8

ISBN 92-95003-11-X

trieste - italy

the
abdus salam
international
centre
for theoretical
physics

united nations
educational, scientific
and cultural
organization

international atomic
energy agency

ictp *lecture notes*

MATHEMATICAL CONTROL THEORY

2002

Number 1

editor
Andrei A. Agrachev

8

ISBN 92-95003-11-X

trieste - italy

the
abdus salam
international
centre
for theoretical
physics

united nations
educational, scientific
and cultural
organization

international atomic
energy agency

ictp *lecture notes*

MATHEMATICAL CONTROL THEORY

2002

Number 2

editor
Andrei A. Agrachev

ICTP *Lecture Notes*

SUMMER SCHOOL ON MATHEMATICAL CONTROL THEORY

3 - 28 September 2001

Editor

Andrei A. Agrachev
Steklov Mathematical Institute, Moscow, Russia
and
SISSA, International School for Advanced Studies,
Trieste, Italy

SUMMER SCHOOL ON MATHEMATICAL CONTROL THEORY
– First edition

Copyright © 2002 by The Abdus Salam International Centre for Theoretical Physics
The Abdus Salam ICTP has the irrevocable and indefinite authorization to reproduce and disseminate these Lecture Notes, in printed and/or computer readable form, from each author.

ISBN 92-95003-11-X

Printed in Trieste by The Abdus Salam ICTP Publications & Printing Section

PREFACE

One of the main missions of the Abdus Salam International Centre for Theoretical Physics in Trieste, Italy, founded in 1964 by Abdus Salam, is to foster the growth of advanced studies and research in developing countries. To this aim, the Centre organizes a large number of schools and workshops in a great variety of physical and mathematical disciplines.

Since unpublished material presented at the meetings might prove of great interest also to scientists who did not take part in the schools the Centre has decided to make it available through a new publication titled ICTP Lecture Note Series. It is hoped that this formally structured pedagogical material in advanced topics will be helpful to young students and researchers, in particular to those working under less favourable conditions.

The Centre is grateful to all lecturers and editors who kindly authorize the ICTP to publish their notes as a contribution to the series.

Since the initiative is new, comments and suggestions are most welcome and greatly appreciated. Information can be obtained from the Publications Section or by e-mail to "`pub_off@ictp.trieste.it`". The series is published in house and also made available on-line via the ICTP web site: "`http://www.ictp.trieste.it`".

M.A. Virasoro
Director

CONTENTS – Number 1

Jerzy Zabczyk
Classical Control Theory 1

Giuseppe Da Prato
Linear Quadratic Control Theory for Infinite Dimensional Systems 59

Bronislaw Jakubczyk
Introduction to Geometric Nonlinear Control; Controllability and Lie Bracket 107

Witold Respondek
Introduction to Geometric Nonlinear Control; Linearization, Observability, Decoupling 169

Matthias Kawski
The Combinatorics of Nonlinear Controllability and Noncommuting Flows 223

A. Bacciotti
Stability Analysis Based on Direct Liapunov Method 313

J.P. Gauthier
A Course on Observability 365

Introduction

This volume is based on the lecture notes of the minicourses given in the frame of the school on Mathematical Control Theory held at the Abdus Salam ICTP from 3 to 28 September 2001.

Mathematical Control Theory is a rapidly growing field which provides strict theoretical and computational tools for dealing with problems arising in electrical and aerospace engineering, automatics, robotics, applied chemistry, and biology etc. Control methods are also involved in questions pertaining to the development of countries in the South, such as wastewater treatment, agronomy, epidemiology, population dynamics, control of industrial and natural bio-reactors. Since most of these natural processes are highly nonlinear, the tools of nonlinear control are essential for the modelling and control of such processes.

At present regular courses in Mathematical Control Theory are rarely included in the curricula of universities, and very few researchers receive enough background in the field. Therefore it is important to organize specific activities in the form of schools to provide the necessary background for those embarking on research in this field.

The school at the Abdus Salam ICTP consisted of several minicourses intended to provide an introduction to various topics of Mathematical Control Theory, including Linear Control Theory (finite and infinite–dimensional), Nonlinear Control, and Optimal Control. The last week of the school was concentrated on applications of Mathematical Control Theory, in particular, those which are important for the development of non-industrialized countries.

The school was intended primarily for mathematicians and mathematically oriented engineers at the beginning of their career. The typical participant was expected to be a graduate student or young post-doctoral researcher interested in Mathematical Control Theory. It was assumed that participants have sufficient background in Ordinary Differential Equations and Advanced Calculus.

The volume contains thirteen contributions divided into two parts. The volume, as well as the school it is based on, pursues primarily educational

and instructive goals. We tried to distribute the material according to the same purposes. The volume starts with Linear Control Systems, then turns to Nonlinear Systems and Optimal Control Theory. Basic elementary courses are intended to help to study subsequent more specific ones. The volume finishes with some real world applications.

We believe that the volume as a whole and its parts can serve for both the self-depended study and the teaching as a kind of contemporary textbook in Mathematical Control Theory.

Andrei Agrachev
May, 2002

CONTENTS – Number 1

Jerzy Zabczyk
Classical Control Theory 1

Giuseppe Da Prato
Linear Quadratic Control Theory for Infinite Dimensional Systems 59

Bronislaw Jakubczyk
Introduction to Geometric Nonlinear Control; Controllability and Lie Bracket 107

Witold Respondek
Introduction to Geometric Nonlinear Control; Linearization, Observability, Decoupling 169

Matthias Kawski
The Combinatorics of Nonlinear Controllability and Noncommuting Flows 223

A. Bacciotti
Stability Analysis Based on Direct Liapunov Method 313

J.P. Gauthier
A Course on Observability 365

CONTENTS – Number 2

Andrei A. Agrachev
Introduction to Optimal Control Theory 451

Hélène Frankowska
Value Function in Optimal Control 515

Jean-Michel Coron
Return Method: Some Applications to Flow Control 655

Ph. Martin, R.M. Murray and P. Rouchon
Flat Systems 705

Olivier Bernard
Mass Balance Modelling of Bioprocesses 769

Olivier Bernard and Jean-Luc Gouzé
State Estimation for Bioprocesses 813

Introduction

This volume is based on the lecture notes of the minicourses given in the frame of the school on Mathematical Control Theory held at the Abdus Salam ICTP from 3 to 28 September 2001.

Mathematical Control Theory is a rapidly growing field which provides strict theoretical and computational tools for dealing with problems arising in electrical and aerospace engineering, automatics, robotics, applied chemistry, and biology etc. Control methods are also involved in questions pertaining to the development of countries in the South, such as wastewater treatment, agronomy, epidemiology, population dynamics, control of industrial and natural bio-reactors. Since most of these natural processes are highly nonlinear, the tools of nonlinear control are essential for the modelling and control of such processes.

At present regular courses in Mathematical Control Theory are rarely included in the curricula of universities, and very few researchers receive enough background in the field. Therefore it is important to organize specific activities in the form of schools to provide the necessary background for those embarking on research in this field.

The school at the Abdus Salam ICTP consisted of several minicourses intended to provide an introduction to various topics of Mathematical Control Theory, including Linear Control Theory (finite and infinite–dimensional), Nonlinear Control, and Optimal Control. The last week of the school was concentrated on applications of Mathematical Control Theory, in particular, those which are important for the development of non-industrialized countries.

The school was intended primarily for mathematicians and mathematically oriented engineers at the beginning of their career. The typical participant was expected to be a graduate student or young post-doctoral researcher interested in Mathematical Control Theory. It was assumed that participants have sufficient background in Ordinary Differential Equations and Advanced Calculus.

The volume contains thirteen contributions divided into two parts. The volume, as well as the school it is based on, pursues primarily educational

and instructive goals. We tried to distribute the material according to the same purposes. The volume starts with Linear Control Systems, then turns to Nonlinear Systems and Optimal Control Theory. Basic elementary courses are intended to help to study subsequent more specific ones. The volume finishes with some real world applications.

We believe that the volume as a whole and its parts can serve for both the self-depended study and the teaching as a kind of contemporary textbook in Mathematical Control Theory.

Andrei Agrachev
May, 2002

Classical Control Theory

Jerzy Zabczyk*

Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028001

*zabczyk@impan.gov.pl

Abstract

The notes introduce basic concepts and results of the classical control theory. The following topics: controllability, observability, minimum energy control, stability and stabilizability as well as linear quadratic control problem and the associated Riccati equations are discussed in details.

Contents

0 Introduction **5**

1 Controllability and Observability **9**
1.1 Preliminaries 9
1.2 The controllability matrix 13
1.3 Rank condition 17
1.4 A classification of control systems 21
1.5 Kalman decomposition 24
1.6 Observability 25

2 Stability and stabilizability **28**
2.1 Stable linear systems 28
2.2 Stable polynomials 33
2.3 Stabilizability and controllability 37

3 Linear quadratic problem **40**
3.1 Introductory comments 40
3.2 Bellman's equation and the value function 41
3.3 The linear regulator problem and the Riccati equation 45
3.4 The linear regulator and stabilization 48

References **55**

0 Introduction

The aim of the lectures is to introduce and motivate basic concepts of the classical control theory. Due to the time limitation several of the important topics like, realisation, control with partial observation, systems on manifolds and infinite dimensional systems will be not covered in the notes. We follow basically our book [31]. For additional information we suggest the reader to consult other sources listed in references, in particular Sontag's book [25]. The reader should try to solve exercises and as a test of good understanding we recommend Exercises 1.7, 1.8, 2.2, 3.7.

A departure point of control theory is the differential equation

$$\dot{y} = f(y, u), \quad y(0) = x \in \mathbb{R}^n, \tag{0.1}$$

with the right-hand side depending on a parameter u from a set $U \subset \mathbb{R}^m$. The set U is called *the set of control parameters*. Differential equations depending on a parameter have been objects of the theory of differential equations for a long time. In particular an important question of continuous dependence of the solutions on parameters has been asked and answered under appropriate conditions. Problems studied in mathematical control theory are, however, of different nature, and a basic role in their formulation is played by the concept of *control*. One distinguishes controls of two types: *open* and *closed loop*. *An open loop control* can be basically an arbitrary function $u(\,\cdot\,) : [0, +\infty) \longrightarrow U$, for which the equation

$$\dot{y}(t) = f(y(t)), u(t)), \quad t \geq 0, \ y(0) = x, \tag{0.2}$$

has a well defined solution.

A closed loop control can be identified with a mapping $k : \mathbb{R}^n \longrightarrow U$, which may depend on $t \geq 0$, such that the equation

$$\dot{y}(t) = f(y(t), k(y(t))), \quad t \geq 0, \ y(0) = x, \tag{0.3}$$

has a well defined solution. The mapping $k(\,\cdot\,)$ is called *feedback*. Controls are called also *strategies* or *inputs*, and the corresponding solutions of (0.2) or (0.3) are *outputs* of the system.

One of the main aims of control theory is to find a strategy such that the corresponding output has desired properties. Depending on the properties involved one gets more specific questions.

Controllability. One says that a state $z \in \mathbb{R}^n$ is *reachable* from x in time T, if there exists an open loop control $u(\cdot)$ such that, for the output $y(\cdot)$, $y(0) = x,\ y(T) = z$. If an arbitrary state z is reachable from an arbitrary state x in a time T, then the system (0.1) is said to be *controllable*. In several situations one requires a weaker property of transferring an arbitrary state into a given one, in particular into the origin. A formulation of effective characterizations of controllable systems is an important task of control theory only partially solved.

Stabilizability. An equally important issue is that of stabilizability. Assume that for some $\bar{x} \in \mathbb{R}^n$ and $\bar{u} \in U$, $f(\bar{x}, \bar{u}) = 0$. A function $k : \mathbb{R}^n \longrightarrow U$, such that $k(\bar{x}) = \bar{u}$, is called a *stabilizing feedback* if $\bar{x}$ is a stable equilibrium for the system

$$\dot{y}(t) = f(y(t), k(y(t))), \quad t \geq 0,\ y(0) = x. \tag{0.4}$$

In the theory of differential equations there exist several methods to determine whether a given equilibrium state is a stable one. The question of whether, in the class of all equations of the form (0.4), there exists one for which $\bar{x}$ is a stable equilibrium is of a new qualitative type.

Observability. In many situations of practical interest one observes not the state $y(t)$ but its function $h(y(t)),\ t \geq 0$. It is therefore often necessary to investigate the pair of equations

$$\dot{y} = f(y, u), \quad y(0) = x, \tag{0.5}$$

$$w = h(y). \tag{0.6}$$

Relation (0.6) is called an *observation equation*. The system (0.5)–(0.6) is said to be *observable* if, knowing a control $u(\cdot)$ and an observation $w(\cdot)$, on a given interval $[0, T]$, one can determine uniquely the initial condition x.

Optimality. Besides the above problems of structural character, in control theory, with at least the same intensity, one asks optimality questions. In the so-called time-optimal problem one is looking for a control which not only transfers a state x onto z but does it in the minimal time T. In other situations the time $T > 0$ is fixed and one is looking for a control $u(\cdot)$ which minimizes the integral

$$\int_0^T g(y(t), u(t))\, dt + G(y(T)),$$

in which g and G are given functions.

We present now some examples to show that the models and problems discussed in control theory have an immediate real meaning.

Example 0.1 *Electrically heated oven.* Let us consider a simple model of an electrically heated oven, which consists of a jacket with a coil directly heating the jacket and of an interior part. Let T_0 denote the outside temperature. We make a simplifying assumption, that at an arbitrary moment $t \geq 0$, temperatures in the jacket and in the interior part are uniformly distributed and equal to $T_1(t), T_2(t)$. We assume also that the flow of heat through a surface is proportional to the area of the surface and to the difference of temperature between the separated media. Let $u(t)$ be the intensity of the heat input produced by the coil at moment $t \geq 0$. Let moreover a_1, a_2 denote the area of exterior and interior surfaces of the jacket, c_1, c_2 denote heat capacities of the jacket and the interior of the oven and r_1, r_2 denote radiation coefficients of the exterior and interior surfaces of the jacket. An increase of heat in the jacket is equal to the amount of heat produced by the coil reduced by the amount of heat which entered the interior and exterior of the oven. Therefore, for the interval $[t, t + \Delta t]$, we have the following balance:

$$c_1(T_1(t+\Delta t) - T_1(t)) \approx u(t)\Delta t - (T_1(t) - T_2(t))a_1 r_1 \Delta t - (T_1(t) - T_0)a_2 r_2 \Delta t.$$

Similarly, an increase of heat in the interior of the oven is equal to the amount of heat radiated by the jacket:

$$c_2(T_2(t + \Delta t) - T_2(t)) = (T_1(t) - T_2(t))a_1 r_2 \Delta t.$$

Dividing the obtained identities by Δt and taking the limit, as $\Delta t \downarrow 0$, we obtain

$$\begin{aligned} c_1 \frac{dT_1}{dt} &= u - (T_1 - T_2)a_1 r_1 - (T_1 - T_0)a_2 r_2, \\ c_2 \frac{dT_2}{dt} &= (T_1 - T_2)a_1 r_1. \end{aligned}$$

Let us remark that, according to the physical interpretation, $u(t) \geq 0$ for $t \geq 0$. Introducing new variables $x_1 = T_1 - T_0$ and $x_2 = T_2 - T_0$, we have

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -\dfrac{r_1 a_1 + r_2 a_2}{c_1} & \dfrac{r_1 a_1}{c_1} \\ \dfrac{r_1 a_1}{c_2} & -\dfrac{r_1 a_1}{c_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} c_1^{-1} \\ 0 \end{bmatrix} u.$$

It is natural to limit the considerations to the case when $x_1(0) \geq 0$ and $x_2(0) \geq 0$. It is physically obvious that if $u(t) \geq 0$ for $t \geq 0$, then also $x_1(t) \geq 0,\ x_2(t) \geq 0,\ t \geq 0$. One can prove this mathematically.

Let us assume that we want to obtain, in the interior part of the oven, a temperature T and keep it at this level infinitely long. Is this possible? Does the answer depend on initial temperatures $T_1 \geq T_0$, $T_2 \geq T_0$?

Example 0.2 *Soft landing.* Let us consider a spacecraft of total mass M moving vertically with the gas thruster directed toward the landing surface. Let h be the height of the spacecraft above the surface, u the thrust of its engine produced by the expulsion of gas from the jet. The gas is a product of the combustion of the fuel. The combustion decreases the total mass of the spacecraft, and the thrust u is proportional to the speed with which the mass decreases. Assuming that there is no atmosphere above the surface and that g is gravitational acceleration, one arrives at the following equations [13]:

$$M\ddot{h} = -gM + u, \tag{0.7}$$

$$\dot{M} = -ku, \tag{0.8}$$

with the initial conditions $M(0) = M_0$, $h(0) = h_0$, $\dot{h}(0) = h_1$; k a positive constant. One imposes additional constraints on the control parameter of the type $0 \leq u \leq \alpha$ and $M \geq m$, where m is the mass of the spacecraft without fuel. Let us fix $T > 0$. The soft landing problem consists of finding a control $u(\,\cdot\,)$ such that for the solutions $M(\,\cdot\,)$, $h(\,\cdot\,)$ of equation (0.7)

$$M(t) \geq m, \quad h(t) \geq 0, \quad t \in [0,T], \quad \text{and} \quad h(T) = \dot{h}(T) = 0.$$

The problem of the existence of such a control is equivalent to the controllability of the system (0.7)–(0.9).

A natural optimization question arises when the moment T is not fixed and one is minimizing the landing time. The latter problem can be formulated equivalently as the *minimum fuel problem.* In fact, let $v = \dot{h}$ denote the velocity of the spacecraft, and let $M(t) > 0$ for $t \in [0,T]$. Then

$$\frac{\dot{M}(t)}{M(t)} = -k\dot{v}(t) - gk, \quad t \in [0,T].$$

Therefore, after integration,

$$M(T) = e^{-v(T)k - gkT + v(0)k} M(0).$$

Thus a soft landing is taking place at a moment $T > 0$ ($v(T) = 0$) if and only if

$$M(T) = e^{-gkT} e^{v(0)k} M(0).$$

Consequently, the minimization of the landing time T is equivalent to the minimization of the amount of fuel $M(0) - M(T)$ needed for landing.

Example 0.3 *Optimal consumption.* The capital $y(t) \geq 0$ of an economy at any moment t is divided into two parts: $u(t)y(t)$ and $(1-u(t))y(t)$, where $u(t)$ is a number from the interval $[0, 1]$. The first part goes for investments and contributes to the increase in capital according to the formula

$$\dot{y} = uy, \quad y(0) = x > 0.$$

The remaining part is for consumption evaluated by the *satisfaction*

$$J_T(x, u(\cdot)) = \int_0^T ((1 - u(t))y(t))^\alpha \, dt + ay^\alpha(T). \tag{0.9}$$

In definition (0.9), the number a is nonnegative and $\alpha \in (0, 1)$. In the described situation one is trying to divide the capital to maximize the satisfaction.

Remark For more information about the electrically heated oven we refer to [2], [24]. The soft landing and optimal consumption models are extensively discussed in [14].

1 Controllability and Observability

1.1 Preliminaries

As we have already mentioned the basic object of classical control theory is a linear system described by a differential equation

$$\frac{dy}{dt} = Ay(t) + Bu(t), \quad y(0) = x \in \mathbb{R}^n, \tag{1.1}$$

and an observation relation

$$w(t) = Cy(t), \quad t \geq 0. \tag{1.2}$$

For completeness of the presentation we recall first basic concepts and notation related to linear differential equations.

Linear transformations $A : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $B : \mathbb{R}^m \longrightarrow \mathbb{R}^n$, $C : \mathbb{R}^m \longrightarrow \mathbb{R}^k$ in (1.1) and (1.2) will be identified with representing matrices and elements of $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^k$ with one column matrices. The set of all matrices with n rows and m columns will be denoted by $\mathbf{M}(n,m)$ and the identity transformation as well as the identity matrix by I. The scalar product $\langle x, y\rangle$ and the norm $|x|$, of elements $x, y \in \mathbb{R}^n$ with coordinates $\xi_1, \dots, \xi_n$ and $\eta_1, \dots, \eta_n$, are defined by

$$\langle x, y\rangle = \sum_{j=1}^{n} \xi_j \eta_j, \quad |x| = \left(\sum_{j=1}^{n} \xi_j^2\right)^{1/2}.$$

The adjoint transformation of a linear transformation A as well as the transpose matrix of A are denoted by A^*. A matrix $A \in \mathbf{M}(n,n)$ is called *symmetric* if $A = A^*$. The set of all symmetric matrices is partially ordered by the relation $A_1 \geq A_2$ if $\langle A_1 x, x\rangle \geq \langle A_2 x, x\rangle$ for arbitrary $x \in \mathbb{R}^n$. If $A \geq 0$ then one says that matrix A is *nonnegative definite* and if, in addition, $\langle Ax, x\rangle > 0$ for $x \neq 0$ that A is *positive definite*. Treating $x \in \mathbb{R}^n$ as an element of $\mathbf{M}(n,1)$ we have $x^* \in \mathbf{M}(1,n)$. In particular we can write $\langle x, y\rangle = x^* y$ and $|x|^2 = x^* x$. The inverse transformation of A and the inverse matrix of A will be denoted by A^{-1}.

If $F(t) = [f_{ij}(t);\ i = 1, \dots, n,\ j = 1, \dots, m] \in \mathbf{M}(n,m)$, $t \in [0,T]$, then, by definition,

$$\int_0^T F(t)\, dt = \left[\int_0^T f_{ij}(t)\, dt;\ i = 1, \dots, n,\ j = 1, \dots, m\right], \tag{1.3}$$

under the condition that elements of $F(\,\cdot\,)$ are integrable.

Derivatives of the 1st and 2nd order of a function $y(t)$, $t \in \mathbb{R}$, are denoted by $\frac{dy}{dt}$, $\frac{d^2y}{dt^2}$ or by $\dot{y}$, $\ddot{y}$ and the nth order derivative, by $\frac{d^{(n)}y}{dt^{(n)}}$.

We will need some basic results on linear equations

$$\frac{dq}{dt} = A(t)q(t) + a(t), \quad q(t_0) = q_0 \in \mathbb{R}^n, \tag{1.4}$$

on a fixed interval $[0,T]$; $t_0 \in [0,T]$, where $A(t) \in \mathbf{M}(n,n)$, $A(t) = [a_{ij}(t);\ i = 1, \dots, n,\ j = 1, \dots, m]$, $a(t) \in \mathbb{R}^n$, $a(t) = (a_i(t);\ i = 1, \dots, n)$, $t \in [0,T]$.

Theorem 1.1 *Assume that elements of the function $A(\,\cdot\,)$ are locally integrable. Then there exists exactly one function $S(t)$, $t \in [0,T]$ with values in $\mathbf{M}(n,n)$ and with absolutely continuous elements such that*

$$\frac{d}{dt}S(t) = A(t)S(t) \quad \textit{for almost all } t \in [0,T], \tag{1.5}$$

$$S(0) = I. \tag{1.6}$$

In addition, a matrix $S(t)$ is invertible for an arbitrary $t \in [0,T]$, and the unique solution of the equation (1.4) *is of the form*

$$q(t) = S(t)S^{-1}(t_0)q_0 + \int_{t_0}^{t} S(t)S^{-1}(s)a(s)\,ds, \quad t \in [0,T]. \tag{1.7}$$

Here is a sketch a proof of the theorem.

Proof. Equation (1.4) is equivalent to the integral equation

$$q(t) = a_0 + \int_{t_0}^{t} A(s)q(s)\,ds + \int_{t_0}^{t} a(s)\,ds, \quad t \in [0,T].$$

The formula

$$\mathcal{L}y(t) = a_0 + \int_{t_0}^{t} a(s)\,ds + \int_{t_0}^{t} A(s)y(s)\,ds, \quad t \in [0,T],$$

defines a continuous transformation from the space of continuous functions $C[0,T;\,\mathbb{R}^n]$ into itself, such that for arbitrary $y(\,\cdot\,), \tilde{y}(\,\cdot\,) \in C[0,T;\,\mathbb{R}^n]$

$$\sup_{t\in[0,T]} |\mathcal{L}y(t) - \mathcal{L}\tilde{y}(t)| \le \left(\int_0^T |A(s)|\,ds\right) \sup_{t\in[0,T]} |y(t) - \tilde{y}(t)|.$$

If $\int_0^T |A(s)|\,ds < 1$, then by the contraction mapping principle the equation $q = \mathcal{L}q$ has exactly one solution in $C[0,T;\,\mathbb{R}^n]$ which is the solution of the integral equation. The case $\int_0^T |A(s)|\,ds \ge 1$ can be reduced to the previous one by considering the equation on appropriately shorter intervals. In particular we obtain the existence and uniqueness of a matrix valued function satisfying (1.5) and (1.6).

To prove the second part of the theorem let us denote by $\psi(t)$, $t \in [0,T]$, the matrix solution of

$$\frac{d}{dt}\psi(t) = -\psi(t)A(t), \quad \psi(0) = I,\ t \in [0,T].$$

Assume that, for some $t \in [0,T]$, $\det S(t) = 0$. Let $T_0 = \min\{t \in [0,T];\, \det S(t) = 0\}$. Then $T_0 > 0$, and for $t \in [0,T_0)$

$$0 = \frac{d}{dt}\left(S(t)S^{-1}(t)\right) = \left(\frac{d}{dt}S(t)\right)S^{-1}(t) + S(t)\frac{d}{dt}S^{-1}(t).$$

Thus

$$-A(t) = S(t)\frac{d}{dt}S^{-1}(t),$$

and consequently

$$\frac{d}{dt}S^{-1}(t) = -S^{-1}(t)A(t), \quad t \in [0, T_0),$$

so $S^{-1}(t) = \psi(t),\ t \in [0, T_0)$.

The function $\det \psi(t),\ t \in [0, T]$, is continuous and

$$\det \psi(t) = \frac{1}{\det S(t)}, \quad t \in [0, T_0),$$

therefore there exists a finite $\lim_{t \uparrow T_0} \det \psi(t)$. This way $\det S(T_0) = \lim_{t \uparrow T_0} S(t) \neq 0$, a contradiction. The validity of (1.6) follows now by elementary calculation. □

The function $S(t),\ t \in [0, T]$ will be called *the fundamental solution* of equation (1.4). It follows from the proof that the fundamental solution of the "adjoint" equation

$$\frac{dp}{dt} = -A^*(t)p(t), \quad t \in [0, T],$$

is $(S^*(t))^{-1},\ t \in [0, T]$.

Exercise 1.1 Show that for $A \in \mathbf{M}(n, n)$ the series

$$\sum_{n=1}^{+\infty} \frac{A^n}{n!}t^n, \quad t \in \mathbb{R},$$

is uniformly convergent, with all derivatives, on an arbitrary finite interval.

The sum of the series from Exercise 1.1 is often denoted by $\exp(tA)$ or $e^{tA},\ t \in \mathbb{R}$. We check easily that

$$e^{tA}e^{sA} = e^{(t+s)A}, \quad t, s \in \mathbb{R},$$

in particular

$$(e^{tA})^{-1} = e^{-tA}, \quad t \in \mathbb{R}.$$

Therefore the solution of (1.1) has the form

$$\begin{aligned} y(t) &= e^{tA}x + \textstyle\int_0^t e^{(t-s)A}Bu(s)\,ds) \\ &= S(t)x + \textstyle\int_0^t S(t-s)Bu(s)\,ds, \quad t \in [0,T], \end{aligned} \tag{1.8}$$

where $S(t) = \exp tA,\ t \geq 0$.

