAV are:

$$\dot{x} = V(\cos\theta, \sin\theta) + W(x,t); \quad \dot{\theta} = u. \tag{4.1}$$

where $x = (x_1, x_2)$, $W(x,t) = (W_1(x,t), W_2(x,t))$, and $W_i(x,t)$; $i = 1, 2$ are functions with bounded derivatives. These equations contain the wind vector field that is not considered in earlier works [5, 8, 6, 7, 9]. Let $q = (x, \theta)$. The initial and final time constraints are: $q(0) = q_0$ and $q(t_f) = q_f$ where t_0 is fixed and t_f is free. The constraint on the piecewise continuous input function $u(\cdot)$ arising from a constraint on the minimum turn-radius for the MAV is:

$$|u(t)| \leq u_{max} = \frac{V}{R_{min}}, \tag{4.2}$$

for all t. Here R_{min} is the minimum turn radius *in the absence of wind* and arises due to mechanical limitations on the aircraft. Even for such a simple model the question of existence of time-optimal trajectories is unknown in the presence of wind. In the absence of wind (that is $W(x,t) = (0,0)$), the well-known Dubins' theorem [8] posits the existence of a time-optimal solution for any initial and final positions and velocities of the aircraft on a plane, when the aircraft is flying with constant speed and has a minimum turn radius constraint. For a non-zero, time-varying wind vector field satisfying certain technical conditions, we provide easily checked sufficient conditions under which a time-optimal solution exists. The verification of these conditions can then be used as the starting point for a numerical algorithm to compute the time-optimal trajectory. In the case of wind that is only a function of time (and independent of the space variable) we show that the solution is unique except for initial and final states taking values on a set of measure zero. For more general wind vector fields, the question of uniqueness is still to be investigated. To prove the existence of a time optimal solution, we do not use perturbation techniques around the zero wind condition for which the solution is known to exist. Instead, we use Filippov's theorem on the existence of a solution in conjunction with Dubin's result for zero wind.

We are given the initial and final positions and orientations for the MAV, and the problem is to find the minimum time path connecting the initial and final states. As the speed of the aircraft increases due to a tail wind, one would expect

the instantaneous minimum turn radius to increase as well. It is easy to check that if $\|\dot{x}\| > V$ for some (x,t) then $\|\dot{x}\| = R'_{min}(x,t)$ max $|\dot{\theta}| = R'_{min}(x,t)\, u_{max}$, along with $V = R_{min}\, u_{max}$ implies $R'_{min}(x,t) > R_{min}$, as required.

As is well known, the minimum time trajectory planning problem can be cast as an optimal control problem for the sake of numerical solution [13]. Direct and indirect methods are usually employed to solve the optimal control problem [13]. Such methods assume the existence of the optimal solution and use gradient-based techniques to find the solution. For the minimum-time problem for the no wind case, we can show using Dubins' theorem [8] that the length of the trajectory (which is proportional to the minimum time) is a discontinuous function of the final state when the initial state is held fixed. To be specific, if the initial position and velocity is fixed, then the length of the minimum time path is a discontinuous function of the final position and velocity. Recall, that the magnitude of the velocity is fixed and only the direction is a variable. We show that the discontinuities are of the first-kind - that is, they are simple jump discontinuities. This implies that numerical methods must be carefully initialized for convergence. In the next section, we study the aspects of the solution for the no wind case paying careful attention to two issues: nonuniqueness of solutions and discontinuity of the solution. For the special case of constant wind vector field, we show using coordinate transformations that the qualitative nature of the solutions is the same as the zero wind case, and hence we can expect both discontinuous and non-unique solutions.

4.2 Discussion of Dubins' Theorem for the Zero Wind Case

Dubins' theorem [8] establishes the existence of a solution to the minimum time optimal control problem for the special case $W(x,t) = (0,0)$ for all $x \in \mathbb{R}^2$ and $t \in \mathbb{R}_+$. This theorem states that for every initial, final positions and velocities the minimum time solution is an arc-line-arc or arc-arc-arc solution. As the minimum time solution is invariant with respect to translations and rotations of the coordinate axis, we can change coordinates so that the initial position is at the origin of $\mathbb{R}^2$ and the final position is at $(l,0)$ on the ordinate axis. The initial and final velocity directions measured with respect to this axis are termed ϕ_0 and ϕ_f respectively in Figures 4.1 - 4.3. To understand the behavior of the minimum time solution as a function of the initial and final states, we considered - without loss of generality - the initial state to be fixed, and varied the final states. As the final state comprises of the final position and the final direction for the velocity vector, we consider each change in turn.

The direction of the velocity induces an orientation on the circles tangent to the velocity vector. In all the figures, we denote the center of the counterclockwise oriented circle by Z_0 and Z_f respectively, while the centers of the clockwise oriented curves are denoted by Y_0 and Y_f respectively. Thus we can distinguish between the Z_0LY_f from the Y_0LZ_f arc-line-arc solutions etc. This distinction is important in what follows. It turns out that for each ϕ_0 and ϕ_f one can compute

a critical separation l_c such that only arc-line-arc solutions can exist for $l > l_c$ (see Appendix 4.A). Below, we will fix the initial and final positions at points T and S respectively; fix the initial velocity direction ϕ_0; and vary the final velocity direction ϕ_f (see Figures 4.1 - 4.3). There are primarily three cases to consider:

1. $l = \|x_f - x_0\| > l_c$.
2. $0 < l \leq l_c$;
3. $l = 0$.

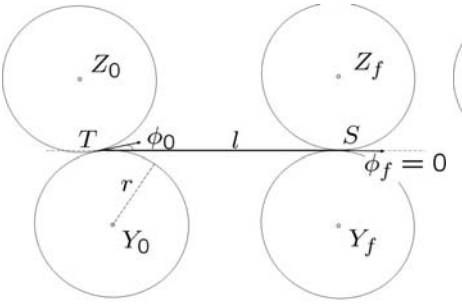
Case 1: Please refer to Figure 4.1. In this case, $l > l_c$; the circles Y_0, Z_0 and Y_f, Z_f do not intersect even at a single point; and hence the trajectories are always of the arc-line-arc type. We further observe that:

- the length of the minimum time solution is a continuously differentiable function of the angle ϕ_f;
- the solution changes from Y_0LZ_f to Z_0LY_f as the angle goes through specific angle ϕ_f^*. The exact value of ϕ_f^* is not as important as its property – it is the angle for which two minimum time solutions exist (see Figure 4.1c). As the solution changes from Y_0LZ_f to Z_0LY_f for $\phi_f > \phi_f^*$, its length changes in a continuous manner as a function of ϕ_f.
- As ϕ_f increases beyond ϕ_f^* the solution is of the type Z_0LY_f (see Figure 4.1d) until $\phi_f = \phi_c$ when we have a LY_f solution (see Figure 4.1e).
- For $\phi_c < \phi_f < 2\pi$, the solution is of the type Y_0LY_f.

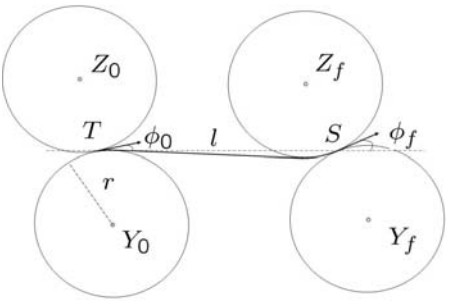
To re-emphasize, when $l > l_c$, the minimum length remains continuous function of ϕ_f.

Case 2: Please refer to Figure 4.2. In this case, $0 < l \leq l_c$. Hence, intersections of the circles Y_0 or Z_0 with either Y_f or Z_f is possible. This leads to two phenomena not observed in the previous case:

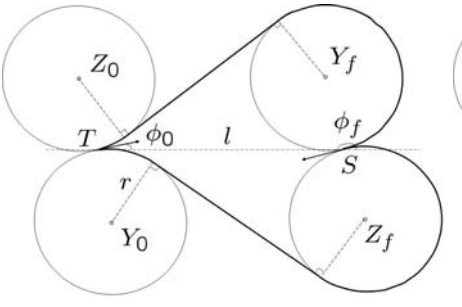
- A discontinuous change in the length of the minimum time trajectory at two critical angles (one of which is shown in Figure 4.2b and the other in Figure 4.2h). At the angle $\phi_f = \bar{\phi}_c$ in Figure 4.2b), the circle Z_f touches Y_0 and hence the Y_0LZ_f solution has a line section of zero length. As ϕ_f increases from $\bar{\phi}_c$, the solution switches to a Z_0LY_f solution shown in Figure 4.2c.
- Appearance of arc-arc-arc solutions for ϕ_f satisfying: $\phi_f \in [\tilde{\phi}_c, \hat{\phi}_c]$. The lengths of the solutions, vary continuously as the arc-arc-arc solutions appear or disappear.
- Just as in Case 1, for $\phi_f = \phi_f^*$ we see the appearance of two arc-arc-arc solutions of equal length as shown in Figure 4.2f. The lengths of the solution changes in a continuous manner as a function of ϕ_f.
- As mentioned in the first item of this case, when $\phi_f = \bar{\phi}_c'$ in Figure 4.2h), the circle Y_f touches Z_0 and hence the Z_0LY_f solution has a line section of zero length. When the angle ϕ_f is decreased from this value, then the length of the minimum time solution changes discontinuously.



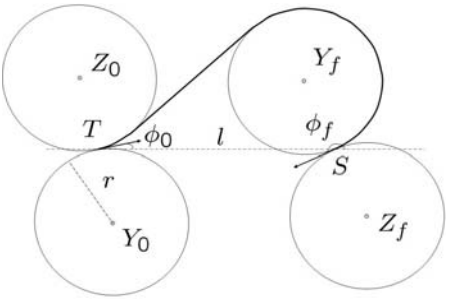
(a) A Y_0L solution for $\phi_f = 0$.



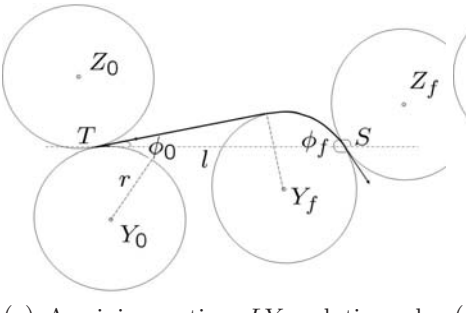
(b) A Y_0LZ_f solution for $0 < \phi_f < \phi_f^*$.



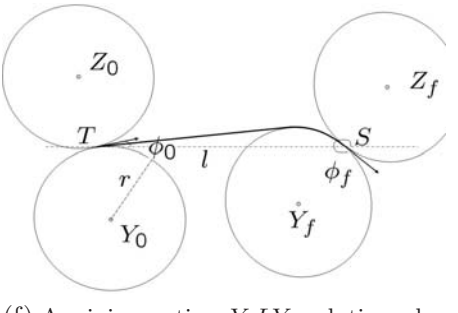
(c) Two minimum time Y_0LZ_f and Z_0LY_f solutions when $\phi_f = \phi_f^*$.



(d) A Z_0LY_f solution for $\phi_f^* < \phi_f < \phi_c$.



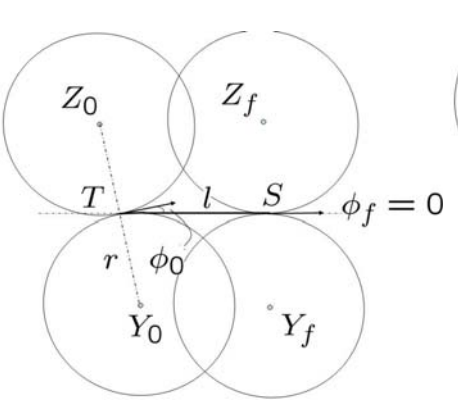
(e) A minimum time LY_f solution when $\phi_f = \phi_c$.



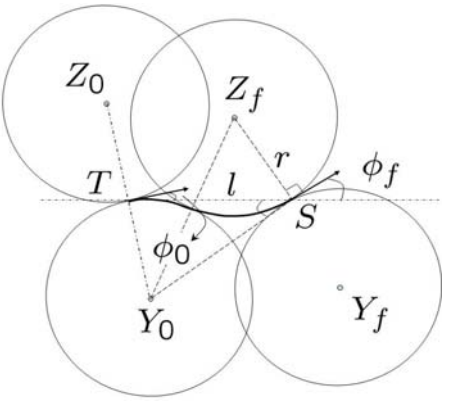
(f) A minimum time Y_0LY_f solution when $\phi_c < \phi_f < 2\pi$.

Fig. 4.1. Variation of the minimum time solution with the final angle ϕ_f for $l \geq r\,(3 + |\sin\phi_0|)$.

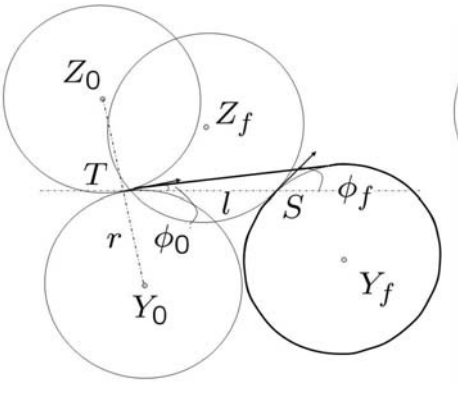
This discussion shows that the minimum time solution for Case 2 is unique and its length is a continuously differentiable function of ϕ_f for fixed l and ϕ_0 is fixed, except for at most four values.



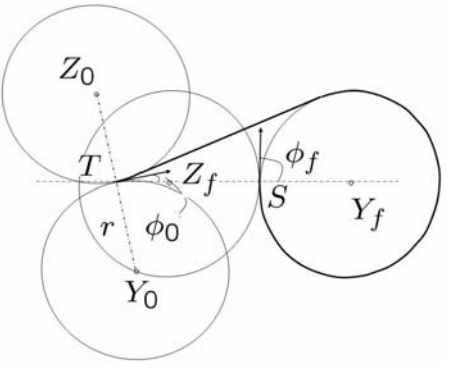
(a) A Y_0L solution for $\phi_f = 0$.



(b) A Y_0Z_f solution for $\phi_f = \bar{\phi}_c$.



(c) A switch to a Z_0LY_f solution for $\phi_f > \bar{\phi}_c$.

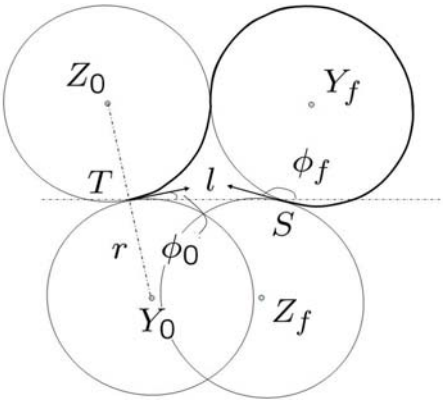


(d) A continuation of the previous case.

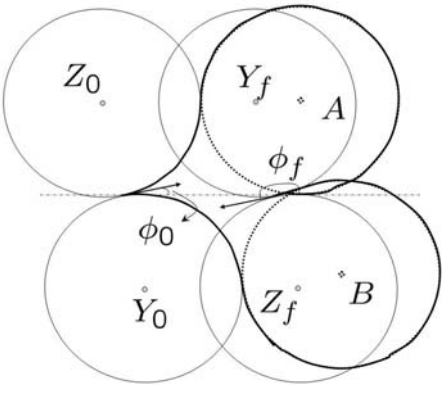
Fig. 4.2. Variation of the minimum time solution with ϕ_f for $0 < l < r(3 + |\sin\phi_0|)$.
Continued in next page

Case 3: Please refer to Figure 4.3. In this case, $l = 0$. As the initial and final positions coincide, there is no angle ϕ_f for which an arc-line-arc solution with a non-zero line segment is possible. The solutions for all angles ϕ_f can be considered to be arc-arc-arc solutions. The solutions are unique except for the case $\phi_f = \phi_0 + \pi$.

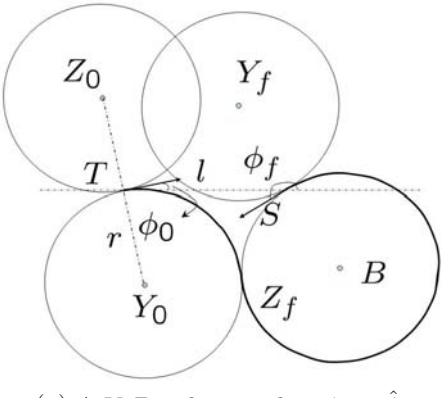
Instead of varying the final angle ϕ_f while holding l fixed, we can vary l with ϕ_f fixed. We can for example reduce l from a large value greater than the critical separation l_c. Then the same arguments presented above still apply. The crux of the matter is that the minimum time solution for the zero wind case is *unique* and is a continuously differentiable function of ϕ_f and l when ϕ_0 is fixed, except for at most four points. These features persist for the important case of non-zero,



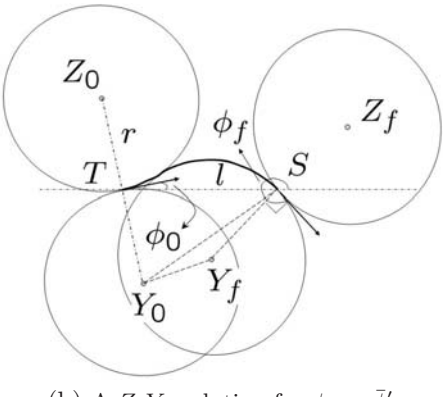
(e) A Z_0Y_f solution when $\phi_f = \tilde{\phi}_c$.



(f) Two arc-arc-arc Z_0AZ_f and Y_0BY_f solutions when $\phi_f = \phi_f^*$.



(g) A Y_0Z_f solution when $\phi_f = \hat{\phi}_c$.



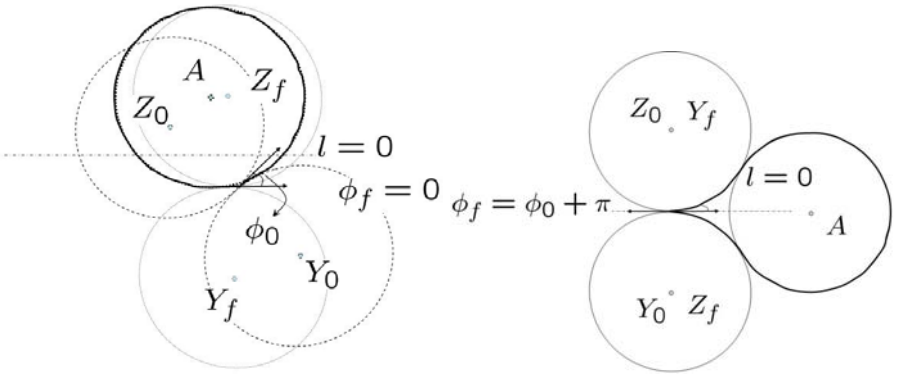
(h) A Z_0Y_f solution for $\phi_f = \bar{\phi}_c'$.

Fig. 4.2. (*continued*)

time-varying wind vector field that is constant in the spatial coordinate, as will be seen later.

4.3 Existence of Minimum Time Solution

Next, we will show that there exists a solution to the minimum time optimal control problem even when the wind is non-constant, time-varying and satisfying two constraints. First, it is clear that there must be a bound on the wind speed so that the aircraft is able to fly upwind. This assumption is needed for *reachability*, that is, for a solution to exist. We will also need a bound on the rate of change of the wind due to the bound on u Specifically, we will assume that W is a Lipschitz continuous function of t and x (which implies it is differentiable almost everywhere by Rademacher's Theorem [11]). Furthermore, suppose that:



(a) An arc-arc-arc Y_0AY_f solution for $l = 0$, $|\phi_f - \phi_0| < \pi$.

(b) Two arc-arc-arc Z_0AZ_f and Y_0AY_f solutions for $l = 0$, $\phi_f = \phi_0 + \pi$.

Fig. 4.3. Variation of the minimum time solution with the final angle ϕ_f for $l = 0$.

Assumption 4.3.1. $\|W\|_\infty = \sup_{x,t} \|W(x,t)\| < V$.

Assumption 4.3.2. *A-2.* $\|\frac{\partial W}{\partial t}\|_\infty + \|\frac{\partial W}{\partial x}\|_\infty (V + \|W\|_\infty) \leq \beta \sqrt{V^2 - \|W\|_\infty^2}\, u_{max}$, *where:* $\|\frac{\partial W}{\partial t}\|_\infty = \sup_{x,t} \|\frac{\partial W}{\partial t}(x,t)\|_2$ *and* $\|\frac{\partial W}{\partial x}\|_\infty = \sup_{x,t} \|\frac{\partial W}{\partial x}(x,t)\|_2$, *and* $0 < \beta < 1$ *is some constant.*

Theorem 4.3.1. *Suppose that $W(x,t)$ is a time-varying (Carathéodory) vector field of wind velocities satisfying assumptions 4.3.1 - 4.3.2. Then there exists a solution to the minimum time optimal control problem for the system (4.1).*

Proof. The theorem is proved using Filippov's Theorem [14] on minimum time optimal control. A special case of this theorem applicable to our problem is given in the Appendix as Theorem 4.B.1. This theorem requires that the set $Q = [-u_{max}, u_{max}]$ be convex which is true. Due to our assumptions on $W(x,t)$, the function $f(q,t,u) = [V\cos\theta + W_1(x,t)\ V\sin\theta + W_2(x,t)\ u]^T$ is differentiable a.e. with an essentially bounded derivative in $q = (x,\theta)$ and continuous in t. The requirement that $q^T f(q,t,u) \leq C(\|q\|^2 + 1)$ for some $C > 0$ is true by the following:

$$\begin{aligned} q^T f(q,t,u) &= V\,\langle x, (\cos\theta, \sin\theta)\rangle + \langle x, W(x,t)\rangle + \theta\, u \\ &\leq V\|x\|_2 + \|W(x,t)\|_2\,\|x\|_2 + |\theta|\, u_{max} \\ &\leq 2\,V\|x\|_2 + |\theta|\, u_{max} \\ &\leq \max\{2V, u_{max}\}\ (\|x\|_2 + |\theta|) \\ &\leq \sqrt{2}\,\max\{2V, u_{max}\}\,\sqrt{\|x\|_2^2 + |\theta|^2} \\ &\leq \sqrt{2}\,\max\{2V, u_{max}\}\ (\|q\|^2 + 1). \end{aligned}$$

As required by Filippov's theorem, we will demonstrate that one solution exists to the trajectory planning problem by modifying a Dubins' solution for the zero-wind case.

Setting the wind vector field to be identically zero, let $r(\tau) = (z(\tau), \theta(\tau))$ be the Dubins' solution (modulo identification) with minimum turn-radius $R \geq R_{min}$ where

$$R = \frac{R_{min}\ (\|W\|_\infty + V)^2}{(1-\beta)\, V^2}. \tag{4.3}$$

We will show later that this value of R allows the modified trajectory to satisfy the constraint $|u(t)| \leq u_{max}$. The variable τ used to denote the Dubins' solution is proportional to the arc-length along this solution with proportionality constant V. Hence, the function $r(\cdot)$ satisfies $r(0) = q_0$ and $r(\tau_f) = q_f$ for some τ_f. The initial condition $\theta_{\text{no wind}}(0)$ for the zero-wind case has to be chosen different from $\theta(0)$ because at the initial time the aircraft is only capable of flying in the direction of $Z = W(x(0), 0) + V\,(\cos\theta(0), \sin\theta(0))$ and not $V\,(\cos\theta(0), \sin\theta(0))$. The angle $\theta_{\text{no wind}}(0)$ is chosen according to: $Z = \|Z\|\,(\cos\theta_{\text{no wind}}(0), \sin\theta_{\text{no wind}}(0))$. Using this solution, we will obtain a solution to the trajectory planning problem when $W(x,t) \neq (0,0)$.

Let T be the unit vector tangent and N be the unit normal vector to the Dubins' solution at a generic point $r(\cdot)$. In other words, $T = \frac{1}{\|\frac{dz}{d\tau}\|}\frac{dz}{d\tau}$. Select N so that it is the *outward pointing* normal at the point $z(\tau)$ (see Figure 4.4). By Dubins' theorem [8], T is a piecewise continuous function.

We will construct a solution that will traverse the same points $r(\cdot)$ at times t, that is, $q(t) = r(\tau)$ for some t and τ. Consider the following differential equation for the time variable t:

$$\frac{dt}{d\tau} = \frac{V}{\sqrt{V^2 - \|W(z(\tau), t(\tau)\|^2 + W_\parallel^2(\tau)} + W_\parallel(\tau)}; \quad t(0) = 0, \tag{4.4}$$

where

$$W_\parallel(\tau) \triangleq W(z(\tau), t(\tau)) \cdot T(z(\tau)).$$

The differential equation (4.4) has a unique solution on $[0, \tau_f]$ because the denominator on the right hand side is strictly greater than zero by Assumption 4.3.1 and hence is Lipschitz continuous almost everywhere as a function of τ. As the right hand side of (4.4) is strictly greater than zero, t is a monotone increasing, Lipschitz function of τ. Denote $t_f \triangleq t(\tau_f)$. By the previous discussion, for every $\hat{t} \in [0, t_f]$, there exists a $\tau \in [0, \tau_f]$ such that $\hat{t} = t(\tau)$. By a slight abuse of notation, we will denote both $\hat{t}$ and t by t in the following.
For any $t \in [0, t_f]$, denote:

$$\begin{aligned}
\bar{V}(t) &\triangleq V(\cos\theta(t(\tau)), \sin\theta(t(\tau))),\\
\bar{V}_\parallel(t) &\triangleq \bar{V}(t) \cdot T(z(\tau)),\\
\bar{V}_\perp(t) &\triangleq \bar{V}(t) \cdot N(z(\tau))\\
W_\perp(t) &\triangleq W(z(\tau), t(\tau)) \cdot N(z(\tau)).
\end{aligned}$$

To obtain the control law for one solution to the trajectory planning problem, set $\theta(t)$ to be such that:

$$\bar{V}_\perp(t) = -W_\perp(t). \tag{4.5}$$

As there is a constraint on $\dot{\theta}$ (see (4.2)), it has to be shown that such a choice of $\theta(t)$ can be made without violating the constraint. This is where the assumption 4.3.2 is needed.

Define $\xi(t) \triangleq \bar{V}_\perp(t) + W_\perp(t)$. At $t = 0$, we can select $\theta(0)$ so that $\xi(0) = 0$ due to the assumption 4.3.1. Thereafter, we need to show that one can choose $u(t)$ so that $\frac{d\xi}{dt} = 0$. This will imply $\xi(t) = 0$ or (4.5) is true. Now, we have the following basic identities for $T(t)$ and $N(t)$:

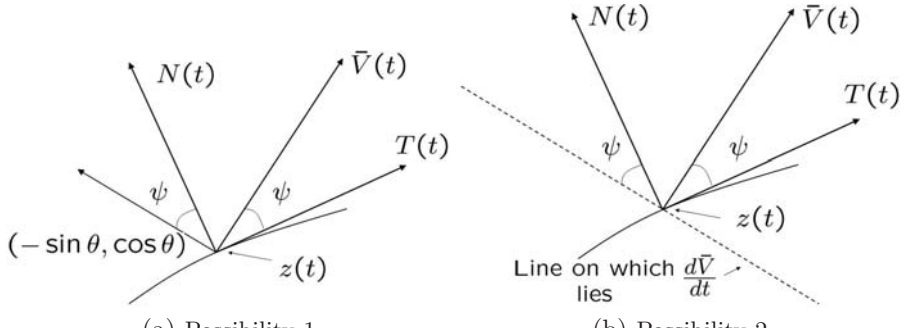
$$\begin{aligned} T(t) \cdot T(t) = 1 &\Longrightarrow T(t) \cdot \frac{dT}{dt} = 0 \\ N(t) \cdot N(t) = 1 &\Longrightarrow N(t) \cdot \frac{dN}{dt} = 0 \\ T(t) \cdot N(t) = 0 &\Longrightarrow T(t) \cdot \frac{dN}{dt} + \frac{dT}{dt} \cdot N(t) = 0 \end{aligned}$$

From the above equations and noting the planar nature of the curves, we get:

$$\frac{dT}{dt} = -\omega\, N \tag{4.6}$$

$$\frac{dN}{dt} = \omega\, T \tag{4.7}$$

where ω is either 0, $\frac{\bar{V}_\parallel + W_\parallel}{R}$ or $-\frac{\bar{V}_\parallel + W_\parallel}{R}$ because $\bar{V}_\parallel + W_\parallel$ is the tangential component of the velocity (that is $\dot{x} \cdot T$) along the Dubins solution. At the end of the proof, it will become clear why ω must take these values. For example, in the case depicted in Figure 4.4, we have $\omega = \frac{\bar{V}_\parallel + W_\parallel}{R}$ because $\frac{dT}{dt}$ is directed in the opposite direction to N.



(a) Possibility 1. (b) Possibility 2.

Fig. 4.4. Relation between $\bar{V}$, N, and T.

Therefore:

$$\begin{aligned}\frac{d\xi}{dt} &= \frac{dW}{dt}\cdot N + W\cdot\frac{dN}{dt} + \frac{d\bar{V}}{dt}\cdot N + \bar{V}\cdot\frac{dN}{dt}\\ &= \omega(W+\bar{V})\cdot T + \frac{dW}{dt}\cdot N + \frac{d\bar{V}}{dt}\cdot N\\ &= \omega(W_{\|}+\bar{V}_{\|}) + \frac{dW}{dt}\cdot N + \frac{d\bar{V}}{dt}\cdot N \end{aligned} \tag{4.8}$$

We now observe that:

$$\frac{d\bar{V}}{dt} = V\left(-\sin\theta,\cos\theta\right)\dot{\theta} = V\left(-\sin\theta,\cos\theta\right)u$$

Let $\mu(t) = \text{sign}\left(\left(-\sin\theta,\cos\theta\right)\cdot N\right)$. Observe that $\mu(\cdot)$ is a piecewise function of t because N is a piecewise continuous function of t. We pick $u(t)$ to be:

$$u(t) = -\mu\frac{1}{\bar{V}_{\|}}\left(\omega\left(W_{\|}+\bar{V}_{\|}\right) + \frac{dW}{dt}\cdot N\right). \tag{4.9}$$

We show that the choice of $u(t)$ leads to $\frac{d\xi}{dt} = 0$. First, observe that (see Figure 4.4):

$$\left|V\left(-\sin\theta,\cos\theta\right)\cdot N\right| = \bar{V}\cdot T = \bar{V}_{\|} \tag{4.10}$$

Substituting (4.9) and (4.10) into (4.8) we get:

$$\begin{aligned}\frac{d\xi}{dt} &= \omega(W_{\|}+\bar{V}_{\|}) + \frac{dW}{dt}\cdot N + V\,u\left(-\sin\theta,\cos\theta\right)\cdot N\\ &= \omega(W_{\|}+\bar{V}_{\|}) + \frac{dW}{dt}\cdot N - \left(\omega\left(W_{\|}+\bar{V}_{\|}\right) + \frac{dW}{dt}\cdot N\right)\\ &= 0\end{aligned}$$

All of the quantities on the right hand side of (4.9) are known at time t. More importantly, $u(\cdot)$ is a measurable function of t because each of the functions on the right hand side is measurable, and the denominator is always greater than zero by Assumption 4.3.1 and $\xi(t) = 0$.

Next, we show that $|u(t)| \le u_{max}$ for R chosen as in (4.3). For this, first observe that the minimum value of $V_{\|}$ is achieved when $W(x,t)$ is directed orthgonal to T and furthermore $|W(x,t)| = \|W\|_{\infty}$. In this case, $\bar{V}_{\|} = \sqrt{V^2 - \|W\|_{\infty}^2}$.

$$\begin{aligned}|u(t)| &= \frac{1}{|\bar{V}_{\|}|}\left|\omega(W_{\|}+\bar{V}_{\|}) + \frac{dW}{dt}\cdot N\right|\\ &\le \frac{(W_{\|}+\bar{V}_{\|})^2}{R\,|\bar{V}_{\|}|} + \frac{\|\frac{dW}{dt}\|\,\|N\|}{\sqrt{V^2-\|W\|_{\infty}^2}}\\ &= \frac{(W_{\|}+\bar{V}_{\|})^2}{R\,|\bar{V}_{\|}|} + \frac{\|\frac{\partial W}{\partial t} + \frac{\partial W}{\partial x}(\bar{V}+W)\|}{\sqrt{V^2-\|W\|_{\infty}^2}}\\ &\le \frac{(W_{\|}+\bar{V}_{\|})^2}{R\,|\bar{V}_{\|}|} + \beta\,u_{max},\end{aligned}$$

where the last inequality is due to Assumption 4.3.2. It is proved in the Appendix 4.C that:

$$\max_{V_\perp = -W_\perp} \frac{(W_\parallel + \bar{V}_\parallel)^2}{|\bar{V}_\parallel|} = \frac{(\|W\|_\infty + V)^2}{V}.$$

This yields:

$$|u(t)| \leq \frac{(\|W\|_\infty + V)^2}{R\,V} + \beta\, u_{max}.$$

As $R = \frac{1}{u_{max}\,(1-\beta)} \frac{(\|W\|_\infty + V)^2}{V} = \frac{R_{min}\,(\|W\|_\infty + V)^2}{(1-\beta)\,V^2}$, we have $|u(t)| \leq u_{max}$.

Now that existence of one solution with a measurable control has been shown, let us see what form $\dot{x}$ takes for this solution:

$$\begin{aligned}
\dot{x} &= \bar{V}(t) + W(\phi(t), t) \\
&= \bar{V}_\parallel(t)\, T + \bar{V}_\perp(t)\, N + W_\parallel(t)\, T + W_\perp(t)\, N \\
&= (\bar{V}_\parallel(t) + W_\parallel(t))\, T. \qquad (4.11)
\end{aligned}$$

Using the time variable τ for the Dubins' solution, we have:

$$\frac{dx}{d\tau} = V\, T(\tau).$$

Comparing the last two equations, we must have

$$\frac{dt}{d\tau} = \frac{V}{\bar{V}_\parallel + W_\parallel}.$$

Using a similar argument that was used to obtain (4.6 - 4.7), we get for the Dubins' solution:

$$\frac{dT}{d\tau} = -\kappa\, N \qquad (4.12)$$

$$\frac{dN}{d\tau} = \kappa\, T \qquad (4.13)$$

where κ is either 0, $\frac{V}{R}$ or $-\frac{V}{R}$, because the Dubins' solution either has lines or curves with constant turn radius R. From Equation (4.4) and (4.12), we obtain:

$$\frac{dT}{dt} = \frac{dT}{d\tau}\,\frac{d\tau}{dt} = -\kappa\,\frac{\bar{V}_\parallel + W_\parallel}{V}\,N$$

which is consistent with Equation (4.6) and the constant ω that was introduced there.

This shows the existence of a solution when the wind speed is not zero.

Now consider the set of pairs all solutions and the final times: $S = \{(q(\cdot), t_f)\,|\,q(0) = q_0;\, q(t_f) = q_f;\, t_f \in \mathbb{R}_+\}$. We know that this set is non-empty. The elements of this set can be ordered according to the final times t_f. There must exist a minimal element $(q^*(\cdot), t_f^*)$ by a special case of Filippov's Theorem [14](See Appendix). ■

Remark 4.3.1. Notice that when the wind is constant with speed W, then one can choose $\beta = 0$ in 4.3.2. Then for the Dubins' solution for the existence part, $R = \frac{1}{u_{max}} \frac{(W+V)^2}{\sqrt{V^2-W^2}}$. In particular, when the constant wind has zero speed, we have $R = \frac{1}{u_{max}} V = R_{min}$.

We present a simple numerical example to illustrate the method used to show existence of one solution. Consider a MAV with $V = 50$, minimum turn radius $R_{min} = 35$ and constant wind $W(x,t) = 25\,(1,1)$. Then according to the remarks after Theorem 4.3.1, we can choose $R = R = R_{min} \frac{(W+V)^2}{V^2} = 102$. The Dubins solution with no wind and $R = 102$ is shown in Figure 4.5a. The path is based on an Arc-Arc-Arc solution from coordinates $(35, 76, 0.5)$ to $(24, 85, \frac{\pi}{4})$. The step-size used was 0.1 and this path took 158 steps to travel. The next two figures are different scenarios of a constant wind. The dark arrow symbolizes the wind direction. In each figure, the spacing is wider where the wind seems to be helping the trajectory of the MAV.

Figure 4.5b represents $W = (25, 25)$, which is wind directed toward the first quadrant. With the same step-size, travel on this path took 546 steps.

4.3.1 Uniqueness of the Solution for a Special Wind Vector Field

Suppose the wind $W(x,t) = W(t)$ depends only on the time variable, while still satisfying the conditions (4.3.1) - (4.3.2). Then, we can the transform coordinates (with the same time variable) as follows:

$$\bar{x} = \varphi(x,t) = x - \int_0^t W(s)\,ds \quad \text{and} \quad \bar{\theta} = \theta. \tag{4.14}$$

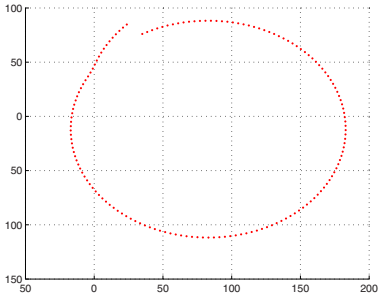
In the new coordinates $(\bar{x}, \bar{\theta})$, the equations take on the form:

$$\dot{\bar{x}} = V(\cos\bar{\theta}, \sin\bar{\theta}); \quad \dot{\bar{\theta}} = u. \tag{4.15}$$

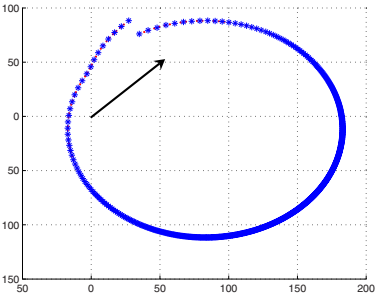
which is of the form considered by Dubins. We had seen earlier that except for $\phi_f = \phi_0 + \pi$ (where we have employed the notation in Figures 4.1, 4.2, 4.3), the minimum time solution is unique. Even for the case $\phi_f = \phi_0 + \pi$, it is natural to identify the two solutions, leaving us with unique minimum time solutions for all combinations of initial, final positions and velocity directions. Therefore, in the coordinates $(\bar{x}, \bar{\theta})$, there exists a unique minimum time solution for all initial, final states. The inverse transform is obviously:

$$x = \vartheta(\bar{x}, t) = \bar{x} + \int_0^t W(s)\,ds \quad \text{and} \quad \theta = \bar{\theta}. \tag{4.16}$$

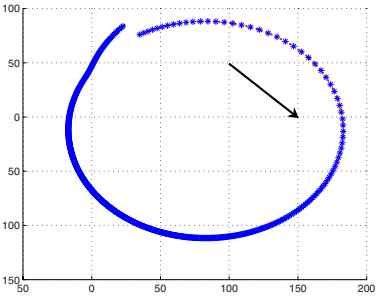
There exists a minimum time solution by Theorem 4.3.1 for the original problem with final states (x_f, θ_f). Let us denote this solution by $q^*(t) = (x^*(t), \theta^*(t))$ with minimum time t_f^*. In the new coordinates, the final states are $(\bar{x}_f, \bar{\theta}_f)) = (x_f - \int_0^{t_f^*} W(s)\,ds, \theta_f)$. There exists a unique Dubins' solution in the new coordinates with final time t_f. If $t_f \neq t_f^*$, then on inverse transformation, the point



(a) Dubins Solution.



(b) Wind in 1st Quadrant.



(c) Wind in 2nd Quadrant.

Fig. 4.5. Illustration of the construction of a solution for non-zero wind.

$(\bar{x}_f, \bar{\theta}_f))$ transforms to $(\bar{x} + \int_0^{t_f} W(s)\, ds, \theta_f) \neq (x_f, \theta_f)$ which is a contradiction. Hence $t_f = t_f^*$, and the minimum time solutions in the initial and transformed coordinates are simply transformations of each other. As the solution in the transformed coordinates is unique, the solution in the original coordinates are also unique.

4.4 Conclusion

In this paper, we have shown that a minimum time solution for the trajectory planning problem for a micro air vehicle with a minimum turn radius constraint in the presence of a non-zero, time-varying wind vector field exists. For a non-zero, time-varying wind vector field, we provide easily checked sufficient conditions under which a time-optimal solution exists. We also shown uniqueness for almost every initial and final conditions for the case of a wind vector field that varies with time, but is constant in the spatial variable for each time instant. These results are of critical importance in proving convergence of numerical algorithms for trajectory planning [10].

Appendices

4.A Some Results on the Dubins Minimum Time Solution

The goal of this section is to study arc-arc solutions that occur when the initial and final positions are sufficiently separated. Specifically, we show that for each value of initial and final velocity vectors, there exists a minimum "separation" – denoted by l_c – between the initial and final positions beyond which there can only exist arc-line-arc solutions. Using the methods of this section it is possible to compute this minimum separation (see Figure 4.8). This is useful in the numerical implementation of Dubins algorithm, as we can reject arc-arc-arc solutions apriori by simply checking the given separation against the critical one.

Choose coordinates so that the initial position is at the origin and the final position is on the positive x-axis. We hold ϕ_0 as fixed and vary the final angle ϕ_f and position $(l, 0)$. For a given ϕ_0 value there exists two circles tangent to the initial velocity (denoted by A and B in Figure 4.7). A similar situation exists for the final position. In the following, we will consider angles ϕ_0 and ϕ_f modulo π radians. It can be seen in Figure 4.7 that rotation by π radians does not change the critical separation beyond which only arc-line-arc solutions exist. For a given ϕ_f value (with ϕ_0 already fixed), there exists a certain critical separation l_c that depends only on ϕ_0 and ϕ_f for which one of the circles tangent to the final velocity vector is tangent to one of the circles tangent to the initial velocity vector, *without any of the other circles intersecting each other.* For example, when $\phi_f \leq \phi_0$ the circles B and D can be made tangent (figure 4.7a) by systematically reducing l from a large value. Similarly, when $\phi_f > \phi_0 > \frac{\pi}{6}$ the circles B and C can be made tangent (figure 4.7b), by reducing l from a large value. It is clear that for values of l greater than the critical value l_c there only exist arc-line-arc solutions as one needs the tangent circles to intersect for an arc-arc-arc solution to exist (Dubins [8]).

The different possibilities when $\phi_0 \geq 0$ are shown in Figure 4.6 (see also Figure 4.7). Case 1 corresponds to combinations of ϕ_0 and ϕ_f such that circles B and D are tangential. Case 2 corresponds to combinations of ϕ_0 and ϕ_f such that circles B and C are tangential. Cases 3, 4 corresponds to combinations of ϕ_0, ϕ_f such that circles A and C are tangential. The difference between the two is in the computation of the critical separation. To be precise, angle $\alpha \leq \phi_f - \phi_0$ for Case 3, while $\phi_f - \phi_0 \leq \alpha$ for Case 4. For each of these cases, the computation of the critical separation has to be done using a different method. These are presented next.

Firstly, consider Case 1 that corresponds to $\phi_f \leq \phi_0$ (see figure 4.7a). In this case, the circles B and D are tangential. In ΔTSD, let $\angle STD = \beta$, $l(TD) = x$ and $l(TS) = l$. It is easy to see that $\angle TSD = \frac{\pi}{2} + \phi_f$. Applying the sine-rule to ΔTSD we get:

$$\frac{\cos(\phi_f)}{x} = \frac{\sin(\beta)}{r} = \frac{\cos(\phi_f + \beta)}{l}.$$

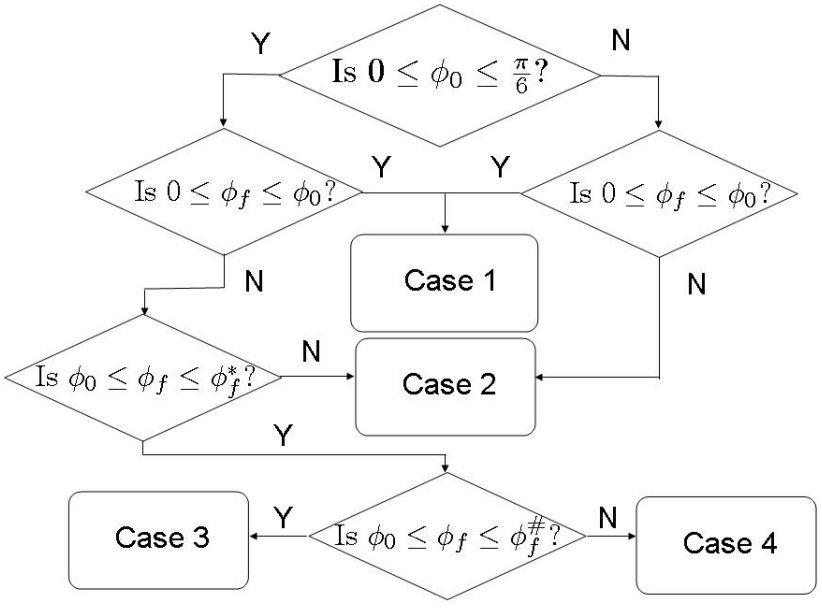


Fig. 4.6. Flow chart for the case $0 \leq \phi_0$.

In ΔTSD, $\angle BTD = \frac{\pi}{2} - \phi_0 - \beta$. Applying the cosine rule:

$$4r^2 = r^2 + x^2 - 2rx\,\sin(\phi_0 + \beta).$$

We thus have two equations in two variables:

$$x\sin(\beta) - r\cos(\phi_f) = 0$$
$$x^2 - 2rx\,\sin(\phi_0 + \beta) - 3r^2 = 0.$$

We can employ an invertible coordinate change: $(x, \beta) \mapsto (y, z)$ given by $x\sin(\beta) = y$ and $x\cos(\beta) = z$. Then: $y = r\cos(\phi_f)$ and

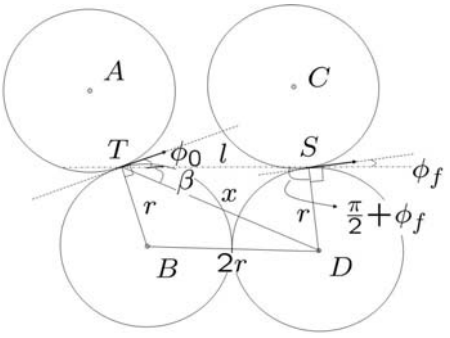
$$z^2 - 2rz\sin(\phi_0) - 2ry\cos(\phi_0) + y^2 - 3r^2 = 0.$$

The latter equation is of the type: $z^2 + bz + c = 0$ with $b,\ c < 0$ which implies that there exist a unique positive and real solution z. This yields a unique solution for the critical separation l_c.

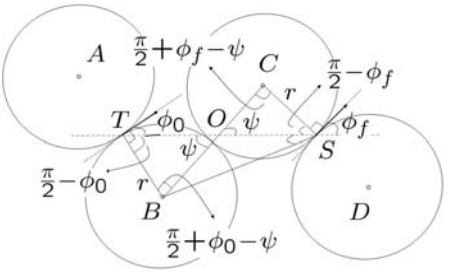
Secondly, consider Case 2 which is easiest to handle (figure 4.7b). In ΔOCS, $\angle CSO = \frac{\pi}{2} - \phi_f$. Let $\angle COS = \psi$ so that $\angle OCS = \frac{\pi}{2} + \phi_f - \psi$. Let $l(OC) = x_1$ and $l(OS) = l_1$. Applying the sine rule for triangles we get:

$$\frac{\sin(\psi)}{r} = \frac{\cos(\phi_f)}{x_1}.$$

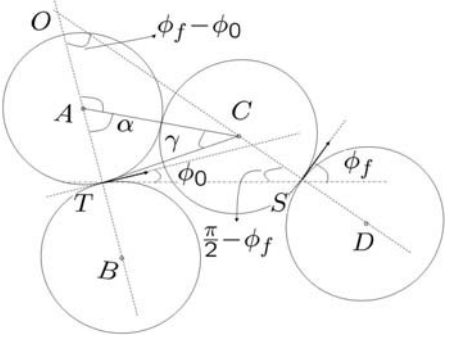
Turning to ΔOBT, $\angle BTO = \frac{\pi}{2} - \phi_0$. Let $\angle BOT = \psi$ so that $\angle OBT = \frac{\pi}{2} + \phi_0 - \psi$. Let $l(OB) = x_2$ and $l(OS) = l_2$. As $x_1 + x_2 = 2r$, we get: $x_1 = \frac{2r\cos(\phi_f)}{\cos(\phi_0) + \cos(\phi_f)}$



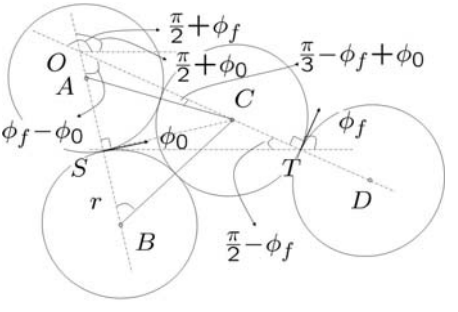
(a) Case 1: Circles B, D are tangential.



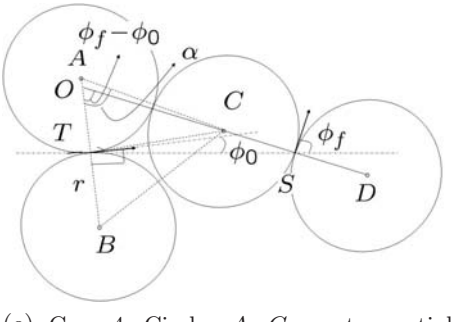
(b) Case 2: Circles B, C are tangential.



(c) Case 3: Circles A, C are tangential with $\phi_f - \phi_0 \leq \alpha$.



(d) Transition from Case 3 to Case 2 at $\phi_f = \phi_f^*$: Going from A and C tangential to B and C tangential.



(e) Case 4: Circles A, C are tangential with $\alpha \leq \phi_f - \phi_0$.

Fig. 4.7. Study of arc-arc Dubins solutions.

and $x_2 = \frac{2r\cos(\phi_0)}{\cos(\phi_0)+\cos(\phi_f)}$. We can further solve for l_1 and l_2 and obtain $l = l_1 + l_2$ from:

$$l_1^2 = x_1^2 + r^2 + 2x_1 r\sin(\phi_f - \psi),$$
$$l_2^2 = x_2^2 + r^2 + 2x_2 r\sin(\phi_0 - \psi).$$

Thirdly, we consider Case 3, which consists of $0 \le \phi_0 \le \frac{\pi}{6}$ and $\phi_0 \le \phi_f \le \phi_f^{\#}$ (see Figure 4.7c). The angle $\phi_f^{\#}$ can be understood as follows. When $\phi_0 > 0$ and $\phi_f = 0$, the circles B and D become tangential first when l is reduced from a large value. As ϕ_f is increased from 0, it is found that when $\phi_f = \phi_0$, the circles A – C and B – D become tangential simultaneously with $l_c = 2r$, if and only if $0 \le \phi_0 \le \frac{\pi}{6}$. When $\phi_f = \phi_0 = \frac{\pi}{6}$, we have A,B and C tangent simultaneously and at the same time B, C and D are tangent simultaneously. If $\phi_0 > \frac{\pi}{6}$, it is not possible to make circles A and C tangent to each other without the other circles intersecting, for any angle ϕ_f[4] In the other case of $0 \le \phi_0 \le \frac{\pi}{6}$, there exists an angle ϕ_f^* for which circles A, B and C become tangential (see Figure 4.7d). Values of $\phi_f < \phi_f^*$ are further sub-divided into Cases 3 and 4 that depend on the relative position of the centers of the circles A, C and D. There exists an angle $\phi_f^{\#} \le \phi_f^*$ such that the centers of A, C and D collinear. We call the case $\phi_f \le \phi_f^{\#}$ as Case 3 and $\phi_f^{\#} \le \phi_f \le \phi_f^*$ as Case 4 (Figure 4.7e).

In ΔOAC:

$$\frac{\sin(\phi_f - \phi_0)}{2r} = \frac{\sin\alpha}{l(OC)} = \frac{\sin(\alpha - \phi_f + \phi_0)}{l(OA)}$$

In ΔATC:

$$\frac{\sin(\alpha+\gamma)}{2r} = \frac{\sin(\alpha)}{l(TC)} = \frac{\sin(\gamma)}{r}.$$

In ΔOTC:

$$\frac{\sin(\phi_f - \phi_0)}{l(TC)} = \frac{\sin(\alpha - \phi_f + \phi_0 + \gamma)}{r + l(OA)}.$$

To find $l(TC)$ which is the critical separation, denote by $l_1 = l(OA)$): Also denote $z = [\alpha,\ \gamma,\ l_1]^T$. Then we have the equations:

$$f(z) = \begin{bmatrix} l_1\sin(\phi_f - \phi_0) - 2r\sin(\alpha + \phi_0 - \phi_f) \\ \sin(\alpha+\gamma) - 2\sin(\gamma) \\ 2r\sin(\alpha+\gamma-\phi_f+\phi_0)\ \sin(\alpha) - (r+l_1)\ \sin(\alpha+\gamma)\ \sin(\phi_f - \phi_0) \end{bmatrix} = 0$$

We can solve for z using Newton's method and then compute $l(TC)$. A similar approach is used to compute $l(TC)$ in Case 4. The results of the computation can be found in Figure 4.8.

[4] To be precise, if $l > 2r$ and $\phi_f = \phi_0 > \frac{\pi}{6}$, then B and C become tangential first leading to Case 2.

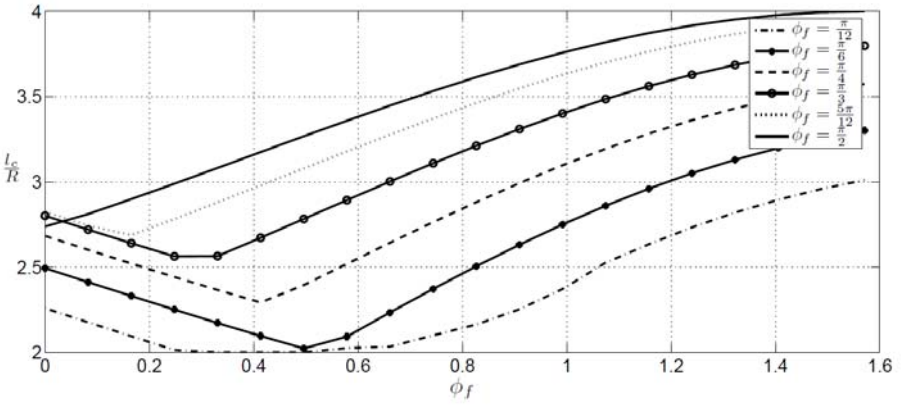


Fig. 4.8. Critical length to Radius ratios.

4.B Basic Results on Optimal Control

The following is a specialization of Filippov's theorem [14] on the existence of minimum time optimal control for the system

$$\dot{q} = f(t, q, u) \tag{4.17}$$

where q and f are n dimensional vectors, u is r dimensional control parameter, which for every t and u takes values in a fixed *convex* set $U \subset \mathbb{R}^n$. The vector function $f(t, q, u)$ is continuous in all variables; differentiable a.e. with respect to q; and $q^T f(t, q, u) \leq C(\|q\|^2 + 1)$ for some $C > 0$, for all t and q, and all $u \in U$. Filippov's theorem [14] asks for continuous differentiability of $f(t, q, u)$ in the q variable, but a study of the proof shows that only differentiability a.e. is needed. By requiring $f(t, q, u)$ to be Lipschitz continuous in x is enough for Filippov's theorem to be true.

By the continuity of f in the u variable, the set $R(t, q) = \{f(t, q, u) \,|\, u \in Q\}$ is a convex set for each t and q. For the system (4.17) satisfying the above conditions, consider the problem:

$$\min T \text{ such that } q(0) = q_0 \text{ and } q(T) = q^*. \tag{4.18}$$

Theorem 4.B.1. *[14] Suppose that the conditions above are satisfied, and that there exists at least one measurable function $\tilde{u}$ with $\tilde{u}(t) \in U$ such that the solution $\bar{q}(t)$ with $u = \bar{u}(t)$ and initial condition $\bar{q}(0) = q_0$ attains q^* for some $t^* > 0$. Then there also exists an optimal control for Problem (4.18), i.e., a measurable function $u(\cdot)$ with $u(t) \in U$.*

4.C Maximum Value of $\frac{(W_\parallel+\bar{V}_\parallel)^2}{|\bar{V}_\parallel|}$

Here, we will prove that:

$$\max_{V_\perp=-W_\perp} \frac{(W_\parallel+\bar{V}_\parallel)^2}{|\bar{V}_\parallel|} = \frac{(\|W\|_\infty+V)^2}{V}.$$

Denote: $x=\bar{V}_\parallel$ and $y=W_\parallel$. Then:

$$f(x)=\frac{(W_\parallel+\bar{V}_\parallel)^2}{|\bar{V}_\parallel|}=\frac{(y+x)^2}{x}.$$

As $W_\parallel=\sqrt{W^2-V^2+\bar{V}_\parallel^2}$ due to the condition $\bar{V}_\perp=-W_\perp$, we have:

$$f(x)=\frac{\|W\|_\infty^2-V^2}{x}+2\sqrt{\|W\|_\infty^2-V^2+x^2}+2x.$$

The critical points where f takes its maximum or minimum values are points where $f'(x)=0$ and the boundary points $x=V$, $x=\sqrt{V^2-\|W\|_\infty^2}$.

Differentiating f with respect to x we get:

$$f'(x)=\frac{V^2-\|W\|_\infty^2}{x^2}+\frac{2x}{\sqrt{\|W\|_\infty^2-V^2+x^2}}+2.$$

For simplicity, denote $z=\frac{V^2-\|W\|_\infty^2}{x^2}$. It is clear from Assumption 4.3.1 that $z>0$. Then:

$$F(z)=f'(x)=z+\frac{2}{\sqrt{1-z}}+2.$$

Setting $F(z)=0$ we get the two points: $z=0$ or $z=-3$. As neither of these points lie in the domain of z, we check the value of $f(x)$ at the boundary points: $x_1=V$ and $x_2=\sqrt{V^2-\|W\|_\infty^2}$. At x_1, $y=\sqrt{\|W\|_\infty^2-V^2+x^2}=\pm\|W\|_\infty$; while at x_2, $y=0$. Comparing the values of $f(x_1)$ and $f(x_2)$, we find $f(x_1)>f(x_2)$. Therefore:

$$\max_{V_\perp=-W_\perp} \frac{(W_\parallel+\bar{V}_\parallel)^2}{|\bar{V}_\parallel|} = \frac{(\|W\|_\infty+V)^2}{V}.$$

Acknowledgement. R. Iyer was supported by ASEE/AFOSR Summer Faculty Fellowship and R. McNeely was supported by an AFRL Graduate Student Fellowship during the summer of 2005.

References

1. Chandler, P., Pachter, M.: Hierachical control for autonomous teams. In: AIAA-2001-4149, Proc. AIAA Guidance, Navigation and Control Conference, Montreal, Canada (August 2001)

2. Schumacher, C., Chandler, P., Rasmussen, S.: Task allocation for wide area search munitions via network flow optimization. In: AIAA-2001-4147, Proc. AIAA Guidance, Navigation and Control Conference, Montreal, Canada (August 2001)
3. Rasmussen, S., Mitchell, J.W., Chandler, P., Schumacher, C., Smith, A.L.: Introduction to the MultiUAV2 simulation and its application to cooperative control research, FrB16.1. In: Proc. American Control Conference, Portland, OR (June 2005)
4. Chaudhry, A.I., Misovec, K.M., D'Andrea, R.: Low Observability Path Planning for an Unmanned Air Vehicle Using Mixed Integer Linear Programming. In: Proc. of the 43rd IEEE Conference on Decision and Control, Paradise Island, The Bahamas (December 2004)
5. Yang, G., Kapila, V.: Optimal path planning for unmanned air vehicles with kinematic and tactical constraints. In: WeA04-6, Proc. 41st IEEE Conf. on Decision and Control, Las Vegas, NV, December 2002, pp. 1301–1306 (2002)
6. Howlett, J.: Path Planning for Sensing Multiple Targets from an Aircraft, Masters Thesis, Dept. of Mechanical Engineering, Brigham Young University (December 2002)
7. Anderson, E.: Extremal Control and Unmanned Air Vehicle Trajectory Generation, Masters Thesis, Department of Electrical and Computer Engineering, Brigham Young University (April 2002)
8. Dubins, L.E.: On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents. American Journal of Mathematics 79, 497–516 (1954)
9. Savla, K., Bullo, F., Frazzoli, E.: On traveling salesperson problems for Dubins' vehicle: Stochastic and dynamic environments. In: Proc. IEEE Conf. on Decision and Control, Seville, Spain, December 2005, pp. 4530–4535 (2005)
10. McNeely, R., Iyer, R., Chandler, P.: Tour planning for an unmanned air vehicle. Journal of Guidance, Control and Dynamics 30(5), 1299–1306 (2007)
11. Clarke, F.H., Yu, S., Stern, R.J., Wolenski, R.R.: Nonsmooth Analysis and Control Theory. Springer, Heidelberg (1998)
12. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: The Mathematical Theory of Optimal Processes. John Wiley & Sons, Inc., New York (1962); Authorized translation from the Russian by K. N. Trirogoff
13. Betts, J.T.: Survey of numerical methods for trajectory optimization. Journal of Guidance, Control and Dynamics 21(2), 193–207 (1998)
14. Filippov, A.F.: On certain questions in the theory of optimal control. SIAM Journal of Control, Ser. A 1(1), 76–84 (1962)

5

A Precise Formulation and Solution of the Drag Racer and Hot Rodder Problems

Kevin R. Kefauver[1] and William S. Levine[2]

[1] Dept. of ME, University of Maryland, College Park, MD 20742, USA
[2] Dept. of ECE, University of Maryland, College Park, MD 20742, USA

Summary. The purpose of this paper is to solve two simplified optimal traction control problems, the drag racer problem and the hot-rodder problem. The control problems are defined and the optimal solutions are given. The solution to each problem is not bang-bang, but includes a singular control.

5.1 Introduction

The drag racer problem is to drive a straight 1/4 mile strip in minimum time from a standing start. Drag racing is a popular sport in the United States with many variants and vehicle classes with serious money at stake [18]. The best racer in 2005, who won the "World Championship," traversed the quarter mile drag strip in 4.496 seconds with a speed of 324.36 mph at the finish. The leading money winner received $1,141,500 for his efforts during the year.

It is intuitively clear that the optimal control for the hot rodder is not bang-bang. If it were, there would be no need for ABS brakes. This also supplies the intuition for the drag racer. The limit on acceleration is not the engine; it is the friction interaction between the wheels and the road. A drag racer that spins his wheels during the race will lose to one that does not.

5.2 Background

The purpose of this section is to introduce the vehicle model used for the analysis, introduce the standard friction models encountered in the literature, and provide a review of the literature related to this problem

5.2.1 Vehicle Model

For longitudinal vehicle motion analysis, the automobile is typically represented by a simple two wheel bicycle model [4], where a torque is applied to one or both tires producing longitudinal force that accelerates or decelerates the vehicle. The model used in this paper will assume that only the rear tire produces the

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 79–92.
springerlink.com

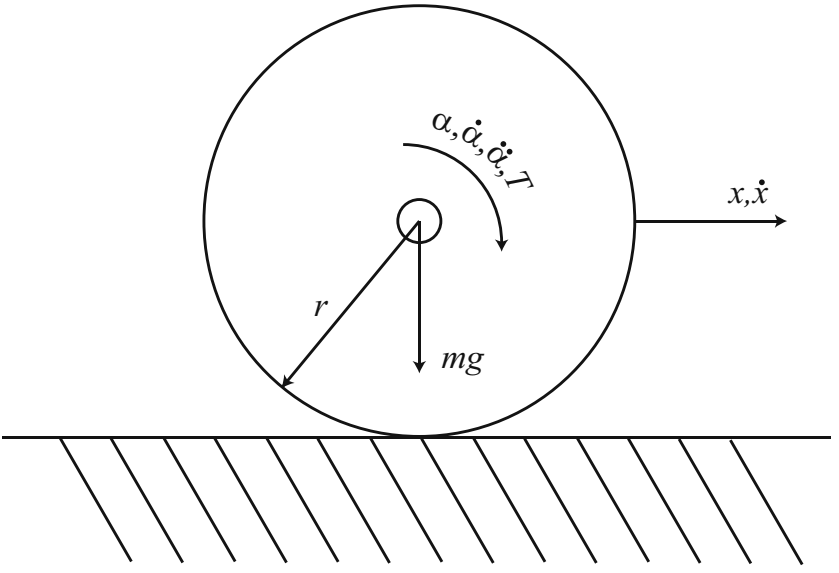


Fig. 5.1. Single wheel longitudinal acceleration model

longitudinal force acting on the vehicle and that the weight transfer associated with the acceleration process can be neglected. Making these assumptions allows us to reduce the longitudinal bicycle model to a quarter vehicle model with a single tire interacting with the ground and a lumped mass representing the quarter vehicle mass acting at the center-of-gravity of the wheel. This type of longitudinal vehicle model is common in the literature and can be found in [8, 9, 13, 14]. It is also a good model for studying the drag racer.

Figure 5.1 is a schematic representing the single wheel longitudinal acceleration model.

Using Figure 5.1, the dynamics describing the acceleration of the wheel can be written as

$$m\ddot{x}(t) = \mu(S(t))F_N \tag{5.1}$$

$$I\ddot{\alpha}(t) = -r\mu(S(t))F_N + T(t) \tag{5.2}$$

where $\dot{x}$ is the longitudinal velocity of the center of the wheel, $\dot{\alpha}$ is the angular velocity of the wheel, T is the torque applied to the wheel, r is the effective radius of the tire at the tire-to-ground interface point, F_N is the normal force acting on the tire, m is the quarter vehicle mass, I is the rotating moment of inertia of the tire, and $\mu(S(t))$ is the coefficient of friction with respect to the slip where the slip S is defined as

$$S(t) = \begin{cases} \frac{r\dot{\alpha}(t)-\dot{x}(t)}{\dot{x}(t)}, & \text{Braking} \\ \frac{r\dot{\alpha}(t)-\dot{x}(t)}{r\dot{\alpha}(t)}, & \text{Accelerating} \end{cases} \tag{5.3}$$

Note that the term slip does not mean that the tire tread is slipping relative to the ground. When a torque is applied to the tire, the tread elements around the tire contact patch deform, but do not immediately move horizontally relative to the ground. Because of the deformation, the effective circumference of the tire changes and the angular velocity of the wheel must change to maintain the no horizontal movement condition at the tire-to-ground interface. As torque is applied to the wheel the slip increases until the tire-to-ground interface can no longer counteract the torque. Since the ground does not counteract the control torque, the tire contact patch has a different longitudinal velocity than the vehicle and the wheel angular velocity either increases or decreases rapidly depending on the direction that the torque is applied. Soon thereafter the wheel spins if accelerating and locks if braking.

5.2.2 Friction Models

The most common tire-to-ground friction model found in the literature is a steady-state model based on the work in [6]. The steady-state friction model assumes that the coefficient of friction between the tire and ground is a function of the slip.

Figure 5.2 is a sample coefficient of friction versus slip curve similar to the ones found in [6], and has some distinct features: the coefficient of friction is a continuous function of the slip, attains its maximum at some slip, S_{max}, where $0 <\mid S_{max} \mid< 1$, and enters the pure sliding mode when $\mid S(t) \mid= 1$. Further, when braking the slip is negative and the coefficient of friction is negative and when accelerating the slip is positive and the coefficient of friction is positive.

Note that it is usually assumed that this type of friction model is only valid for vehicle and wheel motions in the positive direction ($\dot{x} \geq 0$).

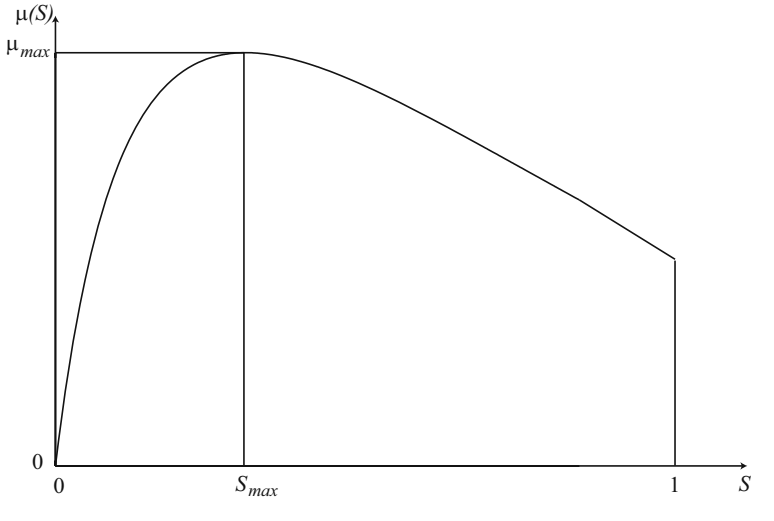


Fig. 5.2. Sample friction versus slip curve for positive slip

The steady-state friction model presented is a good model to perform analysis with because it is relatively simple. Unfortunately, the force generation process is a dynamic, not a static event, so the steady-state friction model does not account for dynamic events. In [11] a dynamic friction model was introduced that is based on the point LuGre friction model. This model uses a first order non-linear filter with the relative velocity between the tire and ground as its input and the friction force as its output. Let $\nu_r = r\dot{\alpha} - \dot{x}$, then the friction force is

$$F_r = (\sigma_0 z + \sigma_1 \dot{z} + \sigma_3 \nu_r) F_n \tag{5.4}$$

where z is the solution of the non-linear differential equation

$$\dot{z} = \nu_r - \sigma_0 \frac{\mid \nu_r \mid}{\mu_c + (\mu_s - \mu_c) \mathrm{e}^{-|\frac{\nu_r}{\nu_s}|^{1/2}}} \tag{5.5}$$

where σ_0, σ_1, σ_2, μ_c, and μ_s are constants relating to the fundamental properties of the tire and ν_s is the Striebeck constant. With properly tuned parameters, this model reproduces the friction versus slip curve of Figure 5.2 in steady-state, and provides good correlation with experimental data. For a detailed derivation of this model see [11].

The main disadvantages of this model are that accurate results require proper tuning of the constant parameters and it is much more complicated than the steady-state friction versus slip curve. Since the slip curve is not known a priori, is not really measurable, and not really correct, it is reasonable to use a simpler model for the problem analysis. We use a friction model that is simpler than both the steady-state friction model and the dynamic friction model. Since the magnitude of I is much less than the magnitude of m in Equations (5.1) and (5.2) the torque applied to the wheel will act to accelerate and decelerate the wheel much faster than it will change the longitudinal velocity of the vehicle. As such, the transition from one slip value to another will happen very quickly, which can be seen in the example worked in [9]. Because of this behavior, we model the system as a hybrid system using the standard stick/slip friction model, where the friction coefficient is μ_{max} when the wheel is sticking to the ground and $\mu_s < \mu_{max}$ when the wheel is sliding with respect to the ground.

Figure 5.3 uses a standard elementary definition of friction where the coefficient of friction is constant when no slip occurs and switches instantaneously to another constant value when slipping begins. This amounts to a conceptually simple hybrid system model for the vehicle with three discrete modes: no slipping, wheel accelerating and slipping, and wheel decelerating and slipping. This model is technically more difficult to optimize because of the non-differentiable dynamics on the transition curves between modes. One justification for the model is that it is easy to solve. Another is that the more detailed smoother model is neither an accurate description of the dynamics nor measurable in real time, which is highlighted by the actual ABS algorithms implemented on vehicles which only use wheel angular acceleration as the feedback variable [7].

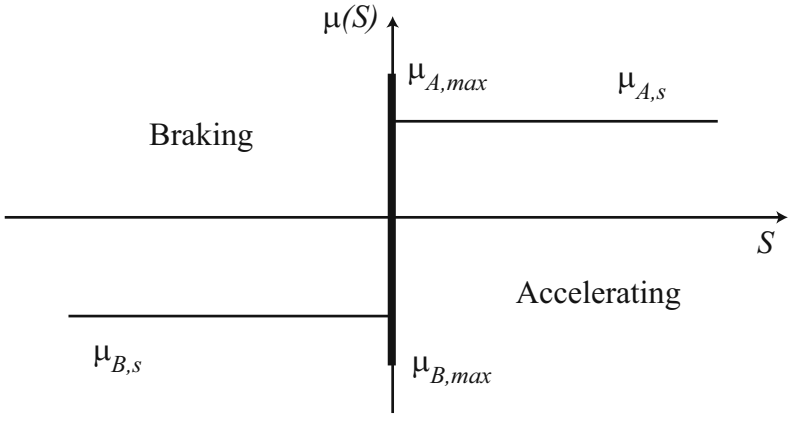


Fig. 5.3. Friction versus slip curve for the optimal control problems

5.2.3 Optimal Traction Control

Many papers exist in the literature that provide algorithms for optimal traction control for the models described in the previous section. The Anti-lock Braking System (ABS) problem was the first problem addressed in the literature. Since acceleration control is a similar problem, the literature began addressing traction control in general, where traction control consists of braking and acceleration control. [5, 8, 10, 12, 13, 14] present various control algorithms, based on the steady-state friction versus slip curve, that try to maximize the coefficient of friction between the tire and the ground. [12] uses sliding mode control techniques to solve the ABS problem. Using sliding mode techniques is common, but the novelty of the algorithm in [12] is that they assume the friction versus slip curve is unknown. [10] introduces an optimal fuzzy logic control algorithm to solve the ABS problem. The fuzzy algorithm is based on the steady-state friction versus slip curve where the fuzzy parameters are optimized using a genetic algorithm technique to maximize the coefficient of friction. [13] introduces a hybrid ABS control technique based on an estimation of the friction versus slip curve. This technique is unique because it exploits the fact that the friction versus slip curve is continuous and has a single maximum. [14] also utilizes the steady-state friction versus slip curve to develop a hybrid traction controller. They decompose the continuous problem into a four node hybrid automaton, where the dynamics of each mode are defined based on the value of the time derivative of the slip and the switching rule is defined by the value of the slip. Controllers are then developed for each node of the automaton. The resulting control algorithm uses the slip to determine the appropriate controller to maximize the coefficient of friction between the tire and ground.

Note that these problems and techniques are not specific to vehicles with rubber-to-ground interfaces. [16] applies fuzzy logic methods to optimize traction control for a train system based on the steady-state friction model and a metal-to-metal interface.

The basic underlying assumption for these algorithms is that the optimal solution of the traction control problem is to maximize this coefficient of friction. In [9] the authors apply Pontryagin's Maximum Principle (PMP) [1, 2, 3] to the steady-state ABS problem to prove optimality of the control. There are two concerns with using the PMP. First, the PMP only provides necessary conditions for optimality. This can be problematic because a control can satisfy the necessary conditions and not be optimal. Second, the PMP only provides an open loop control. For applications, the feedback solution is preferred. Because of these two limitations of the PMP, it is desirable to augment the results of the necessary conditions.

5.2.4 Paper Outline

The purpose of this paper is to provide a precise formulation of the drag racing and hot-rodder problems, to provide feedback solutions to both of them, and to prove these solutions are optimal. These optimal control problems are slightly different from the one given in [9] but the results of this work can be directly applied to solve the problem given in [9]. The rest of the paper will be organized in the following way. First the two optimal control problems will be defined. Then a candidate for the optimal feedback control will be proposed. The corresponding cost-to-go function is not smooth.

5.3 Problem Formulation

The purpose of this section of the paper is to formulate the optimal control problems. The drag racing problem and the hot-rodder problem are similar and both optimal control problems will be given here. The drag racing problem will be developed first, then the hot-rodder problem will be given, and finally some comments about both control problems will be presented. Figure 5.3 depicts the friction model that will be used in solving the drag racing and hot-rodder optimal control problems. Furthermore, the motions for both problems will be constrained to velocities in the positive longitudinal directions.

5.3.1 Simplified Dynamics

Since we are using a simplified friction model, the dynamics governing the longitudinal motion of the vehicle are slightly different from the dynamics given in Equations (5.1) and (5.2). Assume that as long as $T(t) \leq T_{i,s}$ the wheel doesn't slip. Then the dynamics are

$$\ddot{x}(t) = r\ddot{\alpha}(t) \tag{5.6}$$

$$\ddot{\alpha}(t) = \frac{T(t)}{(I + mr^2)} \tag{5.7}$$

where

$$T_{i,s} = \frac{F_n \mu_{i,max}(mr^2 + I)}{mr} \tag{5.8}$$

and $\mu_{i,max}$ is given in Figure 5.3 for $i = A$ for accelerating and $i = B$ for braking.

When the wheel slips, the dynamics are given by

$$\ddot{x}(t) = \frac{F_n}{m}\mu_{i,s} \tag{5.9}$$

$$\ddot{\alpha}(t) = -\frac{rF_n}{I}\mu_{i,s} + \frac{T(t)}{I} \tag{5.10}$$

where

$$\mu_{i,s} = \begin{cases} \mu_{A,s}, & \dot{x} < r\dot{\alpha} \\ \mu_{B,s}, & \dot{x} > r\dot{\alpha} \end{cases} \tag{5.11}$$

and $T(t)$ is the torque applied to the wheel.

5.3.2 Drag Racing Problem

Drag racing is a popular motor sport where the goal of the race is to traverse a quarter mile straight section of track in the shortest time. Two cars race at a time and the car that passes the finish line first wins the race. The drag racing problem can be simplified and studied using optimal control theory. Assume that the race car has only one gear and the dynamics given in Equations (5.6) and (5.7) and Equations (5.9) and (5.10) represent the motion of the vehicle over the time interval, $0 \leq t \leq t_f$. Further assume that the race car engine and brakes can instantaneously produce any torque in the set $[T_{B,max}, T_{A,max}]$ where $T_{A,max}$ is a real positive number representing the maximum engine acceleration torque and $T_{B,max}$ is a real negative number representing the maximum braking torque. Further, assume that at initial time t_0, the initial conditions for the race car are $x(t_0) = 0$, $\dot{x}(t_0) = 0$, $\alpha(t_0) = 0$, and $\dot{\alpha}(t_0) = 0$. The problem ends when $x(t) = 0.25$ miles, which is the terminal condition. Letting t_f represent the time at which $x(t) = 0.25$ miles, the performance criterion is

$$J(T(t)) = \int_{t_0}^{t_f} 1dt. \tag{5.12}$$

The problem is to find a torque, $T(t)$, $t \in [t_0, t_f]$, that minimizes Equation (5.12) subject to the constraints of Equations (5.6) and (5.7) and Equations (5.9) and (5.10) and the boundary conditions. In fact, we solve the more general problem: Given any initial state, find the feedback control that minimizes Equation (5.12) subject to the given dynamics and constraints.

The solution to the drag racing problem is the feedback control defined over the interval of time, $t_0 \leq t \leq t_f$, that satisfies the following equation

$$T(\dot{x}, x, \dot{\alpha}) = \begin{cases} T_{A,max}, & \dot{x} > r\dot{\alpha} \\ T_{A,s}, & \dot{x} = r\dot{\alpha} \\ T_{B,max}, & \dot{x} < r\dot{\alpha} \end{cases} \tag{5.13}$$

where $T_{A,s}$ is given in Equation (5.8). Note that the torques in Equation (5.13) are constant and have the following properties

$$T_{A,max} \geq T_{A,s} \tag{5.14}$$

$$T_{B,max} \leq T_{B,s} \tag{5.15}$$

and Equation (5.13) depends only on x, $\dot{x}$, and $\dot{\alpha}$.

The hybrid maximum principle [17] can be used to show that Equation (5.13) is a candidate open loop control for the drag racing problem with fixed initial condition. Note that there is a region of the state space where the optimal control is not unique. When the wheel is spinning and the drag racer is "close enough" to the target distance, d, any control will be optimal. In this region, there is not enough time to move to the non-sliding state before the target distance is reached. Meanwhile, the torque applied to the wheel does not affect the velocity of the drag racer and hence does not affect the performance. While in this region any torque applied to the wheel is optimal, so our solution still applies.

5.3.3 Hot-Rodder Problem

The hot-rodder problem is very similar to the drag racing problem. Hot rodding is an artificial problem that is more complicated than the drag racing problem. We define hot rodding as the problem of traveling a fixed distance from a standing start to a dead stop in minimum time. Again assume that the car has one gear and the dynamics given in Equations (5.6) and (5.7) and Equations (5.9) and (5.10) define the vehicle's motion over the interval of time, $0 \leq t \leq t_f$. Further, assume that the engine and brake can instantaneously produce the torque, $T(t) \in [T_{B,max}, T_{A,max}]$ and assume the initial conditions for the state variables are $x(t_0) = 0$, $\dot{x}(t_0) = 0$, $\alpha(t_0) = 0$, and $\dot{\alpha}(t_0) = 0$.

Let d represent the fixed distance to be traveled and assume that at time t_f, the state variables have the end conditions $x(t_f) = d$, $\dot{x}(t_f) = 0$, $\dot{\alpha}(t_f) = 0$, and $\alpha(t_f)$ is free. Further, let the cost associated with starting from the initial condition to the final state be, as in the drag-racer problem

$$J(T(t)) = \int_{t_0}^{t_f} 1dt. \tag{5.16}$$

The problem is then to find a torque, $T(t)$, $t \in [t_0, t_f]$, that minimizes Equation (5.16) subject to the constraints of Equations (5.6) and (5.7) and Equations (5.9) and (5.10) and the boundary conditions. Note that the only difference between the drag-racer problem and the hot-rodder problem is the boundary conditions. In particular, the hot rodder has to stop at the end of the course while the drag racer can continue past the end point.

The solution to the hot-rodder problem is more complicated than the solution to the drag racing problem. Because this problem is a stop-go-stop problem, there is an acceleration phase followed by a braking phase.

Let time t_b, $0 < t_b < t_f$, be the time at which the acceleration phase ends and the braking phase begins. Then for the interval of time $0 \leq t < t_b$ the solution is a constant control defined by

$$T(\dot{x}, x, \dot{\alpha}) = \begin{cases} T_{A,max}, & \dot{x} > r\dot{\alpha} \\ T_{A,s}, & \dot{x} = r\dot{\alpha} \\ T_{B,max}, & \dot{x} < r\dot{\alpha} \end{cases} \tag{5.17}$$

and for the interval of time $t_b \leq t \leq t_f$ the solution is a constant control that satisfies

$$T(\dot{x}, x, \dot{\alpha}) = \begin{cases} T_{A,max}, & \dot{x} > r\dot{\alpha} \\ T_{B,s}, & \dot{x} = r\dot{\alpha} \\ T_{B,max}, & \dot{x} < r\dot{\alpha} \end{cases} \tag{5.18}$$

where $T_{A,s}$ and $T_{B,s}$ satisfy Equation (5.8) and the torques given in Equation (5.17) and Equation (5.18) satisfy Equations (5.14) and (5.15).

Since Equations (5.17) and (5.18) give an optimal feedback control, given initial conditions for the hot-rodder problem, it is straightforward to compute the time, t_b, that the vehicle begins the braking phase of the trajectory by integrating Equations (5.6) and (5.7) and applying the boundary conditions and solving for t_b. For example

$$t_b = \sqrt{\frac{d}{\frac{rT_{A,s}}{2(I+mr^2)}(1 + \frac{T_{A,s}}{|T_{B,s}|})}} \tag{5.19}$$

when $x(0) = 0$, $\dot{x}(0) = 0$, and $\dot{\alpha}(0) = 0$.