The majority of the concepts and results discussed for systems (1.1)–(1.2) can be extended to time dependent matrices $A(t) \in \mathbf{M}(n,n)$, $B(t) \in \mathbf{M}(n,n)$, $C(t) \in \mathbf{M}(k,n)$, $t \in [0,T]$, and therefore for systems

$$\frac{dy}{dt} = A(t)y(t) + B(t)u(t), \quad y(0) = x \in \mathbb{R}^n, \tag{1.9}$$

$$w(t) = C(t)y(t), \quad t \in [0,T]. \tag{1.10}$$

1.2 The controllability matrix

An arbitrary function $u(\,\cdot\,)$ defined on $[0,+\infty)$ locally integrable and with values in $\mathbb{R}^m$ will be called a *control, strategy* or *input* of the system (1.1)–(1.2). The corresponding solution of equation (1.1) will be denoted by $y^{x,u}(\,\cdot\,)$, to underline the dependence on the initial condition x and the input $u(\,\cdot\,)$. Relationship (1.2) can be written in the following way:

$$w(t) = Cy^{x,u}(t), \quad t \in [0,T].$$

The function $w(\,\cdot\,)$ is the *output* of the controlled system.

We will assume now that $C = I$ or equivalently that $w(t) = y^{x,u}(t)$, $t \geq 0$.

We say that a control u *transfers* a state a to a state b at the time $T > 0$ if

$$y^{a,u}(T) = b. \tag{1.11}$$

We then also say that the state a can be *steered* to b at time T or that the state b is *reachable* or *attainable* from a at time T.

The proposition below gives a formula for a control transferring a to b. In this formula the matrix Q_T, called the *controllability matrix* or *controllability Gramian*, appears:

$$Q_T = \int_0^T S(r)BB^*S^*(r)\,dr, \quad T > 0.$$

We check easily that Q_T is symmetric and nonnegative definite.

Proposition 1.1 *Assume that for some $T > 0$ the matrix Q_T is nonsingular. Then*

(i) *for arbitrary $a, b \in \mathbb{R}^n$ the control*

$$\hat{u}(s) = -B^* S^*(T-s) Q_T^{-1}(S(T)a - b), \quad s \in [0,T], \tag{1.12}$$

transfers a to b at time T;

(ii) *among all controls $u(\cdot)$ steering a to b at time T the control $\hat{u}$ minimizes the integral $\int_0^T |u(s)|^2\, ds$. Moreover,*

$$\int_0^T |\hat{u}(s)|^2\, ds = \langle Q_T^{-1}(S(T)a - b),\ S(T)a - b\rangle. \tag{1.13}$$

Proof. It follows from (1.12) that the control $\hat{u}$ is smooth or even analytic. From (1.8) and (1.12) we obtain that

$$\begin{aligned} y^{a,\hat{u}}(T) &= S(T)a - \left(\int_0^T S(T-s) B B^* S^*(T-s)\, ds\right)(Q_T^{-1}(S(T)a - b)) \\ &= S(T)a - Q_T(Q_T^{-1}(S(T)a - b)) = b. \end{aligned}$$

This shows (i).

To prove (ii) let us remark that the formula (1.13) is a consequence of the following simple calculations:

$$\begin{aligned} \int_0^T |\hat{u}(s)|^2\, ds &= \int_0^T |B^* S^*(T-s) Q_T^{-1}(S(T)a - b)|^2\, ds \\ &= \left\langle \int_0^T S(T-s) B B^* S^*(T-s)(Q_T^{-1}(S(T)a - b))\, ds, \right. \\ &\qquad\qquad \left. Q_T^{-1}(S(T)a - b) \right\rangle \\ &= \langle Q_T Q_T^{-1}(S(T)a - b), Q_T^{-1}(S(T)a - b)\rangle \\ &= \langle Q_T^{-1}(S(T)a - b), S(T)a - b\rangle. \end{aligned}$$

Now let $u(\cdot)$ be an arbitrary control transferring a to b at time T. We can assume that $u(\cdot)$ is square integrable on $[0,T]$. Then

$$\begin{aligned} \int_0^T \langle u(s), \hat{u}(s)\rangle\, ds &= -\int_0^T \langle u(s), B^* S^*(T-s) Q_T^{-1}(S(T)b - a)\rangle\, ds \\ &= -\left\langle \int_0^T S(T-s) B u(s)\, ds, Q_T^{-1}(S(T)a - b) \right\rangle \\ &= \langle S(T)a - b, Q_T^{-1}(S(T)a - b)\rangle. \end{aligned}$$

Hence

$$\int_0^T \langle u(s), \hat{u}(s)\rangle \, ds = \int_0^T \langle \hat{u}(s), \hat{u}(s)\rangle \, ds.$$

From this we obtain that

$$\int_0^T |u(s)|^2 \, ds = \int_0^T |\hat{u}(s)|^2 \, ds + \int_0^T |u(s) - \hat{u}(s)|^2 \, ds$$

and consequently the desired minimality property. □

Exercise 1.2 Write equation

$$\frac{d^2y}{dt^2} = u, \quad y(0) = \xi_1, \quad \frac{dy}{dt}(0) = \xi_2, \quad \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \in \mathbb{R}^2,$$

as a first order system. Prove that for the new system, the matrix Q_T is nonsingular, $T > 0$. Find the control u transferring the state $\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ at time $T > 0$ and minimizing the functional $\int_0^T |u(s)|^2 \, ds$. Determine the minimal value m of the functional. Consider $\xi_1 = 1$, $\xi_2 = 0$.

Answer. The required control is of the form

$$\hat{u}(s) = -\frac{12}{T^3}\left(\frac{\xi_1 T}{2} + \frac{\xi_2 T^2}{3} - \frac{sT\xi_2}{2} - s\xi_1\right), \quad s \in [0, T],$$

and the minimal value m of the functional is equal to

$$m = \frac{12}{T^3}\left((\xi_1)^2 + \xi_1\xi_2 T - \frac{2T^2}{3}(\xi_2)^2\right).$$

In particular, when $\xi_1 = 1$, $\xi_2 = 0$,

$$\hat{u}(s) = \frac{12}{T^3}\left(s - \frac{T}{2}\right), \quad s \in [0, T], \ m = \frac{12}{T^3}.$$

We say that a state b is *attainable* or *reachable* from $a \in \mathbb{R}^n$ if it is attainable or reachable at some time $T > 0$.

System (1.1) is called *controllable* if an arbitrary state $b \in \mathbb{R}^n$ is attainable from any state $a \in \mathbb{R}^n$ at some time $T > 0$. Instead of saying that system (1.1) is controllable we will frequently say that the *pair* (A, B) is *controllable*.

If for arbitrary $a, b \in \mathbb{R}^n$ the attainability takes place at a given time $T > 0$, we say that the system is *controllable at time* T. Proposition 1.1 gives a sufficient condition for the system (1.1) to be controllable. It turns out that this condition is also a necessary one.

The following result holds.

Proposition 1.2 *If an arbitrary state $b \in \mathbb{R}^n$ is attainable from 0, then the matrix Q_T is nonsingular for an arbitrary $T > 0$.*

Proof. Let, for a control u and $T > 0$,

$$\mathcal{L}_T u = \int_0^T S(r)Bu(T-r)\,dr. \tag{1.14}$$

The formula (1.14) defines a linear operator from $U_T = L^1[0,T;\,\mathbb{R}^m]$ into $\mathbb{R}^n$. Let us remark that

$$\mathcal{L}_T u = y^{0,u}(T). \tag{1.15}$$

Let $E_T = \mathcal{L}_T(U_T)$, $T > 0$. It follows from (1.14) that the family of the linear spaces E_T is nondecreasing in $T > 0$. Since $\bigcup_{T>0} E_T = \mathbb{R}^n$, taking into account the dimensions of E_T, we have that $E_{\widetilde{T}} = \mathbb{R}^n$ for some $\widetilde{T}$. Let us remark that, for arbitrary $T > 0$, $v \in \mathbb{R}^n$ and $u \in U_T$,

$$\langle Q_T v, v\rangle = \left\langle \left(\int_0^T S(r)BB^*S^*(r)\,dr\right)v, v\right\rangle = \int_0^T |B^*S^*(r)v|^2\,dr, \tag{1.16}$$

$$\langle \mathcal{L}_T u, v\rangle = \int_0^T \langle u(r), B^*S^*(T-r)v\rangle\,dr. \tag{1.17}$$

From identities (1.16) and (1.17) we obtain $Q_T v = 0$ for some $v \in \mathbb{R}^n$ if the space E_T is orthogonal to v or if the function $B^*S^*(\,\cdot\,)v$ is identically equal to zero on $[0,T]$. It follows from the analyticity of this function that it is equal to zero everywhere. Therefore if $Q_T v = 0$ for some $T > 0$ then $Q_T v = 0$ for all $T > 0$ and in particular $Q_{\widetilde{T}} v = 0$. Since $E_{\widetilde{T}} = \mathbb{R}^n$ we have that $v = 0$, and the nonsingularity of Q_T follows. □

A sufficient condition for controllability is that the rank of B is equal to n. This follows from the next exercise.

Exercise 1.3 Assume $\operatorname{rank} B = n$ and let B^+ be a matrix such that $BB^+ = I$. Check that the control

$$u(s) = \frac{1}{T}B^+e^{(s-T)A}(b - e^{TA}a), \quad s \in [0,T],$$

transfers a to b at time $T \geq 0$.

1.3 Rank condition

We now formulate an algebraic condition equivalent to controllability. For matrices $A \in \mathbf{M}(n,n)$, $B \in \mathbf{M}(n,m)$ denote by $[A|B]$ the matrix $[B, AB, \ldots, A^{n-1}B] \in \mathbf{M}(n, nm)$ which consists of consecutively written columns of matrices $B, AB, \ldots, A^{n-1}B$.

Theorem 1.2 *The following conditions are equivalent.*

(i) *An arbitrary state $b \in \mathbb{R}^n$ is attainable from 0.*
(ii) *System (1.1) is controllable.*
(iii) *System (1.1) is controllable at a given time $T > 0$.*
(iv) *Matrix Q_T is nonsingular for some $T > 0$.*
(v) *Matrix Q_T is nonsingular for an arbitrary $T > 0$.*
(vi) $\operatorname{rank}[A|B] = n$.

Condition (vi) is called the *Kalman rank condition*, or the rank condition for short.

The proof will use the Cayley-Hamilton theorem. Let us recall that a *characteristic polynomial* $p(\,\cdot\,)$ of a matrix $A \in \mathbf{M}(n,n)$ is defined by

$$p(\lambda) = \det(\lambda I - A), \quad \lambda \in \mathbb{C}. \tag{1.18}$$

Let

$$p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n, \quad \lambda \in \mathbb{C}. \tag{1.19}$$

The Cayley-Hamilton theorem has the following formulation (see [1, 358–359]):

Theorem 1.3 *For arbitrary $A \in \mathbf{M}(n,n)$, with the characteristic polynomial (1.19),*

$$A^n + a_1A^{n-1} + \ldots + a_nI = 0.$$

Symbolically, $p(A) = 0$.

Proof of Theorem 1.2. Equivalences (i)–(v) follow from the proofs of Propositions 1.1 and 1.2 and the identity

$$y^{a,u}(T) = \mathcal{L}_T u + S(T)a.$$

To show the equivalences to condition (vi) it is convenient to introduce a linear mapping l_n from the Cartesian product of n copies $\mathbb{R}^m$ into $\mathbb{R}^n$:

$$l_n(u_0, \ldots, u_{n-1}) = \sum_{j=0}^{n-1} A^jBu_j, \quad u_j \in \mathbb{R}^m,\ j = 0, \ldots, n-1.$$

We prove first the following lemma.

Lemma 1.1 *The transformation $\mathcal{L}_T$, $T > 0$, has the same image as l_n. In particular $\mathcal{L}_T$ is onto if and only if l_n is onto.*

Proof. For arbitrary $v \in \mathbb{R}^n$, $u \in L^1[0,T;\mathbb{R}^m]$, $u_j \in \mathbb{R}^m$, $j = 0,\dots,n-1$:

$$\langle \mathcal{L}_T u, v\rangle = \int_0^T \langle u(s), B^*S^*(T-s)v\rangle\, ds,$$
$$\langle l_n(u_0,\dots,u_{n-1}), v\rangle = \langle u_0, B^*v\rangle + \dots + \langle u_{n-1}, B^*(A^*)^{n-1}v\rangle.$$

Suppose that $\langle l_n(u_0,\dots,u_{n-1}), v\rangle = 0$ for arbitrary $u_0,\dots,u_{n-1} \in \mathbb{R}^m$. Then $B^*v = 0,\dots,B^*(A^*)^{n-1}v = 0$. From Theorem 1.3, applied to matrix A^*, it follows that for some constants $c_0,\dots,c_{n-1}$

$$(A^*)^n = \sum_{k=0}^{n-1} c_k (A^*)^k.$$

Thus, by induction, for arbitrary $l = 0,1,\dots$ there exist constants $c_{l,0},\dots,c_{l,n-1}$ such that

$$(A^*)^{n+1} = \sum_{k=0}^{n-1} c_{l,k}(A^*)^k.$$

Therefore $B^*(A^*)^k v = 0$ for $k = 0,1,\dots$. Taking into account that

$$B^*S^*(t)v = \sum_{k=0}^{+\infty} B^*(A^*)^k v \frac{t^k}{k!}, \quad t \geq 0,$$

we deduce that for arbitrary $T > 0$ and $t \in [0,T]$

$$B^*S^*(t)v = 0,$$

so $\langle \mathcal{L}_T u, v\rangle = 0$ for arbitrary $u \in L^1[0,T;\mathbb{R}^m]$.

Assume, conversely, that for arbitrary $u \in L^1[0,T;\mathbb{R}^n]$, $\langle \mathcal{L}_T u, v\rangle = 0$. Then $B^*S^*(t)v = 0$ for $t \in [0,T]$. Differentiating the identity

$$\sum_{k=0}^{+\infty} B^*(A^*)^k v \frac{t^k}{k!} = 0, \quad t \in [0,T],$$

$0,1,\dots,(n-1)$ times and inserting each time $t = 0$, we obtain that $B^*(A^*)^k v = 0$ for $k = 0,1,\dots,n-1$. And therefore

$$\langle l_n(u_0,\dots,u_{n-1}), v\rangle = 0 \quad \text{for arbitrary } u_0,\dots,u_{n-1} \in \mathbb{R}^m.$$

This implies the lemma. □

Assume that the system (1.1) is controllable. Then the transformation $\mathcal{L}_T$ is onto $\mathbb{R}^n$ for arbitrary $T > 0$ and, by the above lemma, the matrix $[A|B]$ has rank n. Conversely, if the rank of $[A|B]$ is n then the mapping l_n is onto $\mathbb{R}^n$ and also, therefore, the transformation $\mathcal{L}_T$ is onto $\mathbb{R}^n$ and the controllability of (1.1) follows. □

If the rank condition is satisfied then the control $\hat{u}(\cdot)$ given by (1.12) transfers a to b at time T. We now give a different, more explicit, formula for the transfer control involving the matrix $[A|B]$ instead of the controllability matrix Q_T.

Note that if $\operatorname{rank}[A|B] = n$ then there exists a matrix $K \in \mathbf{M}(mn, n)$ such that $[A|B]K = I \in \mathbf{M}(n, n)$ or equivalently there exist matrices K_1, $K_2, \ldots, K_n \in \mathbf{M}(m, n)$ such that

$$BK_1 + ABK_2 + \ldots + A^{n-1}BK_n = I. \tag{1.20}$$

Let, in addition, φ be a function of class C^{n-1} from $[0, T]$ into R such that

$$\frac{d^j\varphi}{ds^j}(0) = \frac{d^j\varphi}{ds^j}(T) = 0, \quad j = 0, 1, \ldots, n-1, \tag{1.21}$$

$$\int_0^T \varphi(s)ds = 1. \tag{1.22}$$

Proposition 1.3 *Assume that* $\operatorname{rank}[A|B] = n$ *and* (1.20)–(1.22) *hold. Then the control*

$$\tilde{u}(s) = K_1\psi(s) + K_2\frac{d\psi}{ds}(s) + \ldots + K_n\frac{d^{n-1}\psi}{ds^{n-1}}(s), \quad s \in [0, T]$$

where

$$\psi(s) = S(s - T)(b - S(T)a)\varphi(s), \quad s \in [0, T] \tag{1.23}$$

transfers a to b at time $T \geq 0$.

Proof. Taking into account (1.21) and integrating by parts $(j-1)$ times, we have

$$\int_0^T S(T-s)BK_j\frac{d^{j-1}}{ds^{j-1}}\psi(s)\,ds = \int_0^T e^{A(T-s)}BK_j\frac{d^{j-1}}{ds^{j-1}}\psi(s)\,ds$$
$$= \int_0^T e^{A(T-s)}A^{j-1}BK_j\psi(s)\,ds$$
$$= \int_0^T S(T-s)A^{j-1}BK_j\psi(s)\,ds,$$
$$j = 1, 2, \ldots, n.$$

Consequently

$$\int_0^T S(T-s)B\tilde{u}(s)ds = \int_0^T S(t-s)[A|B]K\psi(s)ds$$
$$= \int_0^T S(T-s)\psi(s)ds.$$

By the definition of ψ and by (1.22) we finally have

$$y^{a,\tilde{u}}(T) = S(T)a + \int_0^T S(T-s)(S(s-T)(b-S(T)a))\varphi(s)ds$$
$$= S(T)a + (b-S(T)a)\int_0^T \varphi(s)ds = b.$$

□

Remark Note that Proposition 1.3 is a generalization of Exercise 1.3.

Exercise 1.4 Assuming that $U = \mathbb{R}$ prove that the system describing the electrically heated oven from Example 0.1 is controllable.

Exercise 1.5 Let L_0 be a linear subspace dense in $L^1[0,T;\mathbb{R}^m]$. If system (1.1) is controllable then for arbitrary $a, b \in \mathbb{R}^n$ there exists $u(\cdot) \in L_0$ transferring a to b at time T.

Hint. Use the fact that the image of the closure of a set under a linear continuous mapping is contained in the closure of the image of the set.

Exercise 1.6 If system (1.1) is controllable then for arbitrary $T > 0$ and arbitrary $a, b \in \mathbb{R}^n$ there exists a control $u(\cdot)$ of class C^∞ transferring a to b at time T and such that

$$\frac{d^{(j)}u}{dt^{(j)}}(0) = \frac{d^{(j)}u}{dt^{(j)}}(T) = 0 \quad \text{for } j = 0, 1, \ldots.$$

Exercise 1.7 Assuming that the pair (A, B) is controllable, show that the system

$$\begin{aligned} \dot{y} &= Ay + Bv \\ \dot{v} &= u, \end{aligned}$$

with the state space $\mathbb{R}^{n+m}$ and the set of control parameters $\mathbb{R}^m$, is also controllable. Deduce that for arbitrary $a, b \in \mathbb{R}^n$, $u_0, u_1 \in \mathbb{R}^m$ and $T > 0$ there exists a control $u(\,\cdot\,)$ of class C^∞ transferring a to b at time T and such that $u(0) = u_0$, $u(T) = u_1$.

Hint. Use Exercise 1.6 and the Kalman rank condition.

Exercise 1.8 Suppose that $A \in \mathbf{M}(n, n)$, $B \in \mathbf{M}(n, m)$. Prove that the system

$$\frac{d^2y}{dt^2} = Ay + Bu, \quad y(0) \in \mathbb{R}^n, \ \frac{dy}{dt}(0) \in \mathbb{R}^n,$$

is controllable in $\mathbb{R}^{2n}$ if and only if the pair (A, B) is controllable.

Exercise 1.9 Consider system (1.9) on $[0, T]$ with integrable matrix-valued functions $A(t)$, $B(t)$, $t \in [0, T]$. Let $S(t)$, $t \in [0, T]$ be the fundamental solution of the equation $\dot{q} = Aq$. Assume that the matrix

$$Q_T = \int_0^T S(T)S^{-1}(s)B(s)B^*(s)(S^{-1}(s))^*S^*(T)\,ds$$

is positive definite. Show that the control

$$\hat{u}(s) = B^*(S^{-1}(s))^*S^*(T)Q_T^{-1}(b - S(T)a), \quad s \in [0, T],$$

transfers a to b at time T minimizing the functional $u \longrightarrow \int_0^T |u(s)|^2\,ds$.

1.4 A classification of control systems

Let $y(t)$, $t \geq 0$, be a solution of the equation (1.1) corresponding to a control $u(t)$, $t \geq 0$, and let $P \in \mathbf{M}(n, n)$ and $S \in \mathbf{M}(m, m)$ be nonsingular matrices. Define

$$\tilde{y}(t) = Py(t), \quad \tilde{u}(t) = Su(t), \quad t \geq 0.$$

Then

$$\begin{aligned}\frac{d}{dt}\tilde{y}(t) &= P\frac{d}{dt}y(t) = PAy(t) + PBu(t)\\ &= PAP^{-1}\tilde{y}(t) + PBS^{-1}\tilde{u}(t)\\ &= \widetilde{A}\tilde{y}(t) + \widetilde{B}\tilde{u}(t), \quad t \geq 0,\end{aligned}$$

where

$$\widetilde{A} = PAP^{-1}, \quad \widetilde{B} = PBS^{-1}. \tag{1.24}$$

The control systems described by (A, B) and $(\widetilde{A}, \widetilde{B})$ are called *equivalent* if there exist nonsingular matrices $P \in \mathbf{M}(n, n)$, $S \in \mathbf{M}(m, m)$, such that (1.24) holds. Let us remark that P^{-1} and S^{-1} can be regarded as transition matrices from old to new bases in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. The introduced concept is an equivalence relation. It is clear that a pair (A, B) is controllable if and only if $(\widetilde{A}, \widetilde{B})$ is controllable.

We now give a complete description of equivalent classes of the introduced relation in the case when $m = 1$.

Let us first consider a system

$$\frac{d^{(n)}}{dt^{(n)}}z + a_1\frac{d^{(n-1)}}{dt^{(n-1)}}z + \ldots + a_n z = u, \tag{1.25}$$

with initial conditions

$$z(0) = \xi_1, \quad \frac{dz}{dt}(0) = \xi_2, \quad \ldots, \quad \frac{d^{(n-1)}z}{dt^{(n-1)}}(0) = \xi_n. \tag{1.26}$$

Let $z(t)$, $\frac{dz}{dt}(t), \ldots, \frac{d^{(n-1)}z}{dt^{(n-1)}}(t)$, $t \geq 0$, be coordinates of a function $y(t)$, $t \geq 0$, and $\xi_1, \ldots, \xi_n$ coordinates of a vector x. Then

$$\dot{y} = \widetilde{A}y + \widetilde{B}u, \quad y(0) = x \in \mathbb{R}^n, \tag{1.27}$$

where matrices $\widetilde{A}$ and $\widetilde{B}$ are of the form

$$\widetilde{A} = \begin{bmatrix} 0 & 1 & \ldots & 0 & 0\\ 0 & 0 & \ldots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \ldots & 0 & 1\\ -a_n & -a_{n-1} & \ldots & -a_2 & -a_1 \end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix} 0\\ \vdots\\ 0\\ 1 \end{bmatrix}. \tag{1.28}$$

We easily check that on the main diagonal of the matrix $[\widetilde{A}|\widetilde{B}]$ there are only ones and above the diagonal only zeros. Therefore $\operatorname{rank}[\widetilde{A}|\widetilde{B}] = n$ and,

by Theorem 1.2, the pair $(\widetilde{A}, \widetilde{B})$ is controllable. Interpreting this result in terms of the initial system (1.21)–(1.22) we can say that for two arbitrary sequences of n numbers $\xi_1, \ldots, \xi_n$ and $\eta_1, \ldots, \eta_n$ and for an arbitrary positive number T there exists an analytic function $u(t)$, $t \in [0, T]$, such that for the corresponding solution $z(t)$, $t \in [0, T]$, of the equation (1.25)–(1.26)

$$z(T) = \eta_1, \quad \frac{dz}{dt}(T) = \eta_2, \quad \ldots, \quad \frac{d^{(n-1)}z}{dt^{(n-1)}}(T) = \eta_n.$$

Theorem 1.4 states that an arbitrary controllable system with the one dimensional space of control parameters is equivalent to a system of the form (1.25)–(1.26).

Theorem 1.4 *If $A \in \mathbf{M}(n, n)$, $b \in \mathbf{M}(n, 1)$ and the system*

$$\dot{y} = Ay + bu, \;\; y(0) = x \in \mathbb{R}^n \tag{1.29}$$

is controllable then it is equivalent to exactly one system of the form (1.28). *Moreover the numbers $a_1, \ldots, a_n$ in the representation* (1.24) *are identical to the coefficients of the characteristic polynomial of the matrix A:*

$$p(\lambda) = \det[\lambda I - A] = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n, \quad \lambda \in \mathbb{C}. \tag{1.30}$$

Proof. By the Cayley-Hamilton theorem, $A^n + a_1A^{n-1} + \ldots + a_nI = 0$. In particular

$$A^n b = -a_1A^{n-1}b - \ldots - a_n b.$$

Since $\operatorname{rank}[A|b] = n$, therefore vectors $e_1 = A^{n-1}b, \ldots, e_n = b$ are linearly independent and form a basis in $\mathbb{R}^n$. Let $\xi_1(t), \ldots, \xi_n(t)$ be coordinates of the vector $y(t)$ in this basis, $t \geq 0$. Then

$$\frac{d\xi}{dt} = \begin{bmatrix} -a_1 & 1 & 0 & \ldots & 0 & 0 \\ -a_2 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -a_{n-1} & 0 & 0 & \ldots & 0 & 1 \\ -a_n & 0 & 0 & \ldots & 0 & 0 \end{bmatrix} \xi + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u. \tag{1.31}$$

Therefore an arbitrary controllable system (1.29) is equivalent to (1.31) and the numbers $a_1, \ldots, a_n$ are the coefficients of the characteristic polynomial of A. On the other hand, direct calculation of the determinant of $[\lambda I - \widetilde{A}]$ gives

$$\det(\lambda I - \widetilde{A}) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n = p(\lambda), \quad \lambda \in \mathbb{C}.$$

Therefore the pair $(\widetilde{A}, \widetilde{B})$ is equivalent to the system (1.31) and consequently also to the pair (A, b). □

Remark The problem of an exact description of the equivalence classes in the case of arbitrary m is much more complicated; see [27] and [29].

1.5 Kalman decomposition

Theorem 1.2 gives several characterizations of controllable systems. Here we deal with uncontrollable ones.

Theorem 1.5 *Assume that*

$$\operatorname{rank}[A|B] = l < n.$$

There exists a nonsingular matrix $P \in \mathbf{M}(n,n)$ *such that*

$$PAP^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad PB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

where $A_{11} \in \mathbf{M}(l,l)$, $A_{22} \in \mathbf{M}(n-l, n-l)$, $B_1 \in \mathbf{M}(l,m)$. *In addition the pair*

$$(A_{11}, B_1)$$

is controllable.

The theorem states that there exists a basis in $\mathbb{R}^n$ such that system (1.1) written with respect to that basis has a representation

$$\begin{aligned} \dot{\xi}_1 &= A_{11}\xi_1 + A_{12}\xi_2 + B_1 u, \quad && \xi_1(0) \in \mathbb{R}^l, \\ \dot{\xi}_2 &= A_{22}\xi_2, && \xi_2(0) \in \mathbb{R}^{n-l}, \end{aligned}$$

in which (A_{11}, B_1) is a controllable pair. The first equation describes the so-called *controllable part* and the second the *completely uncontrollable* part of the system.

Proof. It follows from Lemma 1.1 that the subspace $E_0 = \mathcal{L}_T(L^1[0,T;\mathbb{R}^m])$ is identical with the image of the transformation l_n. Therefore it consists of all elements of the form $Bu_1 + ABu_1 + \ldots + A^{n-1}Bu_n$, $u_1, \ldots, u_n \in \mathbb{R}^m$ and is of dimension l. In addition it contains the image of B and by the Cayley-Hamilton theorem, it is invariant with respect to the transformation A. Let E_1 be any linear subspace of $\mathbb{R}^n$ complementing E_0 and let $e_1, \ldots, e_l$

and $e_{l+1}, \ldots, e_n$ be bases in E_0 and E_1 and P the transition matrix from the new to the old basis. Let $\widetilde{A} = PAP^{-1}$, $\widetilde{B} = PB$,

$$\widetilde{A}\begin{bmatrix}\xi_1\\ \xi_2\end{bmatrix} = \begin{bmatrix}A_{11}\xi_1 + A_{12}\xi_2\\ A_{21}\xi_1 + A_{22}\xi_2\end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix}B_1 u\\ B_2 u\end{bmatrix},$$

$\xi_1 \in \mathbb{R}^l$, $\xi_2 \in \mathbb{R}^{n-l}$, $u \in \mathbb{R}^m$. Since the space E_0 is invariant with respect to A, therefore

$$\widetilde{A}\begin{bmatrix}\xi_1\\ 0\end{bmatrix} = \begin{bmatrix}A_{11}\xi_1\\ 0\end{bmatrix}, \quad \xi_1 \in \mathbb{R}^l.$$

Taking into account that $B(\mathbb{R}^m) \subset E_0$,

$$B_2 u = 0 \quad \text{for } u \in \mathbb{R}^m.$$

Consequently the elements of the matrices A_{22} and B_2 are zero. This finishes the proof of the first part of the theorem.

To prove the final part, let us remark that for the nonsingular matrix P

$$\operatorname{rank}[A|B] = \operatorname{rank}(P[A|B]) = \operatorname{rank}[\widetilde{A}|\widetilde{B}].$$

Since

$$[\widetilde{A}|\widetilde{B}] = \begin{bmatrix}B_1 & A_{11}B_1 & \ldots & A_{11}^{n-1}B_1\\ 0 & 0 & \ldots & 0\end{bmatrix},$$

so

$$l = \operatorname{rank}[\widetilde{A}|\widetilde{B}] = \operatorname{rank}[A_{11}|B_1].$$

Taking into account that $A_{11} \in \mathbf{M}(l,l)$, one gets the required property. □

Remark Note that the subspace E_0 consists of all points attainable from 0. It follows from the proof of Theorem 1.5 that E_0 is the smallest subspace of R^n invariant with respect to A and containing the image of B, and it is identical to the image of the transformation represented by $[A|B]$.

Exercise 1.10 Give a complete classification of controllable systems when $m = 1$ and the dimension of E_0 is $l < n$.

1.6 Observability

Assume that $B = 0$. Then system (1.1) is identical with the linear equation

$$\dot{z} = Az, \quad z(0) = x. \tag{1.32}$$

The observation relation (1.2) we leave unchanged:

$$w = Cz. \tag{1.33}$$

The solution to (1.32) will be denoted by $z^x(t)$, $t \geq 0$. Obviously

$$z^x(t) = S(t)x, \quad x \in \mathbb{R}^n.$$

The system (1.32)- (1.33), or the pair (A, C), is said to be *observable* if for arbitrary $x \in \mathbb{R}^n$, $x \neq 0$, there exists a $t > 0$ such that

$$w(t) = Cz^x(t) \neq 0.$$

If for a given $T > 0$ and for arbitrary $x \neq 0$ there exists $t \in [0, T]$ with the above property, then the system (1.32)- (1.33) or the pair (A, C) are said to be *observable at time* T. Let us introduce the so-called *observability matrix*:

$$R_T = \int_0^T S^*(r)C^*CS(r)\,dr.$$

The following theorem, dual to Theorem 1.2, holds.

Theorem 1.6 *The following conditions are equivalent.*

(i) *System (1.32)-(1.33) is observable.*
(ii) *System (1.32)-(1.33) is observable at a given time* $T > 0$.
(iii) *The matrix* R_T *is nonsingular for some* $T > 0$.
(iv) *The matrix* R_T *is nonsingular for arbitrary* $T > 0$.
(v) $\operatorname{rank}[A^*|C^*] = n$.

Proof. Analysis of the function $w(\cdot)$ implies the equivalence of (i) and (ii). Besides,

$$\begin{aligned}\int_0^T |w(r)|^2\,dr &= \int_0^T |Cz^x(r)|^2\,dr \\ &= \int_0^T \langle S^*(r)C^*CS(r)x, x\rangle\,dr \\ &= \langle R_T x, x\rangle.\end{aligned}$$

Therefore observability at time $T \geq 0$ is equivalent to $\langle R_T x, x\rangle \neq 0$ for arbitrary $x \neq 0$ and consequently to nonsingularity of the nonnegative, symmetric matrix R_T. The remaining equivalences are consequences of Theorem 1.2

and the observation that the controllability matrix corresponding to (A^*, C^*) is exactly R_T. □

Example 1.1. Let us consider the equation

$$\frac{d^{(n)}z}{dt^{(n)}} + a_1\frac{d^{(n-1)}z}{dt^{(n-1)}} + \ldots + a_n z = 0, \tag{1.34}$$

and assume that

$$w(t) = z(t), \quad t \geq 0. \tag{1.35}$$

Matrices A and C corresponding to (1.34)-(1.35) are of the form

$$A = \begin{bmatrix} 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \\ -a_n & -a_{n-1} & \ldots & -a_2 & -a_1 \end{bmatrix}, \quad C = [1, 0, \ldots, 0].$$

We check directly that rank $[A^*|C^*] = n$ and thus the pair (A, C) is observable.

The next theorem is analogous to Theorem 1.5 and gives a decomposition of system (1.32)-(1.33) into observable and completely unobservable parts.

Theorem 1.7. *Assume that rank* $[A^*|C^*] = l < n$. *Then there exists a nonsingular matrix* $P \in \mathbf{M}(n, n)$ *such that*

$$PAP^{-1} = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \quad CP^{-1} = [C_1, 0],$$

where $A_{11} \in \mathbf{M}(l, l)$, $A_{22} \in \mathbf{M}(n-l, n-l)$ *and* $C_1 \in \mathbf{M}(k, l)$ *and the pair* (A_{11}, C_1) *is observable.*

Proof. The theorem follows directly from Theorem 1.5 and from the observation that a pair (A, C) is observable if and only if the pair (A^*, C^*) is controllable. □

Remark. It follows from the above theorem that there exists a basis in $\mathbb{R}^n$ such that the system (1.1)-(1.2) has representation

$$\begin{aligned} \dot{\xi}_1 &= A_{11}\xi_1 + B_1 u, \\ \dot{\xi}_2 &= A_{21}\xi_1 + A_{22}\xi_2 + B_2 u, \\ \eta &= C_1\xi_1, \end{aligned}$$

and the pair $(A_{11}, C_1$ is observable.

Remark Basic concepts of the chapter are due to R. Kalman [16]. He is also the author of Theorems 1.2, 1.5 and 1.6. Exercise 1.3 as well as Proposition 1.3 are due to R. Triggiani [26].

2 Stability and stabilizability

2.1 Stable linear systems

In this chapter stable linear systems are characterized in terms of associated characteristic polynomials. A formulation of the Routh theorem on stable polynomials is given as well as a complete description of completely stabilizable systems.

Let $A \in \mathbf{M}(n,n)$ and consider linear systems

$$\dot{z} = Az, \quad z(0) = x \in \mathbb{R}^n. \tag{2.1}$$

Solutions of equation (2.1) will be denoted by $z^x(t)$, $t \geq 0$. In accordance with earlier notations we have that

$$z^x(t) = S(t)x = (\exp tA)x, \quad t \geq 0.$$

The system (2.1) is called *stable* if for arbitrary $x \in \mathbb{R}^n$

$$z^x(t) \longrightarrow 0, \quad \text{as } t \uparrow +\infty.$$

Instead of saying that (2.1) is stable we will often say that the matrix A is stable. Let us remark that the concept of stability does not depend on the choice of the basis in $\mathbb{R}^n$. Therefore if P is a nonsingular matrix and A is a stable one, then matrix PAP^{-1} is stable.

In what follows we will need the Jordan theorem [31] on canonical representation of matrices. Denote by $\mathbf{M}(n,m;\mathbb{C})$ the set of all matrices with n rows and m columns and with complex elements. Let us recall that a number $\lambda \in \mathbb{C}$ is called an *eigenvalue* of a matrix $A \in \mathbf{M}(n,n;\mathbb{C})$ if there exists a vector $a \in \mathbb{C}^n$, $a \neq 0$, such that $Aa = \lambda a$. The set of all eigenvalues of a matrix A will be denoted by $\sigma(A)$. Since $\lambda \in \sigma(A)$ if and only if the matrix $\lambda I - A$ is singular, therefore $\lambda \in \sigma(A)$ if and only if $p(\lambda) = 0$, where p is a *characteristic polynomial* of A : $p(\lambda) = \det[\lambda I - A]$, $\lambda \in \mathbb{C}$. The set $\sigma(A)$ consists of at most n elements and is nonempty.

Theorem 2.1 *For an arbitrary matrix $A \in \mathbf{M}(n,n;\mathbb{C})$ there exists a nonsingular matrix $P \in \mathbf{M}(n,n;\mathbb{C})$ such that*

$$PAP^{-1} = \begin{bmatrix} J_1 & 0 & \dots & 0 & 0 \\ 0 & J_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & J_{r-1} & 0 \\ 0 & 0 & \dots & 0 & J_r \end{bmatrix} = \widetilde{A}, \tag{2.2}$$

where $J_1, J_2, \dots, J_r$ are the so-called Jordan blocks

$$J_k = \begin{bmatrix} \lambda_k & \gamma_k & \dots & 0 & 0 \\ 0 & \lambda_k & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_k & \gamma_k \\ 0 & 0 & \dots & 0 & \lambda_k \end{bmatrix}, \quad \gamma_k \neq 0 \text{ or } J_k = [\lambda_k],\ k = 1, \dots, r.$$

In the representation (2.2) *at least one Jordan block corresponds to an eigenvalue $\lambda_k \in \sigma(A)$. Selecting matrix P properly one can obtain a representation with numbers $\gamma_k \neq 0$ given in advance.*

For matrices with real elements the representation theorem has the following form:

Theorem 2.2 *For an arbitrary matrix $A \in \mathbf{M}(n,n)$ there exists a nonsingular matrix $P \in \mathbf{M}(n,n)$ such that* (2.2) *holds with "real" blocks I_k. Blocks I_k, $k = 1, \dots, r$, corresponding to real eigenvalues $\lambda_k = \alpha_k \in \mathbb{R}$ are of the form*

$$[\alpha_k] \quad \text{or} \quad \begin{bmatrix} \alpha_k & \gamma_k & \dots & 0 & 0 \\ 0 & \alpha_k & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \alpha_k & \gamma_k \\ 0 & 0 & \dots & 0 & \alpha_k \end{bmatrix}, \quad \gamma_k \neq 0,\ \gamma_k \in \mathbb{R},$$

and corresponding to complex eigenvalues $\lambda_k = \alpha_k + i\beta_k$, $\beta_k \neq 0$, $\alpha_k, \beta_k \in \mathbb{R}$,

$$\begin{bmatrix} K_k & L_k & \dots & 0 & 0 \\ 0 & K_k & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & K_k & L_k \\ 0 & 0 & \dots & 0 & K_k \end{bmatrix} \quad \text{where } K_k = \begin{bmatrix} \alpha_k & \beta_k \\ -\beta_k & \alpha_k \end{bmatrix}, \; L_k = \begin{bmatrix} \gamma_k & 0 \\ 0 & \gamma_k \end{bmatrix},$$

compare [2].

We now prove the following theorem.

Theorem 2.3 *Assume that* $A \in \mathbf{M}(n,n)$. *The following conditions are equivalent:*

(i) $z^x(t) \longrightarrow 0$ *as* $t \uparrow +\infty$, *for arbitrary* $x \in \mathbb{R}^n$.
(ii) $z^x(t) \longrightarrow 0$ *exponentially as* $t \uparrow +\infty$, *for arbitrary* $x \in \mathbb{R}^n$.
(iii) $\omega(A) = \sup\{\operatorname{Re}\lambda;\ \lambda \in \sigma(A)\} < 0$.
(iv) $\int_0^{+\infty} |z^x(t)|^2\, dt < +\infty$ *for arbitrary* $x \in \mathbb{R}^n$.

For the proof we will need the following lemma.

Lemma 2.1 *Let* $\omega > \omega(A)$. *For arbitrary norm* $\|\cdot\|$ *on* $\mathbb{R}^n$ *there exist constants* M *such that*

$$\|z^x(t)\| \leq M e^{\omega t}\|x\| \quad \text{for } t \geq 0 \text{ and } x \in \mathbb{R}^n.$$

Proof. Let us consider equation (2.1) with the matrix A in the Jordan form (2.2)

$$\dot{x} = \widetilde{A}w, \quad w(0) = x \in \mathbb{C}^n.$$

For $a = a_1 + ia_2$, where $a_1, a_2 \in \mathbb{R}^n$ set $\|a\| = \|a_1\| + \|a_2\|$. Let us decompose vector $w(t)$, $t \geq 0$ and the initial state x into sequences of vectors $w_1(t), \dots, w_r(t)$, $t > 0$ and $x_1, \dots, x_r$ according to the decomposition (2.2). Then

$$\dot{w}_k = J_k w_k, \quad w_k(0) = x_k, \; k = 1, \dots, r.$$

Let $j_1, \dots, j_r$ denote the dimensions of the matrices $J_1, \dots, J_r$, $j_1 + j_2 + \dots + j_r = n$.

If $j_k = 1$ then

$$w_k(t) = e^{\lambda_k t} x_k, \quad t \geq 0.$$

So $\|w_k(t)\| = e^{(\operatorname{Re}\lambda_k)t}\|x_k\|$, $t \geq 0$.

If $j_k > 1$, then

$$w_k(t) = e^{\lambda_k t} \sum_{l=0}^{j_k-1} \begin{bmatrix} 0 & \gamma_k & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \gamma_k \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}^l x_k \frac{t^l}{l!}.$$

So

$$\|w_k(t)\| \le e^{(\operatorname{Re}\lambda_k)t} \|x_k\| \sum_{l=0}^{j_k-1} (M_k)^l \frac{t^l}{l!}, \quad t \ge 0,$$

where M_k is the norm of the transformation represented by

$$\begin{bmatrix} 0 & \gamma_k & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \gamma_k \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Setting $\omega_0 = \omega(A)$ we get

$$\sum_{k=1}^{r} \|w_k(t)\| \le e^{\omega_0 t} q(t) \sum_{k=1}^{r} \|x_k\|, \quad t \ge 0,$$

where q is a polynomial of order at most $\max(j_k - 1)$, $k = 1, \dots, r$. If $\omega > \omega_0$ and

$$M_0 = \sup \left\{ q(t) e^{(\omega_0 - \omega)t}, \quad t \ge 0 \right\},$$

then $M_0 < +\infty$ and

$$\sum_{k=1}^{r} \|w_k(t)\| \le M_0 e^{\omega t} \sum_{k=1}^{r} \|x_k\|, \quad t \ge 0.$$

Therefore for a new constant M_1

$$\|w(t)\| \le M_1 e^{\omega t} \|x\|, \quad t \ge 0.$$

Finally

$$\|z^x(t)\| = \|P w(t) P^{-1}\| \le M_1 e^{\omega t} \|P\| \, \|P^{-1}\| \, \|x\|, \quad t \ge 0,$$

and this is enough to define $M = M_1\|P\|\,\|P^{-1}\|$. □

Proof of the theorem. Assume $\omega_0 \geq 0$. There exist $\lambda = \alpha + i\beta$, $\operatorname{Re}\lambda = \alpha \geq 0$ and a vector $a \neq 0$, $a = a_1 + ia_2$, $a_1, a_2 \in \mathbb{R}^n$ such that

$$A(a_1 + ia_2) = (\alpha + i\beta)(a_1 + ia_2).$$

The function

$$z(t) = z_1(t) + iz_2(t) = e^{(\alpha+i\beta)t}a, \quad t \geq 0,$$

as well as its real and imaginary parts, is a solution of (2.1). Since $a \neq 0$, either $a_1 \neq 0$ or $a_2 \neq 0$. Let us assume, for instance, that $a_1 \neq 0$ and $\beta \neq 0$. Then

$$z_1(t) = e^{\alpha t}(\cos\beta t)a_1 - (\sin\beta t)a_2, \quad t \geq 0.$$

Inserting $t = 2\pi k/\beta$, we have

$$|z_1(t)| = e^{\alpha t}|a_1|$$

and, taking $k \uparrow +\infty$, we obtain $z_1(t) \not\to 0$.

Now let $\omega_0 < 0$ and $\alpha \in (0, -\omega_0)$. Then by the lemma

$$|z^x(t)| \leq Me^{-\alpha t}|x| \quad \text{for } t \geq 0 \text{ and } x \in \mathbb{R}^n.$$

This implies (ii) and therefore also (i).

It remains to consider (iv). It is clear that it follows from (ii) and thus also from (iii). Let us assume that condition (iv) holds and $\omega_0 \geq 0$. Then $|z_1(t)| = e^{\alpha t}|a_1|$, $t \geq 0$, and therefore

$$\int_0^{+\infty} |z_1(t)|^2\,dt = +\infty,$$

a contradiction. The proof is complete. □

Exercise 2.1 The matrix

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix}$$

corresponds to the equation $\ddot{z} + 2\dot{z} + 2z = 0$. Calculate $\omega(A)$. For $\omega > \omega(A)$ find the smallest constant $M = M(\omega)$ such that

$$|S(t)| \leq Me^{\omega t}, \quad t \geq 0.$$

Hint. Prove that $|S(t)| = \varphi(t)e^{-t}$, where

$$\varphi(t) = \frac{1}{2}\left(2 + 5\sin^2 t + (20\sin^2 t + 25\sin^4 t)^{1/2}\right)^{1/2}, \quad t \geq 0.$$

2.2 Stable polynomials

Theorem 2.3 reduces the problem of determining whether a matrix A is stable to the question of finding out whether all roots of the characteristic polynomial of A have negative real parts. Polynomials with this property will be called *stable*. Because of its importance, several efforts have been made to find necessary and sufficient conditions for the stability of an arbitrary polynomial

$$p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n, \quad \lambda \in \mathbb{C}, \tag{2.3}$$

with real coefficients, in term of the coefficients $a_1, \ldots, a_n$. Since there is no general formula for roots of polynomials of order greater than 4, the existence of such conditions is not obvious. Therefore their formulation in the nineteenth century by Routh was a kind of a sensation. Before formulating and proving a version of the Routh theorem we will characterize stable polynomials of degree smaller than or equal to 4 using only the fundamental theorem of algebra. We deduce also a useful necessary condition for stability.

Theorem 2.4

(1) *Polynomials with real coefficients*:

(i) $\lambda + a$,

(ii) $\lambda^2 + a\lambda + b$,

(iii) $\lambda^3 + a\lambda^2 + b\lambda + c$,

(iv) $\lambda^4 + a\lambda^3 + b\lambda^2 + c\lambda + d$

are stable if and only if, respectively

(i)* $a > 0$,

(ii)* $a > 0,\ b > 0$,

(iii)* $a > 0,\ b > 0,\ c > 0$ *and* $ab > c$,

(iv)* $a > 0,\ b > 0,\ c > 0,\ d > 0$ *and* $abc > c^2 + a^2 d$.

(2) *If polynomial* (2.3) *is stable then all its coefficients* $a_1, \ldots, a_n$ *are positive.*

Proof. (1) Equivalence (i)$\Longleftrightarrow$(i)* is obvious.

To prove (ii)$\Longleftrightarrow$(ii)* assume that the roots of the polynomial are of the form $\lambda_1 = -\alpha + i\beta$, $\lambda_2 = -\alpha - i\beta$, $\beta \neq 0$. Then $p(\lambda) = \lambda^2 + 2\alpha\lambda + \beta^2$, $\lambda \in \mathbb{C}$ and therefore the stability conditions are $a > 0$ and $b > 0$. If the roots λ_1, λ_2 of the polynomial p are real then $a = -(\lambda_1 + \lambda_2)$, $b = \lambda_1\lambda_2$. Therefore they are negative if only if $a > 0$, $b > 0$.

To show that (iii)$\Longleftrightarrow$(iii)* let us remark that the fundamental theorem of algebra implies the following decomposition of the polynomial, with real

coefficients α, β, γ:

$$p(\lambda) = \lambda^3 + a\lambda^2 + b\lambda + c = (\lambda + \alpha)(\lambda^2 + \beta\lambda + \gamma), \quad \lambda \in \mathbb{C}.$$

It therefore follows from (i) and (ii) that the polynomial p is stable if only if $\alpha > 0$, $\beta > 0$ and $\gamma > 0$. Comparing the coefficients gives

$$a = \alpha + \beta, \quad b = \gamma + \alpha\beta, \quad c = \alpha\gamma,$$

and therefore $ab - c = \beta(\alpha^2 + \gamma + \alpha\beta) = \beta(\alpha^2 + b)$.

Assume that $a > 0$, $b > 0$, $c > 0$ and $ab - c > 0$. It follows from $b > 0$ and $ab - c > 0$ that $\beta > 0$. Since $c = \alpha\gamma$, α and γ are either positive or negative. They cannot, however, be negative because then $b = \gamma + \alpha\beta < 0$. Thus $\alpha > 0$ and $\gamma > 0$ and consequently $\alpha > 0$, $\beta > 0$, $\gamma > 0$. It is clear from the above formulae that the positivity of α, β, γ implies inequalities (iii)*. To prove (iv)$\Longleftrightarrow$(iv)* we again apply the fundamental theorem of algebra to obtain the representation

$$\lambda^4 + a\lambda^3 + b\lambda^2 + c\lambda + d = (\lambda^2 + \alpha\lambda + \beta)(\lambda^2 + \gamma\lambda + \delta)$$

and the stability condition $\alpha > 0$, $\beta > 0$, $\gamma > 0$, $\delta > 0$.

From the decomposition

$$a = \alpha + \gamma, \quad b = \alpha\gamma + \beta + \delta, \quad c = \alpha\delta + \beta\gamma, \quad d = \beta\delta,$$

we check directly that

$$abc - c^2 - a^2 d = \alpha\gamma\left((\beta - \delta)^2 + ac\right).$$

It is therefore clear that $\alpha > 0$, $\beta > 0$, $\gamma > 0$ and $\delta > 0$, and then (iv)* holds. Assume now that the inequalities (iv)* are true. Then $\alpha\gamma > 0$, and, since $a = \alpha + \gamma > 0$, therefore $\alpha > 0$ and $\delta > 0$. Since, in addition, $d = \beta\delta > 0$ and $c = \alpha\delta + \beta\gamma > 0$, so $\beta > 0$, $\delta > 0$. Finally $\alpha > 0$, $\beta > 0$, $\gamma > 0$, $\delta > 0$, and the polynomial p is stable.

(2) By the fundamental theorem of algebra, the polynomial p is a product of polynomials of degrees at most 2 which, by (1), have positive coefficients. This implies the result. □

Exercise 2.2 Find necessary and sufficient conditions for the polynomial

$$\lambda^2 + a\lambda + b$$

with complex coefficients a and b to have all roots with negative real parts.

Hint. Consider the polynomial $(\lambda^2+a\lambda+b)(\lambda^2+\bar{a}\lambda+\bar{b})$ and apply Theorem 2.4.

We now formulate a theorem which allows us to check, in a finite number of steps, that a given polynomial $p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n$, $\lambda \in \mathbb{C}$, with real coefficients is stable. As we already know, a stable polynomial has all coefficients positive, but this condition is not sufficient for stability if $n > 3$. Let U and V be polynomials with real coefficients given by

$$U(x) + iV(x) = p(ix), \quad x \in \mathbb{R}.$$

Let us remark that $\deg U = n$, $\deg V = n-1$ if n is an even number and $\deg U = n-1$, $\deg V = n$, if n is an odd number. Denote $f_1 = U$, $f_2 = V$ if $\deg U = n$, $\deg V = n-1$ and $f_1 = V$, $f_2 = U$ if $\deg V = n$, $\deg U = n-1$. Let $f_3, f_4, \ldots, f_m$ be polynomials obtained from f_1, f_2 by an application of the Euclid algorithm. Thus $\deg f_{k+1} < \deg f_k$, $k = 2, \ldots, m-1$ and there exist polynomials $\kappa_1, \ldots, \kappa_m$ such that

$$f_{k-1} = \kappa_k f_k - f_{k+1}, \quad f_{m-1} = \kappa_m f_m.$$

Moreover the polynomial f_m is equal to the largest common divisor of f_1, f_2 multiplied by a constant.

The following theorem is due to F. J. Routh [23]. For the proof, see [31].

Theorem 2.5 *A polynomial p is stable if and only if $m = n+1$ and the signs of the leading coefficients of the polynomials $f_1, \ldots, f_{n+1}$ alternate.*

Let us apply the above theorem to polynomials of degree 4,

$$p(\lambda) = \lambda^4 + a\lambda^3 + b\lambda^2 + c\lambda + d, \quad \lambda \in \mathbb{C}.$$

In this case

$$\begin{aligned} U(x) &= x^4 - bx^2 + d = f_1(x), \\ V(x) &= -ax^3 + cx = f_2(x), \quad x \in \mathbb{R}. \end{aligned}$$

Performing appropriate divisions we obtain

$$\begin{aligned} f_3(x) &= \left(b - \frac{c}{a}\right) x^2 - d, \\ f_4(x) &= -\left(c - ad\left(b - \frac{c}{a}\right)^{-1}\right) x, \\ f_5(x) &= d. \end{aligned}$$

The leading coefficients of the polynomials $f_1, f_2, \ldots, f_5$ are

$$1,\ -a,\ \left(b-\frac{c}{a}\right),\ -\left(c-ad\left(b-\frac{c}{a}\right)^{-1}\right),\ d.$$

We obtain therefore the following necessary and sufficient conditions for the stability of the polynomial p:

$$a>0,\ b-\frac{c}{a}>0,\ c-ad\left(b-\frac{c}{a}\right)>0,\ d>0,$$

equivalent to those stated in Theorem 2.4.