As with the drag racing problem, a set of states in the state space exist where the solution is not well behaved. There exists a set of initial states where any control is optimal just like the drag racing problem. There will also exist a set of states where an optimal control does not exist because a trajectory does not exist that satisfies the boundary conditions.

5.3.4 Comments

The two problems given above are different from the pure ABS problem given in [9], because we are minimizing the time and they are minimizing the distance. The PMP can be used to solve the optimal control problem in [9], but it only provides an open-loop control. Further, since the problem has been formulated as a hybrid control problem, the standard PMP does not apply. Non-smooth versions of the maximum principle have been derived, for example see [17], but still only provide an open-loop solution to the control problem and necessary conditions for optimality.

The second thing to note is that the standard sufficiency conditions derived from the HJCB differential equation also do not apply to these problems because they require that the cost-to-go function be differentiable everywhere in the state space [1, 3]. Of course, the integral equation of dynamic programming does still apply. Because of the discontinuity associated with the model, the cost-to-go function is continuous, but not differentiable everywhere. It is straightforward to prove that the cost-to-go function for the ABS problem in [9] is also continuous but not differentiable everywhere. The rest of this section will be used to compute

the cost-to-go function for the drag racing problem as we have defined it and to show that it is continuous but not differentiable. In order to do this, we are first going to perform a coordinate transformation on the dynamics. Let $\underline{y} = [y_1\ y_2\ y_3]^T$, where

$$[y_1\ y_2\ y_3] = [x\ \dot{x}\ (r\dot{\alpha} - \dot{x})] \tag{5.20}$$

then when $\dot{x} = r\dot{\alpha}$

$$\begin{bmatrix} \dot{y_1} \\ \dot{y_2} \\ \dot{y_3} \end{bmatrix} = \begin{bmatrix} y_2 \\ \frac{rT(\underline{y})}{I+mr^2} \\ 0 \end{bmatrix} \tag{5.21}$$

and when $\dot{x} \neq r\dot{\alpha}$

$$\begin{bmatrix} \dot{y_1} \\ \dot{y_2} \\ \dot{y_3} \end{bmatrix} = \begin{bmatrix} y_2 \\ \frac{F_n}{m}\mu_{i,s} \\ \frac{rT(\underline{y})}{I} - F_n\mu_{i,s}(\frac{I+mr^2}{mI}) \end{bmatrix} \tag{5.22}$$

where $\mu_{i,s}$ is given in Equation (5.11).

These state equations have several useful features. The unneeded state, the wheel angle α, is eliminated. The dynamics are, in a sense, decoupled. The third state, y_3, is independent of the other two states. Furthermore, when $y_3 \neq 0$ the dynamics of the other two states depend only on the sign of y_3 and are constant. This means that the control has no immediate effect on y_1 and y_2 whenever $y_3 \neq 0$. Using Equation (5.12), the optimal cost-to-go to the final state from some initial state, $\underline{y}_0$, is defined as

$$J_c^*(\underline{y}) = \inf_T J(T(\underline{y})). \tag{5.23}$$

Now assume that the initial time is $t = 0$. Then the control can be chosen to be $T_{A,s}$. This will keep $y_3 = 0\ \forall t \in [0, t_f]$ and then the dynamics given in Equation (5.21) can be integrated using Equation (5.13) and solved for t_f yielding

$$J_c^*(\underline{y}) = \frac{-y_{2,0} + \sqrt{y_{2,0}^2 + 2\gamma_1(d - y_{1,0})}}{\gamma_1} \tag{5.24}$$

where $d = .25$ miles, $y_{1,0}$ and $y_{2,0}$ are the position and velocity of the wheel at the initial time, and

$$\gamma_1 = \frac{rT_{A,s}}{I + mr^2}. \tag{5.25}$$

We give the argument proving that this is optimal in the following section.

Now consider a starting state with $\dot{x} \leq \dot{r\alpha}\ (y_{3,0} > 0)$. The candidate optimal control is to decelerate the wheel until $y_3(t) = 0$ and then switch to the candidate optimal control for $y_3 = 0$. The cost-to-go for the initial state is

$$J_c^*(\underline{y}) = \frac{-y_{3,0}}{\gamma_2} - \frac{(y_{2,0} + \gamma_3 y_{3,0})}{\gamma_1} + \frac{\sqrt{(y_{2,0} + \gamma_3 y_{3,0})^2 + 2\gamma_1(d - (y_{1,0} - y_{2,0}\frac{y_{3,0}}{\gamma_2} + \gamma_4 y_{3,0}^2))}}{\gamma_1} \tag{5.26}$$

where

$$\gamma_2 = [\frac{rT_{B,max}}{I} - F_n\mu_{A,s}(\frac{I+mr^2}{mI})]$$
$$\gamma_3 = -\frac{F_n\mu_{A,s}}{m\gamma_2} \tag{5.27}$$
$$\gamma_4 = \frac{1}{2}\frac{F_n\mu_{A,s}}{m}(\frac{1}{\gamma_2})^2.$$

A similar derivation produces $J_c^*(\underline{y}_0)$ with $y_{3,0} < 0$, where the constant coefficients have different values ($T_{B,max}$ is replaced by $T_{A,max}$ and $\mu_{A,s}$ is replaced by $\mu_{B,s}$). It is straightforward but tedious to compute $\frac{\partial J_c^*(\underline{y}_0)}{\partial \underline{y}_0}$, and obvious that the result is discontinuous at $y_{3,0}$. Similar analysis can be applied to the hot-rodder problem to show that the optimal cost-to-go function is not differentiable along trajectories.

5.4 Proof

The purpose of this section is to prove that Equations (5.13), (5.17), and (5.18) are the optimal controls for the drag racer and hot-rodder problems. The fact that the right-hand sides of the dynamical equations (5.21) and (5.22) are discontinuous at $y_3 = 0$ (because $\dot{y_2}|_{y_3>0} = \frac{F_n}{m}\mu_{A,s}$ and $\dot{y_2}|_{y_3<0} = \frac{F_n}{m}\mu_{B,s}$) precludes the use of the viscosity solution theory for the HJCB equation developed in [1]. Because of the special form of the dynamics, a direct argument can be used.

5.4.1 Proof for the Drag Racer

Assume that at some initial time $y_3 = 0$. Equation (5.21) then governs the movement of the vehicle. The control, $T(\underline{y})$, can take any value on the interval $[T_{B,s}T_{A,s}]$. The choice of $T_{A,s}$ maximizes the acceleration regardless of y_1 and y_2. This is clearly optimal provided y_3 remains zero until $y_1(t) = d$. It is still possible that this is not the optimal control because there might be a control that drives y_3 away from zero resulting in overall better performance.

Assume that $y_3 > 0$ at some initial time. We want to compare the projection of the actual trajectory of the system onto the y_1 versus y_2 plane with the hypothetical trajectory that would result if y_3 were equal to zero. As long as $y_3 > 0$ the dynamics of y_1 and y_2 is constant with

$$\dot{y}_2(t) = \frac{F_n}{m}\mu_{A,s} < \frac{F_n}{m}\mu_{A,max}. \tag{5.28}$$

Because of the decoupling of $\dot{y}_3$, the control, $T(\underline{y})$, has no effect on the y_1 versus y_2 plane until $y_3 = 0$. We can choose any value for $T(\underline{y})$ from the set $[T_{B,max}T_{A,max}]$. It is clear from Equation (5.22) that, to minimize the time required to drive y_3 to zero, we should choose $T(\underline{y}) = T_{B,max}$. Projecting the trajectory resulting from the control $T(\underline{y}) = T_{B,max}$ onto the y_1 versus y_2 plane

gives a path with $\dot{y_2}$ satisfying Equation (5.28). This is everywhere below and behind the trajectory corresponding to $T(\underline{y}) = T_{A,s}$. Thus if $y_3 > 0$ it is always best to choose a control that first drives y_3 to zero and keeps it there until $y_1(t) = d$.

There is one exception to this. For initial conditions close enough to $y_1 = d$ there is insufficient time to drive y_3 to zero before y_1 hits d. Given an initial $y_{3,0} > 0$, the minimum time to make $y_3 = 0$ is

$$\bar{t} = \frac{y_{3,0}}{\gamma_2}. \tag{5.29}$$

At time $\bar{t}$,

$$y_1(\bar{t}) = \gamma_4 y_{3,0}^2 - \frac{y_{2,0}y_{3,0}}{\gamma_2} + y_{1,0}. \tag{5.30}$$

Thus, if

$$\gamma_4 y_{3,0}^2 - \frac{y_{2,0}y_{3,0}}{\gamma_2} + y_{1,0} \geq d \tag{5.31}$$

all controls give identical performance.

There is still the possibility that an optimal trajectory would include a segment with $y_{3,0} > 0$ at some initial time. Then

$$\dot{y}_2(t) = \frac{F_n}{m}\mu_{B,s} < 0 < \frac{F_n}{m}\mu_{A,max}. \tag{5.32}$$

Repeating the previous analysis completes the proof.

5.4.2 Sketch of Proof for the Hot Rodder

The hot-rodder problem can be proved in a similar way to the drag racing problem, but a coordinate transformation on the dynamics clarifies the proof. Let

$$\begin{aligned} z_1(t) &= x_1(t) - d \\ z_2(t) &= \dot{z}_1(t) \\ z_3(t) &= r\dot{\alpha}(t) - \dot{x}(t). \end{aligned} \tag{5.33}$$

Then

$$\begin{bmatrix} \dot{z_1} \\ \dot{z_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \hat{u}(t) \end{bmatrix} \tag{5.34}$$

where

$$\hat{u}(t) = \begin{cases} \frac{F_n}{m}\mu_{A,s}, & z_3 > 0 \\ \frac{F_n}{m}\mu_{B,s}, & z_3 < 0 \\ \alpha, & z_3 = 0, \text{ where } \frac{F_n}{m}\mu_{B,s} \leq \alpha \leq \frac{F_n}{m}\mu_{A,s} \end{cases} \tag{5.35}$$

and

$$\dot{z}_3(t) = \frac{rT(\underline{z})}{I} - F_n\mu_{i,s}\left(\frac{I + mr^2}{mI}\right). \tag{5.36}$$

This transformation emphasizes that the problem is just the classic problem of finding the minimum time to the origin for the double integral plant with a controllable delay in the switching time of the control $\hat{u}(t)$.

The optimal choice of a feedback control $\hat{u}(z_1, z_2)$ is well known. Since $\mu_{B,max} < \mu_{B,s} < 0 < \mu_{A,s} < \mu_{A,max}$ it is best to drive z_3 to zero as quickly as possible and keep it there until $z_1(t)$ and the race is over.

5.5 Conclusion

The real drag racer and hot-rodder problems are more complicated than the problems solved here. Firstly, the wheel/ground interaction is unknown and changing. The controller must estimate the torque at which the wheels start to spin. This is somewhat simpler than in the ABS case because of the way drag racing is actually done. Two cars race together. All that is necessary is to beat the other car. The winner advances to the next race. The competition is organized as a conventional tournament. Thus a competitor can use the early races to estimate the maximum torque while running at a torque he knows is below the maximum possible. This is a particular example of an iterative learning control problem. We hope to be able to formulate and solve this problem in the future. The real drag racer manages such large accelerations because distortion of the tires results in an effective $\mu_{A,max}$ much greater than one would expect from a less deformable tire [19]. Understanding and incorporating tire distortion in the vehicle model is a very hard problem.

References

1. Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhauser, Boston (1997)
2. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: The Mathematical Theory of Optimal Processes. John Wiley and Sons, New York (1962)
3. Athans, M., Falb, P.: Optimal Control: An Introduction to the Theory and Its Applications. McGraw-Hill, New York (1966)
4. Gillespie, T.: Fundamentals of Vehicle Dynamics. Society of Automotive Engineers, Warrendale (1992)
5. Bosch: Driving-Safety Systems. Society of Automotive Engineers, Warrendale (1999)
6. Pacejka, H.: Tire and Vehicle Dyanmics. Society of Automotive Engineers, Warrendale (2005)
7. Kiencke, U.: Automotive Control Systems For Engine, Driveline and Vehicle. Society of Automotive Engineers, Warrendale (2000)
8. Johansen, T.A., Kalkkul, J., Ludemann, J., Petersen, I.: Hybrid control strategies in ABS. In: Proc. American Control Conference, pp. 1704–1705 (2001)
9. Tsiotras, P., Canudas De Wit, C.: On the optimal braking of wheeled vehicles. In: Proc. American Control Conference, pp. 569–573 (2000)

10. Mirzaei, A., Moallem, M., Mirzaeian, B.: Optimal design of a hybrid controller for antilock braking systems. In: Proc. IEEE/ASME Int'l Conf. on Advanced Intelligent Mechatronics, pp. 905–910 (2005)
11. Canudas De Wit, C., Olsson, H., Astrom, K., Lischinsky, P.: A new model for control of systems with friction. IEEE Trans. on Automatic Control 40, 419–425 (1995)
12. Drakunov, S., Ozguner, U., Dix, P., Ashrafi, B.: ABS control using optimum search via sliding modes. IEEE Trans. on Control Systems Tech. 3, 79–85 (1995)
13. Muragishi, Y., Ono, E.: Application of hybrid control method to braking control system with estimation of tire force characteristics. R&D Review of Toyota CRDL 38, 22–30 (2003)
14. Choi, H., Hong, S.: Hybrid control for longitudinal speed and traction of vehicles. In: IECON 2002, vol. 2, pp. 1675–1680 (2002)
15. de Koker, P., Gouws, J., Pretorius, L.: Fuzzy control algorithm for automotive traction control systems. In: MELECON 1996, vol. 1, pp. 226–229 (1996)
16. Garcia-Rivera, M., Sanz, R., Perez-Rodriguez, J.A.: An antislipping fuzzy logic controller for a railway traction system. In: Proc. Sixth IEEE Int'l Conf. on Fuzzy Systems, vol. 1, pp. 119–124 (1997)
17. Sussmann, H.: A maximum principle for hybrid optimal control problems. In: Proc. IEEE Conference on Decision and Control, pp. 425–430 (1999)
18. Drag Race Results, `http://www.draglist.com/artman/publish/raceresults/article001432.shtml`
19. Hallum, C.: The magic of the drag tire, SAE Paper 942484. Presented at SAE MSEC (1994)

6

An Information Space View of "Time": From Clocks to Open-Loop Control*

Steven M. LaValle[1] and Magnus Egerstedt[2]

[1] Dept. of Computer Science, Univ. of Illinois, Urbana, IL 61801 USA
[2] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Summary. This paper addresses the peculiar treatment that time receives when studying control systems. For example, why is the ability to perfectly observe time assumed implicitly in virtually all control formulations? What happens if this implicit assumption is violated? It turns out that some basic control results fall apart when time cannot be perfectly measured. To make this explicit, we introduce information space concepts that permit imperfect time information to be considered in the same way as imperfect state information. We then argue that classical open-loop control should be reconsidered as perfect *time-feedback* control. Following this, we introduce a notion of *strongly open-loop* control, which does not require perfect time observations. We provide some examples of these concepts and argue that many fascinating directions for future controls research emerge.

6.1 Introduction

"Until the mid 1750s, navigation at sea was an unsolved problem due to the difficulty in calculating longitudinal position. Navigators could determine their latitude by measuring the sun's angle at noon. To find their longitude, however, they needed a portable time standard that would work aboard a ship. However, to find their longitude, they needed a portable time standard that would work aboard a ship. The purpose of a chronometer is to keep the time of a known fixed location, which can then serve as a reference point for determining the ship's position. Conceptually, by comparing local high noon to the chronometer's time, a navigator could use the time difference to determine the ship's present longitude" [7].

As known to anyone who has ever been exposed to a course on control theory, a sharp distinction is made between closed-loop and open-loop control. This distinction is drawn between control laws that involve references to the state of the system and control laws that are specified only in terms of time. The

* The authors thank Roger Brockett and Dan Koditschek for helpful comments and discussions. LaValle is supported in part by the DARPA SToMP program (DSO HR0011-07-1-0002). Egerstedt is supported in part by the National Science Foundation through NSF-CAREER award (grant # 0237971).

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 93–106.
springerlink.com

idea is that time is somehow is readily available to the controller, independent of the state of the system. Of course, time imperfections are often addressed in discrete-time or delayed-measurement models; however, a rich variety of other time uncertainty models can be imagined, rather than just quantizations or delay.

This paper came about because the special treatment of time seemed to the authors to be somewhat arbitrary in the context of robotics, particularly when considering that all information comes from sensors. As a simple example, consider the two systems: $\dot{x} = f(x, u),\ u = g(t)$ and $\dot{x} = f(x, u),\ \dot{z} = 1,\ u = g(z)$. Here one typically asserts that the first case is an open-loop controller and the second is closed-loop. However, they are mathematically the same control law (modulo initial conditions) and both seem to involve some kind of feedback.

The fundamental issues concerning time are not limited to simple mathematical technicalities. For example, in robotics, a number of sensors, including compasses, gyroscopes, accelerometers, wheel-encoders, GPS antennas, IR, sonar and laser-range finders, and cameras (just to name a few) are employed to estimate the internal state of the robot as well as characterize the environment in which it is deployed. However, clocks are also routinely employed, not just for synchronizing the executions, but also, for example, for smoothing the pose estimates or for establishing distances through ultrasonic range sensors. As such, it appears that clocks should be treated as any other sensor because they provide measurements of an important quantity for specifying control laws (whether implicitly or explicitly).

Rather than establish a collection of new results, this paper emphasizes the reconsideration of time and its role in control theory. This may lead to exciting new avenues for research, ultimately providing a better understanding of how time information affects control. Furthermore, systems can be designed that utilize minimal amounts of time information, thereby achieving greater robustness and affordability.

Of course, the study of time within a controls context is certainly not new. For example, in [1] it was pointed out that a temporally driven sampling strategy (so-called Riemann sampling) could be advantageously replaced (in some contexts) by a state-driven strategy (Lebesgue sampling). Similarly, by allowing for time to be controlled, dynamic-time warping has become a standard controls tool, for example for speed regulation in robotics [6, 9, 14]. Moreover, the view that open-loop control is potentially problematic is also not new, as illustrated by the fact that jitter in the clock is known to cause instabilities. As another example, going from open-loop control signals to corresponding closed-loop control signals is a research topic of continued interest in optimal control [4, 5]. Another relevant topic is asynchronous protocols for distributed systems (e.g., [8]). All of these research areas, although disjoint in explicit focus, share the feature of treating time in a non-standard manner.

To formulate time uncertainty in a general way, we extend standard machinery that was developed mainly for state uncertainty: *information spaces*. The earliest ideas are due to Kuhn in the context of sequential games [10], and were found to be a convenient, unified way to express state uncertainty in dynamic

game theory [2], stochastic control [3, 11], and planning algorithms [12]. Our approach is to treat time as "just another state variable" and consequently consider observations of time, parameterized by an internal, continuous index set.

6.2 Illustrative Examples

6.2.1 Brittle Time Sensitivity in Linear Systems

As a first example of the reliance on perfect time measurements, consider the standard, linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t),$$

in which the uncontrolled system is unstable, and (A, B) is a completely controllable pair. Using a static feedback law $u = -Kx$ for stabilizing this system yields $\dot{x}(t) = (A - BK)x(t)$, and $x(t) = e^{(A-BK)t}x_0$, in which $x(0) = x_0$. Now, in the absence of uncertainty we can of course implement this exact control law with an "open-loop" controller

$$u(t) = -Ke^{(A-BK)t}x_0.$$

We now assume that time has to be measured, and instead of t we observe $h(t)$, in which h is an output mapping. We obtain the measurement error $|h(t) - t|$, which is in fact due to a bias in the clock (in that it goes too quickly or too slowly) while it never causes "time" to go backwards; i.e., we assume that $dh(t)/dt > 0$.

Hence, the open-loop controller takes on the form $u(t) = -Ke^{(A-BK)h(t)}x_0$, and subsequently

$$\dot{x}(t) = Ax(t) - BKe^{(A-BK)h(t)}x_0.$$

Now, consider the system

$$\dot{\tilde{x}}(t) = dh(t)/dt(A - BK)\tilde{x}(t),\ \tilde{x}(t) = x_0.$$

This system is globally, asymptotically stable as long as $dh(t)/dt > 0$, which was assumed. The solution is

$$\tilde{x}(t) = e^{\int_0^t (dh(s)dt)ds(A-BK)}x_0 = e^{(A-BK)(h(t)-h(0))}x_0.$$

Under the additional assumption that the initial time measurement is correct, i.e. that $h(0) = 0$, we thus have that $u(t) = -K\tilde{x}(t)$, which in turn implies that

$$\begin{aligned}\dot{x}(t) &= Ax(t) - BK\tilde{x}(t)\\ \dot{\tilde{x}}(t) &= dh(t)/dt(A - BK)\tilde{x}(t).\end{aligned}$$

The first of these equations is an unstable linear system, driven by an input that will decay to zero because $\tilde{x}$ will tend to zero, and, as a result, the x-system is unstable. The only situation in which this will not happen is when $x(t) = \tilde{x}(t)$ for all t, which directly implies that $dh/dt = 1$, i.e., $h(t)$ is equal to $t + c$ for

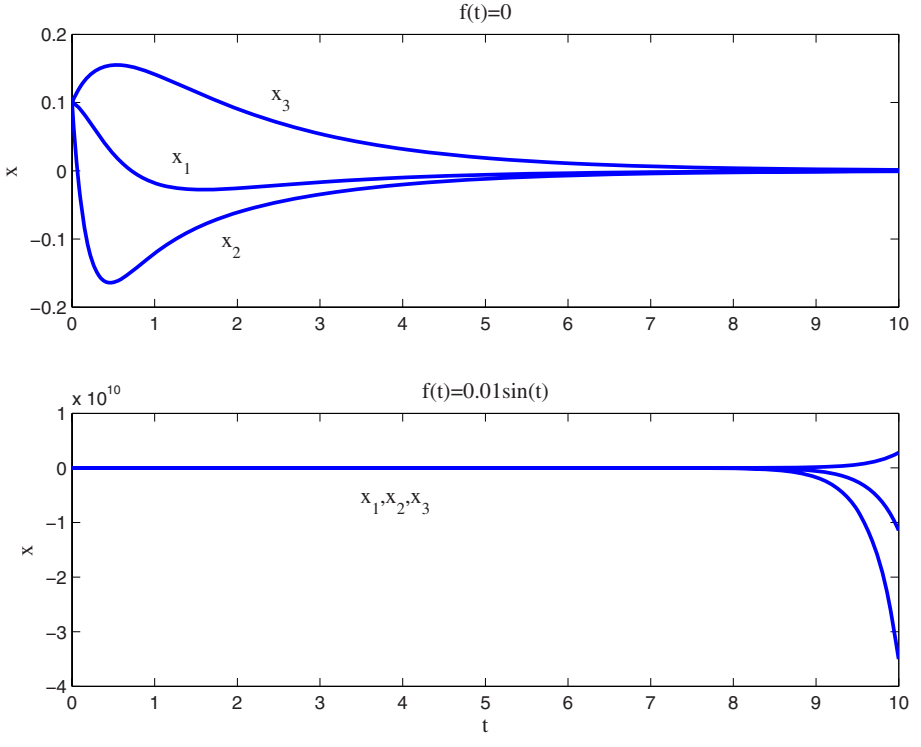


Fig. 6.1. Depicted is the evolution of a third order system. The upper plot corresponds perfect time measurements and the lower plot shows the evolution for the time disturbance $h(t) = t + 0.01\sin(t)$.

an arbitrary constant. However, since $\tilde{x}(0) = x(0) = x_0$ this means that $c = 0$. Hence, the only way in which the "open-loop" system will remain stable is with perfect time measurements. As an example, consider the situation depicted in Figure 6.1, where

$$A = \begin{bmatrix} 2 & 4 & 3 \\ 1 & 0 & 0 \\ 0 & -1 & -1 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 2 \\ -0.5 \end{bmatrix},$$

$$K = \begin{bmatrix} 3.5 & 3.2 & 3.8 \end{bmatrix}.$$

The system starts at $x(0) = (0.1, 0.1, 0.1)^T$. For a measurement model of the form $h(t) = t + f(t)$ for some function f, the upper plot shows the case in which $f(t) = 0$, and the lower shows a small disturbance of $f(t) = 0.01\sin(t)$.

6.2.2 A Nonlinear System with Time Robustness

In contrast to the previous time-sensitive system, consider the nonlinear system defined over $X = \Re$,

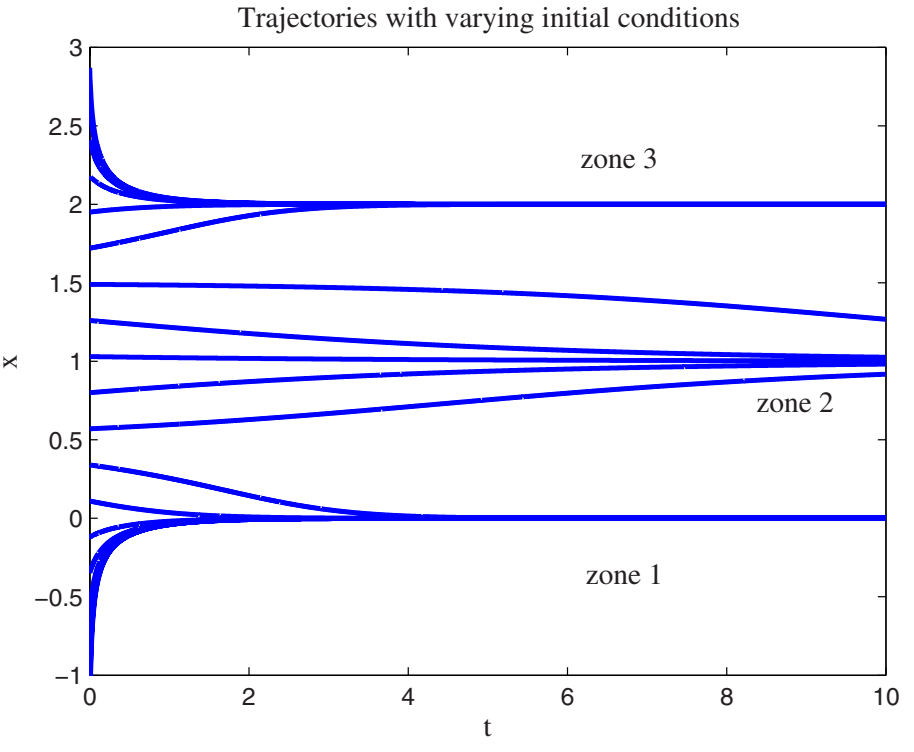


Fig. 6.2.

$$\dot{x} = \prod_{i=1}^{5}(a_i - x),$$

with real constants $a_1 < a_2 < \ldots < a_5$. This system has three stable equilibrium points $\{a_1, a_3, a_5\}$ and two unstable $\{a_2, a_4\}$.

We now assume that the controller consists of an impulse of height d. The control task is to apply this impulse at a distinct time τ; i.e.,

$$x(\tau^+) = x(\tau^-) + d.$$

Under the assumption that $x(0)$ is significantly smaller than a_1, this controller will move the system from the domain of attraction for the first stable equilibrium point (Zone 1 in Figure 6.2) to Zone 2 given certain conditions on τ and d. However, if the controller acts too quickly, this will not happen. On the other hand, if it acts too slowly, it may move the system to Zone 3 instead of Zone 2. As such, this system is robust to incorrectly measured time up to some level, after which the qualitative behavior of the system changes significantly.

This is shown in Figure 6.3, with

$$a_1 = 0,\ a_2 = 0.5,\ a_3 = 1,\ a_4 = 1.5,\ a_5 = 2,$$
$$x(0) = -2, d = 1.7.$$

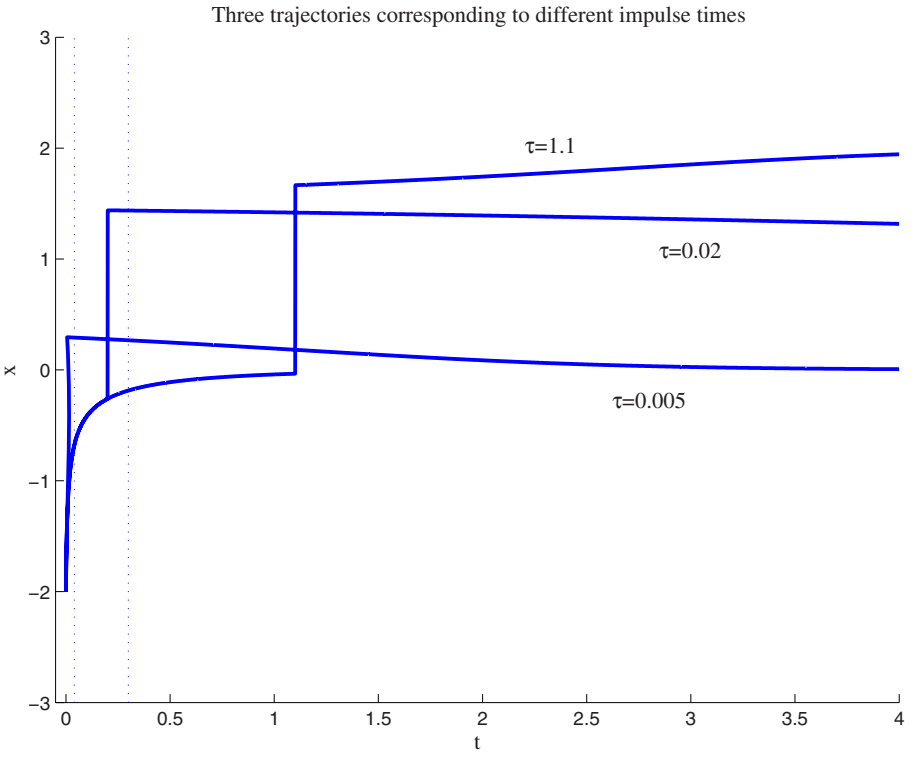


Fig. 6.3.

In the figure, three different scenarios are depicted, in which $\tau = 0.005$ results in a system that asymptotically approaches a_1, $\tau = 0.2$ results in $x \to a_3$, and $\tau = 1.1$ gives $x \to a_5$. Also depicted (with dotted lines as boundaries) are the regions where qualitatively different behaviors are obtained.

The conclusion to be drawn from this example is that, as opposed to the previous example, the controller is robust to small errors associated with the time sensor. However, the controller is not robust to arbitrarily large time measurement errors. This leads to the next example.

6.2.3 Strongly Open-Loop Stabilization

Consider the system

$$\dot{x} = (1 - u)A_1 x + uA_2 x,$$

in which $x \in \Re^n$, $u \in \Re$. Furthermore, assume that $\dot{x} = A_1 x$ is unstable and $\dot{x} = A_2 x$ is asymptotically stable. For any fixed τ, the control law

$$u(t) = \begin{cases} 0 \text{ if } t < \tau \\ 1 \text{ if } t \geq \tau \end{cases}$$

renders the controlled system asymptotically stable. This is indeed true because the controller will switch from the unstable to the stable subsystem at time τ. Moreover, it does not matter when this switch occurs, as long as it does occur, which means that time measurements are really not needed to implement the controller. This is an example of stabilization using a *strongly open-loop* controller, which is defined precisely in Section 6.5.

These simple examples illustrate interesting phenomena that should be carefully studied in increasing degrees of generality. We hope that this would lead to a greater understanding of control systems in general and the particular role that time should play. To begin this quest to improve our understanding, we introduce definitions in the remainder of the paper that permit great modeling flexibility in allowing time measurement uncertainty.

6.3 Basic Definitions

The definitions in this section build on standard control terminology and information space concepts that have previously been included into control formulations (e.g., [2]). Some concepts and terminology are also borrowed from robotics, particularly in the treatment of planning under sensing uncertainty ([12], Chapter 11). Readers familiar with various treatments of uncertainties in control systems will observe similarities with input-output representations, behavioral systems theory [15], quantized control systems [13], and imperfect state information in stochastic control [3, 11]. Throughout the coming presentation, the control systems are particularly inspired by robotics applications.

6.3.1 Inputs, Observations, and History

Let U and Y be the *input space* and *observation space*, respectively; i.e., the spaces in which the inputs and outputs take on values. It is now conceivable that the controllers and sensors are evolving asynchronously, or in other ways are driven by device-specific counters, which may be continuous. As such, we define index sets P and S, which parameterize the inputs and outputs, respectively. If we let $P = S = \mathbb{N}$, then we would obtain a classical discrete-time control model. However, we prefer to let P and S both be closed intervals of $\Re$. For convenience, suppose $P = S = [0, \infty)$. This intuitively corresponds to having a continuous index set that starts at zero and increases monotonically during execution. Let $\tilde{u}_p$ denote a function, $\tilde{u}_p : [0, p] \rightarrow U$, called the *input history*, and let $\tilde{y}_s$ denote a function, $\tilde{y}_s : [0, s] \rightarrow Y$, called the *observation history*. These histories are considered as partial functions on P and S, respectively, because $[0, p] \subset P$ and $[0, s] \subset S$.

The histories naturally "grow" during execution. For example, suppose during execution, $\tilde{y}_s$ is obtained, followed by $\tilde{y}_{s'}$ at some later stage. We then have that $s' > s$, together with the requirement is that $\tilde{y}_{s'}(s'') = \tilde{y}_s(s'')$ for all $s'' \in S$.

Now, given input and output histories associated with a particular evolution of a control system, we let $\eta = (\tilde{u}_p, \tilde{y}_s)$ denote the *history I-state* ("I-state"

is short for "information state"). Moreover, we let the *history I-space* be the set of all possible I-states, η, which includes any permissible $\tilde{u}_p$ and $\tilde{y}_s$ for any $p, s \in [0, \infty)$.

6.3.2 Interacting with a State Space

Now introduce a state space X. Let $\tilde{x}$ denote a *state trajectory*, which is a time-parameterized function into X. We now need to define a transition function, and for this we let $\tilde{w}_t : [0, t] \rightarrow U$ be a time-parameterized *control function*. Let Φ denote the *state transition function*:

$$\tilde{x}(t' + t) = \Phi(\tilde{x}(t'), \tilde{w}_t).$$

For example, Φ could be defined in the usual way if X is a differentiable manifold with an associated *control system* $\dot{x} = f(x, u)$, in which $x \in X$ and $u \in U$. In this case, the state transition function Φ becomes:

$$\tilde{x}(t' + t) = \Phi(\tilde{x}(t'), \tilde{w}_t) = \tilde{x}(t') + \int_{t'}^{t} f(\tilde{x}(\tau), \tilde{w}_t(\tau)) d\tau,$$

assuming appropriate integrability conditions.

6.3.3 The State-Time Space

Even though the way control signals interact with state variables above is not only well-known and supposedly unproblematic, we need to incorporate time more explicitly to understand the role of time in a more direct manner. For this, consider incorporating time into the state space definition to form the *state-time space*. Let $T \subseteq \Re$ be the maximal *time interval* for which the system is defined. Let $Z = X \times T$ denote the *state-time space*. Elements of Z will be denoted as z or (x, t).

We are comfortable with time-parameterized paths through X. However, what is not entirely clear is: What is the best way to parameterize paths through Z? Consider a path $\tau : R \rightarrow Z$, in which R is a closed interval. For example, imagine rigid bodies moving in $\Re^3$ according to the laws of mechanics. To show the motions of the bodies, we could parameterize τ in many ways, much in the same way as varying parameters in a computer-generated animation. We could vary the speed, play parts of it backward, and so on. Since there are many possibilities, we would like to chose one that is most convenient for the coming formulations. Returning to the analogy of an animation, we assume that animations are played forward in "real time". Rather than allow any arbitrary path τ, this means that we require that $d\tau/dr = 1$ (for the case in which Z is a differentiable manifold), in which $r \in R$ is the path parameter.

Since time progresses forward monotonically, any path τ could be reparameterized using t, even though t is a coordinate of Z. This would be convenient, but note that it is somewhat abusive because t is serving both as a coordinate of Z

and the parameter of a path, τ. Nevertheless, we assume that all paths through Z are parameterized by time t and are referred to as *(state-time) trajectories*. A trajectory is denoted by $\tilde{z}$ and refers to a mapping $\tilde{z} : [t_1, t_2] \rightarrow Z$ in which t_1 and t_2 are the starting and ending times, respectively.

If the dimension of X is n, then let z_{n+1} denote the last component of z, i.e., the time component. Under the observation that $\dot{z}_{n+1} = 1$, it is straightforward to convert a system $\dot{x} = f(x, u)$ into $\dot{z} = f'(z, u)$ by simple extension. In general, Φ can be extended in the straightforward way to obtain Φ'.

A trajectory $\tilde{z}$ is called *valid* if there exists some control function $\tilde{w}$ such that $\dot{z} = f'(z, u)$ is satisfied for all times along $\tilde{z}$. Let $\tilde{Z}$ denote the space of all valid trajectories.

6.3.4 Initial Conditions on the State-Time Space

Once the transition has been made to viewing time as something that must be measured and is part of the state-time variable, time-dependent initial conditions become important. In fact, we would like to express various initial conditions on Z; therefore, the history I-state will be expanded to include the initial conditions.

Let $Z_0 \subseteq Z$ denote a given *initial condition*, which indicates the set of possible initial state-times. The expanded history I-state is defined as $\eta = (Z_0, \tilde{u}_p, \tilde{y}_s)$.

Three important special cases of initial conditions are:

1. **Known initial state-time:** Z_0 is a singleton, $\{(x_0, t_0)\}$, which means that the initial time is t_0 and the initial state is x_0. Thus, (6.6) considers only trajectories for which $\tilde{z}(t_0) = (x_0, t_0)$.
2. **Known initial time:** Suppose it is known that t_0 is the starting time. In this case,
$$Z_0 = \{(x, t) \in Z \mid t = t_0\}. \tag{6.1}$$
3. **Known initial state:** Suppose it is known that x_0 is the starting state. In this case,
$$Z_0 = \{(x, t) \in Z \mid x = x_0\}. \tag{6.2}$$

Whenever the initial time is given, we will typically have $t_0 = 0$.

6.3.5 Sensor Mappings

A sensor is defined in terms of a mapping from Z (or trajectories on Z) into Y, the observation space. A sensor may be either: 1. *instantaneous*, which means that the observation depends only on the current state-time, 2. *history-based*, which means that the observation may depend on any portion of the entire state-time trajectory.

An instantaneous *sensor mapping* is defined as $h : Z \rightarrow Y$, for which we write $y = h(z)$ or $y = h(x, t)$. For a history-based sensor mapping, we have $h : \tilde{Z} \rightarrow Y$, and an observation is given as $y = h(\tilde{z})$.

Several simple and important sensors are now defined.

1. **(Perfect Information)** A *perfect state-time sensor* is defined as any injective mapping $h : Z \rightarrow Y$ because (x, t) can be recovered from any observed y. The simplest case is $Y = Z$ and $y = h(z) = z$.
2. **(State-Only)** A *perfect state sensor* can be defined, for example, as $y = h(x, t) = x$. In this case, we may know the current state but remain uncertain about the particular time.
3. **(Clock)** A *perfect time sensor* or (perfect) *clock* is defined as $y = h(x, t) = t$.
4. **(Chronometer)** An important history-based sensor is the *perfect time odometer* or (perfect) *chronometer*, which yields the total time elapsed for a state-time trajectory. Let $\tilde{z}_{t,t'}$ denote a valid space-time trajectory with endpoints t and t', and $t < t'$. The chronometer is defined as $h(\tilde{z}_{t,t'}) = t' - t$.

6.3.6 Disturbances

In this section, we briefly illustrate how to incorporate disturbances into Φ and h; however, to simplify the presentation, we will not include such disturbances in the subsequent sections.

In the case of smooth manifolds, f can be extended to obtain $\dot{x} = f(x, u, \theta)$ in which θ is selected from Θ, a set of possible disturbances (or nature inputs).[3] In general, the disturbance $\theta \in \Theta$ can be incorporated into Φ to obtain $\tilde{x}(t' + t) = \Phi(\tilde{x}(t'), \tilde{w}_t, \theta)$.

Another disturbance parameter can be defined, to interfere with sensors. For example, $y = h(z, \psi)$, in which $\psi \in \Psi$ and Ψ is an *observation disturbance space*. Imperfect versions of previously defined sensors can be made. For example, an *imperfect clock* is defined as $y = h(x, t, \psi) = t + \psi$ and $\Psi = [-\epsilon, \epsilon]$, in which $\epsilon > 0$ represents the maximum error in the time measurement.

6.4 Nondeterministic I-Spaces

In this section, we introduce an *information mapping* and *derived* I-space, which means that history I-states are mapped into a new I-space that provides some interpretation or aggregation of the histories (see [12], Chapter 11, for examples). One possibility is to map η to a posterior pdf $p(z|\eta)$, which would lead to probabilistic I-states (these appear in stochastic control, but usually over X rather than Z). Although there are many possibilities, we exclusively consider a nondeterministic interpretation of the histories: The smallest subset of Z that is consistent with a history I-state η. A set-valued information mapping is thus defined, and is denoted as $Z(\eta) \subseteq Z$. The target of this mapping is a *nondeterministic I-space*, $\mathcal{I}_{ndz} = \text{pow}(Z)$.[4]

[3] Disturbances in the time direction seem absurd and will therefore not be considered.

[4] In most contexts, most elements of $\text{pow}(Z)$ are unnecessary; however, it is simpler to define $\text{pow}(Z)$ than to worry about the precise reachable subset of $\text{pow}(Z)$ (which is an interesting research problem in itself!).

6.4.1 Relating Internal Parameters to Time

The first step is to define the class of mappings that relate the index sets P and S to time. It is assumed that the particular mapping is unknown; otherwise, the precise time could be reconstructed if the mapping is injective. Instead we define Ω as the set of possible mappings from T to P. Similarly, let Λ denote a set of possible mappings from T to S. Although the particular mapping is not given, we assume that the sets Ω and Λ are specified in advance.

Many reasonable definitions are possible for Ω and Λ. Consider defining Ω (the same possibilities exist for Λ). One of the weakest sensible definitions is that Ω contains any mapping $\omega : T \to P$ for which $\omega(t)$ monotonically increases. This at least ensures that a higher index implies later time.

Another possibility is to restrict Ω to differentiable functions and require bounded derivatives and bounded initial error. For example, each $\omega \in \Omega$ must satisfy $|\omega(0)| \leq c_0$ and $|d\omega/dt| \leq c_1$ for some positive constants c_0 and c_1. This restricts the possible times to an interval that widens as time increases. Without assuming differentiability, a similar function space could be obtained using Lipschitz constants instead of bounded derivatives.

6.4.2 The Sensorless Case

We now define the mapping from history I-states into the nondeterministic I-space, $\mathcal{I}_{ndz}$. Assume that Ω and Λ contain only invertible functions (otherwise, the inverses below can be replaced by preimages to obtain slightly more complicated definitions). First consider the sensorless case, in which $\eta = (Z_0, \tilde{u}_p)$. The nondeterministic I-state (smallest consistent subset of Z) given η is

$$\begin{aligned} Z(\eta) = \big\{ & (x,t) \in Z \mid \exists \tilde{z} \in \tilde{Z} \text{ and } \exists \omega \in \Omega \text{ such that} \\ & \tilde{z}(t_1) \in Z_0 \text{ and } \forall t \in [t_1, t_1 + \omega^{-1}(p)), \\ & \tilde{z}(t_1 + t) = \Phi'(\tilde{z}(t_1), \tilde{u}_p(\omega(t))) \big\}, \end{aligned} \tag{6.3}$$

in which t_1 refers to the starting time of $\tilde{z}$.

6.4.3 The Inputless Case

Now consider the case in which $\eta = (Z_0, \tilde{y}_s)$; $\tilde{z}$ is thus completely predictable from any initial $z \in Z$. We have

$$\begin{aligned} Z(\eta) = \big\{ & (x,t) \in Z \mid \exists \tilde{z} \in \tilde{Z} \text{ and } \exists \lambda \in \Lambda \text{ such that} \\ & \tilde{z}(t_1) \in Z_0 \text{ and } \forall t \in [t_1, t_1 + \lambda^{-1}(s)], \\ & h(\tilde{z}(t)) = \tilde{y}_s(\lambda(t)) \big\}. \end{aligned} \tag{6.4}$$

6.4.4 Combining Sensors and Inputs

Now consider the case in which an observation history $\tilde{y}_s$ is also given, yielding $Z(\eta) = Z(Z_0, \tilde{u}_p, \tilde{y}_s)$. The nondeterministic I-state in this case combines the constraints from both (6.3) and (6.4), to obtain:

$$
\begin{aligned}
Z(\eta) = \big\{ (x,t) \in Z \mid \exists \tilde{z} \in \tilde{Z}, \exists \omega \in \Omega, \text{ and } \exists \lambda \in \Lambda \\
\text{ such that } \tilde{z}(t_1) \in Z_0 \text{ and} \\
\forall t \in [t_1, t_1 + \omega^{-1}(p)), \\
\tilde{z}(t_1 + t) = \Phi'(\tilde{z}(t_1), \tilde{u}(\omega(t))) \text{ and} \\
\forall t \in [t_1, t_1 + \lambda^{-1}(s)],\ h(\tilde{z}(t)) = y(\lambda(t)) \big\}.
\end{aligned} \tag{6.5}
$$

6.5 Defining Control Laws

A wide variety of control laws may be defined in terms of *information feedback*. We could define a control law as a mapping from the history I-space into the input space U, but this would be difficult to manage. Therefore, we consider control laws that map from a derived I-space into U. In particular, we consider in this section control laws of the form $\gamma : \mathcal{I} \to U$, in which $\mathcal{I}$ is a particular derived I-space $\mathcal{I}$ that is obtained as a mapping from the history I-space. Many possibilities exist; for example, $\mathcal{I}$ may be Z, X, T, P, or $\mathcal{I}_{ndz}$.

6.5.1 Strongly Open-Loop Control

Consider sensorless information states of the form $\eta = \tilde{u}_p$ (the initial state-time could be any $z \in Z$). Let the derived I-space $\mathcal{I} = P$ be defined by the information mapping $\tilde{u}_p \mapsto p$. A *strongly open-loop* control law is defined as $\gamma : P \to U$.

The trajectory obtained by applying $\tilde{u}_p$ could be any $\tilde{z}$ for which there exists an $\omega \in \Omega$ such that $\tilde{z} : [t_1, t_1 + \omega^{-1}(p)] \to Z$ and

$$
\tilde{z}(t_1 + t) = \tilde{z}(t_1) + \int_{t_1}^{t_1 + t} f'(\tilde{x}(t'), \tilde{u}(\omega(t')))dt' \tag{6.6}
$$

for all $t \in [0, \omega^{-1}(p)]$.

Now suppose there is an initial condition $Z_0 \subseteq Z$ so that $\eta = (Z_0, \tilde{u}_p)$. In this case (6.6) is constrained to consider only those $\tilde{z} \in Z$ for which $\tilde{z}(t_1) \in Z_0$. The last example of Section 6.2 represents strongly open-loop stabilization, even when $Z_0 = Z$ and Ω contains all time-monotonic functions.

6.5.2 Perfect Time-Feedback Control

To obtain *perfect time-feedback control* (otherwise classically known as "open loop" control), consider the derived I-space, $\mathcal{I} = T$, in which the exact time t can be derived from η. In this case, the information mapping is $(Z_0, \tilde{u}_p, \tilde{y}_s) \mapsto t$. The most common case occurs with a perfect time sensor, as defined in Section 6.3.5. The control law is specified as $\gamma : T \to U$. The most common special case is when Z_0 yields a known initial space-time $(x_0, 0)$.

6.5.3 Imperfect Time-Feedback Control

Now consider a derived I-space $\mathcal{I} = \text{pow}(T)$, in which a derived I-state is interpreted as the smallest set of times that are consistent with η. A control law is expressed as $\gamma : \text{pow}(T) \rightarrow U$ (we could restrict the domain to consider only reasonably behaved subsets of T, such as closed intervals). In terms of information requirements, this control law lies somewhere in between strongly open-loop control and perfect time-feedback control.

6.5.4 Other Control Laws

Many other laws may be considered. If it is possible to reconstruct z from η, then *perfect state-time feedback* is possible, defined as $\gamma : Z \rightarrow U$. In this case Z is considered as a derived I-space, $\mathcal{I} = Z$. Note that *perfect state feedback*, $\gamma : X \rightarrow U$ may also be considered, in which case the precise times are unnecessary (a known benefit of state-feedback control). Finally, using (6.5), we can define feedback on the nondeterministic I-space, yielding $\gamma : \mathcal{I}_{ndz} \rightarrow U$. This means that subsets of Z are mapped into U.

6.6 Open Questions and Issues

Using information space concepts, we have proposed new ways to formulate uncertainty in time measurements. We called classical open-loop control "perfect time-feedback control" and introduced the notion of *strongly* open-loop control, which is robust with respect to massive time distortions. We provided several examples that illustrate the associated issues.

We are fascinated by the numerous exciting questions and issues raised by the examples and formulations developed in this paper, which is intended to open doors to new problems, rather than close them with particular results. Here are some points worth considering:

1. What classes of systems support strongly open-loop stabilization? Stability might be a property too strong to demand of an open-loop controller with imperfect time measurements. Instead, other interesting properties such as state containment or invariance should also be studied.
2. Can local (small time perturbations), in contrast to global strongly open-loop control, be characterized in a meaningful manner? Is there a notion of robustness to time perturbations that can be used for designing control laws for robotic systems that are locally immune to time perturbations?
3. Is there a meaningful notion of time observers? In other words, can we estimate t from the history I-state $\eta = (Z_0, \tilde{u}_p, \tilde{y}_s)$? Also, how would an open-loop controller interact with the plant based on the time estimate?
4. What are the relationships between history-based sensors and histories of observations obtained from instantaneous sensors? For example, by measuring angular velocity and having a perfect chronometer, we can simulate an

angular odometer. If the initial angle is given, then we can furthermore simulate a compass. Furthermore, a variety is imperfect versions can be made by replacing the perfect chronometer with a weaker sensor.

5. What other ways can we relate S, P, and T? So far, we related S and P to T via two mappings. We may instead want, for example, to relate S to P, and then P to T. We might even want to consider mappings that are not invertible.
6. Can useful control laws be defined over $\mathcal{I}_{ndz}$? These would choose actions based on particular subsets of $X \times T$.
7. Can control laws that are robust with respect to severe time uncertainty lead to improved approaches to distributed, asynchronous control?

References

1. Åström, K.J., Bernhardsson, B.M.: Comparison of Riemann and Lebesgue sampling for first order stochastic systems. In: Proc. IEEE Conf. Decision and Control, pp. 2011–2016 (2002)
2. Basar, T., Olsder, G.J.: Dynamic Noncooperative Game Theory, 2nd edn. Academic, London (1995)
3. Bertsekas, D.P.: Dynamic Programming and Optimal Control, 2nd edn., vol. I. Athena Scientific, Belmont (2001)
4. Boccadoro, M., Wardi, Y., Egerstedt, M., Verriest, E.: Optimal control of switching surfaces in hybrid dynamical systems. J. of Discrete Event Dynamic Systems 15(4), 433–448 (2005)
5. Bryson, A.E., Ho, Y.C.: Applied Optimal Control. Hemisphere Publishing Corp., New York (1975)
6. Egerstedt, M., Hu, X., Stotsky, A.: Control of mobile platforms using a virtual vehicle approach. IEEE Trans. on Automat. Contr. 46(11), 1777–1782 (2001)
7. Wikipedia: The Free Encyclopedia, Chronometer (February 2007), `http://en.wikipedia.org/wiki/Chronometer`
8. Ganguli, A., Cortés, J., Bullo, F.: Distributed deployment of asynchronous guards in art galleries. In: Proc. American Control Conf., pp. 1416–1421 (2006)
9. Holm, J.K., Lee, D., Spong, M.W.: Time scaling for speed regulation in bipedal locomotion. In: Proc. IEEE International Conf. on Robotics and Automation (2007)
10. Kuhn, H.W.: Extensive games and the problem of information. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games, pp. 196–216. Princeton Univ. Press, Princeton (1953)
11. Kumar, P.R., Varaiya, P.: Stochastic Systems. Prentice Hall, Englewood Cliffs (1986)
12. LaValle, S.M.: Planning Algorithms. Cambridge University Press, Cambridge (2006), `http://planning.cs.uiuc.edu/`
13. Liberzon, D., Hespanha, J.P.: Stabilization of nonlinear systems with limited information feedback. IEEE Trans. on Automat. Contr. 50(6), 910–915 (2005)
14. Pappas, G.J.: Avoiding saturation by trajectory reparameterization. In: Proc. IEEE Conf. Decision and Control (1996)
15. Polderman, J.W., Willems, J.C.: Introduction to Mathematical Systems Theory: A Behavioral Approach. Springer, New York (1998)

7

Control System Design for the Capsubot

Hongyi Li[1], Namkon Lee[2], Norihiro Kamamichi[3], and Katsuhisa Furuta[3]

[1] Shenyang Institute of Automation, China
[2] Department of Advanced Multidisciplinary Engineering, Graduate School of Advanced Science and Technology, Tokyo Denki University, Japan
[3] Department of Robotics and Mechatronics, School of Science and Technology for Future Life, Tokyo Denki University, Japan

Summary. In this paper, a capsule type robot, named the Capsubot, is designed. The robot has no moving part outside of its body, no legs and no wheel. Its motion is purely based on its internal force and friction with the environments. A four-step motion pattern is proposed. A minimal energy solution is stated. A prototype capsule robot, consisting a plastic tube with three independently driven copper coils and a NiFeB magnet rod which can move inside the tube, is built. It is essentially a motion magnet linear motor. The motion generation results are verified experimentally.

7.1 Introduction

In physics, friction is the resistive force that occurs when two surfaces travel along each other while forced together. It causes physical deformations and heat buildup. It is well-know that a rigid body containing a movable internal part can move along a horizontal plane, if the internal part does certain patterned motion inside the body and static friction acts between the body and the plane. There are mainly two types of self-propelled (micro)robots called "capsule robot" or "micro robot". One is the legged robot or the inchworm robot. Robots of this type have external moving parts and move by changing their body shape. Therefore, the complex structure makes it hard to make them small and make them easy to break down, possibly causing injury to human subjects in medical applications [3]. Another type is an external propelled robot, e. g., a passive magnet (robot) is driven by external magnetic fields [4]. They need strong magnetic fields (remote) and are hard to control. The whole system is quite big (because of the need to put a human subject inside its external magnetic fields), therefore expensive. Further, there is proposed a micro capsule robot, not a self-propelled type, for examining an organ by an oral administration. This robot incorporates a camera device, a lighting system, a transmitting and receiving device, a control device, and a power supply device in a capsule body. The transmitting/receiving device is equipped with a stop means which is operated based on a stop control signal wirelessly received from the outside of the human body and stops or delays the movement of the micro capsule robot for a specified examination of part of the organ. However, this microcapsule robot has, as the stop means, a hanging

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 107–123.
springerlink.com

member or expansion member projecting from the robot body, or a adsorption member arranged in the periphery of the robot body. Since the stop means are the moving parts which change the body shape outside of the robot, this robot has drawbacks as mentioned above. Also, this microcapsule robot is not a self-propelled type. This paper describes theoretical and experimental efforts concerning a new type of capsule robot design, named Capsubot. The motion generation of the Capsubot uses internal forces and static friction only, which contrasts to the case that two body parts both interact with the plane in [2]. A related pendulum driven cart systems was experimentally implemented by the authors of the this paper [5]. In section 7.2, a four step patterned motion generation procedure of the capsule robot is proposed. In section 7.3, an optimal control law is derived for the Capsubot locomotion. In section 7.4, simulation results are shown. In section 7.5, a prototype capsule robot is designed. In section 7.6, the pendulum cart is shown as an example.

7.2 One Dimensional Motion Generation

The Capsubot is a two mass system, which is a modification of the two mass system in [2], shown in Figure 7.1. Similar ideas have exploited by several groups [6, 7, 8]. Higuchi et al. (1990) studied precision positioning using piezo impact drive mechanism [6]. Later on, Darby and Pellegrino (1997) designed an inertial stick-slip actuator for shape and vibration control of a space structures [7]. Recently, Chernousko (2005) analyzed motion of such two mass system in anisotropic media [8]. The Capsubot consists of a capsule shell m_1 and a cylinder mass m_2 which can move forward and backward inside the capsule shell, where x_1 and x_2 are the position of m_1 and m_2 with respect to an external fixed frame, μ is the friction coefficient between the shell and the external horizontal surface contacting with it, and u is the internal force input between the shell and the cylinder. Here we don't distinguish between static friction and sliding friction coefficients, though they are slightly different from each other in real experiments. In the experimental part of this paper, the internal force u is generated using a permanent magnet and a coil pair. By observation, the Capsubot cannot move directionally using a sine function type symmetric periodic inputs u. Under such input, the Capsubot can only oscillate around its center of mass.

To move the capsule forward, the required motion consists of four steps:

1. Large backward accelerated motion of m_2 ($\ddot{x}_2 \ll 0$), forward accelerated motion of m_1 ($\ddot{x}_1 > 0$).
2. Small backward decelerated motion of m_2 ($\ddot{x}_2 > 0$), forward decelerated motion of m_1 ($\ddot{x}_1 < 0$).
3. Small backward decelerated motion of m_2 ($\ddot{x}_2 > 0$), m_1 remains stationary ($\dot{x}_1 = 0$).
4. Forward slow motion of m_2 ($\dot{x}_2 \leq \varepsilon$), m_1 remains stationary ($\dot{x}_1 = 0$).

For basic rigid body physics, one can refer to [9]. Let the initial time be 0 and t_i be the terminal time of step i. For step 1 and 2, assume the boundary conditions

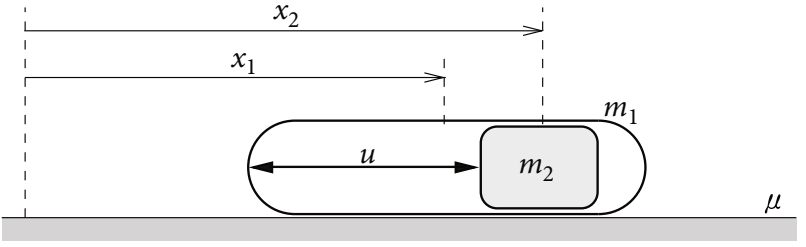


Fig. 7.1. Diagram of the Capsubot

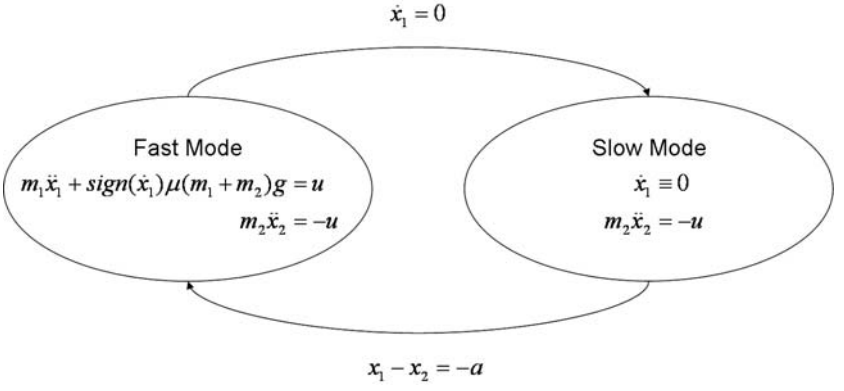


Fig. 7.2. State Transition Diagram

to be $\dot{x}_1(0) = \dot{x}_1(t_2) = \dot{x}_2(0) = 0$, $x_1(0) = 0$ and $x_2(0) = x_1(t_2) - x_2(t_2) = a$. The equation for motion is

$$m_1\ddot{x}_1 + \text{sign}(\dot{x}_1)\mu(m_1 + m_2)g + m_2\ddot{x}_2 = 0 \tag{7.1}$$

With the equation of motion (7.1) for Capsubot, we can put the motion generation into a fast mode (step 1 and 2) and slow mode (step 3 and 4) as shown in the state transition diagram, Figure 7.2. In fast mode, the internal force dominates, so the capsule shell moves against the external surface and the internal cylinder. In slow mode, the external friction dominates, the cylinder moves back to its initial position relative to the shell while keeping the shell stationary.

In fast mode, $\dot{x}_1 > 0$, integrating the motion equation (7.1) once, we get

$$m_1\dot{x}_1 + \mu(m_1 + m_2)gt + m_2\dot{x}_2 = 0 \tag{7.2}$$

Integrating (7.1) twice, we have

$$m_1x_1 + \frac{1}{2}\mu(m_1 + m_2)gt^2 + m_2x_2 - m_2a = 0 \tag{7.3}$$

Substituting the terminal time t_2 of step 2 into equation (7.5), we have the travel distance of the fast mode to be

$$x_1(t_2) = \frac{2m_2 a}{m_1 + m_2} - \frac{1}{2}\mu g t_2^2 \tag{7.4}$$

and the velocity of the internal mass at t_2 from equation (7.2)

$$\dot{x}_2(t_2) = -\frac{\mu(m_1 + m_2)gt_2}{m_2}. \tag{7.5}$$

From equation (7.5), in order to travel a longer distance the duration of the fast mode t_2 should be small.

7.3 Optimization

In order to find a force u to generate the above motion, we rewrite the motion equation in a control form

$$\begin{cases} m_1\ddot{x}_1 + \text{sign}(\dot{x}_1)\mu(m_1 + m_2)g = u \\ m_2\ddot{x}_2 = -u \end{cases} \tag{7.6}$$

where u is the internal force applied between m_1 and m_2. For convenience, we change position the x_2 of cylinder m_2 to the relative position $\overline{x}_2$ with respect to its capsule shell m_1. Let $\overline{x}_2 = x_2 - x_1$. We have

$$\begin{cases} m_1\ddot{x}_1 + \text{sign}(\dot{x}_1)\mu(m_1 + m_2)g = u \\ m_1 m_2\ddot{\overline{x}}_2 - \text{sign}(\dot{x}_1)\mu(m_1 + m_2)m_2 g = -(m_1 + m_2)u \end{cases} \tag{7.7}$$

with initial condition $x_1(0) = 0$, $\overline{x}_2(0) = a$, and $\dot{x}_1(0) = \dot{\overline{x}}_2(0) = 0$. For a fixed terminal time t_2, we need to find a control u so that the system matches the terminal conditions $\dot{x}_1(t_2) = 0$ and $\overline{x}_2(t_2) = -a$. There should not be a unique control u to accomplish the task. It makes senses to find a minimal energy control among all controls which satisfy the above boundary conditions. That is to say, find a control u to minimize the energy function

$$\min_u \eta = \int_0^T u^2\, dt \tag{7.8}$$

so that the system reaches the terminal positions $\dot{x}_1(t_2) = 0$ and $\overline{x}_2(t_2) = -a$. For results on optimal control, one can refer to [1]. Integrating the equation for x_1 in (7.7), from the terminal condition $\dot{x}_1(t_2) = 0$ we have

$$\mu(m_1 + m_2)gt_2 = \int_0^{t_2} u\, d\sigma. \tag{7.9}$$

We can decompose u into average force and variation $u = u_0 + \delta u$ where $u_0 = \mu(m_1 + m_2)g$ and $\int_0^T \delta u\, dt = 0$. The energy function becomes

$$\begin{aligned} \eta &= \textstyle\int_0^{t_2} (u_0 + \delta u)^2\, dt \\ &= \textstyle\int_0^{t_2} u_0^2 + 2u_0\delta u + (\delta u)^2\, dt \\ &= u_0^2 t_2 + \textstyle\int_0^{t_2} (\delta u)^2\, dt. \end{aligned} \tag{7.10}$$

Thus we only need to minimize variation

$$\eta\prime = \int_0^{t_2} (\delta u)^2 \, dt. \tag{7.11}$$

Integrating the equation for $\overline{x}_2$ in (7.7) twice, from the terminal condition $\overline{x}_2(t_2) = -a$ we have

$$\begin{aligned}
&-2m_1m_2a - \tfrac{1}{2}\mu(m_1+m_2)m_2gt_2^2 \\
&= -(m_1+m_2)\int_0^{t_2}\int_0^t u(\sigma)\, d\sigma dt \\
&= -(m_1+m_2)\int_0^{t_2}\int_0^t \mu(m_1+m_2)g + \delta u(\sigma)\, d\sigma dt \\
&= -\tfrac{1}{2}\mu(m_1+m_2)^2gt_2^2 - (m_1+m_2)\int_0^{t_2}\int_0^t \delta u(\sigma)\, d\sigma dt \\
&= -\tfrac{1}{2}\mu(m_1+m_2)^2gt_2^2 - (m_1+m_2)\int_0^{t_2}\int_\sigma^t \delta u(\sigma)\, dt d\sigma \\
&= -\tfrac{1}{2}\mu(m_1+m_2)^2gt_2^2 - (m_1+m_2)\int_0^{t_2}(t_2-\sigma)\delta u(\sigma)\, d\sigma \\
&= -\tfrac{1}{2}\mu(m_1+m_2)^2gt_2^2 + (m_1+m_2)\int_0^{t_2}\sigma\delta u(\sigma)\, d\sigma.
\end{aligned} \tag{7.12}$$

Change of the order of integration is used from line 3 to line 4 in (7.12). We get the constraint equation for δu

$$-2m_1m_2a + \frac{1}{2}\mu(m_1+m_2)m_1gt_2^2 = (m_1+m_2)\int_0^{t_2} t\delta u\, dt. \tag{7.13}$$

In order to find a δu to minimize the cost function $\eta\prime$ with the above constraints (7.9) and (7.13), we need the following lemma.

Lemma 7.3.1. *Given constants T and M, and function $\alpha(t) \not\equiv$ constant, the solution to the following optimization problem*

$$\min_u \eta = \int_0^T u^2\, dt$$

s. t.

$$\int_0^T u\, dt = 0, \quad \textit{and} \quad \int_0^T \alpha(t) u\, dt = M, \tag{7.14}$$

is

$$u^\star = \frac{M(\alpha(t) - \frac{1}{T}\int_0^T \alpha(\sigma)\, d\sigma)}{\int_0^T \alpha^2(\sigma)\, d\sigma - \frac{1}{T}\left(\int_0^T \alpha(\sigma)\, d\sigma\right)^2} \tag{7.15}$$

with minimal cost

$$\eta = \frac{M^2}{\int_0^T \alpha^2(\sigma)\,d\sigma - \frac{1}{T}\left(\int_0^T \alpha(\sigma)\,d\sigma\right)^2}. \tag{7.16}$$

Proof. It follows from the Cauchy-Schwartz inequality. Let α_0 be an arbitrary constant. From the constraints equations, we then have

$$M = \int_0^T (\alpha(t) + \alpha_0)u\,dt \le \left(\int_0^T (\alpha(t) + \alpha_0)^2\,dt \int_0^T u^2\,dt\right)^{\frac{1}{2}}. \tag{7.17}$$

Then

$$\int_0^T u^2\,dt \ge \frac{M^2}{\int_0^T (\alpha(t) + \alpha_0)^2\,dt}. \tag{7.18}$$

Equality holds in (7.18) if and only if

$$u = \beta(\alpha(t) + \alpha_0) \tag{7.19}$$

where β is a constant. To satisfy the first constraint in (7.14), we have

$$\int_0^T \beta(\alpha(t) + \alpha_0)\,dt = 0, \tag{7.20}$$

then

$$\alpha_0 = -\frac{1}{T}\int_0^T \alpha(\sigma)\,d\sigma. \tag{7.21}$$

Plug α_0 into the inequality for the cost function (7.18), we have the minimal cost. And (7.19) becomes

$$u = \beta(\alpha(t) - \frac{1}{T}\int_0^T \alpha(\sigma)\,d\sigma). \tag{7.22}$$

To satisfy the second constraint in (7.14), we have

$$\beta\int_0^T \alpha(t)(\alpha(t) - \frac{1}{T}\int_0^T \alpha(\sigma)\,d\sigma)\,dt = M. \tag{7.23}$$

It gives us

$$\beta = \frac{M}{\int_0^T \alpha^2(\sigma)\,d\sigma - \frac{1}{T}\left(\int_0^T \alpha(\sigma)\,d\sigma\right)^2}. \tag{7.24}$$

This gives us the minimal u. ∎

Note that the variation in function $\alpha(t)$ gives us the opportunity to minimize the total control energy. We identify T with terminal time t_2 of the fast mode, $\alpha(t)$ with t, and M with $-\frac{2m_1m_2a}{m_1+m_2}+\frac{1}{2}\mu m_1 g t_2^2$. Then the minimal energy solution for the fast mode becomes

$$\begin{aligned}\delta u^\star &= \frac{M(t-\frac{1}{t_2}\int_0^{t_2}\sigma\,d\sigma)}{\int_0^{t_2}\sigma^2\,d\sigma-\frac{1}{t_2}\left(\int_0^{t_2}\sigma\,d\sigma\right)^2} \\ &= 3(2t-t_2)(-\tfrac{4m_1m_2a}{(m_1+m_2)t_2^3}+\tfrac{\mu m_1 g}{t_2}),\end{aligned} \tag{7.25}$$

$$\begin{aligned}u^\star &= \mu(m_1+m_2)g+\delta u^\star \\ &= \mu(m_1+m_2)g+3(2t-t_2)(-\tfrac{4m_1m_2a}{(m_1+m_2)t_2^3}+\tfrac{\mu m_1 g}{t_2}).\end{aligned} \tag{7.26}$$

The corresponding minimal cost is

$$\begin{aligned}\eta^\star &= \mu^2(m_1+m_2)^2g^2t_2+\frac{M^2}{\int_0^{t_2}\sigma^2\,d\sigma-\frac{1}{t_2}\left(\int_0^{t_2}\sigma\,d\sigma\right)^2} \\ &= \mu^2(m_1+m_2)^2g^2t_2+\tfrac{12}{t_2^3}(-\tfrac{2m_1m_2a}{m_1+m_2}+\tfrac{1}{2}\mu m_1 g t_2^2)^2 \\ &= \mu^2(4m_1^2+2m_1m_2+m_2^2)g^2t_2-\tfrac{24\mu m_1^2m_2ga}{(m_1+m_2)t_2} \\ &\quad+\tfrac{48m_1^2m_2^2a^2}{(m_1+m_2)^2t_2^3}.\end{aligned} \tag{7.27}$$

In order to increase the energy efficiency, we can minimize energy usage per unit displacement by the selection of the terminal time t_2

$$\begin{aligned}&\min_{t_2}\eta_1=\tfrac{\eta}{x_1(t_2)} \\ &=\frac{\mu^2(4m_1^2+2m_1m_2+m_2^2)g^2t_2-\frac{24\mu m_1^2m_2ga}{(m_1+m_2)t_2}+\frac{48m_1^2m_2^2a^2}{(m_1+m_2)^2t_2^3}}{\frac{2m_2a}{m_1+m_2}-\frac{1}{2}\mu g t_2^2}\end{aligned} \tag{7.28}$$

We need to solve $\frac{d\eta_1}{dt_2}=0$ for t_2, numerically. By observation, η_1 goes to infinity when t_2 goes to 0 and ∞. It can be explained as follows. If the fast mode is too fast, we need to apply a large input which results in large average energy consumption. If the fast mode is too slow, energy is consumed by the external friction force, thus results in large average energy consumption as well.

In step 3 and 4, $\dot{x}_1=0$, we can apply force between m_1 and m_2 to bring m_2 back to its position relative to m_1 at the beginning of step 1, i. e. , $x_2(t_4)-x_2(t_2)=2a$, where t_4 is the end time of step 4. And in order to keep m_1 static the magnitude of the force should be lower than the maximal static friction force between m_1 and the tube, i. e. $\tau\le\mu(m_1+m_2)g$. Here we make no distinction between the static friction coefficient and the sliding friction coefficient, though they are slightly different from each other. We can work out the minimal energy

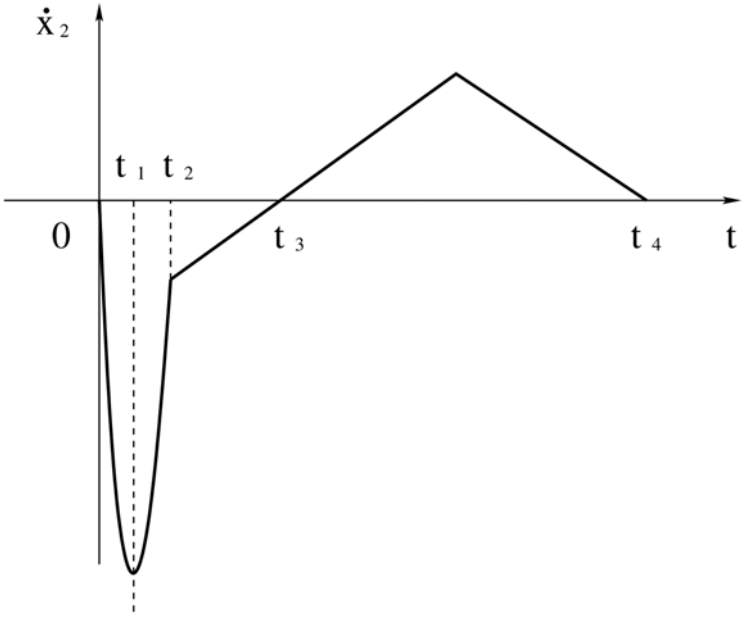


Fig. 7.3. Diagram of the Velocity Profile of x_2

control for the slow mode, but it makes more sense to minimize the time spent on step 3 and 4, i. e. , $\min(t_4 - t_2)$, because the magnitude of control is rather small comparing with that of fast mode. According to Pontryagin's maximal principle, the force applied is a bang-bang control, i. e. positive maximal static friction force followed by negative maximal friction force as show in Figure 7.3, with horizontal axis to be time t, vertical axis velocity $\dot{x}_2$ of cylinder m_2 in its external fixed frame, and t_i to be the terminal time of step i.

Let β be the magnitude of the maximal acceleration applied to m_2. Then $\beta = \frac{\mu(m_1+m_2)g}{m_2}$. And

$$t_3 - t_2 = -\frac{\dot{x}_2(t_2)}{\beta} = t_2, \qquad t_3 = 2t_2. \tag{7.29}$$

Time spent on step 4 satisfies equation

$$\frac{1}{4}\beta(t_4 - t_3)^2 - \frac{1}{2}\beta(t_3 - t_2)^2 = 2a \tag{7.30}$$

$$t_4 = 2t_2 + \sqrt{2t_2^2 + \frac{8a}{\beta}} = 2t_2 + \sqrt{2t_2^2 + \frac{8m_2a}{\mu(m_1+m_2)g}} \tag{7.31}$$

When $t_2 \longrightarrow 0$ we have $t_4 \longrightarrow \sqrt{\frac{8m_2a}{\mu(m_1+m_2)g}}$ and $x_1(t_2) \longrightarrow \frac{2m_2a}{m_1+m_2}$. Thus the average velocity of the capsule

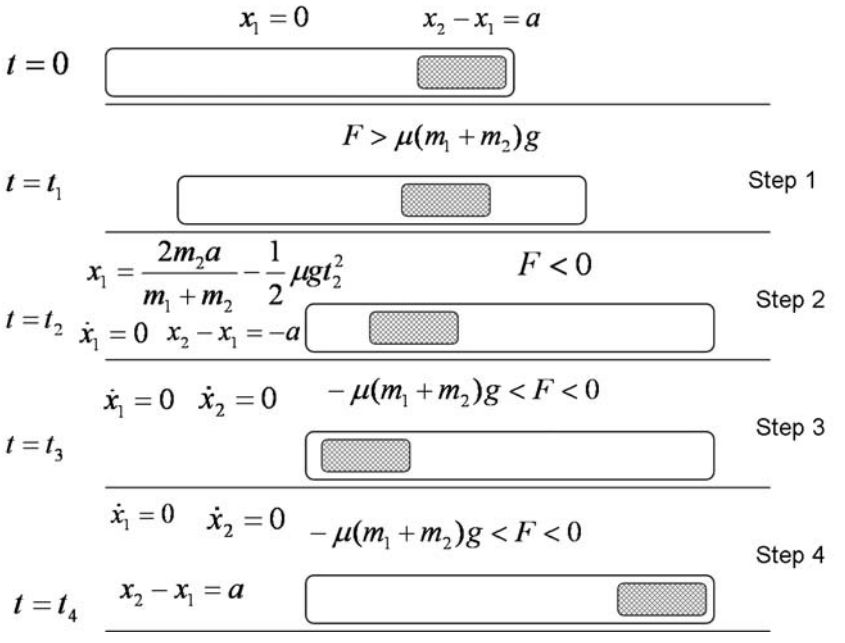


Fig. 7.4. Diagram of the 4 Step Motion

$$\text{ave}(x_1) = \frac{x_1(t_2)}{t_4} \longrightarrow \sqrt{\frac{\mu m_2 g a}{2(m_1 + m_2)}}. \tag{7.32}$$

The complete 4 step motion generation is shown in figure 7.4. Time frames are from up to down. And the motion direction is from left to right. Note that m_1 has to stop first and then slow down m_2 using its static friction in the reverse direction. In order to increase the average velocity of m_1, the ratio $m_1 : m_2$ should as small as possible.

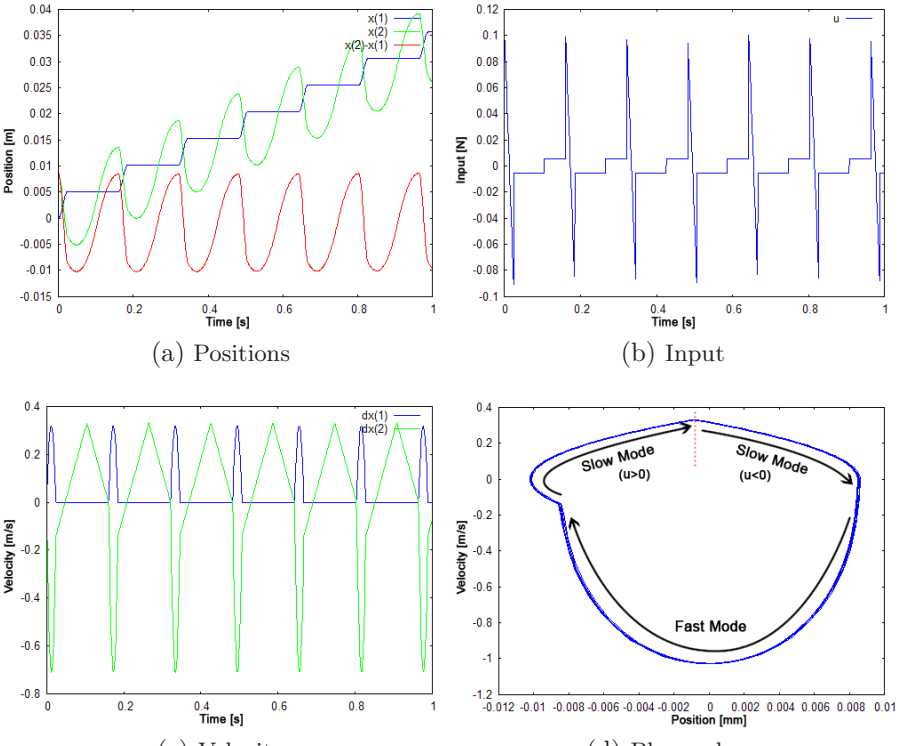
7.4 Simulation

Up to now, the input u has been constructed by optimization. Additionally, we have studied the collision test through simulation by assuming that the inside mass collides with the end of the Capsubot when it comes back. Table 7.1 shows the parameters which are used in the simulation. These parameters are the same as the experimental parameters.

Figure 7.5 shows the results of the simulation. It is confirmed that the capsule moves forward only when the cylinder is accelerated backward in fast mode. However, the capsule stops when the input u is smaller than friction force while the internal mass comes back to original position in slow mode. By doing this repeatedly, the Capsubot can go forward and go backward by changing the steps. From simulation, it is proved that it is proper to do the operation when t_2 is

Table 7.1. Parameters of the Capsubot

m_1	mass of the capsule	0. 0018 kg
m_2	mass of the cylinder	0. 0009 kg
g	gravity acceleration	9. 81 m/s^2
μ	friction coefficient	0. 24
a	half length of the capsule	0. 0085m



(a) Positions (b) Input

(c) Velocity (d) Phase plane

Fig. 7.5. Simulation Results of Optimal Model

0. 024 (s). t_2 and the friction coefficient change by corresponding to a weight of the whole Capsubot and also the friction coefficient. Here, friction coefficient is set as 0. 24. It is obtained through several trials by observing the movement of m_2. It can be confirmed from Figure 7.5(d) which shows the phase plane. Also, it can be confirmed that it is divided into 2 parts–the fast mode and slow mode. The conditions for the two modes are already mentioned before in Figure 7.2. If we could know the friction coefficient, the Capsubot can be controlled with the position information.

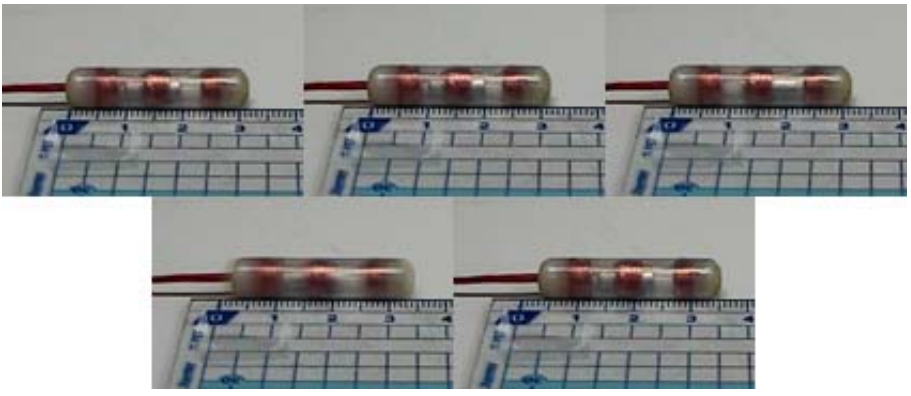


Fig. 7.6. One Step of Capsubot

7.5 Experiments

The Capsubot design in the paper comprises: A cylindrical body(ϕ8mm×L27mm, mass=1. 8gram)having 3 sets coils wound peripherally, a magnet (ϕ4mm× L10mm, mass=0. 9gram) that can move within a predetermined range in the body, and drive means for driving the magnet in a moving direction by energizing the coil, received outside the moving range of the magnet in the body. A driving circuit is designed as in Figure 6 with 3 sets of H-bridge chips are switched on and off, changed current direction through a parallel port of a desktop PC running a Visual C++ driver program under MS windows XP OS. Voltage applied to the coil, using a pulse rate modulation, to drive the magnet of above mentioned four steps repeatedly. The number of pulses in unit time corresponding magnitude of the voltage input to the coil, as shown in Figure 7, where the horizontal axis is time with unit millisecond and the vertical axis to be input voltage in volt.