We leave as an exercise the proof that the Routh theorem leads to an explicit stability algorithm. To formulate it we have to define the so-called *Routh array.*

For arbitrary sequences (α_k), (β_k), the *Routh sequence* (γ_k) is defined by

$$\gamma_k=-\frac{1}{\beta_1}\det\begin{bmatrix}\alpha_1 & \alpha_{k+1}\\ \beta_1 & \beta_{k+1}\end{bmatrix},\quad k=1,2,\ldots$$

If $a_1,\ldots,a_n$ are coefficients of a polynomial p, we set additionally $a_k=0$ for $k>n=\deg p$. The *Routh array* is a matrix with infinite rows obtained from the first two rows:

$$1, a_2, a_4, a_6, \ldots,$$
$$a_1, a_3, a_5, a_7, \ldots,$$

by consecutive calculations of the Routh sequences from the two preceding rows. The calculations stop if 0 appears in the first column. The Routh algorithm can be now stated as the theorem

Theorem 2.6 *A polynomial p of degree n is stable if and only if the $n+1$ first elements of the first columns of the Routh array are positive.*

Exercise 2.3 Show that, for an arbitrary polynomial $p(\lambda)=\lambda^n+a_1\lambda^{n-1}+\ldots+a_n$, $\lambda\in\mathbb{C}$, with complex coefficients $a_1,\ldots,a_n$, the polynomial $(\lambda^n+a_1\lambda^{n-1}+\ldots+a_n)(\lambda^n+\bar{a}_1\lambda^{n-1}+\ldots+\bar{a}_n)$ has real coefficients. Formulate necessary and sufficient conditions for the polynomial p to have all roots with negative real parts.

2.3 Stabilizability and controllability

We say that the system

$$\dot{y} = Ay + Bu, \quad y(0) = x \in \mathbb{R}^n, \tag{2.4}$$

is *stabilizable* or that the pair (A, B) is *stabilizable* if there exists a matrix $K \in \mathbf{M}(m, n)$ such that the matrix $A + BK$ is stable. So if the pair (A, B) is stabilizable and a control $u(\cdot)$ is given in the *feedback* form

$$u(t) = Ky(t), \quad t \geq 0,$$

then all solutions of the equation

$$\dot{y}(t) = Ay(t) + BKy(t) = (A + BK)y(t), \quad y(0) = x, \ t \geq 0, \tag{2.5}$$

tend to zero as $t \uparrow +\infty$.

We say that system (2.4) is *completely stabilizable* if and only if for arbitrary $\omega > 0$ there exist a matrix K and a constant $M > 0$ such that for an arbitrary solution $y^x(t)$, $t \geq 0$, of (2.5)

$$|y^x(t)| \leq Me^{-\omega t}|x|, \quad t \geq 0. \tag{2.6}$$

By p_K we will denote the characteristic polynomial of the matrix $A + BK$. One of the most important results in the linear control theory is given by

Theorem 2.7 *The following conditions are equivalent*:

(i) *System* (2.4) *is completely stabilizable.*

(ii) *System* (2.4) *is controllable.*

(iii) *For arbitrary polynomial* $p(\lambda) = \lambda^n + \alpha_1\lambda^{n-1} + \ldots + \alpha_n$, $\lambda \in \mathbb{C}$, *with real coefficients, there exists a matrix* K *such that*

$$p(\lambda) = p_K(\lambda) \quad \textit{for } \lambda \in \mathbb{C}.$$

Proof. We start with the implication (ii)$\Longrightarrow$(iii) and prove it in three steps.

Step 1. The dimension of the space of control parameters $m = 1$. It follows from §1.4 that we can limit our considerations to systems of the form

$$\frac{d^{(n)}z}{dt^{(n)}}(t) + a_1\frac{d^{(n-1)}z}{dt^{(n-1)}}(t) + \ldots + a_nz(t) = u(t), \quad t \geq 0.$$

In this case, however, (iii) is obvious: It is enough to define the control u in the feedback form,

$$u(t) = (a_1 - \alpha_1)\frac{d^{(n-1)}z}{dt^{(n-1)}}(t) + \ldots + (a_n - \alpha_n)z(t), \quad t \geq 0,$$

and use the result (see §1.4) that the characteristic polynomial of the equation

$$\frac{d^{(n)}z}{dt^{(n)}} + \alpha_1\frac{d^{(n-1)}z}{dt^{(n-1)}} + \ldots + \alpha_n z = 0,$$

or, equivalently, of the matrix

$$\begin{bmatrix} 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 \\ -\alpha_n & -\alpha_{n-1} & \ldots & -\alpha_2 & -\alpha_1 \end{bmatrix},$$

is exactly

$$p(\lambda) = \lambda^n + \alpha_1\lambda^{n-1} + \ldots + \alpha_n\lambda, \quad \lambda \in \mathbb{C}.$$

Step 2. The following lemma allows us to reduce the general case to $m = 1$. Note that in its formulation and proof its vectors from $\mathbb{R}^n$ are treated as one-column matrices.

Lemma 2.2 *If a pair (A, B) is controllable then there exist a matrix $L \in \mathbf{M}(m, n)$ and a vector $v \in \mathbb{R}^m$ such that the pair $(A+BL, Bv)$ is controllable.*

Proof of the lemma. It follows from the controllability of (A, B) that there exists $v \in \mathbb{R}^m$ such that $Bv \neq 0$. We show first that there exist vectors $u_1, \ldots, u_{n-1}$ in $\mathbb{R}^m$ such that the sequence $e_1, \ldots, e_n$ defined inductively

$$e_1 = Bv, \ e_{l+1} = Ae_l + Bu_l \quad \text{for } l = 1, 2, \ldots, n-1 \tag{2.7}$$

is a basis in $\mathbb{R}^n$. Assume that such a sequence does not exist. Then for some $k \geq 0$ vectors $e_1, \ldots, e_k$, corresponding to some $u_1, \ldots, u_k$ are linearly independent, and for arbitrary $u \in \mathbb{R}^m$ the vector $Ae_k + Bu$ belongs to the linear space E_0 spanned by $e_1, \ldots, e_k$. Taking $u = 0$ we obtain $Ae_k \in E_0$. Thus $Bu \in E_0$ for arbitrary $u \in \mathbb{R}^m$ and consequently $Ae_j \in E_0$ for $j = 1, \ldots, k$. This way we see that the space E_0 is invariant for A and contains the image of B. Controllability of (A, B) implies now that $E_0 = \mathbb{R}^n$, and

compare the remark following Theorem 1.5. Consequently $k = n$ and the required sequences $e_1, \dots, e_n$ and $u_1, \dots, u_{n-1}$ exist. Let u_n be an arbitrary vector from $\mathbb{R}^m$.

We define the linear transformation L setting $Le_l = u_l$, for $l = 1, \dots, n$. We have from (2.7)

$$\begin{aligned} e_{l+1} &= Ae_l + BLe_l = (A + BL)e_l \\ &= (A + BL)^l e_1 \\ &= (A + BL)^l Bv, \quad l = 0, 1, \dots, n-1. \end{aligned}$$

Since

$$[A + BL|Bv] = [e_1, e_2, \dots, e_n],$$

the pair $(A + BL, Bv)$ is controllable. □

Step 3. Let a polynomial p be given and let L and v be the matrix and vector constructed in Step 2. The system

$$\dot{y} = (A + BL)y + (Bv)u,$$

in which $u(\cdot)$ is a scalar control function, is controllable. It follows from Step 1 that there exists $k \in \mathbb{R}^n$ such that the characteristic polynomial of $(A + BL) + (Bv)k^* = A + B(L + vk^*)$ is identical with p.

The required feedback K can be defined as

$$K = L + vk^*.$$

We proceed to the proofs of the remaining implications. To show that (iii)$\Longrightarrow$(ii) assume that (A, B) is not controllable, that $\operatorname{rank}[A|B] = l < n$ and that K is a linear feedback. Let $P \in \mathbf{M}(n, n)$ be a nonsingular matrix from Theorem 1.5. Then

$$\begin{aligned} p_K(\lambda) &= \det[\lambda I - (A + BK)] \\ &= \det[\lambda I - (PAP^{-1} + PBKP^{-1})] \\ &= \det \begin{bmatrix} (\lambda I - (A_{11} + B_1 K_1)) & -A_{12} \\ 0 & (\lambda I - A_{22}) \end{bmatrix} \\ &= \det[\lambda I - (A_{11} + B_1 K_1)] \det[\lambda I - A_{22}], \quad \lambda \in \mathbb{C}, \end{aligned}$$

where $K_1 \in \mathbf{M}(m, n)$. Therefore for arbitrary $K \in \mathbf{M}(m, n)$ the polynomial p_K has a nonconstant divisor, equal to the characteristic polynomial of A_{22},

and therefore p_K cannot be arbitrary. This way the implication (iii)$\Longrightarrow$(ii) is true.

Assume now that condition (i) holds but that the system is not controllable. By the above argument we have for arbitrary $K \in \mathbf{M}(m,n)$ that $\sigma(A_{22}) \subset \sigma(A+BK)$. So if for some $M > 0$, $\omega > 0$ condition (2.6) holds then

$$\omega \leq -\sup\{\operatorname{Re}\lambda;\ \lambda \in \sigma(A_{22})\},$$

which contradicts complete stabilizability. Hence (i)$\Longrightarrow$(ii). Assume now that (ii) and therefore (iii) hold. Let $p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \ldots + a_n$, $\lambda \in \mathbb{C}$ be a polynomial with all roots having real parts smaller than $-\omega$ (e.g., $p(\lambda) = (\lambda+\omega+\varepsilon)^n, \varepsilon > 0$). We have from (iii) that there exists a matrix K such that $p_K(\,\cdot\,) = p(\,\cdot\,)$. Consequently all eigenvalues of $A+BK$ have real parts smaller than $-\omega$. By Theorem 2.3, condition (i) holds. The proof of Theorem 2.7 is complete. □

Remark The proof of Theorem 2.7 is due to M. Wonham [28].

3 Linear quadratic problem

3.1 Introductory comments

This chapter starts from a derivation of the dynamic programming equations called Bellman's equations. They are used to solve the linear regulator problem on a finite time interval. A fundamental role is played here by the Riccati-type matrix differential equations. The stabilization problem is reduced to an analysis of an algebraic Riccati equation.

Our considerations will be devoted mainly to control systems

$$\dot{y} = f(y,u), \quad y(0) = x, \tag{3.1}$$

and to *criteria*, called also *cost functionals*,

$$J_T(x,u(\,\cdot\,)) = \int_0^T g(y(t),u(t))\,dt + G(y(T)), \tag{3.2}$$

when $T < +\infty$. If the control interval is $[0,+\infty]$, then the *cost functional*

$$J(x,u(\,\cdot\,)) = \int_0^{+\infty} g(y(t),u(t))\,dt. \tag{3.3}$$

Our aim will be to find a control $\hat{u}(\cdot)$ such that for all admissible controls $u(\cdot)$

$$J_T(x, \hat{u}(\cdot)) \leq J_T(x, u(\cdot)) \tag{3.4}$$

or

$$J(x, \hat{u}(\cdot)) \leq J(x, u(\cdot)). \tag{3.5}$$

There are basically two methods for finding controls minimizing cost functionals (3.2) or (3.3).

One of them *embeds* a given minimization problem into a parametrized family of similar problems. The embedding should be such that the minimal value, as a function of the parameter, satisfies an analytic relation. If the selected parameter is the initial state and the length of the control interval, then the minimal value of the cost functional is called the value function and the analytical relation, Bellman's equation. Knowing the solutions to the Bellman equation one can find the optimal strategy in the form of a closed loop control.

The other method leads to necessary conditions on the optimal, open-loop, strategy formulated in the form of the so-called maximum principle discovered by L. Pontryagin and his collaborators. They can be obtained (in the simplest case) by considering a parametrized family of controls and the corresponding values of the cost functional (3.2) and by an application of classical calculus.

3.2 Bellman's equation and the value function

Assume that the state space E of a control system is an open subset of $\mathbb{R}^n$ and let the set U of control parameters be included in $\mathbb{R}^m$. We assume that the functions f, g and G are continuous on $E \times U$ and E respectively and that g is nonnegative.

Theorem 3.1 *Assume that a real function $W(\cdot,\cdot)$, defined and continuous on $[0,T] \times E$, is of class C^1 on $(0,T) \times E$ and satisfies the equation*

$$\frac{\partial W}{\partial t}(t,x) = \inf_{u \in U}(g(x,u) + \langle W_x(t,x), f(x,u)\rangle), \quad (t,x) \in (0,T) \times E, \tag{3.6}$$

with the boundary condition

$$W(0,x) = G(x), \quad x \in E. \tag{3.7}$$

(i) *If $u(\cdot)$ is a control and $y(\cdot)$ the corresponding absolutely continuous, E-valued, solution of* (3.1), *then*

$$J_T(x, u(\cdot)) \geq W(T, x). \tag{3.8}$$

(ii) *Assume that for a certain function $\hat{v} : [0,T] \times E \longrightarrow U$*

$$\begin{aligned} g(x, \hat{v}(t,x)) + \langle W_x(t,x), f(x, \hat{v}(t,x))\rangle \\ \leq g(x,u) + \langle W_x(t,x), f(x,u)\rangle, \quad t \in (0,T),\ x \in E,\ u \in U, \end{aligned} \tag{3.9}$$

and that $\hat{y}$ is an absolutely continuous, E-valued solution of the equation

$$\begin{aligned} \frac{d}{dt}\hat{y}(t) &= f(\hat{y}(t), \hat{v}(T-t, \hat{y}(t))), \quad t \in [0,T], \\ \hat{y}(0) &= x. \end{aligned} \tag{3.10}$$

Then, for the control $\hat{u}(t) = \hat{v}(T-t, \hat{y}(t))$, $t \in [0,T]$,

$$J_T(x, \hat{u}(\cdot)) = W(x,T).$$

Proof. (i) Let $w(t) = W(T-t, y(t))$, $t \in [0,T]$. Then $w(\cdot)$ is an absolutely continuous function on an arbitrary interval $[\alpha, \beta] \subset (0,T)$ and

$$\begin{aligned} \frac{dw}{dt}(t) &= -\frac{\partial W}{\partial t}(T-t, y(t)) + \left\langle W_x(T-t, y(t)), \frac{dy}{dt}(t)\right\rangle \\ &= -\frac{\partial W}{\partial t}(T-t, y(t)) + \langle W_x(T-t, y(t)), f(y(t), u(t))\rangle \end{aligned} \tag{3.11}$$

for almost all $t \in [0,T]$. Hence, from (3.6) and (3.7)

$$\begin{aligned} W(T-\beta, y(\beta)) - W(T-\alpha, y(\alpha)) = w(\beta) - w(\alpha) = \int_\alpha^\beta \frac{dw}{dt}(t)\, dt \\ = \int_\alpha^\beta \left[-\frac{\partial W}{\partial t}(T-t, y(t)) + \langle W_x(T-t, y(t)), f(y(t), u(t))\rangle\right] dt \\ \geq -\int_\alpha^\beta g(y(t), u(t))\, dt. \end{aligned}$$

Letting α and β tend to 0 and T respectively we obtain

$$G(y(T)) - W(T,x) \geq -\int_0^T g(y(t), u(t))\, dt.$$

This proves (i).

(ii) In a similar way, taking into account (3.9), for the control $\hat{u}$ and the output $\hat{y}$,

$$G(\hat{y}(T)) - W(T,x) = \int_0^T \left[-\frac{\partial W}{\partial t}(T-t, \hat{y}(t)) + \langle W_x(T-t, \hat{y}(t)) \rangle \right] dt$$
$$= \int_0^T g(\hat{y}(t), \hat{u}(t))\, dt.$$

Therefore

$$G(\hat{y}(T)) + \int_0^T g(\hat{y}(s), \hat{u}(s))\, ds = W(T,x),$$

the required identity. □

Remark Equation (3.6) is called *Bellman's equation.* It follows from Theorem 3.1 that, under appropriate conditions, $W(T,x)$ is the minimal value of the functional $J_T(x,\cdot)$. Hence W is the *value function* for the problem of minimizing (3.2).

Let $U(t,x)$ be the set of all control parameters $u \in U$ for which the infimum on the right-hand side of (3.6) is attained. The function $\hat{v}(\cdot,\cdot)$ from part (ii) of the theorem is a *selector* of the multivalued function $U(\cdot,\cdot)$ in the sense that

$$\hat{v}(t,x) \in U(t,x), \quad (t,x) \in [0,T] \times E.$$

Therefore, for the conditions of the theorem to be fulfilled, such a selector not only should exist, but the closed loop equation (3.10) should have a well defined, absolutely continuous, solution.

Remark A similar result holds for a more general cost functional

$$J_T(x, u(\cdot)) = \int_0^T e^{-\alpha t} g(y(t), u(t))\, dt + \bar{e}^{\alpha T} G(y(T)). \tag{3.12}$$

In this direction we propose to solve the following exercise.

Exercise 3.1 Taking into account a solution $W(\cdot,\cdot)$ of the equation

$$\frac{\partial W}{\partial t}(t,x) = \inf_{u \in U} (g(x,u) - \alpha W(t,x) + \langle W_x(t,x), f(x,u) \rangle),$$
$$W(0,x) = G(x), \quad x \in E,\ t \in (0,T),$$

and a selector $\hat{v}$ of the multivalued function

$$U(t,x) = \Big\{u \in U;\, g(x,u) + \langle W_x(t,x), f(x,u)\rangle = \inf_{u\in U}(g(x,u) + \langle W_x(t,x), f(x,u)\rangle)\Big\},$$

generalize Theorem 3.1 to the functional (3.12).

We will now describe an intuitive derivation of equation (3.6). Similar reasoning often helps to guess the proper form of the Bellman equation in situations different from the one covered by Theorem 3.1.

Let $W(t,x)$ be the minimal value of the functional $J_t(x,\,\cdot\,)$. For arbitrary $h > 0$ and arbitrary parameter $v \in U$ denote by $u^v(\,\cdot\,)$ a control which is constant and equal v on $[0,h)$ and is identical with the optimal strategy for the minimization problem on $[h, t+h]$. Let $z^{x,v}(t)$, $t \geq 0$, be the solution of the equation $\dot{z} = f(z,v)$, $z(0) = x$. Then

$$J_{t+h}(x, u^v(\,\cdot\,)) = \int_0^h g(z^{x,v}(s), v)\, ds + W(t, z^{x,v}(h))$$

and, approximately,

$$W(t+h,x) \approx \inf_{v\in U} J_{t+h}(x, u^v(\,\cdot\,)) \approx \inf_{v\in U} \int_0^h g(z^{x,v}(s), v)\, ds + W(t, z^{x,v}(h)).$$

Subtracting $W(t,x)$ we obtain that

$$\frac{1}{h}(W(t+h,x) - W(t,x)) \approx \inf_{u\in U}\Big[\frac{1}{h}\int_0^h g(z^{x,v}(s), v)\, ds + \frac{1}{h}(W(t, z^{x,v}(h)) - W(t,x))\Big].$$

Assuming that the function W is differentiable and taking the limits as $h \downarrow 0$ we arrive at (3.6). □

Exercise 3.2 Show that the solution of the Bellman equation corresponding to the optimal consumption model of Example 0.3, with $\alpha \in (0,1)$, is of the form

$$W(t,x) = p(t)x^\alpha, \quad t \geq 0,\ x \geq 0,$$

where the function $p(\cdot)$ is the unique solution of the following differential equation:

$$\dot{p} = \begin{cases} 1, & \text{for } p \le 1, \\ \alpha p + (1-\alpha)\left(\frac{1}{p}\right)^{\alpha/(1-\alpha)}, & \text{for } p \ge 1, \end{cases}$$
$$p(0) = a.$$

Find the optimal strategy.

Hint. First prove the following lemma.

Lemma 3.1 *Let $\psi_p(u) = \alpha u p + (1-u)^\alpha$, $p \ge 0$, $u \in [0,1]$. The maximal value $m(p)$ of the function $\psi_p(\cdot)$ is attained at*

$$u(p) = \begin{cases} 0, & \text{if } p > 1, \\ \left(\frac{1}{p}\right)^{1/(1-\alpha)}, & \text{if } p \in [0,1]. \end{cases}$$

Moreover

$$m(p) = \begin{cases} 1, & \text{if } p \ge 1, \\ \alpha p + (1-\alpha)\left(\frac{1}{p}\right)^{\alpha/(1-\alpha)}, & \text{if } p \in [0,1]. \end{cases}$$

3.3 The linear regulator problem and the Riccati equation

We now consider a special case of Problems (3.1) and (3.4) when the system equation is linear

$$\dot{y} = Ay + Bu, \quad y(0) = x \in \mathbb{R}^n, \tag{3.13}$$

$A \in \mathbf{M}(n,n)$, $B \in \mathbf{M}(n,m)$, the state space $E = \mathbb{R}^n$ and the set of control parameters $U = \mathbb{R}^m$. We assume that the cost functional is of the form

$$J_T = \int_0^T \left(\langle Qy(s), y(s)\rangle + \langle Ru(s), u(s)\rangle\right) ds + \langle P_0 y(T), y(T)\rangle, \tag{3.14}$$

where $Q \in \mathbf{M}(n,n)$, $R \in \mathbf{M}(m,m)$, $P_0 \in \mathbf{M}(n,n)$ are symmetric, non-negative matrices and the matrix R is positive definite. The problem of minimizing (3.14) for a linear system (3.13) is called the *linear regulator problem* or the *linear-quadratic problem*.

The form of an optimal solution to (3.13) and (3.14) is strongly connected with the following *matrix Riccati equation*:

$$\dot{P} = Q + PA + A^*P - PBR^{-1}B^*P, \quad P(0) = P_0, \tag{3.15}$$

in which $P(s)$, $s \in [0,T]$, is the unknown function with values in $\mathbf{M}(n,n)$. The following theorem takes place.

Theorem 3.2 *Equation* (3.15) *has a unique global solution* $P(s)$, $s \geq 0$. *For arbitrary* $s \geq 0$ *the matrix* $P(s)$ *is symmetric and nonnegative definite. The minimal value of the functional* (3.14) *is equal to* $\langle P(T)x, x\rangle$ *and the optimal control is of the form*

$$\hat{u}(t) = -R^{-1}B^*P(T-t)\hat{y}(t), \quad t \in [0,T], \tag{3.16}$$

where

$$\dot{\hat{y}}(t) = (A - BR^{-1}B^*P(T-t))\hat{y}(t), \quad t \in [0,T], \quad \hat{y}(0) = x. \tag{3.17}$$

Proof. The proof will be given in several steps.

Step 1. For an arbitrary symmetric matrix P_0 equation (3.15) has exactly one local solution and the values of the solution are symmetric matrices.

Equation (3.15) is equivalent to a system of n^2 differential equations for elements $p_{ij}(\cdot)$, $i,j = 1,2,\ldots,n$ of the matrix $P(\cdot)$. The right-hand sides of these equations are polynomials of order 2, and therefore the system has a unique local solution being a smooth function of its argument. Let us remark that the same equation is also satisfied by $P^*(\cdot)$. This is because matrices Q, R and P_0 are symmetric. Since the solution is unique, $P(\cdot) = P^*(\cdot)$, and the values of $P(\cdot)$ are symmetric matrices.

Step 2. Let $P(s)$, $s \in [0,T_0)$, be a symmetric solution of (3.15) and let $T < T_0$. The function $W(s,x) = \langle P(s)x, x\rangle$, $s \in [0,T]$, $x \in \mathbb{R}^n$, is a solution of the Bellman equation (3.6)–(3.7) associated with the linear regular problem (3.13)–(3.14).

The condition (3.7) follows directly from the definitions. Moreover, for arbitrary $x \in \mathbb{R}^n$ and $t \in [0,T]$

$$\begin{aligned}&\inf_{u\in\mathbb{R}^n}\left(\langle Qx,x\rangle + \langle Ru,u\rangle + 2\langle P(t)x, Ax+Bu\rangle\right)\\ &= \langle Qx,x\rangle + \langle (A^*P(t)+P(t)A)x,x\rangle + \inf_{u\in\mathbb{R}^m}\left(\langle Ru,u\rangle + \langle u, 2B^*P(t)x\rangle\right).\end{aligned} \tag{3.18}$$

We need now the following lemma, the proof of which is left as an exercise.

Lemma 3.2 *If a matrix $R \in \mathbf{M}(m,m)$ is positive definite and $a \in \mathbb{R}^m$, then for arbitrary $u \in \mathbb{R}^m$*

$$\langle Ru, u\rangle + \langle a, u\rangle \geq -\frac{1}{4}\langle R^{-1}a, a\rangle.$$

Moreover, the equality holds if and only if

$$u = -\frac{1}{2}R^{-1}a.$$

It follows from the lemma that the expression (3.18) is equal to

$$\langle Q + A^*P(t) + P(t)A^* - P(t)BR^{-1}B^*P(A)x, x\rangle$$

and that the infimum in formula (3.18) is attained at exactly one point given by

$$-R^{-1}B^*P(t)x, \quad t \in [0,T].$$

Since $P(t)$, $t \in [0,T_0)$, satisfies the equation (3.15), the function W is a solution to the problem (3.6)–(3.7).

Step 3. The control $\hat{u}$ given by (3.16) on $[0,T]$, $T < T_0$, is optimal with respect to the functional $J_T(x,\cdot)$.

This fact is a direct consequence of Theorem 3.1.

Step 4. For arbitrary $t \in [0,T]$, $T < T_0$, the matrix $P(t)$ is nonnegative definite and

$$\langle P(t)x, x\rangle \leq \int_0^t \langle Q\tilde{y}^x(s), \tilde{y}^x(s)\rangle\, ds + \langle P_0\tilde{y}^x(t), \tilde{y}^x(t)\rangle, \tag{3.19}$$

where $\tilde{y}^x(\,\cdot\,)$ is the solution to the equation

$$\dot{\tilde{y}} = A\tilde{y}, \quad \tilde{y}(0) = x.$$

Applying Theorem 3.1 to the function $J_t(x,\cdot)$ we see that its minimal value is equal to $\langle P(t)x, x\rangle$. For arbitrary control $u(\,\cdot\,)$, $J_t(x,u) \geq 0$, the matrix $P(t)$ is nonnegative definite. In addition, estimate (3.19) holds because its right-hand side is the value of the functional $J_t(x,\cdot)$ for the control $u(s) = 0$, $s \in [0,t]$.

Step 5. For arbitrary $t \in [0,T_0)$ and $x \in \mathbb{R}^n$

$$0 \leq \langle P(t)x, x\rangle \leq \left\langle \left(\int_0^t S^*(r)QS(r)\, dr + S^*(t)P_0S(t)\right)x, x\right\rangle,$$

where $S(r) = e^{Ar}$, $r \geq 0$.

This result is an immediate consequence of the estimate (3.19).

Exercise 3.3 Show that if, for some symmetric matrices $P = (p_{ij}) \in \mathbf{M}(n,n)$ and $S = (s_{ij}) \in \mathbf{M}(n,n)$,

$$0 \leq \langle Px, x\rangle \leq \langle Sx, x\rangle, \quad x \in \mathbb{R}^n,$$

then

$$-\frac{1}{2}(s_{ii} + s_{jj}) \leq p_{ij} \leq s_{ij} + \frac{1}{2}(s_{ii} + s_{jj}), \quad i,j = 1,\dots,n.$$

It follows from Step 5 and Exercise 3.3 that solutions of (3.15) are bounded in $\mathbf{M}(n,n)$ and therefore an arbitrary maximal solution $P(\cdot)$ in $\mathbf{M}(n,n)$ exists for all $t \geq 0$.

The proof of the theorem is complete. □

Exercise 3.4 Solve the linear regulator problem with a more general cost functional

$$\int_0^T \left(\langle Q(y(t) - a), y(t) - a\rangle + \langle Ru(t), u(t)\rangle\right)\, dt + \langle P_0 y(T), y(T)\rangle,$$

where $a \in \mathbb{R}^n$ is a given vector.

Answer. Let $P(t)$, $q(t)$, $r(t)$, $t \geq 0$, be solutions of the following matrix, vector and scalar equations respectively,

$$\begin{aligned}
\dot{P} &= Q + A^*P + PA - PBR^{-1}B^*P, \quad P(0) = P_0,\\
\dot{q} &= A^*q - PBR^{-1}q - 2Qa, \quad q(0) = 0,\\
\dot{r} &= -\frac{1}{4}\langle R^{-1}q, q\rangle + \langle Qa, a\rangle, \quad r(0) = 0.
\end{aligned}$$

The minimal value of the functional is equal to

$$r(T) + \langle q(T), x\rangle + \langle P(T)x, x\rangle,$$

and the optimal, feedback strategy is of the form

$$u(t) = -\frac{1}{2}R^{-1}q(T-t) - R^{-1}B^*P(T-t)y(t), \quad t \in [0,T].$$

3.4 The linear regulator and stabilization

The obtained solution of the linear regulator problem suggests an important way to stabilize linear systems. It is related to the *algebraic Riccati equation*

$$Q + PA + A^*P - PBR^{-1}B^*P = 0, \quad P \geq 0, \tag{3.20}$$

in which the unknown is a nonnegative definite matrix P. If $\widetilde{P}$ is a solution to (3.20) and $\widetilde{P} \leq P$ for all the other solutions P, then $\widetilde{P}$ is called a *minimal solution* of (3.20). For arbitrary control $u(\,\cdot\,)$ defined on $[0, +\infty)$ we introduce the notation

$$J(x,u) = \int_0^{+\infty} \left(\langle Qy(s), y(s)\rangle + \langle Ru(s), u(s)\rangle\right) ds. \tag{3.21}$$

Theorem 3.3 *If there exists a nonnegative solution P of equation* (3.20) *then there also exists a unique minimal solution $\widetilde{P}$ of* (3.20)*, and the control $\tilde{u}$ given in the feedback form*

$$\tilde{u}(t) = -R^{-1}B^*\widetilde{P}y(t), \quad t \geq 0,$$

minimizes functional (3.21)*. Moreover the minimal value of the cost functional is equal to*

$$\langle \widetilde{P}x, x\rangle.$$

Proof. Let us first remark that if $P_1(t)$, $P_2(t)$, $t \geq 0$, are solutions of (3.15) and $P_1(0) \leq P_2(0)$ then $P_1(t) \leq P_2(t)$ for all $t \geq 0$. This is because the minimal value of the functional

$$J_t^1(x,u) = \int_0^t \left(\langle Qy(s), y(s)\rangle + \langle Ru(s), u(s)\rangle\right) ds + \langle P_1(0)y(t), y(t)\rangle$$

is not greater than the minimal value of the functional

$$J_t^2(x,u) = \int_0^t \left(\langle Qy(s), y(s)\rangle + \langle Ru(s), u(s)\rangle\right) ds + \langle P_2(0)y(t), y(t)\rangle,$$

and by Theorem 3.2 the minimal values are $\langle P_1(t)x, x\rangle$ and $\langle P_2(t)x, x\rangle$ respectively.

If, in particular, $P_1(0) = 0$ and $P_2(0) = P$ then $P_2(t) = P$ and therefore $P_1(t) \leq P$ for all $t \geq 0$. It also follows from Theorem 3.2 that the function $P_1(\,\cdot\,)$ is nondecreasing with respect to the natural order existing in the space of symmetric matrices. This easily implies that for arbitrary $i, j = 1, 2, \ldots, n$ there exist finite limits $\tilde{p}_{ij} = \lim\limits_{t\uparrow+\infty} \tilde{p}_{ij}(t)$, where $(\tilde{p}_{ij}(t)) = P_1(t)$, $t \geq 0$. Taking into account equation (3.15) we see that there exist finite limits

$$\lim_{t\uparrow+\infty} \frac{d}{dt}\tilde{p}_{ij}(t) = \gamma_{ij}, \quad i, j = 1, \ldots, n.$$

These limits have to be equal to zero, for if $\gamma_{i,j} > 0$ or $\gamma_{i,j} < 0$ then $\lim_{t\uparrow+\infty} \tilde{p}_{ij}(t) = +\infty$. But $\lim_{t\uparrow+\infty} \tilde{p}_{ij}(t) = -\infty$, a contradiction. Hence the matrix $\widetilde{P} = (\tilde{p}_{ij})$ satisfies equation (3.20). It is clear that $\widetilde{P} \leq P$.

Now let $\tilde{y}(\,\cdot\,)$ be the output corresponding to the input $\tilde{u}(\,\cdot\,)$. By Theorem 3.2, for arbitrary $T \geq 0$ and $x \in \mathbb{R}^n$,

$$\langle \widetilde{P}x, x\rangle = \int_0^T (\langle Q\tilde{y}(t), \tilde{y}(t)\rangle + \langle R\tilde{u}(t), \tilde{u}(t)\rangle)\, dt + \langle \widetilde{P}\tilde{y}(T), \tilde{y}(T)\rangle, \qquad (3.22)$$

and

$$\int_0^T (\langle Q\tilde{y}(t), \tilde{y}(t)\rangle + \langle R\tilde{u}(t), \tilde{u}(t)\rangle)\, dt \leq \langle \widetilde{P}x, x\rangle.$$

Letting T tend to $+\infty$ we obtain

$$J(x, \tilde{u}) \leq \langle \widetilde{P}x, x\rangle.$$

On the other hand, for arbitrary $T \geq 0$ and $x \in \mathbb{R}^m$,

$$\langle P_1(T)x, x\rangle \leq \int_0^T (\langle Q\tilde{y}(t), \tilde{y}(t)\rangle + \langle R\tilde{u}(t), \tilde{u}(t)\rangle)\, dt \leq J(x, \tilde{u}),$$

consequently, $\langle \widetilde{P}x, x\rangle \leq J(x, \tilde{u})$ and finally

$$J(x, \tilde{u}) = \langle \widetilde{P}x, x\rangle.$$

The proof is complete. □

Exercise 3.5 For the control system

$$\ddot{y} = u,$$

find the strategy which minimizes the functional

$$\int_0^{+\infty} (y^2 + u^2)\, dt$$

and the minimal value of this functional.

Answer. The solution of equation (3.20) in which $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $R = [1]$, is matrix $P = \begin{bmatrix} \sqrt{2} & 1 \\ 1 & \sqrt{2} \end{bmatrix}$. The optimal strategy

is of the form $u = -y - \sqrt{2}(\dot{y})$ and the minimal value of the functional is $\sqrt{2}(y(0))^2 + 2y(0)\dot{y}(0) + \sqrt{2}(\dot{y}(0))^2$.

For stabilizability the following result is essential. We need a new concept of *detectability*. A pair of matrices (A, C) is detectable if there exists a matrix L of proper dimension such that the matrix $A + LC$, is stable.

Theorem 3.4

(i) *If the pair* (A, B) *is stabilizable then equation* (3.20) *has at least one solution.*

(ii) *If* $Q = C^*C$ *and the pair* (A, C) *is detectable then equation* (3.20) *has at most one solution, and if* P *is the solution then the matrix* $A - BR^{-1}B^*P$ *is stable.*

Proof. (i) Let K be a matrix such that the matrix $A+BK$ is stable. Consider a feedback control $u(t) = Ky(t)$, $t \geq 0$. It follows from the stability of $A + BK$ that $y(t) \longrightarrow 0$, and therefore $u(t) \longrightarrow 0$ exponentially as $t \uparrow +\infty$. Thus for arbitrary $x \in \mathbb{R}^n$,

$$J(x, u(\,\cdot\,)) = \int_0^{+\infty} \left(\langle Qy(t), y(t)\rangle + \langle Ru(t), u(t)\rangle\right) dt < +\infty.$$

Since

$$\langle P_1(T)x, x\rangle \leq J(x, u(\,\cdot\,)) < +\infty, \quad T \geq 0,$$

for the solution $P_1(t)$, $t \geq 0$, of (3.15) with the initial condition $P_1(0) = 0$, there exists $\lim\limits_{T\uparrow+\infty} P_1(T) = P$ which satisfies (3.20). (Compare the proof of the previous theorem.)

(ii) We prove first the following lemma.

Lemma 3.3

(i) *Assume that for some matrices* $M \geq 0$ *and* K *of appropriate dimensions,*

$$M(A - BK) + (A - BK)^*M + C^*C + K^*RK = 0. \tag{3.23}$$

If the pair (A, C) *is detectable, then the matrix* $A - BK$ *is stable.*

(ii) *If, in addition,* P *is a solution to* (3.20)*, then* $P \leq M$.

Proof. (i) Let $S_1(t) = e^{(A-BK)t}$, $S_2(t) = e^{(A-LC)t}$, where L is a matrix such that $A - LC$ is stable and let $y(t) = S_1(t)x$, $t \geq 0$. Since

$$A - BK = (A - LC) + (LC - BK),$$

therefore

$$y(t) = S_2(t)x + \int_0^t S_2(t-s)(LC - BK)y(s)\,ds. \tag{3.24}$$

We show now that

$$\int_0^{+\infty} |Cy(s)|^2\,ds < +\infty \quad \text{and} \quad \int_0^{+\infty} |Ky(s)|^2\,ds < +\infty. \tag{3.25}$$

Let us remark that, for $t \geq 0$,

$$\dot{y}(t) = (A - BK)y(t) \quad \text{and} \quad \frac{d}{dt}\langle My(t), y(t)\rangle = 2\langle M\dot{y}(t), y(t)\rangle.$$

It therefore follows from (3.23) that

$$\frac{d}{dt}\langle My(t), y(t)\rangle + \langle Cy(t), Cy(t)\rangle + \langle RKy(t), Ky(t)\rangle = 0.$$

Hence, for $t \geq 0$,

$$\langle My(t), y(t)\rangle + \int_0^t |Cy(s)|^2\,ds + \int_0^t \langle RKy(s), Ky(s)\rangle\,ds = \langle Mx, x\rangle. \tag{3.26}$$

Since the matrix R is positive definite, (3.26) follows from (3.25). By (3.26),

$$|y(t)| \leq |S_2(t)x| + N\int_0^t |S_2(t-s)|(|Cy(s)| + |Ky(s)|)\,ds,$$

where $N = \max(|L|, |B|)$, $t \geq 0$. We need now the following classical result on convolutions of functions due to Young.

Lemma 3.4 *Assume that p, q, r are positive numbers such that $1/p + 1/q = 1 + 1/r$. If functions f, g belong respectively to L^p and L^q, then the convolution $f * g$ belongs to L^r and*

$$||f * g||_r \leq ||f||_p ||g||_q\,.$$

By Young's result and by (3.25),

$$\int_0^{+\infty} |y(s)|^2\,ds \leq N\int_0^{+\infty} |S_2(s)|\,ds\left(\int_0^{+\infty} (|Cy(s)| + |Ky(s)|)^2\,ds\right)^{1/2} + \left(\int_0^{+\infty} |S_2(s)|^2\,ds\right)^{1/2} |x| < +\infty.$$

It follows from Theorem 2.3(iv) that $y(t) \to 0$ as $t \to \infty$. This proves the required result.

Let us also remark that

$$M = \int_0^{+\infty} S_1^*(s)(C^*C + K^*RK)S_1(s)\,ds. \tag{3.27}$$

(ii) Define $K_0 = R^{-1}B^*P$ then $RK_0 = -B^*P$, $PB = -K_0^*R$.

Consequently,

$$P(A - BK) + (A - BK)^*P + K^*RK = -C^*C + (K - K_0)^*R(K - K_0)$$

and

$$M(A - BK) + (A - BK)^*M + K^*RK = -C^*C.$$

Hence if $V = M - P$ then

$$V(A - BK) + (A - BK)^*V + (K - K_0)^*R(K - K_0) = 0.$$

Since the matrix $A - BK$ is stable the above equation has only one solution given by the formula,

$$V = \int_0^{+\infty} S_1^*(s)(K - K_0)^*R(K - K_0)S_1(s)\,ds \geq 0,$$

and therefore $M \geq P$. The proof of the lemma is complete. □

To prove part (ii) of Theorem 3.4 assume that matrices $P \geq 0$, $P_1 \geq 0$ are solutions of (3.20). Define $K = R^{-1}B^*P$. Then

$$\begin{aligned} P(A - BK) + (A - BK)^*P + C^*C + K^*RK \\ = PA + A^*P + C^*C - PBR^{-1}B^*P = 0. \end{aligned} \tag{3.28}$$

Therefore, by Lemma 3.3(ii), $P_1 \leq P$. In the same way $P_1 \geq P$. Hence $P_1 = P$. Identity (3.28) and Lemma 3.3(i) imply the stability of $A - BK$.□

Let us recall that a pair (A, C) is observable if and only if the pair (A^*, C^*) is controllable. As a corollary from Theorem 3.4 we obtain

Theorem 3.5 *If the pair (A, B) is controllable, $Q = C^*C$ and the pair (A, C) is observable, then equation* (3.20) *has exactly one solution, and if P is this unique solution, then the matrix $A - BR^{-1}B^*P$ is stable.*

Theorem 3.5 indicates an effective way of stabilizing linear system (3.13). Controllability and observability tests in the form of the corresponding rank conditions are effective, and equation (3.20) can be solved numerically using methods similar to those for solving polynomial equations. The uniqueness of the solution of (3.20) is essential for numerical algorithms.

The following examples show that equation (3.20) does not always have a solution and that in some cases it may have many solutions.

Example 3.1 If, in (3.20), $B = 0$, then we arrive at the Lyapunov equation

$$PA + A^*P = Q, \quad P \geq 0. \tag{3.29}$$

If Q is positive definite, then equation (3.29) has at most one solution, and if, in addition, matrix A is not stable, then it does not have any solutions.

Exercise 3.6 If Q is a singular matrix then equation (3.20) may have many solutions. For if P is a solution to (3.20) and

$$\widetilde{A} = \begin{bmatrix} 0 & 0 \\ 0 & A \end{bmatrix}, \ \widetilde{Q} = \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}, \quad \widetilde{A} \in \mathbf{M}(k,k), \ k > n,$$

then, for an arbitrary nonnegative matrix $R \in \mathbf{M}(k-n, k-n)$, matrix

$$\widetilde{P} = \begin{bmatrix} R & 0 \\ 0 & P \end{bmatrix}$$

satisfies the equation

$$\widetilde{P}\widetilde{A} + \widetilde{A}^*\widetilde{P} = \widetilde{Q}.$$

Exercise 3.7 Solve the linear regulator problem on finite and infinite intervals when the control system is given by the equation;

$$\dot{y} = Ay + a + Bu, \quad y(0) = x \in \mathbb{R}^n, \tag{3.30}$$

where $a \in \mathbb{R}^n$ is a given vector.

Remark Dynamic programming ideas are presented in the monograph by R. Bellmann [3]. The results of the linear regulator problem are classic. Theorem 3.4 is due to W.M. Wonham [29]. In the proof of Lemma 3.3(i) we follow [30], see also [31].

References

[1] A. V. Balakrishnan, "Semigroup theory and control theory", in *Proc. IFIP Congress*, Tokyo 1965.

[2] S. Barnett, *Introduction to Mathematical Control Theory* (Clarendon Press, Oxford, 1975).

[3] R. Bellman, *Dynamic Programming* (Princeton University Press, 1977).

[4] A. Bensoussan, G. Da Prato, M. Delfour and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems* Birkhäuser, Vol. 1(1992), Vol. 2(1993).

[5] R. W. Brockett, *Finite Dimensional Linear Systems* (Wiley, New York, 1970).

[6] L. Cesari, *Optimization Theory and Applications* (Springer-Verlag, New York, 1963).

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley Interscience, New York, 1983).

[8] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (McGraw-Hill, New York, 1955).

[9] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory* (Lecture Notes in Control and Information Sciences, Springer-Verlag, New York, 1978).

[10] G. Da Prato, J. Math. Pures et Appl. **52**, 353 (1973).

[11] N. Dunford and J. Schwartz, *Linear operators*, Part I (Interscience Publishers, New York, London, 1958).

[12] N. Dunford and J. Schwartz, *Linear operators*, Part II (Interscience Publishers, New York, London, 1963).

[13] H. O. Fattorini, SIAM J. Control, **4**, 686 (1966).

[14] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control* (Springer-Verlag, Berlin, Heidelberg, New York, 1975).

[15] F. R. Gantmacher, *Applications of the Theory of Matrices* (Interscience Publishers Inc., New York, 1959).

[16] R. E. Kalman, "On the general theory of control systems", in: *Automatic and Remote Control*, Proc. First Int. Congress of IFAC, Moscow, 1960, vol. 1 (Izdat. AN SSSR, Moskva, 1961) pp. 481–492.

[17] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory* (Wiley, New York, 1967).

[18] G. Leitman, *An introduction to optimal control* (Mc Graw-Hill, New York, 1966).

[19] J. C. Maxwell, *On governers* (Proc. Royal Society, 1868).

[20] R. Pallu de la Barriere, *Cours d'automatique théorique* (Dunod, Paris, 1966).

[21] L. S. Pontryagin, *Ordinary Differential Equation* (Addison-Wesley, Reading, Mass., 1962).

[22] S. Rolewicz, *Functional Analysis and Control Theory* (Polish Scientific Publishers, Warszawa, and D. Reidel Publishing Company, Dordrecht, Boston, Lancaster, Tokyo, 1987).

[23] E. J. Routh, *Treatise on the Stability of a Given State of Motion* (Macmillan and Co., London, 1877).

[24] T. Schanbacher, SIAM J. Control and Optimization **27**, 457 (1989).

[25] E.D. Sontag, *Mathematical Control Theory*, Springer Verlag, 1990.

[26] R. Triggiani, *Constructive steering control functions for linear systems and abstract rank condition* (to appear).

[27] W. A. Wolvich, *Linear Multivariable Systems* (Springer-Verlag, New York, 1974).

[28] W. M. Wonham, IEEE Trans. Automat. Control **AC-12**, 660 (1967).

[29] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach* (Springer-Verlag, New York, 1979).

[30] J. Zabczyk, Appl. Math. Optimization **3**, 383 (1976).

[31] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhauser, 1996.

Linear Quadratic Control Theory for Infinite Dimensional Systems

Giuseppe Da Prato*

Scuola Normale Superiore, Pisa, Italy

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028002

*DaPrato@sns.it

Abstract

These notes contain a short course on linear quadratic controls problems in Hilbert spaces.

We have described the Dynamic Programming approach based on the solutions of the Riccati Operator Equation and of the Algebraic Riccati Equation, and presented some examples involving Heat and Wave equations.

For the sake of simplicity we have only considered the case of bounded observation and control operators. A necessary prerequisite is the theory of strongly continuous semigroup, that is recalled in the Appendix A.

To have more information and references, the reader can see the books:

A. Bensoussan, G. Da Prato, M. Delfour and S.K. Mitter, *Representation and Control of Infinite Dimensional Systems,* Vol. I, II, Birkhäuser, (1992).
I. Lasiecka and R. Triggiani, *Control theory for partial differential equations,* Vol. I, II, Encyclopedia of Mathematics and its Applications, Cambridge University Press. (1999) .

Contents

1 Control in finite horizon **63**
1.1 Introduction and setting of the problem 63
1.2 Riccati equation 65
1.3 Solution of the control problem 73

2 Control in infinite horizon **75**
2.1 Introduction and setting of the problem 75
2.2 The Algebraic Riccati Equation 77
2.3 Solution of the control problem 80

3 Examples and generalizations **84**
3.1 Parabolic equations 84
3.2 Wave equation 86
3.3 Boundary control problems 88

A Linear Semigroups Theory **91**
A.1 Some preliminaries on spectral theory 91
A.2 Strongly continuous semigroups 93
A.3 The Hille–Yosida theorem 96
A.4 Cauchy problem 101

B Contraction Principle **104**

1 Control in finite horizon

1.1 Introduction and setting of the problem

We are concerned with a dynamical system governed by the following differential equation

$$\begin{cases} y'(t) = Ay(t) + Bu(t), \ t \geq 0, \\ \\ y(0) = x \in H, \end{cases} \tag{1.1.1}$$

where $A : D(A) \subset H \to H$, $B : U \to H$ are linear operators defined on the Hilbert spaces H (*state space*) and U (*control space*). We shall also consider another Hilbert space Y (*observation space*). The inner product and norm in H, U, Y will be denoted by $\langle \cdot, \cdot \rangle$ and $| \cdot |$ respectively.

Given $T > 0$, we want to minimize the cost function

$$J(u) = \int_0^T \left[|Cy(s)|^2 + |u(s)|^2 \right] ds + \langle P_0 y(T), y(T) \rangle, \tag{1.1.2}$$

where $P_0 : H \to H$, $C : H \to Y$ are linear operators defined in H and Y respectively, over all controls $u \in L^2(0, T; U)$ subject to (1.1.1).

Concerning the operators A, B, C and P_0 we shall assume that

Hypothesis 1.1 *(i) A generates a strongly continuous semigroup e^{tA} on H.*

(ii) $B \in L(U, H)$ ([1]).

(iii) $P_0 \in L(H)$ is symmetric and nonnegative.

(iv) $C \in L(H, Y)$.

Under Hypothesis 1.1–(i)–(ii) problem (1.1.1) has a unique *mild* solution y given by the *variation of constants* formula (see Appendix A),

$$y(t) = e^{tA}x + \int_0^t e^{(t-s)A} Bu(s) ds. \tag{1.1.3}$$

[1] Let X, Y be Hilbert spaces. We denote by $L(X, Y)$ the Banach space of all linear bounded operators $T : X \to Y$ endowed with the norm $\|T\| = \sup\{|Tx| : \ x \in X, \ |x| \leq 1\}$. We set $L(X, X) = L(X)$.

A function $u^* \in L^2(0,T;U)$ is called an *optimal control* if

$$J(u^*) \leq J(u), \ \forall\, u \in L^2(0,T;U). \tag{1.1.4}$$

In this case the corresponding solution y^* of (1.1.1) is called an *optimal state* and the pair (u^*, y^*) an *optimal pair*.

Under Hypothesis 1.1 it is easy to see that there is a unique optimal control (since the quadratic form $J(u)$ on $L^2(0,T;U)$ is coercive). However we are interested in showing that the optimal control can be obtained as a *feedback control* (*synthesis problem*). For this reason we shall describe the *Dynamic Programming* approach which consists in the following two steps:

Step 1. We solve the *Riccati operator equation*

$$\begin{cases} P' = A^*P + PA - PBB^*P + C^*C, \\ \\ P(0) = P_0, \end{cases} \tag{1.1.5}$$

where A^*, B^* and C^* are the adjoint operators of A, B and C respectively.

Step 2. We prove that the optimal control u^* is related to the optimal state y^* by the *feedback formula*

$$u^*(t) = -B^*P(T-t)y^*(t), \ \ t \in [0,T], \tag{1.1.6}$$

and moreover that y^* is the mild solution of the *closed loop equation*

$$\begin{cases} y'(t) = [A - BB^*P(T-t)]y(t), \ t \geq 0, \\ \\ y(0) = x \in H. \end{cases} \tag{1.1.7}$$

Finally the optimal cost is given by

$$J^* := \langle P(T)x, x\rangle.$$

Example 1.1.1 Let D be an open subset of $\mathbb{R}^n$ with regular boundary ∂D. Consider the equation

$$\begin{cases} D_t y(t,\xi) = (\Delta_\xi + c)y(t,\xi) + u(t,\xi), \ \text{in } (0,T] \times D, \\ \\ y(t,\xi) = 0, \ \text{on } (0,T] \times \partial D, \\ \\ y(0,\xi) = x(\xi), \ \text{in } D. \end{cases} \tag{1.1.8}$$

We choose $H = U = Y = L^2(D)$, we set $B = C = P_0 = I$ and we denote by A the linear operator in H :

$$\begin{cases} Ay = (\Delta_\xi + c)y \\ \\ D(A) = H^2(D) \cap H_0^1(D). \end{cases} \tag{1.1.9}$$

It is well known that A generates a strongly continuous semigroup on $H = L^2(D)$. ([2])

Setting $y(t) = y(t, \cdot)$, $u(t) = u(t, \cdot)$, we can write (1.1.8) in the abstract form (1.1.1).

In this case the control problem consists in minimizing the cost

$$J(u) = \int_0^T \int_D [|y(t,\xi)|^2 + |u(t,\xi)|^2] dt d\xi + \int_D |y(T,\xi)|^2 d\xi. \tag{1.1.10}$$

Note that the control is distributed on all D.

1.2 Riccati equation

Let us introduce some notation. We set

$$\Sigma(H) = \{T \in L(H) : \ T \text{ is symmetric}\},$$

$$\Sigma^+(H) = \{T \in \Sigma(H) : \ \langle Tx, x\rangle \geq 0, \ \forall\, x \in H\}.$$

$\Sigma(H)$ is a closed subspace of $L(H)$, and $\Sigma^+(H)$ is a cone in $L(H)$.

For any interval $[a, b] \subset \mathbb{R}$, we shall denote by $C([a, b]; \Sigma(H))$ the set of all continuous mappings from $[a, b]$ to $\Sigma(H)$.

$C([a, b]; \Sigma(H))$, endowed with the norm

$$\|F\| = \sup_{t \in [a,b]} \|F(t)\|, \ F \in C([a, b]; \Sigma(H)),$$

is a Banach space.

We shall also need to consider the space $C_s([a, b]; \Sigma(H))$ of all strongly continuous mappings $F : [a, b] \to \Sigma(H)$, that is such that $F(\cdot)x$ is continuous on $[a, b]$ for any $x \in H$. A typical mapping belonging to $C_s([0, T]; \Sigma(H))$ is $F(t) = e^{tA}$.

[2] See e.g. J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod, (1968).

Let $F, \{F_n\} \subset C_s([a,b];\Sigma(H))$. We say that $\{F_n\}$ is strongly convergent to F if

$$\lim_{n\to\infty} F_n(\cdot)x = F(\cdot)x, \ \forall\, x \in H.$$

In this case we shall write

$$\lim_{n\to\infty} F_n = F, \ \text{in } C_s([a,b];\Sigma(H)).$$

If $F \in C_s([a,b];\Sigma(H))$, then the quantity

$$\|F\| = \sup_{t\in[a,b]} \|F(t)\|,$$

is finite by virtue of the Uniform Boundedness Theorem. Endowed with the norm above $C_s([a,b];\Sigma(H))$ is a Banach space that we shall denote by $C_u([a,b];\Sigma(H))$.