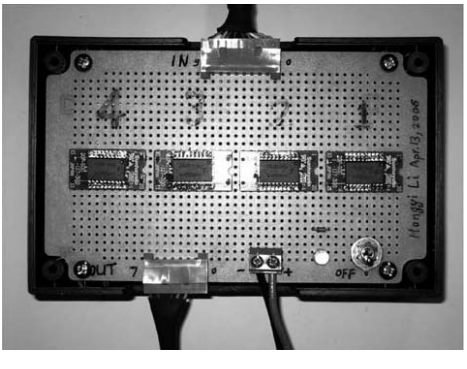


Fig. 7.7. Digital Drive Circuit

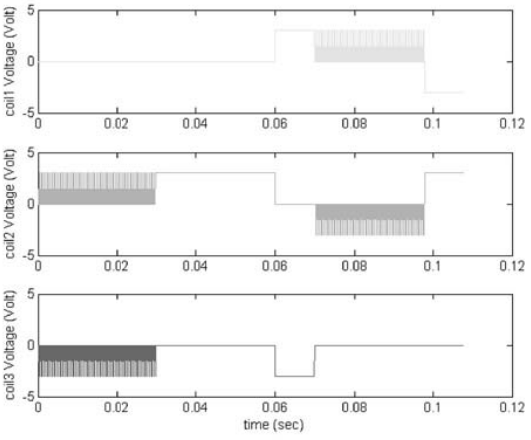


Fig. 7.8. Pulses of the 4 step motion

In experiments, the Capsubot advanced 197 mm in 50 strokes on average, i. e. , 3. 9 mm per stroke(108ms), 36. 5 mm per second, as shown in figure 9. Note that there is not a position sensor for the cylinder inside the capsule shell, only open loop optimal control is implemented here. The control input may cumulate phase error with the real relative position of the internal cylinder with respect to the shell. In order to increase robustness, we deliberately added a few extra pulses at the end of the slow mode. Then the cylinder clash with the shell slightly to make sure that it will return to its initial relative position inside the shell even under some disturbances and inaccurate inputs, voltages measured from ends of the coil after this elementary robustness consideration is shown in figure 8 with horizontal unite to be 25 milliseconds, vertical unit 1 volt. Induction of the coil smoothes the pulse input from the computer end. We observes that the Capsubot

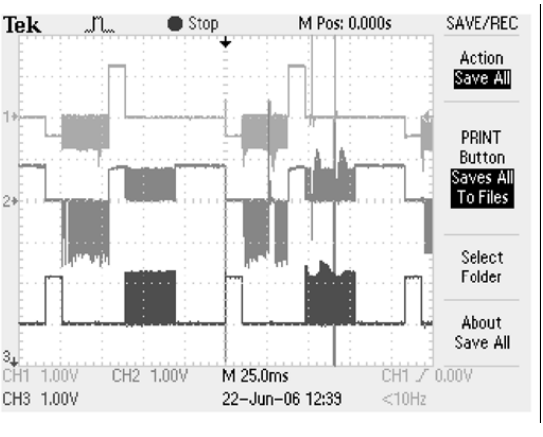


Fig. 7.9. Measured input voltage from the coils

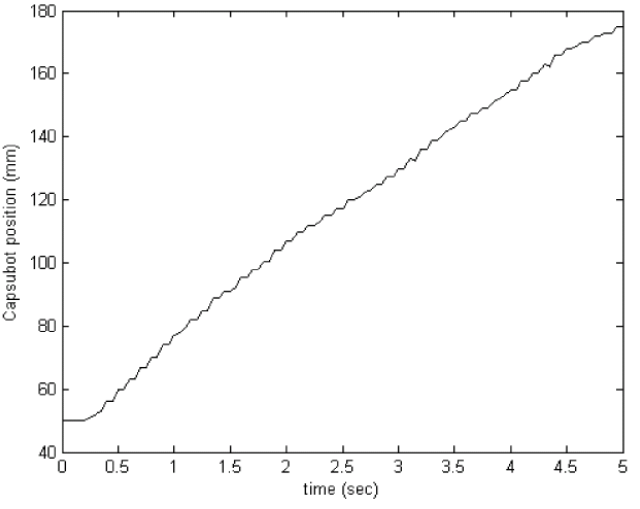


Fig. 7.10. Plot of Capsubot position

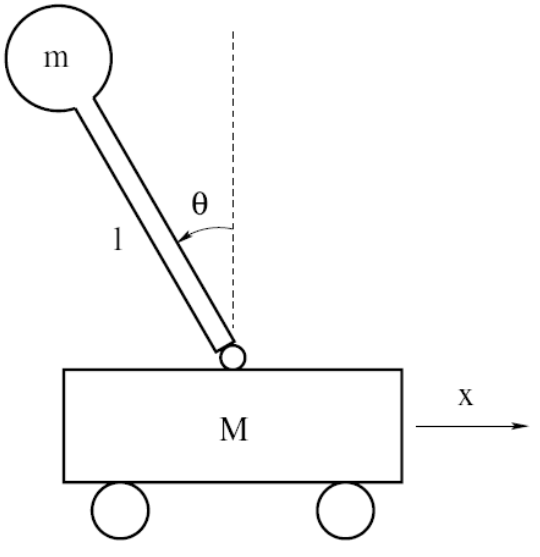


Fig. 7.11. Diagram of the Pendulum Cart

drifts away slightly from its original direction if unconstrained. It is possible to control the spatial direction in the horizontal plane by bundling two Capsubot with independent dives.

Since the capsule robot has no external moving parts and can be made small in size for inserting into a human, it has a large applicability in the medical field, for example, as a self-propelled endoscope. Otherwise, it has good utility in the

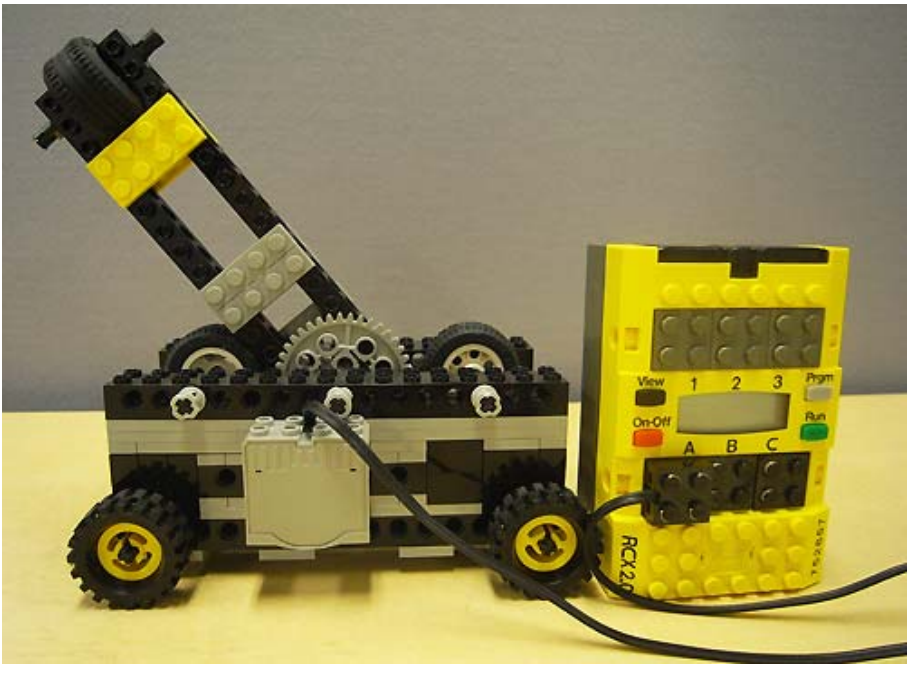


Fig. 7.12. Pendulum Cart System

field where micro robot is needed such as in the case of examining the inside of a narrow conduit or a tube in a factory or laboratory. The test of its utility inside animal body is underway and therefore is not included in this paper.

7.6 Application to the Pendulum Cart

As an example of the theoretical-based experiment, there is the pendulum driven cart. It will help to understand the principle of the Capsubot's motion generation. A diagram of a simplified model of the pendulum cart system is shown in Figure 7.11. A pendulum is mounted on the top of a cart. The cart has four passive wheels which made it possible to move horizontally on the ground with relatively low driving force to counteract the friction force between the cart and the ground. A servo motor is directly mounted to the joint between the pendulum and the cart to swing the pendulum. Let M be the mass of the cart, m the mass of the pendulum, l the distance between the joint and the center of mass of the pendulum, I the moment of inertia of the pendulum, μ the coefficient of friction between the cart and the ground, θ the angle between the pendulum and the vertical, x the coordinate measuring the displacement of the cart relative to a fixed reference frame, and τ the torque applied to the joint by the motor attached to it. Assume no friction between the pendulum and the cart (good ball bearing). To move the cart forward, the required motion of the pendulum consists of four steps:

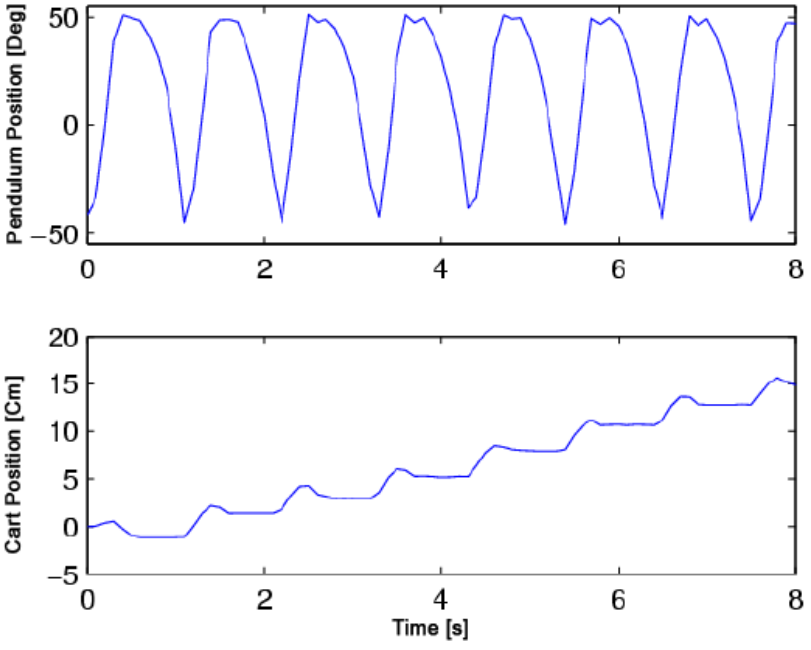


Fig. 7.13. Trajectory of the Pendulum Cart System

1. Counterclockwise motion with high angular acceleration of the pendulum m ($\ddot{\theta} \ll 0$), forward accelerated motion of the cart M ($\ddot{x} > 0$).
2. Counterclockwise motion with low angular acceleration of the pendulum m ($\ddot{\theta} < 0$), forward decelerated motion of the cart M ($\ddot{x} < 0$).
3. Counterclockwise motion with low angular acceleration of the pendulum m ($\ddot{\theta} < 0$), the cart M remains stationary ($\dot{x} = 0$).
4. Clockwise motion with low angular acceleration of the pendulum m ($\ddot{\theta} \leq -\varepsilon$), the cart M remains stationary ($\dot{x} = 0$).

Figure 7.12 shows the Pendulum cart which is made by using Lego Mindstorm 2. 0. It consists of a cart with four passive wheels, a pendulum mounted on top of the cart which could move forward and backward, a DC motor is attached to the hinge joint between that pendulum and the cart, and a RCX controller connected to the motor. The RCX controller is programmed to drive the motor in this experiment.

As comparing with 4 steps of Capsubot, we can see they have the same pattern. By repeating the 4 steps, we get the trajectory of the pendulum cart from the experiment as in Figure 7.14.

A sequence of images from a video clip constitutes a full stoke, from left to right and top to bottom, is shown in Figure 7.14.

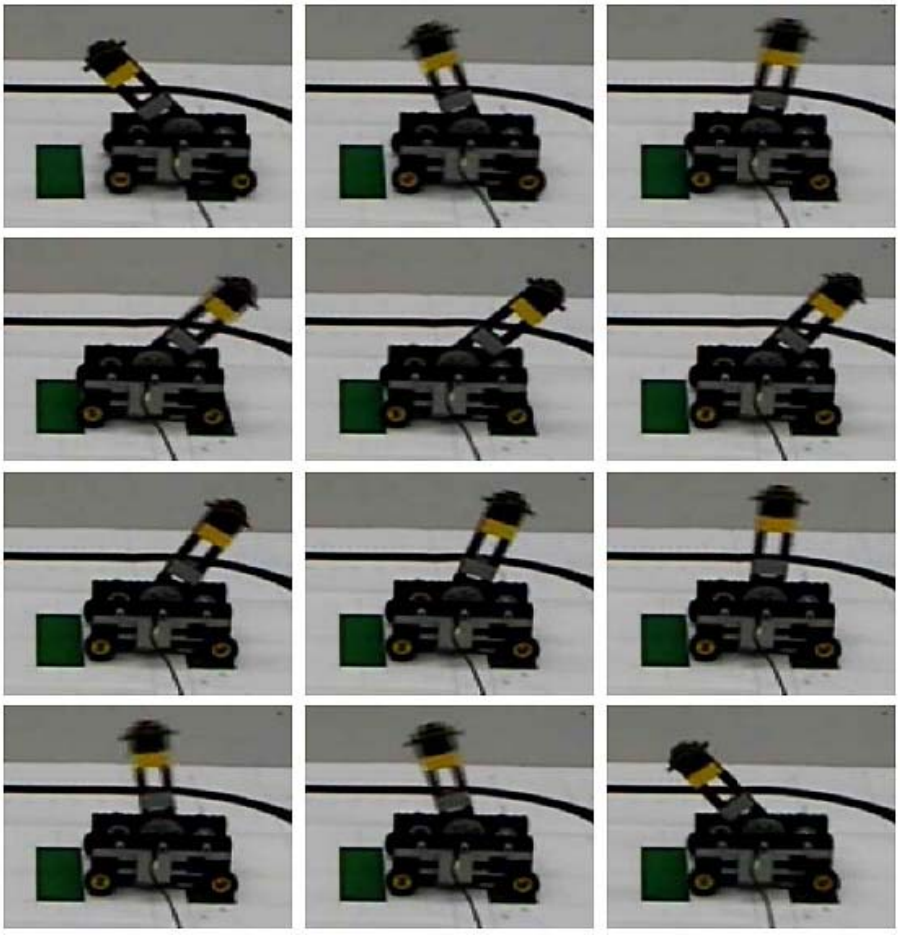


Fig. 7.14. Sequence of the Full Stroke

7.7 Conclusions

A one dimensional motion generation strategy for a capsule-type locomotive system using an internal moving mass is proposed. The control input is determined using an optimal control formulation under the assumption that static friction does not dominate during the first half of the control trajectory (fast mode), but does during the second half (slow mode). By moving the internal mass appropriately, one dimensional motion is achieved. It is then applied to an experimental setup (the Capsubot). Further work will be on real time friction coefficient identification and its experiments inside an animal digestive track.

References

1. Bryson Jr., A.E., Ho, Y.C.: Applied optimal control. Hemisphere Publishing Corporation, New York (1975)
2. Chernous'ko, F.L.: The optimun rectilinear motion of a two-mass system. Journal of Applied Mathematics and Mechanics 66, 1–7 (2002)
3. Kim, B., Lee, S., Park, J.H., Park, J.O.: Design and fabrication of a locomotive mechanism for capsule-type endoscopes using shape memory alloys(SMAs). IEEE/ASME Tran. Mecharonics 10, 77–86 (2005)
4. Sendoh, M., Ishiyama, K., Arai, K.I.: Fabrication of magnetic actuator for use in a capsule endoscope. IEEE Tran. Magnetics 39, 3232–3234 (2003)
5. Li, H., Furuta, K., Chernousko, F.L.: A pendulum-driven cart via internal force and static friction. In: Proc. of Conf. Physics and Control, St. Petersburg, Russia, pp. 15–17 (2005)
6. Higuchi, T., Yamagata, Y., Furutani, K., Kudoh, K.: Precise positioning mechanism utilizing rapid deformations of piezoelectric elements. In: Proc. of IEEE Workshop on Mexico Electro Mechanical Systems, pp. 47–49 (1990)
7. Darby, A.P., Pellegrino, S.: Inertial stick-slip actuator for active control of shape and vibrarion. J. of Intelligent Material systems and Structures 8, 1001–1011 (1997)
8. Chernousko, F.L.: On the motion of a body containing a movable internal mass. Doklady Physics 50, 593–597 (2005)
9. Whittaker, E.T.: A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, 4th edn. Cambridge University Press, Cambridge (1937)
10. Li, H., Furuta, K., Chernousko, F.L.: Motion generation of a Capsubot using internal force and static friction. In: Proc. of IEEE Conf. Decision and Control, San Diego, CA, USA, pp. 6575–6580 (2006)

8

On the Topology of Liapunov Functions for Dissipative Periodic Processes

Christopher I. Byrnes

Department of Electrical and Systems Engineering, Washington University, USA

Dedicated to the memory of my friend and treasured colleague, Wijesura Dayawansa.

Summary. The existence and nature of nonlinear oscillations for periodically forced nonlinear differential equations has historically attracted quite a bit of attention in both the pure and the applied mathematics literature. In control theory, it encompasses the study of the steady-state response of control systems to periodic inputs, generalizing the frequency domain theory that underlies classical control and its many successes. More than fifty years ago, Levinson initiated the study of dissipative periodic processes for planar systems, an approach that has since inspired the development of a general theory of dissipative systems for both lumped and distributed nonlinear systems. In the lumped case, dissipative processes have a dissipative Poincaré map $\mathcal{P}$ and a fair amount of effort has been expended determining the fixed point properties of $\mathcal{P}$, culminating in the use of a remarkable fixed point theorem of F. Browder which showed that general dissipative periodic processes always have harmonic oscillations. An alternative approach to studying dissipative periodic processes using Liapunov theory was developed by the Russian school of nonlinear analysis, pioneered by Pliss, Krasnosel'skiĭ and others. It is fair to say that the largest technical challenge arising in this approach is the lack of a general, user-friendly description of the the level and sublevel sets of these Liapunov functions. In the equilibrium case, the recent solution of the Poincaré Conjecture in all dimensions has resulted in a simple description and useful description of these sets [3], viz. the sublevel sets are always homeomorphic to a disk $\mathbb{D}^n$. Fortunately, the techniques underlying the proofs of the Poincaré conjectures have shed enough light on related classification questions that we can now also describe the topology of the level and sublevel sets of Liapunov functions for dissipative periodic process. Among the results we prove in this paper is that these sublevel sets of a Liapunov function are always homeomorphic to solid tori, $\mathbb{D}^n \times S^1$, and diffeomorphic except perhaps when $n = 3$. Together with recent sufficient conditions for periodic orbits proven by Brockett and the author [4], these descriptions give streamlined proofs of the existence of harmonic oscillations, and some related results. The proof of our main theorem uses the work of Wilson on the topology of Liapunov functions for attractors, the s-cobordism theorem in dimensions greater than five, the validity of the Poincaré Conjecture in dimension three and four, and a smoothing result of Kirby and Siebenmann for five-manifolds.

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 125–139.
springerlink.com

8.1 Introduction

Consider a $\mathcal{C}^\infty$ periodically time-varying ordinary differential equation

$$\dot{x} = f(x,t), \quad f(x,t+T) = f(x,t). \tag{8.1}$$

evolving on $\mathbb{R}^n$. (8.1) defines a time-varying vector field $X \in \mathrm{Vect}(\mathbb{R}^n \times \mathbb{R})$. As a very notable example, the periodically forced van der Pol oscillator

$$\dot{x} = y, \;\; \dot{y} = -x + \mu(1-x^2)y + \alpha p(t), \tag{8.2}$$

where $p(t+T) = p(t)$, was studied classically by van der Pol [38], Cartwright [6], Cartwright and Littlewood [7], Levinson [23] and, in a definitive way, by Holmes and Rand [17], [12]. More generally, we can consider a periodic system (8.1) evolving on $\mathbb{R}^n$ that arises as an additive perturbation of an autonomous system

$$\dot{x} = f(x) + \epsilon p(x,t), \;\; \text{where } p(t+T) = p(t). \tag{8.3}$$

Historically, a central question concerning periodic systems is whether there exists an initial condition $(x_0, 0)$ generating a periodic solution having period T. Such solutions are called "harmonic" solutions. Following the pioneering work of Levinson [22] on dissipative forced systems in the plane, Pliss [30] formulated a general class of periodic vector fields on $\mathbb{R}^n$:

Definition 8.1.1. *The periodic differential equation* (8.1) *is* dissipative *provided there exists* $R > 0$ *such that*

$$\overline{\lim_{t\to\infty}} \|x(t; x_0, t_0)\| < R \tag{8.4}$$

Dissipative systems are also sometimes referred to as being "ultimately bounded".

Remark 8.1.1. The theory of dissipative systems has studied by many authors. Indeed, dissipative systems play a central role in the work of Krasnosel'skiĭ [20], Hale [13], Sell [31] and others for both lumped and distributed parameter systems.

For a dissipative system, we can define a $\mathcal{C}^\infty$ Poincaré map

$$\mathcal{P} : \mathbb{R}^n \to \mathbb{R}^n, \;\; \text{defined via } \; \mathcal{P}(x_0) = x(T; x_0, 0). \tag{8.5}$$

An important consequence [30] of (8.4) is that there exists a closed ball $0 \in B \subset \mathbb{R}^n$ and an $r \in \mathbb{N}$ such that if $x_0 \in B$ then $x(t; x_0, 0) \in B$ for $t \geq rT$; i.e.,

$$\mathcal{P}^r : B \to B. \tag{8.6}$$

Therefore, applying Brouwer's Fixed Point Theorem to $\mathcal{P}^r$ yields the existence of *subharmonic* forced oscillations.

Proposition 8.1.1 (Pliss). *A dissipative system on $\mathbb{R}^n$ has a periodic orbit of period rT, where $r \in \mathbb{N}$.*

According to [30, Theorems 2.5 - 2.6], a vector field X is dissipative if and only if there exists a $\mathcal{C}^\infty$ Lyapunov function

$$V : \mathbb{R}^n \to \mathbb{R} \tag{8.7}$$

and a constant $\rho > 0$ such that:

1. $V(x,t) = V(x,t+T)$,
2. $V(x,t) > 0$ whenever $||x|| > \rho$,
3. $V(x,t) \to +\infty$ for $||x|| \to +\infty$, uniformly in $t \in [0,T]$.
4. $\langle \nabla V, X \rangle < 0$ for $||x|| > \rho$.

Remark 8.1.2. If the unperturbed system

$$\dot{x} = f(x) \tag{8.8}$$

has a compact global attractor $\mathcal{A} \subset \mathbb{R}^n$, then there exists an $\epsilon_{\mathcal{A}} << \infty$ such that for $0 < \epsilon < \epsilon_{\mathcal{A}}$ (8.3) is dissipative. Indeed, there exists [39] -[40] a Liapunov function W for $(f, \mathcal{A})$ and setting

$$V(x,t) = W(x) \tag{8.9}$$

we see that

$$\dot{V}(x) = \langle \nabla W, f \rangle(x) + \epsilon \langle \nabla W, p \rangle(x,t) < 0 \tag{8.10}$$

for

$$W(x) = \rho \geq \text{dist}(0, \mathcal{A}) \text{ and } 0 < \epsilon < \epsilon_{\mathcal{A}} << \infty.$$

In particular, V satisfies conditions (1)-(4) for $0 < \epsilon < \epsilon_{\mathcal{A}}$.

Krasnosel'skiĭ and Perov [20] sharpened Proposition 8.1.1 to obtain the existence of periodic orbits having period T, provided (8.1) has a Liapunov function V satisfying (8.9). Their proof involved a clever sequence of deformations to compute the index of the "translational field"

$$\Phi(t,s)x_0 = x_0 - x(t; x_0, s) \text{ for each } x_0 \in \mathbb{R}^n$$

using the Poincaré-Hopf Index Theorem [26]. In Section 8.3, we give an alternative proof that shows that the Poincaré map

$$\mathcal{P} : W^{-1}(-\infty, c] \to W^{-1}(-\infty, c], \text{ for } c > \rho, \tag{8.11}$$

has a fixed point.

Convention 8.1.1. By smooth, we mean $\mathcal{C}^\infty$. For a smooth n-manifold M with boundary, ∂M is a smooth $(n-1)$-manifold [16]. We denote the *interior*, $M - \partial M$, of M by $\mathring{M}$. For smooth manifolds , we shall write $M \simeq N$ when M and N are diffeomorphic, expressing the situation where M and N are homeomorphic by $M \simeq_h N$.

For example, (8.11) defines a smooth map defined on a smooth manifold with boundary. Using some fundamental results of Wilson [39], the validity of the Poincaré Hypothesis in all dimensions and theorems of B. Mazur [24] and Smale ([33]-[34]), in Section 3 we show:

Proposition 8.1.2. *For any $c > \rho$,*

$$W^{-1}(-\infty, c] \simeq_h \mathbb{D}^n \tag{8.12}$$

In fact, $W^{-1}(-\infty, c] \simeq \mathbb{D}^n$, except perhaps when $n = 4$.

In particular, the Brouwer Fixed Point Theorem applies. In Section 3, we also derive a corollary asserting the existence of periodic orbits for systems with compact attractors and illustrate this result for the van der Pol oscillator with small periodic forcing.

In the general case, there is an enhancement of (8.11). Setting notation, for $t_0 \in \mathbb{R}$, we define

$$W(t_0) : \mathbb{R}^n \to \mathbb{R} \quad \text{via} \quad W(t_0)(x) = V(x, t_0)$$

and consider

$$M(t_0) = W(t_0)^{-1}(-\infty, c], \quad \text{for} \quad c > \rho. \tag{8.13}$$

Then,

$$\mathcal{P}_{t_0} : M(t_0) \to M(t_0), \quad \text{where} \quad \mathcal{P}_{t_0}(x_0) = x(T + t_0; x_0, t_0). \tag{8.14}$$

By Sard's Theorem [37], it follows (see Section 8.2) that, for almost all t_0, $M(t_0)$ is smooth compact manifold with boundary, and hence an aboslute neighborhood retract, which can be shown to be connected using the results of Sections 3 and 5. Choosing any regular value t_0, the argument [30] underlying (8.6) shows that there exists B and an $r \in \mathbb{N}$ such that

$$\mathcal{P}_{t_0}^r : M(t_0) \to B \subset M(t_0) \tag{8.15}$$

In this case, a remarkable theorem due to F. Browder [1],[41] implies that (8.14) always has a fixed point.

Remark 8.1.3. While the theorem in [1] is a result in convex nonlinear analysis, Browder [2]l developed a more general topological approach for maps on infinite dimensional manifolds. In our context, this can be described for a dissipative system (8.1) with Liapunov function V as follows. If there exists a ball $B \subset \mathbb{R}^n$ such that $\mathcal{P} : M(t_0) \to B \subset M(t_0)$ then one might say that $\mathcal{P}|_M(t_0)$ is *topologically trivial*, since

$$\mathcal{P}_*^k : H_k(M) \to H_k(M)$$

satisfies

$$\mathcal{P}_*^k = 0 \text{ for } k \geq 1.$$

On the other hand, (8.15) could be described by saying $\mathcal{P}|_M$ is then one might interpret (8.6) as saying that $\mathcal{P}$ is *topologically nilpotent*, in the sense that the linear transformation $\mathcal{P}_*^k$ is nilpotent for $k \geq 1$ and therefore

$$\text{trace}(\mathcal{P}_*^k) = 0 \text{ for } k \geq 1.$$

In this sense, Browder's Theorem refines Brouwer's Theorem by asserting that topologically nilpotent maps have a fixed point. Indeed by the Lefschetz Fixed Point Theorem,

$$\text{Lef}(\mathcal{P}) = \sum_{k=0}^{k=n+1} \text{trace}(\mathcal{P}_* : H_k(M(t_0)) \to H_k(M(t_0))) = 1 \neq 0,$$

so that $\mathcal{P}$ has a fixed point.

In Section 8.4, we prove several basic facts about the topology of the sublevel sets $V^{-1}(-\infty, c]$ for general dissipative periodic processes. Among the consequences of our main result, Theorem 8.2.1, and the results of [4] is that the map (8.14) always has a fixed point.

Theorem 8.1.1. *For almost all t_0, $M(t_0)$ is a smooth submanifold. If $n \neq 3$,*

$$M(t_0) \simeq \mathbb{D}^n \tag{8.16}$$

and therefore (8.14) *has a fixed point for every t_0. In particular, for all n, a dissipative system* (8.1) *always has a harmonic solution. Moreover,* $\text{Ind}(\mathcal{P}_{t_0}) = 1$ *and therefore:*

1. *If every periodic orbit of period T is hyperbolic, then there is an odd number number of such periodic orbits and*

$$\sum_{\gamma} \text{sign det}(I - D\mathcal{P}_\gamma) = 1. \tag{8.17}$$

2. *If each harmonic orbit is locally asymptotically stable, then there exists just one perioidic orbit having period T.*

Remark 8.1.4. If $n = 3$, we may augment (8.1) with the equation $\dot{x}_4 = -x_4$ and obtain the desired conclusions.

In Section 8.6, Theorem 8.1.1 is illustrated in an explicit example. If (8.9) is satisfied, the first assertion of the theorem is just (8.12). In the general case, Theorem 8.1.1 is an application of the existence theorem and the necessary conditions derived in [4], together with our main result, Theorem 8.2.1. In particular, the proof of Theorem 8.1.1 given in Section 8.2 uses our main result Theorem 8.2.1 and the more general existence results of [4]. The proof of Theorem 8.2.1 given in Section 8.4 is similar in technique to the proof of the Theorem 1.2 in [4], using the work of Wilson on the topology of Liapunov functions for attractors, the s-cobordism theorem in dimensions greater than five, the validity of

the Poincaré Conjecture in dimension three and four, and a smoothing result of Kirby and Siebenmann for five-manifolds. While the conclusion, and the techniques used in the proof, of Theorem 8.2.1 are similar to those for Theorem 1.2 in [4], the proofs are different and the hypotheses are dramatically distinct, as elementary examples show.

In closing, it is a pleasure to thank Roger Brockett, Tom Farrell, David Gilliam, John Morgan, Ron Stern and Shmuel Weinberger for helpful correspondence and suggestions.

8.2 The Main Results

Following [30], we note that the augmented equation

$$\dot{x} = f(x, \tau), \quad \dot{\tau} = 1. \tag{8.18}$$

defines an autonomous vector field X_a on the "toroidal cylinder", $N = \mathbb{R}^n \times S^1$. If (8.1), there exists a periodic Liapunov function (8.7) which defines a smooth map V_a on N. Setting $M = V_a^{-1}(-\infty, c] \subset N$, for $c > \rho$ we note that $M \subset N$ is a smooth *compact* submanifold with boundary.

Turning to the proof of Theorem 8.1.1, consider the smooth mapping

$$\widetilde{J} : V^{-1}(-\infty, c] \to \mathbb{R} \tag{8.19}$$

defined via $\widetilde{J}(x, t) = t$. By Sard's Theorem[37], $\widetilde{J}^{-1}(t_0) = M(t_0)$ is a smooth manifold with boundary, for almost all $t_0 \in \mathbb{R}$. $\widetilde{J}$ descends to a map

$$J : M \to S^1 \tag{8.20}$$

for which $\widetilde{J}^{-1}(t_0) \simeq J^{-1}(\tau_0)$ for an regular value t_0 of $\widetilde{J}$ and where $\tau_0 \in S^1 = \mathbb{R} \bmod \mathbb{Z}$ satisfies $\tau_0 = [t_0]$. In particular, $M(t_0)$ is a smooth *compact* manifold with boundary. We will prove that, for $n \neq 3$, $M(t_0) \simeq \mathbb{D}^n$.

Theorem 8.2.1. *Consider a dissipative system* (8.1) *with a smooth Liapunov function V. The smooth compact submanifold M is positively invariant under X_a and $M \simeq_h \mathbb{D}^n \times S^1$. Moreover, $M \simeq \mathbb{D}^n \times S^1$, except perhaps when $n = 3$.*

When V has the form (8.9), the topological conclusions of Theorem 8.2.1 follows from Theorem 8.12, but the conclusions concerning the differentiable structure of M vary when $n = 3$ or 4, reflecting the different methods of proof. For the sake of clarity, we collect these results as follows.

Corollary 8.2.1. *For a dissipative system* (8.1) *with a smooth Liapunov function V,*

$$\partial M \simeq S^{n-1} \times S^1, \text{ for all } n.$$

If V satisfies (8.9)*, then*

$$M \simeq \mathbb{D}^n \times S^1, \text{ for all } n.$$

Remark 8.2.1. In particular, for $n = 1$, M is diffeomorphic to the annulus $\mathbb{A} \subset \mathbb{R}^2$. In this case, X_a has a periodic orbit by the Poincaré-Bendixson Theorem. We are interested in an enhancement of this situation for $n \geq 2$.

1. If $n = 2$, then M is diffeomorphic to a solid 3-torus, $\mathbb{D}^2 \times S^1$. In [32], Smale asked whether every nonsingular vector field, e.g. X_a, on M had a periodic orbit. In [21], it was shown that smooth nonsingular *aperiodic* vector fields exist on any compact 3-manifold, with or without boundary.
2. Earlier examples due to Fuller [11] show that smooth nonsingular aperiodic vector fields exists on $\mathbb{D}^n \times S^1$, for $n \geq 3$.
3. In [30], a submanifold $M \subset \mathbb{R}^{n+k}$, for $k \geq 1$, which is diffeomorphic to $\mathbb{D}^n \times S^1$ is called a toroidal manifold and the *Principle of the Torus* asserts that, if $X \in \mathrm{Vect}(\mathbb{R}^{n+k})$ leaves a toroidal manifold positively invariant and has a global positive section $\mathcal{S}$ which is diffeomorphic to $\mathbb{D}^{n-1}$, then X has a periodic orbit in M.

Clearly, among the limiting features of the Principle of the Torus is the need to know enough about the long-time behavior of the integral curves of X to check the hypotheses. In [4], an additional hypothesis, reminiscent of Liapunov theory, was shown to imply that the hypotheses in the Principle of the Torus [30] hold. More explicitly, if M is a smooth manifold and $X \in \mathrm{Vect}(M)$, then a closed one-form $\omega \in \Omega^1(M)$ satisfying

$$\langle \omega, X \rangle > 0 \tag{8.21}$$

is said to be a *positive* one-form for X.

Remark 8.2.2. We note that (8.21) can be checked pointwise, *without* knowing the integral curves of X, just as in classical Liapunov theory.

Remark 8.2.3. According to Theorem 8.2.1, for a dissipative system the vector field X_a leaves the toroidal manifold M is a toroidal manifold whenever $n \neq 3$. Moreover, for every $n \geq 1$, the closed one-form $d\tau \in \Omega^1(N)$ satisfies

$$\langle d\tau, X_a \rangle = 1 > 0.$$

Every closed one-form ω defines a period map J_ω. In our case, fixing a base point $x^* \in N$, consider an arbitrary $x \in N$. For any parametrized path in N from x^* to x we may compute the line integral

$$\bar{J}_{d\tau}(x) = \int_{x_0}^{x} d\tau.$$

As a function of the upper limit x, $\bar{J}_{d\tau}$ is only well-defined up to the periods [8] of $d\tau$; i.e., up to the elements of the subgroup

$$\Pi(d\tau) = \{ \int_{\gamma} d\tau : \gamma \in H_1(N) \} \subset \mathbb{R}.$$

By direct computation on the toroidal cylinder, we see that $\Pi(d\tau) = \mathbb{Z} \subset \mathbb{R}$, reflecting the fact that $d\tau$ is an *integral* one-form. Thus, the *period map* $J_{d\tau}$ of ω

$$J_{d\tau} : M \to S^1, \tag{8.22}$$

defined via

$$J_{d\tau}(x) = \int_{x_0}^{x} d\tau \mod \mathbb{Z}, \tag{8.23}$$

is a smooth surjection from M to the circle $S^1 = \mathbb{R} \mod \mathbb{Z}$. In fact, an explicit computation of the line integrals shows:

Lemma 8.2.1. *For a dissipative systems, $J_{d\tau} = J$.*

Theorem 8.2.2. [4] *Suppose N is paracompact smooth manifold and that $M \subset N$ is a smooth submanifold which is diffeomorphic to $\mathbb{D}^n \times S^1$. If $X \in \mathrm{Vect}(N)$ leaves M positively invariant and has a positive integral one-form ω on M. For almost every $\theta \in S^1$, $J_\omega^{-1}(\theta) \subset M$ is a smooth codimension 1 submanifold diffeomorphic to $\mathbb{D}^n$.*

This proves the first two assertions of Theorem 8.1.1. If $n \neq 3$, the existence of a harmonic solution follows from Brouwer's Fixed Point Theorem. For the case $n = 3$, see Remark 8.1.4. The remainder of the Theorem 8.1.1 is just the Lefschetz Fixed Point Formula

$$\sum_{\gamma} \mathrm{sign}\, \det(I - D\mathcal{P}_\gamma) = \mathrm{Lef}(\mathcal{P}) = 1 \tag{8.24}$$

for the smooth map $\mathcal{P}$, as in Theorem 1.1 of [4]:

Theorem 8.2.3. [4] *Suppose N is paracompact smooth manifold and that $M \subset N$ is a smooth submanifold which is diffeomorphic to $\mathbb{D}^{n-1} \times S^1$. If $X \in \mathrm{Vect}(N)$ leaves M positively invariant and has a positive one-form ω on M, then X has a periodic solution γ_0 in M whose homotopy class generates $\pi_1(M)$. If every periodic orbit γ satisfying $[\gamma] = [\gamma_1]$ is hyperbolic, then there is an odd number number of such periodic orbits and* (8.17) *holds. Moreover, if each of these orbits is asymptotically stable, then there exist just one periodic orbit in this homotopy class.*

In the next section, we shall prove Theorems 8.1.1 - 8.2.1 in the setting studied by Krasnosel'skiĭ and Perov.

8.3 Krasnosel'skiĭ's Theorem

Following [20], one says that $W : \mathbb{R}^n \to \mathbb{R}$ is a "guiding function" for (8.1) provided

$$\langle dW, X \rangle > 0 \text{ for } ||\mathrm{x}|| > \rho, \text{ for some } \rho > 0, \tag{8.25}$$

is satisfied. As in [20], we shall also suppose that

$$\lim_{\|x\| \to \infty} \|W(x)\| = +\infty \tag{8.26}$$

is satisfied.

Remark 8.3.1. Assuming that (8.25)-(8.26) are satisfied and defining $V = -W$ and replacing t by $-t$ if necessary [20, p. 49], we can assume that $V(x)$ is nonnegative on $\mathbb{R}^n$ and positive for large values of x. In particular, (8.1) is dissipative and we may consider the Poincaré map (8.11).

Theorem 8.3.1 (Krasnosel'skiĭ-Perov). *If* (8.1) *has a guiding function* W *for which* (8.26) *is satisfied then the index of the field* $x \to x - \mathcal{P}(x)$ *satisfies*

$$\mathrm{Ind}(I - \mathcal{P}) = 1 \tag{8.27}$$

on $W^{-1}(-\infty, c]$ *for* $c > \rho$. *In particular,* (8.1) *has a harmonic solution.*

Proof. Assuming (8.12) holds, (8.1) has a harmonic solution by Brouwer's Fixed Point Theorem. Moreover by the Lefschetz Fixed Point Theorem,

$$1 = \mathrm{Lef}(\mathcal{P}) = \sum_{\gamma} \mathrm{sign}\, \det(I - D\mathcal{P}_{\gamma}) \tag{8.28}$$

Moreover, (8.27) follows from (8.28) and the Poincaré-Hopf Theorem.

From (8.25), it follows that the set, $\mathcal{C}$, of critical points of W is compact. Since (8.25) implies that W is a proper function, Proposition 8.1.2 is a direct corollary of [3, Theorem 4.1]. ∎

Consider the periodically time-varying perturbed system (8.3). In Remark 8.1.2, we observed that if the unperturbed system has a global compact attractor $\mathcal{A}$ then for small perturbations (8.3) is dissipative with a Liapunov function satisfying (8.9). In particular, in light of Remark 8.1.2 and Theorem 8.3.1, we obtain the following result.

Corollary 8.3.1. *Suppose the unperturbed system* $\dot{x} = f(x)$ *has a compact global attractor* $\mathcal{A}$. *There exists an* $\epsilon_{\mathcal{A}} \ll \infty$ *such that for each* $0 < \epsilon < \epsilon_{\mathcal{A}}$ *the perturbed system* (8.3) *has a periodic orbit of period* T.

Example 8.3.1. (*The forced van der Pol Oscillator for small periodic forcing*) Consider the periodically forced van der Pol system (8.2). For the unforced system, we denote by $\mathcal{A}_{\mu} \subset \mathbb{R}^2$ the compact invariant subset consisting of the limit cycle γ_{μ} and the topological disk that γ_{μ} bounds. It is well-known that every trajectory of the unforced van der Pol oscillator is bounded forward in time and tends to $\mathcal{A}_{\mu}$. In particular, $\mathcal{A}_{\mu}$ is the largest global attractor for the unforced van der Pol oscillator. By Remark 8.1.2, there exist $\alpha_{\mu} > 0$ so that (8.2) is dissipative with a guiding function W. By Corollary 8.3.1, the forced van der Pol oscillator (8.2) has a periodic solution of period T for amplitudes $\alpha < \alpha_{\mu} < \infty$.

8.4 Proof of the Main Theorem

Fixing $c > \rho$, consider the submanifold $M^+ = V^{-1}[c, \infty) \subset N$. Clearly,

$$N = M \cup M^+, \;\; M \cap M^+ = V^{-1}(c). \tag{8.29}$$

Consider the vector field $\mathcal{V} = -\mathcal{W}$. $\mathcal{V}$ leaves M^+ invariant and defines a complete vector field on $\mathring{M}^+$with flow

$$\Psi : \mathbb{R} \times \mathring{M}^+ \to \mathring{M}^+. \tag{8.30}$$

Therefore, if $\bar{c} > c$, then

$$\Psi : \mathbb{R} \times V^{-1}(\bar{c}) \simeq \mathring{M}^+.$$

In particular, the inclusion $V^{-1}(\bar{c}) \subset \mathring{M}^+$is a homotopy equivalence.

Proposition 8.4.1. *The inclusion $\iota : V^{-1}(c) \to M^+$ is a homotopy equivalence.*

Proof. The composition $V^{-1}(\bar{c}) \subset \mathring{M}^+ \subset M^+$ is a homotopy equivalence, by [35, p. 297]. Since the level sets of V, for $\bar{c}$ near c, form the leaves of a product neighborhood of $V^{-1}(c) = \partial M^+$ in M^+, the proposition is proved.

Now, $N \subset S^n \times S^1$, where $\mathcal{C} = S^n \times S^1 - N = \{\infty\} \times S^1$ is the "circle at ∞" and V extends to a continuous function $\bar{V} : S^n \times S^1 \to S^1$ such that $\bar{V}^{-1}(\infty) = \mathcal{C}$. In particular,

$$\bar{V}^{-1}(c, \infty] = \mathring{M}^+ \cup \mathcal{C} \tag{8.31}$$

Smoothing $\mathcal{V}$ to 0 on $\mathcal{C}$ defines a smooth, complete vector field $\widetilde{\mathcal{V}}$ on $\bar{V}^{-1}(c, \infty]$ with $\mathcal{C}$ as an attractor. Denote the flow of $\mathcal{V}_1$ by $\widetilde{\Psi}$, so that

$$\widetilde{\Psi} : \mathbb{R} \times \bar{V}^{-1}(c, \infty] \to \bar{V}^{-1}(c, \infty].$$

If $\mathcal{C} \subset T \subset \bar{V}^{-1}(c, \infty]$ is a fixed tubular neighborhood of $\mathcal{C}$ and $K \subset \bar{V}^{-1}(c, \infty]$ is an arbitrary compact subset then there exists $\tau \geq 0$ such that $K \subset T_\tau = \widetilde{\Psi}_{-\tau}(T)$. Therefore, for $\tau \in \mathbb{N} \cup \{0\}$ we have

$$\bar{V}^{-1}(c, \infty] = \cup_{\tau=0}^{\infty} T_\tau \simeq \mathbb{R} \times \mathcal{C} \tag{8.32}$$

is diffeomorphic to a tubular neighborhood of $\mathcal{C}$ via a diffeomorphism which fixes $\mathcal{C}$ [39, Lemma 3.3](see also the proof of [25, Lemma 3]). Accordingly, from (8.32) and [39, Corollary 3.5] we see the following.

Proposition 8.4.2. *$\bar{V}^{-1}[c, \infty]$ is a $K(\mathbb{Z}, 1)$ and $\mathring{M}^+$ has the homotopy type of $S^{n-1} \times S^1$. In particular, $V^{-1}(c)$ has the homotopy type of $S^{n-1} \times S^1$.*

Since $V^{-1}(c)$ is connected, M is connected and therefore the pair $(M, V^{-1}(c))$ is 0-connected. By Proposition 8.4.1 the pair $(M^+, V^{-1}(c))$ is k-connected for every $k \geq 0$, so that by the homotopy excision theorem [14, Theorem 4.23] applied to (8.29) we see that

$$\pi_k(N, M) \simeq \pi_k(M^+, V^{-1}(c)) = (0) \;\text{ for all } k \geq 0. \tag{8.33}$$

Since N is a $K(\mathbb{Z}, 1)$, we have

Proposition 8.4.3. *M is a $K(\mathbb{Z}, 1)$ with $\partial M \sim S^{n-1} \times S^1$.*

Proposition 8.4.4. *If* $n \leq 2$, $M_c \simeq \mathbb{D}^{n-1} \times S^1$. *If* $n = 3$, *then* $M_c \simeq_h \mathbb{D}^3 \times S^1$.

Proof. If $n = 1$, since $\pi_1(M)$ is abelian and M is orientable, $M \simeq \mathbb{A}$ by the classification of surfaces [16]. As a consequence of Perelman's proof of the Poincaré conjecture ([29],[28]) the only compact orientable 3-dimensional $K(\mathbb{Z}, 1)$ is $\mathbb{D}^2 \times S^1$. According to ([10],[36]), up to homeomorphism $\mathbb{D}^3 \times S^1$ is the only $K(\mathbb{Z}, 1)$ in four dimensions.

Suppose $\gamma \subset M$ is closed curve such that $([\gamma]) = \pi_1(M) \simeq \mathbb{Z}$. If $\gamma \subset T \subset M$ is a closed tubular neighborhood of γ, the smooth manifold with boundary

$$W = M - \overset{\circ}{T} \ , \ \partial W = W_1 \dot{\cup} \ W_2 \tag{8.34}$$

is a smooth cobordism between $W_1 = \partial T$ and $W_2 = V^{-1}(c)$. $T \simeq \mathbb{D}^n \times S^1$ since N is orientable and the inclusion $\iota_1 : W_1 \to W$ is a homotopy equivalence since M is a $K(\mathbb{Z}, 1)$ (see, e.g., [4, Lemma 5.4]). In particular, $\pi_1(W) \simeq \mathbb{Z}$. Since $W_2 \sim S^{n-1} \times S^1$, the inclusion $i_2 : W_2 \to W$ is a homotopy equivalence whenever $n \geq 3$ (see, e.g., [4, Lemma 5.5]). Therefore,

Proposition 8.4.5. *If* $n \geq 4$, (W, W_1, W_2) *is an h-cobordism.*

Moreover, since the Whitehead group $\mathrm{Wh}(\pi_1(W))$ vanishes [15], the h-cobordism is an s-cobordism [18].

Proposition 8.4.6. *If* $n \geq 4$, *then* $M \simeq \mathbb{D}^n \times S^1$.

Proof. If $n \geq 5$, $W \simeq \partial T \times [0, 1]$, by the s-cobordism theorem of Barden, Mazur and Stallings [18]. Therefore, the product cobordism $(W, \partial T, \partial M)$ can glued [27, Theorem 1.4] along ∂T with the trivial cobordism $(T, \emptyset, \partial T)$ to obtain a diffeomorphism $M \simeq \mathbb{D}^{n-1} \times S^1$. If $n = 5$, then $W \simeq_h \partial T \times [0, 1]$ by [10, Theorem 7.1A], yielding a homeomorphism $M \simeq_h \mathbb{D}^4 \times S^1$. By [19, TheoremC.2 (ii), p. 275], $\mathbb{D}^4 \times S^1$ has a unique differentiable structure.

The following result completes the proof of Corollary 8.2.1.

Proposition 8.4.7. *If* $n \geq 1$, $V^{-1}(c) \simeq S^{n-1} \times S^1$.

Proof. For $n \leq 3$, this follows from Proposition 8.4.4 and the fact that homeomorphic three manifolds are diffeomorphic. For $n \geq 4$, the claim follows from Proposition 8.4.6.

8.5 An Example

Example 8.5.1. Consider the autonomous system

$$\begin{aligned} \dot{x} &= -(z + w_1)^3 + x + w_1^3 \\ \dot{w_1} &= w_2 \\ \dot{w_2} &= -w_1 \end{aligned} \tag{8.35}$$

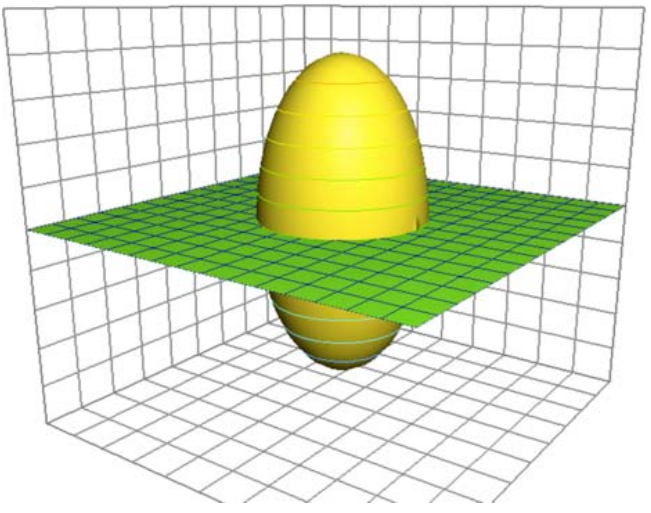


Fig. 8.1. The egg-shaped global attractor for initial data in B

For initial data $w(0) = \begin{pmatrix} A \\ 0 \end{pmatrix}$, this system takes the form of a periodically forced system

$$\dot{x} = f(x,t) = f_A(x,t) = -(x + A\sin t)^3 + x + A^3 \sin^3 t \tag{8.36}$$

evolving on $\mathbb{R}$ with a bifurcation parameter A. We will illustrate the use of Theorem 8.1.1, and epecially the index formula (8.17), in determining the existence, number and stability types of periodic orbits for a range of values of A.

For each fixed $A > 0$, (8.36) evolves on the "toroidal cylinder" $M_A = \{(x,\tau)\} \subset \mathbb{R} \times S^1$ and is dissipative with Liapunov function $V_a(x,\tau) = x^2$ satisfying (8.9). Indeed, for any $A > 0$, (8.36) leaves the annulus

$$\mathbb{A} \simeq \{(x,\tau) : |x| < 2\} \subset M_A \tag{8.37}$$

positively invariant. M_A admits a positively invariant decomposition

$$M_A = M_A^+ \dot{\cup} M_A^- \dot{\cup} M_A^0 \tag{8.38}$$

obtained by fixing the sign of x, or setting $x = 0$, respectively. M_A^0, is a 2π-periodic orbit on M_A which is easily seen to be hyperbolically stable if $A > \sqrt{2/3}$, hyperbolically unstable if $A < \sqrt{2/3}$, and critically stable at the amplitude $A = \sqrt{2/3}$.

In $\mathbb{R}$ this orbit corresponds to an equilibrium of (8.36). It is easy to show that every other periodic orbit in $\mathbb{R}$ has a period commensurate with 2π. A straightforward calculation shows

$$x^2(t) = \exp\left[-2\int_0^t \left(x^2(\tau) + 3w(\tau)x(\tau) + 3w^2(\tau) - 1\right)\, d\tau\right] X^2(0)$$

from which we compute

$$DP^k(z_0) = \exp\,[(3A^2 - 2)k\pi - \int_0^{2k\pi} x^2(t)dt]. \tag{8.39}$$

provided $x(\cdot, x_0)$ is periodic in t of period $2k\pi$ and $x(t) \neq 0$. In particular, if $A < \sqrt{2/3}$, all $2k\pi$-periodic solutions in M_A^+ are hyperbolically stable so that, by the index formula (8.17) on M_A, $k = 1$ and the periodic orbit is unique and asymptotically stable. The same assertion holds for M_A^-.

Summarizing the analysis of this example, we note that:

- For $A > \sqrt{2/3}$, the augmented dynamics on M_A is dissipative and has a hyperbolically stable periodic orbit M_A^0 with period 2π.
- For $A < \sqrt{2/3}$, the augmented dynamics on M_A^+ (resp.,M_A^-) is dissipative and has only hyperbolically stable periodic orbits. Therefore, by Theorem 8.1.1, there is a unique periodic orbit, which has period 2π.
- For $A < \sqrt{2/3}$, the augmented dynamics on M_A has three periodic orbits, each with period 2π: a unique, hyperbolically stable periodic orbit in each of M_A^+, M_A^- and a unique hyperbolically unstable periodic orbit M_A^0.

Remark 8.5.1. For $A_c = \sqrt{2/3}$, form the index formula (8.17) and (8.39) there is unique periodic orbit, viz. $M_{A_c}^0$. Moreover, the Poincaré map P_c can be shown [5] to satisfy:

$$DP_c(0) = 1, \;\; D^2P_c(0) = 0, \;\; D^3P(0) = -12\pi \tag{8.40}$$

and therefore undergoes a pitchfork bifurcation. In particular, (8.17) implies that the resulting three periodic orbits for $\sqrt{2/3} - \epsilon < A < \sqrt{2/3}$ resulting from the local pitchfork bifurcation account for the global behavior for all amplitudes $0 < A < \sqrt{2/3}$.

Following [5], we depict the global attractor of (11.2) for the invariant set $M = \{(z, w) : \|w\|^2 \leq 2/3\}$ in a plot of z on the vertical axis, versus a plane on which the harmonic oscillator

$$\dot{w}_1 = w_2, \;\; \dot{w}_2 = -w_1 \tag{8.41}$$

evolves along periodic orbits of constant amplitude $A^2 = w_1^2 + w_2^2 \leq 2/3$. In particular, the intersection of the egg-shaped surface with the plane is the critically stable perioidic orbit with $A = \sqrt{2/3}$ while the surface is (singularly) foliated by periodic orbits and two equilibria $(0, 0, (\pm 1)$. The compact region bounded by the surface forms the global attractor for M, consisting as it must of those trajectories which are bounded forward and backward in time.

Acknowledgement. Research supported in part by grants from the AFOSR.

References

1. Browder, F.: On a generalization of the Schauder fixed point theorem. Duke Math. 26, 291–303 (1959)
2. Browder, F.: Fixed point theorems on infinite dimensional manifolds. Trans. of the Amer. Math. Soc. 119, 179–194 (1965)
3. Byrnes, C.I.: On Brockett's necessary condition for stabilizability and the topology of Liapunov functions on $\mathbb{R}^n$, CIS (to appear)
4. Byrnes, C.I., Brockett, R.W.: Nonlinear oscillations and vector fields paired with a closed one-form. Submitted to Amer. J. of Math
5. Byrnes, C.I., Gilliam, D.S., Isidori, A., Ramsey, J.: On the steady-state behavior of forced nonlinear systems. In: New trends in nonlinear dynamics and control, and their applications. LNCIS, vol. 295, pp. 119–143. Springer, Heidelberg (2003)
6. Cartwright, M.: Forced oscillations in nearly sinusoidal systems. J. Inst, Elec. Eng. 95, 88–99 (1948)
7. Cartwright, M., Littlewood, J.E.: On nonlinear differential equations of the second order I. J. Lond. Math. Soc. 20, 88–94 (1945)
8. Farber, M.: Topology of closed one-forms. SURV, vol. 108. Amer. Math. Soc., Providence (2004)
9. Freedman, M.H.: The topology of four-dimensional manifolds. J. Diff. Geom. 17, 357–453 (1982)
10. Freedman, M.H., Quinn, F.: Topology of 4-manifolds. Princeton University Press, Princeton (1990)
11. Fuller, F.B.: Note on trajectories in a solid torus. Ann. of Math. 56, 438–439 (1952)
12. Guckenheimer, J., Holmes, P.J.: Nonlinear oscillations, dynamical systems and bifurcations of vector fields. Applied Math. Sciences, vol. 42. Springer, Heidelberg (1983)
13. Hale, J.K.: Asymptotic Behavior of Dissipative Systems. AMS Series: Surv. Series 25 (1988)
14. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge (2001)
15. Higman, G.: The units of group rings. In: Proc. London Math. Soc., vol. 46, pp. 231–248 (1940)
16. Hirsch, M.W.: Differential Topology. Springer, New York (1976)
17. Holmes, P.J., Rand, D.A.: Bifurcations of the forced van der Pol oscillator. Quart. Appl. Math. 35, 495–509 (1978)
18. Kervaire, M.: Le théorèm de Barden-Mazur-Stallings. Comment. Math. Helv. 40, 31–42 (1965)
19. Kirby, R.C., Siebenmann, L.C.: Foundational essays on topological manifolds, smoothings, and triangulations, revised edn. Annals of Math. Studies, vol. AM-88. Princeton University Press, Princeton (1977)
20. M. A. Krasnosel'skiĭ and P. P. Zabreĭko, Geometric Methods of Nonlinear analysis, Springer-Verlag, Berlin, 1984.
21. Kuperberg, G., Kuperberg, K.: Generalized counterexamples to the Seifert conjecture. Ann. of Math. 144, 239–268 (1996)
22. Levinson, N.: Transformation theory of non-linear differential equations of the second order. Ann. of Math 49, 738 (1948)
23. Levinson, N.: Small periodic perturbations of an autonomous system with a stable orbit. Ann. of Math 52, 727–738 (1950)
24. Mazur, B.: On embeddings of spheres. Bull. Amer. Math. Soc. 65, 59–65 (1961)

25. Milnor, J.W.: Differential Topology. In: Saaty, T.L. (ed.) Lectures in Modern Mathematics, vol. II. Wiley, Chichester (1964)
26. Milnor, J.W.: Topology From a Differentiable Viewpoint. University Press of Virgina, Charlotsville (1965)
27. Milnor, J.W.: Lectures on the h-Cobordism Theorem. Princeton University Press, Princeton (1965)
28. Morgan, J.W., Tian, G.: Ricci flow and the Poincaré conjecture, Math. DG/0607607 (2007)
29. Perelman, G.: Finite extinction time for solutions of the Ricci equation on certain three manifolds, Math. DG/0303109 (2003)
30. Pliss, V.A.: Nonlocal Problems in the Theory of Nonlinear Oscillations. Academic Press, New York (1966)
31. Sell, G., You, Y.: Dynamics of Evolutionary Equations. Applied Math Sciences, vol. 143. Springer, New York (2002)
32. Smale, S.: Generalized Poincaré conjecture in dimensions greater than 4. Ann. of Math. 64, 399–405 (1956)
33. Smale, S.: Differentiable and combinatorial structures on manifolds. Ann. of Math. 74, 498–502 (1961)
34. Smale, S.: On the structure of manifolds. Amer. J. of Math. 84, 387–399 (1962)
35. Spanier, E.H.: Algebraic Topology. McGraw-Hill, New York (1966)
36. Stong, R., Wang, Z.: Self-homeomorphisms of 4-manifolds with fundamental group $\mathbb{Z}$. Toplology and its Applications 106, 49–56 (2000)
37. Thom, R.: Quelques propriétiés globales des variétiés différentiables. Comment. Math. Helv. 28, 17–86 (1954)
38. van der Pol, B.: Forced oscillations in a circuit with nonlinear resistance, London, Edinburgh and Dublin. Phil. Mag, vol. 3, pp. 65–80 (1927)
39. Wilson, F.W.: The structure of the level sets of a Lyapunov function. J. of Diff. Eqns. 3, 323–329 (1967)
40. Wilson, F.W.: Smoothing derivatives of functions and applications. Trans. of the Amer. Math. Soc. 139, 413–428 (1969)
41. Yoshizawa, T.: Stability theory and the existence of periodic solutions and almost periodic solutions. Appl. Math. Sci, vol. 14. Springer, New York (1975)

9

Global Controllability of Switched Affine Systems

Daizhan Cheng

Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China
dcheng@iss.ac.cn

To the memory of Dr. Wijesuriya P. Dayawansa

9.1 Introduction

Consider a linear control system

$$\dot{x} = Ax + \sum_{i=1}^{m} b_i u_i := Ax + Bu, \quad x \in \mathbb{R}^n,\ u \in \mathcal{U}^m, \tag{9.1}$$

where $\mathcal{U}$ is the set of piecewise continuous functions. It is well known that (9.1) is controllable, iff the controllability matrix

$$\mathcal{C} = \left[B\ AB \cdots\ A^{n-1}B\right] \tag{9.2}$$

has full row rank, i.e., $rank(\mathcal{C}) = n$ [22].

Consider an affine nonlinear control system

$$\dot{x} = f(x) + \sum_{i=1}^{m} g_i(x) u_i := f(x) + G(x)u, \quad x \in M,\ u \in \mathcal{U}^m, \tag{9.3}$$

where M is an n dimensional manifold (with $\mathbb{R}^n$ as a special case); $f(x), g_i(x) \in V^\infty(M)$, i.e., they are C^∞ vector fields on M. For system (9.3), we define (strong) accessibility Lie algebra, $\mathcal{L}_a$ ($\mathcal{L}_{sa}$), as

$$\mathcal{L}_a = \{f, g_1, \cdots, g_m\}_{LA}, \tag{9.4}$$

$$\mathcal{L}_{sa} = \{ad_f^k g_i \mid i = 1, \cdots, m;\ k = 0, 1, \cdots\}_{LA}. \tag{9.5}$$

Note that in linear case (9.5) is degenerated to (9.3), while $\mathcal{L}_a$ might have one more dimension caused by $f(x)$ as $f(x) \notin \mathcal{L}_{sa}(x)$.

Denote by $R(x_0)$ the reachable set from x_0. That is, $y \in R(x_0)$ means there exist $u_i(t) \in \mathcal{U}$, $i = 1, \cdots, m$, and $t_1 > 0$, such that the trajectory $x(t, u_i, x_0)$

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 141–167.
springerlink.com

satisfies $x(0, u_i, x_0) = x_0$ and $x(t_1, u_i, x_0) = y$. We use $R_T(x_0)$ to denote the reachable set from x_0 at a particular moment $T > 0$.

Unlike linear case, the (strong) accessibility Lie algebra can't determine even the local controllability. A general controllability result deduced from the rank of (strong) accessibility Lie algebra is the following:

Theorem 9.1.1. *[20] For system (9.3) let $x_0 \in M$. If rank$(\mathcal{L}_a(x_0)) = n$, then the reachable set $R(x_0)$ contains a non-empty open set; If rank$(\mathcal{L}_{sa}(x_0)) = n$, then at any moment $T > 0$ the reachable set $R_T(x_0)$ contains a non-empty open set.*

There are many other controllability results for nonlinear systems, particularly, for some nonlinear systems of special forms. For instance, for Lie determined systems [13] provided a systematic analysis with elegant results; some recent results can be found in [5], [10], [15].

In the last decade, the switched systems have attracted a considerable attention from control community. A switched linear system concerned is of the following form.

$$\dot{x} = A^{\sigma(t)}x + B^{\sigma(t)}u, \quad x \in \mathbb{R}^n,\ u \in \mathcal{U}^m, \tag{9.6}$$

where and hereafter $\sigma : [0, \infty) \to \Lambda = \{1, \cdots, N\}$ is assumed to be a right continuous piecewise constant function, called the switching law.

Correspondingly, affine nonlinear switched systems are defined as

$$\dot{x} = f^{\sigma(t)}(x) + g^{\sigma(t)}(x)u, \quad x \in M,\ u \in \mathcal{U}^m. \tag{9.7}$$

Controllability of switched systems becomes another interesting topic. Most of the researches are focused on the controllability of switched linear systems. Say, [19] provided necessary and sufficient condition for controllable subspace. [8] showed that controllable sub-manifold can provide better description for the controllability of switched linear systems. The results on impulsive control systems and the control design can be found in [23]. As for the switched nonlinear systems, the results are very limited. One recent result is about the controllability of bilinear systems [9].

The following example shows that the (strong) controllability rank condition is far no necessary and/or sufficient for global controllability of nonlinear systems.

Example 9.1.1. Consider the following system

$$\begin{aligned} &\dot{x} = f(y), \\ &\dot{y} = u, \quad (x, y) \in \mathbb{R}^2,\ u \in \mathcal{U}, \end{aligned} \tag{9.8}$$

where $(0 < \delta << 1)$

$$f(y) = \begin{cases} e^{-\frac{1}{(y-1)^2}} e^{-\frac{1}{(y-1-\delta)^2}}, & 1 < y < 1 + \delta \\ -e^{-\frac{1}{(y+1)^2}} e^{-\frac{1}{(y+1+\delta)^2}}, & -1 - \delta < y < -1 \\ 0, \quad otherwise. \end{cases}$$

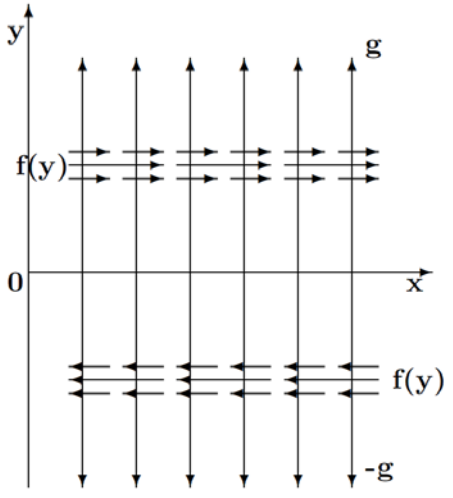


Fig. 9.1. Rank Condition vis Controllability

This is a C^∞ system. Using Theorem 9.2.1, it can be shown that the system is globally controllable, but the rank of its accessibility Lie algebra is 1 except two narrow strips, where the rank is 2.

Next, if we restrict the system on each strip, say, let

$$M = \{(x, y) \in \mathbb{R}^2 \mid 1 < y < 1 + \delta\}.$$

Then the rank condition is satisfied but it is not even locally controllable. (Refer to Figure 9.1)

In this paper, the controllability of affine systems is treated in a unified way, no matter linear or nonlinear, switched or no-switched. In the sequel, only the controllability of system (9.7) is treated formally. But (9.6) is considered as a particular case of (9.7). Moreover, in our approach $N = 1$ is allowed. That is, we also consider (9.3) (including (9.1)) as a particular case of (9.7) with only a single switching model.

The basic idea of this paper is to construct a sequence of distributions, called the controllability distributions, $\Delta_1, \Delta_2, \cdots$, which correspond to $\{B\}$, $\{B, AB\}$, $\{B, AB, A^2B\}$, $\cdots$ of linear case. Then to testify when the trajectory can be driven from the integral manifold of lower dimensional distributions to the integral manifold of higher dimensional distributions.

For statement ease, we introduce some notations:

- $\sqcup_\delta^k := \underbrace{(-\delta, \delta) \times (-\delta, \delta) \times \cdots \times (-\delta, \delta)}_{k} \subset \mathbb{R}^k$, where $0 < \delta << 1$.
- Let X be a vector field on M. Then $\phi_t^X(p)$ is the integral curve of X with initial value $\phi_0^X(p) = p$.
- $T(M)$ is the tangent bundle on M and $T_x(M)$ is the leaf of $T(M)$ at x (i.e., the tangent space of M at x).

- Denote by $V^{\infty}(M)$ (or briefly, $V(M)$) the set of C^{∞} vector fields. All vector fields considered in this paper are C^{∞} (including C^{ω}).
- F is used for the set of drift vector fields of (9.7), i.e., $F = (f^1, f^2, \cdots, f^N)$;
- G^i is the set of input channels of i th switching model, i.e., $G^i = (g_1^i, \cdots, g_m^i)$, and G as the union of G^i, i.e., $G = (G^1, \cdots, G^N)$. We use them in both set and matrix sense.
- $\mathcal{G}$: the Lie algebra generated by all input channels i.e., $\mathcal{G} = \{G\}_{LA}$.
- $\mathcal{I}_{\mathcal{G}}(x_0)$: The maximal integral manifold of $\mathcal{G}$ passing through x_0.
- Let Δ be a distribution. $\bar{\Delta}$ denotes its involutive closure, i.e., the smallest involutive distribution containing Δ.

The rest of this paper is organized as follows: Section 9.2 gives a sufficient condition for the global controllability of system (9.7). Section 9.3 considers the general case where the approach, described in Section 9.2, is applied recursively. Section 9.4 considers the case when the codimension of $\mathcal{G}$ is 1. The it is proved that in certain cases the sufficient condition obtained in Section 9.3 becomes necessary too. Section 9.5 is the concluding remarks. All long proofs are collected in appendixes.

9.2 A Sufficient Condition for Controllability

Definition 9.2.1. *Let $S := \{X_\lambda | \lambda \in \Lambda\} \subset V(M)$ be a (finite or infinite) set of vector fields. A point $x_0 \in M$ is called an interior (point) of S, if there exists a finite set $\{Y_1, \cdots, Y_s\} \subset S$ such that 0 is an interior point of the convex cone generated by $Y_1(x_0), \cdots, Y_s(x_0)$.*

The following proposition tells how to verify the interior of S.

Proposition 9.2.1. *x_0 is an interior of S, iff there exists a finite set $\{Y_1, \cdots, Y_s\} \subset S$ such that*

(i) there exist n vectors from $\{Y_1(x_0), \cdots, Y_s(x_0)\}$, which are linearly independent;

(ii) there exist $c_i > 0$, $i = 1, \cdots, s$ such that

$$\sum_{i=1}^{s} c_i Y_i(x_0) = 0. \tag{9.9}$$

Proof. Necessity of (i) comes from the fact that if (i) fails, it means all $Y_i(0)$ are linearly dependent. So the convex cone generated by them, denoted by $con(Y_1(0), \cdots, Y_s(0))$, has dimension less than n. So it doesn't have any interior point in $\mathbb{R}^n$ topology;

(9.9) with $\sum_{i=1}^{s} c_i = 1$ is necessary and sufficient condition for 0 to be an interior point of the convex cone (with respect to subspace topology, which could be of lower dimension). Since the interior is zero, the condition $\sum_{i=1}^{s} c_i = 1$ can obviously be removed.

As for the sufficiency, condition (i) assures that the cone has dimension n. So as an interior of the cone, zero is also an interior in $\mathbb{R}^n$ topology. ∎

A1: *$\mathcal{G}$ has constant dimension.*

From Frobenius' theorem [14] A1 assures that for each $x \in M$ the maximal integral manifold $\mathcal{I}_{\mathcal{G}}(x)$ exists.

We give a rigorous definition for controllability discussed in this paper.

Definition 9.2.2. *System (9.7) is said to be controllable if for any two given points $x, y \in M$, $x \in R(y)$. It is controllable on a subset, $W \subset M$, (briefly, W is controllable) if for any two given points $x, y \in W$, $x \in R_W(y)$, which means a trajectory can start from y traveling within W to reach x.*

The following Lemma is a key for our main result, and itself is interesting:

Lemma 9.2.1. *Assume A1 and let $x_0 \in M$ be an interior of $\{F, \mathcal{G}\}$. Then there exists a neighborhood W of x_0, which is a controllable subset.*

Proof. In Appendix 9.A.

The disadvantage of Lemma 9.2.1 is obvious. If codimension of $\mathcal{G}$ is k, then, according to Proposition 9.2.1, at least $k+1$ switching models are required. So it can not be applied to non-switched system (9.3). We need to improve it.

Definition 9.2.3. *1. Let Δ be an involutive distribution on M with constant dimension k. A coordinate frame, x, is said to be flat with respect to Δ, if under this coordinate frame*

$$\Delta = Span\{\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_k}\}. \tag{9.10}$$

2. Let x be a coordinate frame flat with respect to an involutive distribution Δ, and a vector field $f \in V(M)$ is expressed in this coordinate frame as $f(x) = (f_1(x), \cdots, f_n(x))^T$. The quotient vector field of $f(x)$ over Δ, denoted by f/Δ, is defined as

$$f(x)/\Delta = f^2(x), \tag{9.11}$$

where $f(x) = \begin{bmatrix} f^1(x) \\ f^2(x) \end{bmatrix}$ and $f^2(x) = (f_{k+1}(x), \cdots, f_n(x))^T$.

From Frobenius' theorem we know that for an involutive distribution with constant dimension, its local flat coordinate chart (a chart with flat coordinate flame) always exists. Moreover, we have

Proposition 9.2.2. *[7] Let Δ be an involutive distribution with constant dimension, and a vector field $f \in V(M)$ be given. Then the quotient vector field f/Δ is uniquely defined.*

Definition 9.2.4. *1. given $Y \in V(M)$, a vector $X(x_0) \in T_{x_0}(M)$ is said to be a G-shifted Y from x_1, denoted by*

$$X(x_0) \in \mathcal{G}(x_0, x_1)_* Y(x_1), \tag{9.12}$$

if there exist a finite set of vector fields $Z_1, \cdots, Z_s \in G$, such that

$$x_0 = \phi_{t_1^0}^{Z_1} \circ \cdots \circ \phi_{t_s^0}^{Z_s}(x_1);$$

and

$$X(x_0) = \left(\phi_{t_1^0}^{Z_1}\right)_* \circ \cdots \circ \left(\phi_{t_s^0}^{Z_s}\right)_* Y(x_1).$$

2. *Let $S \subset V(M)$ be a set of vector fields. We may G-shift all vector fields $f \in S$ from all points $x_1 \in \mathcal{I}_{\mathcal{G}}(x_0)$ to one point x_0, and denote them by $\mathcal{G}_* S(x_0)$. That is*

$$\mathcal{G}_* S(x_0) = \cup \left\{ \mathcal{G}(x_0, x_1)_* Y(x_1) \middle| Y \in S, x_1 \in \mathcal{I}_{\mathcal{G}}(x_0) \right\}. \tag{9.13}$$

The following lemma shows that G-shift is "quotient-independent" of the choice of trajectories.

Proposition 9.2.3. *Let $X(x_0), Z(x_0) \in \mathcal{G}(x_0, x_1)_* Y(x_1)$. Then we have*

$$X(x_0)/\mathcal{G} = Z(x_0)/\mathcal{G}. \tag{9.14}$$

Proof. Assume $X(x_0)$ is obtained through the mapping:

$$x_0 = \phi_{t_1^0}^{Z_1} \circ \cdots \circ \phi_{t_s^0}^{Z_s}(x_1);$$

Note that $Z_i \in G$, so under a flat coordinate frame with respect to $\mathcal{G}$ (i.e., $\mathcal{G} = Span\{\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_k}\}$), we have

$$Z_i = \begin{bmatrix} Z_i^1 \\ 0 \end{bmatrix}.$$

So the Jacobian matrix of $\phi_{t_i}^{Z_i}$ has the form

$$J_\phi = \begin{bmatrix} J_{11} & 0 \\ 0 & I \end{bmatrix}.$$

It follows that

$$\left(\phi_{t_1^0}^{Z_1}\right)_* \circ \cdots \circ \left(\phi_{t_s^0}^{Z_s}\right)_* Y(x_1) = \begin{bmatrix} \phi_*(Y^1(x_1)) \\ Y^2(x_1) \end{bmatrix}$$

That is

$$X(x_0)/\mathcal{G} = Y^2(x_1), \tag{9.15}$$

which means the quotient vector field is independent of the choice of Z_i. ∎

Lemma 9.2.2. *Assume A1 and let $x_0 \in M$. If x_0 is an interior of $\{\mathcal{G}_* F(x_0);\ \mathcal{G}\}$, then there exists a neighborhood W of x_0, which is a controllable subset.*

Proof. In Appendix 9.B.

Now we are ready to state our first main result:

Theorem 9.2.1. *Assume A1 and M is arc-wise connected. Then system (9.7) is controllable if for each maximum integral manifold $\mathcal{I}_{\mathcal{G}}(x_0)$, $\forall x_0 \in M$, there exists a point $y \in \mathcal{I}_{\mathcal{G}}(x_0)$ (equivalently, every point on this integral manifold), which is an interior of $C_y = con\{\mathcal{G}_*(F);\ \mathcal{G}\}$.*

Proof. For any two points $x, y \in M$. Connect them by a path $c(t)$, $0 \leq t \leq 1$ ($c(0) = x$ and $c(1) = y$). According to Lemma 9.2.2 each point $c(t)$ has a controllable neighborhood, called $W(x_t)$, where $x_t = c(t)$. Since $c(t)$ is the continuous image of a compact set $[0, 1]$, it is compact. Now $\{W(x_t)|0 \leq t \leq 1\}$ is a covering of $c(t)$, it has a finite sub-covering $\{W(x_0 = x), W(x_1), \cdots, W(x_j = y)\}$. Without loss of generality, we can assume $W(x_i) \cap W(x_{i+1}) \neq \emptyset$, and $p_i \in W(x_i) \cap W(x_{i+1})$. Then $x \in R(p_i)$, $\forall i$, which means $x \in R(y)$. Similarly, we have $y \in R(x)$. ∎

We give some examples to illustrate this theorem:

Example 9.2.1. 1. Consider the following system

$$\begin{cases} \dot{x}_1 = x_n \\ \dot{x}_2 = x_n^3 \\ \vdots \\ \dot{x}_{n-1} = x_n^{2n-1} \\ \dot{x}_n = u, \quad x \in \mathbb{R}^n,\ u \in \mathcal{U}. \end{cases} \tag{9.16}$$

For notational ease, denote $x = (x^1, x_n)$, where $x^1 = (x_1, x_2, \cdots, x_{n-1})^T$. The integral manifold of G passing x_0 is

$$\mathcal{I}_{\mathcal{G}}(x_0) = \left\{ x \in \mathbb{R}^n | \, x^1 = x_0^1 \right\}.$$

Choosing $y_i = (x_0^1, k_i)$, $z_i = (x_0^1, -k_i)$, $i = 1, \cdots, n-1$, where $k_i > 0$, and $k_i \neq k_j$ $(i \neq j)$. Define vectors $Y_i = f(y_i)$, $Z_i = f(z_i)$, $i = 1, \cdots, n-1$. It is obvious that $g_*(Y_i) = Y_i$ and $g_*(Z_i) = Z_i$. Choose $Y_n = (0, \cdots, 0, 1)^T \in \mathcal{G}$, and $Z_n = -Y_n$. Then as a Vandermonde determined, we have

$$\det \left[Y_1, \cdots, Y_n \right] = \prod_{i=1}^{n-1} k_i \prod_{n \geq j > i \geq 1} (k_j^2 - k_i^2) \neq 0.$$

So, $\{Y_i\}$ are linearly independent. In addition,

$$\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} Z_i = 0.$$

Using Proposition 9.2.1, x_0 is an interior point of $\{\mathcal{G}_*(F);\ \mathcal{G}\}$. Since x_0 is arbitrary, Theorem 9.2.1 tells us that the system (9.16) is controllable.

2. Consider a planar controllable linear system

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = ax_1 + bx_2 + u. \end{cases} \tag{9.17}$$

It is obviously a particular case of system (9.16).

3. Consider the following system

$$\begin{cases} \dot{x}_1 = x_n^{q_1} + \mu_1(x) \\ \dot{x}_2 = x_n^{q_2} + \mu_2(x) \\ \vdots \\ \dot{x}_{n-1} = x_n^{q_{n-1}} + \mu_{n-1}(x) \\ \dot{x}_n = f_n(x) + g_n(x)u. \quad x \in \mathbb{R}^n \end{cases} \tag{9.18}$$

Assume a). $g_n(x) \neq 0$, $\forall x \in \mathbb{R}^n$. b). q_i, $i = 1, \cdots, n-1$ are distinguished positive odd numbers. c). Function $\mu_i(x)$ has the degree of x_n lower than q_i, i.e.,

$$\lim_{x_n \to \infty} \frac{\mu_i(x)}{x_n^{q_i}} = 0, \quad i = 1, \cdots, n-1.$$

Then the system (9.18) is controllable. The proof is sketched as follows: Using pre-feedback

$$u = \frac{1}{g_n(x)}[v - f_n(x)]$$

to simplify f and g first, we have $f_n(x) = 0$ and $g_n(x) = 1$. Then we consider these simplified f and g. Using similar notations as before and let k_i be large enough, since Y_i are linearly independent, we can express Z_i as

$$Z_i = f(z_i) = -\sum_{j=1}^{n} c_{ij} Y_j.$$

Moreover, as $k_i \to \infty$, It is obvious that

$$c_{ii} \to 1, \quad c_{ij} \to 0, \ \ j \neq i. \tag{9.19}$$

Now we have

$$\sum_{i=1}^{n-1} Z_i + \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} c_{ij} Y_j + Y_n + Z_n = \sum_{i=1}^{n-1} Z_i + \sum_{i=1}^{n-1} (c_{ii} + \sum_{j \neq i} c_{ij}) Y_i + Y_n + Z_n = 0.$$

The positivity of the coefficients is from (9.19).

Example 9.2.2. Consider the following switched system

$$\dot{x} = f^{\sigma(t)}(x) + g^{\sigma(t)}(x)u, \quad x \in M,\ u \in \mathcal{U}, \tag{9.20}$$

where $M = \mathbb{R}^3 \backslash \{0\} \subset \mathbb{R}^3$, $\Lambda = \{1, 2\}$ and

$$f^1 = (\sin(x_1), 0, 0)^T; \quad f^2 = (0, 0, 0)^T;$$

$$g^1(x) = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} x := B_1 x; \quad g^2(x) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} x := B_2 x.$$

It is easy to verify that as the subset of $gl(3, \mathbb{R})$ we have

$$\{B_1, B_2\}_{LA} = so(3, \mathbb{R}).$$

That is,

$$G = o(3, \mathbb{R})x.$$

So the maximum integral manifolds of G are spheres

$$\mathcal{I}_{\mathcal{G}}(x_0) = SO(3, \mathbb{R})x_0 = \|x_0\| S^2.$$

Now on each $\|x_0\| S^2$ we can find two points x, y with $0 < x_1 < \pi$, $x_3 > 0$ and $-\pi < y_1 < 0$, $y_3 > 0$ (on upper hemisphere). Then it is easy to see that $f_1(x_1)$ is going out and $f_1(x_2)$ is going into the sphere. A straightforward verification shows that the condition of Theorem 9.2.1 is satisfied. So the system is controllable on $M = \mathbb{R}^3 \backslash \{0\}$.

Remark 9.2.1. Motivated by the above example, we can have the following observation: Let $S \subset M$ be an arc-wise connected open subset of M. Then S is a sub-manifold of M with same dimension. So the controllability result obtained in Theorem 9.2.1 can be easily extended for the semi-global controllability over S.

Example 9.2.3. [6, 18] Consider the following system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -\sin(x_2)\cos(x_2) \\ \sin^2(x_2) \end{bmatrix} + \begin{bmatrix} \sin(x_2)e^{-x_1} \\ \cos(x_2)e^{-x_1} \end{bmatrix} u, \quad x \in \mathbb{R}^2. \tag{9.21}$$

[6] proved that it is not globally linearizable. [18] proved that it is not globally controllable. We claim that it is semi-global controllable on each strip

$$M_k = \{x \in \mathbb{R}^2 | k\pi - \pi/2 < x_2 < k\pi + \pi/2\}, \quad k \in \mathbb{Z}.$$

We prove it for $k = 0$ only. (Argument for other k's is the same.) A basis of $\mathcal{G}$ is $((\sin(x_2), \cos(x_2))^T$. A straightforward computation shows that the integral manifold passing through $x_0 \in M_0$ is

$$\begin{cases} x_1 = ln[\alpha^2 e^{2t} + 1] - t - ln[\alpha^2 + 1] + x_1(0), \\ x_2 = 2\tan^{-1}\left(\frac{\alpha e^t - 1}{\alpha e^t + 1}\right), \quad \text{where } \alpha = \frac{1+\sin(x_2(0))}{\cos(x_1(0))}. \end{cases}$$

Now since $x_2 \to -\pi/2$ $(t \to -\infty)$ and $x_2 \to \pi/2$ $(t \to \infty)$, there exists t_0 such that $x_2(t_0) = 0$. It is easy to see that as $t > t_0$ and $t < t_0$ f points to different sides of $\mathcal{I}_{\mathcal{G}}(x_0)$. So the system is semi-globally controllable on M_0. ■

Theorem 9.2.1 is based on G. It was shown in [8] that a switched system without control may still be controllable. In fact, we can prove the following corollary for switched systems similar to the proof of Theorem 9.2.1.

Corollary 9.2.1. *System (9.7) is controllable if every point in M is an interior point of F.*

Example 9.2.4.

$$\dot{x} = f_{\sigma(t)}(x), \quad x \in \mathbb{R}^n \backslash \{0\}, \tag{9.22}$$

where $\sigma \in \Lambda = \{1, 2, \cdots, n(n-1)+2\}$, and

$$f_i = B_i x, \quad f_{n(n-1)/2+i} = -B_i x, \quad i = 1, 2, \cdots, n(n-1)/2,$$

with $\{B_1, B_2, \cdots, B_{n(n-1)/2}\}$ a basis of $so(n, \mathbb{R})$, and the last two are $\pm I_n$. It is easy to see that at each $x \in M$ the $f_i(x)$, $i = 1, \cdots, n(n-1)$ spend a convex cone on the tangent space of S^{n-1} at x with x as it interior point (with respect to the sub-space topology of S^{n-1}). Adding last two vectors $f_{n(n-1)+1}(x)$ and $f_{n(n-1)+2}(x)$ as vectors at the $\pm$ radius directions make x an interior point of $\{F, \mathcal{G}\}$. So (9.20) is controllable. ■

This example shows that Corollary 9.2.14 is not necessary because it is easy to see that models $i = 1, 2, \cdots, n(n-1)/2$ plus last two are enough to make the system controllable.

9.3 General Case

Before giving main result we would like to give some motivations for considering hierarchical structure of the controllable set. First, consider linear system (9.1). Assume $m = 1$ and put it into Brunovsky Canonical form as

$$\begin{cases} \dot{x_1} = x_2 \\ \vdots \\ \dot{x}_{n-1} = x_n \\ \dot{x}_n = u, \quad x \in \mathbb{R}^n, u \in \mathcal{U}. \end{cases} \tag{9.23}$$

Check the condition of Theorem 9.2.1. Now along $\mathcal{I}_{\mathcal{G}}(x_0)$ the vector field has the form as

$$f = ((x_2)_0, \cdots, (x_{n-1})_0, x_n, 0)\,.$$

So following the moving style as proposed in Section 9.2 the only direction we can go out of the $\mathcal{I}_{\mathcal{G}}(x_0)$ is the x_{n-1} direction. This motivated us that what

proposed in Section 9.2 is only the "first step" of motion. To find all the ways the system can be driven, a sequence of motions should be followed. For linear case, it is not difficult to see that the first motion can reach B, AB. So the second motion will go to A^2B and so on. (Refer to Figure 9.2(a).)

Second motivation is from Lemma 9.2.1 and Lemma 9.2.2. In fact, it is obvious that in Lemma 9.2.1 the movable direction (out of $\mathcal{I}_G(x_0)$) is f. In the proof of Lemma 9.2.2, it was shown that, roughly speaking, in addition to the direction of f the additional direction is $ad_G^k f$ (see the Remark after the proof of Lemma 9.2.2.) In fact, the directions the trajectory can really go are $\left(\phi_t^{\mathcal{G}}\right)_* f$. Using Baker-Campbell-Hausdorff formula [2] for any $X \in \mathcal{G}$ yields

$$\left(\phi_t^X\right)_* f(x_0) = f(x_0) + ad_X f(x_0)t + \frac{1}{2!} ad_X^2 f t^2 + \cdots . \tag{9.24}$$

This equation clearly shows that all the possible direction for the first step of motion is $ad_{\mathcal{G}}^k f$, $k \geq 0$. Similar to linear case, we may consider $[g_*(f)]_* f$ as the next step and keep going on. (Refer to Figure 9.2(b).)

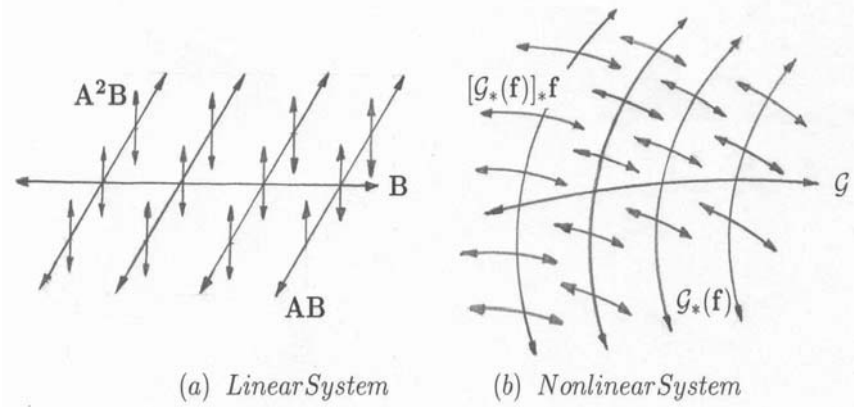


Fig. 9.2. Controllability distributions

Motivated by the aforementioned argument, we will construct a sequence of distributions for the controllability of system (9.7). To begin with, we need some preparations:

1. Let $\Delta_1 \subset \Delta_2 \subset \cdots \subset \Delta_{k^*} = T(M)$ be a sequence of nested involutive distributions of constant dimensions, $\dim(\Delta_i) = n_i$. Using Frobenious' theorem, it is easy to prove that [16] there is a flat coordinate chart (U, x), such that locally
$$\Delta_i = Span\left\{\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_{n_i}}\right\}, \quad i = 1, \cdots, k^*.$$
2. Let Δ be an involutive distribution of dimension k, and x be a flat coordinate frame, such that $\Delta = \left\{\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_k}\right\}$. Assume a vector field, $f(x)$ is locally expressed as $f = (f_1, \cdots, f_n)^T$. Then the projection of f on Δ is defined as

$$f|_{\Delta} = f - icl_*(f/\Delta),$$

where $icl : (\mathcal{I}_{\Delta})^{\perp} \hookrightarrow M$ is the inclusion mapping. Precisely, $f/\Delta = (f_{k+1}(x), \cdots, f_n(x))^T$, $icl_*(f/\Delta) = (0, \cdots, 0, f_{k+1}(x), \cdots, f_n(x))^T$, and $f|_{\Delta} = (f_1(x), \cdots, f_k(x), 0, \cdots, 0)^T$.

Now we can define the following sequence of distributions, called the set of controllability distributions, as:

$$\begin{cases} \Delta_1 := \mathcal{G} \\ \Delta_{k+1} := \bar{\Delta}_k + Span\left\{ ad^j_{\bar{\Delta}_k} F \,\middle|\, j \geq 1 \right\}, \quad k = 1, 2, \cdots. \end{cases} \tag{9.25}$$

Definition 9.3.1. *System (9.7) is said to have a proper set of controllability distributions if the sequence of distributions Δ_i defined in (9.25) satisfy the following conditions:*

1. *There exist k^* such that $\bar{\Delta}_{k^*} = T(M)$;*
2. *$\bar{\Delta}_1, \cdots, \bar{\Delta}_{k^*-1}$ have constant dimensions.*

It is worthwhile noting that for linear systems $\Delta_1 = Span\{B\}$, $\Delta_2 = Span\{B,\ AB\}$, $\cdots$.

Definition 9.3.2. *Let Δ be an involutive distribution on M with constant dimension. A point $x \in \mathcal{I}_{\Delta}(x_0)$ is said to be a deep interior point of $\{(\Delta)_*(F), \Delta\}$ if for any $r > 0$ there exist finite vectors $Y_i(x_0) \in \{(\Delta)_*(F), \Delta\}$, $i = 1, \cdots, s$ such that the whole ball $B_r(0)$ is in the interior of the convex cone: $con\{Y_i(x_0) | \, i = 1, \cdots, s\}$.*

Note that here the radius r can be arbitrary large, but assigned before choosing Y_i.

Now we are ready to state our main result.

Theorem 9.3.1. *Consider system (9.7). Assume it has a proper set of controllability distributions defined in (9.25), which satisfy the following conditions (for each $x_0 \in M$):*

1. *There exists $x_i \in \mathcal{I}_{\bar{\Delta}_i}(x_0)$ (equivalently, for any point $x \in \mathcal{I}_{\bar{\Delta}_i}(x_0)$), such that x_i is a deep interior point of $\left\{(\bar{\Delta}_i)_*(F|_{\bar{\Delta}_{i+1}}), \bar{\Delta}_i\right\}$, with respect to the sub-space topology of $\mathcal{I}_{\bar{\Delta}_{i+1}}(x_0)$, for $i = 1, \cdots, k^* - 2$.*
2. *There exists $x_{k^*-1} \in \mathcal{I}_{\bar{\Delta}_{k^*-1}}(x_0)$ (equivalently, for any point $x \in \mathcal{I}_{\bar{\Delta}_{k^*-1}}(x_0)$), such that x_{k^*-1} is an interior point of $\left\{(\bar{\Delta}_{k^*-1})_*(F), \bar{\Delta}_{k^*-1}\right\}$.*

Then system (9.7) is globally controllable.

The proof is in Appendix 9.C. We now give some examples:

Example 9.3.1. 1. For linear systems Theorem 9.3.1 becomes necessary and sufficient. To see this we have only to prove the necessary. Consider a controllable

linear system. We prove it for single-input case. The proof for multi-input case is exactly the same. Assume the system is already in the canonical form (9.23). First, the set of controllability distributions are easily computed as

$$\begin{aligned}
\Delta_1 &= Span\{b\} = Span\{\delta_n\}\\
\Delta_2 &= Span\{b, Ab\} = Span\{\delta_n, \delta_{n-1}\}\\
&\vdots\\
\Delta_n &= Span\{b, Ab, \cdots, A^{n-1}b\} = T(\mathbb{R}^n),
\end{aligned}$$

where δ_i is the i-th column of the identity matrix I_n. Now the restriction of $f = Ax$ on Δ_2 is $f|_{\Delta_2} = (0, \cdots, 0, x_n, 0)^T$ and it is obvious that corresponding to the sub-space topology of $\mathcal{I}_{\Delta_2}$ the $f|_{\Delta_2}$ can approach to $\pm\infty$. That is, obviously, any point on $\mathcal{I}_{\Delta_1}$ is a deep interior of $\{(\Delta_1)_*(F|_{\Delta_2}), \Delta_1\}$. Continuing the same argument shows that the conditions of Theorem 9.3.1 are verified. We conclude that the conditions of Theorem 9.3.1 are necessary and sufficient for linear systems.

2. Consider the following system

$$\begin{cases}
\dot{x}_1 = \sin(x_2)\\
\dot{x}_2 = x_3\\
\vdots\\
\dot{x}_{n-1} = x_n\\
\dot{x}_n = u.
\end{cases} \tag{9.26}$$

Similar to 1, one sees easily that the conditions of Theorem 9.3.1 are verified. So (9.26) is globally controllable. (Note that only at last step for $x \in \mathcal{I}_{\Delta_{n-1}}$, it is only an interior (not deep interior) of $\{(\Delta_{n-1})_*(F), \Delta_{n-1}\}$. But it is enough.) ■

Example 9.3.2. Consider the switched linear system (9.6). It is easy to see that

$$\Delta_1 = Span\{\, B_i \,|\, 1 \le i \le N\};$$

$$\Delta_2 = \Delta_1 + Span\{\, A_{i_1} B_{i_2} \,|\, 1 \le i_1, i_2 \le N\};$$

and so on. As argued in linear (non-switched) case, $x \in \mathcal{I}_{\Delta_i}$ is always a deep interior of $\big\{(\Delta_i)_*(F|_{\Delta_{i+1}}), \Delta_i\big\}$. (In fact, if a straight line is not in a subspace, we can always find two points which are on the opposite directions and as far from the subspace as we wish.) So the only condition we have to check is: Whether there exists a k^* such that $\Delta_{k^*} = \mathbb{R}^n$. This argument leads to the following known result [19]. ■

Proposition 9.3.1. *The switched linear system is controllable, iff*

$$\Delta_n = Span\{B_{i_1^1}, A_{i_1^2} B_{i_2^2}, \cdots, A_{i_1^n} A_{i_2^n} \cdots A_{i_{n-1}^n} B_{i_n^n} \Big| \, 1 \le i_k^j \le N\} = \mathbb{R}^n. \tag{9.27}$$

Example 9.3.3. Consider the following system

$$\begin{cases} \dot{x}_1 = x_2^{q_1} \\ \vdots \\ \dot{x}_{n-1} = x_n^{q_{n-1}} \\ \dot{x}_n = u, \end{cases} \tag{9.28}$$

where $q_1, \cdots, q_{n-1}$ are positive odd integers. Arguing as for linear systems, it is easy to prove that this system is controllable.

Moreover, if $x_i^{q_i}$ are replaced by $x_i^{q_i} + LOT(x_i)$, the conclusion remains true. (LOT: lower order terms.) ■

Example 9.3.4. Consider the following system

$$\dot{x} = \begin{bmatrix} \cos(x_2 + x_3) \\ x_1 \\ x_4^3 \\ \sin(x_2 + x_3) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} u. \tag{9.29}$$

Calculate that

$$\Delta_1 = \mathcal{G} = Span\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}; \quad \Delta_2 = Span\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right\}.$$

Then it is easy to see that Δ_1 is in the deep interior of $\{(\Delta_1)_* f|_{\Delta_2}, \Delta_1\}$. Moreover, it is also easy to check $\bar{\Delta}_3 = T(\mathbb{R}^4)$. So we have only to check whether Δ_2 is in the interior of $\{(\Delta_2)_* f, \Delta_2\}$.

At $p_1 = (0, \pi/2, 0, 0)^T$ and $p_2 = 0$ we have

$$f(p_1) = (0, 0, 0, 1)^T, \quad f(p_2) = (1, 0, 0, 0)^T.$$

Choosing $g_{1,2} = \pm(0, 1, 0, 0)^T$ and $g_{3,4} = \pm(0, 0, 1, 0)^T$ $g_5 = (1, 0, 0, 1)^T$ and setting $Z = g_1 = (0, 1, 0, 0)^T$. $\phi^Z_{\pi/2}(p_2) = p_1$, we have

$$\left(\phi^Z_{\pi/2}\right)_* f(p_2) = (1, 0, 0, 0)^T,$$

then we have

$$f(p_1) + \left(\phi^Z_{\pi/2}\right)_* f(p_2) + (-g_5) + g_1 + g_2 + g_3 + g_4 = 0.$$

Since all $\pm g_i \in \mathcal{G}$, using Proposition 9.2.1, one sees that Δ_2 is in the interior of $T(\mathbb{R}^4)$. We conclude that the system (9.29) is controllable. ■

9.4 $Codim(\mathcal{G}) = 1$

This section considers the case when $Codim(\mathcal{G}) = 1$. We will do two things for this particular case. First, simplify the sufficient condition. Secondly, show that under certain mild restriction on systems the condition is also necessary.

Appendix 9.D gives a brief review on some related concepts, such as tensor field, orientation etc., which are used in the sequel.

Theorem 9.4.1
1. Assume for each $x_0 \in M$ there exist $\sigma_1 \in \Omega^1(M)$, $\sigma_G \in \Omega^{n-1}(M)$, where σ_G is an orientation of $\mathcal{I}_{\mathcal{G}}(x_0)$, and $\sigma_1(Z) = 0$, $\forall Z \in \mathcal{G}$, such that $\sigma = \sigma_1 \wedge \sigma_G$ is an orientation of M. Moreover, on each $\mathcal{I}_{\mathcal{G}}(x)$ there are two vector fields $f_1,\ f_2 \in F$ and two points $x_1,\ x_2 \in \mathcal{I}_G(x_0)$ such that

$$i_{f_1}(\sigma)|_{\mathcal{G}}(x_1) = c_1\sigma_G(x_1), \quad i_{f_2}(\sigma)|_{\mathcal{G}}(x_2) = c_2\sigma_G(x_2), \tag{9.30}$$

where $f_1,\ f_2 \in F$ and $c_1c_2 < 0$. Then the system (9.7) is globally controllable.
2. If, in addition, each $\mathcal{I}_{\mathcal{G}}(x_0)$ separates M, then (9.30) is also necessary.

Proof. In Appendix 9.E.

Some conditions in Theorem 9.4.1 are related. We refer to [12] for verification and possible simplification of conditions of Theorem 9.4.1.

Corollary 9.4.1. *Assume $M = \mathbb{R}^n$. Let $\mathcal{G}(x_0)$, $\forall x_0 \in M$, be an $n-1$ hyperplane, and $Z_1, \cdots, Z_{n-1}$ be a basis of $\mathcal{G}$. Then a necessary and sufficient condition for the system (9.7) to be globally controllable is that on each $\mathcal{I}_{\mathcal{G}}(x_0)$, $x_0 \in M$, there are two vector fields, $f_1,\ f_2 \in F$, and two points $x_1, x_2 \in \mathcal{I}_{\mathcal{G}}(x_0)$ (same f_i and/or x_i are allowed), such that*

$$\det(f_1, Z_1, \cdots, Z_{n-1})(x_1)\det(f_2, Z_1, \cdots, Z_{n-1})(x_2) < 0. \tag{9.31}$$

Proof. In Appendix 9.F.

Example 9.4.1. Consider the following system

$$\dot{x} = \begin{bmatrix} x_2^2 + x_3 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u_1 + \begin{bmatrix} 0 \\ \sin(x_4) \\ \cos(x_4) \\ g_4^2 \end{bmatrix} u_2. \tag{9.32}$$

It is easy to check that

$$\mathcal{G} = \left\{ \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_4} \right\}.$$

A basis of $\mathcal{G}$ is

$$Z_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}; \quad Z_2 = \begin{bmatrix} 0 \\ \sin(x_4) \\ \cos(x_4) \\ 0 \end{bmatrix}; \quad Z_3 = [Z_1, Z_2] = \begin{bmatrix} 0 \\ \cos(x_4) \\ -\sin(x_4) \\ 0 \end{bmatrix}.$$

Mover, $\mathcal{I}_{\mathcal{G}}(x_0) = \left\{ x \in \mathbb{R}^4 \middle| \, x_4 = x_4(x_0) \right\}$, and

$$\det(f, Z_1, Z_2, Z_3) = x_2^2 + x_3.$$

Obviously, (9.31) is satisfied for some $x_2, x_3 \in \mathcal{I}_G(x_0)$. So system (9.32) is controllable.

9.5 Conclusion

The paper provides a sufficient condition for global controllability of switched affine nonlinear systems. It is applicable to switched linear systems and non-switched affine nonlinear systems. The condition is a generalization of $\{B, AB, \cdots, A^{n-1}B\}$ type of criteria. For linear (switched or not) systems it becomes a necessary and sufficient condition. In $Codim(\mathcal{G}) = 1$ case, with mild additional restriction, it is also necessary and sufficient.

The author's conjecture is: with possible mild restrictions the condition is also necessary in general.

References

1. Abraham, R., Marsden, J.E.: Foundations of Mechanics, 2nd edn. Benjamin/Cummings, Massachusetts (1978)
2. Abraham, R., Marsden, J.E., Ratiu, T.: Manifolds, Tensor Analysis and Applications. Springer, New York (1988)
3. Boothby, W.M.: An Introduction to Differential Manifolds and Riemannian Geometry, 2nd edn. Academic Press, London (1986)
4. Brockett, R.W.: Nonlinear systems and differential geometry. IEEE Trans. Aut. Contr. 64(1), 61–72 (1976)
5. Cheng, B., Zhang, J.: Robust controllability for a class of uncertain linear time-invariant MIMO systems. IEEE Trans. Aut. Contr. 49(11), 2022–2027 (2004)
6. Cheng, D., Tarn, T.J., Isidori, A.: Global external linearization of nonlinear systems via feedback. IEEE Trans. Aut. Contr. 30(8), 808–811 (1985)
7. Cheng, D.: Design for noninteractiong decomposition of nonlinear systems. IEEE Trans. Aut. Contr. 33(11), 1070–1074 (1988)
8. Cheng, D., Chen, H.: Accessibility of Switched Linear Systems. In: Proc. 42nd IEEE CDC 2003, Mauii, pp. 5759–5764 (2003)
9. Cheng, D.: Controllability of switched bilinear Systems. IEEE Trans. Aut. Contr. 50(4), 511–515 (2005)
10. Clarke, F.H., Stern, R.J.: Lyapunov and feedback characterizations of state constrained controllability and stabilization. Sys. Contr. Lett. 54(8), 747–752 (2005)

11. Hermann, R.: Differential Geometry and the Calculus of Variations. Academic Press, London (1968)
12. Hirsch, M.W.: Differential Topology. Springer, Heidelberg (1976)
13. Jurdjevic, V.: Geometric Control Theory. Cambridge Univ. Press, Cambridge (1997)
14. Olver, P.J.: Applications of Lie Groups to Differential Equations, 2nd edn. Springer, New York (1993)
15. Respondek, J.: Controllability of dynamical systems with constraints. Sys. Contr. Lett. 54(4), 293–314 (2005)
16. Respondek, W.: On decomposition of nonlinear control systems. Sys. Contr. Lett. 1, 301–308 (1982)
17. Spivak, M.: A Comprehensive Introduction to Differential Geometry. Publish or Perish Inc, Berkeley (1979)
18. Sun, Y., Guo, L.: On controllability of some classes of affine nonlinear systems. In: Glas, T., Hendeby, G. (eds.) Forever Ljung in System Identification, pp. 127–146. Student literature, Lund (2006)
19. Sun, Z., Ge, S.S., Lee, T.H.: Controllability and reachability criteria for switched linear systems. Automatica 38, 775–786 (2002)
20. Sussmann, H.J., Jurdjevic, V.: Controllabbility of nonlinear systems. Journal of Differential Equations 12(1), 95–116 (1972)
21. Susmann, H.J.: A general theorem on local controllability. SIAM J. Contr. Opt. 25, 158–194 (1987)
22. Wonham, W.M.: Linear Multivariable Control, A Geometric Approach. Springer, Berlin (1974)
23. Xie, G., Wang, L.: Necessary and sufficient conditions for controllability and observability of switched impulsive control systems. IEEE Trans. Aut. Contr. 49(6), 960–966 (2004)

Appendices

9.A Proof of Lemma 9.2.1

To prove Lemma 9.2.1, we need some preparations:

Lemma 9.A.1. *Let $G = \{g_1, \cdots, g_s\}$ and $\mathcal{G} = \{G\}_{LA}$ be of constant dimension m and $x_0 \in \mathcal{I}_G(x_0)$. Then there exist m vector fields $\{X_1, \cdots, X_m\} \subset G$ (since $m \geq s$, some X_i may be duplicated), and $t_1^0, \cdots, t_m^0$, such that the mapping*

$$\Phi(t_1, \cdots, t_m) := \phi_{-t_1^0}^{X_1} \circ \cdots \circ \phi_{-t_m^0}^{X_m} \circ \phi_{t_m^0+t_m}^{X_m} \circ \cdots \circ \phi_{t_1^0+t_1}^{X_1}(x_0) \tag{9.33}$$

is a diffiomorphism from a neighborhood $0 \in \sqcup_\delta^m \subset \mathbb{R}^m$ ($0 < \delta << 1$) to a neighborhood $x_0 \in W \subset \mathcal{I}_G(x_0)$.

Proof. Choose any $X_1 \in G$ such that $X_1(x_0) \neq 0$. Construct a mapping

$$\Phi_1(t) := \phi_t^{X_1}(x_0),$$

which is a diffiomprphism from $0 \in \sqcup_\delta^1 \subset \mathbb{R}^1$ onto its image. Since $dim(\mathcal{I}_\mathcal{G}(x_0)) = m$, there exists a $t_1^0 \in \sqcup_\delta^1$ and a vector field $X_2 \in G$ such that at point $p_1 = \Phi_1(t_1^0)$, $X_1(p_1)$ and $X_2(p_1)$ are linearly independent. (Otherwise, along the integral curve of X_1 $\dim(\mathcal{I}_G(x_0)) = 1$, which is a contradiction.) Next, we construct a mapping

$$\Phi_2(t_1, t_2) := \phi_{t_2}^{X_2} \circ \phi_{t_1^0+t_1}^{X_1}(x_0),$$

which is a diffiomorphism from $0 \in \sqcup_\delta^2 \subset \mathbb{R}^2$ onto its image. Similar argument shows that there exists a point $p_2 = \Phi_2(t_1^0 + \delta t_1^0, t_0^2)$, and a $X_3 \in G$, such that $X_1(p_2)$, $X_2(p_2)$, and $X_3(p_2)$ are linearly independent. For notational ease, we still use t_1^0 for $t_1^0 + \delta t_1^0$. Continuing this procedure, we can finally construct a mapping as

$$\Phi_m(t_1, \cdots, t_m) := \phi_{t_m^0+t_m}^{X_m} \circ \phi_{t_{m-1}^0+t_{m-1}}^{X_{m-1}} \circ \cdots \circ \phi_{t_1^0+t_1}^{X_1}(x_0) \tag{9.34}$$

which is a diffiomorphism from $0 \in \sqcup_\delta^m \subset \mathbb{R}^m$ to a neighborhood of $p_m = \phi_{t_m^0}^{X_m} \circ \phi_{t_{m-1}^0}^{X_{m-1}} \circ \cdots \circ \phi_{t_1^0}^{X_1}(x_0)$ in $\mathcal{I}_\mathcal{G}(x_0)$. (Note that according to our construction $t_m^0 = 0$.) Define

$$\Phi^c(x) := \phi_{-t_1^0}^{X_1} \circ \cdots \circ \phi_{-t_m^0}^{X_m}(x),$$

which is a diffiomorphism from a neighborhood of p_m back to a neighborhood of x_0. Composing it with $\Phi_m(t_1, \cdots, t_m)$ yields

$$\Phi(t_1, \cdots, t_m) := \Phi^c \circ \Phi_m(t_1, \cdots, t_m)$$

which is a diffiomorphism from $0 \in \sqcup_\delta^m \subset \mathbb{R}^m$ to $x_0 \in W \subset \mathcal{I}_\mathcal{G}(x_0)$. ∎

Remark 9.A.1. Basically, the above Lemma is a mimic of the proof of Chow's Theorem [11].

Lemma 9.A.2. *If there exists a finite set $\{Y_1, \cdots, Y_s\} \subset S$ such that conditions (i) and (ii) of Proposition 9.2.1 are satisfied, then for any $Z \in S$, we can assume it is a vector in (9.9), i.e., $Z = Y_i$ for some i.*

Proof. Since $Y_1, \cdots, Y_n$ are linearly independent, we can express Z as

$$Z(x_0) = \sum_{i=1}^{n} d_i Y_i(x_0). \tag{9.35}$$

Adding (9.35) to (9.9) yields

$$Z(x_0) + \sum_{i=1}^{n} (c_i - d_i) Y_i(x_0) + \sum_{i=n+1}^{s} c_i Y_i(x_0) = 0. \tag{9.36}$$

Now if all the coefficients are positive we are done. Since in (9.9) we can choose $c_i > 0$ as large as we wish the conclusion follows. ∎

Definition 9.A.1. *A mapping $\phi_t^X(x) : M \times \mathbb{R}^1 \to M$ is called (physically) realizable (by system (9.7)) if*

$$X \in \left\{ f^i + \sum_{j=1}^{m} g_j^i u_j \,\middle|\, i = 1, \cdots, N; u_j \in \mathcal{U} \right\}$$

and $t \geq 0$.

Now we are ready to prove Lemma 9.2.1.

Proof of Lemma 9.2.1. According to Proposition 9.2.1, we can find $Y_1, \cdots, Y_s \in \{F, \mathcal{G}\}$, such that (9.9) holds. Moreover, there is a subset of n vector fields, say, $Y_1, \cdots, Y_n$, which are linear independent at x_0 (and hence independent over a neighborhood $W \ni x_0$). Construct a mapping

$$\Theta(t_1, \cdots, t_n) := \phi_{t_1}^{Y_1} \circ \phi_{t_2}^{Y_2} \circ \cdots \circ \phi_{t_n}^{Y_n}(x_0), \tag{9.37}$$

which is a diffeomorphism from $0 \in \sqcup_\delta^n \subset \mathbb{R}^n$ to a neighborhood $x_0 \in W \subset M$. Note that the region of $\Theta(t_1, \cdots, t_n)(x_0)$ contains a neighborhood of x_0. So if each segment of its integral curves is realizable, then the reachable set $R(x_0)$ of x_0 contains a neighborhood of x_0. The following is devoted to modify Θ to make it realizable.

Consider Y_i, which should be in one of the following three categories:

Case 1. $Y_i \in F$:
According to Lemma 9.A.2, we can assume that this Y_i is an element in (9.9). Say, it is the Y_i in (9.9). Then we construct the following mapping

$$\psi_i(t) := \begin{cases} \phi_t^{Y_i}(x_0), & t \geq 0, \\ \phi_{-\frac{c_1}{c_i}t}^{Y_1} \circ \cdots \circ \phi_{-\frac{c_{i-1}}{c_i}t}^{Y_{i-1}} \circ \phi_{-\frac{c_{i+1}}{c_i}t}^{Y_{i+1}} \circ \cdots \phi_{-\frac{c_n}{c_i}t}^{Y_n}(x_0), & t < 0. \end{cases} \tag{9.38}$$

We prove that

$$\left.\frac{d\psi_i(t)}{dt}\right|_{t=0} = Y_i(0). \tag{9.39}$$

It is obvious that

$$\left.\frac{d\psi_i(t)}{dt}\right|_{t=0^+} = Y_i(0).$$

So we have only to consider the case of $t \to 0^-$. Denote

$$\psi_i^-(t) := \phi_{-\frac{c_1}{c_i}t}^{Y_1} \circ \cdots \circ \phi_{-\frac{c_{i-1}}{c_i}t}^{Y_{i-1}} \circ \phi_{-\frac{c_{i+1}}{c_i}t}^{Y_{i+1}} \circ \cdots \phi_{-\frac{c_n}{c_i}t}^{Y_n}(x_0).$$

Using the fact that $\phi_0^X(p) = p$, $(\phi_0^X)_* = identity$, and an easily proved formula

$$\frac{\partial}{\partial t_2}\phi_{t_1}^X \circ \phi_{t_2}^Y(x) = (\phi_{t_1}^X)_* Y(\phi_{t_1}^X \circ \phi_{t_2}^Y(x)) \tag{9.40}$$

we can prove that

$$\left.\frac{\partial}{\partial t}\psi_i^-(t)\right|_{t=0} = -\sum_{j=1, j\neq i}^{s} \frac{c_j}{c_i} Y_j(x_0) = Y_i(x_0).$$

The last equality is from (9.9). We hence proved (9.39).

Recall Θ of (9.37). If there is a integral segment of $Y_i \in F$ as its component , we replace $\phi_{t_i}^{Y_i}$ by $\psi_i(t_i)$. Then we have a new Θ. For notational easy, we still denote it by Θ. We claim that new $\Theta(t_1, \cdots, t_n)$ is still a diffiomorphism from $0 \in \sqcup_\delta^n \subset \mathbb{R}^n$ to a neighborhood $x_0 \in W \subset M$. ($\sqcup_\delta^n$ may be shrank if necessary.) Using (9.39), it is easy to prove that we still have the Jacobian matrix of Θ at zero as

$$J_\Theta(0) = [Y_1(x_0), Y_2(x_0), \cdots, Y_n(x_0)].$$

The linear independence of $\{Y_i\}$ implies that Θ is a local diffiomorphism. The advantage of new Θ is: any segment of integral curves in Θ, which involves $Y_i \in F$, is now realizable because of its corresponding non-negative time. (Recall (9.38) and note that all $c_i > 0$.)

Case 2. $Y_i \in G$:
Say $Y_i = g_j^i$. Then we have

$$\phi_{t_i}^{Y_i} = \phi_{t_i}^{\frac{(f^i + g_j^i u) - f^i}{u}}. \tag{9.41}$$

When $t_i \geq 0$ choose corresponding control $u > 0$ and when $t_i < 0$ let $u < 0$. Then we have

$$\phi_{t_i}^{Y_i} = \phi_{t_i/u}^{(f^i + g_j^i u) - f^i}. \tag{9.42}$$

Now we define

$$\psi_i(t) := \phi_{t/u}^{(f^i+g_j^i u)}, \tag{9.43}$$

where $t/u \geq 0$. So $\psi_i(t)$ is also physically realizable. Replacing $\phi_{t_i}^{Y_i}$ in Θ by this $\psi_i(t_i)$ as $Y_i \in G$, we can prove that as $|u|$ large enough now Θ is still a local diffiamorphism. It is because

$$\left.\frac{\partial \psi_i(t)}{\partial t_i}\right|_{t=0} = Y_i(0) + f^i(0)/u.$$

As $|u|$ large enough, similar to Case 1, we can prove that the Jacobian matrix for modified Θ is still non-singular at $t = 0$.

Case 3. $Y_i \in \mathcal{G}\backslash G$:
Assume that the region $W = \Theta(\sqcup_\delta^n)$ and $\Theta : \sqcup_\delta^n \to W$ is a diffiomorphism. We claim that W is a reachable set of x_0. Let $x_1 \in W$. It suffices to show that $x_1 \in R(x_0)$. After components of Case 1 and Case 2 have been treated, we have only to treat the components of Case 3. Let $Y_i \in \mathcal{G}\backslash G$. Then there exist $t^0 = (t_1^0, \cdots, t_n^0) \in \sqcup_\delta^n$, such that

$$x_1 = \Theta(t^0) = \phi_{t_1^0}^{Y_1} \circ \phi_{t_2^0}^{X_2} \circ \cdots \circ \phi_{t_n^0}^{Y_n}(x_0). \tag{9.44}$$

Since $\sqcup_\delta^n$ is an open set, we can find $\epsilon > 0$ such that if $t = (t_1, \cdots, t_n)$ satisfying $\|t\| < \epsilon$, then $t^0 + t \in \sqcup_\delta^n$. Define

$$\tilde{\Theta}(t) = \phi_{t_1^0+t_1}^{Y_1} \circ \phi_{t_2^0+t_2}^{X_2} \circ \cdots \circ \phi_{t_n^0+t_n}^{Y_n}(x_0), \quad \|t\| < \epsilon. \tag{9.45}$$

Checking the Jacobian matrix, it is easy to see that $\tilde{\Psi}$ is a diffiomorphism from a neighborhood $0 \in \sqcup_\epsilon^n \subset \mathbb{R}^n$ to a neighborhood of $x_1 \in W_1 \subset W$. Denote by

$$p = \phi_{t_{i+1}^0}^{Y_{i+1}} \circ \cdots \circ \phi_{t_n^0}^{Y_n}(x_0); \quad q = \phi_{t_i^0}^{Y_i}(p).$$

Since $Y_i \in \mathcal{G}$, by Chow's Theorem, we can find $Z_1, \cdots, Z_j \in G$ and $d_1, \cdots, d_j \in \mathbb{R}$, such that

$$q = \phi_{d_1}^{Z_1} \circ \phi_{d_2}^{Z_2} \circ \cdots \circ \phi_{d_j}^{Z_j}(p)$$

Using Lemma 9.A.2, we can construct a $\Phi(\tau_1, \cdots, \tau_m)$ as in (9.33) which is a local diffiomorphism from $\sqcup_\delta^m$ to a neighborhood of q (with respect to the inherited topology from M to the sub-manifold $\mathcal{I}_{\mathcal{G}}(q)$). Now define

$$\psi_i(\tau_1, \cdots, \tau_m) := \Phi(\tau_1, \cdots, \tau_m) \circ \phi_{d_1}^{Z_1} \circ \phi_{d_2}^{Z_2} \circ \cdots \circ \phi_{d_j}^{Z_j}(p)$$

Note that the region of $\psi_i(\tau_1, \cdots, \tau_m)(p)$ over $\sqcup_\delta^m$ contains a neighborhood of q (under the sub-manifold topology). While $\phi_{t_i^0+t_i}^{Y_i}(p)$, $t_i \in \sqcup_\epsilon^1$ contains a special integral curve of Y_i on the sub-manifold. As $\epsilon > 0$ small enough, we conclude that

$$\left\{ \phi_{t_i^0+t_i}^{Y_i}(p) \,\middle|\, t_i \in \sqcup_\epsilon^1 \right\} \subset \{ \psi_i(\tau) |\, \tau \in \sqcup_\delta^m \}$$

Note that though $\psi_i(\tau_1,\cdots,\tau_m)$ is defined at q, i.e. $\psi_i(\tau_1,\cdots,\tau_m)(q)$ is a local diffiomorphism, by continuity of the Jacobian matrix one sees easily that as a neighborhood W_q of q being small enough, $\psi_i(\tau_1,\cdots,\tau_m)(q')$, $q'\in W_q$, is also a diffiomprphis. In our case since $\sqcup_\epsilon^n$ can be arbitrary small, $\psi_i(\tau_1,\cdots,\tau_m)$ is diffiomorphism for each $q'\in W_q$.

Now we replace $\phi^{Y_i}_{t_i^0+t_i}$ in $\tilde{\Theta}$ by $\psi_i(\tau_1,\cdots,\tau_m)$. Then according to the above argument one sees easily that the region of new $\tilde{\Theta}$ over $\sqcup_\epsilon^{i-1}\times\sqcup_\delta^m\times\sqcup_\delta^{n-i}$ cover the region of the old $\tilde{\Theta}$ over $\sqcup_\epsilon^n$. Therefore, it covers a neighborhood of x_1. After this modification we can replace all $Y_i\in\mathcal{G}\backslash G$ by some $Z\in G$. Then using the technique in Case 2, we can further modify $\tilde{\Theta}$ to make it realizable. We conclude that W is a reachable set of x_0.

Note that unlike Case 1 and Case 2, in Case 3 for each $x_1\in W$ we define $\tilde{\Theta}=\tilde{\Theta}_{x_1}$, which depends on x_1 and is a submersion (onto mapping) to a neighborhood of x_1. It is not a diffiomorphim any more.

Finally, we have to show that W is a controllable neighborhood of x_0. To see that let $x_1\in W$. Construct $\tilde{\Theta}(t)=\tilde{\Theta}_{x_1}(t)$ which start from x_0 and mapped over an neighborhood of x_1. Replace all the constant time variables t_i^0 by $-t_i^0$ to get $\tilde{\Theta}^c(t)$. Then it maps from x_1 back to a neighborhood of x_0. Note that in $\tilde{\Theta}(t)$ there is no vector field of Case 3, then so is $\tilde{\Theta}^c(t)$. Using the tricks used in Case 1 and Case 2 for handling negative time, we can show that $x_0\in R(x_1)$, which completes the proof. ■

9.B Proof of Lemma 9.2.2

Exactly follow the same procedure proposed in the proof of Lemma 9.2.1, one sees easily that we can find an open neighborhood W of x_0 and for each $x_1\in W$ construct a $\tilde{\Theta}:\sqcup_\delta^\ell\to M$, which is composed by a set of integral curves, with the region coving a neighborhood $x_1\in W_1\subset W$. Note that where $\ell\geq n$ might be a very large integer. Precisely, $\tilde{\Theta}$ is composed of the integral curves of two categories of vector fields:

First group of vector fields are as $f^k+g_j^k$ with non-negative time, which is realizable;

Second group of vector fields are as $\tilde{f}^i\in\mathcal{G}(x_0,x_i)_*(f^i)$ with non-negative time. That is, there exist $Z_1,\cdots,Z_j\in G$ such that

$$\tilde{f}^i=\left(\phi^{Z_1}_{t_1^0}\right)_*\circ\cdots\circ\left(\phi^{Z_j}_{t_j^0}\right)_* f^i \tag{9.46}$$

Without loss of generality and for notational ease, we assume $j=1$, that is

$$\tilde{f}^i=\left(\phi^{Z}_{t^0}\right)_* f^i, \tag{9.47}$$

where $Z\in G$. Now we have to make $\phi^{\tilde{f}^i}_{t_i}(\tilde{x}_0)$ realizable, where $\tilde{x}_0$ is a point near x_0 (corresponding to the "p" in the proof of Lemma 9.2.1).

A simple computation shows that

$$\phi_{t^0}^{\tilde{f}_i}(\tilde{x}_0) = \phi_{t_i}^{(\phi_{t^0}^Z)_* f^i}(\tilde{x}_0) = \phi_{t^0}^Z \circ \phi_{t_i}^{f^i}\left(\phi_{-t^0}^Z(\tilde{x}_0)\right) = \phi_{t^0}^Z \circ \phi_{t_i}^{f^i} \circ \phi_{-t^0}^Z(\tilde{x}_0). \quad (9.48)$$

As discussed in the Case 2 of the proof of Lemma 9.2.1, we can easily reply the integral of $Z \in G$ in (9.48) by some realizable vector fields. Then we proved that any point x_1 in W is reachable from x_0.

Note that after the treatment of (9.48), there is no component involving $\tilde{f}^i$. So showing $x_0 \in R(x_1)$ is exactly the same as in the proof of Lemma 9.2.1. ■

Remark 9.B.1. The physical meaning of (9.48) is: instead of going along the direction of $\tilde{f}^i$, we can go along $-Z$ from $\tilde{x}_0$ to $\tilde{x}_1$, then along f from $\tilde{x}_1$ to A, and then along Z from A to B. (Refer to Figure 9.2) Now let's see what is the direction of $\tilde{f}^i$? In fact, if we set $t_i = t_0 := t$ and let the trajectory continue to go from B along $-\tilde{f}^i$ for time t to C, then it is well known [4]that the vector from $\tilde{x_0}$ to C is $ad_Z \tilde{f}^i + O(t^2)$. (Ref. Figure 9.3) We conclude that

$$(\phi_t^Z)_* f^i \approx f^i + t \cdot ad_z f^i. \quad (9.49)$$

For more than one shifting (see (9.46)) $ad_Z^k f^i$, $k > 1$ appear. This tells us what direction $\mathcal{G}_* F$ can really go.

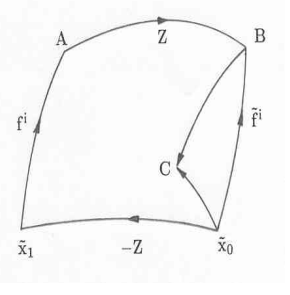


Fig. 9.3. Integral Curve of $\mathcal{G}(x_0, x_i)_*(F)$

9.C Proof of Theorem 9.3.1

First considering $\bar{\Delta}_{k^*-1}$ as $\mathcal{G}$ and similar to the proofs of Lemmas 9.2.1 and 9.2.2, it is easy to find for each $x_0 \in M$ a mapping Θ as in (9.37), such that $\Theta : \sqcup_\delta^n \to W \subset M$ is a diffiomorphism, where W is a neighborhood of x_0. Note that this Θ is composed of integral curves of the vector fields in $\bar{\Delta}_{k^*-1}$ or F. Using same argument as in Case 1 and Case 3 of the proof of Lemma 9.2.1 and the proof of Lemma 9.2.2, we can find, for each $x_1 \in W$, $Z_1, \cdots, Z_\ell \in \Delta_{k^*-1} \cup F$, such that the modified $\tilde{\Theta} : \sqcup_\delta^\ell \to W_{x_1} \subset W$ starts from x_0 and using the integral curves of Z_i and finally maps onto a neighborhood of x_1. Moreover, for $Z_i \in F$ the corresponding time is non-negative. As reversing the time and modifying $f \in F$, a new $\tilde{\Theta}$, denoted by $\tilde{\Theta}^c : \sqcup_\delta^\ell \to W_{x_0} \subset W$, goes backward from x_1 to a neighborhood of x_0. Roughly speaking, x_0 is locally controllable by $(F; \Delta_{k^*-1})$.

Next, we restrict to each leave $\mathcal{I}_{\bar{\Delta}_{k^*-1}}(x_0)$, and show that for $Z \in \Delta_{k^*-1}$ the corresponding component ϕ_t^Z in $\tilde{\Theta}$ can be replaced by the integral curves of vector fields in $\Delta_{k^*-2} \cup F$. First, using Baker-Campbell-Hausdorff formula, we can prove easily that

$$\Delta_{k^*-1} \subset Span\left\{ (\bar{\Delta}_{k^*-2})_* F\big|_{\Delta_{k^*-1}}; \bar{\Delta}_{k^*-2} \right\}. \tag{9.50}$$

According to (9.50) and the argument in the proof of Lemma 9.2.2, ϕ_t^Z can be replaced by the integral curves of vector fields in $\left\{\bar{\Delta}_{k^*-2} \cup F|_{\Delta_{k^*-1}}\right\}$. Repeating the same argument as in Case 1 and Case 3 of the prove of Lemma 9.2.1 and the proof of Lemma 9.2.2, one sees that we can do this.

In the resulting mapping we have integral curves of vector fields as $Z \in \bar{\Delta}_{k^*-2}$ and $f|_{\Delta_{k^*-1}}$, $f \in F$. For $Z \in \bar{\Delta}_{k^*-2}$ we know how to replace it by a set of $Z_i \in \Delta_{k^*-2}$ as in the first step. As for $f|_{\Delta_{k^*-1}}$, we can formally replace it by its corresponding f. Then the problem caused by this replacement is that the components of the vector field belong to $f|_{\bar{\Delta}_{k^*}} = f$ has been changed from $(0, \cdots, 0)^T$ to $f^{k^*} = (f_{n_{k^*-1}+1}, f_{n_{k^*-1}+2}, \cdots, f_n)^T$. This makes the destination point drift. Fortunately from Lemma A9.3.1, f^{k^*} is coordinate-independent of $\bar{\Delta}_{k^*-2}$. (Refer to the following remark.) So when the system goes along the trajectories of $\bar{\Delta}_{k^*-2}$ and $F|_{\bar{\Delta}_{k^*-1}}$ the part of variables corresponding to f^{k^*} drift freely. By the definition of deep interior, the motion along $\bar{\Delta}_{k^*-2}$ and $f|_{\Delta_{k^*-1}}$ can be within a time period as short as we wish. This makes the drift be as small as possible. So it doesn't affect the property that the region of corresponding transfer mapping covers a neighborhood of x and vise versa. By induction, we can keep on reducing the subscript of the distributions of $\{\Delta_l\}$ till Δ_1. Using the argument for Case 2 of the proof of Lemma 9.2.1, we can easily convert the vector fields in Δ_1 to being realizable. ■

Remark 9.C.1. Assume (9.7) has a proper set of controllability distributions, we may express (9.7) in a coordinate chart which is flat with respect to $\bar{\Delta}_i$. For notational ease, denote

$$\{x^1, \cdots, x^i\} = \{x_1, x_2, \cdots, x_{n_i}\}, \quad i = 1, 2, \cdots, k^*.$$

Then (9.7) can be expressed locally as

$$\begin{cases} \dot{x}^1 = f^{\sigma(t),1}(x) + G^{\sigma(t),1}(x)u \\ \dot{x}^2 = f^{\sigma(t),2}(x) \\ \dot{x}^3 = f^{\sigma(t),3}(x^2, x^3, \cdots, x^{k^*}) \\ \dot{x}^4 = f^{\sigma(t),4}(x^3, x^4, \cdots, x^{k^*}) \\ \vdots \\ \dot{x}^{k^*-1} = f^{\sigma(t),k^*-1}(x^{k^*-2}, x^{k^*-1}, x^{k^*}) \\ \dot{x}^{k^*} = f^{\sigma(t),k^*}(x^{k^*-1}, x^{k^*}). \end{cases} \tag{9.51}$$

We may call it the controllability canonical form of affine control systems.

9.D Tensor Field and Orientation

We give a brief review on tensor field, orientation, etc., and refer to, e.g., [3], [17] for details.

Definition 9.D.1. *Let V be an n dimensional vector space. A k-th order covariant tensor φ is a k-th fold multi-linear mapping: $\varphi : \underbrace{V \times \cdots \times V}_{k} \to \mathbb{R}$.*

Definition 9.D.2. *Let M be an n dimensional manifold. A k-th order covariant tensor field φ is a k-th fold multi-linear mapping: $\varphi : \underbrace{T(M) \times \cdots \times T(M)}_{k} \to C^\infty(M)$, such that at each point $x \in M$ $\varphi(x) : \underbrace{T_x(M) \times \cdots \times T_x(M)}_{k} \to \mathbb{R}$ is a k-th order covariant tensor.*

Definition 9.D.3. *1. A k-th order covariant tensor field φ is said to be symmetric if for any k vector fields $X_1, \cdots, X_k$ we have*

$$\varphi(X_1, \cdots, X_i, \cdots, X_j, \cdots, X_k) = \varphi(X_1, \cdots, X_j, \cdots, X_i, \cdots, X_k). \quad (9.52)$$

2. φ is said to be skew-symmetric if for any k vector fields $X_1, \cdots, X_k$ we have

$$\varphi(X_1, \cdots, X_i, \cdots, X_j, \cdots, X_k) = -\varphi(X_1, \cdots, X_j, \cdots, X_i, \cdots, X_k). \quad (9.53)$$

3. A k-th skew-symmetric covariant tensor field is briefly called a k-form. The set of k-forms on M is denoted by $\Omega^k(M)$.

4. Let $\mathbf{S}_k$ be the k-th order permutation group and φ a k-th order covariant tensor. Then an alternating mapping $\mathcal{A}$ can convert it to a skew-symmetric one. Precisely,

$$\mathcal{A}(\varphi)(X_1, \cdots, X_k) = \frac{1}{k!} \sum_{\sigma \in \mathbf{S_k}} sign(\sigma) \varphi(X_{\sigma(1)}, \cdots, X_{\sigma(k)}). \quad (9.54)$$

5. Let $\alpha \in \Omega^s(M)$, $\beta \in \Omega^t(M)$. Then the tensor product of $\alpha \otimes \beta \in \Omega^{s+t}$ is defined as

$$\alpha \otimes \beta(X_1, \cdots, X_{s+t}) = \alpha(X_1, \cdots, X_s)\beta(X_{s+1} \cdots, X_{s+t}). \quad (9.55)$$

The wedge product of $\alpha \wedge \beta \in \Omega^{s+t}$ is defined as

$$\alpha \wedge \beta = \frac{(s+t)!}{s!t!} \mathcal{A}(\alpha \otimes \beta). \quad (9.56)$$

Definition 9.D.4. *An n dimensional manifold is orientable if it is possible to define a C^∞ n-form ω on M which is not zero at any point, in which case M is said to be oriented by ω.*

Definition 9.D.5. *Let $f \in V(M)$. It deduces a mapping $i_f : \Omega^{k+1} \to \Omega^k$, which is defined for each $\omega \in \Omega^{k+1}$ as*

$$i_f(\omega)(\cdot) := \omega(f, \cdot). \quad (9.57)$$

9.E Proof of Theorem 9.4.1

(Sufficiency) According to Theorem 9.2.1 we have only to prove that $\mathcal{G}(x_2,x_1)_*f_1(x_1)$ and $f_2(x_2)$ are on the two sides of $\mathcal{I}_{\mathcal{G}}(x_2)$.

Since $\mathcal{I}_{\mathcal{G}}(x)$ is orientable, we can find a set of coherently oriented coordinate charts (U_μ,x^μ), such that [3]

$$\sigma_G=\eta^\mu(x)x_1^\mu\wedge x_2^\mu\wedge\cdots\wedge x_{n-1}^\mu,$$

where $\eta^\mu(x)>0$, $x\in U_\mu$. Let a canonical local basis of $\mathcal{G}$ be chosen as

$$Z_i=\frac{\partial}{\partial x_i^\mu},\quad x\in U_\mu,\ i=1,\cdots,n-1,$$

which are chart-depending. We have

$$\sigma_G(Z_1,\cdots,Z_{n-1})(x)>0,\quad \forall x\in M. \tag{9.58}$$

(Precisely, $x\in U_\mu$ and $Z_i=Z_i^\mu$, which are defined over U_μ.) Now we have

$$i_{f_1}\sigma(Z_1,\cdots,Z_{n-1})=\sigma_1(f_1)\sigma_G(Z_1,\cdots,Z_{n-1})=c_1(x)\sigma_G(Z_1,\cdots,Z_{n-1}); \tag{9.59}$$

So we have

$$\sigma_1(f_1)(x_1)=c_1(x_1). \tag{9.60}$$

Similarly, we have

$$\sigma_1(f_2)(x_2)=c_2(x_2). \tag{9.61}$$

Recall that

$$\mathcal{G}(x_2,x_1)_*f_1(x_1)=icl_*[f_1(x_1)/\mathcal{G}]+\eta,\quad \eta\in\mathcal{G}.$$

Then using the fact that $\sigma_1(Z)=0$, $Z\in\mathcal{G}$, we have

$$\sigma_1[\mathcal{G}(x_2,x_1)_*f_1(x_1)](x_2)=\sigma_1(f_1)=c_1(x_1). \tag{9.62}$$

Now since

$$\sigma[\mathcal{G}_*f_1(x_2,x_1),Z_1,\cdots,Z_{n-1}](x_2)=c_1(x_1)\sigma_G(Z_1,\cdots,Z_{n-1})$$

and

$$\sigma[f_2,Z_1,\cdots,Z_{n-1}](x_2)=c_2(x_2)\sigma_G(Z_1,\cdots,Z_{n-1})$$

have opposite signs, $[\mathcal{G}(x_2,x_1)_*f_1(x_1)](x_2)$ and $f_2(x_2)$ must on the different sides of $\mathcal{I}_{\mathcal{G}}(x_2)$. To see this define

$$Y(t)=tY_1+(1-t)Y_2,\quad t\in[0,1],$$

where $Y_1=[\mathcal{G}(x_2,x_1)_*f_1(x_1)](x_2)$ and $Y_2=f_2(x_2)$. If both Y_1 and Y_2 are on the same side of $\mathcal{I}_{\mathcal{G}}(x_2)$, then $Y(t)$ is always linearly independent of $\mathcal{G}(x_2)$. But $\sigma[Y(t),Z_1,\cdots,Z_{n-1}](x_2)$ has different signs at $t=0$ and $t=1$, so there exits a $0<t^*<1$, such that

$$\sigma[Y(t^*), Z_1, \cdots, Z_{n-1}](x_2) = 0,$$

which is a contradiction. (Because an orientation can never be zero over a set of n linearly independent vector fields.)

(Necessity)Assume there is an $\mathcal{I}_{\mathcal{G}}(x)$ such that (9.30) fails. According to our previous argument we know that for all $y, z \in \mathcal{I}_{\mathcal{G}}(x)$ and $f_1, f_2 \in F$ we have $\sigma_1(f_1(y))$ and $\sigma_1(f_2(z))$ have same sign. That is, all the $f \in F$ point to a same side of $\mathcal{I}_{\mathcal{G}}(x)$. Since $\mathcal{I}_{\mathcal{G}}(x)$ separates M, it is easy to see that the half M which is pointed by $f's$ is an invariant set. So the system is not controllable. ■

9.F Proof of Corollory 9.4.1

First, (9.31) is equivalent to the following statement: Choosing $\sigma = dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$, then

$$\sigma(f_1, Z_1, \cdots, Z_{n-1})(x_1)\sigma(f_2, Z_1, \cdots, Z_{n-1})(x_2) < 0. \tag{9.63}$$

Since (9.63) is independent to the choice of orientation σ, we have only to prove that there exists a particular orientation such that (9.31) implies (9.30).

Choosing a particular coordinate frame $z = z(x)$ such that

$$\mathcal{G} = Span\left\{\frac{\partial}{\partial z_2}, \cdots, \frac{\partial}{\partial z_n}\right\}$$

Then we simply choose $\sigma = dz_1 \wedge dz_2 \wedge \cdots \wedge dz_n$, and set $\sigma_1 = dz_1$ and $\sigma_G = dz_2 \wedge \cdots \wedge dz_n$. Let $f_i(z) = (f_{i1}(z), \cdots, f_{in}(z))^T$, $i = 1, 2$. Then

$$i_{f_1}\sigma = f_{11} dz_2 \wedge \cdots \wedge dz_n + dx_1 \wedge \eta, \quad where \ \ \eta \in \Omega^{n-2}(M).$$

So

$$i_{f_1}\sigma|_{\mathcal{G}} = f_{11} dz_2 \wedge \cdots \wedge dz_n.$$

Similarly,

$$i_{f_2}\sigma|_{\mathcal{G}} = f_{21} dz_2 \wedge \cdots \wedge dz_n.$$

Note that if $Z_i \in \mathcal{G}$, then $Z_i = (0, z_2^i, \cdots, z_n^i)^T$. So (9.63) implies that $f_{11}(x_1)$ and $f_{21}(x_2)$ have different sign, which implies (9.30). ■

10

Control and Observation of the Matrix Riccati Differential Equation

G. Dirr[1], U. Helmke[2], and J. Jordan[3]

[1] Institute of Mathematics, University of Würzburg, 97074 Würzburg, Germany
dirr@mathematik.uni-wuerzburg.de
[2] Institute of Mathematics, University of Würzburg, 97074 Würzburg, Germany
helmke@mathematik.uni-wuerzburg.de
[3] Institute of Mathematics, University of Würzburg, 97074 Würzburg, Germany
jordan@mathematik.uni-wuerzburg.de

Summary. We explore controllability and observability properties of the matrix Riccati differential equation as a flow on the Grassmann manifold. Using the known classification of transitive Lie group actions on Grassmann manifolds, we derive necessary and sufficient conditions for accessibility of Riccati equations. This also leads to new sufficient Lie–algebraic conditions for controllability of generalized double bracket flows. Observability of Riccati equations with linear fractional output functions is characterized via a generalized Hautus–Popov test, thus making contact with earlier work by Dayawansa, Ghosh, Martin and Rosenthal on perspective observability of linear systems.

10.1 Introduction

The role of matrix Riccati differential equations in H^∞-optimization, linear quadratic optimal control and stochastic filtering is well-known; see e.g. [17]. Such equations also arise in a very natural way in perspective observability problems and state estimation tasks from computer vision, cf. [5, 13]. In this paper, we explore controllability and observability properties for the controlled matrix Riccati differential equation

$$\begin{aligned} \dot{K} &= -KA_{11}(u) + A_{22}(u)K - KA_{12}(u)K + A_{21}(u) \\ Y &= (C_{21} + C_{22}K)(C_{11} + C_{12}K)^{-1} \end{aligned} \tag{10.1}$$

with linear fractional outputs. Here, the coefficient matrices $A_{ij}(u)$ are assumed to depend affinely on a vector of control variables $u \in \mathbb{R}^m$ and $C_{11}, C_{12}, C_{21}, C_{22}$ are matrices of suitable size. Using the known classification of transitive Lie group actions on Grassmann manifolds, we characterize accessibility of (10.1) in terms of the associated system Lie algebra. We also derive sufficient observability conditions for the uncontrolled system ($A(u) = A$), and more generally for linear induced flows on Grassmann manifolds.

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 169–184.
springerlink.com

The motivation for studying such control problems for the Riccati equation comes from several different directions. In [3], for instance, Brockett noticed that variants of the double bracket flow

$$\dot{X} = [\Omega(u), X] + \big[[S(u), X], X\big] \tag{10.2}$$

on symmetric matrices X with real skew–symmetric and symmetric control terms $\Omega(u)$ and $S(u)$, respectively, can simulate arbitrary finite-state automata. Thus, for $\Omega \neq 0$, system (10.2) provides a straightforward extension of Brockett's double bracket equation, studied extensively since his pioneering paper [2]; see e.g. [12] and the references therein. Here, the isospectral flow (10.2) is of interest to us, as it has been shown in [11, 6] that it yields a natural generalization of the matrix Riccati differential equation (10.1). Thus the control problems of (10.1) and (10.2) are intimately related.

Proceeding in a different direction, control problems for the Riccati equation are also of interest in numerical analysis. In fact, as has been first observed by Ammar and Martin [1], the Riccati equation (10.1) defines a continuous–time version of the celebrated power method, and thus can be interpreted as a continuous–time eigenvalue method for computing dominant eigenspaces of a matrix A. Therefore, the systematic analysis and design of such eigenvalue methods via control theoretic methods becomes a challenging task for future research; we refer to [18, 10] and [14] for first steps in this direction.

Moreover, machine learning and computer vision feature several exciting observation problems for Riccati differential equations. Here, we mention the pioneering work by Ghosh, Jankovic, Wu [7] and Dayawansa, Ghosh, Martin, Wang [5] on the closely related problem of perspective observability of linear systems, as well as the subsequent papers [8, 9, 15]. In such camera vision problems, one wants to estimate the 'projective state', i.e. the state up to a scalar multiple of a linear system

$$\dot{x}(t) = Ax(t) \tag{10.3}$$

from perspective output measurements. This is a non-trivial problem in computer vision and observability conditions analogous to the celebrated Hautus-Popov criterion are available for systems with complex coefficients; see [8, 15]. Since the induced flow of (10.3) on the projective space is equivalent (in local coordinates) to the Riccati equation (10.1), we are back again to our problem.

Our approach to perspective observability differs from that by the above mentioned authors in a crucial point. Instead of trying to solve the problem by (sophisticated) linear algebra techniques, we will mainly employ basic tools from dynamical systems theory. Explicitly, we characterize the ω–limit sets of linear induced flows on certain Grassmann manifolds and deduce observability conditions from that information. This leads also to new results in the more complicated case of real coefficients.

It follows, that the combination of control and perspective estimation problems connects with interesting dynamical systems tasks for the Riccati equation, that are not fully understood yet. The interaction between geometry and

dynamics has been a central theme in Dayawansa's work and it is a pleasure to acknowledge the special impact that Daya's research had on this circle of ideas.

10.2 Controllability of the Riccati Differential Equation

Let $\mathbb{K}$ denote either the field of real numbers $\mathbb{R}$ or the field of complex numbers $\mathbb{C}$. For $1 \leq m \leq n-1$, we consider the *matrix Riccati differential equation*

$$\dot{K} = -KA_{11} + A_{22}K - KA_{12}K + A_{21} \tag{10.4}$$

with $K \in \mathbb{K}^{(n-m)\times m}$, $A_{11} \in \mathbb{K}^{m\times m}$, $A_{22} \in \mathbb{K}^{(n-m)\times(n-m)}$ and $A_{12} \in \mathbb{K}^{m\times(n-m)}$, $A_{21} \in \mathbb{K}^{(n-m)\times m}$. We naturally associate with (10.4) a *linear induced flow* on an appropriate *Grassmann manifold*. To this end, let $\mathbb{G}_m(\mathbb{K}^n)$ denote the real or complex Grassmann manifold, i.e. the set of all m-dimensional $\mathbb{K}$-linear subspaces of $\mathbb{K}^n$. As a special case, let $\mathbb{P}(\mathbb{K}^n) := \mathbb{G}_1(\mathbb{K}^n)$ denote the real or complex projective space, i.e. the space of all (real or complex) lines through the origin in $\mathbb{K}^n$. Moreover, for any matrix $X \in \mathbb{K}^{n\times m}$ let

$$\langle X \rangle := \{Xu \mid u \in \mathbb{K}^m\}$$

denote the linear span of the columns of X. Now, for

$$A := \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{K}^{n\times n}, \tag{10.5}$$

one has the linear induced flow $V \to \mathrm{e}^{tA}V$ on $\mathbb{G}_m(\mathbb{K}^n)$. The corresponding vector field is denoted by $L_A : \mathbb{G}_m(\mathbb{K}^n) \to \mathrm{T}\,\mathbb{G}_m(\mathbb{K}^n)$ and the linear induced system

$$\dot{V} = L_A(V) \tag{10.6}$$

on $\mathbb{G}_m(\mathbb{K}^n)$ is called the *extended Riccati differential equation* of (10.4). Its flow $V \to \mathrm{e}^{tA}V$ on $\mathbb{G}_m(\mathbb{K}^n)$ has the following well-known relation to the Riccati flow.

Proposition 10.2.1 ([1, 16, 12]). *Let $K(t)$ be a solution of* (10.4)*. Then*

$$V(t) := \left\langle \begin{bmatrix} \mathrm{I}_m \\ K(t) \end{bmatrix} \right\rangle \tag{10.7}$$

is a solution of the extended Riccati vector field $\dot{V} = L_A(V)$ on $\mathbb{G}_m(\mathbb{K}^n)$. More precisely, one has the identity $V(t) := \mathrm{e}^{tA}V(0)$ for all $t \in \mathbb{R}$.

In the sequel, we assume that $A \in \mathbb{K}^{n\times n}$ depends affinely on *control parameters* as $A = A(u) := A_0 + \sum_{j=1}^{k} u_j A_j$, with $A_0, A_1, \ldots A_k \in \mathbb{K}^{n\times n}$ and piecewise constant input functions $u \in \mathbb{R}^k$. This leads to the *controlled Riccati differential equation*

$$\dot{K} = -KA_{11}(u) + A_{22}(u)K - KA_{12}(u)K + A_{21}(u). \tag{10.8}$$

By Proposition 10.2.1, equation (10.8) readily extends to a control system on $\mathbb{G}_m(\mathbb{K}^m)$, cf. [6]. We are interested in necessary and sufficient conditions that guarantee accessibility of the controlled Riccati differential equation, i.e. in conditions when the reachable sets of (10.8) have non-empty interior. Our first main result [6] characterizes accessibility in terms of the so-called *system Lie algebra* $\mathcal{L} \subset \mathbb{K}^{n\times n}$, i.e. by the Lie subalgebra generated by $A(u)$, $u \in \mathbb{R}^k$. Controllability of the Riccati equation is a more subtle issue, since available controllability results for linear induced flows on Grassmann manifolds cannot be carried over directly. In fact, the trajectories of the linear induced flow may leave the maximal domain of the coordinate chart which relates the Riccati flow and the linear induced one, cf. (10.7). At the end of this subsection, we therefore state a sufficient condition which yields at least controllability of the linear induced systems.

Now, accessibility of the controlled Riccati equation (10.8) is equivalent to transitivity of the system group action of the associated control system on $\mathbb{G}_m(\mathbb{K}^n)$. Therefore, Theorem 10.2.1 follows from Völklein's [19] classification of transitive Lie subgroup actions on $\mathbb{G}_m(\mathbb{K}^n)$. Here, we restrict to the complex case, for further details and a similar result in the real case we refer to [6].

Theorem 10.2.1 (Accessibility). *Let $A(u) \in \mathbb{C}^{n\times n}$ be partitioned as in (10.5) and let $\mathcal{L}$ be the system Lie algebra generated by $A(u) := A_0 + \sum_{j=1}^k u_j A_j$, $u \in \mathbb{R}^k$. The controlled Riccati differential equation* (10.8) *is* accessible *if and only if*

(a) $\mathcal{L} = \mathfrak{z} \oplus \mathcal{L}_0$, where $\mathfrak{z}$ is any real Lie subalgebra of $\mathbb{C}\mathrm{I}_n$ and $\mathcal{L}_0$ is conjugate to $\mathfrak{sl}_n(\mathbb{C})$ or $\mathfrak{su}_n$, provide either n is odd or $1 < m < n$ holds.
(b) $\mathcal{L} = \mathfrak{z} \oplus \mathcal{L}_0$, where $\mathfrak{z}$ is any real Lie subalgebra of $\mathbb{C}\mathrm{I}_n$ and $\mathcal{L}_0$ is conjugate to $\mathfrak{sl}_n(\mathbb{C}), \mathfrak{su}_n$, $\mathfrak{sl}_{n/2}(\mathbb{H})$, $\mathfrak{sp}_{n/2}(\mathbb{C})$ or $\mathfrak{sp}_{n/2}$, if n is even and $m = 1$ or $m = n-1$.

Here, $\mathfrak{sl}_n(\mathbb{C})$ and $\mathfrak{sp}_{n/2}(\mathbb{C})$ denote the Lie subalgebras of all complex and complex Hamiltonian matrices with trace zero, respectively. Moreover, $\mathfrak{su}_n$ and $\mathfrak{sp}_{n/2}$ are the Lie subalgebras of $\mathfrak{sl}_n(\mathbb{C})$ and $\mathfrak{sp}_{n/2}(\mathbb{C})$, which consist of all skew-Hermitian matrices and skew-Hermitian Hamiltonian matrices, respectively. Finally, $\mathbb{H} \cong \mathbb{C}^2$ denotes the algebra of all quaternions and $\mathfrak{sl}_{n/2}(\mathbb{H})$ is defined as a subalgebra of $\mathfrak{sl}_n(\mathbb{C})$, cf. [4]. Together with a well-known controllability criterion for right–invariant control systems on compact Lie groups, Theorem 10.2.1 implies the following sufficient controllability condition [6].

Theorem 10.2.2 (Controllability). *Let $A(u) \in \mathbb{C}^{n\times n}$ be partitioned as in (10.5) and let $\mathcal{L}$ be the system Lie algebra generated by $A(u) := A_0 + \sum_{j=1}^k u_j A_j$, $u \in \mathbb{R}^k$. Assume that the drift term A_0 is skew-Hermitian. The extended Riccati equation* (10.8) *is controllable on $\mathbb{G}_m(\mathbb{C}^n)$ if $\mathcal{L}$ satisfies one of the conditions (a) or (b) of Theorem 10.2.1.*

We conclude this section by connecting the above results to the generalized double bracket equation (10.2), cf. [3, 6]. For $A_0, A_1, \ldots, A_k \in \mathbb{C}^{n\times n}$, let $\Omega(u)$ and $S(u)$ denote the skew-Hermitian and Hermitian part of $A(u) := A_0 + \sum_{j=1}^k u_j A_j$, respectively. The *generalized double bracket equation*

$$\dot{X} = [\Omega(u), X] + \big[[S(u), X], X\big] \tag{10.9}$$

then defines an isospectral flow on the manifold $\mathrm{Grass}_{m,n}(\mathbb{C})$ of all Hermitian rank-m projection operators which act on $\mathbb{C}^n$. In [6] it is shown that solutions of (10.9) uniquely correspond to solutions of the linear induced flow $\mathrm{e}^{tA}V$ on $\mathbb{G}_m(\mathbb{C}^n)$. Thus the above ideas, based on Völklein's classification of transitive Lie subgroup actions on $\mathbb{G}_m(\mathbb{K}^n)$, also yield accessibility and controllability conditions for (10.9). We state only one such result which displays the close connection to the theory of Riccati equations.

Theorem 10.2.3 ([6]). *Let $\mathcal{L}$ denote the system Lie algebra generated by $A(u)$, $u \in \mathbb{R}^k$. The generalized double bracket equation* (10.9) *is accessible on the Grassmannian* $\mathrm{Grass}_{m,n}(\mathbb{C})$ *if and only if*

(a) $\mathcal{L} = \mathfrak{z} \oplus \mathcal{L}_0$, where $\mathfrak{z}$ is any real Lie subalgebra of $\mathbb{C}\mathrm{I}_n$ and $\mathcal{L}_0$ is equal to $\mathfrak{sl}_n(\mathbb{C})$ or conjugate to $\mathfrak{su}_n$, if n is odd or $1 < k < n$.
(b) $\mathcal{L} = \mathfrak{z} \oplus \mathcal{L}_0$, where $\mathfrak{z}$ is any real Lie subalgebra of $\mathbb{C}\mathrm{I}_n$ and $\mathcal{L}_0$ is equal to $\mathfrak{sl}_n(\mathbb{C})$ or conjugate to $\mathfrak{su}_n$, $\mathfrak{sl}_{n/2}(\mathbb{H})$, $\mathfrak{sp}_{n/2}(\mathbb{C})$ or $\mathfrak{sp}_{n/2}$, if n is even and $k = 1$ or $k = n - 1$.

10.3 Riccati Equations and Perspective Observability

We now discuss observability results for the matrix Riccati differential equation (10.1) without inputs ($A(u) = A$) and linear fractional outputs, i.e.

$$\begin{aligned} \dot{K} &= -KA_{11} + A_{22}K - KA_{12}K + A_{21} \\ Y &= (C_{21} + C_{22}K)(C_{11} + C_{12}K)^{-1} \end{aligned} \tag{10.10}$$

Here, $C_{11} \in \mathbb{K}^{m\times m}$, $C_{12} \in \mathbb{K}^{m\times(n-m)}$, $C_{21} \in \mathbb{K}^{(p-m)\times m}$ and $C_{22} \in \mathbb{K}^{(p-m)\times(n-m)}$ with $p \geq m+1$. System (10.10) is called observable on a nontrivial interval $I \subset \mathbb{R}$ if for any two initial values $K_1, K_2 \in \mathbb{K}^{(n-m)\times n}$ the implication

$$Y_1(t) = Y_2(t) \text{ for almost all } t \in I \implies K_1 = K_2 \tag{10.11}$$

holds, where $Y_i(t) := \big(C_{21}+C_{22}K_i(t)\big)\big(C_{11}+C_{12}K_i(t)\big)^{-1}$ for $i = 1, 2$. Of course, one has to assume that any solution of (10.10) is such that $C_{11} + C_{12}K(t)$ is invertible for almost all $t \in I$. We call the Riccati equation (10.10) *output regular*, if for all m-dimensional linear subspaces $V \subset \mathbb{K}^n$ and for almost all t the condition $\dim[C_{11}\ C_{12}]\mathrm{e}^{\mathrm{tA}}\mathrm{V} = \mathrm{m}$ holds. Here, $A \in \mathbb{K}^{n\times n}$ is given by (10.5). From now on, we pass to the extended Riccati equation on $\mathbb{G}_m(\mathbb{K}^n)$

$$\dot{V} = L_A(V), \quad Y = CV, \tag{10.12}$$

or, equivalently, to the *linear perspective system* on full column rank matrices $X \in \mathbb{K}^{n\times m}$

$$\dot{X} = AX, \quad Y = C\langle X\rangle, \tag{10.13}$$

whose output is the image space of $\langle X\rangle$ under

$$C := \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \in \mathbb{K}^{p\times n}. \tag{10.14}$$

Our basic assumption is $\operatorname{rk} C = p \geq m+1$. Equation (10.13) can be regarded as a real analytic system that acts on $\mathbb{G}_m(\mathbb{K}^n)$ via

$$\langle X\rangle \mapsto \mathrm{e}^{tA}\langle X\rangle \in \mathbb{G}_m(\mathbb{K}^n). \tag{10.15}$$

The corresponding output map is given by

$$\langle X\rangle \mapsto Y = C\langle X\rangle \in \mathbb{G}_p(\mathbb{K}^n). \tag{10.16}$$

Observe, that the image of (10.16) is not entirely contained in $\mathbb{G}_p(\mathbb{K}^n)$. However, $C\langle X\rangle$ belongs to $\mathbb{G}_p(\mathbb{K}^n)$ for almost all $\langle X\rangle \in \mathbb{G}_m(\mathbb{K}^n)$. Now, the pair (C, A) is called *perspectively observable* on $\mathbb{G}_m(\mathbb{K}^n)$ if for any two subspaces $V_1, V_2 \in \mathbb{G}_m(\mathbb{K}^n)$ the implication

$$\text{(OC)} \qquad C\mathrm{e}^{tA}V_1 = C\mathrm{e}^{tA}V_2 \text{ for almost all } t\in\mathbb{R} \implies V_1 = V_2 \tag{10.17}$$

holds. Due to the previous remark on the image of the output map (10.16), we do not require equality for *all* $t \in \mathbb{R}$ in the above definition, but for almost all $t \in \mathbb{R}$. The following result shows, that perspective observability can be decided on arbitrary subintervals of $\mathbb{R}$, is an immediate consequence of the identity theorem for meromorphic functions.

Lemma 10.3.1. *Let $I \subset \mathbb{R}$ be any subinterval. Then the pair (C, A) is perspectively observable on $\mathbb{G}_m(\mathbb{K}^n)$ if and only if for any two subspaces $V_1, V_2 \in \mathbb{G}_m(\mathbb{K}^n)$ the implication*

$$\text{(OC)}_I \qquad C\mathrm{e}^{tA}V_1 = C\mathrm{e}^{tA}V_2 \text{ for almost all } t\in I \implies V_1 = V_2 \tag{10.18}$$

holds. In particularly, an output regular Riccati equation (10.10) *is observable if and only if (C, A) is perspectively observable.*

We begin our analysis of the perspective observation problem by first establishing a basic linear algebra characterization.

Lemma 10.3.2. *The pair (C, A) is not perspectively observable on $\mathbb{G}_m(\mathbb{K}^n)$ if and only if there exists a $(m+1)$-dimensional linear subspace $W \in \mathbb{G}_{m+1}(\mathbb{K}^n)$ such that $\dim C\mathrm{e}^{tA}W \leq m$ for all $t \in \mathbb{R}$.*

Proof. Suppose $V_1 = \langle X_1\rangle, V_2 = \langle X_2\rangle \in \mathbb{G}_m(\mathbb{K}^n)$ with $V_1 \neq V_2$ and $C\mathrm{e}^{tA}V_1 = C\mathrm{e}^{tA}V_2$ for almost all $t \in \mathbb{R}$. Then, choose $x_2 \in V_2$ with $x_2 \notin V_1$ and let $W := \langle X_1, x_2\rangle$. Thus, $C\mathrm{e}^{tA}W = \langle C\mathrm{e}^{tA}X_1, C\mathrm{e}^{tA}x_2\rangle = C\mathrm{e}^{tA}V_1$ holds for almost all $t \in \mathbb{R}$ and $C\mathrm{e}^{tA}W$ is at most m-dimensional for all $t \in \mathbb{R}$.

Conversely, let $W \subset \mathbb{K}^n$ be a $(m+1)$-dimensional linear subspace such that $r := \max_{t\in\mathbb{R}} \dim C\mathrm{e}^{tA}W \leq m$. Choose $t^* \in \mathbb{R}$ with $\dim C\mathrm{e}^{t^*A}W = r$. Then there exist linear subspaces $V_1 = \langle X_1\rangle, V_2 = \langle X_2\rangle \in \mathbb{G}_m(\mathbb{K}^n)$ with $V_1 \neq V_2$ and $\dim C\mathrm{e}^{t^*A}V_1 = \dim C\mathrm{e}^{t^*A}V_2 = r$. By analyticity of $t \mapsto C\mathrm{e}^{tA}X_i$, $i = 1, 2$, we conclude $\dim C\mathrm{e}^{tA}V_i = r$, $i = 1, 2$ for almost all $t \in \mathbb{R}$. Since r is the maximal dimension of $C\mathrm{e}^{tA}W$, this shows $C\mathrm{e}^{tA}V_1 = C\mathrm{e}^{tA}W = C\mathrm{e}^{tA}V_2$ for almost all $t \in \mathbb{R}$. Thus, (C, A) is not perspectively observable on $\mathbb{G}_m(\mathbb{K}^n)$. ∎

From Lemma 10.3.2 we conclude that a pair (C, A) is perspectively observable on $\mathbb{G}_m(\mathbb{K}^n)$ if and only if the condition

$$(*) \quad \dim Ce^{tA}W = m+1 \text{ for almost all } t \in \mathbb{R}$$

holds for all $W \in \mathbb{G}_{m+1}(\mathbb{K}^n)$. In the following proposition, we show that perspective observability is already guaranteed under the significantly weaker assumption that $(*)$ is only satisfied on certain *ω-limit sets*. Recall, that the ω-limit set of $W \in \mathbb{G}_{m+1}(\mathbb{K}^n)$ is the compact subset of $\mathbb{G}_{m+1}(\mathbb{K}^n)$ defined as

$$\omega(W) := \{ \lim_{k\to\infty} e^{t_k A} W \mid t_k > 0, \lim_{k\to\infty} t_k = \infty \}.$$

Proposition 10.3.1. *The pair (C, A) is perspectively observable on $\mathbb{G}_m(\mathbb{K}^n)$ if and only if for each $W \in \mathbb{G}_{m+1}(\mathbb{K}^n)$ there exists $W_\infty \in \omega(W)$ with $\dim CW_\infty = m+1$.*

Proof. Assume that there exists a subspace $W \in \mathbb{G}_{m+1}(\mathbb{K}^n)$ such that $\dim CW_\infty < m+1$ for all $W_\infty \in \omega(W)$. Then, choose any $W_0 \in \omega(W)$. By the compactness of $\mathbb{G}_{m+1}(\mathbb{K}^n)$, a standard result from dynamical systems theory asserts that all ω-limit sets of (10.15) are non-empty, compact, connected and invariant under $e^{tA}, t \in \mathbb{R}$. Thus, one has $\dim Ce^{tA}W_0 < m+1$ for all $t \in \mathbb{R}$. Hence, by Lemma 10.3.2, the pair (C, A) is not perspectively observable on $\mathbb{G}_m(\mathbb{K}^n)$.

Conversely, let $W = \langle Z \rangle \subset \mathbb{K}^n$ be any $(m+1)$-dimensional subspace and let $W_\infty \in \omega(W)$ with $\dim CW_\infty = m+1$. Thus there exists a sequence t_n such that $t_n \to \infty$ and $e^{t_n A}W \to W_\infty$ for $n \to \infty$. Hence, an easy continuity argument implies $\dim Ce^{t_n A}W = \dim CW_\infty = m+1$ for sufficiently large $n \in \mathbb{N}$. From the analyticity of the map $t \mapsto Ce^{tA}Z$ we conclude $\dim Ce^{tA}W = m+1$ for almost all $t \in \mathbb{R}$. Lemma 10.3.2 thus implies perspective observability for (C, A) on $\mathbb{G}_m(\mathbb{K}^n)$. ∎

The subsequent result serves to rephrase the ω-limit set condition of Proposition 10.3.1 as a pure rank condition, cf. Theorem 10.4.2 and 10.4.4.

Lemma 10.3.3. *Let $0 \le r \le n$ be given and let $A \in \mathbb{K}^{n\times n}$ and $C \in \mathbb{K}^{p\times n}$ be arbitrary. The following conditions are equivalent:*

(a) All k-dimensional A-invariant subspaces E with $0 \le k \le r$ satisfy the equality $\dim CE = \dim E$.
(b) For all monic $\pi \in \mathbb{K}[x]$ with $\deg \pi = r$ one has

$$\operatorname{rk} \begin{bmatrix} \pi(A) \\ C \end{bmatrix} = n. \tag{10.19}$$

For $\mathbb{K} = \mathbb{C}$, conditions (a) and (b) are equivalent to

(c) All r-dimensional A-invariant subspaces E satisfy the equality $\dim CE = \dim E$.

For $\mathbb{K} = \mathbb{R}$, conditions (a) and (b) are equivalent to

(d) All $(r-1)$- and r-dimensional A-invariant subspaces E satisfy the equality $\dim CE = \dim E$.

Proof. First, assume

$$\operatorname{rk} \begin{bmatrix} \pi(A) \\ C \end{bmatrix} < n \tag{10.20}$$

holds for some monic $\pi \in \mathbb{K}[x]$ with $\deg \pi = r$. Then there exists a nonzero vector $x \in \mathbb{R}^n$ such that $\pi(A)x = 0$ and $Cx = 0$. Thus, $E := \langle x, Ax, \ldots, A^{r-1}x \rangle$ is an A-invariant subspace with $\dim CE < \dim E \leq r$. Hence, condition (a) is violated.