Let A, B, C and P_0 be given linear operators such that Hypothesis 1.1 is fulfilled. This section is devoted to solve the following Riccati equation

$$\begin{cases} P' = A^*P + PA - PBB^*P + C^*C, \\ \\ P(0) = P_0, \end{cases} \tag{1.2.1}$$

We first notice that if $A \in L(H)$ then it is easy to see that (1.2.1) is equivalent to the following integral equation

$$\begin{aligned} P(t)x &= e^{tA^*}P_0e^{tA}x + \int_0^t e^{sA^*}C^*Ce^{sA}x\,ds \\ &- \int_0^t e^{(t-s)A^*}P(s)BB^*P(s)e^{(t-s)A}x\,ds, \quad x \in H. \end{aligned} \tag{1.2.2}$$

Now, since the mapping

$$[0,T] \to \Sigma(H), \ t \to e^{tA^*}Te^{tA},$$

belongs to $C_s([a,b];\Sigma(H))$, equation (1.2.2) is meaningful in $C_s([a,b];\Sigma(H))$ and we will try to solve it in this space.

Definition 1.2.1 *(i) A mild solution of equation* (1.2.1) *in the interval* $[0,T]$ *is a function* $P \in C_s([a,b];\Sigma(H))$ *that verifies the integral equation* (1.2.2).

(ii) A weak solution of equation (1.2.1) *in the interval* $[0,T]$ *is a function* $P \in C_s([a,b];\Sigma(H))$ *such that* $P(0) = P_0$ *and for any* $x, y \in D(A)$, $\langle P(\cdot)x, y\rangle$ *is differentiable in* $[0,T]$ *and verifies the equation*

$$\begin{aligned} \frac{d}{dt}\langle P(t)x, y\rangle &= \langle P(t)x, Ay\rangle + \langle P(t)Ax, y\rangle \\ &- \langle B^*P(t)x, B^*P(t)y\rangle + \langle Cx, Cy\rangle. \end{aligned} \tag{1.2.3}$$

Proposition 1.2.2 *Let* $P \in C_s([a,b];\Sigma(H))$. *Then* P *is a mild solution of equation* (1.2.1) *if and only if* P *is a weak solution of equation* (1.2.1).

Proof. If P is a mild solution of equation (1.2.1), then for any $x, y \in H$ we have

$$\langle P(t)x, y\rangle = \langle P_0 e^{tA}x, e^{tA}y\rangle + \int_0^t \langle Ce^{sA}x, Ce^{sA}y\rangle ds$$

$$- \int_0^t \langle P(s)BB^*P(s)e^{(t-s)A}x, e^{(t-s)A}y\rangle ds.$$

Now if $x, y \in D(A)$ it follows that $\langle P(t)x, y\rangle$ is differentiable with respect to t and, by a simple computation, that (1.2.3) holds. Conversely if P is a weak solution, then it is easy to check that for all $x, y \in D(A)$

$$\frac{d}{ds}\langle P(s)e^{(t-s)A}x, e^{(t-s)A}y\rangle = \langle Ce^{(t-s)A}x, Ce^{(t-s)A}y\rangle$$

$$-\langle B^*P(t)e^{(t-s)A}x, B^*P(t)e^{(t-s)A}y\rangle.$$

Integrating from 0 to t we see that (1.2.2) holds for any $x \in D(A)$. Since $D(A)$ is dense in H the conclusion follows. □

It is convenient to introduce the following approximating problem

$$\begin{cases} P_n' = A_n^*P_n + P_nA_n - P_nBB^*P_n + C^*C, \\ P_n(0) = P_0, \end{cases} \tag{1.2.4}$$

where $A_n = n^2R(n,A) - nI$ is the Yosida approximation of A and $R(n,A)$ is the resolvent of A. Problem (1.2.4) is equivalent to the following integral

equation

$$P_n(t)x = e^{tA_n^*}P_0e^{tA_n}x + \int_0^t e^{sA_n^*}C^*Ce^{sA_n}xds$$
$$- \int_0^t e^{(t-s)A_n^*}P_n(s)BB^*P_n(s)e^{(t-s)A_n}xds, \quad x \in H. \tag{1.2.5}$$

We now solve problem (1.2.1). We first prove the local existence of a solution. We recall that by the Hille–Yosida Theorem (see Appendix A) for any $T > 0$ there exists $M_T > 0$ such that

$$\|e^{tA}\| \le M_T, \quad \|e^{tA_n}\| \le M_T, \ \forall\, t \in [0,T], \ n \in \mathbb{N}.$$

Lemma 1.2.3 *Assume that Hypothesis 1.1 holds, fix $T > 0$, set*

$$\rho = 2M_T^2\|P_0\| \tag{1.2.6}$$

and let τ be such that

$$\tau \in [0,T], \quad \tau\left(\|C\|^2 + \rho^2\|B\|^2\right) \le \|P_0\|, \quad 2\rho\tau M_T^2\|B\|^2 \le \frac{1}{2}. \tag{1.2.7}$$

Then problems (1.2.1) *and* (1.2.5) *have unique mild solutions P and P_n in the ball*

$$B_{\rho,\tau} = \{F \in C_u([0,\tau];\Sigma(H)) : \ \|F\| \le \rho\}.$$

Moreover

$$\lim_{n\to\infty} P_n = P, \text{ in } C_s([a,b];\Sigma(H)). \tag{1.2.8}$$

Proof. Equation (1.2.2) (resp. the integral version of equation (1.2.4)) can be written in the form

$$P = \gamma(P)(\text{resp. } P_n = \gamma_n(P_n)),$$

where for $x \in H$

$$\gamma(P)(t)x = e^{tA^*}P_0e^{tA}x$$
$$+ \int_0^t e^{(t-s)A^*}[C^*C - P(s)BB^*P(s)]e^{(t-s)A}xds$$

and

$$\begin{aligned}\gamma_n(P)(t)x &= e^{tA_n^*}P_0e^{tA_n}x \\ &+ \int_0^t e^{(t-s)A_n^*}[C^*C - P_n(s)BB^*P_n(s)]e^{(t-s)A_n}xds.\end{aligned}$$

Choose now ρ and τ such that (1.2.6) and (1.2.7) hold. We show that γ and γ_n are 1/2–contractions on the ball $B_{\rho,\tau}$ of $C_u([0,\tau];\Sigma(H))$. Let in fact $P \in B_{\rho,\tau}$. It follows that

$$|\gamma(P)(t)x| \leq M_T^2\left[\|P_0\| + \tau\|C\|^2 + \tau\rho^2\|B\|^2\right]|x| \leq 2M_T^2\|P_0\||x|,$$

and analogously

$$|\gamma_n(P)(t)x| \leq 2M_T^2\|P_0\||x|.$$

It follows that

$$\|\gamma(P)(t)\| \leq \rho, \quad \|\gamma_n(P)(t)\| \leq \rho, \quad \forall\, t \in [0,\tau],\ n \in \mathbb{N},\ P \in B_{\rho,\tau},$$

so that γ and γ_n map $B_{\rho,\tau}$ into $B_{\rho,\tau}$.

For $P, Q \in B_{\rho,\tau}$ we have

$$\begin{aligned}&\gamma(P)(t)x - \gamma(Q)(t)x \\ &= \int_0^t e^{(t-s)A^*}[PBB^*(Q-P) + (Q-P)BB^*Q](s)e^{(t-s)A}xds,\end{aligned}$$

and a similar formula holds for $\gamma_n(P)(t)x - \gamma_n(Q)(t)x$. It follows that

$$\|\gamma(P)(t) - \gamma(Q)(t)\| \leq 2\rho\tau M_T^2\|B^2\|\|P-Q\| \leq \frac{1}{2}\,\|P-Q\|,$$

$$\|\gamma_n(P)(t) - \gamma_n(Q)(t)\| \leq 2\rho\tau M_T^2\|B^2\|\|P-Q\| \leq \frac{1}{2}\,\|P-Q\|.$$

Thus γ and γ_n are 1/2–contractions on $B_{\rho,\tau}$ and there exists unique mild solutions P and P_n in $B_{\rho,\tau}$. Finally (1.2.8) follows from a generalization of the classical Contraction Mapping Principle (see Appendix B). □

We now prove global uniqueness.

Lemma 1.2.4 *Assume that Hypothesis 1.1 holds, let $T > 0$ and let P, Q be two mild solutions of problem* (1.2.1) *in* $[0,T]$. *Then* $P = Q$.

Proof. Set

$$\alpha = \sup_{t\in[0,T]} \max\left\{\|P(t)\|, \|Q(t)\|\right\}.$$

α is finite by the Uniform Boundedness Theorem. Choose $\rho > 0$ and $\tau \in [0,T]$ such that

$$\rho = 2M_T^2\alpha, \quad \tau\left(\|C\|^2 + \rho^2\|B\|^2\right) \le \alpha, \ \ 2\rho\tau M_T^2\|B\|^2 \le \frac{1}{2}.$$

By Lemma 1.2.3 it follows that $P(t) = Q(t)$ for any $t \in [0,\tau]$. It is now sufficient to repeat this argument in the interval $[\tau, 2\tau]$ and so on. □

The main result of this section is the following theorem.

Theorem 1.2.5 *Assume that Hypothesis 1.1 holds. Then problem* (1.2.1) *has a unique mild solution* $P \in C_s([0,+\infty);\Sigma_+(H))$. *Moreover for each* $n \in \mathbb{N}$ *problem* (1.2.5) *has a unique mild solution* $P_n \in C([0,+\infty);\Sigma_+(H))$ *and*

$$\lim_{n\to\infty} P_n = P \text{ in } C_s([0,T];\Sigma_+(H)),$$

for any $T > 0$.

Proof. Fix $T > 0$, set $\beta = M_T^2\left(\|P_0\| + T\|C\|^2\right)$, and choose $\rho > 0$ and $\tau > 0$ such that

$$\rho = 2M_T^2\beta, \quad \tau\left(\|C\|^2 + \rho^2\|B\|^2\right) \le \beta, \ \ 2\rho\tau M_T^2\|B\|^2 \le \frac{1}{2}.$$

By Lemma 1.2.3 there exists a unique solution P (resp. P_n) of (1.2.1) (resp. (1.2.5)) in $[0,\tau]$ and $P_n \to P$ in $C_s([0,\tau];\Sigma(H))$. We now prove that

$$P_n(t) \ge 0, \ \ \forall\, t \in [0,\tau]. \tag{1.2.9}$$

This will imply

$$P(t) \ge 0, \ \ \forall\, t \in [0,\tau]. \tag{1.2.10}$$

To this end we notice that P_n is the solution of the following linear problem in $[0,\tau]$

$$P_n' = L_n^*P_n + P_nL_n + C^*C, \ \ P_n(0) = P_0,$$

where $L_n = A_n - \frac{1}{2}BB^*P_n$. Denote by $U_n(t,s),\ 0 \le s \le t \le \tau$, the evolution operator associated to L_n^*, that is the solution to

$$D_t U_n(t,s) = L_n^*(t)U_n(t,s),\ U_n(s,s) = I,\ \ 0 \le s \le t \le \tau.$$

Then we can write the solution $P_n(t)$ as

$$P_n(t) = U_n(t,0)P_0U_n^*(t,0) + \int_0^t U_n(t,s)C^*CU_n^*(t,s)ds.$$

Thus (1.2.9) and (1.2.10) follow immediately.

Note that, arguing as in Lemma 1.2.3, we have

$$\|P(t)\| \le \rho\, I = 2M_T^2\beta\, I$$

We now prove that we have a better estimate

$$P(t) \le \beta\, I,\ \ \forall\, t \in [0,\tau]. \tag{1.2.11}$$

This inequality will allow us to repeat the previous argument in the interval $[\tau, 2\tau]$ and so on. In this way the theorem will be proved. We have in fact

$$\langle P(t)x, x\rangle = \langle P_0 e^{tA}x, e^{tA}x\rangle + \int_0^t |Ce^{sA}x|^2 ds$$

$$-\int_0^t |B^*P(s)e^{(t-s)A}x|^2 ds \le \beta|x|^2.$$

Since $P(t) \ge 0$ this implies (1.2.11). The proof is complete. □

We now prove continuous dependence with respect to data. Consider a sequence of Riccati equations

$$\begin{cases} (P^k)' = (A^k)^*P^k + P^kA^k - P^kB^k(B^k)^*P^k + (C^k)^*C^k, \\ \\ P^k(0) = P_0^k, \end{cases} \tag{1.2.12}$$

under the following assumption.

Hypothesis 1.2 *(i) For any $k \in \mathbb{N}$, (A^k, B^k, C^k, P_0^k) fulfil Hypothesis 1.1.*

(ii) For all $T > 0$ and all $x \in H$,

$$\lim_{k\to\infty} e^{tA^k}x = e^{tA}x,\ \textit{uniformly in}\ [0,T].$$

(iii) The sequences $\{B^k\}, \{(B^k)^*\}, \{C^k\}, \{(C^k)^*\}, \{P_0^k\}$ *are strongly convergent to* B, B^*, C, C^*, P_0, *respectively.*

Theorem 1.2.6 *Assume that Hypotheses 1.1 and 1.2 hold. Let* P *(resp.* P^k*) be the mild solution to* (1.2.1) *(resp.* (1.2.12)*). Then, for any* $T > 0$ *we have*

$$\lim_{n\to\infty} P^k = P \text{ in } C_s([0,T];\Sigma_+(H)).$$

Proof. Fix $T > 0$. By the Uniform Boundedness Theorem there exists positive numbers p, b and c such that

$$\|P_0^k\| \le p, \ \ \|(C^k)^*C^k\| \le c, \ \|B^k(B^k)^*\| \le p, \ \forall\, k \in \mathbb{N}.$$

Set $\beta = M_T^2(p + cT)$ and choose ρ and $\tau \in [0,T]$ such that

$$\rho = 2\beta M_T^2, \ \ \tau(c + \rho^2 b) \le \beta, \ \ 2\tau M_T^2\|B\|^2 \le \frac{1}{2}.$$

Then, arguing as we did in the proof of Lemma 1.2.3, we can show that $P^k(\cdot)x \to P(\cdot)x$ for any $x \in H$. Finally, proceeding as in the proof of Theorem 1.2.5, we prove that this argument can be iterated in the interval $[\tau, 2\tau]$ and so on. □

We conclude this section by proving an important monotonicity property of the solutions of the Riccati equation (1.2.1).

Proposition 1.2.7 *Consider the Riccati equations:*

$$\begin{cases} P_i' = A^*P_i + P_iA - P_iB_iB_i^*P_i + C_i^*C_i, \\ \\ P_i(0) = P_{i,0}, \ i = 1,2. \end{cases} \tag{1.2.13}$$

Assume that $(A, B_i, C_i, P_{i,0})$ *verify Hypothesis 1.1, and, in addition, that*

$$P_{1,0} \le P_{2,0}, \ \ C_1^*C_1 \le C_2^*C_2, \ \ B_2B_2^* \le B_1B_1^*.$$

Then we have

$$P_1(t) \le P_2(t), \ \forall\, t \ge 0. \tag{1.2.14}$$

Proof. Due to Theorem 1.2.5 it is sufficient to prove (1.2.14) when A is bounded. Set $Z = P_2 - P_1$, then, as easily checked, Z is the solution to the linear problem

$$\begin{cases} Z' = X^*Z + ZX - P_2[B_2B_2^* - B_1B_1^*]P_2 + C_2^*C_2 - C_1^*C_1, \\ \\ Z(0) = P_{2,0} - P_{1,0}, \end{cases} \tag{1.2.15}$$

where

$$X = A - \frac{1}{2}\, B_1B_1^*(P_1 + P_2).$$

Let $V(t,s)$ be the evolution operator associated with X^*, that is the solution to the problem

$$D_tV(t,s) = X(t)^*V(t,s),\ V(s,s) = I,\ \ 0 \le s \le t \le \tau.$$

Then we have

$$\begin{aligned} Z(t) &= V(t,0)(P_{2,0} - P_{1,0})V^*(t,0) \\ &+ \int_0^t V(t,s)[C_2^*C_2 - C_1^*C_1]V^*(t,s)ds \\ &+ \int_0^t V(t,s)P_1(s)[B_1B_1^* - B_2B_2^*]P_1(s)V^*(t,s)ds, \end{aligned}$$

so that $Z(t) \ge 0$ and the conclusion follows. □

1.3 Solution of the control problem

In this section we consider the control problem (1.1.1)–(1.1.2). We assume that Hypothesis 1.1 is fulfilled and we denote by $P \in C_s([0,T];\Sigma^+(H))$ the mild solution of the Riccati equation (1.2.1). We first consider the closed loop equation

$$\begin{cases} y'(t) = Ay(t) - BB^*P(T-t)y(t),\ t \in [0,T], \\ \\ y(0) = x \in H. \end{cases} \tag{1.3.1}$$

We say that $y \in C([0,T];H)$ is a mild solution of equation (1.3.1) if it is a solution of the following integral equation

$$y(t) = e^{tA}x - \int_0^t e^{(t-s)A}BB^*P(T-s)y(s)ds.$$

Proposition 1.3.1 *Assume that Hypothesis 1.1 is fulfilled and let* $x \in H$. *Then equation* (1.3.1) *has a unique mild solution* $y \in C([0,T];H)$.

Proof. It follows by using standard successive approximations. □

We now prove a basic identity.

Proposition 1.3.2 *Assume that Hypothesis 1.1 is fulfilled and let* $u \in L^2(0,T,U)$ $x \in H$. *Let* y *be the solution of the state equation* (1.1.1) *and let* P *be the mild solution of the Riccati equation* (1.2.1). *Then the following identity holds*

$$J(u) = \int_0^T |u(s) + B^*P(T-s)y(s)|^2 ds + \langle P(T)x, x\rangle. \tag{1.3.2}$$

Proof. Let P_n be the mild solution of the approximated Riccati equation (1.2.5), and let y_n be the solution of the problem

$$\begin{cases} y_n'(t) = A_n y(t) + Bu(t), \ t \in [0,T], \\ \\ y(0) = x \in H. \end{cases}$$

Now, by computing the derivative

$$\frac{d}{ds}\,\langle P_n(T-s)y_n(s), y_n(s)\rangle$$

and completing the squares, we obtain the identity

$$\frac{d}{ds}\,\langle P_n(T-s)y_n(s), y_n(s)\rangle$$

$$= |u_n(s) + B^*P_n(T-s)y_n(s)|^2 - |Cy_n(s)|^2 - |u(s)|^2.$$

Integrating from 0 to T and letting n tend to infinity we obtain (1.3.2). □

We are now ready to prove the following result.

Theorem 1.3.3 *Assume that Hypothesis 1.1 is fulfilled and let* $x \in H$. *Then there exists a unique optimal pair* (u^*, y^*). *Moreover*

(i) $y^* \in C([0,T];H)$ *is the mild solution to the closed loop equation* (1.3.1).

(ii) $u^* \in C([0,T];U)$ *is given by the feedback formula*

$$u^*(t) = -B^*P(T-t)y^*(t), \ t \in [0,T]. \tag{1.3.3}$$

(iii) The optimal cost $J(u^*)$ is given by

$$J(u^*) = \langle P(T)x, x\rangle. \tag{1.3.4}$$

Proof. We first remark that by identity (1.3.2) it follows that

$$J(u^*) \geq \langle P(T)x, x\rangle, \tag{1.3.5}$$

for any control $u \in C([0,T];U)$. Let now y^* be the mild solution to (1.3.1) and let u^* be given by (1.3.3). Setting in (1.3.2) $u = u^*$ and taking into account (1.3.5) it follows that (u^*, y^*) is an optimal pair and that (1.3.4) holds.

It remains to prove uniqueness. Let $(\overline{u}, \overline{y})$ be another optimal pair. Setting in (1.3.2) $u = \overline{u}$ and $y = \overline{y}$ we obtain

$$\int_0^T |\overline{u}(s) + B^*P(T-s)\overline{y}(s)|^2 ds = 0,$$

so that $\overline{u}(s) = -B^*P(T-s)\overline{y}(s)$ for almost every $s \in [0,T]$. But this implies that $\overline{y}$ is a mild solution of (1.3.1) so that $\overline{y} = y^*$ and consequently $\overline{u} = u^*$. □

2 Control in infinite horizon

2.1 Introduction and setting of the problem

As in Section 1 we are concerned with a dynamical system governed by the following state equation

$$\begin{cases} y'(t) = Ay(t) + Bu(t), \ t \geq 0, \\ y(0) = x \in H. \end{cases} \tag{2.1.1}$$

We shall assume that

Hypothesis 2.1 *(i) A generates a strongly continuous semigroup e^{tA} on H.*

(ii) $B \in L(U,H)$.

(iii) $C \in L(H,Y)$.

We want to minimize the cost function

$$J_\infty(u) = \int_0^{+\infty} \left[|Cy(s)|^2 + |u(s)|^2\right] ds, \tag{2.1.2}$$

over all controls $u \in L^2(0,+\infty,U)$ subject to (1.1.1).

We say that the control $u \in L^2(0,+\infty;U)$ is *admissible* if $J_\infty(u) < +\infty$.

An admissible control $u^* \in L^2(0,+\infty;U)$ is called an *optimal control* if

$$J_\infty(u^*) \le J_\infty(u),\ \forall\, u \in L^2(0,+\infty;U).$$

In this case the corresponding solution y^* of (2.1.1) is called an *optimal state* and the pair (u^*,y^*) an *optimal pair*.

An admissible controls can fail to exist, as the following simple example shows.

Example 2.1.1 Let $H = U = Y = \mathbb{R}$, $B = 0, A = C = 1$. Then for any $u \in L^2(0,+\infty;U)$ we have $y(t) = e^t x$ and if $x \neq 0$

$$J_\infty(u) = \int_0^{+\infty} \left[|e^s y(s)|^2 + |u(s)|^2\right] ds = +\infty.$$

If for any $x \in H$ an admissible control exists, we say that (A,B) is *stabilizable with respect to the observation operator* C or, for brevity, that (A,B) is C–stabilizable. In this case is still possible to solve problem (2.1.1)–(2.1.2) following the following steps,

Step 1. We show that the *minimal nonnegative solution* $P_{min}(t)$ to the Riccati equation

$$P' = A^*P + PA - PBB^*P + C^*C,$$

that is the solution to (1.2.1) corresponding to $P_0 = 0$, converges, as $t \to +\infty$ to a solution P^∞_{min} to the *algebraic Riccati equation*:

$$A^*X + XA - XBB^*X + C^*C = 0 \tag{2.1.3}$$

Step 2. We show that the optimal control u^* is given by the feedback formula

$$u^*(t) = -B^*P^\infty_{min}y^*(t),\quad t \ge 0, \tag{2.1.4}$$

where y^* is the mild solution of the *closed loop equation*

$$\begin{cases} y'(t) = [A - BB^*P_{min}^{\infty}]y(t), \ t \geq 0, \\ \\ y(0) = x \in H. \end{cases} \tag{2.1.5}$$

Example 2.1.2 (i). Assume that A is of negative type. Then (A, B) is C–stabilizable since the control $u(t) = 0$ is clearly admissible.

(ii). Assume that $B = I$. Then (A, B) is C–stabilizable. In fact let M, ω be such that $\|e^{tA}\| \leq Me^{\omega t}, \ t \geq 0$. Choose $u(t) = -(\omega + 1)e^{t(A-\omega-1)}, \ t \geq 0$. Then $y(t) = e^{t(A-\omega-1)}, \ t \geq 0$ so that $J_{\infty}(u) < +\infty$.

(iii). Assume that there is $\alpha > 0, \beta > 0, K > 0$ such that

$$\|e^{t(A-2\alpha BB^*)}\| \leq Ke^{-\beta t}, \ t \geq 0. \tag{2.1.6}$$

Then (A, B) is C–stabilizable. In fact setting $u(t) = -2\alpha B^* e^{t(A-2\alpha BB^*)}, \ t \geq 0$, one has $y(t) = e^{t(A-2\alpha BB^*)}, \ t \geq 0$, and so $J_{\infty}(u) < +\infty$.

2.2 The Algebraic Riccati Equation

We assume here that Hypothesis 2.1 holds and consider the system (2.1.1).

We consider the Riccati equation

$$P' = A^*P + PA - PBB^*P + C^*C, \tag{2.2.1}$$

and the corresponding stationary equation

$$A^*X + XA - XBB^*X + C^*C = 0. \tag{2.2.2}$$

In the sequel we shall consider only nonnegative solutions of (2.2.1) and (2.2.2).

Definition 2.2.1 *We say that $X \in \Sigma^+(H)$ is a weak solution of* (2.2.2) *if*

$$\langle Xx, Ay\rangle + \langle Ax, Xy\rangle - \langle B^*Xx, B^*Xy\rangle + \langle Cx, Cy\rangle = 0 \tag{2.2.3}$$

for all $x, y \in D(A)$.

Definition 2.2.2 *We say that $X \in \Sigma^+(H)$ is a stationary solution of* (2.2.1) *if it coincides with the mild solution of* (2.2.1) *with initial condition $P(0) = X$.*

Recalling Proposition 1.2.2 the following results follows immediately.

Proposition 2.2.3 *Let $X \in \Sigma^+(H)$, then the following statements are equivalent*

(i) X is a weak solution of (2.2.2).

(ii) X is a stationary solution of (2.2.1).

We are going to study existence of a solution of the Algebraic Riccati equation. To this purpose it is useful to consider the solution of the Riccati equation (2.2.1) with initial condition 0. This solution will be denoted by P_{min}. It is the minimal nonnegative solution of (2.2.1). In fact if $P_0 \in \Sigma^+(H)$ and P is the mild solution of (2.2.1) such that $P(0) = P_0$, then by Proposition 1.2.7 we have

$$P_{min}(t) \leq P(t), \quad \forall\, t \geq 0.$$

In particular if X is a solution of (2.2.2), then

$$P_{min}(t) \leq X, \quad \forall\, t \geq 0.$$

We now prove the following properties of P_{min}.

Proposition 2.2.4 *(i) For any $x \in H$, $\langle P_{min}(\cdot)x, x\rangle$ is non decreasing.*

(ii) Assume that for some $R \in \Sigma^+(H)$, we have

$$P_{min}(t) \leq R, \quad \forall\, t \geq 0.$$

Then for all $x \in H$ the limit

$$P^{\infty}_{min}x = \lim_{t\to+\infty} P_{min}(t)x, \tag{2.2.4}$$

exists, and P^{∞}_{min} is a solution of (2.2.2).

In other words there exists a nonnegative solution of (2.2.2) if and only if P_{min} is bounded.

Proof. Let $\varepsilon > 0$, $t \geq 0$ and let P be the solution of (2.2.1) such that $P(0) = P_{min}(\varepsilon)$. By Proposition 1.2.7 we have

$$P(t) = P_{min}(t+\varepsilon) = P(t) \geq P_{min}(t),$$

and (i) is proved. Assume now $P_{min}(t) \leq R$; since $P_{min}(t)$ is nondecreasing and bounded we can set

$$\gamma(x) = \lim_{t \to +\infty} \langle P_{min}(t)x, x \rangle, \ \forall\, x \in H.$$

For $x, y \in H$ we have

$$2\langle P_{min}(t)x, y \rangle$$

$$= \langle P_{min}(t)(x+y), (x+y) \rangle - \langle P_{min}(t)x, x \rangle - \langle P_{min}(t)y, y \rangle.$$

So the limit

$$\Gamma(x, y) = \lim_{t \to +\infty} \langle P_{min}(t)x, y \rangle, \ \ \forall\, x, y \in H,$$

exists and the following operator $P^{\infty}_{min} \in \Sigma^{+}(H)$ can be defined

$$\lim_{t \to +\infty} \langle P_{min}(t)x, y \rangle = \langle P^{\infty}_{min}x, y \rangle, \ \ \forall\, x, y \in H.$$

It follows that

$$\lim_{t \to +\infty} \langle [P^{\infty}_{min} - P_{min}(t)]x, x \rangle = 0, \ \ \forall\, x \in H,$$

which is equivalent to

$$\lim_{h \to +\infty} [P^{\infty}_{min} - P_{min}(t)]^{1/2} x = 0, \ \ \forall\, x \in H$$

This implies that

$$\lim_{t \to +\infty} [P^{\infty}_{min} - P_{min}(t)]x = 0, \ \ \forall\, x \in H$$

so that (2.2.4) holds.