Conversely, assume that (a) is false. Then there exists a k-dimensional A-invariant subspace E with $\dim CE < \dim E = k$. In particular, there exists $x_0 \in E$, $x_0 \neq 0$, with $Cx_0 = 0$. Let π' denote the characteristic polynomial of the restricted operator $A|E$. Then, $\pi := x^{r-k}\pi'$ satisfies $\deg \pi = r$ and $\pi(A)x = A^{r-k}\pi'(A)x = 0$ for all $x \in E$. Thus, $\pi(A)x_0 = Cx_0 = 0$, which shows that (b) is not satisfied.

If $\mathbb{K} = \mathbb{C}$, using the complex Jordan normal form of A, one sees that any k-dimensional A-invariant subspace E can be complemented to a r-dimensional A-invariant subspace E'. Therefore, condition (c) is equivalent to (a) and hence to (b). For $\mathbb{R} = \mathbb{C}$, and real Jordan normal form of A, we deduce that any k-dimensional A-invariant subspace E can be complemented either to a $(r-1)$-dimensional or a r-dimensional A-invariant subspace E'. Thus, condition (d) is equivalent to (a) and, therefore, to (b). ∎

10.4 Perspective Observability – Projective Case

In the previous section we have seen that a test for perspective observability on $\mathbb{G}_m(\mathbb{K}^n)$ can be reduced to analysing the ω-limit sets of (10.15) on $\mathbb{G}_{m+1}(\mathbb{K}^n)$. Shayman derived in [16] a complete description of the phase portrait of Riccati differential equations on real Grassmann manifolds. His analysis included the description of ω–limit sets, but was restricted to a certain subclass of matrices. Here, we extend Shayman's characterization of ω–limit sets by allowing for a larger class of matrices. For expository purposes, we still impose a generic eigenvalue condition on A and focus on the case $m = 2$, as in this case the necessary notation remains elementary. In principle, however, the same techniques apply to any eigenvalue configuration as well as to the case $m \geq 2$.

10.4.1 The Complex Projective Space

Let $\lambda_1, \ldots, \lambda_r$ denote the r distinct eigenvalues of $A \in \mathbb{C}^{n\times n}$. Then A is called *strongly cyclic* if the following conditions are satisfied: (i) $\operatorname{Re}\lambda_i \neq \operatorname{Re}\lambda_j$ for $i \neq j$ and (ii) the geometric multiplicity of each eigenvalue λ_i is one. Thus there is in particular only one Jordan block corresponding to each eigenvalue of A.

Theorem 10.4.1. *Assume that $A \in \mathbb{C}^{n\times n}$ is strongly cyclic. Then, the ω-limit sets of* (10.15) *on $\mathbb{G}_2(\mathbb{C}^n)$ are the complex 2-dimensional A-invariant subspace. More precisely, for each $W \in \mathbb{G}_2(\mathbb{C}^n)$ there exists a unique complex 2-dimensional A-invariant subspace E such that $\omega(W) = \{E\}$.*

Proof. Without loss of generality, we assume that A is in Jordan normal from, i.e.

$$A = \begin{bmatrix} J(\lambda_1) & 0 & \cdots & 0 \\ 0 & J(\lambda_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & J(\lambda_r) \end{bmatrix} \quad \text{with} \quad J(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & 0 \\ & \ddots & 1 \\ 0 & & \lambda_i \end{bmatrix} \in \mathbb{C}^{n_i\times n_i}. \tag{10.21}$$

Here the eigenvalues λ_i are assumed to be in decreasing order with respect to their real parts, i.e. $\mathrm{Re}\lambda_1 > \cdots > \mathrm{Re}\lambda_r$. For $W \in \mathbb{G}_2(\mathbb{C}^n)$ choose a generating matrix $Z \in \mathbb{C}^{n\times 2}$ such that $W = \langle Z\rangle$. Moreover, let

$$Z = \begin{bmatrix} x & y \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ x_{i_1} & 0 \\ \vdots & y_{i_2} \\ \vdots & \vdots \\ x_r & y_r \end{bmatrix} \quad \text{with } x_i, y_i \in \mathbb{C}^{n_i}, i = 1, \ldots, r \tag{10.22}$$

be in *block echelon form* corresponding to (10.21) with *echelon indices* i_1 and i_2, i.e. $x_{i_1} \neq 0$, $x_i = 0$ for all $i < i_1$ and $y_{i_2} \neq 0$, $y_i = 0$ for all $i < i_2$. Thus, either $i_1 < i_2$ or $i_1 = i_2$, with x_{i_1}, y_{i_1} linear independent. In either case, one has

$$\mathrm{e}^{tA} Z \begin{bmatrix} \mathrm{e}^{-t\lambda_{i_1}} & 0 \\ 0 & \mathrm{e}^{-t\lambda_{i_2}} \end{bmatrix} = \begin{bmatrix} \vdots & \vdots \\ 0 & \vdots \\ \mathrm{e}^{tJ(0)} x_{i_1} & 0 \\ \vdots & \mathrm{e}^{tJ(0)} y_{i_2} \\ \vdots & \vdots \\ \mathrm{e}^{tJ(\lambda_r - \lambda_{i_1})} x_r & \mathrm{e}^{tJ(\lambda_r - \lambda_{i_2})} y_r \end{bmatrix}$$

Since $\mathrm{e}^{tJ(\lambda_i - \lambda_{i_1})} x_i \to 0$ and $\mathrm{e}^{tJ(\lambda_j - \lambda_{i_2})} y_j \to 0$ for $t \to \infty$ and $i > i_1$, $j > i_2$, it suffices to consider the ω-limit set of $\langle Z'(t)\rangle$ with

$$Z'(t) := \begin{bmatrix} \vdots & \vdots \\ 0 & \vdots \\ \mathrm{e}^{tJ(0)} x_{i_0} & 0 \\ 0 & \mathrm{e}^{tJ(0)} y_{j_0} \\ \vdots & 0 \\ \vdots & \vdots \end{bmatrix}$$

Moreover, it is straightforward to see that one has $t^{-s_1}\mathrm{e}^{tJ(0)}x_{i_1} \to e_1 \in \mathbb{C}^{n_{i_1}}$ for an appropriate power t^{s_1}, $0 \le s_1 \le n_{i_1}$. Thus, for $i_1 < i_2$, we obtain $\mathrm{e}^{tA}W = \langle \mathrm{e}^{tA}Z\rangle \to \langle Z'(t)\rangle \to E_{i_1} \oplus E_{i_2}$ for $t \to \infty$, where $E_{i_1} := \ker(A - \lambda_{i_1} I)$ and $E_{i_2} := \ker(A - \lambda_{i_2} I)$. For $i_1 = i_2$, we assume without loss of generality $i_1 = i_2 = 1$ and

$$\begin{bmatrix} x_1 & y_1 \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & 1 & * & \cdots & \cdots & \cdots & * \\ 0 & \cdots & 0 & 0 & \cdots & 1 & * & \cdots & * \end{bmatrix}^{\top}.$$

From linear independence of x_1, y_1 we obtain

$$\left\langle \mathrm{e}^{tJ(0)} \begin{bmatrix} x_1 & y_1 \end{bmatrix} \right\rangle = \left\langle \mathrm{e}^{tJ(0)} \begin{bmatrix} x_1 & y_1 \end{bmatrix} \begin{bmatrix} \frac{t^{s-1}}{(s-1)!} & \frac{t^{s'-1}}{(s'-1)!} \\ \frac{t^{s-2}}{(s-2)!} & \frac{t^{s'-2}}{(s'-2)!} \end{bmatrix} \right\rangle \to \langle e_1, e_2 \rangle$$

for $t \to \infty$, where the indices $s, s' \in \mathbb{N}$ are defined by $x_{1,\,s} = y_{1,\,s'} = 1$ and $x_{1,\,k} = 0$ for $k > s$, $y_{1,\,k} = 0$ for $k > s'$. By linear independence of x_1, y_1, we have $s \neq s'$ and thus

$$\det \begin{bmatrix} \frac{t^{s-1}}{(s-1)!} & \frac{t^{s'-1}}{(s'-1)!} \\ \frac{t^{s-2}}{(s-2)!} & \frac{t^{s'-2}}{(s'-2)!} \end{bmatrix} = \frac{t^{s+s'-3}(s'-s)}{(s-1)!(s'-1)!} \neq 0$$

for $t > 0$. Therefore, $\mathrm{e}^{tA}W = \langle \mathrm{e}^{tA}Z\rangle \to \langle Z(t)\rangle \to \ker(A - \lambda_{i_1} I)^2$ for $t \to \infty$. Thus each ω-limit set contains only a single point, namely a complex 2-dimensional A-invariant subspace. ∎

Theorem 10.4.2. *Assume that $A \in \mathbb{C}^{n\times n}$ is strongly cyclic. The following statements are equivalent.*

(a) The pair (C, A) is perspectively observable on $\mathbb{P}(\mathbb{C}^n)$.
(b) $\dim CE = \dim E$ holds for each complex 2-dimensional A-invariant subspace E.
(c) For all $\alpha, \beta \in \mathbb{C}$ one has

$$\mathrm{rk} \begin{bmatrix} A^2 + \alpha A + \beta I_n \\ C \end{bmatrix} = n. \tag{10.23}$$

Proof. The equivalence of (a) and (b) follows from Proposition 10.3.1 and Theorem 10.4.1. Lemma 10.3.3, for $\mathbb{K} = \mathbb{C}$ and $r = 2$, yields the equivalence of (b) and (c). ∎

10.4.2 The Real Projective Space

Let $\lambda_1, \ldots, \lambda_r \in \mathbb{C}$ denote the distinct real or complex eigenvalues of $A \in \mathbb{R}^{n\times n}$. Then, A is called *strongly regular* if the following conditions are satisfied: (i) $r = n$ and (ii) $\mathrm{Re}\lambda_i \neq \mathrm{Re}\lambda_j$ except for $i = j$ or $\lambda_i = \overline{\lambda_j}$. Moreover, A is said to satisfy the *irrationality condition* (I) if $\mathrm{Im}\lambda_i/\mathrm{Im}\lambda_j \notin \mathbb{Q}$ whenever λ_i and λ_j belong to distinct complex conjugate pairs of eigenvalues.

Theorem 10.4.3. *Assume that $A \in \mathbb{R}^{n\times n}$ is strongly regular. Then any ω-limit set of (10.15) on $\mathbb{G}_2(\mathbb{R}^n)$ is equal to one of the following types:*

(a) $\omega(W) = \{E\}$, *where E is a real 2-dimensional A-invariant subspace.*
(b) $\omega(W) = \{E \oplus \langle \cos(\nu t)z + \sin(\nu t)w\rangle \mid t \in \mathbb{R}\}$, *where E is a 1-dimensional real eigenspace of A and $z + \mathrm{i}w$ is a complex eigenvector of A to a complex eigenvalue $\mu + \mathrm{i}\nu$, $\nu \neq 0$.*
(c) $\omega(W) = \mathrm{cl}\big\{\langle \cos(\nu_1 t)z_1 + \sin(\nu_1 t)w_1\rangle \oplus \langle \cos(\nu_2 t)z_2 + \sin(\nu_2 t)w_2\rangle \mid t \in \mathbb{R}\big\}$, *where $z_1 + \mathrm{i}w_1$ and $z_2 + \mathrm{i}w_2$ are eigenvectors of A to complex eigenvalues $\mu_1 + \mathrm{i}\nu_1$, $\nu_1 \neq 0$ and $\mu_2 + \mathrm{i}\nu_2$, $\nu_2 \neq 0$. Here, $\mathrm{cl}\{...\}$ denotes the closure within $\mathbb{G}_2(\mathbb{R}^n)$.*

If A satisfies additionally the irrationality condition (I) then (c) can be replaced by

($\tilde{c}$) $\omega(W) = \{V_1 \oplus V_2 \mid V_i \subset E_i, \dim V_i = 1, i = 1, 2\}$, *where E_i, $i = 1, 2$ are real 2-dimensional irreducible A-invariant subspace, i.e. $E_i = \langle z_i, w_i\rangle$ with z_i and w_i as in (c).*

Conversely, any set of the above type (a)-(c) acts as an ω-limit set.

Proof. Let A be in real Jordan normal form, i.e.

$$A = \begin{bmatrix} J(\lambda_1) & 0 & \cdots & 0 \\ 0 & J(\lambda_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & J(\lambda_{n-l}) \end{bmatrix}, \quad J(\lambda_i) = \begin{cases} \lambda_i & \text{for } \lambda_i \in \mathbb{R}, \\ \begin{bmatrix} \mu_i & \nu_i \\ -\nu_i & \mu_i \end{bmatrix} & \text{for } \lambda_i = \mu_i + \mathrm{i}\nu_i \in \mathbb{C} \setminus \mathbb{R}, \end{cases} \tag{10.24}$$

where the collection $\lambda_1, \ldots, \lambda_{n-l}$ contains all real eigenvalues of A, as well as one representative for each pair of complex conjugate eigenvalues. Moreover, let $\lambda_1, \ldots, \lambda_{n-l}$ be arranged in decreasing order with respect to their real parts, i.e. $\mathrm{Re}\lambda_1 > \cdots > \mathrm{Re}\lambda_{n-l}$. For $W \in \mathbb{G}_2(\mathbb{R}^n)$ then choose $Z \in \mathbb{R}^{n\times 2}$ such that $W = \langle Z\rangle$. Let

$$V = \begin{bmatrix} 0 & 0 \\ x_{i_1} & 0 \\ \vdots & y_{i_2} \\ \vdots & \vdots \\ x_{n-l} & y_{n-l} \end{bmatrix} \tag{10.25}$$

be in *block echelon form* with *echelon indices* i_1 and i_2, i.e. $x_{i_1} \neq 0$, $x_i = 0$ for all $i < i_1$ and $y_{i_2} \neq 0$, $y_i = 0$ for all $i < i_2$. Here, x_i, y_i, $i = 1, \ldots, n-l$, are either real numbers or vectors in $\mathbb{R}^2$. Thus, $i_1 = i_2$ can only occur for $x_{i_1}, y_{i_1} \in \mathbb{R}^2$ as the definition of the block echelon form implies that x_{i_1}, y_{i_2} are linear independent whenever $i_1 = i_2$. As in Theorem 10.4.1, it suffices to analyse the ω-limit set of $\langle Z'(t)\rangle$ with

$$Z'(t) = \begin{bmatrix} \vdots & \vdots \\ 0 & \vdots \\ \mathrm{e}^{tJ(0)}x_{i_1} & 0 \\ 0 & \mathrm{e}^{tJ(0)}y_{i_2} \\ \vdots & 0 \\ \vdots & \vdots \end{bmatrix}$$

For $i_1 = i_2$, the real 2-dimensional subspace

$$E = \begin{bmatrix} 0 & 0 \\ x_{i_1} & y_{i_1} \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, \qquad x_{i_1}, y_{i_1} \in \mathbb{R}^2 \tag{10.26}$$

is A-invariant. Thus $\langle Z'(t)\rangle = E$ and, therefore, $\omega(W) = E$. If $i_1 < i_2$, then one has to distinguish three different cases.

Case 1. x_{i_1} and y_{i_2} are real:
Then, by the same argument as before, $\langle Z'(t)\rangle = E_{i_1} \oplus E_{i_2}$, where E_{i_1} and E_{i_2} are real eigenspaces to the real eigenvalues λ_{i_1} and λ_{i_2}, respectively. Hence, $\omega(W) = E_{i_1} \oplus E_{i_2}$.

Case 2. $x_{i_1} \in \mathbb{R}$ and $y_{i_2} \in \mathbb{R}^2$ or vice versa:
Then, $Z'(t)$ is periodic of the form

$$Z'(t) = \begin{bmatrix} 0 & 0 \\ x_{i_1} & 0 \\ 0 & 0 \\ 0 & D_{i_2}(t)y_{i_2} \\ 0 & 0 \end{bmatrix}, \quad D_{i_2}(t) = \begin{bmatrix} \cos(\nu_{i_2}t) & \sin(\nu_{i_2}t) \\ -\sin(\nu_{i_2}t) & \cos(\nu_{i_2}t) \end{bmatrix},$$

where $x_{i_1} \neq 0$ belongs to the 1-dimensional real eigenspace E_{i_1} and $y_{i_2} \neq 0$ to the real 2-dimensional irreducible A-invariant subspace E_{i_2} which corresponds to the complex pair $(\lambda_{i_2}, \overline{\lambda}_{i_2})$, $\lambda_{i_2} = \mu_{i_2} + \mathrm{i}\nu_{i_2}$. Thus $E_{i_2} = \langle z_{i_2}, w_{i_2}\rangle$, where $z_{i_2} + \mathrm{i}w_{i_2}$ is a complex eigenvector to λ_{i_2}. We obtain

$$\omega(W) = \big\{E_{i_1} \oplus \langle \cos(\nu_{i_2}t)z_{i_2} + \sin(\nu_{i_2}t)w_{i_2}\rangle \mid t \in \mathbb{R}\big\}.$$

Case 3. $x_{i_1}, y_{i_2} \in \mathbb{R}^2$:
Then, the same arguments as in Case 2 yield

$$Z'(t) = \begin{bmatrix} 0 & 0 \\ D_{i_1}(t)x_{i_1} & 0 \\ 0 & 0 \\ 0 & D_{i_2}(t)y_{i_2} \\ 0 & 0 \end{bmatrix}, \; D_\alpha(t) = \begin{bmatrix} \cos(\nu_{i_\alpha}t) & \sin(\nu_{i_\alpha}t) \\ -\sin(\nu_{i_\alpha}t) & \cos(\nu_{i_\alpha}t) \end{bmatrix}, \; \alpha = 1, 2,$$

where $x_{i_1} \neq 0$ and $y_{i_2} \neq 0$ belong to real 2-dimensional irreducible A-invariant subspaces $E_{i_1} = \langle z_{i_1}, w_{i_1} \rangle$ and $E_{i_2} = \langle z_{i_2}, w_{i_2} \rangle$, respectively, with $z_{i_1} + \mathrm{i}w_{i_1}$ and $z_{i_2} + \mathrm{i}w_{i_2}$ defined as in Case 2. Thus, $Z'(t)$ is periodic if $\nu_{i_1}/\nu_{i_2} \in \mathbb{Q}$, and defines a dense winding on a 2-dimensional torus if $\nu_{i_1}/\nu_{i_2} \notin \mathbb{Q}$. In either case, one has

$$\omega(W) = \mathrm{cl}\big\{ \langle \cos(\nu_{i_1} t) z_{i_1} + \sin(\nu_{i_1} t) w_{i_1} \rangle \oplus \langle \cos(\nu_{i_2} t) z_{i_2} + \sin(\nu_{i_2} t) w_{i_2} \rangle \;\big|\; t \in \mathbb{R} \big\}.$$

If the irrationality condition $\nu_{i_1}/\nu_{i_2} \notin \mathbb{Q}$ is satisfied, we obtain the simplified expression as

$$\omega(W) = \{ V_1 \oplus V_2 \mid V_\alpha \subset E_{i_\alpha},\, \dim V_\alpha = 1,\, \alpha = 1, 2 \}.$$

This completes the proof. ∎

Theorem 10.4.4. *Assume that $A \in \mathbb{R}^{n \times n}$ is strongly regular and satisfies the irrationality condition (I). The following statements are equivalent*

(a) The pair (C, A) is perspectively observable on $\mathbb{P}(\mathbb{R}^n)$.
(b) For each 1- and 2-dimensional A-invariant subspace E one has the equality $\dim CE = \dim E$.
(c) For all $\alpha, \beta \in \mathbb{R}$ one has

$$\mathrm{rk} \begin{bmatrix} A^2 + \alpha A + \beta I_n \\ C \end{bmatrix} = n. \tag{10.27}$$

Proof. (a) $\Longrightarrow$ (b): Assume that there exists a 1- or 2-dimensional A-invariant subspace E with $\dim CE < \dim E$. Case 1: If $\dim E = 2$, then $\dim C\mathrm{e}^{tA}E = \dim CE \leq 1$. Thus, (C, A) is not perspectively observable on $\mathbb{P}(\mathbb{R}^n)$ by Lemma 10.3.2. Case 2: If $\dim E = 1$, then choose any other 1- or 2-dimensional irreducible A-invariant subspace E' and consider $E \oplus E'$. For $\dim E' = 1$, continue as in Case 1. For $\dim E' = 2$ with $E' = \langle x_1, x_2 \rangle$, let $W := E \oplus \langle x_1 \rangle$. Then, one has $\dim C\mathrm{e}^{tA}W = \dim C\big(E \oplus \langle \mathrm{e}^{tA}x_1 \rangle\big) = \dim \langle C\mathrm{e}^{tA}x_1 \rangle \leq 1$ for all $t \in \mathbb{R}$. Again, (C, A) is not perspectively observable on $\mathbb{P}(\mathbb{R}^n)$ by Lemma 10.3.2.

(b) $\Longrightarrow$ (a): Assume that (C, A) is not perspectively observable on $\mathbb{P}(\mathbb{R}^n)$. By Proposition 10.3.1, there exists $W \in \mathbb{G}_2(\mathbb{R}^n)$ such that $\dim CW_\infty \leq 1$ for all $W_\infty \in \omega(W)$. Hence, using Theorem 10.4.3, we have to distinguish between three cases. Case 1: $\omega(W)$ is of type (a). Then, $\omega(W)$ is a real 2-dimensional A-invariant subspace, which contradicts (b). Case 2: $\omega(W)$ is of type (b). One has

$$\dim C\big(E \oplus \langle \cos(\nu t) z + \sin(\nu t) w \rangle\big) \leq 1$$

for all $t \in \mathbb{R}$, where $E = \langle x \rangle$ is a real 1-dimensional eigenspace and $E' := \langle z, w \rangle$ is a real 2-dimensional irreducible A-invariant subspace. Then, by choosing $t = 0$ and $t = (2\nu)^{-1}\pi$, respectively, we conclude that Cx, Cz and Cx, Cw are linear dependent pairs. Thus, either $Cx = 0$ or Cz, Cw are linear dependent. This, however, contradicts (b). Case 3: $\omega(W)$ in of type (c). Then the irrationality condition (I) implies

$$\omega(W) = \big\{ \langle \cos(\nu t)z + \sin(\nu t)w \rangle \oplus \langle \cos(\nu' t')z' + \sin(\nu' t')w' \rangle \;\big|\; t, t' \in \mathbb{R} \big\},$$

where $E := \langle z, w \rangle$ and $E' := \langle z', w' \rangle$ are real 2-dimensional irreducible A-invariant subspaces, cf. Theorem 10.4.3. The same line of arguments as in Case 2 yields a contradiction to (b). Hence, we conclude that (b) implies (a).

(b) $\Longleftrightarrow$ (c): Lemma 10.3.3, for $\mathbb{K} = \mathbb{R}$ and $r = 2$, yields the desired equivalence. ∎

We close with some examples. The first one is taken from Dayawansa et al. [5]. It shows that a system can be perspectively observable over $\mathbb{R}$ without being perspectively observable over $\mathbb{C}$.

Example 1. Let

$$A := \begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & -2 & 2 \end{bmatrix}, \quad C := \begin{bmatrix} 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}. \tag{10.28}$$

Note that $x_1^\top := (1, \mathrm{i}, 0, 0)$ and $x_2^\top := (0, 0, 1, \mathrm{i})$ cannot be distinguished via their perspective outputs $C\mathrm{e}^{tA}\langle x_1 \rangle$ and $C\mathrm{e}^{tA}\langle x_2 \rangle$, respectively. Thus (10.28) is not perspectively observable over $\mathbb{C}$. Perspective observability over $\mathbb{R}$ follows by inspection, see also [5].

The second example shows that the above Theorem 10.4.4 is false if the irrationality condition (I) on A does not hold. Nevertheless, we conjecture that the irrationality condition can be significantly relaxed while still guaranteeing perspective observability on $\mathbb{P}(\mathbb{R}^n)$.

Example 2. Let

$$A := \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 1 & 2 \end{bmatrix}, \quad C := \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \tag{10.29}$$

Here, one has no 1-dimensional A-invariant subspaces and exactly two 2-dimensional ones, namely $E := \langle e_1, e_2 \rangle$ and $E' := \langle e_3, e_4 \rangle$, where $e_i \in \mathbb{R}^4$ denotes the i-th standard basis vector. Clearly, (C, A) meets condition (b) of Theorem 10.4.4. However, for $V := \langle e_1 \rangle$ and $V' := \langle e_3 \rangle$ we obtain $C\mathrm{e}^{tA}V = C\mathrm{e}^{tA}V'$ for all $t \in \mathbb{R}$. Thus, the pair (C, A) is not perspectively observable on $\mathbb{P}(\mathbb{R}^n)$.

The last example shows that condition (b) of Theorem 10.4.4 cannot be simplified by requiring $\dim CE = \dim E$ just for the 2-dimensional A-invariant subspaces. This is due to the fact that a 1-dimensional eigenspace E of A may have no 1-dimensional complement V such that $E \oplus V$ is A-invariant. This difficulty does not occur in the complex case.

Example 3. Let

$$A := \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad C := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{10.30}$$

Then, $\dim CE = \dim E$ is satisfied for all 2-dimensional A-invariant subspaces. Yet, the pair (C, A) is not perspectively observable. To see this, choose for example $V = \langle e_1 \rangle$ and $V' = \langle e_1 + e_3 \rangle$.

We now explain how the previous results relate to earlier work on perspective observability. Ghosh and Rosenthal proved in [8, 15] that any linear perspective system (10.13) is perspectively observable on the *complex* Grassmann manifold $\mathbb{G}_m(\mathbb{C}^n)$, if and only if the full rank condition

$$\operatorname{rk} \begin{bmatrix} \prod_{i=0}^{m} (A - \lambda_i I_n) \\ C \end{bmatrix} = n \tag{10.31}$$

holds for all complex coefficients $\lambda_0, \cdots, \lambda_m \in \mathbb{C}$. This generalizes an earlier work by Dayawansa et al. [5], who established the same result for $m = 1$. Their approach is quite different to ours as they do not exploit structural information on the ω–limit sets of the (extended) Riccati equation (10.6). Moreover, perspective observability is shown to hold in the real case $\mathbb{K} = \mathbb{R}$, provided A has only real eigenvalues and condition (10.31) holds for arbitrary *real* parameters $\lambda_0, \cdots, \lambda_m \in \mathbb{R}$. In the complex case, condition (10.31) is equivalent to our rank condition (10.23). This is not true in the real case. Thus the results in [5, 8, 15] show that Theorem 10.4.2 holds without any strong cyclicity assumption on A. The situation is more complicated in the real case and not yet fully understood. In [9] it is claimed, that

$$\operatorname{rk} \begin{bmatrix} A^2 + \alpha A + \beta I_n \\ C \end{bmatrix} = n \tag{10.32}$$

for all $\alpha, \beta \in \mathbb{R}$ is necessary and sufficient for perspective observability on real projective spaces. Our above Example 2 shows that this is false and, in fact, Theorem 10.4.4 cannot hold without imposing a constraint on the eigenvalues of A.

Acknowledgement. This work was supported by the grant HE 1858/12-1 of the DFG-Schwerpunktprogramm 1305.

References

1. Ammar, G.S., Martin, C.F.: The geometry of matrix eigenvalue methods. Acta Applic. Math. 5, 239–278 (1986)
2. Brockett, R.W.: Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems. In: Proc. of the 27th IEEE Conference on Decision and Control, Austin, Texas, pp. 799–803 (1988); Linear Algebra Appl., vol. 146, pp. 79–91 (1991)

3. Brockett, R.W.: Smooth dynamical systems which realize arithmetical and logical operations. In: Nijmeijer, H., Schumacher, J. (eds.) Three Decades of Mathematical System Theory. LNCIS, vol. 135, pp. 19–30. Springer, Berlin (1989)
4. Bröcker, T., tom Dieck, T.: Representations of Compact Lie Groups. Springer, New York (1995)
5. Dayawansa, W.P., Ghosh, B.K., Martin, C.F., Wang, X.: A necessary and sufficient condition for the perspective observability problem. Syst. Control Lett. 25, 159–166 (1995)
6. Dirr, G., Helmke, U.: Accessibility of a class of generalized double bracket flows. Communications in Information and Systems (2008) (to appear)
7. Ghosh, B.K., Jankovic, M., Wu, Y.: Perspective problems in system theory and its application to machine vision. J. Math. Syst. Estim. Control 4, 3–38 (1994)
8. Ghosh, B.K., Rosenthal, J.: A generalized Popov-Belevitch-Hautus test of observability. IEEE Trans. Automat. Contr. 40, 176–180 (1995)
9. Ghosh, B.K., Martin, C.F.: Homogeneous dynamical systems theory. IEEE Trans. Automat. Contr. 47, 462–472 (2002)
10. Helmke, U., Fuhrmann, P.A.: Controllability of matrix eigenvalue algorithms: the inverse power method. Syst. Control Lett. 41, 57–66 (2000)
11. Helmke, U.: Isospectral flows on symmetric matrices and the Riccati equation. System and Control Letters 16, 159–166 (1991)
12. Helmke, U., Moore, J.: Optimization and Dynamical Systems. Springer, London (1994)
13. Hespanha, J.: State estimation and control for systems with perspective outputs, Technical Report, UC Santa Barbara (2002)
14. Jordan, J.: Reachable sets of numerical iteration schemes: A system semigroup approach, Ph. D. Thesis, University of Würzburg (2008)
15. Rosenthal, J.: An observability criterion for dynamcal systems governed by Riccati differential equations. IEEE Trans. Automat. Contr. 41, 434–436 (1996)
16. Shayman, M.: Phase portrait of the Matrix Riccati Equation. SIAM J. Contr. and Optim. 24, 1–65 (1986)
17. Sontag, E.D.: Mathematical Control Theory. Springer, New York (1990)
18. van Dooren, P., Sepulchre, R.: Shift policies in QR-like algorithms and feedback control of self-similar flows. In: Bondel, V., et al. (eds.) Open Problems in Mathematical Systems and Control Theory, pp. 245–249. Springer, London (1999)
19. Völklein, H.: Transitivitätsfragen bei linearen Gruppen. Arch. Math. 36, 23–34 (1981)

11

Nonlinear Locally Lipschitz Output Regulation in Presence of Non-hyperbolic Zero-Dynamics

Alberto Isidori[1,2] and Lorenzo Marconi[2]

[1] DIS, Università di Roma "La Sapienza", Italy
albisidori@dis.uniroma1.it
[2] CASY-DEIS, Università di Bologna, Italy
lorenzo.marconi@unibo.it

This work is dedicated to the memory of Wijesuriya P. Dayawansa, with admiration for all his outstanding scientific achievements and deep sorrow for the loss of a friend and advisor.

Summary. The purpose of this work is to complement some recent advances in the field of robust stabilization of nonlinear systems contained in [11]. By exploiting a general tool design presented in that paper and briefly summarized here, we address the problem of output regulation for nonlinear systems whose zero dynamics posses compact attractors which are asymptotically but not necessarily exponentially stable. The crucial requirement of designing a locally Lipschitz regulator makes the problem at hand particularly challenging, and conventional tools based on pure high-gain feedback cannot be used. A few academic examples are also presented which illustrate the proposed method.

11.1 Introduction and Problem Statement

Internal model-based regulators have been extensively studied in the control literature, specifically in the context of problems of output regulation. Many techniques have been proposed in the recent years, which have yielded conceptual frameworks and constructive design procedures to the purpose of handling a large variety of advanced applications (see [2, 15, 3, 7]). As classically explained in the literature on output regulation ([2]), the need of embedding internal models in output feedback regulators is motivated by the request of generating exactly the *steady state* control input necessary to fulfill the regulation objective (typically expressed in terms of zeroing a regulation error). In this respect, internal models are meant generate "only" those control inputs which are associated to *steady state behaviors* of the so-called zero dynamics (see [1] and [2]).

Recently, [11] has shown the effectiveness of the the conceptual framework of the theory of output regulation in the un-conventional setting of robust output feedback stabilization, by means of locally Lipschitz regulators, of nonlinear systems possessing zero dynamics which are asymptotically but not necessarily

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 185–198.
springerlink.com

exponentially stable. The idea in [11] (see also [13]) is to design regulators embedding "redundant" internal models able to reproduce not only the control inputs associated to a steady-state behavior (which, in "simple" stabilization problems, could be trivially zero) but also reproducing behaviors which characterize the zero dynamics close to the asymptotic attractor. In this framework, the internal model have been discovered to have a possible role not only in steady-state but also during the transient phase. All the results presented in [11] are focused on output feedback stabilization of compact attractors for nonlinear systems and no special attention was given to the problem of output regulation. In this work we complement the results of [11], by showing how the main idea and tools proposed in that paper can be successfully used also in dealing with output regulation problems involving nonlinear minimum-phase systems possessing a non-hyperbolic zero dynamics. More specifically, we consider the class of smooth nonlinear systems described in normal form as

$$\begin{aligned} \dot{z} &= f(w, z, y_1) && z \in \mathbb{R}^m \\ \dot{y}_1 &= y_2 \\ &\vdots && y_i \in \mathbb{R} \\ \dot{y}_{r-1} &= y_r \\ \dot{y}_r &= b(w, z, y) + a(w, z, y)u \end{aligned} \tag{11.1}$$

with control input $u \in \mathbb{R}$ and measurable output $y_m = y_1$, in which $y = \text{col}(y_1, \ldots, y_r)$. The exogenous input w is thought as generated by an autonomous smooth system of the form

$$\dot{w} = s(w) \qquad w \in W \subset \mathbb{R}^s \tag{11.2}$$

in which W is a compact set invariant for (11.2). The "high frequency gain" $a(\cdot)$ in (11.1) is assumed to be nonzero and, without loss of generality, positive for all $(w, z, y) \in W \times \mathbb{R}^m \times \mathbb{R}^r$. The previous system will be studied under the following "minimum-phase" assumption.

Assumption 11.1.1. *There exists a compact set $\mathcal{A} \subset W \times \mathbb{R}^m$ which is Locally Asymptotically Stable (LAS) for the system*

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(w, z, 0) \end{aligned} \tag{11.3}$$

with a domain of attraction $\mathcal{D}(\mathcal{A})$. ◁

In this framework we address the problem of semiglobal stabilization by means of a *locally Lipschitz* output–feedback regulator.

Problem 11.1.1. Given arbitrary compact sets $\mathcal{Z} \in \mathcal{D}(\mathcal{A})$ and $\mathcal{Y} \in \mathbb{R}^r$, design a *locally Lipschitz* regulator of the form

$$\begin{aligned} \dot{\eta} &= \Phi(\eta, y_m) \qquad \eta \in \mathbb{R}^\nu \\ u &= \Upsilon(\eta, y_m) \end{aligned} \tag{11.4}$$

such that, in the closed-loop system (11.1), (11.2), (11.4), for some choice of compact sets $\mathcal{B}$ and $\mathcal{N}$ of $\mathbb{R}^\nu$, the set $\mathcal{A} \times \{0\} \times \mathcal{B}$ is[3] LAS($\mathcal{Z} \times \mathcal{Y} \times \mathcal{N}$) ◁

It is interesting to note how, in the previous framework, it is possible to address problems of *robust output-feedback stabilization* and *output regulation*, which can not be easily handled with the available design tools.

In the first case (*robust output feedback stabilization*) ([8], [9]), a possible representative and meaningful scenario to be considered in order to appreciate the challenging aspects of the proposed problem, is the one in which the variable w models constant parametric uncertainties whose values range in the set W (in which case system (11.2) simplifies as $\dot{w} = 0$), in which the set $\mathcal{A}$ collapses to $W \times \{0\}$ (namely the origin $z = 0$ is an equilibrium for the system $\dot{z} = f(w, z, 0)$ which is LAS for any possible value of the uncertainties), and in which the function $b(\cdot)$ in (11.1) is such that $b(w, 0, 0) = 0$ for all $w \in W$ (in which case the origin of system (11.1) with $u \equiv 0$ is an equilibrium point). Under these circumstances, the addressed problem boils down to a "classical" problem of stabilizing an equilibrium point (the origin) of a system in presence of parametric uncertainties. Even in this simplified scenario, though, the problem at hand is far from being easily solvable due to two main features characterizing the previous framework: the first one is the absence of local exponential stability properties of the set $\mathcal{A}$ (only assumed to be LAS) while the second one is the requirement of a locally Lipschitz regulator. As a matter of fact, the set $\mathcal{A}$ being only LAS (and not LES) it is not possible to use, off-the-shelf, high-gain linear arguments in order to deal with the stability of system (11.1) due to the fact that the local asymptotic gain (see [8]) of the "inverse dynamics" (11.3) is not, in general, linear. On the other hand, the fact that the regulator is required to be locally Lipschitz does not allow one to use control laws which are only continuous at the origin which, having in mind small gain arguments and results on gain assignment for nonlinear systems (see [10], [9]), one would adopt to handle the nonlinearity of the local asymptotic gain of (11.3).

The story becomes even more challenging if one looks at the previous framework as a problem of *output regulation* (see [2], [14], [15], [5], [7]) in which the variable w may assume the meaning of reference signals to be tracked and/or of disturbances to be rejected generated by the autonomous system (11.2) which, in the output regulation literature, is usually referred to as the exosystem. In this context the measurable output y_m, which must be asymptotically steered to zero, has the role of *regulation error* and the set $\mathcal{A}$ assumes the meaning of *steady state locus* (by using the terminology introduced in [2]). The latter, according to recent developments in the field (see [14]), is usually expressed as the graph of a map, namely it is assumed the existence of a smooth function $\pi : \mathbb{R}^s \to \mathbb{R}^m$, possibly *set-valued* (see [2]), such that

[3] In the following, for a smooth system $\dot{x} = f(x)$, $x \in \mathbb{R}^n$, a compact set $\mathcal{A}$ is said to be LAS($\mathcal{X}$) (respectively LES($\mathcal{X}$)), with $\mathcal{X} \subset \mathbb{R}^n$ a compact set, if it is locally asymptotically (respectively exponentially) stable with a domain of attraction containing $\mathcal{X}$.

$$\mathcal{A} = \{(w, z) \in W \times \mathbb{R}^m \ : \ z = \pi(w)\}\,.$$

Not surprisingly, it turns out that the output regulation problem hides the same challenging design aspects highlighted above for the stabilization problem which are even worsened by the fact that the function $b(\cdot)$ in (11.1) is not, in general, vanishing on the desired attractor $\mathcal{A} \times \{0\}$ which thus is not necessarily forward invariant for (11.1) with $u \equiv 0$. In this respect, what it is required from the controller (11.4) is also the ability to asymptotically generate a non-zero steady state control input, namely to offset the term $b(w, z, 0)/a(w, z, 0)$ with $(w, z) \in \mathcal{A}$, by only processing the regulation error. Indeed, this distinguishing feature of the problem of output regulation is what motivates the key concept of internal model and the need of designing internal model-based controllers (see [2], [14], [7]). It is worth noting how, in a local setting, the problem of handling non-hyperbolic zero dynamics in output regulation problems has been addressed in [6].

This work is organized as follows. In the next section we present and slightly extend the results presented in [11] which are instrumental in the solution of the problem formulated above. Then in Section 3 the main result is stated and proved. Section 4 concludes with the presentation of two academic examples showing the effectiveness of the proposed theory both for stabilization and regulation of nonlinear minimum-phase systems with non-hyperbolic zero dynamics.

11.2 The Basic Tool

In this section we briefly review, and slightly extend, the framework and results presented in [11] which are instrumental in the analysis of the next section. The main goal is to introduce a theoretical tool for handling the presence of non-hyperbolic zero dynamics in the problem of robust stabilization of nonlinear systems. More specifically we consider a system of the form

$$\begin{array}{ll} \dot{x} = f(x, y) & x \in \mathbb{R}^n \\ \dot{y} = a(x, y)\ [\kappa\, Ay + B(q(x, y) + v)] & y \in \mathbb{R}^r \end{array} \tag{11.5}$$

with measurable output

$$y_m = Cy \qquad y_m \in \mathbb{R}$$

in which the triplet (A, B, C) is assumed to have relative degree r and A is Hurwitz, κ is a positive design parameter, v is a control input and $a(x, y)$ a smooth real valued function with $a(x, y) > 0$ for all x and y. System (11.5) is assumed to evolve on a closed invariant set of the form $\mathcal{C} \times \mathbb{R}^r$ in which $\mathcal{C}$ is a closed set of $\mathbb{R}^n$. As said, we are interested to study the previous system under a non-hyperbolic minimum-phase assumption which, in particular, asks for the existence of a compact set $\mathcal{A}$ which is LAS (but not necessarily LES) for the zero dynamics of (11.5), given by

$$\dot{x} = f(x, 0)\,. \tag{11.6}$$

In this setting the stabilization problem consists of properly choosing the value of the parameter κ and of designing a *locally Lipschitz* output feedback controller for the control input v so that $\mathcal{A}\times\{0\}$ is asymptotically stable for the closed-loop system. We refer the reader to [11] for a thorough discussion about the difficulties which characterize the previous problem and which make not possible to use, off-the-shelf, existing tools. In this work we only limit ourselves to observe that the absence of a local exponential stability property for (11.6) and the fact that the function $q(x,y)$, coupling the x and the y dynamics, is not necessarily vanishing on $\mathcal{A}\times\{0\}$ (namely that the set $\mathcal{A}\times\{0\}$ is not necessarily forward invariant for the system (11.5) with $v\equiv 0$), does not allow one to solve the problem at hand by simply increasing the value of κ and by choosing $v\equiv 0$. Indeed, in the case the two previous pathologies are dropped, a large value of κ succeeds in solving the problem as formalized in the next theorem (see [16], [14]).

Theorem 11.2.1. *Let $\mathcal{A}$ be LES($\mathcal{X}$) for the system $\dot{x}=f(x,0)$ and $q(x,0)\equiv 0$ for all $x\in\mathcal{A}$. Then for any compact set $\mathcal{Y}\subset\mathbb{R}^r$ there exists a $\kappa^\star>0$ such that for all $\kappa\geq\kappa^\star$ the set $\mathcal{A}\times\{0\}$ is LES($\mathcal{X}\times\mathcal{Y}$) for (11.5) with $v\equiv 0$.*

In the case in which $\mathcal{A}$ is "only" LAS and $q(x,y)$ is not vanishing on $\mathcal{A}\times\{0\}$ a not trivial design of the control input v is needed. In this respect the following definition of local exponential reproducibility of a triplet plays a crucial role.

Definition 11.2.1. *(LER, ISS-LER) A triplet $(g(\cdot),h(\cdot),\mathcal{A})$, where $g:\mathbb{R}^m\to\mathbb{R}^m$ and $h:\mathbb{R}^m\to\mathbb{R}$ are smooth functions and $\mathcal{A}\subset\mathbb{R}^m$ is a compact set, is said to be* Locally Exponentially Reproducible *(LER), if there exists a compact set $\mathcal{R}\supseteq\mathcal{A}$ which is LES for $\dot{z}=g(z)$ and, for any bounded set $\mathcal{Z}$ contained in the domain of attraction of $\mathcal{R}$, there exist an integer p,* locally Lipschitz *functions $\varphi:\mathbb{R}^p\to\mathbb{R}^p$, $\gamma:\mathbb{R}^p\to\mathbb{R}$, and $\psi:\mathbb{R}^p\to\mathbb{R}^p$, with ψ a complete vector field, and a smooth function $T:\mathbb{R}^m\to\mathbb{R}^p$, such that*

$$h(z)+\gamma(T(z))=0\qquad\forall\, z\in\mathcal{R}\,, \tag{11.7}$$

and for all $\xi_0\in\mathbb{R}^p$ and $z_0\in\mathcal{Z}$ the solution $(\xi(t),z(t))$ of

$$\begin{aligned}\dot{z}&=g(z) & z(0)&=z_0\\ \dot{\xi}&=\varphi(\xi)+\psi(\xi)\,h(z) & \xi(0)&=\xi_0\end{aligned} \tag{11.8}$$

satisfies,

$$|(\xi(t),z(t))|_{\operatorname{graph} T|_{\mathcal{R}}}\leq\beta(t,|(\xi_0,z_0)|_{\operatorname{graph} T|_{\mathcal{R}}}) \tag{11.9}$$

where $\beta(\cdot,\cdot)$ is a locally exponentially class-$\mathcal{KL}$ function.

Furthermore the triplet in question is said to be robustly Locally Exponentially Reproducible *(ISS-LER) if it is LER and, in addition, for all locally essentially bounded $v(t)$, for all $\xi_0\in\mathbb{R}^p$ and $z_0\in\mathcal{Z}$ the solution $(\xi(t),z(t))$ of*

$$\begin{aligned}\dot{z}&=f(z) & z(0)&=z_0\\ \dot{\xi}&=\varphi(\xi)+\psi(\xi)[q(z)+v(t)] & \xi(0)&=\xi_0\end{aligned} \tag{11.10}$$

satisfies

$$|(\xi(t), z(t))|_{\text{graph } T|_{\mathcal{R}}} \leq \beta(t, |(\xi_0, z_0)|_{\text{graph } T|_{\mathcal{R}}}) + \ell(\sup_{\tau \leq t} |v(\tau)|) \tag{11.11}$$

where $\beta(\cdot,\cdot)$ is a locally exponentially class-$\mathcal{KL}$ function and ℓ is a class-$\mathcal{K}$ function. ◁

We refer the reader to the work [11] for comments about this definition and for the presentation of meaningful sufficient conditions under which a triplet can be claimed to be ISS-LER and thus LER (see also Section 11.4 with illustrative examples). In this work we limit ourselves to recall a result, stated and proved in [11], which presents a sufficient condition under which a triplet can be claimed ISS-LER (and thus LER).

Proposition 11.2.1. *Let $g : \mathbb{R}^m \to \mathbb{R}^m$ and $h : \mathbb{R}^m \to \mathbb{R}$ be given smooth functions and $\mathcal{A} \subset \mathbb{R}^m$ be a given compact set which is LAS for $\dot{z} = g(z)$. Assume, in addition, that there exist a $\tilde{m} > 0$, a compact set $\mathcal{S}$ such that $\mathcal{A} \subset \text{int}\mathcal{S}$ and a locally Lipschitz function $\Upsilon : \mathbb{R}^{\tilde{m}} \to \mathbb{R}$ such that the following identity holds*

$$L^{\tilde{m}}_{g(z)} h(z) = \Upsilon(\, h(z),\, L_{g(z)}h(z), \ldots,\, L^{\tilde{m}-1}_{g(z)} h(z)\,) \qquad \forall\, z \in \mathcal{S}\,. \tag{11.12}$$

Then the triplet $(g, h, \mathcal{A})$ is ISS-LER. In particular (φ, ψ, γ) can be taken as the functions $\varphi : \mathbb{R}^{\tilde{m}} \to \mathbb{R}^{\tilde{m}}$, $\psi : \mathbb{R}^{\tilde{m}} \to \mathbb{R}^{\tilde{m}}$, $\gamma : \mathbb{R}^{\tilde{m}} \to \mathbb{R}$ defined as

$$\varphi(\xi) = \begin{pmatrix} \xi_2 + c_0 L \xi_1 \\ \xi_3 + c_1 L^2 \xi_1 \\ \vdots \\ \xi_{\tilde{m}} + c_{\tilde{m}-2} L^{\tilde{m}-1} \xi_1 \\ \Upsilon_c(\xi_1,\, \xi_2,\, \ldots,\, \xi_{\tilde{m}}) + c_{m-1} L^{\tilde{m}} \xi_1 \end{pmatrix}, \qquad \psi(\xi) = \begin{pmatrix} -c_0 L \\ -c_1 L^2 \\ \vdots \\ -c_{\tilde{m}-2} L^{\tilde{m}-1} \\ -c_{\tilde{m}-1} L^{\tilde{m}} \end{pmatrix}$$

and $\gamma(\xi) = \xi_1$, where L is a positive design parameter to be taken sufficiently large, c_i, $i = 0, \ldots, \tilde{m}-1$, are such the polynomial $s^{\tilde{m}} + c_{\tilde{m}-1} s^{\tilde{m}-1} + \ldots + c_1 s + c_0$ is Hurwitz, and $\Upsilon_c(\cdot)$ is any bounded function such that $\Upsilon_c \circ \tau(z) = \Upsilon(z)$ for all $z \in \mathcal{S}$ where $\tau : \mathbb{R}^m \to \mathbb{R}^{\tilde{m}}$ is defined as

$$\tau(z) = \left(h(z)\; L_{g(z)}h(z) \cdots\; L^{\tilde{m}-1}_{g(z)} h(z) \right)^{\mathrm{T}}.$$

Remark 11.2.1. If $\mathcal{A}$ reduces to a singleton $a \in \mathbb{R}^m$ it turns out that a *sufficient* condition for the existence of a locally Lipschitz function Υ and a compact set $\mathcal{S}$ containing a in its interior such that condition (11.12) is fulfilled, is that the pair $(g(z), h(z))$ has a linear approximation at $z = a$ which is observable. As a matter of fact, under the previous conditions, the function

$$\begin{aligned} Y : \mathbb{R}^m &\to \mathbb{R}^m \\ z &\mapsto \tilde{z} := \left(h(z)\; L_{g(z)}h(z) \ldots\; L^{m-1}_{g(z)} h(z) \right)^{\mathrm{T}} \end{aligned}$$

is a local diffeomorphism at $z = a$ and thus there exists an open set $\tilde{\mathcal{S}} \supset a$ and a smooth function $Y^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ such that

$$Y^{-1} \circ Y(z) = z \qquad \forall\, z \in \tilde{S}\,.$$

This implies the existence of a smooth function $\Upsilon : \mathbb{R}^m \to \mathbb{R}$ such that

$$L^m_{g(z)} h(z)\Big|_{z=Y^{-1}(\tilde{z})} = \Upsilon(h(z), L_{g(z)}h(z), \ldots, L^{m-1}_{g(z)}h(z))$$

which, in turn, implies (11.12) with $\mathcal{S}$ any compact set included in $\tilde{\mathcal{S}}$ and containing a in its interior. As a consequence, by Proposition 11.2.1, it immediately follows that if a is LAS for $\dot{z} = g(z)$ and the linear approximation of $(g(z), h(z))$ at $z = a$ is observable, then the triplet $(g(z), h(z), a)$ is ISS-LER. ◁

With the definition ISS-LER at hand we are able to state the following result presenting a sufficient condition under which the stabilization problem formulated above can be solved.

Theorem 11.2.2. *Let $\mathcal{A}$ be LAS($\mathcal{X}$) for the system $\dot{x} = f(x, 0)$ for some compact set $\mathcal{X} \subset \mathcal{C}$. Assume, in addition, that the triplet $(f(x, 0), q(x, 0), \mathcal{A})$ is LER. Then there exist a locally Lipschitz regulator of the form*

$$\dot{\eta} = \Phi_k(\eta, y_{\rm m}) \qquad v = \Upsilon_k(\eta, y_{\rm m}) \qquad \eta \in \mathbb{R}^\nu\,, \tag{11.13}$$

a compact set $\mathcal{R} \supseteq \mathcal{A}$, a continuous function $\tau : \mathcal{R} \to \mathbb{R}^\nu$, and, for any compact set $\mathcal{Y} \subset \mathbb{R}^r$ and $\mathcal{N} \subset \mathbb{R}^\nu$, a positive constant $\kappa^\star$, such that for all $\kappa \geq \kappa^\star$ the set

$$\text{graph}\tau \times \{0\} = \{(x, y, \eta) \in \mathcal{R} \times \mathbb{R}^r \times \mathbb{R}^\nu \quad : \quad y = 0\,,\ \eta = \tau(x)\}$$

is LES($\mathcal{X} \times \mathcal{Y} \times \mathcal{N}$) for (11.5), (11.13) and the set

$$\text{graph}\ \tau|_{\mathcal{A}} \times \{0\} = \{(x, y, \eta) \in \mathcal{A} \times \mathbb{R}^r \times \mathbb{R}^\nu \quad : \quad y = 0\,,\ \eta = \tau(x)\}$$

is LAS($\mathcal{X} \times \mathcal{Y} \times \mathcal{N}$) for (11.5), (11.13).

Proof. The proof of the theorem can be found in [11] (see Theorem 2) with the only mild difference that, in the framework of [11], the "high frequency gain" $a(x, y)$ is considered unitary. The control structure is chosen to be of the form

$$\begin{aligned} \dot{\eta} &= \varphi(\eta) - \psi(\eta)[\gamma(\eta) + \kappa B^{\rm T} A y] \\ v &= \gamma(\eta) \end{aligned} \tag{11.14}$$

in which $(\varphi(\cdot), \psi(\cdot), \gamma(\cdot))$ are the locally Lipschitz functions which are associated to the triplet $(f(x, 0), q(x, 0), \mathcal{A})$ in the definition of local exponential reproducibility and κ is the same of (11.5). The same arguments of Theorem 2 of [11] can be used off-the-shelf in case of not unitary function $a(x, y)$ by simply replacing the change of variable (51) of [11] with

$$\eta \ \to\ \chi := \phi_\psi\Big(\int_0^{B^{\rm T} y} \frac{ds}{a(x, y', s)}, \eta\Big) \tag{11.15}$$

in which $a(x, y', y_r) = a(x, y)$, $y' = \text{col}(y_1, \ldots, y_{r-1})$. We refer the reader to the proof of Theorem 2 in [11] for details. ◁

By bearing in mind Remark 11.2.1, the previous theorem immediately leads to formulate the following corollary.

Corollary 11.2.1. *Let the singleton $a \in \mathbb{R}^n$ be LAS($\mathcal{X}$) for the system $\dot{x} = f(x,0)$ for some compact set $\mathcal{X} \subset \mathcal{C}$. Suppose the linear approximation of $(f(x,0), q(x,0))$ at $x = a$ is observable. Then there exist a locally Lipschitz regulator of the form (11.13), a vector $b \in \mathbb{R}^\nu$ and, for any compact set $\mathcal{Y} \subset \mathbb{R}^r$ and $\mathcal{N} \subset \mathbb{R}^\nu$, a positive constant $\kappa^\star$, such that for all $\kappa \geq \kappa^\star$ the singleton $(a,0,b) \in \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^\nu$ is LAS($\mathcal{X} \times \mathcal{Y} \times \mathcal{N}$) for (11.5),(11.13).*

11.3 Robust Stabilization and Regulation of Non-hyperbolic Minimum-Phase Nonlinear Systems

In this section we formulate and prove the main result of the work by providing a solution to the problem of output regulation formulated in Section 1 (see also [12] for a preliminary presentation). In order to simplify the notation, in the following we shall drop in (11.1) the dependence from the variable w which, in turn, will be thought as embedded in the variable z. This, with a mild abuse of notation, will allow us to rewrite system (11.1) and (11.2) in the more compact form

$$\begin{aligned} \dot{z} &= f(z, y_1) && z \in W \times \mathbb{R}^m \\ \dot{y}_1 &= y_2 \\ &\vdots && y_i \in \mathbb{R} \\ \dot{y}_{r-1} &= y_r \\ \dot{y}_r &= b(z,y) + a(z,y)u \end{aligned} \tag{11.16}$$

and system (11.3) as $\dot{z} = f(z,0)$. Furthermore, since W is invariant for (11.2), system (11.16) evolves on the closed cylinder $W \times \mathbb{R}^{m+r}$ and it is natural to regard these dynamics restricted to $W \times \mathbb{R}^{m+r}$ and endow the latter with the relative topology. This will be done from now on by referring to system (11.16).

Proposition 11.3.1. *Let $\mathcal{A}$ be LAS for the system $\dot{z} = f(z,0)$ with domain of attraction $\mathcal{D}(\mathcal{A})$. Let the triplet $(f(z,0), b(z,0)/a(z,0), \mathcal{A})$ be ISS-LER. Then there exist an integer ν, a continuous function $T : \mathcal{A} \to \mathbb{R}^\nu$ and, for any compact set $\mathcal{Z} \subset \mathcal{D}(\mathcal{A})$, $\mathcal{Y} \subset \mathbb{R}^r$ and $\mathcal{N} \subset \mathbb{R}^\nu$, a controller of the form (11.4) such that the set*

$$\text{graph}T \times \{0\} = \{(z,y,\eta) \in \mathcal{A} \times \mathbb{R}^r \times \mathbb{R}^\nu \quad : \quad y = 0\,,\ \eta = T(z)\}$$

is LAS($\mathcal{Z} \times \mathcal{Y} \times \mathcal{N}$).

Proof. We consider the change of variables

$$\begin{aligned} y_i &\mapsto \tilde{y}_i = g^{-(i-1)} y_i\,, \quad i = 1, \ldots, r-1\,, \\ y_r &\mapsto \tilde{y}_r = y_r + g^{r-1} a_0 y_1 + \ldots + g a_{r-2} y_{r-1}\,, \end{aligned}$$

where $g > 1$ is a design parameter and a_i, $i = 0, \ldots, r-2$, are such that all the roots of the polynomial $\lambda^{r-1} + a_{r-2}\lambda^{r-1} + \ldots + a_1\lambda + a_0 = 0$ have negative real

part. Denoting by $\tilde{y} = \mathrm{col}(\tilde{y}_1, \ldots, \tilde{y}_{r-1})$, system (11.16) in the new coordinates reads as

$$\begin{aligned} \dot{z} &= f(z, \bar{C}\tilde{y}) \\ \dot{\tilde{y}} &= g\bar{A}\tilde{y} + \bar{B}\tilde{y}_r \\ \dot{\tilde{y}}_r &= \bar{a}(z, \tilde{y}, \tilde{y}_r)\,[u + \bar{q}(z, \tilde{y}, \tilde{y}_r)] \end{aligned} \tag{11.17}$$

in which $(\bar{A}, \bar{B}, \bar{C})$ is a suitably defined triplet with $\bar{A}$ a Hurwitz matrix, and $\bar{a}(\cdot)$, $\bar{q}(\cdot)$ are smooth functions (dependent on g) such that $\bar{a}(z, \tilde{y}, \tilde{y}_r) = a(z, y)$ and $\bar{q}(z, 0, 0) = b(z, 0)/a(z, 0)$. As $(f(z, 0), \bar{q}(z, 0, 0), \mathcal{A})$ is ISS-LER, there exists a compact set $\mathcal{R} \supseteq \mathcal{A}$ which is LES for $\dot{z} = f(z, 0)$ with $\mathcal{D}(\mathcal{R}) \supseteq \mathcal{D}(\mathcal{A})$. Furthermore, by definition, also the triplet $(f(z, 0), \bar{q}(z, 0, 0), \mathcal{R})$ is ISS-LER. We consider now the zero dynamics, with respect to the input u and output $\tilde{y}_r$, of (11.17) given by

$$\begin{aligned} \dot{z} &= f(z, \bar{C}\tilde{y}) \\ \dot{\tilde{y}} &= g\bar{A}\tilde{y}\,. \end{aligned} \tag{11.18}$$

It can be proved (by means of arguments which, for instance, can be found in [14]), that for any compact set $\tilde{\mathcal{Y}} \in \mathbb{R}^{r-1}$ there exists a $g^\star > 0$ such that for all $g \geq g^\star$ the set $\mathcal{R} \times \{0\}$ is LES($\mathcal{Z} \times \tilde{\mathcal{Y}}$) for (11.18). Fix, once for all, $g \geq g^\star$. By the previous facts, by the fact that the triplet $(f(z, 0), \bar{q}(z, 0, 0), \mathcal{R})$ is ISS-LER and by Proposition 6 of [11], it follows that the triplet $(\mathrm{col}(f(z, \bar{C}\tilde{y}),\, g\bar{A}\tilde{y}),\, \bar{q}(z, \tilde{y}, 0),\, \mathcal{R} \times \{0\})$ is LER. Now fix

$$u = -\kappa\tilde{y}_r + v \tag{11.19}$$

where κ is a positive design parameters and v is a residual control input. From the previous results, it follows that system (11.17) with (11.19) fits in the framework of Theorem 11.2.2, by which it is possible to conclude that there exists a controller of the form

$$\dot{\xi} = \Phi'_k(\xi, \tilde{y}_r) \qquad v = \Upsilon'_k(\xi, \tilde{y}_r) \qquad \xi \in \mathbb{R}^p\,, \tag{11.20}$$

a continuous function $\tau' : \mathcal{R} \times \{0\} \to \mathbb{R}^p$ and, for any compact set $\tilde{\mathcal{Y}}_r \subset \mathbb{R}$ and $\mathcal{N}' \subset \mathbb{R}^p$, a positive constant $\kappa^\star$, such that for all $\kappa \geq \kappa^\star$ the set

$$\mathrm{graph}\tau' \times \{0\} = \{((z, \tilde{y}), \tilde{y}_r, \xi) \in (\mathcal{R} \times \{0\}) \times \mathbb{R} \times \mathbb{R}^p \quad : \quad \tilde{y}_r = 0\,,\ \xi = \tau'(z, \tilde{y})\}$$

is LES($\mathcal{Z} \times \mathcal{Y} \times \tilde{\mathcal{Y}}_r \times \mathcal{N}'$) for (11.17), (11.20) and graph $\tau'|_{\mathcal{A} \times \{0\}} \times \{0\}$ is LAS($\mathcal{Z} \times \mathcal{Y} \times \tilde{\mathcal{Y}}_r \times \mathcal{N}'$).

The previous facts have shown how to solve the problem at hand by means of a *partial state feedback* regulator (namely a regulator processing the measurable output $y_m = y_1$ and its first r time derivatives) of the form

$$\begin{aligned} \dot{\xi} &= \Phi'_k(\xi, \tilde{y}_r) \\ u &= -\kappa\tilde{y}_r + \Upsilon'_k(\xi, \tilde{y}_r)\,. \end{aligned} \tag{11.21}$$

In order to obtain a pure output feedback regulator of the form (11.4), we follow [16] and design a "dirty derivatives" observer-based regulator

$$\begin{aligned}
\dot{\hat{y}}_i &= \hat{y}_{i+1} + K^i \lambda_i (\hat{y}_1 - y_m) \qquad i = 1, \dots r-1 \\
\dot{\hat{y}}_r &= \hat{K}^r \lambda_r (\hat{y}_1 - y_m) \\
\dot{\xi} &= \Phi'_k(\xi, \hat{\tilde{y}}_r) \\
u &= -\kappa \hat{\tilde{y}}_r + \Upsilon'_k(\xi, \hat{\tilde{y}}_r)
\end{aligned} \tag{11.22}$$

where K is a positive design parameters, the λ_i's are such that the polynomial $s^r + \lambda_r s^{r-1} + \ldots + \lambda_2 s + \lambda_1$ is Hurwitz and where

$$\hat{\tilde{y}}_r = \mathrm{sat}_\ell(\hat{y}_r + g^{r-1} a_0 \hat{y}_1 + \ldots + g a_{r-2} \hat{y}_{r-1})$$

in which $\mathrm{sat}_\ell(s)$ is the saturation function such that $\mathrm{sat}_\ell(s) = s$ if $|s| \leq \ell$ and $\mathrm{sat}_\ell(s) = \ell\, \mathrm{sgn}(s)$ otherwise. Letting $y = \mathrm{col}(y_1, \dots, y_r)$, $\hat{y} = \mathrm{col}(\hat{y}_1, \dots, \hat{y}_r)$, and defining the change of variables

$$\hat{y} \;\mapsto\; e := D_K(y - \hat{y})$$

in which $D_K = \mathrm{diag}(K^{r-1}, \dots, K, 1)$, it turns out that the overall closed-loop system (11.17), (11.22) reads as

$$\begin{aligned}
\dot{x} &= \varphi(x) + \Delta_1(x, e) \\
\dot{e} &= KHe + \Delta_2(x, e)
\end{aligned} \tag{11.23}$$

where

$$x := \mathrm{col}(z, \tilde{y}, \tilde{y}_r, \xi)\,,$$

$\dot{x} = \varphi(x)$ is a compact representation of (11.17), (11.21), H is a Hurwitz matrix and $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$ are defined as

$$\Delta_1(\cdot) = \begin{pmatrix} 0 \\ 0 \\ \bar{a}(\cdot)\left[\kappa(\tilde{y}_r - \hat{\tilde{y}}_r) + \Upsilon'_k(\xi, \hat{\tilde{y}}_r) - \Upsilon'_k(\xi, \tilde{y}_r)\right] \\ \Phi'_k(\xi, \hat{\tilde{y}}_r) - \Phi'_k(\xi, \tilde{y}_r) \end{pmatrix}$$

and

$$\Delta_2(\cdot) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \bar{a}(\cdot)\left[\kappa(\tilde{y}_r - \hat{\tilde{y}}_r) + \Upsilon'_k(\xi, \hat{\tilde{y}}_r) - \Upsilon'_k(\xi, \tilde{y}_r)\right] \end{pmatrix}$$

By construction the set $\mathrm{graph}\tau' \times \{0\}$ is LES($\mathcal{X}$) for the system $\dot{x} = \varphi(x)$ with $\mathcal{X} := (\mathcal{X} \times \mathcal{Y}) \times \tilde{\mathcal{Y}}_r \times \mathcal{N}'$ and, by construction, it turns out that for any $\ell > 0$, $\Delta_1(x, 0) \equiv 0$ and $\Delta_2(x, 0) \equiv 0$ for all $x \in \mathrm{graph}\tau' \times \{0\}$. Furthermore, for any compact $\tilde{\mathcal{Y}} \in \mathbb{R}^{r-1}$, $\tilde{\mathcal{Y}}_r \in \mathbb{R}$ and $\hat{\mathcal{Y}} \in \mathbb{R}^r$, there exists a compact set $\mathcal{E} \subset \mathbb{R}^r$ (dependent on K) such that if $\tilde{y}(0) \in \tilde{\mathcal{Y}}$, $\tilde{y}_r \in \tilde{\mathcal{Y}}_r$ and $\hat{y}(0) \in \hat{\mathcal{Y}}$ then $e(0) \in \mathcal{E}$ for all $K > 0$. From these facts, by definition of saturation function and by Theorem 11.2.1, it follows that for any $\mathcal{E} \in \mathbb{R}^r$ there exists a $K^\star > 0$ such that for all $K \geq K^\star$ the set

$$\text{graph}\tau' \times \{0\} \times \{0\} = \{((z,\tilde{y}),\tilde{y}_r,\xi,e) \in (\mathcal{R}\times\{0\})\times\mathbb{R}\times\mathbb{R}^p\times\mathbb{R}^r \ : \\ \tilde{y}_r = 0\,,\ \xi = \tau'(z,\tilde{y})\,,\ e=0\}$$

is LES($\mathcal{Z}\times\mathcal{Y}\times\tilde{\mathcal{Y}}_r\times\mathcal{N}'\times\mathcal{E}$).

From the previous results, the fact (see also Remark 11.2.1) that $\mathcal{A}$ is LAS for the system $\dot{z} = f(z,0)$, and the fact that on $\text{graph}\tau'\times\{0\}\times\{0\}$ the closed-loop dynamics is described by (11.18), the desired result follows by properly adapting the ω-limit set arguments used at the end of the proof of Theorem 2 in [11].

11.4 Examples

11.4.1 Example 1

We consider the system presented in [1] given by

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= -z_1 e^{z_1 z_2} + y \\ \dot{y} &= \mu(z_2 + u) \end{aligned} \tag{11.24}$$

with input $u \in \mathbb{R}$ and measurable output $y \in \mathbb{R}$, in which μ is a constant uncertainty whose value ranges in the closed-set $M = [\underline{\mu},\infty)$ with $\underline{\mu}$ a positive known value. For this system we address the problem of stabilizing the origin by means of a locally Lipschitz output feedback regardless the value of μ. The zero dynamics of the system, described by

$$\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= -z_1 e^{z_1 z_2}\,, \end{aligned} \tag{11.25}$$

have the origin which is globally asymptotically stable but not locally exponentially (the linear approximation at the origin has a pair of purely imaginary eigenvalues). As shown in [1], the static output feedback $u = -ky$, for any positive k fails to stabilize the origin of the closed-loop system, because the resulting linear approximation at the origin is unstable. In particular, for any $r > 0$, there is $k^* > 0$ such that, if $k > k^*$, the closed-loop system obtained with $u = -ky$ has an unstable equilibrium at the origin and a stable limit cycle entirely contained in the sphere of radius r centered at the origin.

Nevertheless, we show how the theory proposed in the paper (and specifically Theorem 11.2.2) can be used to design a *dynamic* output feedback regulator stabilizing the origin. In particular, by letting

$$u = -\kappa y + v\,,$$

it turns out that system (11.24) can be written in the form (11.5) with $x = (\mu, z_1, z_2)^{\mathrm{T}}$, $B = -A = 1$, $a(x,y) = \mu$, $f(x,y) = \text{col}(0, z_2, -z_1 e^{z_1 z_2} + y)$, $q(x,y) = z_2$. In particular the minimum-phase assumption in Section 11.2 is fulfilled with $\mathcal{A} = \mu\times\{0\}\times\{0\}$. In order to apply Theorem 11.2.2, the local exponential reproducibility of the triplet $(f(x,0), q(x,0), \mathcal{A})$ must be checked. To this purpose

we show that Proposition 11.2.1 applies. For, note that $L_{f(x,0)}q(x,0) = -z_1 e^{z_1 z_2}$ and that the transformation

$$(z_1, z_2) \;\rightarrow\; (z_2, -z_1 e^{z_1 z_2}) = (q(x,0), L_{f(x,0)}q(x,0))$$

is a local diffeomorphism at the origin, namely system (11.25) with output z_2 is locally observable at the origin. From the previous fact it turns out that there exists a smooth function $\Upsilon : \mathbb{R}^2 \to \mathbb{R}$ and a compact set $\mathcal{S}$, including $\mathcal{A}$ in its interior, such that

$$L^2_{f(x,0)}q(x,0) = \Upsilon(q(x,0), L_{f(x,0)}q(x,0)) \qquad \forall\, x \in \mathcal{S}\,.$$

Hence, by Proposition 11.2.1, the triplet $(f(x,0), q(x,0), \mathcal{A})$ is ISS-LER (and thus LER) and Theorem 11.2.2 guarantees the existence of the locally Lipschitz output feedback dynamic regulator. In particular, by bearing in mind (11.14), the regulator has the form

$$\begin{aligned} \dot\eta_1 &= \eta_2 + c_0 L\kappa y \\ \dot\eta_2 &= \Upsilon_c(\eta_1, \eta_2) + c_1 L^2 \kappa y \\ v &= \eta_1 \end{aligned}$$

where c_0, c_1 are arbitrary coefficients such that $s^2 + c_1 s + c_0$ is Hurwitz, Υ_c is an arbitrary bounded function such that $\Upsilon_c \circ \tau(z) = \Upsilon(z)$ for all $z \in \mathcal{S}$ with $\tau(z) = (z_2, -z_1 e^{z_1 z_2})^{\mathrm{T}}$, and L and $\kappa(L)$ are positive numbers to be taken sufficiently large according to the desired domain of attraction of the closed-loop system and to the value of $\underline{\mu}$.

11.4.2 Example 2

We consider the set-point control problem in which the output x_1 of the system in $\mathbb{R}^3$

$$\begin{aligned} \dot z &= -z|z| + x_1 \\ \dot x_1 &= x_2 \\ \dot x_2 &= z + u \end{aligned} \tag{11.26}$$

with control input $u \in \mathbb{R}$, is required to track a constant set point w. The problem can be cast as the regulation problem formulated in Section 11.3. In particular, let $y_1 = x_1 - w$ and $y_2 = x_2$ so that system (11.26) reads as

$$\begin{aligned} \dot z &= -z|z| + w + y_1 \\ \dot y_1 &= y_2 \\ \dot y_2 &= z + u \end{aligned}$$

with the trivial dynamics of the set-point w governed by the exosystem $\dot w = 0$. The zero dynamics of the overall system (with respect to the input u and output y_1) are described by

$$\dot w = 0\,, \qquad \dot z = -z|z| + w$$

and elementary arguments can be used to prove that they are ultimately bounded. More specifically, denoting by $W = [\underline{w}, \bar{w}]$ the closed interval containing the value of w, it turns out that the set

$$\mathcal{A} = \{(w, z) \in W \times \mathbb{R} \ : \ z|z| = w\}$$

is asymptotically stable with domain of attraction $\mathcal{D}(\mathcal{A}) = W \times \mathbb{R}$, but not locally exponentially stable. The minimum-phase assumption of Section 11.3 is thus satisfied and, by letting $f(w, z, y_1) = -z|z| + w + y_1$ and $b(w, z, y) = z$, it follows by Proposition 11.3.1 that an output feedback regulator exists if the triplet $(f(x, z, 0), b(x, z, 0), \mathcal{A})$ is ISS-LER. To check if this is the case, we note that

$$L^2_{f(x,z,0)} b(w, z, 0) = -|b(w, z, 0)| L_{f(w,z,y)} b(w, z, y) \qquad \forall\, (w, z) \in \mathbb{R} \times \mathbb{R}$$

which implies, by Proposition 11.2.1, that the triplet in question is ISS-LER, and, by Proposition 11.3.1, that the problem at hand has a solution. In particular, by going through the proof of Proposition 11.3.1, by bearing in mind (11.14) and the specific expression of the triplet (φ, ψ, γ) in Proposition 11.2.1, it turns out that the regulator has the form

$$\begin{aligned} \dot{\xi}_1 &= \xi_2 + c_0 L \kappa \hat{\tilde{y}}_2 \\ \dot{\xi}_2 &= \Upsilon_c(\xi_1, x_2) + c_1 L^2 \kappa \hat{\tilde{y}}_2 \\ u &= -\kappa \hat{\tilde{y}}_2 + \xi_1 \end{aligned}$$

in which Υ_c is any smooth bounded function such that

$$\Upsilon_c(\xi_1, \xi_2) = -|\xi_1|\xi_2 \quad \forall (\xi_1, \xi_2) \ : \ \xi_1 \in \left[\min_{(w,z)\in\mathcal{A}} \{z\} - \epsilon, \max_{(w,z)\in\mathcal{A}} \{z\} + \epsilon \right], \ |\xi_2| \le \epsilon$$

with ϵ a positive number, c_0, c_1 are arbitrary coefficients such that $s^2 + c_1 s + c_0$ is Hurwitz, L is a design parameter, and $\hat{\tilde{y}}_2$ is given by

$$\hat{\tilde{y}}_2 = \mathrm{sat}_\ell(\hat{y}_2 + g\hat{y}_1)$$

where g ia a positive design parameter, sat_ℓ is a saturation function, and $(\hat{y}_1, \hat{y}_2)$ are the state variables of the "dirty derivatives observer"

$$\begin{aligned} \dot{\hat{y}}_1 &= \hat{y}_2 + K\lambda_0(\hat{y}_1 - y_1) \\ \dot{\hat{y}}_2 &= K^2\lambda_1(\hat{y}_1 - y_1) \end{aligned}$$

with c_0, c_1 arbitrary coefficients such that $s^2 + c_1 s + c_0$ is Hurwitz and K a positive design parameter. The high-gain design parameters g, $L(g)$, $\kappa(L)$, and $K(\kappa)$ must be taken sufficiently large according to the desired domain of attraction for the closed-loop system.

Acknowledgement. The authors wish to thank L. Praly for many useful suggestions in the preparation of the manuscript.

References

1. Byrnes, C.I., Isidori, A.: Bifurcation analysis of the zero dynamics and the practical stabilization of nonlinear minimum-phase systems. Asian Journal of Control 4(2), 171–185 (2002)
2. Byrnes, C.I., Isidori, A.: Limit sets, zero dynamics and internal models in the problem of nonlinear output regulation. IEEE Trans. on Automat. Contr. 48, 1712–1723 (2003)
3. Delli Priscoli, F., Marconi, L., Isidori, A.: New approach to adaptive nonlinear regulation. SIAM Journal on Control and Optimization 45, 829–855 (2006)
4. Delli Priscoli, F., Marconi, L., Isidori, A.: Nonlinear observers as nonlinear internal models. Systems & Control Letters 55, 640–649 (2006)
5. Huang, J., Lin, C.F.: On a robust nonlinear multivariable servomechanism problem. IEEE Trans. on Automat. 39, 1510–1513 (1994)
6. Huang, J.: Output regulation of nonlinear systems with nonhyperbolic zerodynamics. IEEE Trans. on Automat. Control 40, 1497–1500 (1995)
7. Huang, J.: Nonlinear Output Regulation Theory and Applications. SIAM Series: Advances in Design and Control. Cambridge University Press, Cambridge (2007)
8. Isidori, A.: Nonlinear Control Systems II, 1st edn. Springer, New York (1999)
9. Jiang, Z.P.: Control of Interconnected Nonlinear Systems: A Small-Gain Viewpoint. LNCIS. Springer, Heidelberg (2004)
10. Jiang, Z.P., Teel, A.R., Praly, L.: Small-gain theorem, gain assignment and applications. In: Proc. of the 33rd IEEE Conference on Decision and Control (1994)
11. Marconi, L., Praly, L., Isidori, A.: Robust asymptotic stabilization of nonlinear systems with non-hyperbolic zero dynamics. Accepted for publication on IEEE Trans. on Automat. Contr. (March 2009)
12. Marconi, L., Praly, L., Isidori, A.: Robust asymptotic stabilization of nonlinear systems with non-hyperbolic zero dynamics: Part II. In: 47th IEEE Conference on Decision and Control, Cancun, Mexico (2008)
13. Marconi, L., Praly, L.: Essential and redundant internal models in nonlinear output regulation. In: Astolfi, A., Marconi, L. (eds.) Analysis and Design of Nonlinear Control Systems. Springer, Berlin (2007)
14. Marconi, L., Praly, L., Isidori, A.: Output stabilization via nonlinear Luenberger observers. SIAM Journal on Control and Optimization 45, 2277–2298 (2007)
15. Marconi, L., Praly, L.: Uniform practical nonlinear output regulation. IEEE Trans. on Automat. Contr. 53(5), 1184–1202 (2008)
16. Teel, A.R., Praly, L.: Tools for semiglobal stabilization by partial state and output feedback. SIAM Journal on Control and Optimization 33, 1443–1485 (1995)

12

On the Observability of Nonlinear and Switched Systems

Wei Kang[1,⋆], Jean-Pierre Barbot[2], and Liang Xu[3]

[1] Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA, USA
[2] ECS-EA3649 , ENSEA, 6 Avenue du Ponceau, 95014 Cergy-Pontoise and Equipe Projet ALIEN, INRIA, France
[3] Naval Research Laboratory, Monterey, CA, USA

Summary. In this paper, new concept of observability are introduced for both nonlinear systems and switched systems. The new definitions are applicable to a much broader family of problems of estimation including unmeasured state variables, unknown input, and unknown parameters in control systems. It is also taken into account the notion of partial observability which is useful for complex or networked systems. For switched systems, the relationship between the observability and hybrid time trajectories is analyzed. It is proved that a switched system might be observable even when individual subsystems are not. Another topic addressed in this paper is the measure of observability, which is able to quantitatively define the robustness and the precision of observability. It is shown that a system can be perfectly observable in the traditional sense, but in the case of high dimensions, it is practically unobservable (or extremely weekly observable). Moreover, computational algorithm for nonlinear systems is developed to compute the observability with precision. Several examples are given to illustrate the fundamentals and the usefulness of the results.

12.1 Introduction

It is well known that the definition of linear observability is universal for all linear systems, but there exist many different definitions of observability in the literature of nonlinear control systems. While impossible to exhaust all previous results, some well know definitions include weakly local observability [15], algebraic observability [10], infinitesimal observability [12], unboundedness observability [1], ... It is a technical question at the crossway of several factors such as the generality of the concept, easy to check, and practical observer design.

In this paper, the concept of observability for nonlinear systems and switched systems is investigated with new definitions in an extended context. Relative to the classical concept of observability, the proposed definitions are applicable to a much broader family of problems of estimation including unmeasured state variables, unknown input, and unknown parameters in control systems. Moreover, the proposed definition takes into account the notion of partial observability

⋆ This author is funded in part by Naval Research Laboratory.

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 199–216.
springerlink.com

which has potential applications in the context of complex or networked systems, for which achieving complete observability of the entire system is difficult, if not impossible. Another new concept introduced in this paper is the measure of observability, which is able to tell the robustness and the precision of observability. An interesting example shows that a system can be perfectly observable in the traditional sense, but in the case of high dimensions, it is practically unobservable (or extremely weekly observable) when the observability is measured by the new definition. A definition is not useful if it cannot be computationally implemented. For this reason, a computational algorithm is developed to compute the observability with precision. This algorithm is applicable to both linear and nonlinear systems.

In this paper, the notation Z represents a variable or function associated with a control system. The goal of the research is to develop a framework of observability that takes into the consideration of several factors that are not considered all together in existing definitions:

- Partial observability (observability of a part of the system variables, not all the state variables);
- Measure of observability that quantitatively determines the degree of observability, from strongly observable to weekly observable;
- Computational algorithms for both linear and nonlinear systems.

In the second part of the paper, it is introduced the notion of $Z(T_N)$-observability for switched systems. This new definition takes into consideration of four factors not considered all together in classical definitions:

- Partial state observability
- Observability with partial model of a dynamical system
- Systems with algebraic constraints
- Partial time observability

The concepts in this paper are applicable to several other related problems, including left invertibility as defined in [23], and the observability of unknown inputs and/or parameters. However, it is not specifically addressed due to the limit of space. One of the main advantages of this approach is to relax the assumptions that require all individual subsystems be observable (see for example [5, 6]). The efficiency of these definitions are highlighted with three examples, including one networked system and two circuit systems.

12.2 Z-Observability and Its Measure

Consider a general nonlinear control system

$$\begin{aligned} \dot{\xi} &= f(t, \xi, u),\ \xi \in \mathbb{R}^n,\ u \in \mathbb{R}^m \\ y &= h(t, \xi, u) \end{aligned} \tag{12.1}$$

Suppose $z = Z(t, \xi, u)$ is a variable to be estimated. In the following, U represents an open and connected set in the time-state-control space $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m$. For the

input in any time interval $[t_0, t_1]$, we assume that $u(t)$ is bounded and C^∞ at all but finite many points in $[t_0, t_1]$. At each discontinuous point $t_i \in [t_0, t_1]$, the following limits exist

$$\lim_{t \to t_i^+} u(t), \ \lim_{t \to t_i^-} u(t)$$

In a trajectory, $\xi(t)$ is absolutely continuous. If $(\xi(t), u(t))$ satisfies the differential equation in (15.1) for all t in $[t_0, t_1]$ except for finite many points, then $(t, \xi(t), u(t))$ is called a trajectory. In this note, equations involving $u(t)$ always mean " equal almost everywhere," and the notation is "a.e. in $[t_0, t_1]$."

Definition 12.2.1. *The function $z = Z(t, \xi, u)$ is said to be Z-observable in U with respect to the system (15.1) if for any two trajectories, $(t, \xi^i(t), u^i(t))$, $i = 1, 2$, in U defined on a same interval $[t_0, t_1]$, the equality*

$$h(\xi^1(t), u^1(t)) = h(\xi^2(t), u^2(t)), \quad \textit{a.e. in } [t_0, t_1]$$

implies

$$Z(t, \xi^1(t), u^1(t)) = Z(t, \xi^2(t), u^2(t)), \ \textit{a.e. in } [t_0, t_1]$$

Now, suppose for any trajectory $(t, \xi(t), u(t))$ in U, there always exists an open set $U_1 \subset U$ so that $Z(t, \xi, u)$ is Z-observable in U_1. Then, $z = Z(t, \xi, u)$ is said to be locally Z-observable in U.