It remains to show that P^{∞}_{min} is a solution of (2.2.2). For this we denote by P_h the solution of (2.2.1) for which $P_h(0) = P_{min}(h)$, i.e. $P_h(t) = P_{min}(h+t)$. Since

$$\lim_{h \to +\infty} P_{min}(h)x = P^{\infty}_{min}x, \ \forall\, x \in H,$$

by Theorem 1.2.6, we have

$$\lim_{h \to +\infty} P_h(\cdot)x = P^{\infty}_{min}x \text{ in } C([0,T]; H), \ \forall\, x \in H, \ T > 0.$$

Moreover P^{∞}_{min} is a solution of (2.2.1) (hence stationary). $\square$

Remark 2.2.5 Assume that there exists a solution $X \in \Sigma^+(H)$ of (2.2.2). Then by Proposition 2.2.4 the solution P_{min}^{∞} defined by (2.2.4) exists. By the above proposition it follows that

$$P_{min}^{\infty} \leq X,$$

for all solutions $X \in \Sigma^+(H)$ of (2.2.2). Thus P_{min}^{∞} is the minimal solution of the algebraic Riccati equation (2.2.2).

We now prove that if (A, B) is C–stabilizable, then a nonnegative solution of the algebraic Riccati equation exists.

Proposition 2.2.6 *Assume that Hypothesis 2.1 is fulfilled and that (A, B) is C–stabilizable. Then there exists a minimal solution P_{min}^{∞} of equation (2.2.2).*

Proof. We first recall that by the basic identity (1.3.2) we have

$$\begin{aligned} &\langle P_{min}(t)x, x\rangle + \int_0^t |u(s) + B^* P_{min}(t-s)y(s)|^2 ds \\ &= \int_0^t [|Cy(s)|^2 + |u(s)|^2] ds, \end{aligned} \tag{2.2.5}$$

for any $x \in H$ and any $u \in L^2(0, +\infty; U)$, where y is the solution to (2.1.1). Let $x \in H$ and let u_x be a control in $L^2(0, +\infty; U)$ such that the corresponding solution of (2.1.1) is such that $Cy \in L^2(0, +\infty; Y)$. By (2.2.5) it follows that

$$\sup_{t \geq 0} \langle P_{min}(t)x, x\rangle \leq \int_0^{+\infty} [|Cy(s)|^2 + |u(s)|^2] ds = J_{\infty}(u_x) < +\infty$$

for any $x \in H$. By the Uniform Boundedness Theorem it follows that $P_{min}(t)$ is bounded above, so that by Proposition 2.2.4 there exists a solution of equation (2.2.2). □

2.3 Solution of the control problem

We now consider the control problem (2.1.1)–(2.1.2) and prove the following result.

Theorem 2.3.1 *Assume that Hypothesis 2.1 is fulfilled, that (A, B) is C–stabilizable, and let $x \in H$. Then there exists a unique optimal pair (u^*, y^*). Moreover*

(i) $y^* \in C([0,+\infty);H)$ *is the mild solution to the closed loop equation* (2.1.5).

(ii) $u^* \in C([0,+\infty);U)$ *is given by the feedback formula*

$$u^*(t) = -B^*P^\infty_{min}y^*(t),\ t \geq 0. \tag{2.3.1}$$

(iii) The optimal cost $J_\infty(u^*)$ *is given by*

$$J_\infty(u^*) = \langle P^\infty_{min}x, x\rangle. \tag{2.3.2}$$

Proof. Let $u \in L^2([0,+\infty);U)$ and let y be the corresponding solution of the state equation (2.1.1). By the identity (2.2.5) we have

$$\langle P_{min}(t)x, x\rangle \leq \int_0^t [|Cy(s)|^2 + |u(s)|^2]ds \leq J_\infty(u).$$

It follows that

$$J_\infty(u) \geq \langle P_{min}(t)x, x\rangle,\ \forall\, u \in L^2([0,+\infty);U),\ t \geq 0,$$

and so

$$J_\infty(u) \geq \langle P^\infty_{min}x, x\rangle,\ \forall\, u \in L^2([0,+\infty);U). \tag{2.3.3}$$

Let now y_∞ be the solution of the problem,

$$\begin{cases} y'_\infty(s) = Ay_\infty(s) - BB^*P^\infty_{min}y_\infty(s),\ s \geq 0, \\ \\ y_\infty(0) = x, \end{cases}$$

and set

$$u_\infty(s) = -B^*P^\infty_{min}y_\infty(s),\ s \geq 0. \tag{2.3.4}$$

We are going to show that (u_∞, y_∞) is an optimal pair.

For this purpose it is useful to introduce, for any $t > 0$, the following auxiliary optimal control problem over the finite time horizon $[0,t]$:

to minimize

$$J_t(u) = \int_0^t [|Cy(s)|^2 + |u(s)|^2]ds \tag{2.3.5}$$

over all controls $u \in L^2(0,t;U)$ subject to (2.1.1). By Theorem 1.3.3 we know that there exists a unique optimal pair (u_t, y_t) for problem (2.3.5), where y_t is the mild solution to the closed loop equation

$$\begin{cases} y_t'(s) = Ay_t(s) - BB^*P_{min}(t-s)y_t(s), \ s \in [0,t], \\ \\ y_t(0) = x, \end{cases}$$

and u_t is given by the feedback formula

$$u_t(s) = -B^*P_{min}(t-s)y_t(s), \ s \in [0,t].$$

Moreover the optimal cost is given by

$$\langle P_{min}(t)x, x\rangle = \int_0^t [|Cy_t(s)|^2 + |u_t(s)|^2]ds. \tag{2.3.6}$$

Now we proceed into two steps.

Step 1. We have

$$\lim_{t\to+\infty} y_t(s) = y_\infty(s), \ s \geq 0. \tag{2.3.7}$$

$$\lim_{t\to+\infty} u_t(s) = u_\infty(s), \ s \geq 0. \tag{2.3.8}$$

In fact fix $T > t$ and set $z_t = y_t - y_\infty$; then z_t is the mild solution to the problem:

$$\begin{cases} z_t'(s) &= [A - BB^*P_{min}(t-s)]z_t(s) \\ \\ &+ BB^*[P_{min}(t-s) - P_{min}^\infty]y_\infty(s) \\ \\ z_t(0) &= 0. \end{cases} \tag{2.3.9}$$

Denote by $U(r,s)$ the evolution operator corresponding to $A - BB^*P_{min}(t-\cdot)$ then for $x \in H$

$$\begin{cases} U(r,\sigma)x = e^{(r-\sigma)A}x - \displaystyle\int_\sigma^r e^{(r-\rho)A}BB^*P_{min}(t-\rho)U(\rho,\sigma)xd\rho, \\ \\ U(\sigma,\sigma) = I. \end{cases}$$

It follows that

$$\|U(r,\sigma)\| \leq Me^{(r-\sigma)\omega} + M\|B\|^2\|P_{min}^{\infty}\| \int_{\sigma}^{r} e^{(r-\rho)\omega}\|U(\rho,\sigma)\|d\rho.$$

By Gronwall's Lemma we have

$$\|U(r,\sigma)\| \leq Me^{(r-\sigma)[\omega+M\|B\|^2\|P_{min}^{\infty}\|]}, 0 \leq \sigma \leq r \leq T. \tag{2.3.10}$$

We now return to problem (2.3.9) which we write in the form

$$z_t(s) = \int_0^s U(s,\sigma)BB^*[P_{min}(t-\sigma) - P_{min}^{\infty}]y_{\infty}(\sigma)d\sigma.$$

By (2.3.10) and the dominate convergence theorem we obtain $z_t(s) \to 0$ as $t \to +\infty$. So (2.3.7) and then (2.3.8) follow.

Step 2. Conclusion.

By (2.3.6) we have for $t \geq T$

$$\langle P_{min}^{\infty}x, x\rangle \geq \int_0^T [|Cy_t(s)|^2 + |u_t(s)|^2]ds \tag{2.3.11}$$

and, as $t \to +\infty$,

$$\langle P_{min}^{\infty}x, x\rangle \geq \int_0^T [|Cy_{\infty}(s)|^2 + |u_{\infty}(s)|^2]ds. \tag{2.3.12}$$

But, since T is arbitrary, we find

$$\langle P_{min}^{\infty}x, x\rangle \geq \int_0^{+\infty} [|Cy_{\infty}(s)|^2 + |u_{\infty}(s)|^2]ds, \tag{2.3.13}$$

and thus

$$\langle P_{min}^{\infty}x, x\rangle \geq J_{\infty}(u_{\infty}), \tag{2.3.14}$$

so that u_{∞} is optimal. Moreover formula (2.3.1) with $u^* = u_{\infty}$, $y^* = y_{\infty}$ follows from (2.2.5).

It remains to to show uniqueness. Let $(\hat{u}, \hat{y})$ be another optimal pair, then $J_{\infty}(\hat{u}) = \langle P_{min}^{\infty}x, x\rangle$. Fix $T > 0$. By applying (2.2.5) with $t \geq T$ we obtain

$$\int_0^T |\hat{u}(s) + B^*P_{min}(t-s)\hat{y}(s)|^2 ds \leq J_{\infty}(\hat{u}) - \langle P_{min}(t)x, x\rangle$$

$$\leq \langle [P_{min}^{\infty} - P_{min}(t)x, x\rangle.$$

As $t \to +\infty$ we have

$$\int_0^T |\hat{u}(s) + B^* P_{min}^{\infty} \hat{y}(s)|^2 ds = 0,$$

that yields $\hat{u}(s) = -B^* P_{min}^{\infty} \hat{y}(s)$. Consequently $\hat{y} = y^*$ and $\hat{u} = u^*$. □

3 Examples and generalizations

3.1 Parabolic equations

We consider here Example 1.1.1. Let D be an open subset of $\mathbb{R}^n$ with regular boundary ∂D. Consider the state equation

$$\begin{cases} D_t y(t,\xi) = (\Delta_\xi + c) y(t,\xi) + u(t,\xi), \text{ in } (0,T] \times D, \\ \\ y(t,\xi) = 0, \text{ on } (0,T] \times \partial D, \\ \\ y(0,\xi) = x(\xi), \text{ in } D, \end{cases} \tag{3.1.1}$$

where Δ_ξ represents the Laplace operator:

$$\Delta_\xi x = \sum_{k=1}^{n} D_{\xi_k}^2 x,$$

and c is a real constant.

Let $H = U = Y = L^2(D)$, $B = C = P_0 = I$ and define the linear operator by A in H :

$$\begin{cases} Ay = (\Delta_\xi + c) y \\ \\ D(A) = H^2(D) \cap H_0^1(D). \end{cases} \tag{3.1.2}$$

Since A is self–adjoint, it is the infinitesimal generator of a strongly continuous semigroup on $H = L^2(D)$ ([3]). Moreover ([4]) there exists a complete orthonormal system $\{e_k\}$ in $L^2(D)$ and a sequence $\{\lambda_k\}$ of positive numbers such that $\{\lambda_k\} \uparrow +\infty$ with

$$Ae_k = -\lambda_k e_k, \quad k \in \mathbb{N}. \tag{3.1.3}$$

[3] See e.g. J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod, (1968).

[4] See e.g. S. A. Agmon, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, (1965).

Setting $y(t) = y(t,\cdot)$, $u(t) = u(t,\cdot)$, we write (3.1.1) in the abstract form (1.1.1).

We want to minimize the cost

$$J(u) = \int_0^T \int_D [|y(t,\xi)|^2 + |u(t,\xi)|^2] dt d\xi + \int_D |y(T,\xi)|^2 d\xi, \tag{3.1.4}$$

that we shall write on the form:

$$J(u) = \int_0^T \left[|y(s)|^2 + |u(s)|^2\right] ds + |y(T)|^2.$$

By Theorem 1.3.3 there exists a unique optimal pair (u^*, y^*) where y^* is the solution of the closed loop equation

$$\begin{cases} D_t y(t,\xi) = (\Delta_\xi + c) y(t,\xi) - P(T-t) y(t,\cdot)(\xi), \text{ in } (0,T] \times D, \\ y(t,\xi) = 0, \text{ on } (0,T] \times \partial D, \\ y(0,\xi) = x(\xi), \text{ in } D. \end{cases} \tag{3.1.5}$$

Moreover u^* is given by

$$u^*(t,\xi) = -P(T-t) y(t,\cdot)(\xi),$$

and the Riccati equation reeds as follows

$$P' = 2AP - P^2 + I, \quad P(0) = I. \tag{3.1.6}$$

For any $t \geq 0$ we can find explicitly $P(t)$ as

$$P(t) e_k = p_k(t) e_k, \quad k \in \mathbb{N},$$

where p_k is the solution to the ordinary differential equation

$$p_k'(t) = -2\lambda_k p_k(t) - p_k^2(t) + 1, \quad p_k(0) = 1.$$

Let us consider now the infinite horizon problem, $T = +\infty$. We want to minimze the cost

$$J_\infty(u) = \int_0^{+\infty} \int_D [|y(t,\xi)|^2 + |u(t,\xi)|^2] dt d\xi + \int_D |y(T,\xi)|^2 d\xi. \tag{3.1.7}$$

By Example 2.1.2–(ii) (A, I) is I–stabilizable, and consequently by Theorem 2.3.1 there exists a unique optimal pair (u^*, y^*) where y^* is the solution of the closed loop equation

$$\begin{cases} D_t y(t,\xi) = (\Delta_\xi + c)y(t,\xi) - P_\infty y(t,\cdot)(\xi), \text{ in } (0,+\infty)\times D, \\ y(t,\xi) = 0, \text{ on } (0,+\infty)\times \partial D, \\ y(0,\xi) = x(\xi), \text{ in } D. \end{cases} \tag{3.1.8}$$

Moreover u^* is given by

$$u^*(t,\xi) = -P_\infty y(t,\cdot)(\xi),$$

and the Algebraic Riccati equation reeds as follows

$$2AP - P^2 + I = 0 \tag{3.1.9}$$

Consequently

$$P^\infty_{min} = \sqrt{A^2 + I} + A$$

and

$$P^\infty_{min} e_k = (\sqrt{\lambda_k^2 + 1} - \lambda_k)e_k, \ k \in \mathbb{N}.$$

Notice that

$$\sup_{k\in\mathbb{N}}\{\sqrt{\lambda_k^2 + 1} - \lambda_k\} < +\infty,$$

so that P^∞_{min} is bounded.

3.2 Wave equation

Let D be an open subset of $\mathbb{R}^n$ with regular boundary ∂D. Consider the state equation

$$\begin{cases} D_t^2 y(t,\xi) = \Delta_\xi y(t,\xi) + u(t,\xi), \text{ in } (0,T]\times D, \\ y(t,\xi) = 0, \text{ on } (0,T]\times \partial D, \\ y(0,\xi) = x_0(\xi), \ D_t y(0,\xi) = x_1(\xi), \text{ in } D. \end{cases} \tag{3.2.1}$$

We want to minimze the cost

$$J(u) = \int_0^T \int_D [|\nabla_\xi y(t,\xi)|^2 + |y_t(t,\xi)|^2 + |u(t,\xi)|^2] dt d\xi$$
$$+ \int_D [|\nabla_\xi y(T,\xi)|^2 + |y_t(T,\xi)|^2] d\xi. \tag{3.2.2}$$

Setting $y(t) = y(t,\cdot),\ u(t) = u(t,\cdot)$, we write (3.2.1) as

$$\begin{cases} y''(t) = Ay(t) + u(t) \\ \\ y(0) = x_0,\ y'(0) = x_1, \end{cases} \tag{3.2.3}$$

where A is defined by (3.1.2). Now, setting $y'(t) = z(t)$, $Y(t) = \begin{pmatrix} y(t) \\ z(t) \end{pmatrix}$, and $X = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix}$, we reduce the problem to a first order problem

$$\begin{cases} Y'(t) = \mathcal{A}Y(t) + \mathcal{B}u(t) \\ \\ Y(0) = X, \end{cases} \tag{3.2.4}$$

where

$$\mathcal{A} = \begin{pmatrix} 0 & 1 \\ A & 0 \end{pmatrix},$$

and

$$\mathcal{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Now we choose $H = Y = H_0^1(D) \oplus L^2(D)$, $U = L^2(D)$ and

$$\mathcal{C} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus

$$D(\mathcal{A}) = (H^2(D) \oplus H_0^1(D)) \oplus H_0^1(D),$$

and $\mathcal{A}$ generates a strongly continuous semigroup of contractions on H ([5]) given by:

$$e^{t\mathcal{A}} = \begin{pmatrix} \cos(\sqrt{-A}\, t) & \frac{1}{\sqrt{-A}} \sin(\sqrt{-A}\, t) \\ \\ -\sqrt{-A}\, \sin(\sqrt{-A}\, t) & \cos(\sqrt{-A}\, t). \end{pmatrix} \tag{3.2.5}$$

[5] See e.g. J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod, (1968).

By $\cos(\sqrt{-A}\,t)$ and $\sin(\sqrt{-A}\,t))$ we mean the linear operators

$$\cos(\sqrt{-A}\,t)e_k = \cos(t\sqrt{\lambda_k})e_k,\ \sin(\sqrt{-A}\,t)e_k = \sin(t\sqrt{\lambda_k})e_k, k \in \mathbb{N}.$$

Finally the cost can be written as

$$J(u) = \int_0^T \int_D [Y(t)|_H^2 + |u(t)|^2_{L^2(D)}]dt + |Y(T)|_H^2. \tag{3.2.6}$$

By Theorem 1.3.3 there exists a unique optimal pair (u^*, y^*).

Finally we can show that $(\mathcal{A}, \mathcal{B})$ is $\mathcal{C}$–stabilizable. For this we shall fulfill the conditions of Example 2.1.2–(iii) by proving that for all $\alpha < \sqrt{\lambda_0}$ the condition (2.1.6) holds. We have in fact, by a direct computation

$$e^{t(A-2\alpha BB^*)} = e^{-t\alpha}\begin{pmatrix} \cos(Et) + \frac{\alpha}{E}\sin(Et) & \frac{1}{E}\sin(Et) \\ \frac{\alpha^2+E^2}{E}\sin(Et) & -\frac{\alpha}{E}\sin(Et) + \cos(Et) \end{pmatrix}, \tag{3.2.7}$$

where $E = \sqrt{-A - \alpha^2}I$, so that (2.1.6) is fulfilled.

3.3 Boundary control problems

Let us consider the following state equation

$$\begin{cases} D_t y(t,\xi) = D_\xi^2 y(t,\xi), \text{ in } (0,T] \times [0,1], \\ y(t,0) = u_0(t),\ y(t,1) = u_1(t), \text{ on } (0,T], \\ y(0,\xi) = x(\xi), \text{ in } D. \end{cases} \tag{3.3.1}$$

Here the control is given on the boundary of D.

We want to minimize the cost

$$J(u) = \int_0^T \int_0^1 |y(t,\xi)|^2 dt\, d\xi + \int_0^T [|u_0(t)|^2 + |u_1(t)|^2]dt + \int_0^1 |y(T,\xi)|^2 d\xi. \tag{3.3.2}$$

In order to reduce this problem to the standard form (1.1.1), it is convenient to introduce the *Dirichlet mapping*

$$\delta : \mathbb{R}^2 \to L^2(0,1),\ (\alpha_0, \alpha_1) \to \delta(\alpha_0, \alpha_1),$$

where

$$\delta(\alpha_0, \alpha_1)(\xi) = (\alpha_1 - \alpha_0)\xi + \alpha_0.$$

Notice that $\delta(\alpha_0, \alpha_1)$ is the unique harmonic function on $[0, 1]$ that holds α_0 at $\{0\}$ and α_1 at $\{1\}$.

Let us now proceed formally by setting

$$z(t, \xi) = y(t, \xi) - \delta(u_0(t), u_1(t))$$

$$= y(t, \xi) - (u_1(t) - u_0(t))\xi - u_0(t), \ \xi \in [0, 1], t \geq 0,$$

so that $z(t, 0) = z(t, 1) = 0$. Then

$$D_t z(t, \xi) = D_t y(t, \xi) - \delta u'(t),$$

where $u(t) = (u_0(t), u_1(t))$, and we can write problem (3.3.1) as

$$\begin{cases} D_t z(t, \xi) = D_\xi^2 z(t, \xi) - \delta u'(t), \text{ in } (0, T] \times [0, 1], \\ z(t, 0) = z(t, 1) = 0, \text{ on } (0, T], \\ y(0, \xi) = x(\xi), \text{ in } D. \end{cases} \tag{3.3.3}$$

Now this problem can be written in the abstract form

$$\begin{cases} z'(t) = Az(t) - \delta u'(t), \\ z(0) = z(t, 1) = x - \delta u(0), \end{cases}$$

where A denotes the operator (3.1.2) (with $D = [0, 1]$). Using the variation of constants formula we find

$$z(t) = e^{tA}(x - \delta u(0)) - \int_0^t e^{(t-s)A} \delta u'(s) ds,$$

and, integrating by parts, we find (always formally),

$$y(t) = e^{tA} x - \int_0^t A e^{(t-s)A} \delta u(s) ds. \tag{3.3.4}$$

We show now that this formula is meaningful.

Notice that for any $s \in [0,T]$, the function of ξ

$$\delta u(s)(\xi) = (u_1(s) - u_0(s))\xi + u_0(s),$$

is obviously of class C^∞; however does not belong to the domain of A since does not vanishes at 0 and 1. One can show however that([6])

$$\delta u(s) \in D((-A)^\varepsilon),\ \forall\ s \geq 0,\ \varepsilon \in [0,1/4).$$

This implies that, for a suitable constant $c > 0$ we have

$$|Ae^{(t-s)A}\delta u(s)| \leq \frac{c}{(t-s)^\rho},\ t > s \geq 0,$$

with $\rho < 3/4$, so that formula (3.3.4) is meaningful.

Equation (3.3.4) can be considered as the mild form of the state equation

$$\begin{cases} y'(t) = Ay(t) - A\delta u(t),\ t \geq 0, \\ \\ y(0) = x \in H, \end{cases} \tag{3.3.5}$$

so that $B = -A\delta$. This is not meaningful because the intersection of the range of δ with the domain of A is $\{0\}$. However one is able to give a meaning to the Riccati equation by writing

$$B = (-A)^{1-\gamma}[(-A)^\gamma \delta] := (-A)^{1-\gamma}\delta_\gamma,$$

where $\gamma \in (0,1/4)$ and consequently the operator δ_γ is bounded. In this way the term PBB^*P can be written as

$$P(-A)^{1-\gamma}\delta_\gamma\delta_\gamma^*[P(-A)^{1-\gamma}]^*.$$

Now the idea is to try to write an equation for $P(-A)^{1-\gamma}$.

To go further see the books quoted in the preface.

[6] See e.g. J.L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod, (1968).

APPENDICES

A Linear Semigroups Theory

In all this appendix X represents a complex Banach space (norm $|\cdot|$), and $L(X)$ the Banach algebra of all linear bounded operators from X into X endowed with the sup norm:

$$\|T\| = \sup\{|Tx| : \ x \in X, \ |x| \leq 1\}.$$

A.1 Some preliminaries on spectral theory

Let $A : D(A) \subset X \to X$ be a linear closed operator. We say that $\lambda \in \mathbb{C}$ belongs to the *resolvent set* $\rho(A)$ of A if $\lambda - A$ is bijective e and $(\lambda - A)^{-1} \in L(X)$; in this case the operator $R(\lambda, A) := (\lambda - A)^{-1}$ is called the *resolvent* of A at λ. The complementary set $\sigma(A)$ of $\rho(A)$ is called the *spectrum* of A.

Example A.1.1 Let $X = C([0,1])$ be the Banach space of all continuous functions on $[0,1]$ endowed with the sup norm, and let $C^1([0,1])$ be the subspace of $C([0,1])$ of all functions u that continuously differentiable. Let us consider the two following linear operators on X :

$$D(A) = C^1([0,1]), \ Au = u' \ , \forall \ u \in D(A),$$

$$D(B) = \{u \in C^1([0,1]); \ u(0) = 0\}, \ Bu = u' \ \ \forall \ u \in D(B).$$

We have

$$\rho(A) = \emptyset, \ \sigma(A) = \mathbb{C}.$$

In fact, given $\lambda \in \mathbb{C}$, the mapping $\lambda - A$ is not injective since, for all $c \in \mathbb{C}$ the function $u(\xi) = ce^{\lambda\xi}$ belongs to $D(A)$ and $(\lambda - A)u = 0$.

For as the operator B is concerned, we have

$$\rho(B) = \mathbb{C}, \ \sigma(A) = \emptyset.$$

and

$$(R(\lambda, B)f)(\xi) = -\int_0^\xi e^{\lambda(\xi-\eta)} f(\eta) d\eta, \forall \ \lambda \in \mathbb{C}, \forall \ f \in X, \forall \ \xi \in [0,1].$$

In fact $\lambda \in \rho(B)$ if and only if the problem

$$\begin{cases} \lambda u(\xi) - u'(\xi) = f(\xi) \\ \\ u(0) = 0 \end{cases}$$

has a unique solution $f \in X$.

Let us prove the important *resolvent identity.*

Proposition A.1.2 *If $\lambda, \mu \in \rho(A)$ we have*

$$R(\lambda, A) - R(\mu, A) = (\mu - \lambda)R(\lambda, A)R(\mu, A) \tag{A.1.1}$$

Proof. For all $x \in X$ we have

$$(\mu - \lambda)R(\lambda, A)x = (\mu - A + A - \lambda)R(\lambda, A)x = (\mu - A)R(\lambda, A)x - x$$

Applying $R(\mu, A)$ to both sides of the above identity, we find

$$(\mu - \lambda)R(\mu, A)R(\lambda, A)x = R(\lambda, A)x - R(\mu, A)x$$

and the conclusion follows. □

Proposition A.1.3 *Let A be a closed operator. Let $\lambda_0 \in \rho(A)$, and $|\lambda - \lambda_0| < \frac{1}{\|R(\lambda_0, A)\|}$. Then $\lambda \in \rho(A)$ and*

$$R(\lambda, A) = R(\lambda_0, A)(1 + (\lambda - \lambda_0)R(\lambda_0, A))^{-1} \tag{A.1.2}$$

Thus $\rho(A)$ is open and $\sigma(A)$ is closed. Moreover

$$R(\lambda, A) = \sum_{k=1}^{\infty} (-1)^k (\lambda - \lambda_0)^k R^{k+1}(\lambda_0, A), \tag{A.1.3}$$

and so $R(\lambda, A)$ is analytic on $\rho(A)$.