In linear control theory, the control input is assumed to be a known variable. In our notation, u is included in y as a variable. The classical observability of linear control systems is equivalent to the Z-observability of all state variables in ξ. For nonlinear uniformly observable systems, all the state variables are Z-observable under Definition 12.2.1 with $Z(t, \xi, u) = x$. However, a system could be unobservable under classical definitions of observability, while part of the state variables is Z-observable. This partial observability is especially important for large complex systems with subsystems, in which the observability of the entire system is either impossible or unnecessary. For example, consider a simple cascaded system

$$\begin{aligned} &\dot{\xi}_1 = f_1(\xi_1, \xi_2, \xi_3), \ \dot{\xi}_2 = f_2(\xi_1, \xi_2, \xi_3), \\ &\dot{\xi}_3 = \xi_4, \qquad\qquad \dot{\xi}_4 = -\xi_4 - \xi_5 + \sin(t)u, \\ &Y = \begin{bmatrix} \xi_3 \\ u \end{bmatrix} \end{aligned}$$

The overall system is not observable. However, the variable

$$Z = \begin{bmatrix} \xi_3 \\ \xi_4 \end{bmatrix}$$

is Z-observable. It is also important to notice that, in Definition 12.2.1, the input u is a variable treated equally as the state variable x in the definition. So, it automatically handles both the problems of left invertibility and parameter identification.

The local Z-observability is defined differently from some typical definitions of local properties. In fact, the meaning of local in Definition 12.2.1 is not around a single point in the state space. Instead, the local neighborhood U_1 is selected to be around a trajectory. A local observer around a single point would be too restrictive for many applications. On the other hand, it is possible to design nonlinear observers applicable to non-local trajectories, but with local initial estimation (the observer converges if the initial estimation error is small), like the observer in [17].

Some sufficient conditions for Z-observability were proved in [16]. Basically, if $z = Z(\xi, u)$ is Z-observable, then a variable is observable if it equals a function of z and its derivatives in the direction along the trajectory of the control system. For example,

$$\begin{aligned} \dot{\xi}_1 &= -\xi_1 - \xi_2 + \xi_1^2 \\ \dot{\xi}_2 &= -\xi_2 + u, \\ \dot{\xi}_3 &= \xi_4 \\ \dot{\xi}_4 &= u \end{aligned} \tag{12.2}$$

Let

$$y = \xi_1$$

be the measured output. In this case, $\xi_2 = \dot{y} + y - y^2$ is Z-observable; $u = \dot{\xi}_2 + \xi_2$ is Z-observable. However, ξ_3 and ξ_4 are not observable.

While Definition 12.2.1 is straightforward, checking the observability can be cumbersome in its algebraic derivations, especially for complex nonlinear systems. More importantly, it does not provide information about the robustness of observability. For instance, consider the system (12.2). If the measured variable is

$$y = \begin{bmatrix} \xi_1 \\ u \end{bmatrix}$$

then ξ_2 is more robustly observable relatively to the case of $y = \xi_1$. With the information of u, we expect more accurate estimation in the presence of noise. The problem is: how to quantitatively measure the robustness of observability? In the following, we introduce another definition to take into consideration of robustness and computational accuracy of observability. In this definition, we assume that variables along trajectories are associated with metrics. For instance, $h(t, \xi(t), u(t))$, as a function of t, can be measured by L_2 or L_∞ norm; $Z(t, \xi(t), u(t))$ can be measured by its function norm, or by the norm of its initial value $Z(t_0, \xi(t_0), u(t_0))$. A metric used for a specific variable, for instance z, is denoted by $||\cdot||_Z$.

Definition 12.2.2. *Given positive numbers $\epsilon > 0$ and $\delta > 0$. The variable $z = Z(t, \xi, u)$ is said to be observable in U with precision (ϵ, δ) if for any two trajectories, $(t, \xi^i(t), u^i(t))$, $i = 1, 2$, in U defined on a same interval $[t_0, t_1]$, the inequality*

$$||h(\xi^1(t), u^1(t)) - h(\xi^2(t), u^2(t))||_Y \leq \epsilon$$

implies

$$||Z(t, \xi^1(t), u^1(t)) - Z(t, \xi^2(t), u^2(t))||_Z \leq \delta.$$

Basically, this definition implies that, if the error of y is bounded by ϵ, the goal of the estimation error for z should be within the limit of δ. Or equivalently, if a system is NOT observable with precision (ϵ, δ), then there always exists a trajectory for which, even though the error of output is bounded by ϵ, the estimation error is larger than δ. In this case, there is no estimator can guarantee an error smaller than δ if no further information is provided. The error tolerance in this definition represents a fundamental limitation on the observability of a system configuration.

In the following, we give a comparison of Definition 12.2.2 to traditional definitions of observability. It is shown that a well designed Z-observable system may not be observable with a precision. It is our opinion that traditional definitions, as well as Definition 12.2.1 in this paper, are not adequate for practical applications because it does not provide a measure on the limitation of estimation robustness and accuracy. Before we can carry out a comparison, it is necessary to develop methods for the determination of the observability in the sense of Definition 12.2.2. Rather than cumbersome algebraic derivation, in the following we introduce a computational approach for approximate observability so that handling general nonlinear systems becomes possible.

The computation of precision can be formulated as a problem of dynamical optimization in the following way.

Problem 12.2.1. (Calculation of δ for a given ϵ)
Given a positive number $\epsilon > 0$ and a nominal trajectory $(\xi_0(t), u_0(t))$, $t \in [t_0, t_1]$. Define

$$J(\xi(\cdot), u(\cdot)) = ||Z(\xi(\cdot), u(\cdot)) - Z(\xi_0(\cdot), u_0(\cdot))||_Z$$

Then, the minimum observability error tolerance, $\bar{\delta}$, associated with the nominal trajectory is the solution of the following dynamical optimization

$$\max_{(\xi,u)} J \tag{12.3}$$

subject to

$$\begin{array}{ll} (\xi(t), u(t)) \in U, & t \in [t_0, t_1] \\ \dot{\xi} = f(t, \xi, u), & \xi \in \mathbb{R}^n, \quad u \in \mathbb{R}^m \\ ||h(t, \xi(t), u(t)) - h(t, \xi_0(t), u_0(t))||_Y \leq \epsilon, & t \in [t_0, t_1] \end{array} \tag{12.4}$$

A variable is observable with a precision (ϵ, δ) if $\bar{\delta}$ from Problem 12.2.1 is less than or equal to δ for all nominal trajectories in U. However, if one of the $\bar{\delta}$ is greater than δ, then the variable is not observable in U with the given precision.

The formulation in Problem 12.2.1 is useful only if it can be solved. Obviously, an analytic solution of the dynamical optimization is very difficult to find, if not impossible, especially in the case of nonlinear systems. However, there exist numerical approaches that can be used to find its approximate solution. The numerical approach adopted in this paper is the Pseudospectral (PS) optimal control method [11], [14]. In this approach, Problem 12.2.1 is transformed through discretization into a finite dimensional nonlinear programming, which can be

numerically solved. In a PS method, a function such like $(\xi_1(t), \ldots, \xi_n(t))^T$ is approximated by its value at a set of nodes $t_0 = -1 < t_1 < \cdots < t_N = 1$. From approximation theory, specially designed nodes are able to improve the accuracy. For instance, Legendre-Gauss-Lobatto (LGL) node points are adopted in this paper. They are defined by $t_0 = -1$, $t_N = 1$, and the critical points of Legendre polynomials [8]. These nodes form a partition of $[-1, 1]$. For an arbitrary interval $[t_0, t_1]$, the LGL nodes can be easily mapped onto it using a linear function. In the discretization, the state variables are approximated by the vectors $\bar{\xi}^{Nk} \in \mathbb{R}^n$, i.e.

$$\bar{\xi}^{Nk} = \begin{bmatrix} \bar{\xi}_1^{Nk} \\ \bar{\xi}_2^{Nk} \\ \vdots \\ \bar{\xi}_n^{Nk} \end{bmatrix}$$

is an approximation of $\xi(t_k)$. Similarly, $\bar{u}^{Nk}$ is the approximation of $u(t_k)$. Thus, a discrete approximation of the function $\xi_i(t)$ is the vector

$$\bar{\xi}_i^N = \begin{bmatrix} \bar{\xi}_i^{N1} & \bar{\xi}_i^{N2} & \cdots & \bar{\xi}_i^{NN} \end{bmatrix}$$

A continuous approximation is defined by its polynomial interpolation, denoted by $\xi_i^N(t)$, i.e.

$$\xi_i(t) \approx \xi_i^N(t) = \textstyle\sum_{k=0}^{N} \bar{\xi}_i^{Nk} \phi_k(t),$$

where $\phi_k(t)$ is the Lagrange interpolating polynomial [8]. In this paper, the discrete variables are denoted by letters with an upper bar, such as $\bar{\xi}_i^{Nk}$ and $\bar{u}^{Nk}$. If k in the superscript and/or i in the subscript are missing, it represents the corresponding vector or matrix in which the indices run from minimum to maximum. For example,

$$\begin{aligned} \bar{\xi}_i^N &= \begin{bmatrix} \bar{\xi}_i^{N0} & \bar{\xi}_i^{N1} & \cdots & \bar{\xi}_i^{NN} \end{bmatrix} \\ \bar{\xi}^{Nk} &= \begin{bmatrix} \bar{\xi}_1^{Nk} \\ \bar{\xi}_2^{Nk} \\ \vdots \\ \bar{\xi}_r^{Nk} \end{bmatrix} \\ \bar{\xi}^N &= \begin{bmatrix} \bar{\xi}_1^{N0} & \bar{\xi}_1^{N1} & \cdots & \bar{\xi}_1^{NN} \\ \bar{\xi}_2^{N0} & \bar{\xi}_2^{N1} & \cdots & \bar{\xi}_2^{NN} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{\xi}_r^{N0} & \bar{\xi}_r^{N1} & \cdots & \bar{\xi}_r^{NN} \end{bmatrix} \end{aligned} \tag{12.5}$$

Similarly,

$$\bar{u}^N = \begin{bmatrix} \bar{u}^{N0} & \bar{u}^{N1} & \cdots & \bar{u}^{NN} \end{bmatrix}$$

For differentiation, the derivative of $\xi_i^N(t)$ at a LGL node t_k is easily computed by the following matrix multiplication [8]

$$\begin{bmatrix} \dot{\xi}_i^N(t_0) & \dot{\xi}_i^N(t_1) & \cdots & \dot{\xi}_i^N(t_N) \end{bmatrix}^T = D(\bar{\xi}_i^N)^T$$

where the $(N+1)\times(N+1)$ differentiation matrix D can be computed off-line independently from the function to be approximated. If the cost function includes integration, such as L_2 norm, it can be approximated by the Gauss-Lobatto integration rule,

$$\int_{-1}^{1} C(\xi(t), u(t))dt \approx \sum_{k=0}^{N} C(\bar{\xi}^{Nk}, \bar{u}^{Nk})w_k$$

where w_k are the LGL weights [8]. Now, we are ready to define the discretization to Problem 12.2.1. A discretization depends on the metric used in the problem definition. As an example, we assume $||\cdot||_Z$ is the norm of its initial value, i.e. $||Z((t_0), \xi(t_0), u(t_0))||_2$, and the norm $||\cdot||_Y$ is $\max_{t\in[t_0,t_1]}\{||y(t)||_2\}$.

Problem 12.2.1$^{\mathbf{N}}$: Find $\bar{\xi}^{Nk} \in \mathbb{R}^n$ and $\bar{u}^{Nk} \in \mathbb{R}^m$, $k = 0, 1, \ldots, N$, that maximizes

$$\bar{J}^N(\bar{\xi}^N, \bar{u}^N) = ||Z(\bar{\xi}^{N0}, \bar{u}^{N0})||_2$$

subject to

$$\begin{aligned}
&(t_k, \xi^{Nk}, u^{Nk}) \in U, \qquad k = 0, \cdots, N\\
&\begin{cases}
\bar{\xi}_1^N D^T = f_1(t^N, \bar{\xi}^N, u^N)\\
\bar{\xi}_2^N D^T = f_2(t^N, \bar{\xi}^N, u^N)\\
\qquad\vdots\\
\bar{\xi}_n^N D^T = f_n(t^N, \bar{\xi}^N, u^N)
\end{cases}\\
&||h((t_k, \xi^{Nk}, u^{Nk}) - y(t_k)||_2 \le \epsilon
\end{aligned}$$

In this formulation, the notation $f_i(t^N, \bar{\xi}^N, u^N)$ is slightly abused. Its value is a row vector

$$f_i(t^N, \bar{\xi}^N, u^N) = \left[\, f_i(t^{N0}, \bar{\xi}^{N0}, u^{N0}) \cdots,\ f_i(t^{NN}, \bar{\xi}^{NN}, u^{NN}) \,\right]$$

Problem 12.2.1$^{\mathbf{N}}$ is a nonlinear programming with constraints. It can be solved using computational algorithms, such as sequential quadratic programming. Some commercially available software packages can be found to handle these problems, such as SNOPT.

In the following example, we illustrate that the traditional concept of observability is ineffective for systems with large dimensions. It justifies the necessity to quantitatively measure the robustness of observability, as defined in Definition 12.2.2. Consider the following linear system

$$\begin{aligned}
&\dot{\xi}_1 = x_2\\
&\dot{x}_2 = x_3\\
&\vdots\\
&\dot{x}_n = -\sum_{i=1}^{n} \binom{n}{i-1} x_i\\
&y = x_1
\end{aligned} \tag{12.6}$$

Under a traditional definition of observability, this system is perfectly observable for any choice of n. However, if a precision is applied to the observability, it is a completely different story when the dimension is large. Let us assume that the true initial state is

$$x_0 = \begin{bmatrix} 0 & 0 & \cdots & 1 \end{bmatrix}^T$$

The goal is to estimate x_0. The precision for the output is $\epsilon = 10^{-6}$. The time interval is $[0, 15]$. Problem $12.2.1^N$ is solved to compute the error tolerance, δ, for the estimation of x_0. Table 12.1 lists the result for $n = 2, 3, \cdots, 9$.

Table 12.1. Error Tolerance

n	2	3	4	5	6	7	8	9
δ	4.70×10^{-6}	2.67×10^{-5}	1.53×10^{-4}	8.89×10^{-4}	5.20×10^{-3}	3.01×10^{-2}	1.75×10^{-1}	1.02
ϵ	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}

The value of δ listed in this table is the tight lower bound for observability, i.e. the system is not observable for any δ less than the one listed in the table; and this particular initial state is observable with precision (ϵ, δ) if δ is bigger than the one in the table (computational error may require a δ slightly bigger). From the table, the estimation of x_0 can be as accurate as 4.70×10^{-6} if the dimensional is 2. It agrees with the traditional theory of observability. However, when the dimension is increased, the estimation becomes less accurate. At $n = 8$, the estimation error can be as big as 0.175, or 17.5% relative to the true x_0. When $n = 9$, the estimation error is 1.02. In this case, the relative error is more than 100%! Thus, the system is practically unobservable, although it is perfectly observable under a tradidtional definition! In Figure 12.1, the dotted curve represents a trajectory of the system ($n = 9$); the continous curve is the true trajectory. The outputs of both trajectories agree to each other (Figure on top), but the initial states are significantly different.

A definition is useful only if it is numerically implementable. The solvability of Problem $12.2.1^N$ extends the applicability of Definition 12.2.2 to general nonlinear systems. For example, consider the following nonlinear system

$$\begin{aligned}
\dot{\xi}_1 &= -\xi_1 \frac{\xi_2}{2\Delta r} + \kappa \frac{\xi_2 - 2\xi_1}{\Delta r^2} \\
\dot{\xi}_2 &= -\xi_2 \frac{\xi_3 - \xi_1}{2\Delta r} + \kappa \frac{\xi_1 + \xi_3 - 2\xi_2}{\Delta r^2} \\
&\vdots \\
\dot{\xi}_k &= -\xi_k \frac{\xi_{k+1} - \xi_{k-1}}{2\Delta r} + \kappa \frac{\xi_{k-1} + \xi_{k+1} - 2\xi_k}{\Delta r^2} \\
&\vdots \\
\dot{\xi}_n &= -\xi_n \frac{-\xi_{n-1}}{2\Delta r} + \kappa \frac{\xi_{n-1} - 2\xi_n}{\Delta r^2} \\
y &= \xi_{i_0}
\end{aligned} \tag{12.7}$$

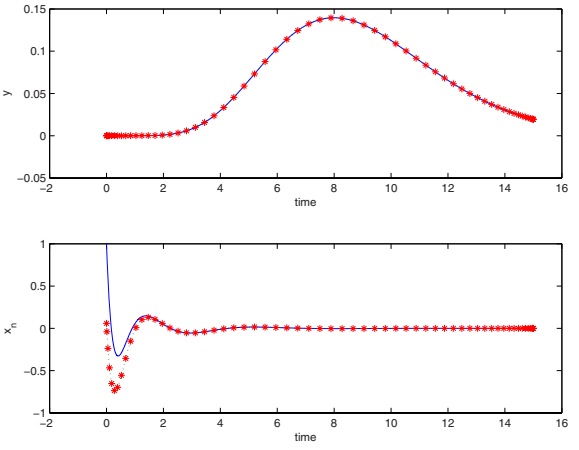


Fig. 12.1. Estimation error ($n = 9$)

This is, in fact, the discretization of damped Burgers equation using finite difference. In the simulation, we assume $n = 9$, $\kappa = 0.14$, $\Delta r = \dfrac{2\pi}{n+1}$. The observation is $y = \xi_6$. Checking the observability for this system using algebraic tools, such as Lie bracket, is not easy, if not impossible. However, for the observability with a precision, we can solve Problem $12.2.1^N$. As an example, we checked the observability of the following initial state

$$\xi_0 = \begin{bmatrix} 0 & 0 & 5 & 0 & \cdots & 0 \end{bmatrix}^T$$

Assume that the precision of output is 10^{-10}, which is a high accuracy requirement. However, it is proved that the initial state is not approximately observable with a precision $\delta = 0.01$. In other words, it is possible that another initial state with a distance to ξ_0 greater than or equal to 0.01 so that its output is within the 10^{-10} range of that of ξ_0.

12.3 Observability for Switched Systems

Switched systems, or hybrid systems in a general set-up, can be found in many industrial applications, such as MEMS [19], Power Electronics [20], Robotic [7],... In this section, we deal with new observation concepts and some related results for switched systems without jump. Let us consider the following class of systems:

$$\begin{aligned} \dot{\xi} &= f_q(t, \xi, u), \quad q \in Q, \, \xi \in \mathbb{R}^n, \, u \in \mathbb{R}^m \\ y &= h_q(t, \xi, u) \end{aligned} \tag{12.8}$$

where Q is a finite index set, $f_q \, \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}^m$ is sufficiently smooth, all dwell time intervals, $[t_{i,0}, t_{i,1}]$, between two switchings of the structure (i.e. change of q) satisfy $(t_{i,1} - t_{i,0}) > \tau_{min}$ for some $\tau_{min} > 0$ (this assumption excludes Zeno

phenomena [2]). For the input u in any time interval $[t_{i,0}, t_{i,1}[\subseteq [t_{ini}, t_{end}[$, we assume that $u(t)$ is bounded and C^∞.

The system (12.8) assumes no information and condition on the switching sequence. If this sequence is determined by a known state function without discrete state memory, the following representation is applicable.

$$\begin{array}{l} \dot{\xi} = f_q(t, \xi, u),\ q \in Q,\ \xi \in \mathbb{R}^n,\ u \in \mathbb{R}^m \\ y = h_q(t, \xi, u),\ q = \sigma(\xi) \end{array} \tag{12.9}$$

If all subsystems in (12.9) are continuously state observable, then there exist algebraic and geometrical conditions as well as observer design methods in the literature [5, 4, 6, 9] for the recover of the discrete variable q or [22, 24] for the resolution of left invertibility. In this section, the assumption on the full state observability of individual subsystems is removed.

For switched systems, the concept of observability and methods of observer design are strongly related to the dwell time and the sequence of switching, thus it is important to recall (in our context) the following definition of hybrid time trajectory [18] (see also [13]).

Definition 12.3.1. *A hybrid time trajectory is a finite or infinite sequence of intervals $T_N = \{I_i\}_{i=0}^N$, such that*

- *$I_i = [t_{i,0}, t_{i,1}[$, for all $0 \leq i < N$;*
- *For all $i < N$ $t_{i,1} = t_{i+1,0}$*
- *$t_{0,0} = t_{ini}$ and $t_{N,1} = t_{end}$*

Moreover, we define $\langle T_N \rangle$ as the ordered list of q associated to T_N (i.e. $\{q_0, ..., q_N\}$ with q_i the value of q during the time interval I_I).

In this paper, all hybrid time trajectories T_N and $< T_N >$ are assumed to satisfy $\tau_{min} > 0$. State trajectories that do not admit such a time trajectory are not considered. Now, we are ready to define a new concept of observability for switched systems:

Definition 12.3.2. *Consider a system (12.8) and a variable $z = Z(t, \xi, u)$. Let $(t, \xi^1(t), u^1(t))$ be a trajectory in U with a hybrid time trajectory T_N and $\langle T_N \rangle$. Suppose for any trajectory, $(t, \xi^2(t), u^2(t))$, in U with the same T_N and $\langle T_N \rangle$, the equality*

$$h(t, \xi^1(t), u^1(t)) = h(t, \xi^2(t), u^2(t)), \quad \text{a.e. in } [t_{ini}, t_{end}]$$

implies

$$Z(t, \xi^1(t), u^1(t)) = Z(t, \xi^2(t), u^2(t)), \ \text{a.e. in } [t_{ini}, t_{end}]$$

Then we say that $z = Z(t, \xi, u)$ is $Z(T_N)$-observable along the trajectory $(t, \xi^1(t), u^1(t))$.

For a fixed hybrid time trajectory T_N and $\langle T_N \rangle$, if $z = Z(t, \xi, u)$ is $Z(T_N)$-observable along all trajectories in U, then, $z = Z(t, \xi, u)$ is said to be $Z(T_N)$-observable in U.

Suppose for any trajectory $(t, \xi(t), u(t))$ in U, there always exists an open set $U_1 \subset U$ so that $(t, \xi(t), u(t))$ is contained in U_1 and $Z(t, x, u)$ is $Z(T_N)$-observable in U_1. Then, $z = Z(t, x, u)$ is said to be locally $Z(T_N)$-observable

The previous definition deals with the case of "synchronous switched system" because the definition of $Z(T_N)$-observability is based upon a fixed hybrid time trajectory for all initial states and input functions. In order to remove this restriction, we propose hereafter the concept of $Z(T)$-observability.

Definition 12.3.3. *The variable $z = Z(t, \xi, u)$ of system (12.8) or (12.9) is said to be $Z(T)$-observable in U with respect to a fixed time interval $T = [t_{ini}, t_{end}]$ if for any two trajectories, $(t, \xi^i(t), u^i(t))$, $i = 1, 2$, in U defined on the interval, the equality*

$$h(t, \xi^1(t), u^1(t)) = h(t, \xi^2(t), u^2(t)), \quad a.e.\ in\ [t_{ini}, t_{end}] \tag{12.10}$$

implies

$$Z(t, \xi^1(t), u^1(t)) = Z(t, \xi^2(t), u^2(t)),\ \ a.e.\ in\ [t_{ini}, t_{end}] \tag{12.11}$$

Suppose for any trajectory $(t, \xi(t), u(t))$ in U, there always exists an open set $U_1 \subset U$ so that $(t, \xi(t), u(t))$ is contained in U_1 and $Z(t, \xi, u)$ is $Z(T)$-observable in U_1. Then, $z = Z(t, \xi, u)$ is said to be locally $Z(T)$-observable in U.

A straightforward application of previous definitions implies the following lemma.

Lemma 12.3.1. *Consider the system (12.8) (or respectively (12.9)) and a time interval T. Suppose for every two trajectories satisfying (12.10), there exists a common hybrid time trajectory T_N and $< T_N >$ shared by both trajectories; and $Z(t, \xi, u)$ is $Z(T_N)$-observable along these trajectories. Then the system is $Z(T)$-observable.*

The following lemma implies that the difference of the subdynamics in a switched system provide extra information for observability. In fact, a variable can be observable along a time trajectory while it is unobservable on each individual subsystem. In the following, the dimension of z variable is denoted by n_z. A linear projection P is defined by

$$P : \begin{bmatrix} z_1 \\ \vdots \\ z_{n_z} \end{bmatrix} \rightarrow \begin{bmatrix} \delta_1 & 0 & 0 \cdots & 0 \\ 0 & \delta_2 & 0 \cdots & 0 \\ . & . & . & . \\ 0 & 0 & 0 \cdots & \delta_{n_z} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_{n_z} \end{bmatrix}$$

where δ_i, $i = 1, 2, \cdots, n_z$, is zero or one. The complement of P is called $\bar{P}$ (projecting z to the variables eliminated by P).

Lemma 12.3.2. *Consider the system (12.8) and a fixed hybrid time trajectory T_N and $\langle T_N \rangle$. Let U be an open set in time-state-control space. Suppose*

$Z(t,\xi(t),u(t)) \in \mathbb{R}^{N_z}$ is always continuous under any admissible control input. Suppose there exists a sequence of projections P_i, $i = 0, 1, \cdots, N$, such that

(1) given any $0 \leq i \leq N$, $P_i Z(t,\xi,u)$ is Z-observable in U on the subinterval $t \in [t_{i,0}, t_{i,1}[$;

(2) Rank $\left[P_0^T \ \cdots \ P_N^T \right] = dim(Z) = n_z$;

(3) $\dfrac{d\bar{P}_i Z(t,\xi(t),u(t))}{dt} = 0$ for $t \in [t_{i,0}, t_{i,1}[$ and $(t,\xi(t),u(t)) \in U$.
Then, $z = Z(t,\xi,u)$ is $Z(T_N)$-observable in U with respect to the hybrid time trajectory T_N and $\langle T_N \rangle$.

Proof. From the assumptions, within each I_i the unobservable parts of the function Z are constants; and they are observable in some other time intervals. More specifically, $P_i Z(t,\xi(t),u(t))$ is uniquely determined in the time interval I_i. Each variable in the unobservable part, $\bar{P}_i Z(t,\xi(t),u(t))$, is a constant until at some subinterval $[t_{j0}, t_{j1}]$ where it becomes observable. Because of the continuity, this unobservable constant in I_i can be recovered by the boundary value of the corresponding variable in the interval I_j. As a result, all components of z are uniquely determined, i.e. it is $Z(T_N)$-observable.

Remark 12.3.1. In Lemma 12.3.2, the assumption on the continuity of $Z(t,\xi,u)$ excludes switched systems with state jump if $Z = \xi$.

Remark 12.3.2. For a synchronous hybrid system, the hybrid time trajectory T_N and $\langle T_N \rangle$ has their influences on the observability property in the way similar to an input. Therefore, like the definition of universal input (see [12]), it is possible to define universal hybrid time trajectories as the those which preserve the $Z(T_N)$-observability when $Z = \xi$.

Lemma 12.3.3. *Consider (12.9). Let U be an open set in time-state-control space. Suppose $Z(t,\xi(t),u(t))$ is continuous under any admissible control input. Suppose for every two trajectories satisfying (12.10), there exists a hybrid time trajectory T_N and $< T_N >$ shared by both trajectories and two projections P and Q so that along both trajectories*

1. *$Rank[P^T Q^T] = dim(Z) = n_z$;*
2. *$PZ(t,x,u)$ is $Z(T_N)$-observable;*
3. *at each transient instant t_{i0} the value of $QZ(t_{i0})$ is available;*
4. *Define $\vartheta := QZ(t,\xi(t),u(t))$. For each $[t_{i,0}, t_{i,1}]$, ϑ satisfies a known differential equation of the form $\dot{\vartheta} = g(q,\vartheta,t,u,y,\dot{y},...)$.*

Then, $Z(t,\xi,u)$ of the system (12.9) is $Z(T)$-observable in U.

Proof. It is assumed that $PZ(t,\xi(t),u(t))$ is $Z(T_N)$-observable. From assumption (3) and (4), $QZ(t,\xi(t),u(t))$ is uniquely determined by its initial value at t_i. Because of (1), the variable $Z(t,\xi,u)$ is $Z(T_N)$-observable along the trajectories. Because this is true for all trajectories, Lemma 12.3.1 implies that $Z(t,\xi,u)$ is $Z(T)$-observable.

Remark 12.3.3. In Lemma 12.3.3, we assume more information than the measurement of y. In the assumption (3), the value at initial time of subintervals is a piece of information in addition to the value of y. So, the claimed Z(T)-observability in this lemma slightly abused the definition. Nevertheless, the information in (3) is available in many switched systems with a flat switching manifold. This point is exemplified in the next section.

12.4 Examples

12.4.1 Network Systems: *Z*-Observability

This section is dedicated to a networked system, more specifically to the influence of the small world effect with respect to the Z-observability. In [21], D.S. Watts and S.H. Strogatz gave an example of small-world effect for a network with twenty vertices (see Figure 12.2).

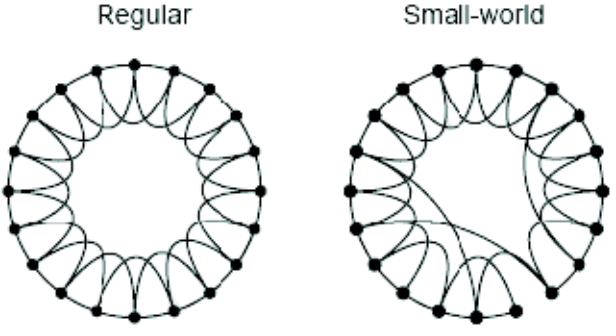


Fig. 12.2. Networkes of Watts and Strogatz

Hereafter, it is considered the same network where each vertex has the following dynamics:

$$\begin{array}{l} \dot{\xi}_{i,1} = \xi_{i,2} + \sum_{j\neq i,j=1}^{20} a_{i,j}\xi_{j,1} \\ \dot{\xi}_{i,2} = -\xi_{i,1} + m_i \end{array} \tag{12.12}$$

where $\xi_{i,1}$ and $\xi_{i,2}$ are the two state variables of the i^{th} vertex. The dynamic at each vertex is similar. It is an oscillator coupled with four other ones. Moreover, only four $a_{i,j} = 0.25$ all other one are equal to zero. And, m_i is the fault. Only some variables $\xi_{i,1}$ are measurable.

First of all let us consider the following regular network (i.e. $a_{i,i-2} = a_{i,i-1} = a_{i,i+1} = a_{i,i+2} = 0.25$ all other $a_{i,j} = 0$, where $a_{i,21} = a_{i,1}$, $a_{i,22} = a_{i,2}$, $a_{i,0} = a_{i,19}$ and $a_{i,-1} = a_{i,18}$). Consider the case of no fault in the system. With one output, such as

$$y = \xi_{i_0,1}$$

for a fixed i_0, the system is not observable, and the dimension of the observable space is only 20; with two successive outputs, such as

$$y = \left[\xi_{i_0-1,1} \; \xi_{i_0,1} \right]^T$$

the dimension of the observable space is 36; with three successive outputs the dimension of the observable space is 38; and only for four and more successive outputs the regular system is observable. However, in the presence of fault m_i, the system is still unobservable using four successive outputs.

On the other hand, if a fault detection problem is considered vertex by vertex, then the state of a vertex can be observable even if the entire system is not. Define $Z(t, \xi, m)$ for the i^{th} vertex as follows

$$Z(t, \xi, m) = (\xi_{i,2}, m_i)^T$$

The outputs are defined by

$$\xi_{i-2,1} = y_{-2}, \xi_{i-1,1} = y_{-1}, \xi_{i,1} = y_0, \xi_{i+1,1} = y_1, \xi_{i+2,1} = y_2 \tag{12.13}$$

the following differential algebraic equations are obtains:

$$\begin{array}{l} Dy_0 \;\; = \xi_{i,2} + 0.25(y_{-2} + y_{-1} + y_1 + y_2) \\ D^2 y_0 = -y_0 + m_i + 0.25(Dy_{-2} + Dy_{-1} + Dy_1 + Dy_2) \end{array} \tag{12.14}$$

Therefore, $Z(t, \xi, m) = (\xi_{i,2}, m_i)^T$ is Z-observable. However, the entire system is not observable.

Now consider the same network with the same vertex dynamics, but with additional small world effect (see Figure 12.2). Assuming no fault, then with one output the dimension of the observable space is 36; with two successive outputs the system is observable, i.e. $Z = \xi$ is Z-observable. This seems to be a beneficial effect on observability of small world network architecture, i.e. less outputs are required to recover the full state of the system. Nevertheless, from this architecture and without further assumption on m_i, it is impossible to detect the fault on all vertices with the five successive outputs defined in (12.13). This shows that the sensors repartition is of first importance for fault detection in network with small world structure. Moreover the choice of sensor localization must be done in accordance to the concept of Z-observability if the detection of m_i is requested and the observation of the full state is impossible.

12.4.2 Multi-cell Chopper: $Z(T_N)$-Observability

In a multi-cell chopper system, each cell has a switching frequency that is imposed by technological constraints such as transistor's limitation, power dissipation, etc. Meanwhile, the load voltage has a switching frequency as well. It is known that a multi-cell chopper has the property that the switching frequency of the load voltage is, by the theory, the switching frequency of the cells multiplied by the number of cells; and the maximum value of $\frac{dV}{dt}$, the rate of voltage variation, divided by the number of cells is the discontinuous gap of each cell. In addition, the hybrid time trajectory of the overall system is a combination of the hybrid time trajectories of all cells, i.e. the partition points of the time

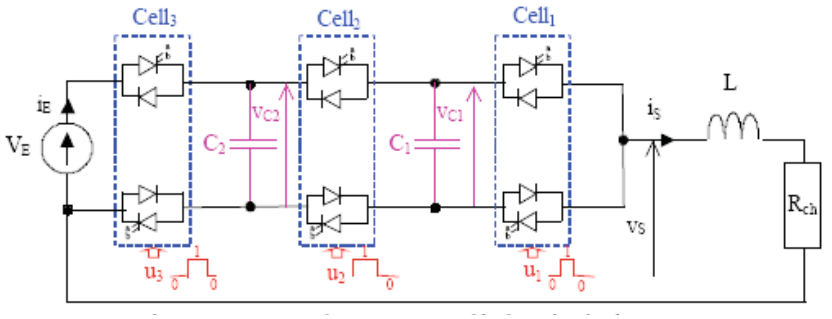


Fig. 12.3. Multi-Cell Chopper

interval is the union of all switching time points from individual cells. Because the hybrid time trajectory T_N and $< T_N >$ is fixed, the system is considered as being synchronous. Thus, the concept of $Z(T_N)$-observability is appropriate for studying the observability of multi-cell chopper systems. Nevertheless, this kind of multi-cell chopper systems (see Figure 12.3) is more difficult to control and to estimate when comparing to regular chopper systems. This is due to the extra difficulties to observe and to control each voltage capacitors, in this case V_{C1} and V_{C2}. More precisely, the observation problem is to recover the value of V_{C1} and V_{C2} from the measurement of i_s and V_s. The answer to this problem is not so obvious because, in any time interval I_i, the voltage of neither capacitors is observable. In fact, for any system configuration, the accessible information is always a combination of the voltages of both capacitors. This is a main reason that leads us to deal with $Z(T_N)$-observability by using lemma 12.3.2. The goal is to recover the voltage of both capacitors by taking advantage of appropriate hybrid time trajectories.

Hereafter, $x_1 = V_{C1}$, $x_2 = V_{C2}$, $x_3 = I_S$, and $x = \left[x_1 \ x_2 \ x_3 \right]^T$ are considered as the state variables of the system. The parameters L and R_{ch} are system uncertainties, which are considered unknown. Let us assume that we can measure the state of each cell and the power voltage V_E. The measured outputs are

$$\begin{aligned} y_1 &= V_S = (Cell_1 V_{C1} + Cell_2(V_{C2} - V_{C1}) + Cell_3(V_E - V_{C2}) \\ y_2 &= I_S \end{aligned}$$

The status of the cells is represented by the parameter $Cell_i$. It is 1 if the upper switch is "on" and the lower switch is "off"; and 0 for the opposite. Thus the status of the system can be represented by a vector of zeros and ones. For instance, (0,0,1) implies that $cell_1$ and $cell_2$ are zeros and $cell_3$ equals one. Consequently, depending on the status of Cells, the multi-cell chopper has 2^3 different statuses. The dynamics of the system is:

$$\begin{aligned} \dot{x}_1 &= \frac{(Cell_2 - Cell_1)}{C1} x_3 \\ \dot{x}_2 &= \frac{(Cell_3 - Cell_2)}{C2} x_3 \\ \dot{x}_3 &= \frac{V_S - R_{Ch} x_3}{L} \end{aligned} \tag{12.15}$$

The problem is to exhibit some particular hybrid time trajectory T_N and $\langle T_N \rangle$ such that the system (12.15) is $Z(T_N)$-observable for $Z(t,x) = \left[x_1, x_2 \right]^T$. It can be verified that $Z(t,x) = \left[x_1, x_2 \right]^T$ is not $Z(T_N)$-observable with any hybrid time trajectory under discrete status (0,0,0) or (1,1,1). However, if a trajectory of the system satisfies that the status is (1,0,0) in period I_1 and (1,1,0) in period I_2, then $Z(t,x) = \left[x_1, x_2 \right]^T$ is $Z(T_N)$-observable along such a trajectory. More precisely, in I_1, we define

$$P_1 = \left[1 \ 0 \right]$$

Obviously, we have $\bar{P_1} Z = x_1$ and $\dot{x}_1 = 0$. Similarly in I_2 we define

$$P_2 = \left[0 \ 1 \right]$$

Then, $\bar{P_2} = x_2$ and $\dot{x}_2 = 0$. Because

$$Rank \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = 2$$

this problem satisfies the assumptions in Lemma 12.3.2. Therefore, $Z(t,x) = \left[x_1, x_2 \right]^T$ is $Z(T_N)$-observable. It is important to note that the dynamics of $\dot{x}_3$ is totally unknown and this is not an obstacle for the $Z(T_N)$-observability of the function $Z(t,x) = (x_1, x_2)^T$. In [3], based on the concept of $Z(T_N)$-observability, some observer designs for multi-cell chopper systems, where Z is a function of continuous states as well as the load voltage, were proposed; simulation results were shown to highlight the usefulness of this approach.

12.4.3 Analogical Switched System: $Z(T)$ Observability

Let us consider the following system (see Figure 12.4):

$$\begin{aligned} \dot{x}_1 &= \frac{2(q-1.5)V_S - x_1}{R_1 C_1} \\ \dot{x}_2 &= \frac{x_1 - x_2}{R_2 C_2} \\ y &= x_1 \end{aligned} \tag{12.16}$$

Where $x = (x_1, x_2)^T = (v_{c_1}, v_{c_2})^T$ with $q = 1$ if $x_2 > 0$ and $q = 2$ if $x_2 \leq 0$. The system (12.16) is not a synchronous system. Moreover it is not $Z(T_N)$-observable with $Z = x$ for any hybrid time trajectory. This is due to the fact that, in any time interval $\tau_i = [T_{i,0}, T_{i,1}]$, x_2 is not observable through the measurement of x_1. Nevertheless, $\dot{y}$ uniquely determines q

$$q = \frac{\dot{y}_1 R_1 C_1 + y}{2V_S} + 1.5$$

So, it is possible to determine the switching instant and consequently the partition points t_i in the time trajectory. At these points, x_2 equals zero. Assume

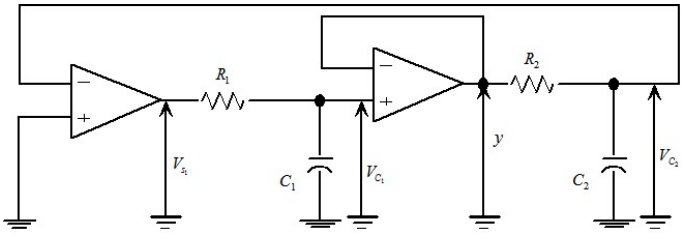


Fig. 12.4. Analogical switched circuit

that T is sufficiently large[4]. Define $P = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and $Q = \begin{bmatrix} 0 & 1 \end{bmatrix}$. From lemma 12.3.3, we conclude that the system is $Z(T)$-observable.

Remark 12.4.1. In this case the information on x_2 can be interpreted as a particular case of observation under sampling; and the behavior knowledge between the sampling points is uniquely determined by the dynamics model of x_2. (see condition (4) of lemma 12.3.3).

12.5 Conclusions

Through some definitions and examples, it is revealed that partial observability is a useful concept for complex systems, networked or switched, in which complete observability is either impossible or unnecessary. The concept of observability with precision is proved to be a measure of the robustness of observability. It refines the conventional concept of observability. In fact, it is able to tell whether a variable is practically observable in the presence of measurement error. The concept has great potential not only because of its practical perspective, but also because the concept is numerically verifiable for general nonlinear systems.

References

1. Angeli, S., Sontag, E.: Forward completeness, unboundedness observability and their Lyapunov characterisations. Systems & Control Lett. 38, 209–217 (1999)
2. Ames, A., Zheng, H., Gregg, R., Sastry, S.: Is there life after Zeno? In: Taking executions past the breaking (zeno) point, IEEE Proc. of American Control Conference (2006)
3. Bejarano, F., Ghanes, M., Barbot, J.P.: Observability analysis and observer design for a class of hybrid systems: application to multi-cell chopper (submitted for publication) (2008)
4. Balluchi, A., Benvenuti, L., Di Benedetto, M.D., Sangiovanni-Vincentelli, A.L.: Design of observers for hybrid systems. In: Tomlin, C.J., Greenstreet, M.R. (eds.) HSCC 2002. LNCS, vol. 2289, p. 76. Springer, Heidelberg (2002)
5. Bemporad, A.G., Ferrari, T., Morari, M.: Observability and controllability of piecewise affine and hybrid systems. IEEE Trans. on Automat. Contr. 45(10), 1864–1876 (2000)

[4] Including at least one switching instant t_i.

6. Boutat, D., Benali, A., Barbot, J.P.: About the observability of piecewise dynamical systems. In: Proc. of IFAC, NOLCOS (2004)
7. Brogliato, B.: Nonsmooth mechanics models. In: Dynamics and Control Series, Communications and Control Engineering. Springer, Heidelberg (1999)
8. Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Method in Fluid Dynamics. Springer, New York (1988)
9. De Santis, E., Di Benedetto, M., Girasole, G.: Digital idle speed control of automative engines using hybrid models. In: Proc. of IFAC World Congress, Prague (2005)
10. Diop, S., Fliess, S.: On nonlinear observability. In: Proc. of ECC 1991, vol. 1, pp. 152–157 (1991)
11. Fahroo, F., Ross, I.M.: Costate estimation by a Legendre Pseudospectral Method. In: Proc. of the AIAA Guidance, Navigation and Control Conference, Boston, MA (August 1998)
12. Gauthier, J.P., Kupka, I.A.K.: Deterministic Observation Theory & Applications. Cambridge University Press, Cambridge (2002)
13. Goebel, R., Hespanha, J., Teel, A.R., Cai, C., Sanfelice, R.: Hybrid systems: Generalized solutions and robust stability. In: Proc. of IFAC, NOLCOS (2004)
14. Gong, Q., Kang, Q., Ross, I.M.: A pseudospectral method for the optimal control of constrained feedback linearizable systems. IEEE Trans. on Automat. Contr. 51(7), 1115–1129 (2006)
15. Hermann, R., Krener, A.J.: Nonlinear controllability and observability. IEEE Trans. on Automat. Contr. 22(9), 728–740 (1977)
16. Kang, W., Barbot, J.P.: Discussions on observability and invertibility. In: Proc. of IFAC, NOLCOS, Pretoria, South Africa, August 2007, pp. 14–22 (2007)
17. Krener, A.J., Kang, W.: Locally Convergent Nonlinear Observers. SIAM J. on Control and Optimization 42(1), 155–177 (2003)
18. Lygeros, J., Johansson, H.K., Simć, S.N., Zhang, J., Sastry, S.S.: Dynamical Properies of Hubrid Automata. IEEE Trans. on Autom. Control 48(1), 2–17 (2003)
19. Maithripala, D., Berg, J., Dayawansa, W.P.: Nonlinear dynamic output feedback stabilization of electrostatically actuated MEMS. In: Proc. of the IEEE Conference on Decision and Control, Maui, pp. 61–66 (2003)
20. Meynard, T., Foch, H.: Multi-level choppers for high voltage applications. EPE Journal 1 (1992)
21. Watts, D.S., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 440–443 (1998)
22. Vu, L., Liberzon, D.: Invertibility of switched linear systems. Automatica 44(4), 949–958 (2008)
23. Respondek, W.: Right and left invertibility of nonlinear control system. In: Sussman, H. (ed.) Nonlinear Controllability and Optimal Control, pp. 133–176. Marcel Dekker, New York (1990)
24. Tanwani, A., Liberzon, D.: Invertibility of nonlinear switched systems. In: Proc. of the 47th IEEE Conference on Decision and Control, Cancun, Mexico (2008)

13

Feedback Stabilization of Solitons and Phonons Using the Controlled Lax Form

R. Palamakumbura[1], D.H.S. Maithripala[2], J.M. Berg[3], and M. Holtz[4]

[1] Dept. of Engineering Mathematics, Faculty of Engineering, Univ. of Peradeniya, Peradeniya, Sri Lanka
[2] Dept. of Mechanical Engineering, Faculty of Engineering, Univ. of Peradeniya, Peradeniya, Sri Lanka
[3] Dept. of Mechanical Engineering & Nano Tech Center, Texas Tech University, Lubbock, Texas, USA
[4] Dept. of Physics & Nano Tech Center, Texas Tech University, Lubbock, Texas, USA

Summary. We consider the problem of asymptotically stabilizing a desired family of soliton solutions of a completely integrable system. We proceed from the Lax form of the system, to which we add a suitable control term. We present a framework for making a Liouville torus, on which a particular multi-soliton lies, a global attractor. The method applies tools from nonlinear control theory to a controlled Lax form, augmented by a set of output functions. In principle there may be a large number of such outputs, but in practice we observe that it is often sufficient to use only a small subset. For the periodic Toda lattice the number of outputs should be equal to the number of lattice elements, but for lattices of up to 50 elements we show reasonable performance for as few as four output functions. We also apply the method to a discretization of the Korteweg-de Vries equation for which a complete set of independent invariant functions may be arbitrarily large, and observe reasonable performance using only three outputs. Finally we apply our results for the Toda lattice to an important potential application—the control of thermal transport at the nanoscale.

13.1 Introduction

The field of nonlinear science has experienced explosive growth since the 1970s, fueled by a steady increase in computing power and by the broad range of disciplines and phenomena to which its methods apply. The subject of this paper is the feedback control of solitary waves and solitons in discrete arrays, or in discrete approximations to distributed systems. Solitons, chaos, and reaction-diffusion phenomena comprise the three main branches of the field of nonlinear science [11, 20, 23, 24, 25, 33, 42]. A solitary wave is a localized traveling solution of a wave equation. A soliton is a solitary wave that retains its shape and speed after colliding with another solitary wave. The control problem discussed here is that of making a particular soliton solution, or specifically a class of soliton solutions, a globally asymptotic attractor. Solitary wave and soliton

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 217–238.
springerlink.com

solutions arise from nonlinear wave equations due to the balance of dispersion and nonlinear effects [32, 33, 40, 42, 43, 48]. Some of the well known systems that support soliton solutions are the Korteweg-de Vries (KdV) equation, the self-induced transparency equation, the sine-Gordon equation, the Toda lattice, the Boussinesq equation, the Born-Infeld Equation, and the nonlinear Schrödinger equation. In this work we specifically treat the Toda lattice and KdV equation.

We briefly survey some of the physical phenomena modeled by these dynamical systems, to give a sense of the range of potential applications of soliton control:

A large class of hyperbolic systems reduce to the KdV equation, which may be used variously to represent lossless propagation of shallow water waves, ion-acoustic waves in plasma, magnetohydrodynamic waves in plasma, longitudinal dispersive waves in elastic rods, pressure waves in liquid-gas bubble mixtures, rotating flow down a tube, and continuous approximation to an anharmonic lattice and thermally exited phonons in a crystal. The nonlinear Schrödinger equation is a model for stationary two-dimensional focusing of a plane wave, one-dimensional self-modulation of a monochromatic wave, self-trapping phenomena of nonlinear optics, propagation of a heat pulse in a solid, Langmuir waves in plasmas, and energy transport in an α-helix protein. It is also related to the Ginzburg-Landau equation of superconductivity. The KdV and nonlinear Schrödinger equations are often used generically for describing nonlinear wave phenomena with small nonlinearity and dispersion, the former at low frequencies and the latter at high frequencies.

The Born-Infeld equation is a nonlinear version of Maxwell's equations that permits modeling electrons as singularities. The Boussinesq equation describes the bi-directional propagation of shallow water waves. The sine-Gordon equation models the dynamics of domain walls in ferromagnetic and ferroelectric materials, the propagation of splay waves on lipid membranes, self-induced transparency of short optical pulses, the propagation of quanta of magnetic flux along long Josephson transmission lines, and slinky modes in reversed-field pinch plasma confinement machines. The Toda lattice is used to model thermal transport in crystal lattices, conduction in nonlinear transmission lines, energy transport in an α-helix protein, and supersonic sound propagation through the atmosphere. Solitons of the Toda lattice have been used to code secure communication system.

More detail on these equations and their applications may be found in [40] and the references therein, as well as in several more recent texts [11, 20, 23, 24, 27, 32, 42]. In contrast to the large body of work on the *analysis* of soliton behavior, the subject of soliton *control* has only been touched upon. Preliminary results have been reported in applications to optical soliton lasers [18, 31, 7] and the electrical soliton oscillator [39]. Theoretical results have been reported in the use of wave-like small forcing to create soliton solutions in the KdV equation and Toda lattice from stationary (vacuum) conditions [14, 22].

In addition to their relevance as models of important physical phenomena, soliton-generating equations may be used as templates for desired dynamical behavior in large-scale systems. We have previously shown how control may

be used to *impose* the dynamics of the Toda lattice or KdV equation upon a large array. Arbitrary desired dynamical behaviors are then approximated as multi-solitons of the superimposed system. Thus control of solitons may be seen as a precursor to a general theory of dynamic pattern generation in large arrays. Along these lines, we have considered the problem of embedding soliton-generating dynamics in large arrays of microelectromechanical systems (MEMS) for micro-propulsion, for secure information transmission, or for efficient optical pattern generation [35, 36, 38, 9].

Another potential application of soliton control is to thermal management of solid-state devices at the nanoscale, through the control of *phonons*. Phonons are vibrations of the crystal lattice, and are an important mechanism for the flow of heat in epitaxial dielectrics and semiconductors. To a surprising degree, models such as the KdV and the Toda lattice display many salient phenomena observed in these crystal structures [26]. The ability to control these vibrations—including the ability to stabilize phonons—would provide unprecedented new device design capabilities, for example, in the mitigation of self-heating in devices such as heterostructure field-effect transistors (HFETs). Here confinement of electrical conduction to planar layers with thickness on the order of tens of nanometers creates intense local heating, often leading to failure [1]. We present a preliminary investigation in [37]. Phonon control would also be beneficial for the design of solid-state thermoelectric energy convertors, which rely on epitaxial heterostructures that must be electrically conductive yet thermally insulating. While phonons are not typically solitons, the models and control formulation developed for soliton control also proves relevant to a (simplified) phonon control problem.

In the above-mentioned applications, the control objective is not to stabilize an equilibrium or even a relative equilibrium; nor is it to drive the system to a specific trajectory. Rather the object is to drive the behavior of the system to one of a *class* of solutions. For example, in a secure communication pulse generator we may wish to produce a single soliton of specific amplitude, but with arbitrary phase. In a soliton laser we may wish to pass any single soliton solution, regardless of phase and amplitude, while filtering out multiple soliton or non-soliton solutions. In the nanoscale heat mitigation problem it may be sufficient to shift energy from standing wave solutions to any traveling wave solution with sufficiently large group velocity. Previously we considered this type of problem for the periodic Toda lattice and discretized periodic KdV equation. The family of solutions addressed there was a multi-soliton specified up to an arbitrary phase shift. It was shown, based on heuristic arguments, that this family could indeed be globally asymptotically stabilized using an actuation input for each lattice element and a feedback signal that required knowledge of the full state of the array [35, 36, 38].

The goal of stabilizing a desired family of solutions may be formalized by posing the problem in an appropriately defined quotient space. In the present paper we describe progress towards a rigorous framework for development of control structures. We revisit our earlier results, and show that they can be

considered a specialization of this more general approach. This insight gives guidance for improving and extending those results.

Our construction takes advantage of the structure of *completely integrable* systems. All the soliton-producing equations mentioned previously are of this type. Completely integrable systems have a rich and well-studied structure. It is known that for a completely integrable system with $2N$ states there exists a choice of local coordinates, called the *action-angle variables*, that generically transforms the system into N decoupled undamped second-order oscillators with frequency a function of the action variables, whose solutions are completely characterized by a constant amplitude (the "action") and linearly increasing phase (the "angle"). The possible trajectories of each decoupled oscillator can therefore be thought of as forming concentric circles. Thus the complete solution space is foliated by (possibly degenerate) toruses, called Liouville toruses, where the solutions that lie on a particular Liouville torus differ from each other only in phase. The Liouville toruses are natural objects with which to construct the equivalence classes that define a particular control problem. The objective then becomes that of making a desired set of Liouville toruses a global attractor. This can be formalized by requiring the action variables to asymptotically converge to desired set-points, which may be zero, or a specified positive constant. The action variables are also convenient to characterize solitons. When solitons exist, they correspond to a one-dimensional Liouville torus, that is, to a circle or a single action-angle pair. Unfortunately for this approach, the action variables are difficult to write explicitly.

In principle the action-angle variables may be found using the *Lax form*. The Lax form is a first-order differential equation for the Lax operator, denoted $L_\lambda(t)$. Here $L_\lambda(t)$ evolves on a manifold $\mathcal{L}$ and depends analytically on a complex parameter λ. For finite-dimensional systems, like the periodic Toda lattice, the Lax operator has a square matrix representation. For infinite-dimensional systems, like the KdV equation, $L_\lambda(t)$ is a differential or difference operator. The Lax equation is isospectral, that is, the eigenvalues of $L_\lambda(t)$ are constant along any solution of the Lax equation. The *spectral curve* is the algebraic curve $\Gamma = \{(\lambda, \mu) \in \mathcal{C}^2 \mid \det(L_\lambda(t) - \mu I) = 0\}$. By the isospectral property, the spectral curve is independent of time. The parameters that define Γ are the *moduli*. The action variables of the system are the moduli of Γ. In the infinite-dimensional case they are given by certain functions of the *scattering data* associated with the Lax operator, while in finite dimensions they are given by the *spectral data*.

In the absence of control, the trajectory of a completely integrable system must remain on a single Liouville torus. Thus control is necessary to asymptotically stabilize a desired torus or set of toruses. To the standard Lax form we add a control term. The control term must be constructed carefully. For one, it must lie in the tangent space to $\mathcal{L}$ at $L_\lambda(t)$. For another, it should be a physically meaningful object. However in the presence of the control term, the spectrum of $L_\lambda(t)$ is no longer invariant, and the procedure sketched above does not apply. Thus it is not clear how to obtain the action variables directly for use

as output functions. Instead we pick more easily computable quantities. In the finite dimensional case, these are the N (not necessarily independent) functions $\text{trace}\left\{L_\lambda(t)^i\right\}$, $i = 1, ..., N$, which are determined by the eigenvalues of $L(t)$, and hence fully characterize the Liouville toruses. We have demonstrated their use to stabilize a single torus [35, 36, 38], however the extension to families of toruses is unclear.

For a large network, the tasks of sensing of information, computation of control action, and commanding of individual actuators can be daunting, even for macroscale systems. These problems are greatly compounded in microsystems, such as MEMS arrays, where fabrication constraints restrict sensors and actuators to be of simple design and limit functionality. Furthermore in microsystems the difficulties associated with addressing individual elements, combined with parasitic effects due to long interconnections, makes local control strongly preferable to centralized control. Thus feedback approaches are desirable that distribute relatively simple controllers throughout the structure. The problem becomes even more severe at the nanoscale, as for phonon control, where the necessary control structures may need to be only a few tens of atomic layers in size. While the details of computation, sensing, and actuation are beyond the scope of the present paper, we are at all times motivated by the need to distribute the control action, and to minimize sensing and data transfer requirements. We first addressed this in [37], where it was shown that convergence to a neighborhood of the desired class of solutions could be accomplished with a combination of local control and a small number of global connections. We continue to stress the importance of this issue here, where in addition to the fully sensed and actuated cases, we consider reduced sets of output functions, and controls that are applied only on *blocks* of array elements, interspersed with blocks of uncontrolled elements.

The present paper considers the use of feedback to stabilize a desired soliton solution of a completely integrable system in Lax form. The system may represent physical behavior in a discrete lattice, physical behavior in a continuum, or artificially imposed discrete dynamics. Section 13.2 reviews the relevant properties of the Lax form. Section 13.2.1 outlines a general control procedure for finite-dimensional systems, which is demonstrated on the Toda lattice in Section 13.3.1. We then consider the effects of reducing the number of output functions in Section 13.3.1. Section 13.3.2 presents control of the discretized KdV equation with only a few output functions. Finally, Section 13.4 treats phonon control and the nanoscale thermal transport problem.

13.2 Completely Integrable Systems and the Lax Form

A dynamical system is said to be *completely integrable* if the solution of the system can be constructed through explicit integration of the equations of motion. A pioneering result in the study of such systems is the Arnold-Liouville theorem [2], which states that a $2N$-dimensional Hamiltonian system is completely integrable if there exist N functionally independent and mutually involutive

integrals of motion. The Arnold-Liouville theorem further states that generically the flow of such a system is linear on an associated N-torus of the state space. The Liouville toruses are the level sets of the integrals of motion of the system. These results follow from being able to find a special set of local co-ordinates called the action-angle variables of the system. In these co-ordinates the action variables are conserved and the angle variable evolves linearly with time. Thus the Liouville toruses are in fact characterized by the action variables. In general not all conserved quantities are functionally independent. Thus the dimension of the Liouville torus on which a general solution lies is equal to the number of functionally independent conserved quantities and hence to the number of nonzero action variables.

The control problem considered in this paper is that of stabilizing a particular torus or sets of toruses. Finding and characterizing the action variables play a crucial role in this work, and is greatly facilitated by use of the *Lax form* [3, 8, 32, 33]. Starting from a completely integrable system

$$\dot{p} = X(p), \tag{13.1}$$

where $p \in \mathcal{P}$, the Zakharov-Shabat construction yields a consistent Lax pair, L_λ and B_λ. The pair depends analytically on a parameter $\lambda \in \mathcal{C}$, called the spectral parameter, such that the *Lax equation*,

$$\dot{L}_\lambda(t) = [B_\lambda(t), L_\lambda(t)], \tag{13.2}$$

is equivalent to the equations of motion of (13.1). Here $L_\lambda(t)$ and $B_\lambda(t)$ are operators that depend on the state variables and $[\cdot, \cdot]$ denotes the commutator. This equivalence holds true identically in λ. That is, with a given solution $p(t)$ of the equation (13.1) we can associate a family of solutions $L_\lambda(t)$ of the Lax equation (13.2). This family of solutions is parameterized by λ. It can be shown that the solution $L_\lambda(t)$ of (13.2) is an isospectral deformation of $L_\lambda(0)$, that is, the conserved quantities of the system (13.1) are contained in the eigenvalues of $L_\lambda(t)$, or, equivalently in the case of finite dimensional systems, the traces $H_i = \text{trace}(L^i_\lambda(t))$ for $i = 1, 2, \cdots$. The action variables are functions of the functionally independent H_i. The importance of the Lax formulation is that it allows one to compute the action-angle variables of the system. For infinite dimensional system they are given by the *scattering data* of the L_λ operator and in finite dimensional systems they are given by the *spectral data* of L_λ [3, 8, 32, 33]. Following [3, 33] we briefly review these aspects below. First we will consider finite dimensional systems where the Lax operator L_λ is a matrix.

Consider the family of solutions $L_\lambda(t)$ of the Lax equation (13.2) associated with a particular solution $p(t)$ of (13.1). The equations for the eigenvalues of $L_\lambda(t)$ describe an algebraic curve

$$\Gamma = \{(\lambda, \mu) \in \mathcal{C}^2 | \det(L_\lambda(t) - \mu I) = 0\}. \tag{13.3}$$

The desingularized version of this curve is referred to as the *spectral curve* associated with the solution $p(t)$ of (13.1) and can be shown to be a compact Riemann

surface. To be exact, when desingularized, Γ is a complex manifold with a complex structure. From now on whenever we refer to a spectral curve Γ we will always consider it to be the de-singularized curve. The flow of a Lax equation is an isospectral deformation of the initial state $L_\lambda(0)$, i.e. eigenvalues of $L_\lambda(t)$ are independent of t for each λ. Thus the spectral curve Γ is independent of time t. The parameters that define the spectral curve are referred to as the *moduli* of the spectral curve. The number of moduli necessary to characterize the spectral curve is equal to the genus g of the spectral curve. It can be shown that the moduli of the spectral curve are in one-to-one correspondence with the action variables of the torus on which $p(t)$ evolves [3]. Thus the dimension of this Liouville torus is also g. This defines a unique correspondence between the Liouville toruses that foliate the solution space, $\mathcal{P}$, and spectral curves of the form (13.3). In fact the torus can be identified with one of the connected components of a real slice of the Jacobian $J(\Gamma)$ of the spectral curve Γ. For additional details we refer to several excellent overviews [3, 8, 32, 33], and the references therein.

13.2.1 The Controlled Lax Form

Consider the evolution equation

$$\dot{p} = X(p) + g(p, u), \tag{13.4}$$

where u is a control. Using the Zakharov-Shabat construction we introduce the *controlled Lax form* with a spectral parameter λ:

$$\dot{L}_\lambda(t) = [B_\lambda(t), L_\lambda(t)] + U_\lambda(t), \tag{13.5}$$

such that (13.5) holds identically in λ if and only if (13.4) holds. This construction allows us to see how the control affects the spectral curve. Let $\psi_\lambda(t)$ be a normalized eigenvector of $L_\lambda(t)$ with eigenvalue μ_λ. That is

$$(L_\lambda - \mu_\lambda)\psi_\lambda = 0.$$

Differentiating we obtain

$$(L_\lambda - \mu_\lambda)(\dot{\psi}_\lambda - B_\lambda\psi_\lambda) = \dot{\mu}_\lambda\psi_\lambda - U_\lambda\psi_\lambda.$$

Taking the inner product of both sides with the normalized eigenvector ψ_λ and noting that $(L_\lambda - \mu_\lambda)$ is self-adjoint we have

$$\dot{\mu}_\lambda = \langle\langle \psi_\lambda, U_\lambda\psi_\lambda \rangle\rangle. \tag{13.6}$$

For each $\lambda \in \mathcal{C}$ the pair (λ, μ_λ) is a point on the spectral curve Γ, and in fact the collection of all such points are the spectral curve. Thus equation (13.6) describes how each point changes under the control U_λ and hence how the spectral curve Γ changes under the influence of the controls.

The control problem motivating this work is the stabilization of classes of solutions of integrable systems, enabling the design of soliton generators and

filters. Potential applications may be found in soliton lasers and phonon control, in information transmission over long distances, and in nanoscale thermal management. Equivalence classes $\mathcal{S}_1$ of solutions of (13.1) are defined by considering two solutions to be equivalent if they lie on the same Liouville torus in $\mathcal{P}$. All solutions evolving on the same Lioville torus have the same qualitative behaviour [12]. In finite-dimensional systems, stabilizing an element of $\mathcal{S}_1$ amounts to stabilizing a point in the moduli space of spectral curves. This can be accomplished by driving the conserved quantities to the corresponding values. The conserved quantities are readily computable as $H_i = \text{trace}(L^i_\lambda)$. We demonstrate this procedure in Section 13.3, where we stabilize an arbitrary n-soliton solution modulo a phase shift.

In what follows we propose to classify solutions based on spectral curves and use controls to asymptotically stabilize these classes of solutions. In the presence of a non-zero control, the spectral curve Γ of the controlled Lax equation (13.5) will no longer be time independent. This allows one to formulate the control problem as one on the space of moduli of spectral curves. We consider two control problems, both based on stabilizing equivalence classes of solutions of the uncontrolled Lax equation.

13.3 Stabilization of Solitons

Let $\tilde{M}$ be the Liouville torus that corresponds to a desired class of solutions in $\mathcal{S}_1$. The control problem that we consider in this section is that of finding a control U_λ that will asymptotically drive the solution of (13.4) to any solution of the uncontrolled evolution equation (13.1) that lies on $\tilde{M}$. That is, given any initial condition $p(0)$ we set out to find a control $u = \alpha(p)$, with the properties i) $\alpha(p)|_{\tilde{M}} = 0$, and ii) $\lim_{t\to\infty} p(t) \subset \tilde{M}$. Recall that as a consequence of integrability, the solution of the uncontrolled evolution equation (13.1) that starts at $p(0) \in \tilde{M}$ will remain on the torus $\tilde{M}$. In other words, we use feedback control to make a given torus $\tilde{M}$ an attractor of the controlled evolution equation (13.4).

In the presence of a non-zero control, the spectral curve Γ of the controlled Lax equation (13.5) will no longer be time-independent. We denote the spectral curve corresponding to the controlled solution $L(t)$ by $\Gamma(t)$. Since all solutions of the uncontrolled evolution equation (13.1) that lie on $\tilde{M}$ are completely characterized by the spectral curve $\tilde{\Gamma}$ the problem can now be transformed to that of finding a control U_λ so that $\lim_{t\to\infty} \Gamma(t) = \tilde{\Gamma}$ and we arrive at a stabilization problem defined on the space of the spectral curves associated with (13.2), or, more precisely, on the moduli space of the spectral curves associated with (13.2).

In the presence of controls, explicitly writing out the moduli in terms of the system variables is, in general, very hard. Since the moduli are a function of the conserved quantities of the system it suffices to drive the conserved quantities to their respective values. This is much easier, because they are given by the computable quantities $H_i = \text{trace}\left\{L^i\right\}$. However, since not all H_i are independent, this may represent an over-parameterization of the controls. In the next section

we demonstrate this approach on the N-periodic Toda lattice. The same method is extended to the periodic KdV equation in section 13.3.2.

13.3.1 Cnoidal Waves of the N-Periodic Toda Lattice

In this section we show how to stabilze any n-Cnoidal wave solution of the Toda lattice, modulo a phase shift. The Toda lattice is a discrete one-dimensional lattice originally proposed as an anharmonic model for crystal lattice vibrations. The controlled N-periodic Toda lattice in the Flaschka variables is described by

$$\dot{a}_n = a_n(b_n - b_{n+1}) \tag{13.7}$$

$$\dot{b}_n = 2(a_{n-1}^2 - a_n^2) + \frac{1}{2}u_j, \tag{13.8}$$

where $a_{n+N} = a_n$ and $b_{n+N} = b_n$ for all n. Here we give the controls the physical interpretation of intermolecular forces. The associated controlled Lax form (13.5) is obtained by introducing the $N \times N$ matrices L and B given by

$$L = \begin{bmatrix} b_1 & a_1 & 0 & \cdots & \cdots & a_N \\ a_1 & b_2 & a_2 & 0 & \cdots & 0 \\ 0 & a_2 & b_3 & a_3 & 0 & \cdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \\ a_N & 0 & \cdots & \cdots & a_{N-1} & b_N \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -a_1 & 0 & \cdots & \cdots & a_N \\ a_1 & 0 & -a_2 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & -a_3 & 0 & \cdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \\ -a_N & 0 & \cdots & \cdots & a_{N-1} & 0 \end{bmatrix},$$

and $D = \mathrm{diag}([u_1, u_2, \cdots, u_N])$.

Let $e_n = H_n - \tilde{H}_n$ and taking the derivative of e_n along the solutions of (13.7)—(13.8) we have

$$\dot{e}_n = \dot{H}_n = \left(\mathrm{trace}(L^{n-1}[B, L]) + \frac{1}{2}\mathrm{trace}(L^{n-1}U)\right).$$

Since $\mathrm{trace}(L^m)$ are integrals of motion of the uncontrolled system, $\mathrm{trace}(L^{n-1}[B, L]) = 0$ and we have $\dot{e}_n = \frac{n}{2}\,\mathrm{trace}(L^{n-1}U)$. Thus we have a linear time varying system

$$\dot{e} = A(t)u, \tag{13.9}$$

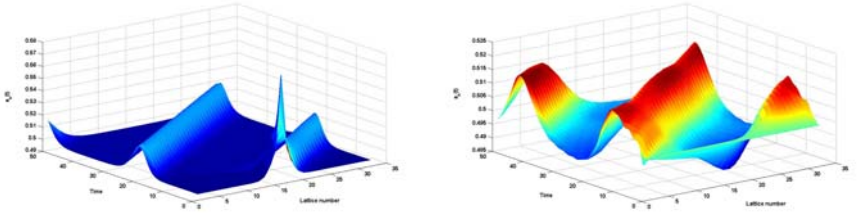
where $e = [e_1, e_2, \cdots, e_N]^T$, $u = [u_1, u_2, \cdots, u_N]^T$ and A is an $N \times N$ matrix with all entries on the first row equal to one-half and the j^{th} row consisting of $j/2$ times the diagonal entries of the matrix L^{j-1} for $j = 2, 3, \cdots, N$. Therefore we have reduced the problem to that of stabilizing a reduced-order linear time varying system. In what follows we will provide a control that globally stabilizes the origin of (13.9) provided certain detectability-like conditions are satisfied.

Define a Liapunov function $W(e) = \frac{1}{2}e^T K e$ where K is a positive definite symmetric $N \times N$ matrix. Then $\dot{W}(e) = e^T K \dot{e} = e^T K A u$, and therefore

$$u = -A^T K e, \tag{13.10}$$

globally asymptotically stabilizes the origin provided that the system (13.9) is zero-state detectable with output $y = -A^T K e$.

For $N = 32$, Figure 13.1 shows the effectiveness of this control where the initial condition was chosen to be a stationary or vaccum solution. The desired class of solutions were chosen to be the one-cnoidal wave solutions with modulus $k = 0.999$ and spatial periodicity equal to 32.



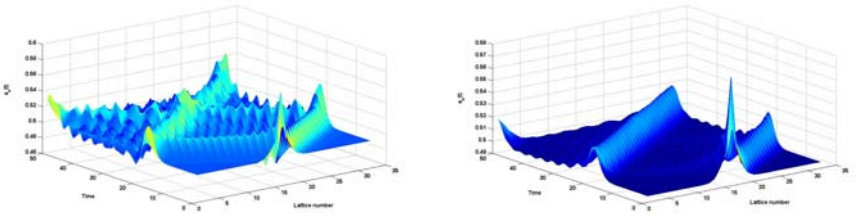
(a) (b)

Fig. 13.1. Fully actuated and sensed 32-Periodic Toda Lattice. Figure (a) corresponds to a Gaussian distribution and Figure (b) corresponds to a vacuum solution.

The Periodic Toda Lattice Under Local Feedback Control

The full-state feedback controller described by (13.10) uses complete system information to compute the control input at each lattice point. This level of measurement is typically not feasible. Therefore we consider the possibility of approximating the full-state feedback controller using more limited information.

The local feedback controller is suggested by the Fermi-Pasta-Ulam numerical experiment, in which the dynamic behavior of a chain of masses connected by springs with cubic nonlinearities was simulated (see e.g. [32]) using one of the first digital computers. It was expected that the system would settle at a thermal equilibrium in which every mode had an equal amount of energy. Instead, the simulations revealed nearly periodic solutions. Gardner et. al. [15, 16] explained this result by pointing out that the simulated solution is in the proximity of a multi-soliton. Indeed, it has been observed by physicists [45, 46] that typical solutions of the Toda lattice (and other equations) are generally near a low-order multi-soliton, and hence that a small number of parameters suffice to characterize it "nearly completely." Our argument is accordingly that the first few integral invariants $H_m, m = 1, \cdots$ should approximately characterize a typical solution of (13.7)–(13.8), and that driving this set of invariants asymptotically to the desired values should suffice to drive the state close to the desired solution. Therefore, in the equations representing error dynamics of the controlled Toda lattice, we disregard all but the first few equations for the purposes of controller development. In particular, the simulations shown in Fig. 13.2 include only the first four terms.



(a) (b)

Fig. 13.2. Figure (a) corresponds to the uncontrolled lattice and figure (b) corresponds to the lattice with local feedback.

Fix a small integer k and consider the first k error dynamic equations,

$$\dot{\epsilon}(t) = \check{A}(t)u(t), \tag{13.11}$$

where, $\epsilon = [H_1(t) - \tilde{H}_1, \cdots, H_k(t) - \tilde{H}_k]'$, $u = [u_1, u_2, \cdots, u_N]^T$ and $\check{A}_{i,j} = i(L^{i-1})_{j,j}$, $1 \le i \le k, 1 \le j \le N$. Now consider the control law,

$$u = -\check{A}^T(t)\Lambda\epsilon. \tag{13.12}$$

Because the $(i,j)^{\text{th}}$ element of $\check{A}$ only involves elements $L(p,q), j-k \le p,q \le j+k$, of L, it is clear that the controller only uses information from the nearest k neighbors on each side to construct the control input at each node, modulo the global information $\epsilon_i, 1 \le i \le k$.

With reasonable initial conditions, e.g. zero, one soliton, two solitons, half sine waves, exponentials etc., and networks of up to 50 elements, simulations are successful with $k = 4$ and exponentially decreasing diagonal elements of Λ (i.e. $\Lambda(j+1, j+1) = (1/N)\Lambda(j,j)$). A theoretical framework for formal error analysis is under development, for example to answer the question of how many error terms must be included in the control law to get within a given δ of the desired solution. researchers in this field sometimes refer to "temperature" in this context, as the number of soliton modes needed to approximate a given solution [46]. It is likely that there is a close relationship between the "temperature" of the initial condition and the number k of the integrals that need to be controlled. Heuristically this suggests that "low temperature" patterns can be approached closely using a controller of low complexity, in the sense that the number of computations needed is a polynomial of the network size with a low order exponent.

13.3.2 The KdV Solitons

Consider the controlled KdV equation

$$q_t = 6qq_x - q_{xxx} + u(x,t). \tag{13.13}$$

with corresponding Lax formulation given by, $\frac{\partial L}{\partial t} = [B, L] + W$. The Lax operator L and the corresponding operator B are given by

$$L = -D^2 + q, \qquad B = -4D^3 + 3(qD + Dq).$$

where D is the spatial derivative operator $\frac{\partial}{\partial x}$. Here it is observed that L is the Schrödinger operator with potential $q(x,t)$. The eigenvalues of the Schrödinger operator are constant along the solutions of the KdV.

We remark that the control operator W may be chosen as

$$W = [B(w), L(w)]$$

where

$$L(w) = -D^2 + w, \qquad B_j = \alpha D^{2j+1} + \sum_{i=1}^{j} (b_i D^{2j-1} + D^{2j-1} b_i),$$

and b_i are polynomials in w and its derivatives to be chosen such that $[B(w), L(w)]$ is a zero order operator.

It is well known that the central difference spatial descretization preserves the integrals to second order. Suppose that we wish to employ k control elements to control the KdV flow from an arbitrary initial state to a soliton. The control problem may be stated as follows: Given the controlled KdV equation (13.13) where $x \in [0,1]$ and $u(x,t)$ is piecewise constant in x (more precisely $u(x,t) = \gamma_i(t),\ x \in [i/k, i+1/k]$), develop a feedback control law to asymptotically stabilize a desired one-soliton solution of the uncontrolled KdV.

In general, the space of spectral curves of the infinite Toda lattice is infinite dimensional. Thus the control approach described at the beginning of this chapter would in general result in an infinite dimensional control problem. Recall from section 13.3.1 that driving a few of the conserved quantities to the values corresponding to a desired solution approximately stabilizes a given class of solutions. Following the same idea for the KdV, we use only a few conserved quantities. These can be written in the form $H_n(q) = \int_0^1 \psi_n(q, \partial q, \cdots, \partial^{n-2} q) dx$ for $n \geq 2$ where ∂ denotes differentiation with respect to x. For an arbitrary m we designate the controls

$$u = -A^T(q, q^{(1)}, \cdots, q^{(2m-2)}) K e, \tag{13.14}$$

where $A(q, q^{(1)}, \cdots, q^{(2m-2)}) = (\delta\psi_1, \cdots, \delta\psi_m)^T$ with $\delta\psi_n = \sum_{j=0}^{n-2} (-1)^j \partial^j \frac{\partial \psi_n}{\partial q^{(j)}}$. A detailed derivation may be found in [35, 38]. Observe that (13.14) is equivalent to (13.12).

Now on the principle that the information contained in the initial data can be approximately captured by the first few integrals, we pick the parameters as $m = 3, \quad x \in [0, 15]$, $K = \mathrm{diag}(1,1,1)$, $q_0(t,x) = a^2 \mathrm{sech}^2(a(x - a^2 t/3)/\sqrt{12})$, $a = 3.5$. For numerical purposes we first discretized the system with respect to the spatial variable x using a symmetric difference method with step size 0.2 and integrated the ordinary differential equations by using the Runge-Kutta method

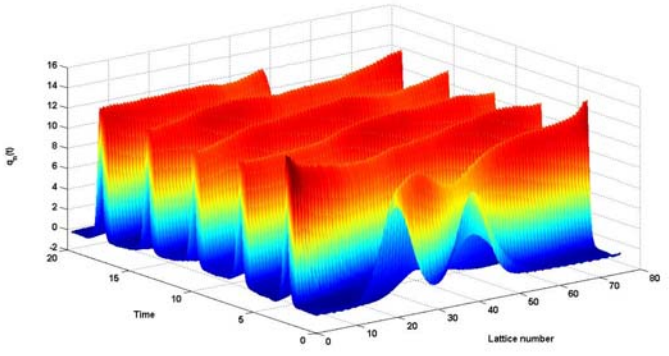


Fig. 13.3. Discretized KDV with local feedback control. The initial condition corresponds to a linear superposition of two Gaussian distributions.

with Matlab. Initial states are chosen to be the superposition of two Gaussian distributions. As seen in Figure 13.3, even using only three output functions, this control law provides excellent convergence to the soliton solution.

13.4 Nanoscale Thermal Transport

In this section we describe a potential application of the theory outlined above—nanoscale thermal transport. Heat conduction at nanometer length scales is inherently different from its macroscale counterpart. At the macroscale, heat conduction is typically well described by Fourier's law, which states that heat flux is proportional to the local temperature gradient, and by the diffusion equation, which does not admit wave-like solutions. Microscopically however, thermal energy takes the form of lattice vibrations, and thermal transport may be thought of as momentum transfer from a lattice element to its neighbors via inter-atomic forces. In strong contrast to macroscopic behavior, the lattice vibrations display behaviors such as traveling waves, standing waves, and solitons [20]. Progress has been made in connecting deterministic microscopic dynamics with macroscopic thermal phenomena using probabilistic models [6, 26].

The characteristic length scales of modern semiconductor devices are such that functional layers may only be a few tens of lattice constants thick. Furthermore the layers are high-quality crystalline materials. The thermal transport mechanisms are closer to those demonstrated by an ideal lattice than to those of a bulk solid. Important examples are the heterojunction bipolar transistor (HBJT) and the heterostructure field effect transistor (HFET) used in high power and microwave applications. Electrical conduction in an HFET, for example, is constrained to the planar interface between two dissimilar semiconducting materials. This layer is extremely thin, typically some tens of nanometers, and is sometimes referred to as a "two-dimensional electron gas" (2DEG). The highly

localized Joule heating associated with the 2DEG may degrade performance or cause material damage or device failure [1]. Optimal operation calls for innovative thermal management approaches, that recognize and exploit the unique behaviors possible in the thin epitaxial films that constitute these devices.

Vibrational dynamics are generally studied using harmonic oscillators in either a quantum mechanical or classical framework. In lattices this approach leads to quantized vibrational modes, known as *phonons*. An instructive microscopic model is the diatomic linear chain with nearest-neighbor harmonic interactions. This lattice supports solutions with distinct *optic* and *acoustic* phonon branches, corresponding to very different dispersion curves. At large wavelengths the optical phonon dispersion curve is flat, corresponding to group velocities that approach zero. At long wavelengths (the Brillouin zone center) the acoustic phonon dispersion curve has a relatively steep slope, and therefore a much higher group velocity. The *speed of sound* in the lattice is the group velocity of the long wavelength acoustic phonons. In contrast, the monatomic harmonic lattice dispersion curve has only an acoustic branch, but the group velocity tends to zero as the wavelength decreases (the Brillouin zone edge). The resulting small wavelength *standing wave* acoustic phonons of the monatomic model have a behavior similar to the optical phonons of the diatomic model.

In this section we describe two relevant control problems. The first is the problem of moving energy from slow modes to fast modes. The intended application of these results is the design of crystal structures for thermal management in electronic and optoelectronic epitaxial devices. The second is the design of a filter that removes modes above a given cut-off point. This would be potentially useful in, for example, a soliton laser. As a starting point we consider a linear lattice. It can be shown that the controls are functions of the action variables and hence moduli of the associated spectral curve. In principle this allows us to generalize the above results to general integrable systems. Such a generalization is desirable since the Toda lattice and KdV are one-dimensional models for thermal transport in crystal lattices that incorporate more realistic anharmonic effects. A very preliminary approach to these problems was discussed in [37].

13.4.1 The Harmonic Lattice

Consider the controlled N-periodic linear lattice given by

$$\ddot{q}_n = \omega_0^2 \left(q_{n+1} - 2q_n + q_{n-1}\right) + u_n, \tag{13.15}$$

where $q_{n+N} = q_n$ for all n. The lattice equation (13.15) can be written as

$$\ddot{q} = \Omega q + u. \tag{13.16}$$

where $q = [q_1, q_2, \cdots, q_N]^T$, $u = [u_1, u_2, \cdots, u_N]^T$ and

$$\Omega = \omega_0^2 \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -2 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & -2 \end{bmatrix}$$

This is a periodic Jacobi matrix and can be diagonalized. Let T be the modal matrix of Ω and let $T^T \Omega T = D = \text{diag}([-\omega_1^2, -\omega_2^2, \cdots, -\omega_N^2])$ where we have numbered the modes in increasing order. In the transformed co-ordinates $q = Tz$ the equation (13.16) reduces to the uncoupled form

$$\ddot{z} = Dz + W, \tag{13.17}$$

where $W = \text{diag}([w_1, w_2, \cdots, w_N]) = T^T u$. Writing the individual terms explicitly gives the modal equations

$$\ddot{z}_n = -\omega_n^2 z_n + w_n \qquad \text{for } n = 1, 2, \cdots, N. \tag{13.18}$$

The energy in the nth mode is given by

$$H_n = \frac{1}{2}(\dot{z}_n^2 + \omega_n^2 z_n^2). \tag{13.19}$$

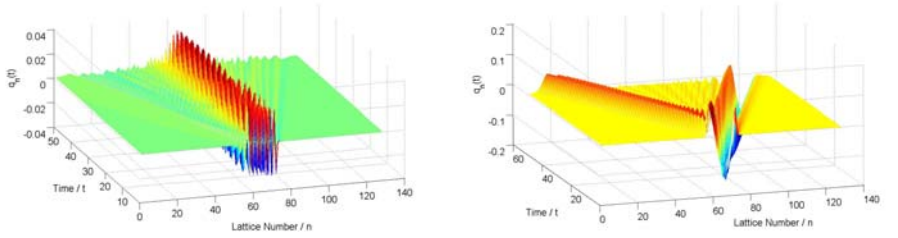
In the uncontrolled lattice it is clear that each of the H_n are invariants of the motion of the lattice. Furthermore the action variables of the system are also given by $\rho_n = \sqrt{2H_n}$. The Liouville toruses of the lattice are completely characterized by the action variable and are thus of the form

$$\mathcal{M} = \{(q_n, \dot{q}_n) \in \mathcal{R}^n \times \mathcal{R}^n | H_n = c_n \text{ where } c_n \in \mathcal{R}^+, \ \forall \ n = 1, 2, \cdots, N\}.$$

The dimension of the torus corresponds to the number of nonzero action variables and hence to the number of nonzero c_n.

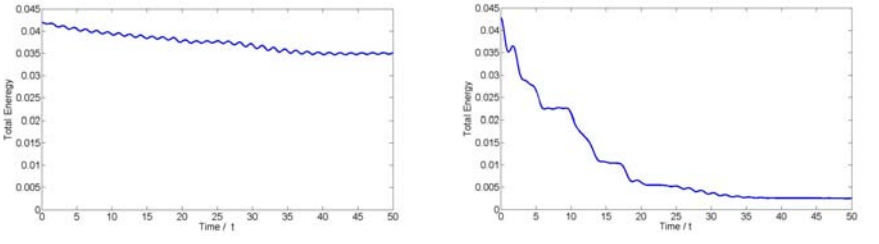
The Controlled Slow Phonon Decay Problem

In a monatomic lattice the long wavelength modes are usually referred to as zone center acoustic phonons while the short wavelength modes are referred to as zone edge acoustic phonons. (Note that by artificially inducing a diatomic lattice-like symmetry breaking—by treating every other lattice element as different—we can "fold" the Brillouin zone, and identify zone edge acoustic phonons with zone center optical phonons.) Initial perturbations that give rise to zone edge phonons persist for longer and contribute towards local heating. This is shown in the simulation of a 128 element lattice where an initial perturbation is provided to the middle 16 elements of the lattice. Figure 13.4 shows the time evolution of the lattice and Fig. 13.5 shows the total energy in the middle 16 elements of the lattice. Figure 13.4(b) show how the energy created by the fast moving



(a) Zone edge phonon (b) Zone center phonon

Fig. 13.4. Time evolution of zone edge and zone center acoustic phonon initiated in the middle 16 elements of a 128 element lattice.

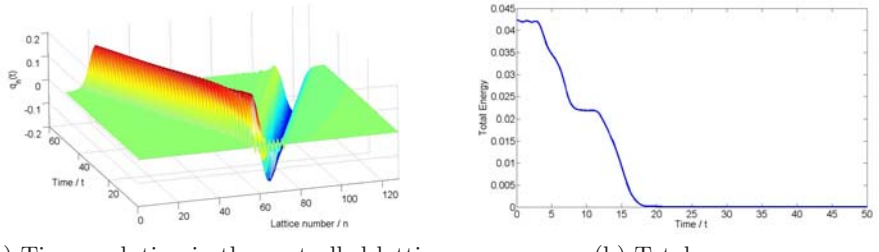


(a) Zone edge phonon (b) Zone center phonon

Fig. 13.5. The total energy in the middle 16 elements of the lattice with initial conditions corresponding to zone edge and zone center acoustic phonons.

zone center phonons dispersing to the left and right. Fig. 13.5(b) shows that the energy in the middle 16 elements dissipates rapidly. In contrast, Fig. 13.5(a) shows that the energy created by the zone edge phonons dissipates relatively slowly.

We are interested in using feedback control to transform the slow moving zone-edge modes into fast moving zone-center modes. Specifically, for any given



(a) Time evolution in the controlled lattice. (b) Total energy.

Fig. 13.6. The effect of the controls (13.21) in dissipation of energy corresponding to an initial zone-edge phonon.

initial condition the solution of the lattice should asymptotically converge to a solution with the following properties:

1. The amplitude of the last m modes of the desired solution should be zero, that is, energy is removed from the slow moving high wave number modes.
2. Since implementation of any controller will likely be through manipulation of the interatomic potential forces, it will not be possible for the controller to directly remove any thermal energy. Therefore we require that the energy removed from the last m modes be transfered to the lower wave number modes. For now, we arbitrarily designate one of these, the jth mode (where $1 \leq j \leq N-m$), as the energy recipient. That is, H_j for the desired solution should be $H_j(0) + \sum_{k=N-m+1}^{N} H_k(0)$.

Such a desired solution lies on a $(N-m)$-Liouville torus

$$\tilde{\mathcal{M}}_E = \left\{ (q_n, \dot{q}_n) \in \mathcal{R}^n \times \mathcal{R}^n \;\middle|\; \begin{array}{l} H_j = H_j(0) + \sum_{k=N-m+1}^{N} H_k(0), \\ H_k = H_k(0) \;\forall\; k = 1, \cdots, N-m \text{ and } i \neq j, \\ H_k = 0 \;\forall\; k = N-m+1, \cdots, N \end{array} \right\}.$$

We call this the problem of *controlled decay of slow phonons*.

Thus the control problem reduces to that of asymptotically stabilizing $\tilde{\mathcal{M}}_E$ using energy preserving controls. Unlike the control problem addressed in section 13.3.1 and 13.3.2 in this case the torus $\tilde{\mathcal{M}}_E$ is not known *a priori*, but rather is determined by the initial conditions. Observe that the Liouville toruses are completely charcterized by the values of H_n and that driving the last m number of them to zero while conserving the total energy and the H_i for $1 \leq i \leq N-m$ and $i \neq j$ achieves the control objective.

In the following we derive the controls that solve the phonon control problem for the linear lattice. The total energy of the entire lattice is given by $H = \sum_{n=1}^{N} H_n$. and the conservation of total energy in the lattice requires that

$$\dot{H} = \sum_{n=1}^{N} \dot{z}_n w_n = 0. \tag{13.20}$$

The effect of the controls on the energy in the nth mode is $\dot{H}_n = \dot{z}_n w_n$. Transforming the energy from all the last m modes to the jth mode ($1 \leq j \leq N-m$) while satisfying (13.20) requires that

$$\begin{aligned} w_n &= -\dot{z}_n \dot{z}_j^2 H_n \quad \text{for } n = N-m+1, \cdots, N \\ w_j &= \sum_{N-m+1}^{N} \dot{z}_j \dot{z}_i^2 H_i, \\ w_k &= 0 \quad \text{for } k = 1, \cdots, N-m \text{ and } k \neq j. \end{aligned} \tag{13.21}$$

The effect of the above control in dissipating the energy of a zone edge slow mode created in the middle 16 elements of the lattice is shown in figure 13.6. The controls transforms all the energy in this mode to the zone center mode.

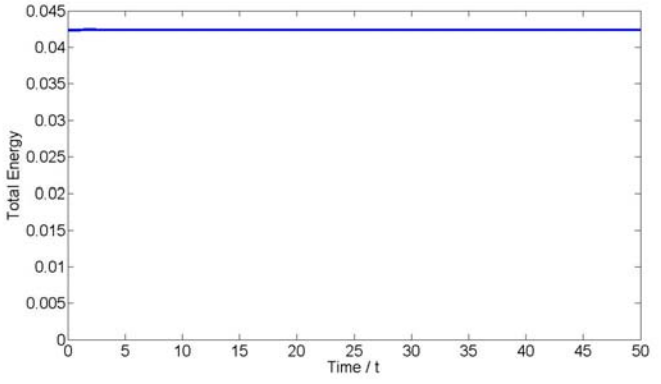
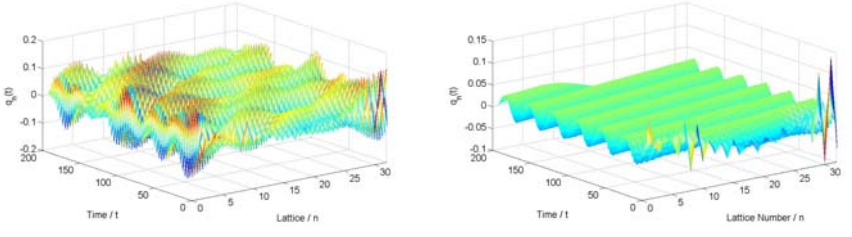


Fig. 13.7. Total lattice energy with energy-conserving energy transfer control for an initial zone-edge phonon.



(a) Unfiltered lattice. (b) Filtered lattice.

Fig. 13.8. The modal filter.

Since these have high group velocities they travel out of the region thereby rapidly cooling the middle section. Figure 13.7 shows that total lattice energy is conserved. Convergence to the desired solution is almost global. Only for initial conditions corresponding to $H_j(0) = 0$ do the solutions remains in the initial torus. However this is an unstable condition, and an arbitrarily small perturbation, such as might be due to thermal noise, will move the trajectories from the initial torus, and result in convergence to the desired torus.

Low Pass Filtering

If the energy preservation requirement is relaxed, we obtain a phonon filter by implementing the controls

$$w_k = \begin{cases} -\dot{z}_k H_k & \text{for } k = N - m + 1, \cdots, N \\ 0 & \text{for } k = 1, \cdots, N - m \end{cases} \tag{13.22}$$

The controlled lattice leaves invariant all the first $N-m$ modes while dissipating the energy in all the last m modes. That is the control (13.22) drives the solutions to the invariant torus

$$\tilde{\mathcal{M}}_F = \left\{ (q_n, \dot{q}_n) \in \mathcal{R}^n \times \mathcal{R}^n \;\middle|\; \begin{array}{l} H_k = c_k \in \mathcal{R} \;\forall\; k = 1, \cdots, N-m, \\ H_k = 0 \;\forall\; k = N-m+1, \cdots, N, \end{array} \right\},$$

where the c_k are determined by the initial conditions as $c_k = H_k(0)$. Figure 13.8 shows the results for a 32 element lattice where the initial condition was a linear combination of the last four modes and the first mode. Figure 13.8(a) shows the time evolution of the uncontrolled lattice while Figure 13.8(b) shows the time evolution of the controlled lattice with $m = 4$.

13.5 Conclusions

We have shown how standard tools of nonlinear control may be applied to the problem of stabilizing a desired family of solutions of a completely integrable system in Lax form. A control input is constructed to satisfy mathematical and physical constraints. A set of output functions is defined to completely characterize the target family. Then the objective of making the desired family a global attractor is equivalent to driving the output functions to specified constant values. In the case where the desired family is a specified soliton of arbitrary phase, the attractor is a Liouville torus of genus one. Results for this case were illustrated on the Toda lattice and the KdV equation. It was also shown that acceptable performance could be obtained using a reduced set of outputs and controls.

The results were also applied to a linear model of thermal transport at the nanoscale. Here the target families of solutions are particular lattice vibration modes, representing phonons. This linear model necessarily neglects effects such as interactions between modes and phonon decay, but the preliminary results show that energy may be transferred from slow-moving to fast-moving modes. We are currently studying application of the controlled Lax form to more realistic phonon models.

Acknowledgement. This research partially supported by NSF grants ECS0524713, CCF0523983, ECS0220314, ECS0218245, and CTS0210141.

References

1. Ahmad, I., Kasisomayajula, V., Holtz, M., Berg, J.M., Kurtz, S.R., Tigges, C.P., Allerman, A.A., Baca, A.G.: Self heating study of an AlGaN/GaN-based heterostructure field-effect transistor using ultraviolet micro-Raman scattering. Applied Physics Letters 86, 173503 (2005)
2. Arnold, V.I.: Mathematical Methods of Classical Mechanics, 2nd edn. Springer, New York (1989)

3. Babelon, O., Bernard, D., Talon, M.: Introduction to Classical Integrable Systems. Cambridge Monographs on Mathematical Physics. Cambridge University Press, Cambridge (2003)
4. Brockett, R.W.: Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. Linear Algebra Appl. 146, 79–91 (1991)
5. Brockett, R.W.: A synchronization mechanism for pulse streams and the synchronization of traveling waves. In: Proc. of the 40th IEEE Conference on Decision and Control (2004)
6. Cahill, D.G., Ford, W.K., Goodson, K.E., Mahan, G.D., Majumdar, A., Maris, H.J., Merlin, R., Phillpot, S.R.: Nanoscale thermal transport. Applied Physics Reviews 93(2), 793–818 (2003)
7. Carruthers, T.F., Duling III, I.N.: Dispersion management in a harmonically modelocked fiber soliton laser. Optics Letters 25(3), 153–155 (2000)
8. Dickey, L.A.: Soliton Equations and Hamiltonian Systems, 2nd edn. Advanced Series in Mathematical Physics. World Scientific, Singapore (2003)
9. Edd, J., Payen, S., Sitti, M., Stoller, M.L., Rubinsky, B.: Biomimetic propulsion mechanism for a swimming surgical micro-robot. In: Int. Conf. on Intelligent Robots and Systems, Las Vegas, USA (2003)
10. Flaschka, H.: The Toda Lattice II, The Existence of Integrals. Phys. Rev. B 9, 1924–1925 (1974)
11. Fillippov, A.: The Versatile Soliton. Birkhauser, Boston (2000)
12. Fomenko, A.T.: Integrability and Nonintegrability in Geometry and Mechanics, Mathematics and its Applications. Kluwer Academic Publishers, Netherlands (1988)
13. Foxman, J.A., Robbins, J.M.: Singularities, Lax degeneracies and Maslov indices of the periodic Toda chain, vol. 1 (November 1, 2004) arXiv:math-ph/0411018
14. Friedland, L., Shagalov, A.G.: Emergence and control of multiphase nonlinear waves by synchronization. Physical Review Letters 90(7), 074101 (2003)
15. Gardner, C.S., Greene, J.M., Kruskal, M.D., Miura, R.M.: Method for solving the Korteweg-de Vries equation. Phys. Rev. Lett. 19, 1095–1097 (1967)
16. Gardner, C.S.: The Korteweg-de Vries Equation and Generalizations IV: The Korteweg-de Vries Equation as a Hamiltonian System. J. Math. Phys. 12, 1548–1551 (1971)
17. Genedelman, O.V., Savin, A.V.: Normal heat conductivity of the one-dimensional with periodic potential of nearest neighbor interaction. Physical Review Letters 84(11), 2381–2384 (2005)
18. Grudinin, A.B., Gray, S.: Passive harmonic mode locking in soliton fiber lasers. Journal of Opt. Soc. Am. B 14(1), 144–154 (1997)
19. Hu, B., Li, B., Zhao, H.: Heat conduction in one-dimensional chains. Physical Review E 57(3), 2992–2995 (1998)
20. Infeld, E., Rowlands, G.: Nonlinear Waves, Solitons and Chaos, 2nd edn. Cambridge University Press, Cambridge (2000)
21. Jovanovic, M.R., Bamieh, B.: Lyapunov-based distributed control of systems on lattices. IEEE Trans. Automat. Contr. 50(4), 422–433 (2005)
22. Khasin, M., Friedland, L.: Multiphase control of a nonlinear lattice. Physical Review E 68, 66214 (2003)
23. Lakshman, M.: Nonlinear dynamics: Integrability, Chaos and Patterns. Springer, Heidelberg (2002)
24. Lakshman, M.: Nonlinear dynamics: Challenges and perspectives, Pramana. Journal of Physics 64(4) (April 2005)

25. Lam, L.: Introduction to Nonlinear Physics. Springer, Heidelberg (2003)
26. Lepria, S., Livib, R., Politib, A.: Thermal conduction in classical low dimensional lattices. Physics Reports 377(1), 1–80 (2003)
27. Lomdahl, P.S., Layne, S.P., Bigio, I.J.: Solitons in Biology. Los Almos Science (Spring 1984)
28. Mellen, N., Kiemel, T., Cohen, A.H.: Correlational analysis of fictive swimming in the lamprey reveals strong functional intersegmental coupling. J. Neurophysiol. 73, 1020–1030 (1995)
29. Miller, J.D., Navaratna, M., Dayawansa, W.P.: Modelling transient dynamics in a small world network of oscillators. In: Proc. of the IEEE Conf. on Decision and Control, Nassau, Bahamas (December 2004)
30. Miranda, R.: Algebraic curves and Riemann surfaces. Graduate Studies in Mathematics, vol. 5, pp. 1–115. American Mathematical Society (1995); La Rivista del Nuovo Cimento 27 (10-11), pp. 1–115 (2004)
31. Mitschke, F.M., Mollenauer, L.F.: Stabilizing the soliton laser. IEEE Journal of Quantum Electronics QE-22(12), 2242–2250 (1986)
32. Newell, A.C.: Solitons in Mathematics and Physics. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia (1985)
33. Mitschke, F.M., Mollenauer, L.F.: The symmetries of solitons. IEEE Journal of Quantum Electronics, Bulletin (New Series) of the American Mathematical Society 34(4), 339–403 (1997)
34. d'Ovidio, F., Bohr, H.G., Lindgård, P.-A.: Solitons on H-bonds in proteins. J. Phys. Condens. Matter 15, s1699–s1707 (2003)
35. Palamakumbura, R., Maithripala, S., Dayawansa, W.P., Inaba, H.: Control of travelling pulses in MEMS arrays: Numerical evidence of practical asymptotic stabilization. In: Proc. of the American Control Conference, Portland (June 2005)
36. Palamakumbura, R., Maithripala, S., Dayawansa, W.P., Inaba, H.: Stabilization of travelling pulses in actuator arrays. In: Electronic proceedings of the ICIA2005 conference, Colombo, Sri Lanka (December 2005)
37. Palamakumbura, R., Maithripala, D.H.S., Holtz, M., Berg, J.M., Dayawansa, W.P.: Induced thermal transport in the Toda Lattice using localized passive control. In: Electronic Proc. of the ICIIS Conference, Peradeniya, Sri Lanka (August 2006)
38. Palamakumbura, R.: Control of Travelling Pulses in Actuator Arrays, Ph.D. Dissertation, Texas Tech. University (May 2006)
39. Ricketts, D.S., Li, X., Ham, D.: Electrical soliton oscillator. IEEE Trans. Microwave Theory and Techniques 50(1), 373–382 (2006)
40. Scott, A.C., Chu, F.Y.F., McLaughlin, D.W.: The soliton: A new concept in applied science. Proc. of the IEEE 61(10) (1973)
41. Scott, A.C.: Davydov's soliton. Physics Reports 217(1), 1–67 (1992)
42. Scott, A.C.: Nonlinear Science: Emergence and Dynamics Of Coherent Structures. Oxford University Press, Oxford (1999)
43. Toda, M.: Theory of Nonlinear Lattices. Springer, Heidelberg (1981)
44. Toda, M.: Nonlinear Lattice and Soliton Theory. IEEE Trans. on Circuits and Systems cas-30(8), 542–554 (1983)
45. Theodorakopoulos, N., Mertens, F.G.: Dynamics of the Toda lattice: A soliton-phonon phase-shift analysis. Physical Review B 28(6), 3512–3519 (1983)
46. Theodorakopoulos, N., Bacalis, N.C.: Thermal Solitons in the Toda chain. Physical Review B 46(17), 10706–10709 (1992)

47. Theodorakopoulos, N., Peyrard, M.: Solitons and Nondissipative Diffusion. Physical Review Letters 83(12), 2293–2296 (1999)
48. Wadati, M.: Introduction to Solitons, Pramana. Journal of Physics, Indian Academy of Sciences 57, 841–847 (2001)
49. Watts, D.J., Strogatz, S.H.: Collective dynamics of small world networks. Nature 393(4), 440–442 (1998)
50. Yoshida, F., Sakuma, T.: Statistical mechanics of the Toda lattice based on soliton dynamics. Physical Review A 25(5), 2750–2762 (1982)

14

Global Asymptotic Controllability of Polynomial Switched Systems and Their Switching Laws

P.C. Perera

Dept. of Engineering Mathematics, Faculty of Engineering, Univ. of Peradeniya, Peradeniya, Sri Lanka

Summary. A fundamental requirement for the design of feedback control systems is the knowledge of structural properties of the plant under consideration. These properties are closely related to the generic properties such as controllability. A sufficient condition for controllability of polynomial switched systems is established here. Control design for nonlinear switched control systems is known to be a nontrivial problem. In this endeavor, an indirect approach is taken to resolve the control design problem pertaining to polynomial switched systems satisfying the aforementioned sufficient condition for controllability; it is shown that trajectories of a related controllable polynomial system can be approximated arbitrarily closely by those of the polynomial switched control system of our interest.

14.1 Introduction

A switched system is a hybrid dynamical system consisting of a family of continuous-time subsystems and a rule that describes how the subsystems switch among them. Many such systems encountered in actual practice exhibit switching among several subsystems which depends on various physical phenomena. In recent years, much research has focused on switched control systems. In fact, switched control systems deserve investigation for both theoretical and practical reasons. Switching among different system structures is an essential feature of many engineering applications such as power systems as investigated in [10]. Furthermore, switched control systems have numerous applications in control of mechanical systems such as aircrafts and satellites [4]. Above all, the control techniques obtained by switching among different systems provide improved performance over a fixed controller [5, 6].

In many engineering problems, restrictions on the switching signal cannot be specified a priori. Also, the need for supervisory switching arises for many reasons. One of them is that many physical systems can be represented by switching or interpolating between locally valid models and controllers that encompass switching are often a natural method for dealing with such systems. Another important motivation for designing switching control strategies is to ensure robust control performance in the presence of component failure. For example, if an operating condition changes (a sensor failure, a change in sampling rate, or even

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 239–251.
springerlink.com

a controller failure), then a more appropriate control action may be initiated by the supervisor. Most importantly, switching between number of control structures automatically results in control systems that are no longer constrained by the limitations of linear design and it is therefore not surprising that switching based control strategies can result in algorithms that offer significant performance improvements over traditional linear control.

Switching among different system structures is an essential feature of many engineering applications such as power systems[10]. A system of the above type can be modelled as a switched control system which is a hybrid dynamical system consisting of a family of subsystems and a rule that describes how the subsystems switch among them. Such switched systems are called supervisory-based switched control systems. Switching based control strategies employed in a switched system can result in algorithms that offer significant performance improvements over traditional linear control.

A fundamental requirement for the design of feedback control systems is the knowledge of the structural properties of the switched control system under consideration. These properties are closely related to the concepts of controllability, observability, stability and stabilizability. The main aim of this endeavor is to present an indirect route to the controllability problem and show that a given switched system consisting of time-invariant polynomial subsystems satisfying a certain condition, it is possible to construct a controllable non-switched polynomial system in such a way that the trajectories of the latter may be arbitrarily well approximated (in the sense of C^∞ norm on finite time intervals) by the trajectories of the polynomial switched system. Since much is known about homogeneous polynomial systems vis-a-vis their controllability, constructing control inputs, and stabilizability, this technique provides a theoretical avenue to study finer details of control and stabilization of switched control systems consisting of time-invariant polynomial subsystems.

In section 14.2, our attention is focused to the definitions and preliminaries. First, the general form of the polynomial subsystems of the switched control system of our interest, is presented. Then, the Accessibility and the Strong Accessibility Lie Algebras in the context of nonlinear systems are defined. Using the Strong Accessibility Lie Algebra of each of the polynomial subsystems, a sequence of Lie Algebras is recursively obtained. The notations pertaining to the above sequence are also given in this section.

In section 14.3, a continuous-time time-invariant polynomial system associated to the polynomial switched system is constructed. It should be noted that this associated polynomial system is not unique. Then, a sufficient condition for the global controllability of the aforementioned associated polynomial system is established. Since the polynomial system is globally controllable under the sufficient condition established, there exist control inputs using which the state can be steered between any two arbitrary points in the state space. The control inputs for the associated homogeneous polynomial system are known. A similar work pertaining to linear counterpart was attempted in [7].

Then, in section 14.4, the attention is focused in approximating the trajectories of the associated polynomial system arbitrarily closely by those of the polynomial switched system of our interest. Moreover, this approximation can be used to deduce the switching strategy for controlling a nonlinear switched system consisting of time-invariant polynomial subsystems.

14.2 Definitions and Preliminaries

Notation 14.2.1. *Let the set consisting of the first positive n integers be denoted by* $\underline{n}$. *For the sake of brevity, this notation is used in this article throughout.*

Here, our attention is focused to the nonlinear switched system consisting of subsystems of the form

$$\dot{x} = f_i(x) + \sum_{j=1}^{m} b_j u_j \quad \text{for} \quad i \in \underline{k}, \tag{14.1}$$

where $x, b_j \in \mathbb{R}^n$, $u_j \in \mathbb{R}$ and $f_i : \mathbb{R}^n \to \mathbb{R}^n$ is a polynomial vector field for every $i \in \underline{k}$. Without loss of generality, it can be assumed that $b_j = e_j$ for all $j \in \underline{m}$.

Definition 14.2.1. *The accessibility Lie algebra* $\mathcal{A}$ *of the system (14.1) is the smallest subalgebra that contains* $\{f, b_1, ..., b_m\}$. *We write this as*

$$\mathcal{A} = \{f, b_1, ..., b_m\}_{LA}.$$

Definition 14.2.2. *The* **strong accessibility Lie algebra** $\mathcal{S}$ *of the system (14.1) is the smallest subalgebra of* $\mathcal{A}$ *that contains* $\{b_1, ..., b_m\}$ *satisfying* $[f, X] \in \mathcal{S}$ *for all* $X \in \mathcal{S}$. *In other words,* $\mathcal{S}$ *is the smallest Lie ideal in* $\mathcal{A}$ *containing* $\{b_1, ..., b_m\}$. *We write this as*

$$\mathcal{S} = \{ad_f^s b_p | p \in \underline{m}, s \geq 0\}_{LA}.$$

For the sake of notational simplicity, in this endeavor, $\mathcal{S}$ *is expressed as*

$$\mathcal{S} = SAL(\{f, b_1, ..., b_m\}).$$

If $dim(\mathcal{S})|_{x_0} = n$ *for some* $x_0 \in M$, *then the system (14.1) is said to be* **locally strongly accessible** *from* x_0. *This means for any neighborhood* V *of* x_0 *and* $T > 0$, *the set* $R^V(x_0, T)$ *contains a nonempty open set of* M *for any* $T > 0$ *sufficiently small. If* $dim(\mathcal{S}) = n$ *for all* $x \in M$ *then the system is said to be* **locally strongly accessible**.

In one of the landmark papers [3] in nonlinear control theory, the following sufficient condition on controllability of a homogeneous systems is established.

Theorem 14.2.1. *Suppose* $M = \mathbb{R}^n$ *and* $B = \{b_1, ..., b_m\}$ *with* $b_l \in \mathbb{R}^n$ *for* $(l \in \underline{m})$ *and* $f(x)$ *is a polynomial vector field of odd degree, i.e.,*

$$f(x) = \sum_{|j|=2p+1} a_j x^j,$$

where $p \in \mathbb{N}$, $a_j, x \in \mathbb{R}^n$ *for all* j *and* $j = (j_1, ..., j_n)$ *is a multi-index. Then the system* $\sum(f, B)$ *given by*

$$\sum(f, B) : \dot{x} = f(x) + \sum_{i=1}^{m} b_i(x)u_i, u = (u_1, ..., u_m)^T \in \mathbb{R}^m$$

is controllable on $\mathbb{R}^n$ *if*

$$rank(L) = n$$

where $L \subseteq \mathbb{R}^n$ *is the set of constant vector fields of strong accessibility Lie algebra.*

It can be noticed that this condition is equivalent to the full rank condition of the strong accessibility Lie algebra at the origin as in the case of linear systems. Then the $\pm$ symmetry of the system aids time reversibility in exactly the same way as it aids a linear system.

Notation 14.2.2. *Let the set which consists of the drift vector field* f_i *and the input vector fields* b_j *for* $j \in \underline{m}$ *of the* i^{th} *subsystem of the nonlinear switched system given in (14.1), be denoted by* S_i. *Thus,*

$$S_i = \{f_i, b_1, ..., b_m\} \text{ for } i \in \underline{k}.$$

Also, let the Strong Accessibility Lie Algebra of the i^{th} *subsystem be denoted by* $\mathcal{S}_i^{(0)}$. *Thus,*

$$\mathcal{S}_i^{(0)} = SAL(S_i) \text{ for } i \in \underline{k}.$$

Definition 14.2.3. *Define the subspace* $\mathcal{D}_0$ *of* $\mathbb{R}^n$ *by* $\mathcal{D}_0 = \bigoplus_{i=1}^{k} span\left\{\mathcal{S}_i^{(0)}\right\}$. *Let the constant vector fields in* $\bigcup_{i=1}^{k} \mathcal{S}_i^{(0)}$ *be* L_0. *Also let* $\mathcal{S}_i^{(1)} = SAL(S_i \cup L_0)$ *for* $i \in \underline{k}$. *Define* $\mathcal{D}_1 = \bigoplus_{i=1}^{k} span\left\{\mathcal{S}_i^{(1)}\right\}$.

Let the constant vector fields in $\bigcup_{i=1}^{k} \mathcal{S}_i^{(1)}$ *be* L_1. *Then, iteratively,*

$$\mathcal{D}_l = \bigoplus_{i=1}^{k} span\left\{\mathcal{S}_i^{(l)}\right\} \text{ for } l = 0, 1, 2, ...$$

where $\mathcal{S}_i^{(l)} = SAL(S_i \cup L_{l-1})$.

It is obvious that $\{\dim(\mathcal{D}_l)\}_{l=0}^{\infty}$, is a nondecreasing sequence of positive integers. Moreover, n is an upper bound for this sequence for any nonlinear switched system. Moreover, if $\dim\left(\mathcal{D}_{\hat{l}+1}\right) = \dim\left(\mathcal{D}_{\hat{l}}\right) = \hat{n} \in \mathbb{N}$ for some $\hat{l} \in \mathbb{N}$, it follows that $\dim(\mathcal{D}_l) = \hat{n}$ for all $l \geq \hat{l}$. Within the context of what is to follow, it is worth noticing that $\mathcal{D}_n$ consists of constant vector fields of the Lie algebra generated by the family of vector fields of the polynomial switched system.

In the next section, the time-invariant non-switched polynomial system related to the nonlinear switched system given in (14.1) is constructed. The global controllability of the polynomial system is also established.

14.3 Controllability of Related Polynomial System

Focus the attention to the collection of subsystems of the form

$$\dot{x} = f_i(x) + \sum_{j=1}^{m} b_j u_j \quad \text{for} \quad i \in \underline{k},$$

of the switched control system under consideration. In subsection 14.3.1, a related non-switched polynomial system to the polynomial switched system of our interest, is constructed. A sufficient condition on controllability of the polynomial system is established in subsection 14.3.2.

14.3.1 Construction of Related Polynomial System

Here, it is attempted to construct a polynomial system of the form

$$\dot{x} = \sum_{i=1}^{k} p_i(x) f_i(x) + \sum_{j=1}^{m} b_j u_j,$$

where

$$p_i(x) = \prod_{j=1}^{m} x_j^{\beta_i^{(j)}}$$

with $\beta_i^{(j)}$ is nonnegative and even for all $i \in \underline{k}$ and $j \in \underline{m}$.

It is of paramount importance to notice that $p_i(x)$ $(i \in \underline{k})$ are state feedback functions with the property that each is positive semi-definite and

$$\sum_{i=1}^{k} p_i(x) > 0 \text{ for } x \in \mathbb{R}\backslash\{0\}.$$

The latter condition is needed to well approximate trajectories of the related polynomial system arbitrarily closely by those of the switched system consisting of time-invariant polynomial subsystems satisfying a certain rank condition.

In this endeavor, it is required to obtain the general form of the drift vector fields $f_i : \mathbb{R}^n \to \mathbb{R}^n$ $(i \in \underline{k})$ of the subsystems of the switched system given in (14.1). Since $f_i : \mathbb{R}^n \to \mathbb{R}^n$ for $i \in \underline{k}$, it yields that

$$f_i(x) = [f_{(i,1)}, f_{(i,2)}, ..., f_{(i,q)}, ..., f_{(i,n)}]^T \text{ for all } i \in \underline{k}, \tag{14.2}$$

where

$$f_{(i,q)}(x) = \sum_{r=1}^{J_{i,q}} p_{i,q,r}(x) \text{ for } i \in \underline{k} \text{ and } q \in \underline{n}, \tag{14.3}$$

with

$$p_{i,q,r}(x_1, x_2, ..., x_n) = \gamma_{iqr} x_1^{\alpha_{iqr}^{(1)}} x_2^{\alpha_{iqr}^{(2)}} ... \, x_n^{\alpha_{iqr}^{(n)}}, \tag{14.4}$$

where $\gamma_{iqr} \in \mathbb{R}^n$. The indices $\alpha_{iqr}^{(\hat{r})}$ are such that

$$\sum_{\hat{r}=1}^{n} \alpha_{iqr}^{(\hat{r})}, \tag{14.5}$$

is odd for all $i \in \underline{k}$ and $r \in \underline{J_{i,q}}$ with $q > m$.

Let $\{q_1, q_2, ..., q_{n-m}\}$ be a permutation of $\{m+1, m+2, ..., n\}$. Suppose there exist $q_1 \in \underline{n} \backslash \underline{m}$, $i_1 \in \underline{k}$ and $r \in \underline{J_{i_1,q_1}}$ for which

$$ad_{e_m}^{\alpha_{i_1 q_1 r}^{(m)}} ad_{e_{m-1}}^{\alpha_{i_1 q_1 r}^{(m-1)}} ... \, ad_{e_1}^{\alpha_{i_1 q_1 r}^{(1)}} \, p_{i_1,q_1,r}(x)$$

is a constant vector field in $\mathbb{R}^n$.

Also, suppose there exist $q_2 \in \{\underline{n} \backslash \underline{m}\} \setminus \{q_1\}$, $i_2 \in \underline{k}$ and $r \in \underline{J_{i_2,q_2}}$ for which

$$ad_{e_{q_1}}^{\alpha_{i_2 q_2 r}^{(q_1)}} ad_{e_m}^{\alpha_{i_2 q_2 r}^{(m)}} ad_{e_{m-1}}^{\alpha_{i_2 q_2 r}^{(m-1)}} ... \, ad_{e_1}^{\alpha_{i_2 q_2 r}^{(1)}} \, p_{i_2,q_2,r}(x)$$

is a constant vector field in $\mathbb{R}^n$.

Iteratively, suppose there exist $q_j \in \{\underline{n} \backslash \underline{m}\} \setminus \{q_1, ..., q_{j-1}\}$, $i_j \in \underline{k}$ and $r \in \underline{J_{i_j,q_j}}$ for which

$$\begin{aligned} ad_{e_{q_{j-1}}}^{\alpha_{i_j q_j r}^{(q_{j-1})}} ad_{e_{q_1}}^{\alpha_{i_j q_j r}^{(q_1)}} ad_{e_m}^{\alpha_{i_j q_j r}^{(m)}} ad_{e_{m-1}}^{\alpha_{i_j q_j r}^{(m-1)}} ... \\ ... \, ad_{e_1}^{\alpha_{i_j q_j r}^{(1)}} \, \left[p_{i_j,q_j,r}(x)\right] \end{aligned} \tag{14.6}$$

is a constant vector field in $\mathbb{R}^n$ for $j \in \underline{n-m}$.

In the following theorem, a related non-switched polynomial system of the switched system consisting of the subsystems given in (14.1) is constructed while establishing its global controllability.

14.3.2 Controllability of Related Polynomial System

The following theorem asserts that if the subsystems of the polynomial switched system satisfies a certain rank condition, the related non-switched polynomial system can be constructed in such a way that it is globally controllable.

Theorem 14.3.1. *Suppose a nonlinear switched control system consisting of subsystems of the form given in (14.1) has drift vector fields which are in the form given in (14.2), (14.3), (14.4) and (14.5) satisfy the condition (14.6). Then there exist a set of even degree polynomials $p_i(x)$ such that*

$$\dot{x} = \sum_{i=1}^{k} p_i(x) f_i(x) + \sum_{j=1}^{m} b_j u_j \tag{14.7}$$

is globally controllable in $\mathbb{R}^n$.

Proof. Let

$$p_i(x) = \prod_{j=1}^{m} x_j^{\beta_i^{(j)}}$$

with $\beta_i^{(j)}$ is even for all $i \in \underline{k}$ and $j \in \underline{m}$. Since, $\{b_1, ..., b_m\} \subset \bigcup_{i=1}^{k} \mathcal{S}_i^{(0)}$, we have

$$m = dim\left(span\left\{b_1, ..., b_m\right\}\right) \leq dim\left(\mathcal{D}_0\right).$$

Since $f_i(x)$ for $i \in \underline{k}$ are in the form given by (14.2), (14.3) and (14.4) and satisfy (14.6), it follows that

$$ad_{e_m}^{\alpha_{i_1 q_1 r}^{(m)}} ad_{e_{m-1}}^{\alpha_{i_1 q_1 r}^{(m-1)}} ... \, ad_{e_1}^{\alpha_{i_1 q_1 r}^{(1)}} \, p_{i_1, q_1, r}(x)$$

is a constant vector field in $\mathbb{R}^n$ for some pair of integers i_1 and q_1 such that $q_1 \in \underline{n} \backslash \underline{m}$ and $i_1 \in \underline{k}$, straightforward calculations yield that

$$\begin{aligned} &ad_{e_m}^{\alpha_{i_1 q_1 r}^{(m)} + \beta_{i_1}^{(m)}} ad_{e_{m-1}}^{\alpha_{i_1 q_1 r}^{(m-1)} + \beta_{i_1}^{(m-1)}} ... \\ &\quad ... \, ad_{e_1}^{\alpha_{i_1 q_1 r}^{(1)} + \beta_{i_1}^{(1)}} \left(p_{i_1}(x) p_{i_1, q_1, r}(x)\right) = \gamma e_{q_1} \end{aligned}$$

is a constant vector field in $\mathbb{R}^n$ for the pair of integers i_1 and q_1 such that $q_1 \in \underline{n} \backslash \underline{m}$ and $i_1 \in \underline{k}$ where $\gamma \in \mathbb{R}$.

Since, $\{b_1, ..., b_m, e_{q_1}\} \subset \bigcup_{i=1}^{k} \mathcal{S}_i^{(1)}$, we have

$$dim\left(span\left\{b_1, ..., b_m, e_{q_1}\right\}\right) = m + 1 \leq dim\left(\mathcal{D}_1\right).$$

Recursively, after $n - m$ steps, it follows that

$$\begin{aligned} &dim\left(span\left\{b_1, ..., b_m, e_{q_1}, e_{q_2}, ..., e_{q_{n-m}}\right\}\right) \\ &\qquad = m + (n - m) \leq dim\left(\mathcal{D}_{n-m}\right). \end{aligned}$$

Thus,

$$\begin{aligned} &dim\left(span\left\{b_1, ..., b_m, e_{q_1}, e_{q_2}, ..., e_{q_{n-m}}\right\}\right) \\ &\qquad = n \leq dim\left(\mathcal{D}_{n-m}\right) \leq dim\left(\mathbb{R}^n\right) = n. \end{aligned}$$

Thus, it follows that

$$span\left\{b_1, ..., b_m, e_{q_1}, e_{q_2}, ..., e_{q_{n-m}}\right\} = \mathbb{R}^n.$$

Since the constant vector fields in the Strong Accessibility Lie Algebra of the system given in (14.7) span the entire state space, by theorem 14.2.1, it turns out that (14.7) is globally controllable in $\mathbb{R}^n$. ■

Remark 14.3.1. Every related non-switched polynomial system of the nonlinear switched system consisting of the subsystems given in (14.1), has the form of (14.7).

Example 14.3.1. Consider the polynomial switched system consisting of subsystems SU_1, SU_2, SU_3 and SU_4 given by

$$SU_1 : \begin{aligned} \dot{x}_1 &= 4x_2x_3 - x_5^4 + u \\ \dot{x}_2 &= x_4^3 \\ \dot{x}_3 &= x_1x_5^2 \\ \dot{x}_4 &= -x_2x_3^2 \\ \dot{x}_5 &= 0 \end{aligned} \tag{14.8}$$

$$SU_2 : \begin{aligned} \dot{x}_1 &= 2x_2^3 + 6x_4 + u \\ \dot{x}_2 &= x_4^5 \\ \dot{x}_3 &= x_1^4x_2 - 2x_4^3x_5^2 \\ \dot{x}_4 &= 0 \\ \dot{x}_5 &= x_1x_2x_3^2x_5 \end{aligned} \tag{14.9}$$

$$SU_3 : \begin{aligned} \dot{x}_1 &= 3x_3^3 + u \\ \dot{x}_2 &= 2x_1^3 + x_2x_5^2 \\ \dot{x}_3 &= 0 \\ \dot{x}_4 &= 4x_2x_4x_5 - 5x_5^3 \\ \dot{x}_5 &= x_3^2x_4 \end{aligned} \tag{14.10}$$

$$SU_4 : \begin{aligned} \dot{x}_1 &= 5x_3x_4 + x_3x_5^4 + u \\ \dot{x}_2 &= x_3x_4^2 + 4x_5^3 \\ \dot{x}_3 &= 0 \\ \dot{x}_4 &= 6x_1x_4x_5 \\ \dot{x}_5 &= 0 \end{aligned} \tag{14.11}$$

Straightforward calculations yield that $e_2 \in \mathcal{S}_3^{(0)}$. Thus, $\mathcal{D}_0 = span\{e_1, e_2\}$. Moreover, $L_0 = \{e_1, e_2\}$. Then, it follows that $e_3 \in \mathcal{S}_2^{(1)}$. Thus, it yields $\mathcal{D}_1 = span\{e_1, e_2, e_3\}$ and $L_1 = \{e_1, e_2, e_3\}$. By following similar steps, we get $e_4 \in \mathcal{S}_1^{(2)}$, $\mathcal{D}_2 = span\{e_1, e_2, e_3, e_4\}$, $L_2 = \{e_1, e_2, e_3, e_4\}$ and $e_5 \in \mathcal{S}_3^{(3)}$, $\mathcal{D}_3 = span\{e_1, e_2, e_3, e_4, e_5\} = \mathbb{R}^5$, $L_3 = \{e_1, e_2, e_3, e_4, e_5\}$.

Let $SU_i : \dot{x} = f_i(x) + u$ for $i \in \{1, 2, 3, 4\}$. Choosing $p_1(x) = 0.01x_1^2$, $p_2(x) = 1$, $p_3(x) = 0.07x_1^2$ and $p_4(x) = 0$, it yields that

$$\dot{x} = \sum_{i=1}^{4} p_i(x) f_i(x) + [1\ 0\ 0\ 0\ 0]^t u$$

is globally controllable. ■

14.4 Asymptotic Global Controllability of a Polynomial Switched Systems

To generate the control signals for a polynomial switched system with drift vector fields in the form (14.2), (14.3), (14.4) and (14.5) satisfying the condition (14.6), it is required to approximate the trajectories of the associated polynomial system arbitrarily closely by those of the switched system of our interest. This phenomenon is called the global asymptotic controllability.

The following theorem implies that the above arbitrary close approximation is, in fact, possible.

Theorem 14.4.1. *Let $p_i : \mathbb{R}^n \to [0, \infty)$ and $u_i : \mathbb{R} \to \mathbb{R}^m$ be smooth functions. Define the time varying vector field Z on $\mathbb{R}^n$ as*

$$Z(x,t) = \sum_{i=1}^{k} p_i(x) f_i x + b_1 u_1(t) + b_2 u_2 + \cdots b_m u_m.$$

where f_i for $i \in \underline{k}$ and b_j for $j \in \underline{m}$ are as above. Then, for each $a \in \mathbb{R}^n$ and each $T > 0$ for which $\phi_T^Z(a)$ is defined, $\phi_T^Z(a)$ is in the closure of the reachable set of the switching system $\dot{x} = f_i x + \sum_{j=1}^{m} b_j u_j,\ i \in \underline{k}$ from the initial state a.

The following technical lemmas are required for the proof of the above theorem.

Lemma 14.4.1. *(Convergence of the Euler Approximation): Let Z be a smooth vector field on $\mathbb{R}^n$ and let $a, b \in \mathbb{R}^n$ and $T > 0$ are related by $b = \phi_T^Z(a)$. For each $N \in \mathbb{N}$, define a sequence of points $\left\{b_j^N\right\}_{j=0}^{N}$ by $b_0^N = a$ and $b_{j+1}^N = b_j^N + \frac{T}{N} Z\left(b_j^N\right)$ for $j \in \underline{N-1}$. Then for each $\epsilon > 0$ there exists $\hat{N}$ such that,*

$$\left\| b_j^N - \phi_{\frac{jT}{N}}^Z(a) \right\| < \epsilon,$$

for all $N > \hat{N}$ and for all $j \in \underline{N}$. (See [8]).

Now suppose $X_1, \cdots, X_k$ be smooth vector fields on $\mathbb{R}^n$, and let $Z = \sum_{i=1}^{k} X_i$. Let $a, b \in \mathbb{R}^n$ and $T > 0$ are related by $b = \phi_T^Z(a)$, and for each $N \in \mathbb{N}$ define a sequence of points $\left\{b_j^N\right\}_{j=0}^{N}$ as in the previous lemma. Furthermore, define points $\{y_{i,j}^N\}_{0 \le i \le k,\ 0 \le j \le N}$ and $\{x_{i,j}^N\}_{0 \le i \le k,\ 0 \le j \le N}$ by,

$$\begin{aligned} y^N_{i,j} &= b^N_j \ + \ \tfrac{T}{kN}\textstyle\sum_{l=1}^{i} X_l(b^N_j), \ \ 0 \le i \le k, \ 0 \le j \le N \\ x^N_{0,0} &= a, \\ x^N_{0,j} &= x^N_{k,j-1}, \ 0 \le j \le N \\ x^N_{i,j} &= x^N_{i-1,j} + \tfrac{T}{kN} X_i(x^N_{i-1,j}), \ \ 0 \le i \le k, \ 1 \le j \le N. \end{aligned}$$

The sequence of points $\{x^N_{i,j}\}$ corresponds to a concatenation of the Euler approximations of the flows of vector fields $X_1, \cdots, X_k$ in sequence. The next lemma states that for large N the concatenated Euler sequence converges to the standard Euler sequence of $\sum_{i=1}^{k} X_i$.

Lemma 14.4.2. *For each ϵ there exists $\hat{N}$ such that $\left\|x^N_{i,j} - y^N_{i,j}\right\| < \epsilon$, $i \in \underline{k}$, $j \in \underline{N}$ and $N > \hat{N}$. In particular, $\left\|x^N_{k,N} - b\right\| < \epsilon$ for sufficiently large N.*

Proof. The second assertion follows trivially from the first assertion and the convergence of the Euler approximation. Let us now proceed to prove the first assertion.

Let us fix a compact neighborhood V of the set $\{\phi^X_t(a) | 0 \le t \le T\}$, and fix a constant M such that $\|X_i(x)\| < M$ and $\|DX_i(x)\| < M$ for all $x \in V$ and $i = 1, ..., k$. In the remainder of the proof N will be assumed to be large enough such that $\{y^N_{i,j}\}$ sequence is contained in V. Existence of such N follows from the convergence of the Euler approximations.

Let us define $\alpha^N_{i,j} = x^N_{i,j} - y^N_{i,j}$. Let us observe that,

$$\begin{aligned} x^N_{i,j+1} &= x^N_{i,j} + \frac{T}{kN} X_{i+1}\left(x^N_{i,j}\right) \\ &= y^N_{i,j} + \alpha^N_{i,j} + \frac{T}{kN} X_{i+1}\left(y^N_{i,j} + \alpha^N_{i,j}\right) \\ &= y^N_{i,j} + \alpha^N_{i,j} + \frac{T}{kN} X_{i+1}\left[b^N_j + \ \frac{T}{kN}\sum_{l=1}^{i} X_l\left(b^N_j\right) + \alpha^N_{i,j}\right] \\ &= y^N_{i,j} + \alpha^N_{i,j} + \frac{T}{kN} X_{i+1}\left(b^N_j\right) + \frac{T}{kN}\left[\int_0^1 DX_{i+1}\left\{b^N_j + \right.\right. \\ &\qquad \left.\left. \theta\frac{T}{kN}\sum_{l=1}^{i} X_l\left(b^N_j\right) + \theta\alpha^N_{i,j}\right\} d\theta\right] \times \left\{\frac{T}{kN}\sum_{l=1}^{i} X_l\left(b^N_j\right) + \alpha^N_{i,j}\right\}. \end{aligned}$$

Inductively suppose that $\{x^N_{i,j}\}$ is in V. From the last line of the previous equality it now follows that

$$\|\alpha^N_{i+1,j}\| \le \|\alpha^N_{i,j}\| + \frac{TM}{kN}\left(i\frac{TM}{kN}\right) + \frac{TM}{kN}\|\alpha^N_{i,j}\|.$$

From this recursive inequality we obtain,

$$\|\alpha^N_{k,j}\| \le \left(1 + \frac{TM}{kN}\right)^k \|\alpha^N_{0,j}\| + k^2\left(1 + \frac{TM}{kN}\right)^{k-1}\left(\frac{TM}{kN}\right)^2. \tag{14.12}$$

Since $\alpha_{k,j}^N = \alpha_{0,j+1}^N$ the inequality (14.12) can be restated as,

$$\|\alpha_{0,j+1}^N\| \leq \left(1+\frac{TM}{kN}\right)^k \|\alpha_{0,j}^N\| + k^2 \left(1+\frac{TM}{kN}\right)^{k-1} \left(\frac{TM}{kN}\right)^2. \tag{14.13}$$

Noting that $\alpha_{0,0}^N = x_{0,0}^N - y_{0,0}^N = a - a = 0$, from this recursive inequality we obtain,

$$\begin{aligned}
\|\alpha_{0,j}^N\| &\leq k^2 \left(1+\frac{TM}{kN}\right)^{k-1} \left(\frac{TM}{kN}\right)^2 \left[\sum_{l=1}^{j} \left(1+\frac{TM}{kN}\right)^{k(l-1)}\right] \\
&= k^2 \left(1+\frac{TM}{kN}\right)^{k-1} \left(\frac{TM}{kN}\right)^2 \left\{\frac{\left(1+\frac{TM}{kN}\right)^{kj}-1}{\left(1+\frac{TM}{kN}\right)^{k}-1}\right\} \\
&= k^2 \left(1+\frac{TM}{kN}\right)^{k-1} \left(\frac{TM}{kN}\right)^2 \left\{\frac{\left(1+\frac{TM}{kN}\right)^{kj}-1}{k\frac{TM}{kN}}\right\} \\
&\leq \frac{TM}{N} \left(1+\frac{TM}{kN}\right)^{k-1} \left(1+\frac{TM}{kN}\right)^{kj} \\
&\leq \frac{TM}{N} e^{\frac{TM}{N}} \left(1+\frac{TM}{kN}\right)^{kj}
\end{aligned}$$

Since

$$\lim_{N\to\infty} \left(1+\frac{TM}{kN}\right)^{kN} = e^{TM},$$

it follows that there exists a constant $L > 0$ which is independent of N such that,

$$\|\alpha_{0,j}^N\| \leq \frac{L}{N} \; for \; 0 \leq j \leq N. \tag{14.14}$$

From inequalities (14.12) and (14.14), and from the inductive hypothesis, it now follows that the sequence $\left\{x_{ij}^N\right\}$ remains in V for large enough N, and in addition, the first assertion of the lemma is true. This concludes the proof of the lemma. ■

Next technical lemma is needed to show that one may absorb functions p_i $(i \in \underline{k})$ in the related non-switched polynomial system into the time parameters between switching in the nonlinear switched system.

Notation 14.4.1. *When W_j for $j \in \underline{N}$ are noncommutative variables, let the product $W_1W_2...W_N$ be denoted by $\prod_{j=1}^{N} W_j$.*

Lemma 14.4.3. *Let X_i $(i \in \underline{k})$ be smooth vector fields on $\mathbb{R}^n$ and α_i $(i \in \underline{k})$ be nonnegative valued smooth functions on $\mathbb{R}^n$. Let $a, b \in \mathbb{R}^n$ and $T > 0$ are related by $b = \phi_T^{\sum_{i=1}^k \alpha_i X_i}(a)$. Then for all $\epsilon > 0$ there exists $\hat{N} \in \mathbb{N}$ and $\{\tau_{i,j}^N\}_{1 \le i \le k,\ 1 \le j \le N} \subset [0, \infty)$ such that*

$$\left\| \prod_{j=1}^{N} \left\{ \prod_{i=1}^{k} \phi_{\tau_{i,j}^N}^{X_i} \right\} (a) - b \right\| < \epsilon,$$

for all $N > \hat{N}$.

Proof. The conclusion directly follows from lemma 14.4.1 and lemma 14.4.2. ∎

Proof of Theorem 14.4.1: From Lemma 14.4.3 it follows that there exists $\hat{N} \in \mathbb{N}$ such that

$$\prod_{j=1}^{N} (\prod_{k=1}^{m} \phi_{T/N}^{\alpha_k Y_k})(a) - b)\| < \epsilon$$

for all $N > \hat{N}$. Since α_i are scalars, it is always possible to re scale time and represent the scaling effect of α_i, and hence possible to restate this inequality in the form stated in the Lemma 14.4.3. This concludes the proof. ∎

Suppose a nonlinear switched control system consisting of subsystems of the form given in (14.1) has drift vector fields which are in the form given in (14.2), (14.3), (14.4) and (14.5) satisfy the condition (14.6). Then, as in theorem 14.3.1, it can be chosen polynomials $p_i(x)$ $(i \in \underline{k})$ such that the homogeneous system given in (14.7) is globally controllable. Now, suppose the state $x \in \mathbb{R}^n$ is needed to be steered to an arbitrary point $\hat{x} \in \mathbb{R}^n$ in the state space. First, the state is steered from x to $\hat{x}$ by means of the associated homogeneous system. The switching signals for controlling a homogeneous system are known. This paves the way to use the results established in lemma 14.4.1, lemma 14.4.2 and theorem 14.4.1, to show that the switching signals for the time-invariant polynomial system satisfying a certain rank condition can be devised in steering the state between any two arbitrary points $x \in \mathbb{R}^n$ and $\hat{x} \in \mathbb{R}^n$ by approximating the trajectories of the related time-invariant non-switched homogeneous system with those of the aforementioned switched system. It is worth recalling that the feedback functions $p_i(x)$ $(i \in \underline{k})$ employed in constructing the related polynomial should be selected subject to the restrictions that each is positive semi-definite and $\sum_{i=1}^{k} > 0$ on $x \in \mathbb{R} \backslash \{0\}$. This restriction is required in approximating the trajectories of the switched system consisting of polynomial subsystems arbitrarily closely (in the sense of C^∞ norm) by those of the related non-switched homogeneous system.

As noted earlier, this will provide a theoretical avenue to analyze the properties such as controllability, observability, stability, and stabilizability which are closely related to the structural properties of the class of switched systems of our interest.

References

1. Ezzine, J., Haddad, A.H.: Controllability and observability of hybrid systems. Int. J. of Control 49(6), 2045–2055 (1989)
2. Hermann, R.: On the accessibility problem in control theory. In: Int. Symp. on Nonlinear Differential Equations and Nonlinear Mechanics, pp. 325–332. Academic, New York (1963)
3. Jurdjevic, V., Kupka, I.: Polynomial control systems. Mathematische Annalen 272(3), 361–368 (1985)
4. Li, Z.G., Wen, C.Y., Soh, Y.C.: Switched controllers and their applications in bilinear systems. Automatica 37(3), 477–481 (2001)
5. Morse, A.S.: Supervisory control of famillies of linear set-point controllers, Part 1: Exact matching. IEEE Trans. on Automat. 41(10), 1413–1431 (1996)
6. Narendra, K.S., Balakrishnan, J.: Adaptive control using multiple models. IEEE Trans. on Automat. Contr. 42(2), 171–187 (1997)
7. Perera, P.C., Dayawansa, W.P.: Asymptotic feedback controllability of switched control systems. In: Proc. of the American Control Conference, pp. 239–244 (2004)
8. Stoer, J., Bulrisch, R.: Introduction to Numerical Analysis. Springer, New York (1980)
9. Szigeti, F.: A differential-algebraic condition for controllability and observability of time varying linear systems. In: Proc. of the IEEE Conf. on Decision and Control, pp. 3088–3090 (1992)
10. Williams, S.M., Hoft, R.G.: Adaptive frequency domain control of PPM switched power line conditioner. IEEE Trans. on Power Electronics 6(4), 665–670 (1991)
11. Zhendong, S., Ge, S.S., Lee, T.H.: Controllability and reachability criteria for switched linear systems. Automatica 38, 775–786 (2002)

15

Semi-global Output Feedback Stabilization of a Class of Non-uniformly Observable and Non-smoothly Stabilizable Systems

Bo Yang[1] and Wei Lin[2]

[1] Dept. of Mathematics and Statistics, Texas Tech Univ., Lubbock, TX 79409, USA
[2] Dept. of Electrical Engineering and Computer Science, Case Western Reserve Univ., Cleveland, OH 44106, and affiliated with HIT Graduate School, Shenzhen, China

In memory of our friend and teacher Dayawansa

Summary. The problem of semi-global stabilization by output feedback is investigated for nonlinear systems which are non-uniformly observable and non-smoothly stabilizable. Previously, it was shown that under certain restrictive conditions, global stabilization of such nonlinear systems was achievable by nonsmooth output feedback. The main contribution of this paper is to prove that in the context of semi-global control, most of the restrictive growth conditions required in the previous work can be relaxed or removed. In particular, we show that without imposing any growth condition, it is possible to achieve semi-global stabilization by nonsmooth output feedback for a chain of odd power integrators perturbed by a triangular vector field — a significant class of nonlinear systems that are known difficult to be controlled via output feedback, due to the lack of smooth stabilizability and uniform observability. Extensions to non-strictly triangular systems are also discussed in the two-dimensional case. Examples are given to demonstrate the key features of the new semi-global output feedback control schemes.

15.1 Introduction

The purpose of this paper is to address the problem of semi-global stabilization by output feedback, for a class of highly nonlinear systems that are neither uniformly observable nor smoothly stabilizable. Specifically, we are interested in the question of when semi-global stabilization by nonsmooth output feedback can be achieved for the triangular system

$$\begin{aligned}
\dot{x}_1 &= x_2^{p_1} + f_1(x_1) \\
&\vdots \\
\dot{x}_{n-1} &= x_n^{p_{n-1}} + f_{n-1}(x_1, \cdots, x_{n-1}) \\
\dot{x}_n &= u + f_n(x_1, \cdots, x_n) \\
y &= x_1,
\end{aligned} \tag{15.1}$$

B. Ghosh et al. (Eds.): Emergent Prob. in Nonlinear Sys. & Ctrl., LNCIS 393, pp. 253–283.
springerlink.com

considered previously, for instance, in [3, 22], where $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$, $u \in \mathbb{R}$ and $y \in \mathbb{R}$ are the system state, input and output, respectively. The mappings $f_i : \mathbb{R}^i \to \mathbb{R},\ i = 1, \cdots, n$, are C^1 functions with $f_i(0, \cdots, 0) = 0$ and $p_1, \cdots, p_{n-1}$ are odd positive integers.

The problem of semi-global stabilization by nonsmooth output feedback can be formulated as follows. Given a bound $r > 0$, find, if possible, a continuous but non-differentiable dynamical output compensator, which may depend on r, of the form

$$\begin{aligned} \dot{\hat{z}} &= \eta(\hat{z}, y), \qquad \hat{z} \in \mathbb{R}^{n-1}, \qquad \eta \in C(\mathbb{R}^n, \mathbb{R}^{n-1}) \\ u &= \theta(\hat{z}, y), \qquad\qquad\quad \theta \in C(\mathbb{R}^n, \mathbb{R}) \end{aligned} \tag{15.2}$$

with $\eta(0, 0) = 0$ and $\theta(0, 0) = 0$, such that the following properties hold:

- Semi-global attractivity: all the trajectories of the closed-loop system (15.1)-(15.2) starting from the compact set $\Gamma_x \times \Gamma_{\hat{z}} \triangleq [-r, r]^n \times [-r, r]^{n-1} \subset \mathbb{R}^n \times \mathbb{R}^{n-1}$ converge to the origin;
- Local asymptotic stability: the closed-loop system (15.1)-(15.2) is locally asymptotically stable at the origin $(x, \hat{z}) = (0, 0)$.

It has been known that for nonlinear control systems, global stabilizability by state feedback plus global observability is usually not sufficient for achieving global stabilizability by output feedback. As a matter of fact, counter-examples were given in [21] illustrating that even for a simple feedback linearizable or minimum-phase system in the plane, which is uniformly observable and globally stabilizable by smooth state feedback, global output feedback stabilization may still not be possible. The impossibility of this task indicates that in the nonlinear case, semi-global, instead of global, stabilization by output feedback is perhaps the more realistic control objective to be pursued.

While there are numerous papers in the literature devoted to the topic of semi-global output feedback stabilization, a major progress was reported in the work [26], where it was shown that for nonlinear systems, uniform observability [9] and global stabilizability by smooth state feedback implies semi-global stabilizability by smooth output feedback. As a consequence, semi-global stabilization by smooth output feedback was shown to be possible for minimum-phase nonlinear systems including system (15.1) with $p_1 = \cdots = p_{n-1} = 1$ [27], as well as for a class of non-minimum-phase nonlinear systems [11].

Note that when $p_i \equiv 1,\ i = 1, \cdots, n-1$, the triangular system (15.1) is feedback linearizable and hence globally stabilizable by smooth state feedback [10]. In addition, according to the characterization given in [9], the system is also uniformly observable. These two conditions are, however, violated when some of $p_i > 1$. Indeed, the triangular system (15.1) is no longer uniformly observable [9, 26], simply because the system state of (15.1) can only be represented as a *non-smooth*, or worse, *singular* function of the system input, output, and their derivatives (see, for instance, Example 15.4.1). Furthermore, system (15.1) is not smoothly stabilizable, even locally, for the reason that its linearized system may have uncontrollable modes associated with eigenvalues on the right-half plane,

as illustrated by the simple planar system $\dot{x}_1 = x_2^3 + x_1,\ \dot{x}_2 = u,\ y = x_1$ [13]. The loss of uniform observability and smooth stabilizability makes the output feedback stabilization of system (15.1), on one hand, a challenging problem from a viewpoint of control theory. On the other hand, from an application perspective, the triangular system (15.1) contains, for instance, a class of underactuated mechanical systems [25] as a special case. These two factors are the primary motivation of our investigations in this paper.

Although the nonlinear system (15.1) is neither uniformly observable nor smoothly stabilizable, it is *globally* stabilizable by *nonsmooth state* feedback, as shown in [22]. More recently, it has been further proved in [23], among the other things, that global stabilization of (15.1) by nonsmooth output feedback is possible as long as $f_i(x_1, \cdots, x_i), i = 1, \cdots, n$ satisfy some restrictive growth conditions. In addition, the *local* stabilization of the triangular system (15.1) can always be achieved by *nonsmooth output* feedback. The latter was accomplished by means of the theory of homogeneous systems [1], particularly, using the homogeneous approximation [4]-[6], [13]-[15] and the robust stability of homogeneous systems [8, 24]. In view of the results obtained in [22, 23], one might naturally make the conjecture that the triangular system (15.1) is semi-globally stabilizable by nonsmooth output feedback, without requiring any growth condition.

It turns out that this conjecture is true and semi-global stabilization by nonsmooth output feedback is indeed possible, for a significant class of non-uniformly observable and non-smoothly stabilizable systems such as (15.1). One of the main contributions of this paper is the following theorem.

Theorem 15.1.1. *For the triangular system (15.1), there exists a nonsmooth dynamic output compensator of the form (15.2), such that the closed-loop system (15.1)-(15.2) is semi-globally asymptotically stable.*

The significance of Theorem 15.1.1 over the existing results can be summarized as follows. On the one hand, it generalizes the semi-global output feedback stabilization results in [26, 27] for nonlinear systems that are required to be uniformly observable and smoothly stabilizable, to a wider class of non-uniformly observable and non-smoothly stabilizable systems such as (15.1) and (15.60). On the other hand, it shows that the local output feedback stabilization result in [23] (see Theorem 3.5) can be extended, without requiring any growth condition on (15.1), to the semi-global case, which is certainly a substantial progress from both theoretical and practical viewpoints. Finally, compared with the previous global output feedback stabilization results [28, 29, 23], all the restrictive conditions such as $p_1 = \cdots = p_{n-1} = p$ and a high-order global Lipschitz-like condition in [28], or the growth requirements imposed on system (15.1) in [23] have been removed. The price we paid is that only semi-global rather than global stabilizability is achieved.

In Section 15.4, we shall prove Theorem 15.1.1 by explicitly constructing a non-smooth, dynamic output feedback compensator of the form (15.2) for the triangular system (15.1). While the design of Hölder continuous state feedback control laws is reminiscent of the work [22], the nonsmooth observer construction is new and carried out in a subtle manner. It integrates the idea of the recursive

observer design in [23] with the technique of the saturated state estimates in [16]. To prove the semi-global stability of the closed-loop system, we employ a Lyapunov function that is motivated by the work [30] yet much simpler than the one used in [26]. Such a Lyapunov function, together with a delicate choice of the level sets, makes it possible to simplify the analysis and synthesis of our semi-global, nonsmooth output feedback control scheme. In the case when $p_i = 1,\ 1 \leq i \leq n-1$ in (15.1), our design method leads to Proposition 15.3.1 in Section 15.3, which has refined the existing results given, for instance, in [10, 27], by relaxing the uniformly observability condition.

The paper is organized as follows. Section 15.2 introduces useful notations and a number of technical lemmas and inequalities to be used in the sequel. In Section 15.3, we revisit the problem of semi-global stabilization via output feedback for a class of C^0 nonlinear systems that are smoothly stabilizable but *not necessarily uniformly observable.* In particular, a non-separation based paradigm is presented, demonstrating how a semi-global output feedback control law can be constructed step-by-step, for the C^0 triangular system (15.1) with $p_1 = \cdots = p_n = 1$, *without requiring uniform observability* (see, e.g., Remark 3.1). The new ingredients include the construction of a simple control Lyapunov function that substantially simplifies the analysis of semi-global stability, and the introduction of a recursive algorithm for assigning the observer gains. Motivated by the idea exploited in Section 15.3, we present in Section 15.4 a novel semi-global output feedback control scheme that can be viewed as a non-smooth enhancement of the result obtained in Section 15.3. The new output feedback control method makes it possible to construct both non-smooth observers and state feedback controllers recursively, achieving semi-global stabilization for the non-uniformly observable and non-smoothly stabilizable system (15.1). Section 15.5 discusses how the result in Section 15.4 can be extended to a class of planar systems in the Hessenberg form [12, 2], which goes beyond the strict-triangular structure and makes the semi-global stabilization by output feedback a far more difficult task. An illustrative example and some discussions are also included in Section 15.5, explaining why it is hard to establish a semi-global stabilization result for n-dimensional non-triangular systems. Concluding remarks are drawn in Section 15.6. The appendix contains the proofs of three propositions used in Section 15.4.

15.2 Preliminaries

In this section, we introduce several useful notations and technical lemmas that will be crucial in subsequent developments. Throughout this paper, the following notations will be used.

- $\forall \sigma > 0$ and $a \in \mathbb{R}$, a function $[a]^\sigma$ is defined as

$$[a]^\sigma = \begin{cases} -|a|^\sigma & \text{if} \quad a < 0 \\ |a|^\sigma & \text{if} \quad a \geq 0. \end{cases}$$

Clearly, $[[a]^{\sigma}]^{\frac{1}{\sigma}} \equiv a$ and if σ is a ratio of two odd positive numbers, then $[a]^{\sigma} \equiv a^{\sigma}$. Moreover, if $\sigma \geq 1$, $[a]^{\sigma}$ is a C^1 function and its derivative is $\sigma|a|^{\sigma-1}$.

- B_M represents the compact set $[-M, M]^n \triangleq [-M, M] \times \cdots \times [-M, M] \subset \mathbb{R}^n$.
- $\min\{a, b\}$ and $\max\{a, b\}$ stand for the minimum and maximum of two real numbers a and b, respectively.
- $f(\cdot)\big|_{\Gamma}$ denotes the restriction of a function $f(\cdot)$ on the set Γ.
- K represents a generic positive real number for which we use the convention $K + K = K$ and $K \times K = K$.

The next four lemmas are the key for the analysis and synthesis of semi-globally stabilizing, nonsmooth output feedback controllers proposed in this paper.

Lemma 15.2.1. *Given positive real numbers $a, b, m, n, \gamma, \delta$, the following inequality holds:*

$$\gamma a^m b^n \leq \delta a^{m+n} + \frac{n}{m+n}\left(\frac{m+n}{m}\right)^{-m/n} \gamma^{(m+n)/n} \delta^{-m/n} b^{m+n}.$$

Lemma 15.2.2. *Let $a_1, \cdots, a_n$ and p be positive real numbers. Then,*

$$(a_1 + \cdots + a_n)^p \leq \max\{n^{p-1}, 1\}(a_1^p + \cdots + a_n^p).$$

Lemma 15.2.3. *Let a and b be any real numbers and $\sigma \in (0, 1]$. Then,*

$$|a - b| \leq 2^{1-\sigma}\left|[a]^{\frac{1}{\sigma}} - [b]^{\frac{1}{\sigma}}\right|^{\sigma}.$$

Lemma 15.2.4. *For all $a, b \in \mathbb{R}$ and any odd positive integer p, given an arbitrarily small $\varepsilon > 0$, there exists a real number $K_0 > 0$, which depends only on p and ε, such that*

$$|a^p - b^p| \leq K_0|a - b|^p + \varepsilon|a|^p.$$

The proofs of Lemmas 15.2.1—15.2.4 are straightforward and hence left to the reader as an exercise.

Definition 15.2.1. *A saturation function with the threshold $M > 0$ is defined as*

$$\mathrm{sat}_M(a) = \begin{cases} -M & \text{if} \quad a < -M \\ a & \text{if} \quad |a| \leq M \\ M & \text{if} \quad a > M. \end{cases}$$

Clearly, the saturation function thus defined is continuous, bounded by M and has the following properties.

Lemma 15.2.5. *Assume that $p \geq 1$ is an odd integer. Then,*

$$|a - \mathrm{sat}_M(b)| \leq 2\min\{|a^p - b^p|^{\frac{1}{p}}, M\}, \quad \forall a \in [-M, M], \ \forall b \in \mathbb{R}.$$

Proof. When $p = 1$, the inequality above reduces to

$$|a - \text{sat}_M(b)| \leq 2\min\{|a - b|, M\}, \tag{15.3}$$

which can be proved straightforwardly by using Definition 15.2.1. When $p > 1$, Lemma 15.2.5 can be easily derived from (15.3). In fact, observe that $[\text{sat}_M(b)]^p = \text{sat}_{M^p}(b^p)$. Then, it follows from Lemma 15.2.3 that

$$\begin{aligned} |a - \text{sat}_M(b)| &\leq 2^{1-\frac{1}{p}} |a^p - \text{sat}_{M^p}(b^p)|^{\frac{1}{p}} \\ &\leq 2^{1-\frac{1}{p}} \left[2\min\{|a^p - b^p|, M^p\}\right]^{\frac{1}{p}} = 2\min\{|a^p - b^p|^{\frac{1}{p}}, M\}. \end{aligned}$$

■

The following lemma characterizes a useful property of smooth functions on a compact set.

Lemma 15.2.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 mapping and $\Gamma = [-N, N]^n$ a compact set in $\mathbb{R}^n$ with $N > 0$ being a real number. Then, $\forall \sigma_i \in (0, 1]$, $i = 1, \cdots, n$, there exists a constant $K \geq 1$ depending on N, so that $\forall (a_1, \cdots, a_n) \in \Gamma$, $\quad \forall (b_1, \cdots, b_n) \in \Gamma$,*

$$|f(a_1, \cdots, a_n) - f(b_1, \cdots, b_n)| \leq K(|a_1 - b_1|^{\sigma_1} + \cdots + |a_n - b_n|^{\sigma_n}).$$

Proof. Without the loss of generality, consider the case of $n = 1$. When $\sigma_1 = 1$, it is clear that for any $K \geq \max\{f'(\xi) : \xi \in [-N, N]\}$,

$$|f(a_1) - f(b_1)| \leq K|a_1 - b_1|, \qquad \forall a_1 \in [-N, N], \quad \forall b_1 \in [-N, N], \tag{15.4}$$

which is a consequence of the mean-value theorem. When $0 < \sigma_1 < 1$, in view of (15.4) and the smoothness of $f([\cdot]^{\frac{1}{\sigma_1}})$, one deduces from Lemma 15.2.3 that there exists a constant $K > 0$ such that

$$\begin{aligned} |f(a_1) - f(b_1)| &= \left| f([[a_1]^{\sigma_1}]^{\frac{1}{\sigma_1}}) - f([[b_1]^{\sigma_1}]^{\frac{1}{\sigma_1}}) \right| \\ &\leq \frac{K}{2} \left| [a_1]^{\sigma_1} - [b_1]^{\sigma_1} \right| \leq K|a_1 - b_1|^{\sigma_1}. \end{aligned}$$

■

15.3 Existing Results Revisited and a Direct Design Method

In order to tackle the semi-global stabilization problem for the highly nonlinear system (15.1) by output feedback, it is essential to understand how the problem was solved in the simple case when $p_i = 1,\ 1 \leq i \leq n$ in (15.1). In this case, the controlled plant (15.1) reduces to

$$\begin{aligned} \dot{x}_1 &= x_2 + f_1(x_1) \\ &\vdots \\ \dot{x}_n &= u + f_n(x_1, \cdots, x_n), \\ y &= x_1, \end{aligned} \tag{15.5}$$

which is smoothly stabilizable and uniformly observable if the functions $f_i(\cdot), i = 1, \cdots, n$, are smooth. By [26, 27] or [11, 10], semi-global stabilization of the smooth system (15.5) is achievable by output feedback. This was done based on the high-gain observer [17, 9], the idea of saturating the estimated state [16], and the technique of dynamic extension. Notably, uniform observability plays a crucial role. It basically implies that after a dynamic extension [26, 11], the augmented system is globally diffeomorphic to a sort of "feedback linearizable form" where all the nonlinearities appear in the last equation involving $u(\cdot)$.

Unfortunately, such a global diffeomorphism does not exist anymore for the non-feedback linearizable system (15.1). Moreover, due to the lack of uniform observability and smooth stabilizability, the semi-global design methods [26, 27, 11, 10] cannot be applied to deal with the system (15.1). Lastly, the proof of semi-global stability of the closed-loop system in [26] was very complicated, because of the use of intricate Lyapunov functions that are only defined on a subset of the state space. Consequently, stability analysis in [26] was less transparent and required a series of subtle estimations, making the arguments of [26, 27] hard to be adopted for the nonsmoothly stabilizable and nonuniformly observable system (15.1).

To motivate how an effective control scheme can be developed for the system (15.1), we revisit in this section the simple system (15.5). The purpose is to develop a "direct" semi-global design method *without requiring uniform observability nor making a change of coordinates* — the idea that has not been fully exploited yet for system (15.5). It turns out that the new control strategy can relax the uniform observability condition, thus resulting in a refined semi-global stabilization result for system (15.5). By comparison, our method is simpler than the one in [26, 27] (in terms of the stability analysis) but a bit more complicated than the method in [11, 10] (in terms of step-by-step design). However, two advantages of our design method are: 1) it is applicable to the C^0 system (15.5) that is not necessarily uniformly observable; 2) it can be carried over, with a subtle twist, to the case of the highly nonlinear system (15.1) with $p_i \geq 1$.

The following proposition is the main result of this section and can be proved by using the proposed semi-global design method.

Proposition 15.3.1. *For the C^0 triangular system (15.5), assume that for $i = 1, \cdots, n$, $\forall (x_1, \cdots, x_i) \in \Gamma_i$, $\forall (y_1, \cdots, y_i) \in \Gamma_i$,*

$$|f_i(x_1, \cdots, x_i) - f_i(y_1, \cdots, y_i)| \leq K(|x_1 - y_1| + \cdots + |x_i - y_i|), \tag{15.6}$$

where $\Gamma_i = [-N, N]^i$ is a compact set in $\mathbb{R}^i$, $N > 0$ is a real number and $K \geq 1$ is a constant depending on N. Then, system (15.5) is semi-globally stabilizable by the C^0 dynamic output compensator (15.2).

The following remark illustrates how Proposition 15.3.1 refines the existing semi-global stabilization results in the case of the triangular system (15.5), without the uniform observability assumption.

Remark 15.3.1. For the triangular system (15.5), semi-global output feedback stabilization was commonly done in two steps [10, 11, 27] . First, a global change of coordinates (diffeomorphism) is used to transformed system (15.5) into a "feedback linearizable form" in which the nonlinearity appears only in the last equation, i.e., moving all the nonlinearity $f_i(\cdot)(i = 1, \cdots, n-1)$ in (15.5) to the last equation of the system where the control signal $u(\cdot)$ is present. Then, a semi-global output feedback controller was designed for the resulted system. The advantage of such design methods is that semi-global output feedback design is much easier (one step design) for the resulted system than for the original system (15.5). The drawback is, however, that the functions $f_i(\cdot), i = 1, \cdots, n$ are required to be smooth enough to guarantee uniform observability of system (15.5) and the existence of a global diffeomorphism. In contrast, Proposition 15.3.1 relaxes the smoothness requirement and hence the uniform observability condition, by designing an output feedback compensator for the original system (15.5) directly, step-by-step. For example, it is easy to verify that our semi-global design method is still valid for the C^0 yet smoothly stabilizable system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 + |e^{x_2} - 1| \\ \dot{x}_3 &= u \\ y &= x_1, \end{aligned} \tag{15.7}$$

while the methods in [10, 11, 26, 27] no longer work, because the non-differentiable nonlinearity $|e^{x_2} - 1|$ renders system (15.7) non-uniformly observable nor differmorphic to a "feedback linearizable form".

Proof. To begin with, we observe that by adding an integrator, it is easy to get recursively a smooth state feedback controller (via domination)

$$u^*(x_1, \cdots, x_n) = -\xi_n \beta_n(x_1, \cdots, x_n) \tag{15.8}$$

such that the closed-loop system (15.5)-(15.8) satisfies

$$\dot{V}_c(x) \leq -2(\xi_1^2 + \cdots + \xi_n^2) + \xi_n(u - u^*(x_1, \cdots, x_n)), \tag{15.9}$$

where $V_c(x) = \frac{1}{2}(\xi_1^2 + \cdots + \xi_n^2)$ is a Lyapunov function and $\xi_i = x_i - x_i^*,\ i = 1, \cdots, n$,

$$x_1^* = 0, \quad x_2^* = -\xi_1 \beta_1(x_1), \quad \cdots, \quad x_n^* = -\xi_{n-1} \beta_{n-1}(x_1, \cdots, x_{n-1})$$

with $\beta_i(\cdot) > 0$ being known smooth functions.

From inequality (15.9), it is clear that the smooth state feedback controller $u = u^*(x_1, \cdots, x_n)$ globally stabilizes the nonlinear system (15.5) at the origin $x = 0$.

Since $V_c(\cdot)$ is positive definite and proper, one can define the level set $\Omega_x = \{x \in \mathbb{R}^n | V_c(x) \leq r_0 + 1\}$, where $r_0 > 0$ is a constant such that $\Gamma_x = [-r, r]^n \subset \{x \in \mathbb{R}^n | V_c(x) \leq r_0\}$. Then, denote

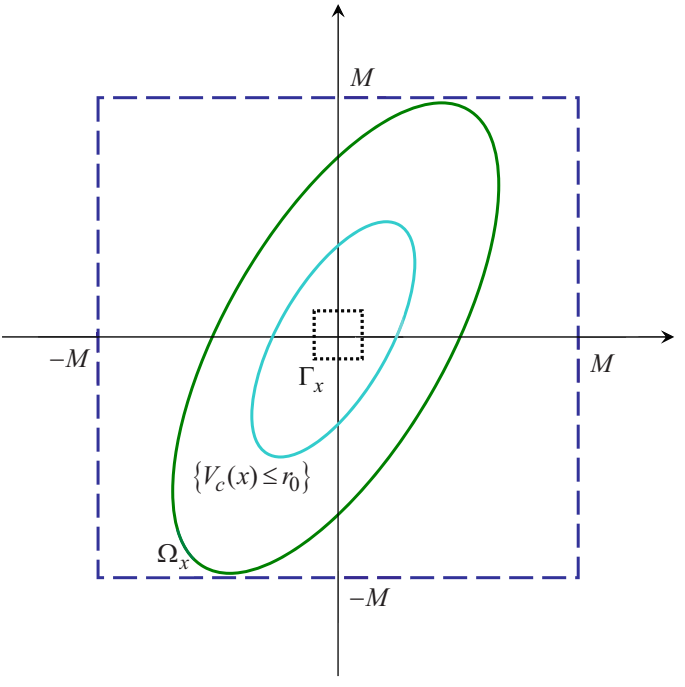


Fig. 15.1. The level set on x-space and the saturation threshold M.

$$M = \max_{x \in \Omega_x} ||x||_\infty > 0$$

as a saturation threshold, where $||\cdot||_\infty$ stands for L_∞-norm of vectors. The relations among the compact sets thus introduced are illustrated in Figure 15.1.

Because the states $(x_2, \cdots, x_n)$ of (15.5) are not measurable, the controller (15.8) is not realizable. To get an implementable controller, we shall design an $(n-1)-th$ order observer to estimate, instead of the states $(x_2, \cdots, x_n)$, the unmeasurable variables $(z_2, \cdots, z_n)$ defined by

$$z_2 = x_2 - L_2 x_1, \quad \cdots, \quad z_n = x_n - L_n x_{n-1}, \tag{15.10}$$

where $L_i \ge 1,\ i = 2, \cdots, n$ are gains to be assigned later.

From (15.10) it follows that

$$\begin{aligned}
\dot{z}_2 &= \big[x_3 + f_2(x_1, x_2)\big] - L_2\big[x_2 + f_1(x_1)\big] \\
\dot{z}_3 &= \big[x_4 + f_3(x_1, x_2, x_3)\big] - L_3\big[x_3 + f_2(x_2, x_3)\big] \\
&\vdots \\
\dot{z}_n &= \big[u + f_n(x_1, \cdots, x_n)\big] - L_n\big[x_n + f_{n-1}(x_1, \cdots, x_{n-1})\big].
\end{aligned} \tag{15.11}$$

In view of (15.11), we design the implementable C^0 dynamic compensator

$$\begin{aligned}
\dot{\hat{z}}_2 &= \big[\hat{x}_3 + \hat{f}_2(\cdot)\big] - L_2\big[\hat{x}_2 + f_1(\cdot)\big] \\
\dot{\hat{z}}_3 &= \big[\hat{x}_4 + \hat{f}_3(\cdot)\big] - L_3\big[\hat{x}_3 + \hat{f}_2(\cdot)\big] \\
&\vdots \\
\dot{\hat{z}}_n &= \big[u + \hat{f}_n(\cdot)\big] - L_n\big[\hat{x}_n + \hat{f}_{n-1}(\cdot)\big],
\end{aligned} \tag{15.12}$$

where

$$\hat{x}_2 = \hat{z}_2 + L_2 x_1, \quad \hat{x}_3 = \hat{z}_3 + L_3\hat{x}_2, \quad \cdots, \quad \hat{x}_n = \hat{z}_n + L_n\hat{x}_{n-1}, \tag{15.13}$$

$$\hat{f}_i(\cdot) \triangleq f_i(x_1, \mathrm{sat}_M(\hat{x}_2), \cdots, \mathrm{sat}_M(\hat{x}_i)), \quad i = 2, \cdots, n. \tag{15.14}$$

Using the certainty equivalence principle, we obtain the realizable controller

$$u = \hat{u}^*(\cdot) \triangleq u^*(x_1, \mathrm{sat}_M(\hat{x}_2), \cdots, \mathrm{sat}_M(\hat{x}_n)). \tag{15.15}$$

Let $e_i = z_i - \hat{z}_i = x_i - L_i x_{i-1} - \hat{z}_i,\ i = 2, \cdots, n$ be the estimate errors. Then, by the identity $x_i - \hat{x}_i = e_i + L_i(x_{i-1} - \hat{x}_{i-1})$, we have

$$x_i - \hat{x}_i = e_i + L_i e_{i-1} + \cdots + L_i \cdots L_3 e_2. \tag{15.16}$$

Consequently, the error dynamics is given by

$$\begin{aligned}
\dot{e}_2 &= \Big[e_3 + L_3 e_2 + \Big(f_2(\cdot) - \hat{f}_2(\cdot)\Big)\Big] - L_2 e_2 \\
\dot{e}_3 &= \Big[e_4 + L_4 e_3 + L_4 L_3 e_2 + \Big(f_3(\cdot) - \hat{f}_3(\cdot)\Big)\Big] - \Big[L_3 e_3 + L_3^2 e_2 + L_3\Big(f_2(\cdot) - \hat{f}_2(\cdot)\Big)\Big] \\
&\vdots \\
\dot{e}_n &= \Big[f_n(\cdot) - \hat{f}_n(\cdot)\Big] - \Big[L_n e_n \cdots + L_n^2 L_{n-1} \cdots L_3 e_2 + L_n\Big(f_{n-1}(\cdot) - \hat{f}_{n-1}(\cdot)\Big)\Big].
\end{aligned} \tag{15.17}$$

By definition, $\hat{f}_i(\cdot) \equiv f_i(x_1, \mathrm{sat}_M(\hat{x}_2), \cdots, \mathrm{sat}_M(\hat{x}_i)),\ i = 2, \cdots, n$ are smooth functions of the variables $x_1, \mathrm{sat}_M(\hat{x}_2), \cdots, \mathrm{sat}_M(\hat{x}_i)$ which, except x_1, satisfy $|\mathrm{sat}_M(\hat{x}_j)| \le M,\ j = 2, \cdots, i$ no matter how large L_i's are. These observations, together with Lemmas 15.2.5 and the condition (15.6), imply the existence of a fixed constant $K \ge 1$, which depends on M and is independent of all the L_i's, such that on the set $B_M \times \mathbb{R}^{n-1} = \{(x, \hat{z}) \in \mathbb{R}^n \times \mathbb{R}^{n-1} | (x_1, \cdots, x_n) \in [-M, M]^n\}$, the following estimations hold ($\sigma_1 = \cdots = \sigma_i = 1$):

$$\begin{aligned}
|f_i(\cdot) - \hat{f}_i(\cdot)|\Big|_{B_M \times \mathbb{R}^{n-1}} &\le \frac{K}{n}\Big(|x_2 - \mathrm{sat}_M(\hat{x}_2)| + \cdots + |x_i - \mathrm{sat}_M(\hat{x}_i)|\Big) \\
&\le \frac{K}{n}\Big(|x_2 - \hat{x}_2| + \cdots + |x_i - \hat{x}_i|\Big).
\end{aligned}$$

In view of (15.16), we deduce from (15.18)

$$|f_i(\cdot) - \hat{f}_i(\cdot)|\Big|_{B_M \times \mathbb{R}^{n-1}} \leq n(|e_i| + L_i|e_{i-1}| + \cdots + L_i \cdots L_3|e_2|), \qquad i = 2, \cdots, n. \tag{15.18}$$

Using the same argument, we have (by Lemmas 15.2.5 and the condition (15.6))

$$|\hat{u}^*(\cdot) - u^*(\cdot)|\Big|_{B_M \times \mathbb{R}^{n-1}} \leq K \min\Big\{|e_n| + L_n|e_{n-1}| + \cdots + L_n \cdots L_3|e_2|, 1\Big\}, \tag{15.19}$$

where $K \geq 1$ is a generic constant, which depends on M and is independent of all the L_i's.

With the aid of (15.19), it concluded from (15.9) that on the set $B_M \times \mathbb{R}^{n-1}$,

$$\begin{aligned} \dot{V}_c &\leq -2(\xi_1^2 + \cdots + \xi_n^2) + K|\xi_n| \min\{|e_n| + L_n|e_{n-1}| + \cdots + L_n \cdots L_3|e_2|, 1\} \qquad (15.20) \\ &\leq -(\xi_1^2 + \cdots + \xi_n^2) + \min\{C_n e_n^2 + C_{n-1}(L_n)e_{n-1}^2 + \cdots + C_2(L_n, \cdots, L_3)e_2^2, C_n\}, \end{aligned}$$

where $C_n \geq 1$ is a constant independent of L_i's, $C_{n-1}(L_n) \geq C_n$, $\cdots, C_2(L_n, \cdots, L_3) \geq C_n$ are polynomial functions of their arguments. The inequality (15.20) can be obtained by completing the squares, as done in [28].

Now, choose $V_e(e) = \frac{1}{2}(e_n^2 + \cdots + e_2^2)$ for the error dynamics (15.17). On the set $B_M \times \mathbb{R}^{n-1}$,

$$\begin{aligned} \dot{V}_e &\leq \big|e_n[f_n(\cdot) - \hat{f}_n(\cdot)]\big| - L_n e_n^2 + \big|e_n(L_n^2 e_{n-1} + \cdots + L_n^2 L_{n-1} \cdots L_3 e_2)\big| \qquad (15.21) \\ &\quad + \big|e_n L_n[f_{n-1}(\cdot) - \hat{f}_{n-1}(\cdot)]\big| + \cdots + \big|e_2[f_2(\cdot) - \hat{f}_2(\cdot)]\big| + \big|e_2(e_3 + L_3 e_2)\big| - L_2 e_2^2 \\ &\leq -\big(L_n - \bar{C}_n\big)e_n^2 - \big(L_{n-1} - \bar{C}_{n-1}(L_n)\big)e_{n-1}^2 - \cdots - \big(L_2 - \bar{C}_2(L_n, \cdots, L_3)\big)e_2^2, \end{aligned}$$

where $\bar{C}_n \geq 1$ is a constant independent of L_i's, $\bar{C}_{n-1}(L_n) \geq 1$, $\cdots, \bar{C}_2(L_n, \cdots, L_3) \geq 1$ are fixed polynomial functions of their arguments. The estimation (15.21) can be deduced by (15.18) and the completion of the square, as done in [28].

From (15.21), it is easy to see that the gain assignments

$$\begin{aligned} L_n &= L_n(L) \triangleq \bar{C}_n + LC_n \geq 1 \\ L_{n-1} &= L_{n-1}(L) \triangleq \bar{C}_{n-1}(L_n) + LC_{n-1}(L_n) \geq 1 \\ &\vdots \qquad (15.22) \\ L_2 &= L_2(L) \triangleq \bar{C}_2(L_n, \cdots, L_3) + LC_2(L_n, \cdots, L_3) \geq 1 \end{aligned}$$

with $L > 0$ being a parameter to be determined later, render

$$\dot{V}_e\Big|_{B_M \times \mathbb{R}^{n-1}} \leq -LW_e, \tag{15.23}$$

where

$$W_e \triangleq C_n e_n^2 + C_{n-1}(L)e_{n-1}^2 + \cdots + C_2(L)e_2^2 \tag{15.24}$$

and $C_{n-1}(L) \geq C_n, \cdots, C_2(L) \geq C_n$ are fixed positive polynomial functions of L.

Motivated by [26], we define

$$\mu(L) = \frac{1}{2}[(2r + L_2(L)r)^2 + \cdots + (2r + L_n(L)r)^2].$$

In view of the identity $e_i = x_i - L_i x_{i-1} - \hat{z}_i,\ i = 2, \cdots, n$,

$$\mu(L) \geq \max_{(x,\hat{z}) \in \Gamma_x \times \Gamma_{\hat{z}}} V_e(x, \hat{z}) > 0.$$

As done in [30], choose the Lyapunov function

$$V(x, \hat{z}) = V_c(x) + \frac{\ln(1 + V_e(e))}{\ln(1 + \mu(L))}, \tag{15.25}$$

which is different from the complicated one used in [26], for the closed-loop system (15.5) and (15.12)-(15.15). Moreover, define the corresponding level set

$$\Omega = \{(x, \hat{z}) \in \mathbb{R}^n \times \mathbb{R}^{n-1} | V(x, \hat{z}) \leq r_0 + 1\}. \tag{15.26}$$

Then, it is easy to verify the following facts (see Figure 15.2):

- For every $L > 0$, $V(x, \hat{z})$ is a positive definite and proper function and Ω is a compact set in $\mathbb{R}^n \times \mathbb{R}^{n-1}$ (i.e., $(x, \hat{z})$-space). Once $L > 0$ is fixed, $V(x, \hat{z})$ and Ω are fixed too;
- $\forall L \geq 1,\ \ \Omega \supset \Gamma_x \times \Gamma_{\hat{z}}$.
  This can be deduced from the relationship that $V_c(x) \leq r_0$ and $\frac{\ln(1+V_e(e))}{\ln(1+\mu(L))} \leq 1,\ \forall (x, \hat{z}) \in \Gamma_x \times \Gamma_{\hat{z}}$.
- $\forall L \geq 1,\ \ B_M \times \mathbb{R}^{n-1} \supset \Omega$.
  This follows from the inclusion $B_M \supset \Omega_x$ and the fact that $V(x, \hat{z}) \leq r_0 + 1$ implies $V_c(x) \leq r_0 + 1$.

To prove the semi-global asymptotic stability, it remains to show that one can take advantage of the uniform boundedness of Ω with respect to L and choose sufficiently large $L > 0$, such that $\dot{V}\big|_{\Omega} < 0$.

In view of the relationship $B_M \times \mathbb{R}^{n-1} \supset \Omega$, we deduce from (15.23) and (15.20) that $\forall L > 0$,

$$\dot{V}\Big|_{\Omega} = \dot{V}_c\Big|_{\Omega} + \frac{1}{\ln(1 + \mu(L))} \frac{\dot{V}_e}{1 + V_e}\Big|_{\Omega} \leq -(\sum_{i=1}^{n} \xi_i^2) + \min\{W_e, C_n\} - \frac{L}{\ln(1 + \mu(L))} \frac{W_e}{1 + V_e}. \tag{15.27}$$

Observe that $C_i(L) \geq C_n \geq 1,\ i = 2, \cdots, n$ implies

$$\frac{W_e}{1 + V_e} \geq \frac{C_n \cdot \frac{W_e}{C_n}}{1 + \frac{W_e}{C_n}} \geq \frac{C_n}{2} \min\{\frac{W_e}{C_n}, 1\} = \frac{1}{2} \min\{W_e, C_n\}.$$

We have,

$$\dot{V}\Big|_{\Omega} \leq -(\sum_{i=1}^{n} \xi_i^2) - \Big[\frac{L}{2\ln(1 + \mu(L))} - 1\Big] \min\{W_e, C_n\}. \tag{15.28}$$

By construction, $\mu(L) > 0$ is a *polynomial* function of L. Thus, there is a constant $L^* > 0$ such that

$$\frac{L}{2\ln(1+\mu(L))} \geq 2, \qquad \forall L \geq L^*.$$

Choosing $L = L^*$, we have immediately,

$$\dot{V}\Big|_{\Omega} \leq -(\sum_{i=1}^{n} \xi_i^2) - \min\{W_e, C_n\}.$$

That is, the C^0 triangular system (15.5) that is *not necessarily uniformly observable* is semi-globally stabilizable via output feedback. ■

As we shall see in the next section, the advantage of our direct design method over the existing work will become more obvious when dealing with the highly nonlinear system (15.1). Indeed, because of the loss of smooth stabilizability and uniform observability, the existing semi-global design methods [10, 11, 26, 27] cannot be be applied to system (15.1). However, the idea discussed in this section provides a possible way of controlling the nonlinear system (15.1) via output feedback. This is exactly the contribution of the next section.

15.4 Semi-global Output Feedback Stabilization of Triangular Systems

Inspired by the philosophy exploited in the previous section, we now present a recursive, semi-global output feedback design method that leads to the the main result of this paper — Theorem 15.1.1. The establishment of Theorem 15.1.1 is constructive and carried out by generalizing the output feedback control scheme

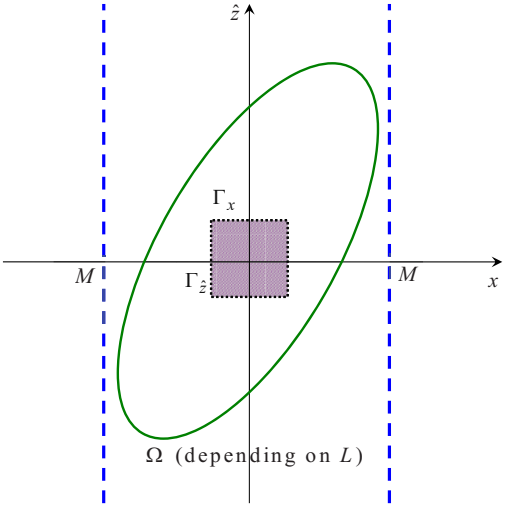


Fig. 15.2. The level set on $(x, \hat{z})$-space.

developed in Section 15.3, with a subtle twist, to the non-uniformly observable and non-smoothly stabilizable system (15.1). In particular, we design explicitly a semi-globally stabilizing, nonsmooth dynamic output compensator based on the tool of adding a power integrator [22], the recursive nonsmooth observer design algorithm [23], and the idea of saturating the estimated states [16].

Proof of Theorem 15.1.1. First of all, using the nonsmooth state feedback control scheme [22], we can design a globally stabilizing state feedback controller as follows.

Let $\xi_1 = x_1$ and choose $V_1(x_1) = \frac{1}{2}\xi_1^2$. Since $f_1(x_1)$ is a C^1 function with $f_1(0) = 0$, there exists a smooth function $\rho_1(\cdot) \geq 0$ such that $|f_1(x_1)| \leq |x_1|\rho_1(x_1)$. Hence,

$$\dot{V}_1 \leq \xi_1 x_2^{*p_1} + \xi_1(x_2^{p_1} - x_2^{*p_1}) + \rho_1(x_1)\xi_1^2.$$

Setting $x_2^{*p_1} = -\xi_1\beta_1(x_1) \triangleq -\xi_1[(n+1) + \rho_1(x_1)]$ yields

$$\dot{V}_1 \leq -(n+1)\xi_1^2 + \xi_1(x_2^{p_1} - x_2^{*p_1}).$$

Next, let $\xi_2 = x_2^{p_1} - x_2^{*p_1}$ and choose

$$V_2(x_1, x_2) = V_1(x_1) + \int_{x_2^*}^{x_2} (s^{p_1} - x_2^{*p_1})^{2-\frac{1}{p_1}} \mathrm{d}s,$$

which is C^1, positive definite and proper [22].

Observe that $|f_2(x_1, x_2)| \leq (|x_1| + |x_2|)\bar{\rho}_2(x_1, x_2)$, where $\bar{\rho}_2(\cdot) \geq 0$ is a smooth function. With this in mind, an argument similar to the one in [22] gives

$$\dot{V}_2 \leq -n\xi_1^2 + \xi_2^{2-\frac{1}{p_1}} x_3^{*p_2} + \xi_2^{2-\frac{1}{p_1}}(x_3^{p_2} - x_3^{*p_2}) + \rho_2(x_1, x_2)\xi_2^2,$$

where $\rho_2(\cdot) \geq 0$ is a smooth function.

Clearly, one can find a smooth function $\beta_2(\cdot, \cdot) \geq 0$ such that

$$\beta_2(x_1, x_2^{p_1}) \geq n + \rho_2(x_1, x_2), \qquad \forall (x_1, x_2).$$

Then, setting $x_3^{*p_2 p_1} = -\xi_2\beta_2(x_1, x_2^{p_1})$ yields

$$\dot{V}_2 \leq -n(\xi_1^2 + \xi_2^2) + \xi_2^{2-\frac{1}{p_1}}(x_3^{p_2} - x_3^{*p_2}).$$

Following the inductive argument in [22], one can obtain a set of virtual controllers

$$\begin{array}{lll}
x_1^* & = 0, & \xi_1 = x_1 - x_1^*, \\
x_2^{*p_1} & = -\xi_1\beta_1(x_1), & \xi_2 = x_2^{p_1} - x_2^{*p_1}, \\
 & \vdots & \vdots \\
x_n^{*p_{n-1}\cdots p_1} & = -\xi_{n-1}\beta_{n-1}(x_1, x_2^{p_1}, \cdots, x_{n-1}^{p_{n-2}\cdots p_1}), & \xi_n = x_n^{p_{n-1}\cdots p_1} - x_n^{*p_{n-1}\cdots p_1}, \\
 & u^{*p_{n-1}\cdots p_1} = -\xi_n\beta_n(x_1, x_2^{p_1}, \cdots, x_n^{p_{n-1}\cdots p_1}) & \triangleq \beta(x_1, x_2^{p_1}, \cdots, x_n^{p_{n-1}\cdots p_1})
\end{array} \tag{15.29}$$

with $\beta_i : \mathbb{R}^i \to (0, +\infty)$, $i = 1, \cdots, n$, and $\beta : \mathbb{R}^n \to \mathbb{R}$ being smooth functions, and a set of C^1 Lyapunov functions

$$
\begin{aligned}
V_1(x_1) &= \frac{1}{2}\xi_1^2 \\
V_2(x_1,x_2) &= V_1(x_1) + \int_{x_2^*}^{x_2} (s^{p_1} - x_2^{*p_1})^{2-\frac{1}{p_1}}\,\mathrm{d}s \\
&\vdots \\
V_c(x_1,\cdots,x_n) &= V_{n-1}(x_1,\cdots,x_{n-1}) + \int_{x_n^*}^{x_n} (s^{p_{n-1}\cdots p_1} - x_n^{*p_{n-1}\cdots p_1})^{2-\frac{1}{p_{n-1}\cdots p_1}}\,\mathrm{d}s,
\end{aligned}
\tag{15.30}
$$

which are positive definite and proper, such that (see [22])

$$
\dot{V}_c(x) \le -2\Big(\sum_{i=1}^{n}\xi_i^2\Big) + \xi_n^{2-\frac{1}{p_{n-1}\cdots p_1}}(u-u^*). \tag{15.31}
$$

Similar to the feedback linearizable case, we define the level set $\Omega_x = \{x \in \mathbb{R}^n | V_c \le r_0 + 1\}$, where $r_0 > 0$ is a constant such that $\Gamma_x \subset \{x \in \mathbb{R}^n | V_c \le r_0\}$. Moreover, denote $M = \max_{x\in\Omega_x} ||x||_\infty$ as a saturation threshold.

As done in Section 15.3, to get an implementable controller a reduced-order observer must be designed for the estimation of the unmeasurable states $(x_2,\cdots,x_n)$ of system (15.1). Motivated by the nonsmooth observer design in [23], in what follows we shall construct a reduced-order observer to estimate, instead of $(x_2,\cdots,x_n)$, the unmeasurable variables $(z_2,\cdots,z_n)$ defined by

$$
\begin{aligned}
z_2 &= x_2^{p_1} - L_2 x_1 \qquad \Leftrightarrow \qquad x_2^{p_1} = z_2 + L_2 x_1 \\
&\vdots \\
z_n &= x_n^{p_{n-1}} - L_n x_{n-1} \quad \Leftrightarrow \qquad x_n^{p_{n-1}} = z_n + L_n x_{n-1},
\end{aligned}
\tag{15.32}
$$

where $L_i \ge 1,\ 2 \le i \le n$ are the observer gains to be assigned later.

By (15.32), the z-dynamics is described by

$$
\begin{aligned}
\dot{z}_2 &= p_1 x_2^{p_1-1}\big[x_3^{p_2} + f_2(x_1,x_2)\big] - L_2\big[x_2^{p_1} + f_1(x_1)\big] \\
\dot{z}_3 &= p_2 x_3^{p_2-1}\big[x_4^{p_3} + f_3(x_1,x_2,x_3)\big] - L_3\big[x_3^{p_2} + f_2(x_1,x_2)\big] \\
&\vdots \\
\dot{z}_n &= p_{n-1} x_n^{p_{n-1}-1}\big[u + f_n(x_1,\cdots,x_n)\big] - L_n\big[x_n^{p_{n-1}} + f_{n-1}(x_1,\cdots,x_{n-1})\big].
\end{aligned}
\tag{15.33}
$$

In view of (15.33), we design, similar to what we did in the preceding section (see (15.11)-(15.12)), the realizable observer

$$
\begin{aligned}
\dot{\hat{z}}_2 &= -L_2\big[\hat{x}_2^{p_1} + f_1(x_1)\big] \\
\dot{\hat{z}}_3 &= -L_3\big[\hat{x}_3^{p_2} + \hat{f}_2(\cdot)\big] \\
&\vdots \\
\dot{\hat{z}}_n &= -L_n\big[\hat{x}_n^{p_{n-1}} + \hat{f}_{n-1}(\cdot)\big],
\end{aligned}
\tag{15.34}
$$

where for $i = 2,\cdots,n$, (let $\hat{x}_1 = x_1$)

$$\hat{x}_i^{p_{i-1}} = \hat{z}_i + L_i \hat{x}_{i-1} \qquad \text{or} \qquad \hat{z}_i = \hat{x}_i^{p_{i-1}} - L_i \hat{x}_{i-1} \tag{15.35}$$

$$\hat{f}_i(\cdot) \triangleq f_i(x_1, \text{sat}_M(\hat{x}_2), \cdots, \text{sat}_M(\hat{x}_i)). \tag{15.36}$$

By the certainty equivalence principle, we replace the unmeasurable state $(x_2, \cdots, x_n)$ in the virtual controller u^* by the saturated state estimate $(\hat{x}_2, \cdots, \hat{x}_n)$, which is generated by the observer (15.34)-(15.36). In doing so, we obtain the implementable controller

$$u^{p_n \cdots p_1} = \hat{\beta}(\cdot) \triangleq \beta(x_1, [\text{sat}_M(\hat{x}_2)]^{p_1}, \cdots, [\text{sat}_M(\hat{x}_n)]^{p_{n-1} \cdots p_1}). \tag{15.37}$$

Define the estimate errors

$$e_i = z_i - \hat{z}_i = x_i^{p_{i-1}} - L_i x_{i-1} - \hat{z}_i, \qquad i = 2, \cdots, n. \tag{15.38}$$

Clearly,

$$x_i^{p_{i-1}} - \hat{x}_i^{p_{i-1}} = e_i + L_i(x_{i-1} - \hat{x}_{i-1}), \qquad i = 2, \cdots, n. \tag{15.39}$$

Therefore, the error dynamics can be expressed as

$$\begin{aligned}
\dot{e}_2 &= p_1 x_2^{p_1 - 1}\left[x_3^{p_2} + f_2(\cdot)\right] - L_2 e_2 \\
\dot{e}_3 &= p_2 x_3^{p_2 - 1}\left[x_4^{p_3} + f_3(\cdot)\right] - L_3 e_3 - (x_2 - \hat{x}_2) - L_3\left[f_2(\cdot) - \hat{f}_2(\cdot)\right] \\
&\vdots \\
\dot{e}_n &= p_{n-1} x_n^{p_{n-1} - 1}\left[u + f_n(\cdot)\right] - L_n e_n - (x_{n-1} - \hat{x}_{n-1}) - L_n\left[f_{n-1}(\cdot) - \hat{f}_{n-1}(\cdot)\right].
\end{aligned} \tag{15.40}$$

To analyze the error dynamics, we introduce several useful propositions that can be proved by straightforward but tedious calculations. The detailed proofs are included in the appendix.

First of all, using (15.39) and Lemma 15.2.3 repeatedly, it is easy to show the existence of a generic constant $K \geq 1$, which depends on M and is independent of all the L_i's, such that on the set $B_M \times \mathbb{R}^{n-1}$ $(x \in B_M,\ (\hat{x}_2, \cdots \hat{x}_n) \in \mathbb{R}^{n-1})$, the following estimations hold on the set $B_M \times \mathbb{R}^{n-1}$ for $i = 2, \cdots, n$: (see (A.10) in [23])

$$|x_i - \hat{x}_i| \leq K(|e_i|^{\frac{1}{p_{i-1}}} + L_i^{\frac{1}{p_{i-1}}} |e_{i-1}|^{\frac{1}{p_{i-1} p_{i-2}}} + \cdots + L_i^{\frac{1}{p_{i-1}}} \cdots L_3^{\frac{1}{p_{i-1} \cdots p_2}} |e_2|^{\frac{1}{p_{i-1} \cdots p_1}}). \tag{15.41}$$

Proposition 15.4.1. *There exists a constant $\varepsilon_0 > 0$, which depends on M and is independent of all the L_i's, such that on the set $B_M \times \mathbb{R}^{n-1}$,*

$$\dot{V}_c\Big|_{B_M \times \mathbb{R}^{n-1}} \leq -3\varepsilon_0(x_1^2 + \cdots + x_n^{2p_{n-1} \cdots p_1}) + \xi_n^{2 - \frac{1}{p_{n-1} \cdots p_1}}(u - u^*). \tag{15.42}$$

Proposition 15.4.2. *There exists a generic constant $K \geq 1$, which depends on M and is independent of all the L_i's, such that on the set $B_M \times \mathbb{R}^{n-1}$, the following estimations hold: $(i = 2, \cdots, n)$*

$$|f_i(\cdot)-\hat{f}_i(\cdot)|\Big|_{B_M\times\mathbb{R}^{n-1}} \le K[|e_i|^{\frac{1}{p_{i-1}}}+\sum_{j=2}^{i-1}(L_i^{\frac{1}{p_{i-1}}}\cdots L_{j+1}^{\frac{1}{p_{i-1}\cdots p_j}}|e_j|^{\frac{1}{p_{i-1}\cdots p_{j-1}}})], \quad (15.43)$$

$$|f_i(\cdot)|\Big|_{B_M\times\mathbb{R}^{n-1}} \le K(|x_1|^{\frac{1}{p_{i-1}\cdots p_1}}+\cdots+|x_{i-1}|^{\frac{1}{p_{i-1}}}+|x_i|). \quad (15.44)$$

Proposition 15.4.3. *Given $\varepsilon_0>0$, there is a generic constant $K\ge 1$, which depends on M and is independent of all the L_i's, such that on the set $B_M\times\mathbb{R}^{n-1}$,*

$$|u-u^*| \le \varepsilon_0(|x_1|^{\frac{1}{p_{n-1}\cdots p_1}}+\cdots+|x_{n-1}|^{\frac{1}{p_{n-1}}}+|x_n|)+K\min\Big\{1,\ |e_n|^{\frac{1}{p_{n-1}}} +L_n^{\frac{1}{p_{n-1}}}|e_{n-1}|^{\frac{1}{p_{n-1}p_{n-2}}}+\cdots+L_n^{\frac{1}{p_{n-1}}}\cdots L_3^{\frac{1}{p_{n-1}\cdots p_2}}|e_2|^{\frac{1}{p_{n-1}\cdots p_1}}\Big\} \quad (15.45)$$

$$|u| \le K\Big[(|x_1|^{\frac{1}{p_{n-1}\cdots p_1}}+\cdots+|x_{n-1}|^{\frac{1}{p_{n-1}}}+|x_n|)+|e_n|^{\frac{1}{p_{n-1}}} +L_n^{\frac{1}{p_{n-1}}}|e_{n-1}|^{\frac{1}{p_{n-1}p_{n-2}}}+\cdots+L_n^{\frac{1}{p_{n-1}}}\cdots L_3^{\frac{1}{p_{n-1}\cdots p_2}}|e_2|^{\frac{1}{p_{n-1}\cdots p_1}}\Big]. \quad (15.46)$$

In view of the relationship $B_M\times\mathbb{R}^{n-1}\supset\Omega_x\times\mathbb{R}^{n-1}\supset\Omega$ (see Figure 15.2), it is not difficult to deduce from Proposition 15.4.1 and (15.45) that (by the Young's inequality)

$$\dot{V}_c\Big|_{\Omega} \le -2\varepsilon_0(x_1^2+\cdots+x_n^{2p_{n-1}\cdots p_1}) \quad (15.47)$$
$$+\min\Big\{C_n e_n^{2p_{n-2}\cdots p_1}+C_{n-1}(L_n)e_{n-1}^{2p_{n-3}\cdots p_1}+\cdots+C_2(L_n,\cdots,L_3)e_2^2,\ C_n\Big\},$$

where $\varepsilon_0>0, C_n\ge 1$ are constants independent of L_i's while

$$C_{n-1}(L_n)\ge C_n,\ \cdots,\ C_2(L_n,\cdots,L_3)\ge C_n$$

are fixed polynomial functions of their arguments. They can be obtained in a manner similar to the one in [23].

For the error dynamics (15.40), consider the Lyapunov function

$$V_e=\frac{1}{2p_{n-2}\cdots p_1}e_n^{2p_{n-2}\cdots p_1}+\cdots+\frac{1}{2p_1}e_3^{2p_1}+\frac{1}{2}e_2^2. \quad (15.48)$$

By Propositions 15.4.2 and 15.4.3, the derivative of V_e along the trajectories of (15.40) on the set $B_M\times\mathbb{R}^{n-1}$ satisfies

$$\begin{aligned}\dot{V}_e &\le |e_n|^{2p_{n-2}\cdots p_1-1}\Big(p_{n-1}|x_n|^{p_{n-1}-1}\big(|u|+|f_n(\cdot)|\big)+|x_{n-1}-\hat{x}_{n-1}| \\ &\quad +L_n|f_{n-1}(\cdot)-\hat{f}_{n-1}(\cdot)|\Big) \\ &\quad -L_n e_n^{2p_{n-2}\cdots p_1}+\cdots+|e_2|\Big(p_1|x_2|^{p_1-1}\Big[|x_3|^{p_2}+|f_2(\cdot)|\Big]\Big)-L_2e_2^2 \\ &\le K(x_1^2+\cdots+x_n^{2p_{n-1}\cdots p_1})-[L_n-\bar{C}_n]e_n^{2p_{n-2}\cdots p_1} \\ &\quad -[L_{n-1}-\bar{C}_{n-1}(L_n)]e_{n-1}^{2p_{n-3}\cdots p_1}-\cdots-[L_2-\bar{C}_2(L_n,\cdots,L_3)]e_2^2, \end{aligned} \quad (15.49)$$

where K and $\bar{C}_n$ are positive constants independent of L_i's, while $\bar{C}_{n-1}(L_n)>0,\cdots,\bar{C}_2(L_n,\cdots,L_3)>0$ are fixed polynomial functions of their arguments, which can be computed using a similar argument as done in [23].

From (15.49), it is clear that by choosing

$$\begin{aligned} L_n &= L_n(L) \triangleq \bar{C}_n + LC_n \geq 1 \\ L_{n-1} &= L_{n-1}(L) \triangleq \bar{C}_{n-1}(L_n) + LC_{n-1}(L_n) \geq 1 \\ &\vdots \\ L_2 &= L_2(L) \triangleq \bar{C}_2(L_n, \cdots, L_3) + LC_2(L_n, \cdots, L_3) \geq 1 \end{aligned} \tag{15.50}$$

with $L > 0$ being a parameter to be determined later, one has

$$\dot{V}_e\Big|_{B_M \times \mathbb{R}^{n-1}} \leq K(x_1^2 + \cdots + x_n^{2p_{n-1}\cdots p_1}) - LW_e, \tag{15.51}$$

where

$$W_e \triangleq C_n e_n^{2p_{n-2}\cdots p_1} + C_{n-1}(L) e_{n-1}^{2p_{n-3}\cdots p_1} + \cdots + C_2(L) e_2^2 \tag{15.52}$$

and $C_{n-1}(L) \geq C_n, \cdots, C_2(L) \geq C_n$ are fixed polynomial functions of L.

For the closed-loop system (15.1) and (15.34)-(15.37), we choose the Lyapunov function

$$V(x, \hat{z}) = V_c(x) + \frac{\ln(1 + V_e(e))}{\ln(1 + \mu(L))}, \tag{15.53}$$

where (note that $e_i = z_i - \hat{z}_i = x_i^{p_{i-1}} - L_i x_{i-1} - \hat{z}_i,\ i = 2, \cdots, n$)

$$\begin{aligned} \mu(L) &\triangleq \frac{1}{2}\Big[(r^{p_1} + L_2(L)r + r)^2 + (r^{p_2} + L_3(L)r + r)^{2p_1} + \\ &\qquad \cdots + (r^{p_{n-1}} + L_n(L)r + r)^{2p_{n-2}\cdots p_1}\Big] \\ &\geq \max_{(x,\hat{z}) \in \Gamma_x \times \Gamma_{\hat{z}}} V_e > 0. \end{aligned}$$

Associated with $V(x, \hat{z})$, define the level set

$$\Omega = \{(x, \hat{z}) \in \mathbb{R}^n \times \mathbb{R}^{n-1} | V(x, \hat{z}) \leq r_0 + 1\}. \tag{15.54}$$

Similar to Section 15.3, for every $L > 0$, $V(\cdot)$ is a positive definite and proper function and Ω is a compact set in $\mathbb{R}^n \times \mathbb{R}^{n-1}$. Moreover, by construction, $B_M \times \mathbb{R}^{n-1} \supset \Omega \supset \Gamma_x \times \Gamma_{\hat{z}},\ \forall L > 0$. Hence, it follows from (15.51) and (15.47) that $\forall L > 0$,

$$\begin{aligned} \dot{V}\Big|_{\Omega} &= \dot{V}_c\Big|_{\Omega} + \frac{1}{\ln(1 + \mu(L))} \frac{\dot{V}_e}{1 + V_e}\Big|_{\Omega} \\ &\leq -\Big[2\varepsilon_0 - \frac{K}{\ln(1 + \mu(L))(1 + V_e)}\Big](x_1^2 + \cdots + x_n^{2p_{n-1}\cdots p_1}) \\ &\quad -\Big[\frac{L}{2\ln(1 + \mu(L))} - 1\Big] \min\{W_e, C_n\}. \end{aligned} \tag{15.55}$$

The remaining part of the proof is to find a suitable constant L such that $\dot{V}\big|_{\Omega} \leq 0$. Recall that $\mu(L)$ is a fixed polynomial function of L and the constants K and ε_0 are independent of L. Hence, there exists a constant $L^* > 0$ such that

$$\frac{K}{\ln(1+\mu(L))} \leq \varepsilon_0 \quad \text{and} \quad \frac{L}{2\ln(1+\mu(L))} \geq 2, \qquad \forall L \geq L^*.$$

Choosing $L = L^*$, we have

$$\dot{V}\Big|_{\Omega} \leq -\varepsilon_0(x_1^2 + \cdots + x_n^{2p_{n-1}\cdots p_1}) - \min\{W_e, C_n\}.$$

In summary, system (15.1) is semi-globally stabilizable by the nonsmooth dynamic output compensator (15.34)-(15.37), although it is non-uniformly observable and non-smoothly stabilizable. ■

The significance of Theorem 15.1.1 can be seen easily from the following example.

Example 15.4.1. Consider the three-dimensional nonlinear system

$$\begin{aligned} \dot{x}_1 &= x_2^3 + x_1 e^{x_1} \\ \dot{x}_2 &= x_3 \\ \dot{x}_3 &= u \\ y &= x_1, \end{aligned} \tag{15.56}$$

which is not smoothly stabilizable for the well-known reason that the linearized system

$$(A, B) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right)$$

is uncontrollable and the uncontrollable mode is associated with an eigenvalue on the open right-half plane. In addition, the nonlinear system (15.56) is neither uniformly observable because

$$x_2 = (\dot{y} - ye^y)^{1/3} \qquad \text{and} \qquad x_3 = \frac{\ddot{y} - \dot{y}e^y - ye^y\dot{y}}{3(\dot{y} - ye^y)^{2/3}}.$$

Indeed, both functions are non-smooth. Moreover, x_3 is a singular function of $(y, \dot{y}, \ddot{y})$, i.e. it is not well-defined on the manifold characterized by $\dot{y} = ye^y$. Therefore, the semi-global design method proposed in [26, 27] is invalid.

On the other hand, the work [23] gave only a *local* output feedback stabilization result due to the non-homogeneous term $x_1 e^{x_1}$. However, according to Theorem 15.1.1, we now know that the non-uniformly observable and non-smoothly stabilizable system (15.56) is semi-globally stabilizable via Hölder continuous output feedback, for instance, by the output dynamic compensator (15.34)-(15.37). ■

We conclude this section by illustrating, via an example, a possible extension of Theorem 15.1.1 to a class of cascade systems involving zero dynamics.

Example 15.4.2. Consider the cascade system

$$\begin{aligned}\dot{\zeta} &= -(1+\sin^2 v)\zeta + \eta_1\\ \dot{\eta}_1 &= \eta_2^3\\ \dot{\eta}_2 &= v + \ln(1+\zeta^2)\eta_2^2\\ y &= \eta_1.\end{aligned} \tag{15.57}$$

Although (15.57) is not precisely in the form (15.1) due to the presence of the zero-dynamics and the term $\ln(1+\zeta^2)\eta_2^2$, the output feedback design method proposed in Theorem 15.1.1 is still applicable. Indeed, it can be proved that the cascade system (15.57), which is non-smoothly stabilizable and non-uniformly observable, is semi-globally stabilizable by the non-smooth dynamic output compensator (15.34)-(15.37) as long as the parameter L is large enough.

To see why this is the case, we notice that (15.57) is stabilizable by partial state feedback, i.e., via the states η_1 and η_2, because the ζ-subsystem satisfies the ISS condition. Moreover, given the prescribed compact domain $\Gamma \triangleq \{(\zeta,\eta_1,\eta_2) \mid |\zeta| \le r, |\eta_1| \le r, |\eta_2| \le r\}$ with $r > 0$, one can find, using the nonsmooth partial-state feedback design method [20], a Lyapunov function of the form

$$V_c(\zeta,\eta_1,\eta_2) = \frac{\zeta^2}{2} + \frac{1}{2}(\eta_1^2 + \xi_2^2), \qquad \xi_2 = \eta_2^3 + a_1\eta_1,$$

the associated level set

$$\Omega_\eta = \{(\zeta,\eta_1,\eta_2) | V_c \le r_0 + 1\} \supset \{(\zeta,\eta_1,\eta_2) | V_c \le r_0\} \supset \Gamma$$

and a partial-state feedback controller $u^{*3} = -a_2\xi_2 = -[a_2(\eta_2^3 + a_1\eta_1)]$, such that

$$\dot{V}_c\Big|_{\Omega_\eta} \le -2(\zeta^2 + \eta_1^2 + \xi_2^2) - \xi_2^{2-\frac{1}{3}}(u - u^*),$$

where both a_1 and a_2 are positive constants.

Using the similar feedback design method in this section, one can construct a dynamic output compensator of the form

$$\dot{\hat{z}}_2 = -L\hat{\eta}_2^3, \qquad \text{with} \qquad \hat{\eta}_2^3 = \hat{z}_2 + L\eta_1, \tag{15.58}$$

$$u = -\Big(a_2([\mathrm{sat}_M(\hat{\eta}_2)]^3 + a_1\eta_1)\Big)^{\frac{1}{3}} \tag{15.59}$$

where $M = \max_{(z,\eta_1,\eta_2)\in\Omega_\eta} ||(z,\eta_1,\eta_2)||_\infty$.

The simulation shown in Figure 15.3 demonstrates the transient response of the closed-loop system with the initial condition $(\zeta,\eta_1,\eta_2,\hat{z}_2) = (2,-1,-2,1)$ and $a_1 = 2,\ a_2 = 7,\ M = 5,\ L = 5$. ■

15.5 A Class of Non-triangular Systems in the Plane

In this section, we present a generalization of Theorem 15.1.1 to a class of planar systems beyond a strict-triangular form. The nonlinear system under consideration is described by equations of the form

$$\begin{aligned}\dot{\eta}_1 &= \eta_2^p + \eta_2^{p-1}\phi_{p-1}(\eta_1) + \cdots + \eta_2\phi_1(\eta_1) + \phi_0(\eta_1)\\ \dot{\eta}_2 &= v\\ y &= \eta_1,\end{aligned} \tag{15.60}$$

where $(\eta_1, \eta_2) \in \mathbb{R}^2$, $v \in \mathbb{R}$ and $y \in \mathbb{R}$ are the system state, input and output, respectively, p is an odd positive integer and the mappings $\phi_i : \mathbb{R} \to \mathbb{R}$, $i = 0, \cdots, p-1$ are C^1 with $\phi_i(0) = 0$.

It is of interest to note that the form (15.60) is representative of a class of two-dimensional affine systems. In fact, Jakubczyk and Respondek [12] proved that every smooth affine system in the plane, i.e.,

$$\dot{\xi} = f(\xi) + g(\xi)u,$$

is feedback equivalent to the system (15.60) without η_2^{p-1} term if $g(0)$ and $\mathrm{ad}_f^p g(0)$ are linearly independent. A more general characterization was given in [2] later on, showing that the equation (15.60) is indeed a special case of the so-called "p-normal form" or Hessenberg form [2]. In other words, system (15.60) is a *normal form* of two-dimensional affine systems when $\mathrm{rank}[g(0), \mathrm{ad}_f^p g(0)] = 2$.

A distinguished feature of the planar system (15.60) is that it is in general neither uniformly observable nor smoothly stabilizable when $p > 1$, because on the one hand, the state of (15.60) can only be represented as a Hölder continuous rather than smooth function of the system input, output, and their derivatives; on the other hand, the linearized system of (15.60) may have uncontrollable modes associated with eigenvalues on the right-half plane. These points can be seen easily, for instance, from the simple Example 15.5.1. In addition, system (15.60) is not in a triangular form. The lack of uniform observability, smooth stabilizability and triangular structure makes the semi-global stabilization of

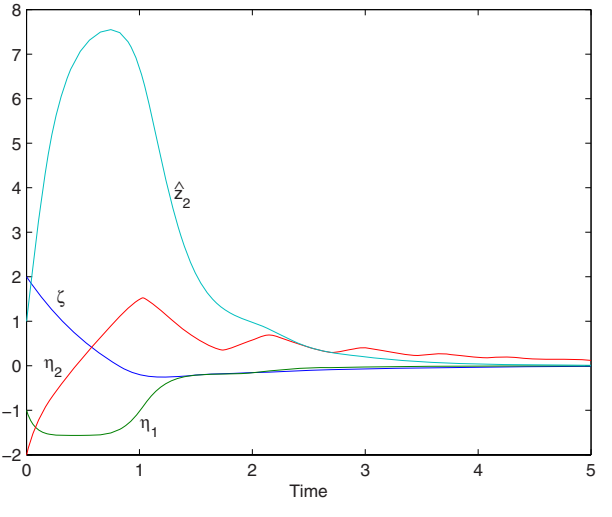


Fig. 15.3. The transient response of the closed-loop system (15.57)-(15.58)-(15.59)

(15.60) via output feedback extremely difficult. In fact, all the existing output feedback stabilization results in the literature (e.g., [26, 27, 11]) and Theorem 15.1.1 can not be applied to the planar system (15.60).

Despite of the aforementioned difficulties, it was shown in [23] that the planar system (15.60) is locally stabilizable by non-smooth output feedback. In the global case, output feedback stabilization of (15.60) was only possible provided that some very restrictive growth conditions are fulfilled [29]. In what follows, we prove that similar to the local case, semi-global stabilization of the planar system (15.60) can be achieved by non-smooth output feedback, without requiring any growth condition such as Assumption 3.1 in [29].

Theorem 15.5.1. *There exists a nonsmooth dynamic output compensator of the form (15.2), which semi-globally stabilizes the planar systems (15.60).*

Proof. Motivated by the work [29], we introduce a rescaling transformation with a suitable dilation for the planar system (15.60), which turns out to be crucial for handling the non-triangular structure of (15.60). To be precise, let $x_1 = \eta_1, \quad x_2 = \frac{\eta_2}{L}, \quad u = \frac{v}{L^{1+p}}$, where $L \geq 1$ is a rescaling factor to be assigned later.

Under the x-coordinate, the original system (15.60) is rewritten as

$$\begin{aligned} \dot{x}_1 &= L^p x_2^p + L^{p-1} x_2^{p-1} \phi_{p-1}(x_1) + \cdots + L x_2 \phi_1(x_1) + \phi_0(x_1) \\ \dot{x}_2 &= L^p u \\ y &= x_1. \end{aligned} \tag{15.61}$$

Similar to the design in [22], one can construct the globally stabilizing, non-smooth state feedback controller $\tilde{x}_3^* = -[\xi_2 \beta_2(x_1, x_2^p)]^{1/p}$ and the Lyapunov function $V_c(x_1, x_2) = \frac{1}{2}x_1^2 + \int_{x_2^*}^{x_2} (s^p - x_2^{*p})^{2-\frac{1}{p}} \mathrm{d}s$, which is positive definite and proper [22], such that

$$\dot{V}_c(x_1, x_2) \leq L^p[-3(x_1^2 + \xi_2^2) + \xi_2^{2-\frac{1}{p}}(u - \tilde{x}_3^*)], \tag{15.62}$$

where $x_2^{*p} = -\beta_1(x_1)x_1$ and $\xi_2 = x_2^p - x_2^{*p} = x_2^p + \beta_1(x_1)x_1$ with $\beta_i : R^i \to (0, +\infty)$, $i = 1, 2$ being positive smooth functions independent of L.

In the coordinate of $x = (x_1, x_2)$, we define the level set $\Omega_x = \{(x_1, x_2) | V_c(x_1, x_2) \leq r_0 + 1\}$, where $r_0 > 0$ is a constant such that

$$\Gamma_\eta \subset \Gamma_x \triangleq \{(x_1, x_2) \mid |x_1| = |\eta_1| \leq r, |x_2| = |\frac{\eta_2}{L}| \leq r\} \subset \{(x_1, x_2) | V_c(x_1, x_2) \leq r_0\}.$$

Moreover, denote $M = \max_{x \in \Omega_x} ||x||_\infty$ as a saturation threshold, which is independent of L (see Figure 15.4).

Clearly, there exists a constant $B_2 > 0$ independent of L, satisfying

$$0 < [\beta_2(x_1, x_2^p)]^{\frac{1}{p}} \leq B_2, \qquad \forall (x_1, x_2) \in \Omega_x. \tag{15.63}$$

Therefore, on the level set Ω_x, the controller $\tilde{x}_3^*$ can be replaced by

$$x_3^* = -B_2 \xi_2^{\frac{1}{p}} \triangleq [u^*(x_1, x_2^p)]^{1/p}.$$

Because of (15.63)-(15.62), the derivative of V_c on Ω_x satisfies

$$\dot{V}_c\Big|_{\Omega_x} \leq L^p[-3(x_1^2 + \xi_2^2) + \xi_2^{2-\frac{1}{p}}(u - x_3^*)]. \tag{15.64}$$

Next, we construct a reduced-order observer to estimate, instead of x_2 itself, the unmeasurable variable $z_2 = x_2^p - Lx_1$. In view of z_2's dynamics

$$\dot{z}_2 = pL^p x_2^{p-1} u - L\big[L^p x_2^p + L^{p-1} x_2^{p-1} \phi_{p-1}(x_1) + \cdots + \phi_0(x_1)\big], \tag{15.65}$$

we design the one-dimensional observer

$$\dot{\hat{z}}_2 = -L\big[L^p \hat{x}_2^p + L^{p-1} \hat{x}_2^{p-1} \phi_{p-1}(x_1) + \cdots + \phi_0(x_1)\big] \quad \text{with} \quad \hat{x}_2 = (\hat{z}_2 + Lx_1)^{\frac{1}{p}}. \tag{15.66}$$

By the certainty equivalence principle, we replace the unmeasurable state x_2 in the virtual controller x_3^* by the saturated state estimate $\hat{x}_2$ from the observer (15.66). In this way, we obtain

$$u = \Big(u^*(x_1, [\mathrm{sat}_M(\hat{x}_2)]^p)\Big)^{1/p} = -B_2\Big([\mathrm{sat}_M(\hat{x}_2)]^p + \beta_1(x_1)x_1\Big)^{1/p}, \tag{15.67}$$

or equivalently,

$$v = L^{1+p}\Big[u^*(\eta_1, \mathrm{sat}_{M^p}(\hat{z}_2 + L\eta_1))\Big]^{\frac{1}{p}} = -B_2 L^{1+p}\Big[\mathrm{sat}_{M^p}(\hat{z}_2 + L\eta_1) + \beta_1(\eta_1)\eta_1\Big]^{\frac{1}{p}}. \tag{15.68}$$

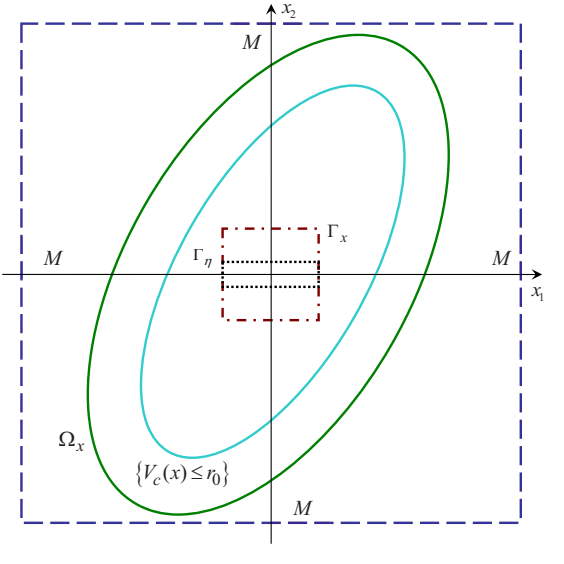


Fig. 15.4. The level set on (x_1, x_2) or (η_1, η_2)-space.

Finally, we prove that the non-smooth output feedback controller thus constructed, i.e., (15.66)-(15.68), semi-globally stabilizes system (15.60). To this end, let

$$e_2 = z_2 - \hat{z}_2 = x_2^p - \hat{x}_2^p = \frac{\eta_2^p}{L^p} - L\eta_1 - \hat{z}_2 \tag{15.69}$$

be the estimate error. Then, the error dynamics are characterized by

$$\dot{e}_2 = L^p[px_2^{p-1}u - Le_2 - (x_2^{p-1} - \hat{x}_2^{p-1})\phi_{p-1}(x_1) - \cdots - L^{-(p-2)}(x_2 - \hat{x}_2)\phi_1(x_1)]. \tag{15.70}$$

For the closed-loop system (15.60)-(15.66)-(15.68), consider the Lyapunov function

$$V(\eta, \hat{z}_2) = V_c(x) + V_e(e) = V_c(\eta_1, \frac{\eta_2}{L}) + \frac{\ln(1+e_2^2)}{\ln(1+(r^p + Lr + r)^2)}.$$

The corresponding level set can be defined as

$$\Omega = \{(\eta, \hat{z}_2) \in \mathbb{R}^3 \big| V(\eta, \hat{z}_2) \leq r_0 + 1\}.$$

In view of the definition of e_2 in (15.69), it is easy to verify the level set Ω contains the prescribed attractive domain $\Gamma_\eta \times \Gamma_{\hat{z}}$ uniformly with respect to $L \geq 1$, because $(\eta_1, \eta_2, \hat{z}_2) \in \Gamma_\eta \times \Gamma_{\hat{z}}$ implies $V(\eta_1, \eta_2, \hat{z}_2) \leq r_0 + 1,\ \forall L \geq 1$.

From the definition of M, it is clear that

$$|x_1| = |\eta_1| \leq M, \quad |x_2| = |\frac{\eta_2}{L}| \leq M, \quad \forall (\eta_1, \eta_2, \hat{z}_2) \in \Omega.$$

The relationships above are illustrated in Figure 15.5.

By Lemma 15.2.6, the boundedness of (x_1, x_2)-coordinate on the level set Ω implies there exists a generic constant $K > 0$ independent of L such that

$$\begin{aligned} \Big| \big[u^*(x_1, [\mathrm{sat}_M(\hat{x}_2)]^p)\big]^{\frac{1}{p}} - \big[u^*(x_1, x_2^p)\big]^{\frac{1}{p}} \Big|\Big|_\Omega &\leq 2^{1-\frac{1}{p}} |u^*(x_1, [\mathrm{sat}_M(\hat{x}_2)]^p) \\ - u^*(x_1, x_2^p)|^{\frac{1}{p}}\Big|_\Omega &\leq K \min\{|e_2|^{\frac{1}{p}}, 1\}. \end{aligned} \tag{15.71}$$

In view of (15.64), (15.71) and Young's inequality, we have

$$\dot{V}_c\Big|_\Omega \leq L^p\Big[-2(x_1^2 + \xi_2^2) + K \min\{e_2^2, 1\}\Big], \tag{15.72}$$

where $K > 0$ is a generic constant independent of L.

Similarly, by smoothness of $\phi_i(x_1)$ and $\phi_i(0) = 0$, it is easy to see that $\forall L \geq 1$,

$$|L^{-(i-p-1)}\phi_i(x_1)|\Big|_\Omega \leq K|x_1|^{1-\frac{i}{p}}, \qquad i = 0, \cdots, p-1, \tag{15.73}$$

where $K > 0$ is a generic constant independent of L.

Hence, by the Young's inequality, a direct but tedious calculation gives

$$\dot{V}_e\Big|_{\Omega} \le \frac{L^p\Big[2p|x_2|^{p-1}|e_2||u| - 2Le_2^2 + 2C|e_2|^{2-\frac{1}{p}}|x_1|^{\frac{1}{p}} + \cdots + 2C|e_2|^{2-\frac{p-1}{p}}|x_1|^{1-\frac{1}{p}}\Big]}{(1+e_2^2)\ln(1+(r^p+Lr+r)^2)}$$
$$\le \frac{L^p\big[-(2L-K)e_2^2 + (x_1^2+\xi_2^2)\big]}{(1+e_2^2)\ln(1+(r^p+Lr+r)^2)}$$
$$\le L^p\Big[-\frac{2L-K}{\ln(1+(r^p+Lr+r)^2)}\frac{e_2^2}{1+e_2^2} + \frac{x_1^2+\xi_2^2}{\ln(1+(r^p+Lr+r)^2)}\Big] \tag{15.74}$$

with $K>0$ being a generic constant independent of L.

Observe that $\frac{e_2^2}{1+e_2^2} \ge \frac{1}{2}\min\{e_2^2, 1\}$. Then, putting (15.72) and (15.74) together results in

$$\dot{V}\Big|_{\Omega} \le L^p\Big[-(2-\frac{1}{\ln(1+(r^p+Lr+r)^2)})(x_1^2+\xi_2^2)$$
$$-\big(\frac{L}{\ln(1+(r^p+Lr+r)^2)} - K\big)\min\{e_2^2,1\}\Big] \tag{15.75}$$

with $K>0$ being a generic constant independent of L.

Since $1+(r^p+Lr+r)^2$ is a quadratic polynomial of L, $\lim_{L\to+\infty}\frac{L}{\ln(1+(r^p+Lr+r)^2)} = +\infty$. As a consequence, there exists a sufficiently large L such that

$$2-\frac{1}{\ln(1+(r^p+Lr+r)^2)} \ge 1 \qquad \text{and} \qquad \frac{L}{\ln(1+(r^p+Lr+r)^2)} - K \ge 1,$$

which, in turn, yields

$$\dot{V}\Big|_{\Omega} \le L^p\Big[-(x_1^2+\xi_2^2) - \min\{e_2^2,1\}\Big].$$

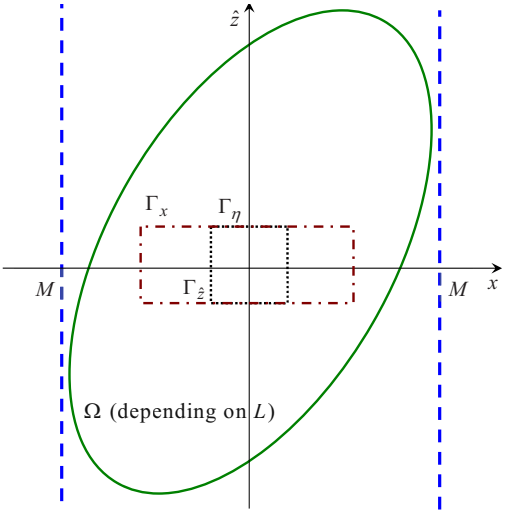


Fig. 15.5. The level set on $(x_1, x_2, \hat{z})$ or $(\eta_1, \eta_2, \hat{z})$-space.

This completes the proof of Theorem 15.5.1. ■

The effectiveness of Theorem 15.5.1 can be demonstrated by the following example.

Example 15.5.1. Consider the non-triangular system

$$\begin{aligned} \dot{\eta}_1 &= \eta_2^3 + \eta_2\eta_1 + \eta_1 \\ \dot{\eta}_2 &= v \\ y &= \eta_1. \end{aligned} \tag{15.76}$$

Due to the lack of a triangular structure, the semi-global design method proposed in Section 15.3 is no longer valid. However, the planar system (15.76) is of the form (15.60). By Theorem 15.5.1, it is semi-globally stabilizable by non-smooth output feedback.

In fact, one can design the dynamic output compensator

$$\begin{aligned} \dot{\hat{z}}_2 &= -L^4[\hat{z}_2 + L\eta_1] - L^2(\hat{z}_2 + L\eta_1)^{\frac{1}{3}} - L\eta_1 \\ v &= -B_2L^4\Big[\mathrm{sat}_{M^3}(\hat{z}_2 + L\eta_1) + \eta_1\frac{21+\eta_1^2}{4}\Big]^{\frac{1}{3}} \end{aligned} \tag{15.77}$$

which semi-globally stabilizes the non-triangular system (15.76), where $B_2,\ M > 0$ and $L \geq 1$ are appropriate constants. This conclusion can be verified by using the Lyapunov function

$$V(\eta_1, \eta_2, \hat{z}_2) = \frac{1}{2}\eta_1^2 + \int_{x_2^*}^{\frac{\eta_2}{L}} (s^3 - x_2^{*3})^{\frac{5}{3}} + \frac{(\frac{\eta_2^3}{L^3} - L\eta_1 - \hat{z}_2)^2}{1 + (r^p + Lr + L)^2}$$

with $L \geq 1$ and $x_2^* = [-\frac{\eta_1}{4}(\eta_1^2 + 21)]^{\frac{1}{3}}$.

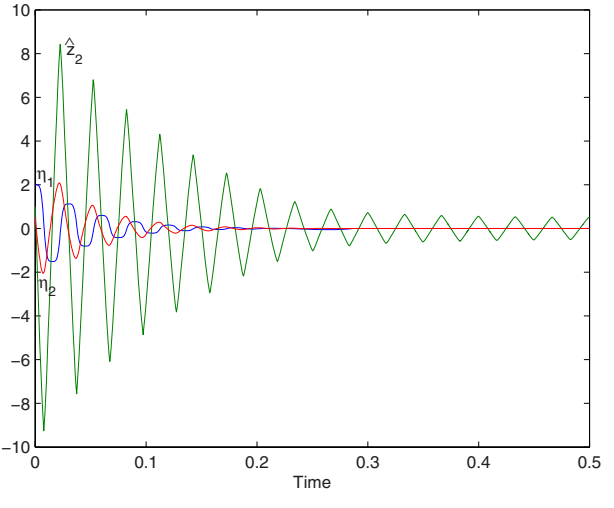


Fig. 15.6. The transient response of the closed-loop system (15.76)-(15.77)

The simulation shown in Figure 15.6 demonstrates the transient response of the closed-loop system (15.76)-(15.77) with the initial condition $(\eta_1, \eta_2, \hat{z}_2) = (2,\ 1,\ 0.5)$ and $B_2 = 7,\ M = 5,\ L = 3$. ■

We end this section with a remark that achieving semi-global stabilization of higher-dimensional nonlinear systems in the Hessenberg form or the "p-normal form" by output feedback [2] is a far more difficult and complex problem than in the two-dimensional case. Due to the loss of a strict-triangular structure, the semi-global control scheme developed in Section 15.4 cannot adopted directly. On the other hand, the approach presented in this section for the planar system (15.60) is an ad hoc one and does not appear to be extendable to the n-dimensional case, because of some inherent difficulties in the construction of an $(n-1)$-dimensional, reduced-order nonlinear observer. This problem is currently under our investigation and we hope to be able to say more about it in the future.

15.6 Conclusion

In this paper, we have proved that *without requiring uniform observability and smooth stabilizability* by state feedback, it is still possible to achieve semi-global stabilization via non-smooth output feedback, for a significant class of nonlinear systems such as (15.1) and (15.60). This was made possible by developing a non-smooth semi-global output feedback control scheme, which significantly extended the semi-global stabilization results obtained in [26, 27] and integrated the recursive nonsmooth observer design algorithm [23] with the idea of saturating the estimated state [16, 26].

In the case when the nonlinear system is smoothly stabilizable but not necessarily uniformly observable (for instance, the class of C^0 triangular systems (15.5) satisfying the condition (15.6)), the result of Section 15.3 has provided a refinement of the existing work [26, 27, 10, 11] by relaxing the uniform observability condition. As a byproduct, the idea used in this paper has also led to a simple and intuitive Lyapunov argument that enables us to establish semi-global stabilizability for multi-input, multi-output nonlinear systems [30].

Acknowledgement. This work was supported in part by the NSF under grant ESC-0400413, the Herbold Faculty Fellow Award, and the 111 project (B08015).

Appendices

15.A Proof of Proposition 15.4.1

Using (15.29), one has

$$(\xi_{n-1}^2+\xi_n^2)\Big|_{B_M\times\mathbb{R}^{n-1}} = \xi_{n-1}^2+(x_n^{p_{n-1}\cdots p_1}+\xi_{n-1}\beta_{n-1}(\cdot))^2$$
$$\geq \frac{1}{2}\xi_{n-1}^2+\frac{1}{2\beta_{n-1}^2(\cdot)+1}x_n^{2p_{n-1}\cdots p_1}. \qquad (15.78)$$

Note that on the set $B_M\times\mathbb{R}^{n-1}$, $\beta_i(\cdot),\ i=1,\cdots,n-1$ are bounded by a constant independent of all L_i's. Using the above argument repeatedly, it is concluded that there is a constant $\varepsilon_0>0$ satisfying

$$(\xi_1^2+\xi_2^2+\cdots+\xi_n^2)\Big|_{B_M\times\mathbb{R}^{n-1}} \geq \frac{3}{2}\varepsilon_0(x_1^2+x_2^{2p_1}+\cdots+x_n^{2p_{n-1}\cdots p_1}). \qquad (15.79)$$

Thus, (15.42) follows from (15.31) and (15.79). ■

15.B Proof of Proposition 15.4.2

By definition, $\hat{f}_i(\cdot)\equiv f_i(x_1,\mathrm{sat}_M(\hat{x}_2),\cdots,\mathrm{sat}_M(\hat{x}_i)),\ i=2,\cdots,n$ are smooth functions of the variables $x_1,\mathrm{sat}_M(\hat{x}_2),\cdots,\mathrm{sat}_M(\hat{x}_i)$ which, except x_1, satisfy $|\mathrm{sat}_M(\hat{x}_j)|\leq M,\ j=2,\cdots,i$ uniformly with respect to L_i's. These observations, together with Lemmas 15.2.5-15.2.6, imply the existence of a fixed constant $K\geq 1$, which depends on M and is independent of all the L_i's, such that on the set $B_M\times\mathbb{R}^{n-1}=\{(x,\hat{z})\in\mathbb{R}^n\times\mathbb{R}^{n-1}|(x_1,\cdots,x_n)\in[-M,M]^n\}$, the following estimations hold: (pick $\sigma_j=\frac{1}{p_{i-1}\cdots p_j},\ j=2,\cdots,i$)

$$|f_i(\cdot)-\hat{f}_i(\cdot)|\Big|_{B_M\times\mathbb{R}^{n-1}} \leq \frac{K}{n}\Big(\sum_{j=2}^{i}|x_j-\mathrm{sat}_M(\hat{x}_j)|^{\frac{1}{p_{i-1}\cdots p_2}}\Big) \qquad (15.80)$$
$$\leq \frac{K}{n}\Big(|x_2-\hat{x}_2|^{\frac{1}{p_{i-1}\cdots p_2}}+\cdots+|x_{i-1}-\hat{x}_{i-1}|^{\frac{1}{p_{i-1}}}$$
$$+|x_i-\hat{x}_i|\Big).$$

The inequality (15.43) can be deduced from (15.80) by using (15.39) repeatedly.

Since $f_i(0,\cdots,0)=0$, employing Lemma 15.2.6 with $b_1=\cdots=b_i=0$ and $\sigma_j=\frac{1}{p_{i-1}\cdots p_j}, j=1,\cdots,i$ yields (15.44). ■

15.C Proof of Proposition 15.4.3

Recall that $u^{p_n\cdots p_1} \equiv \beta(x_1, [\mathrm{sat}_M(\hat{x}_2)]^{p_1}, \cdots, [\mathrm{sat}_M(\hat{x}_n)]^{p_{n-1}\cdots p_1})$ and $u^{*p_n\cdots p_1} \equiv \beta(x_1, x_2^{p_1}, \cdots, x_n^{p_{n-1}\cdots p_1})$ are smooth functions of the arguments $(x_1, [\mathrm{sat}_M(\hat{x}_2)]^{p_1}, \cdots, [\mathrm{sat}_M(\hat{x}_n)]^{p_{n-1}\cdots p_1})$ and $(x_1, x_2^{p_1}, \cdots, x_n^{p_{n-1}\cdots p_1})$, respectively. Moreover, on the set $B_M \times \mathbb{R}^{n-1}$, all these arguments are bounded by a positive constant independent of L_i. Hence, by Lemmas 15.2.3 and 15.2.6, there exists $K_0 > 0$, which is independent of L_i's and satisfies (on the set $B_M \times \mathbb{R}^{n-1}$)

$$\begin{aligned} |u - u^*| &\le 2|u^{p_{n-1}\cdots p_1} - u^{*p_{n-1}\cdots p_1}|^{\frac{1}{p_{n-1}\cdots p_1}} \\ &\le K_0\Big(|x_2^{p_1} - [\mathrm{sat}_M(\hat{x}_2)]^{p_1}| + \cdots \\ &\quad + |x_n^{p_{n-1}\cdots p_1} - [\mathrm{sat}_M(\hat{x}_n)]^{p_{n-1}\cdots p_1}|\Big)^{\frac{1}{p_{n-1}\cdots p_1}}. \end{aligned} \tag{15.81}$$

This, together with Lemmas 15.2.5 and 15.2.2, implies that

$$\begin{aligned} |u - u^*|\Big|_{B_M\times\mathbb{R}^{n-1}} &\le K_0\Big(|x_2^{p_1} - \hat{x}_2^{p_1}| + \cdots \\ &\quad + |x_n^{p_{n-1}\cdots p_1} - \hat{x}_n^{p_{n-1}\cdots p_1}|\Big)^{\frac{1}{p_{n-1}\cdots p_1}} \\ &\le K_0\Big(|x_2^{p_1} - \hat{x}_2^{p_1}|^{\frac{1}{p_{n-1}\cdots p_1}} + \cdots \\ &\quad + |x_n^{p_{n-1}\cdots p_1} - \hat{x}_n^{p_{n-1}\cdots p_1}|^{\frac{1}{p_{n-1}\cdots p_1}}\Big). \end{aligned} \tag{15.82}$$

By Lemmas 15.2.4 and 15.2.2, there exists $K_1 > 0$ independent of L_i's such that for $i = 3, \cdots, n$,

$$|x_i^{p_{i-1}\cdots p_1} - \hat{x}_i^{p_{i-1}\cdots p_1}|^{\frac{1}{p_{n-1}\cdots p_1}} \le \frac{\varepsilon_0}{K_0}|x_i|^{\frac{1}{p_{n-1}\cdots p_i}} + K_1|x_i^{p_{i-1}} - \hat{x}_i^{p_{i-1}}|^{\frac{1}{p_{n-1}\cdots p_{i-1}}}. \tag{15.83}$$

Therefore,

$$\begin{aligned} |u - u^*|\Big|_{B_M\times\mathbb{R}^{n-1}} &\le \varepsilon_0(|x_3|^{\frac{1}{p_{n-1}\cdots p_3}} + \cdots + |x_{n-1}|^{\frac{1}{p_{n-1}}} + |x_n|) \\ &\quad + K_0K_1\Big(|x_2^{p_1} - \hat{x}_2^{p_1}|^{\frac{1}{p_{n-1}\cdots p_1}} + \cdots \\ &\quad + |x_n^{p_{n-1}} - \hat{x}_n^{p_{n-1}}|^{\frac{1}{p_{n-1}}}\Big). \end{aligned} \tag{15.84}$$

In view of the identity that $x_i^{p_{i-1}} - \hat{x}_i^{p_{i-1}} = e_i + L_i(x_{i-1} - \hat{x}_{i-1})$ and (15.41), we deduce from Lemma 15.2.2 that on the set $B_M \times \mathbb{R}^{n-1}$,

$$\begin{aligned} |u - u^*| &\le \varepsilon_0(|x_1|^{\frac{1}{p_{n-1}\cdots p_1}} + \cdots + |x_{n-1}|^{\frac{1}{p_{n-1}}} + |x_n|) \\ &\quad + K(|e_n|^{\frac{1}{p_{n-1}}} + L_n^{\frac{1}{p_{n-1}}}|e_{n-1}|^{\frac{1}{p_{n-1}p_{n-2}}} + \cdots \\ &\quad + L_n^{\frac{1}{p_{n-1}}}\cdots L_3^{\frac{1}{p_{n-1}\cdots p_2}}|e_2|^{\frac{1}{p_{n-1}\cdots p_1}}), \end{aligned} \tag{15.85}$$

where $K > 0$ is a constant independent of L_i's.

On the other hand, both u and u^* are bounded on the set $B_M \times \mathbb{R}^{n-1}$. Using this fact, we deduce (15.45) immediately from (15.85).

By a similar argument, the inequality (15.46) can be proved as well. ■

References

1. Bacciotti, A.: Local Stabilizability of Nonlinear Control Systems. World Scientific, Singapore (1992)
2. Cheng, D., Lin, W.: On p-normal forms of nonlinear systems. IEEE Trans. Automat. Contr. 48, 1242–1248 (2003)
3. Coron, J.M., Praly, L.: Adding an integrator for the stabilization problem. Syst. Contr. Lett. 17, 89–104 (1991)
4. Dayawansa, W.P.: Recent advances in the stabilization problem for low dimensional systems. In: Proc. of the 2^{nd} IFAC NOLCOS, Bordeaux, pp. 1–8 (1992)
5. Dayawansa, W.P., Martin, C.F., Knowles, G.: Asymptotic stabilization of a class of smooth two dimensional systems. SIAM. J. Contr. Optimiz. 28, 1321–1349 (1990)
6. Dayawansa, W.P.: Personal Communication (2002)
7. Hahn, W.: Stability of Motion. Springer, New York (1967)
8. Hermes, H.: Homogeneous coordinates and continuous asymptotically stabilizing feedback controls. In: Elaydi, S. (ed.) Differential Equations Stability and Control, pp. 249–260. Marcel Dekker, New York (1991)
9. Gauthier, J.P., Hammouri, H., Othman, S.: A simple observer for nonlinear systems: Applications to Bioreactors. IEEE Trans. Automat. Contr. 37, 875–880 (1992)
10. Isidori, A.: Nonlinear Control Systems II. Springer, New York (1999)
11. Isidori, A.: A tool for semiglobal stabilization of uncertain non-minimum-phase nonlinear systems via output feedback. IEEE Trans. Automat. Contr. 45, 1817–1827 (2000)
12. Jakubczyk, B., Respondek, W.: Feedback equivalence of planar systems and stability. In: Kaashoek, M.A., van Schuppen, J.H., Ran, A.C.M., et al. (eds.) Robust Control of Linear Systems and Nonlinear Control, pp. 447–456. Birkhäuser, Basel (1990)
13. Kawski, M.: Stabilization of nonlinear systems in the plane. Syst. Contr. Lett. 12, 169–175 (1989)
14. Kawski, M.: Homogeneous stabilizing feedback laws. Control Theory and Advanced Technology 6, 497–516 (1990)
15. Kawski, M.: Geometric homogeneity and applications to stabilization. In: Proc. of the 3rd IFAC NOLCOS, Lake Tahoe, pp. 164–169 (1995)
16. Khalil, H.K., Esfandiari, F.: Semi-global stabilization of a class of nonlinear systems using output feedback. IEEE Trans. Automat. Contr. 38, 1412–1415 (1995)
17. Khalil, H.K.: High-gain observers in nonlinear feedback control. In: Nijmeijer, H., Fossen, T.I. (eds.) New Directions in Nonlinear Observer Design, pp. 249–268. Springer, New York (1999)
18. Krener, A.J., Kang, W.: Backstepping design of nonlinear observers. SIAM J. Contr. Optimiz. 42, 155–177 (2003)
19. Krener, A.J., Xiao, M.: Observers for linearly unobservable nonlinear systems. Syst. Contr. Lett. 46, 281–288 (2002)
20. Lin, W., Pongvuthithum, R.: Global stabilization of cascade systems by C^0 partial state feedback. IEEE Trans. Automat. Contr. 47, 1356–1362 (2002)

21. Mazenc, F., Praly, L., Dayawansa, W.P.: Global stabilization by output feedback: Examples and counterexamples. Syst. Contr. Lett. 23, 119–125 (1994)
22. Qian, C., Lin, W.: A continuous feedback approach to global strong stabilization of nonlinear systems. IEEE Trans. Automat. Contr. 46, 1061–1079 (2001)
23. Qian, C., Lin, W.: Recursive observer design and nonsmooth output feedback stabilization of inherently nonlinear systems. In: Proc. of the 43rd IEEE CDC, Nassau, Bahamas, pp. 4927–4932 (2004)
24. Rosier, L.: Homogeneous Lyapunov functions for homogeneous continuous vector field. Syst. Contr. Lett. 19, 467–473 (1992)
25. Rui, C., Reyhanoglu, M., Kolmanovsky, I., et al.: Non-smooth stabilization of an underactuated unstable two degrees of freedom Mechanical system. In: Proc. of the 36th IEEE CDC, San Diego, pp. 3998–4003 (1997)
26. Teel, A., Praly, L.: Global stabilizability and observability imply semi-global stabilizality by output feedback. Syst. Contr. Lett. 22, 313–325 (1994)
27. Teel, A., Praly, L.: Tools for semiglobal stabilization by partial state and output feedback. SIAM J. Contr. Optimiz. 33, 1443–1488 (1995)
28. Yang, B., Lin, W.: Homogeneous observers, iterative design and global stabilization of high order nonlinear systems by smooth output feedback. IEEE Trans. Automat. Contr. 49, 1069–1080 (2004)
29. Yang, B., Lin, W.: Robust output feedback stabilization of uncertain nonlinear systems with uncontrollable and unobservable linearization. IEEE Trans. Automat. 50, 619–630 (2005)
30. Yang, B., Lin, W.: On semiglobal stabilizability of MIMO nonlinear systems by output feedback. Automatica 42, 1049–1054 (2006)

Author Index

Lecture Notes in Control and Information Sciences

Edited by M. Thoma, F. Allgöwer, M. Morari

Further volumes of this series can be found on our homepage:
springer.com

Vol. 393: Ghosh, B.K.; Martin, C.F.; Zhou, Y.:
Emergent Problems in Nonlinear Systems and Control
285 p. [978-3-642-03626-2]

Vol. 392: Bandyopadhyay, B.; Deepak, F.; Kim, K.-S.:
Sliding Mode Control Using Novel Sliding Surfaces
137 p. [978-3-642-03447-3]

Vol. 391: Khaki-Sedigh, A.; Moaveni, B.:
Control Configuration Selection for Multivariable Plants
232 p. [978-3-642-03192-2]

Vol. 390: Chesi, G.; Garulli, A.; Tesi, A.; Vicino, A.:
Homogeneous Polynomial Forms for Robustness Analysis of Uncertain Systems
197 p. [978-1-84882-780-6]

Vol. 389: Bru, R.; Romero-Vivó, S. (Eds.):
Positive Systems
398 p. [978-3-642-02893-9]

Vol. 388: Jacques Loiseau, J.; Michiels, W.; Niculescu, S-I.; Sipahi, R. (Eds.):
Topics in Time Delay Systems
418 p. [978-3-642-02896-0]

Vol. 387: Xia, Y.; Fu, M.; Shi, P.:
Analysis and Synthesis of Dynamical Systems with Time-Delays
283 p. 2009 [978-3-642-02695-9]

Vol. 386: Huang, D.; Nguang, S.K.:
Robust Control for Uncertain Networked Control Systems with Random Delays
159 p. 2009 [978-1-84882-677-9]

Vol. 385: Jungers, R.:
The Joint Spectral Radius
144 p. 2009 [978-3-540-95979-3]

Vol. 384: Magni, L.; Raimondo, D.M.; Allgöwer, F. (Eds.):
Nonlinear Model Predictive Control
572 p. 2009 [978-3-642-01093-4]

Vol. 383: Sobhani-Tehrani E.; Khorasani K.;
Fault Diagnosis of Nonlinear Systems Using a Hybrid Approach
360 p. 2009 [978-0-387-92906-4]

Vol. 382: Bartoszewicz A.; Nowacka-Leverton A.;
Time-Varying Sliding Modes for Second and Third Order Systems
192 p. 2009 [978-3-540-92216-2]

Vol. 381: Hirsch M.J.; Commander C.W.; Pardalos P.M.; Murphey R. (Eds.)
Optimization and Cooperative Control Strategies: Proceedings of the 8th International Conference on Cooperative Control and Optimization
459 p. 2009 [978-3-540-88062-2]

Vol. 380: Basin M.
New Trends in Optimal Filtering and Control for Polynomial and Time-Delay Systems
206 p. 2008 [978-3-540-70802-5]

Vol. 379: Mellodge P.; Kachroo P.;
Model Abstraction in Dynamical Systems: Application to Mobile Robot Control
116 p. 2008 [978-3-540-70792-9]

Vol. 378: Femat R.; Solis-Perales G.;
Robust Synchronization of Chaotic Systems Via Feedback
199 p. 2008 [978-3-540-69306-2]

Vol. 377: Patan K.
Artificial Neural Networks for the Modelling and Fault Diagnosis of Technical Processes
206 p. 2008 [978-3-540-79871-2]

Vol. 376: Hasegawa Y.
Approximate and Noisy Realization of Discrete-Time Dynamical Systems
245 p. 2008 [978-3-540-79433-2]

Vol. 375: Bartolini G.; Fridman L.; Pisano A.; Usai E. (Eds.)
Modern Sliding Mode Control Theory
465 p. 2008 [978-3-540-79015-0]

Vol. 374: Huang B.; Kadali R.
Dynamic Modeling, Predictive Control
and Performance Monitoring
240 p. 2008 [978-1-84800-232-6]

Vol. 373: Wang Q.-G.; Ye Z.; Cai W.-J.;
Hang C.-C.
PID Control for Multivariable Processes
264 p. 2008 [978-3-540-78481-4]

Vol. 372: Zhou J.; Wen C.
Adaptive Backstepping Control of Uncertain
Systems
241 p. 2008 [978-3-540-77806-6]

Vol. 371: Blondel V.D.; Boyd S.P.;
Kimura H. (Eds.)
Recent Advances in Learning and Control
279 p. 2008 [978-1-84800-154-1]

Vol. 370: Lee S.; Suh I.H.;
Kim M.S. (Eds.)
Recent Progress in Robotics:
Viable Robotic Service to Human
410 p. 2008 [978-3-540-76728-2]

Vol. 369: Hirsch M.J.; Pardalos P.M.;
Murphey R.; Grundel D.
Advances in Cooperative Control and
Optimization
423 p. 2007 [978-3-540-74354-5]

Vol. 368: Chee F.; Fernando T.
Closed-Loop Control of Blood Glucose
157 p. 2007 [978-3-540-74030-8]

Vol. 367: Turner M.C.; Bates D.G. (Eds.)
Mathematical Methods for Robust and Nonlinear
Control
444 p. 2007 [978-1-84800-024-7]

Vol. 366: Bullo F.; Fujimoto K. (Eds.)
Lagrangian and Hamiltonian Methods for
Nonlinear Control 2006
398 p. 2007 [978-3-540-73889-3]

Vol. 365: Bates D.; Hagström M. (Eds.)
Nonlinear Analysis and Synthesis Techniques for
Aircraft Control
360 p. 2007 [978-3-540-73718-6]

Vol. 364: Chiuso A.; Ferrante A.;
Pinzoni S. (Eds.)
Modeling, Estimation and Control
356 p. 2007 [978-3-540-73569-4]

Vol. 363: Besançon G. (Ed.)
Nonlinear Observers and Applications
224 p. 2007 [978-3-540-73502-1]

Vol. 362: Tarn T.-J.; Chen S.-B.;
Zhou C. (Eds.)
Robotic Welding, Intelligence and Automation
562 p. 2007 [978-3-540-73373-7]

Vol. 361: Méndez-Acosta H.O.; Femat R.;
González-Álvarez V. (Eds.):
Selected Topics in Dynamics and Control of
Chemical and Biological Processes
320 p. 2007 [978-3-540-73187-0]

Vol. 360: Kozlowski K. (Ed.)
Robot Motion and Control 2007
452 p. 2007 [978-1-84628-973-6]

Vol. 359: Christophersen F.J.
Optimal Control of Constrained
Piecewise Affine Systems
190 p. 2007 [978-3-540-72700-2]

Vol. 358: Findeisen R.; Allgöwer
F.; Biegler L.T. (Eds.): Assessment and Future
Directions of Nonlinear
Model Predictive Control
642 p. 2007 [978-3-540-72698-2]

Vol. 357: Queinnec I.; Tarbouriech
S.; Garcia G.; Niculescu S.-I. (Eds.):
Biology and Control Theory: Current Challenges
589 p. 2007 [978-3-540-71987-8]

Vol. 356: Karatkevich A.:
Dynamic Analysis of Petri Net-Based Discrete
Systems
166 p. 2007 [978-3-540-71464-4]

Vol. 355: Zhang H.; Xie L.:
Control and Estimation of Systems with
Input/Output Delays
213 p. 2007 [978-3-540-71118-6]

Vol. 354: Witczak M.:
Modelling and Estimation Strategies for Fault
Diagnosis of Non-Linear Systems
215 p. 2007 [978-3-540-71114-8]

Vol. 353: Bonivento C.; Isidori A.; Marconi L.;
Rossi C. (Eds.)
Advances in Control Theory and Applications
305 p. 2007 [978-3-540-70700-4]

Vol. 352: Chiasson, J.; Loiseau, J.J. (Eds.)
Applications of Time Delay Systems
358 p. 2007 [978-3-540-49555-0]

Vol. 351: Lin, C.; Wang, Q.-G.; Lee, T.H., He, Y.
LMI Approach to Analysis and Control of
Takagi-Sugeno Fuzzy Systems with Time Delay
204 p. 2007 [978-3-540-49552-9]

Vol. 350: Bandyopadhyay, B.; Manjunath, T.C.;
Umapathy, M.
Modeling, Control and Implementation of Smart
Structures 250 p. 2007 [978-3-540-48393-9]

Vol. 349: Rogers, E.T.A.; Galkowski, K.;
Owens, D.H.
Control Systems Theory
and Applications for Linear
Repetitive Processes
482 p. 2007 [978-3-540-42663-9]

Vol. 347: Assawinchaichote, W.; Nguang,
K.S.; Shi P.
Fuzzy Control and Filter Design
for Uncertain Fuzzy Systems
188 p. 2006 [978-3-540-37011-6]