Proof. The equation $\lambda x - Ax = y$ is equivalent to

$$(\lambda - \lambda_0)x + (\lambda_0 - A)x = y,$$

and, setting $z = (\lambda_0 - A)x$, to

$$z + (\lambda - \lambda_0)R(\lambda_0, A)z = y.$$

Since $\|(\lambda - \lambda_0)R(\lambda_0, A)\| < 1$ it follows

$$z = (1 + (\lambda - \lambda_0)R(\lambda_0, A))^{-1} y,$$

that yields the conclusion. □

A.2 Strongly continuous semigroups

A *strongly continuous semigroup* on X is a mapping $T : [0, \infty) \to L(X),\ t \to T(t)$ such that

(i) $T(t+s) = T(t)T(s), \forall\ t, s \geq 0,\ T(0) = I.$

(ii) $T(\cdot)x$ is continuous for all $x \in X$.

Remark A.2.1 $\|T(\cdot)\|$ is locally bounded by the uniform boundedness theorem.

The *infinitesimal generator* A of $T(\cdot)$ is defined by

$$\begin{cases} D(A) = \left\{ x \in X : \exists \lim\limits_{h \to 0+} \Delta_h x \right\} \\ Ax = \lim\limits_{h \to 0+} \Delta_h x, \end{cases} \tag{A.2.1}$$

where

$$\Delta_h = \frac{T(h) - I}{h}, h > 0.$$

Proposition A.2.2 *$D(A)$ is dense in X.*

Proof. For all $x \in H$ and $a > 0$ we set

$$x_a = \frac{1}{a} \int_0^a T(s)x ds.$$

Since $\lim_{a \to 0} x_a = x$, it is enough to show that $x_a \in D(A)$. We have in fact for any $h \in (0, a)$,

$$\Delta_h x_a = \frac{1}{ah} \left[\int_a^{a+h} T(s)x ds - \int_0^h T(s)x ds \right],$$

and, consequently $x_a \in D(A)$ since

$$\lim_{h \to 0} \Delta_h x_a = \Delta_a x.$$

□

Exercise A.2.3 *Prove that $D(A^2)$ is dense in X.*

We now study the derivability of the semigroup $T(t)$. Let us first notice that, since

$$\Delta_h T(t)x = T(t)\Delta_h x,$$

if $x \in D(A)$ then $T(t)x \in D(A), \forall\, t \geq 0$ and $AT(t)x = T(t)Ax$.

Proposition A.2.4 *Assume that $x \in D(A)$, then $T(\cdot)x$ is differentiable $\forall\, t \geq 0$ and*

$$\frac{d}{dt}\, T(t)x = AT(t)x = T(t)Ax \tag{A.2.2}$$

Proof. Let $t_0 \geq 0$ be fixed and let $h > 0$. Then we have

$$\frac{T(t_0 + h)x - T(t_0)x}{h} = \Delta_h T(t_0)x \overset{h\to 0}{\to} AT(t_0)x.$$

This shows that $T(\cdot)x$ is right differentiable at t_0. Let us show left differentiability, assuming $t_0 > 0$. For $h \in]0, t_0[$ we have

$$\frac{T(t_0 - h)x - T(t_0)x}{h} = T(t_0 - h)\Delta_h x \overset{h\to 0}{\to} T(t_0)Ax,$$

since $\|T(t)\|$ is locally bounded by Remark A.2.1. □

Proposition A.2.5 *A is a closed operator.*

Proof. Let $(x_n) \subset D(A)$, and let $x, y \in X$ be such that

$$x_n \to x, \quad Ax_n = y_n \to y$$

Then we have

$$\Delta_h x_n = \frac{1}{h}\int_0^h T(t)y_n dt.$$

As $h \to 0$ we get $x \in D(A)$ and $y = Ax$, so that A is closed. □

We end this section by studying the asymptotic behaviour of $T(\cdot)$. We define the *type* of $T(\cdot)$ as

$$\omega_0 = \inf_{t>0} \frac{\log \|T(t)\|}{t}.$$

Clearly $\omega_0 \in [-\infty, +\infty)$.

Proposition A.2.6 *We have*

$$\omega_0 = \lim_{t\to+\infty} \frac{\log \|T(t)\|}{t}. \tag{A.2.3}$$

Proof. It is enough to show that

$$\limsup_{t\to\infty} \frac{\log \|T(t)\|}{t} \leq \omega_0.$$

Let $\varepsilon > 0$ and $t_\varepsilon > 0$ be such that

$$\frac{\log \|T(t_\varepsilon)\|}{t_\varepsilon} < \omega_0 + \varepsilon.$$

Set

$$t = n(t)t_\varepsilon + r(t),\ n(t) \in \mathbb{N}, r(t) \in [0, t_\varepsilon).$$

Since $\|T(\cdot)\|$ is locally bounded, there exists $M_\varepsilon > 0$ such that

$$\|T(t)\| \leq M_\varepsilon,\ t \in [0, t_\varepsilon].$$

We have

$$\frac{\log \|T(t)\|}{t} = \frac{\log \|T(t_\varepsilon)^{n(t)} T(r(t))\|}{t}$$

$$\leq \frac{n(t)\ \log \|T(t_\varepsilon)\| + \log \|T(r(t))\|}{n(t)t_\varepsilon + r(t)} \leq \frac{\log \|T(t_\varepsilon)\| + \frac{M_{t_\varepsilon}}{n(t)}}{t_\varepsilon + \frac{r(t)}{n(t)}}.$$

As $t \to +\infty$, we obtain

$$\limsup_{t\to\infty} \frac{\log \|T(t)\|}{t} \leq \frac{\log \|T(t_\varepsilon)\|}{t_\varepsilon} \leq \omega_0 + \varepsilon.$$

□

Corollary A.2.7 *Let T be of type ω_0. Then for all $\varepsilon > 0$ there exists $N_\varepsilon \geq 1$ such that*

$$\|T(t)\| \leq N_\varepsilon e^{(\omega_0+\varepsilon)t}, \forall\ t \geq 0 \tag{A.2.4}$$

Proof. Let $t_\varepsilon, n(t), r(t)$ as in the previous proof. Then we have

$$\|T(t)\| \le \|T(t_\varepsilon)\|^{n(t)} \|T(r(t))\| \le e^{t_\varepsilon n(t)(\omega_0+\varepsilon)} M_{t_\varepsilon} \le M_{t_\varepsilon} e^{(\omega_0+\varepsilon)t}.$$

and the conclusion follows.□

In what follows we shall denote by $\mathcal{G}(M,\omega)$ the set of all strongly continuous semigroups T such that

$$\|T(t)\| \le Me^{\omega t}, t \ge 0$$

Example A.2.8 Let $X = L^p(\mathbb{R}), p \ge 1$, $(T(t)f)(\xi) = f(\xi - t), f \in L^p(\mathbb{R})$. Then we have $\|T(t)\| = 1$ and so $\omega_0 = 0$.

Example A.2.9 Let $X = L^p(0,T), T > 0, p \ge 1$, and let

$$(T(t)f)(\xi) = \begin{cases} f(\xi - t) \text{ if } \xi \in [t,T] \\ 0 \text{ if } \xi \in [0,t[\end{cases}$$

Then we have $T(t) = 0$ if $t \ge T$ and so $\omega_0 = -\infty$.

Exercise A.2.10 Let $A \in L(X)$ compact and let $\{\lambda_i\}_{i\in\mathbb{N}}$ be its eigenvalues. Set $T(t) = e^{tA}$. Then we have

$$\omega_0 = \sup_{i\in\mathbb{N}} \text{ Re } \lambda_i.$$

A.3 The Hille–Yosida theorem

We assume here that $T \in \mathcal{G}(M,\omega)$. We denote by A its infinitesimal generator.

Proposition A.3.1 *We have*

$$\rho(A) \supset \{\lambda \in \mathbb{C} \text{ Re } \lambda > \omega\} \tag{A.3.1}$$

$$R(\lambda, A)y = \int_0^\infty e^{-\lambda t} T(t)y dt, \ y \in X, \ \text{ Re } \lambda > \omega \tag{A.3.2}$$

Proof. Set

$$\Sigma = \{\lambda \in \mathbb{C}; \ \text{Re}\, \lambda > \omega\}$$

$$F(\lambda)y = \int_0^\infty e^{-\lambda t}T(t)ydt,\ y \in X,\ \text{ Re } \lambda > \omega.$$

This is meaningful since $T \in \mathcal{G}(M,\omega)$. We have to show that, given $\lambda \in \Sigma$ and $y \in X$ the equation $\lambda x - Ax = y$ has a unique solution given by $x = F(\lambda)y$.

Existence

Let $\lambda \in \Sigma, y \in X$, $x = F(\lambda)y$. Then we have

$$\Delta_h x = \frac{1}{h}(e^{\lambda h} - 1)x - \frac{1}{h}e^{\lambda h}\int_0^h e^{-\lambda t}T(t)ydt$$

and so, as $h \to 0$,

$$\lim_{h\to 0+} \Delta_h x = \lambda x - y = Ax$$

that is x is a solution of the equation $\lambda x - Ax = y$.

Uniqueness

Let $x \in D(A)$ be a solution of the equation $\lambda x - Ax = y$. Then we have

$$\begin{aligned}\int_0^\infty e^{-\lambda t}T(t)(\lambda x - Ax)dt &= \lambda\int_0^\infty e^{-\lambda t}T(t)xdt \\ &- \int_0^\infty e^{-\lambda t}\frac{d}{dt}T(t)xdt = x,\end{aligned}$$

so that $x = F(\lambda)y$.

We are now going to prove the *Hille–Yosida* theorem.

Theorem A.3.2 *Let $A : D(A) \subset X \to X$ be a closed operator. Then A is the infinitesimal generator of a strongly continuous semigroup belonging to $\mathcal{G}(M,\omega)$ if and only if*

$$\begin{aligned}&(i)\ \ \rho(A) \supset \{\lambda \in \mathbb{R};\ \lambda > \omega\} \\ &(ii)\ \ \|R^n(\lambda, A)\| \le \tfrac{M}{(\lambda-\omega)^n},\ \forall\, n \in \mathbb{N}\ \ \forall\, \lambda > \omega \qquad \text{(A.3.3)} \\ &(iii)\ \ D(A)\ \ \textit{is dense in}\ \ X.\end{aligned}$$

Given a linear operator A fulfilling $(A.3.3)$ it is convenient to introduce a sequence of linear operators (called the *Yosida approximations* of A). They are defined as

$$A_n = nAR(n,A) = n^2 R(n,A) - n \tag{A.3.4}$$

Lemma A.3.3 *We have*

$$\lim_{n\to\infty} nR(n,A)x = x, \ \ \forall\, x \in X, \tag{A.3.5}$$

and

$$\lim_{n\to\infty} A_n x = Ax, \ \ \forall\, x \in D(A). \tag{A.3.6}$$

Proof. Since $D(A)$ is dense in X and $\|nR(n,A)\| \le \frac{Mn}{n-\omega}$, to prove (A.3.5) it is enough to show that.

$$\lim_{n\to\infty} nR(n,A)x = x, \ \ \forall\, x \in D(A).$$

In fact for any $x \in D(A)$ we have

$$|nR(n,A)x - x| = |R(n,A)Ax| \le \frac{M}{n-\omega}|Ax|,$$

and the conclusion follows.

Finally if $x \in D(A)$ we have

$$A_n x = nR(n,A)Ax \to Ax,$$

and $(A.3.6)$ follows.□

Proof of Theorem $A.3.2$. **Necessity**. (i) follows from Proposition A.3.1 and (iii) from Proposition A.2.2. Let us show (ii). Let $k \in \mathbb{N}$ and $\lambda > \omega$. It follows

$$\frac{d^k}{d\lambda^k} R(\lambda,A)y = \int_0^\infty (-t)^k e^{-\lambda t} T(t)y dt, \ \ y \in X,$$

from which

$$\left\| \frac{d^k}{d\lambda^k} R(\lambda,A) \right\| \le M \int_0^\infty t^k e^{-\lambda t + \omega t} dt$$

that yields the conclusion.

Sufficiency.

Step 1. We have

$$\|e^{tA_n}\| \leq Me^{\frac{\omega nt}{n-\omega}}, \quad \forall\, n \in \mathbb{N}. \tag{A.3.7}$$

In fact, by the identity

$$e^{tA_n} = e^{-nt}e^{tn^2R(n,A)} = e^{-nt}\sum_{k=0}^{\infty}\frac{n^{2k}t^kR^k(n,A)}{k!},$$

it follows

$$\|e^{tA_n}\| \leq Me^{-nt}\sum_{k=0}^{\infty}\frac{n^{2k}t^k}{(n-\omega)^k k!}.$$

Step 2. There exists $C > 0$ such that, for all $m, n > 2\omega$, and $x \in D(A^2)$,

$$\|e^{tA_n}x - e^{tA_m}x\| \leq Ct\frac{|m-n|}{(m-\omega)(n-\omega)}\,\|A^2x\|. \tag{A.3.8}$$

Setting $u_n(t) = e^{tA_n}x$, we have

$$\frac{d}{dt}(u_n(t) - u_m(t)) = A_n(u_n(t) - u_m(t)) - (A_m - A_n)u_m(t)$$

$$= A_n(u_n(t) - u_m(t)) - (n-m)A^2R(m,A)R(n,A)u_m(t).$$

It follows

$$u_n(t) - u_m(t) = (n-m)A^2R(m,A)R(n,A)\int_0^t e^{(t-s)A_n}u_m(s)ds$$

$$= (n-m)R(m,A)R(n,A)\int_0^t e^{(t-s)A_n}e^{sA_m}A^2x.$$

Step 3. For all $x \in X$ there exists the limit

$$\lim_{n\to\infty} e^{tA_n}x =: T(t)x \tag{A.3.9}$$

and $T : [0, \infty) \to L(X), t \to T(t)$ is strongly continuous.

From the second step it follows that the sequence $(u_n(t)$ is Cauchy, uniformly in t on compact subsets of $[0, +\infty[$, for all $x \in D(A^2)$. Since $D(A^2)$ is dense in X (see Exercise A.2.3) this holds for all $x \in X$. Finally it is easy to check that $T(\cdot)$ is strongly continuous.

Step 4. If $x \in D(A)$, then $T(\cdot)x$ is differentiable and

$$\frac{d}{dt}\, T(t)x = T(t)Ax = AT(t)x.$$

In fact let $x \in D(A)$, and $v_n(t) = \frac{d}{dt}u_n(t)$. Then

$$v_n(t) = e^{tA_n} A_n x$$

Since $x \in D(A)$ there exists the limit

$$\lim_{n\to\infty} v_n(t) = e^{tA}Ax.$$

This implies that u is differentiable and $u'(t) = v(t)$ so that $u \in C^1([0, +\infty); X)$. Moreover

$$A(nR(n, A)u_n(t)) = u_n'(t) \to v(t).$$

Since A is closed and $nR(n, A)u_n(t) \to u(t)$ it follows that $u(t) \in D(A)$ and $u'(t) = Au(t)$.

Step 5. A is the infinitesimal generator of $T(\cdot)$.

Let B be the infinitesimal generator of $T(\cdot)$. By Step 4 $B \supset A$ ([7]). It is enough to show that if $x \in D(B)$ then $x \in D(A)$. Let $x \in D(B)$, $\lambda_0 > \omega$, setting $z = \lambda_0 x - Bx$ we have

$$z = (\lambda_0 - A)R(\lambda_0, A)z$$

$$= \lambda_0 R(\lambda_0, A)z - BR(\lambda_0, A)z = (\lambda_0 - B)R(\lambda_0, A)z.$$

Thus $x = R(\lambda_0, B)z = R(\lambda_0, A)z \in D(A)$.□

Remark A.3.4 To use the Hille-Yosida theorem requires to check infinite conditions. However if $M = 1$ it is enough to ask (ii) only for $n = 1$. In such a case $T \in \mathcal{G}(1, \omega)$. If $\omega \leq 0$ we say that $T(\cdot)$ is a *contraction semigroup*.

[7] That is $D(B) \supset D(A)$ and $Ax = Bx \ \forall\, x \in D(A)$

Example A.3.5 Let $X = C_0([0,\pi])$ the Banach space of all continuous functions in $[0,\pi]$ that vanish at 0 and π. Let A be the linear operator in X defined as

$$\begin{cases} D(A) = \{y \in C^2([0,\pi]); y(0) = y''(0) = y(\pi) = y''(\pi) = 0\} \\ \\ Ay = y'', \ \forall \ y \in D(A) \end{cases}$$

It is easy to check that $\sigma(A) = \{-n^2; \ n \in \mathbb{N}\}$. Moreover any element of $\sigma(A)$ is a simple eigenvalue whose corresponding eigenvector is given by

$$\varphi_n(\xi) = \sin \ n\xi, \ \ \forall \ n \in \mathbb{N}.$$

We have

$$A\varphi_n = -n^2\varphi_n$$

Moreover if $\lambda \in \rho(A)$ and $f \in C_0([0,\pi])$, $u = R(\lambda, A)f$ is the solution of the problem

$$\begin{cases} \lambda u(\xi) - u''(\xi) = f(\xi) \\ \\ u(0) = u(\pi) = 0. \end{cases}$$

By a direct verification we find

$$\begin{aligned} u(\xi) = &\frac{\sinh(\sqrt{\lambda}\xi)}{\sqrt{\lambda}\sinh(\sqrt{\lambda}\pi)} \int_\xi^\pi \sinh[\sqrt{\lambda}(\pi - \eta)] f(\eta) d\eta \\ &+ \frac{\sinh[\sqrt{\lambda}(\pi - \xi)]}{\sqrt{\lambda}\sinh(\sqrt{\lambda}\pi)} \int_0^\xi \sinh[\sqrt{\lambda}\eta] f(\eta) d\eta. \end{aligned} \tag{A.3.10}$$

From $(A.3.10)$ it follows that

$$\|R(\lambda, A)\| \leq \frac{1}{\lambda}, \ \forall \ \lambda > 0. \tag{A.3.11}$$

Therefore the assumptions of the Hille-Yosida theorem are fulfilled.

A.4 Cauchy problem

Let $A \in \mathcal{G}(M, \omega)$, and let $T(\cdot)$ be the semigroup generated by A.

We are here concerned with the following problem

$$\begin{cases} u'(t) = Au(t) + g(t), \ t \in [0,T] \\ \\ u(0) = x, \end{cases} \tag{A.4.1}$$

where $x \in H$ and $g \in C([0,T];X)$.

We say that $u : [0,T] \to X$ is a *strict solution* of problem $(A.4.1)$ if

(i) $u \in C^1([0,T];X)$.

(ii) $u(t) \in D(A), \forall\, t \in [0,T]$.

(iii) $u'(t) = Au(t) + g(t), \ \forall\, t \in [0,T], \ u(0) = x$

We first cosider the homogeneous problem

$$\begin{cases} u'(t) = Au(t), \ t \in [0,T] \\ \\ u(0) = x, \end{cases} \tag{A.4.2}$$

Theorem A.4.1 *Let $x \in D(A)$. Then problem $(A.4.2)$ has a unique strict solution given by $u(t) = T(t)x$.*

Proof. Existence follows from the Hille–Yosida theorem. Let us prove uniqueness. Let v be a strict solution of $(A.4.2)$. Let us fix $t > 0$ and set

$$f(s) = T(t-s)v(s), \quad s \in [0,t].$$

$f(s)$ is differentiable for $s \in [0,t)$, since

$$\begin{aligned} \frac{1}{h}(f(s+h) - f(s)) &= \frac{1}{h}(T(t-s+h)v(s+h) - T(t-s)v(s)) \\ &+ T(t-s-h)\frac{v(s+h)-v(s)}{h} \\ &+ \frac{T(t-s-h)v(s) - T(t-s)v(s)}{h}. \end{aligned} \tag{A.4.3}$$

As $h \to 0$ we find

$$\begin{aligned} f'(s) &= T(t-s)v'(s) - T'(t-s)v(s) \\ &= T(t-s)Av(s) - AT(t-s)v(s) = 0. \end{aligned}$$

Therefore f is costant and $T(t)x = v(t)$. $\square$

We now consider problem $(A.4.1)$.

Theorem A.4.2 *Let $x \in D(A)$ and $g \in C^1([0,T];X)$. Then there is a unique strict solution $u(\cdot)$ di (A.4.1) given by*

$$u(t) = T(t)x + \int_0^t T(t-s)g(s)ds. \tag{A.4.4}$$

Proof. Uniqueness can be proved as in Theorem $A.4.1$. Let us prove existence. We shall prove that the function $u(\cdot)$, defined by $(A.4.4)$ is a solution of $(A.4.1)$ and

$$u \in C^1([0,T];X) \cap C([0,T];D(A)).$$

First it is easy to check that $u \in C^1([0,T];X)$ and

$$u'(t) = T(t)g(0) + \int_0^t T(t-s)g'(s)ds. \tag{A.4.5}$$

Let us prove now that $v(t) \in D(A)$. We have in fact

$$\begin{aligned}
&\frac{1}{h}(T(h)u(t) - u(t)) \\
&= \frac{1}{h}\left[\int_0^t T(s+h)g(t-s)ds - \int_0^t T(s)g(t-s)ds\right] \\
&+\frac{1}{h}\int_t^{t+h} T(s)g(t-s+h))ds - \frac{1}{h}\int_0^h T(s)g(t-s))ds.
\end{aligned}$$

As $h \to 0$ we have

$$\begin{aligned}
&\lim_{h\to 0}\frac{1}{h}(T(h)u(t) - u(t)) \\
&= \int_0^t T(s)g'(t-s)ds + T(t)g(0) - g(t).
\end{aligned} \tag{A.4.6}$$

From $(A.4.5)$ and $(A.4.6)$ it follows that $u \in C([0,T];D(A))$ and the conclusion follows. □

Let $x \in H$ and $f \in C([0,T];H)$. The the function u defined by

$$u(t) = T(t)x + \int_0^t T(t-s)g(s)ds \tag{A.4.7}$$

clearly belongs to $C([0,T];H)$. We say that u is a *mild solution* of $(A.4.1)$.

B Contraction Principle

Let $T > 0$, and let $\{\gamma_n\}$ be a sequence of mappings from $C_u([0,T];\Sigma(H))$ into itself such that

$$\|\gamma_n(P) - \gamma_n(Q)\| \le \alpha\|P - Q\|, \ \forall\, P, Q \in C_u([0,T];\Sigma(H)), \ n \in \mathbb{N},$$

where $\alpha \in [0,1)$.

Moreover assume that there exists a mapping γ from the space $C_u([0,T];\Sigma(H))$ into itself such that

$$\lim_{n\to\infty} \gamma_n^m(P) = \gamma^m(P) \text{ in } C_s([0,T];\Sigma(H)), \tag{B.0.1}$$

for all $P \in C_u([0,T];\Sigma(H))$ and all $m \in \mathbb{N}$, where γ^m and γ_n^m are defined by recurrence as

$$\gamma^1 = \gamma, \ \gamma^{m+1}(P) = \gamma(\gamma^m(P)),$$

$$\gamma_n^1 = \gamma_n, \ \gamma_n^{m+1}(P) = \gamma_n(\gamma_n^m(P)),$$

for $m = 2, 3, ...$ and $P \in C_s([0,T];\Sigma(H))$. It is easy to check that

$$\|\gamma(P) - \gamma(Q)\| \le \alpha\|P - Q\|, \ \forall\, P, Q \in C_u([0,T];\Sigma(H)).$$

Then, by the classical Contraction Mapping Principle, there exists unique P_n and P in $C_u([0,T];\Sigma(H))$ such that

$$\gamma_n(P_n) = P_n \text{ rm and } \gamma(P) = P.$$

However, since we do not assume that

$$\gamma_n(P) \to \gamma(P) \text{ in } C_u([0,T];\Sigma(H)$$

we cannot conclude that $P_n \to P$ in $C_u([0,T];\Sigma(H)$, but a weaker result holds.

Lemma B.0.3 *Under the previous hypotheses on the sequence of mappings* $\{\gamma_n\}$,

$$P_n \to P \text{ in } C_s([0,T];\Sigma(H)).$$

Proof. Set

$$P^0 = 0, \quad P_n^0 = 0,$$

and define

$$P^m = \gamma^m(P^0), \quad P_m^m = \gamma_n^m(P^0), \ m \in \mathbb{N}.$$

By the classical Contraction Mapping Principle, we have

$$\lim_{m\to\infty} P^m = P, \quad \lim_{m\to\infty} P_n^m = P_n \text{ in } C_u([0,T];\Sigma(H)), \ n \in \mathbb{N}.$$

Moreover

$$\|P - P^m\| \leq \sum_{k=m}^{\infty} \alpha^k \|\gamma(P^0)\|, \quad \|P_n - P_n^m\| \leq \sum_{k=m}^{\infty} \alpha^k \|\gamma_n(P^0)\|.$$

Now fix $x \in H$, then for all $t \in [0,T]$

$$\begin{aligned} |P(t)x - P_n(t)x| \quad &\leq \quad |P(t)x - P^m(t)x| + |P^m(t)x - P_n^m(t)x| \\ &+ \quad |P_n^m(t)x - P_n(t)x|. \end{aligned} \tag{B.0.2}$$

Given $\varepsilon > 0$ there exists $m_\varepsilon \in \mathbb{N}$ such that

$$\sum_{k=m}^{\infty} \alpha^k \left[\|\gamma(P^0)\| + \|\gamma_n(P^0)\|\right] \leq \frac{\varepsilon}{2}, \tag{B.0.3}$$

for all $m \geq m_\varepsilon$ and all $n \in \mathbb{N}$. By (B.0.2) and (B.0.3) it follows that

$$|P(t)x - P_n(t)x| \leq \frac{\varepsilon}{2} + |P^{m_\varepsilon}(t)x - P_n^{m_\varepsilon}(t)x|, \ \forall\, t \in [0,T].$$

Now (B.0.1) yields the conclusion. □

Introduction to Geometric Nonlinear Control; Controllability and Lie Bracket

Bronisław Jakubczyk*

Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028003

*B.Jakubczyk@impan.gov.pl

Abstract

We present an introduction to the qualitative theory of nonlinear control systems, with the main emphasis on controllability properties of such systems. We introduce the differential geometric language of vector fields, Lie bracket, distributions, foliations etc. One of the basic tools is the orbit theorem of Stefan and Sussmann. We analyse the basic controllability problems and give criteria for complete controllability, accessibility and related properties, using certain Lie algebras of vector fields defined by the system. A problem of path approximation is considered as an application of the developed theory. We illustrate our considerations with examples of simple systems or systems appearing in applications. The notes start from an elementary level and are self-contained.

Contents

1 Controllability and Lie bracket **111**
1.1 Control systems and controllability problems 111
1.2 Vector fields and flows 115
1.3 Lie bracket and its properties 117
1.4 Coordinate changes and Lie bracket 121
1.5 Vector fields as differential operators 125

2 Orbits, distributions, and foliations **130**
2.1 Distributions and local Frobenius theorem 130
2.2 Submanifolds and foliations 133
2.3 Orbits of families of vector fields 136
2.4 Integrability of distributions and foliations 139

3 Controllability and accessibility **143**
3.1 Basic definitions 143
3.2 Taylor linearization 144
3.3 Lie algebras of control system 146
3.4 Accessibility criteria 148

4 Controllability and path approximation **154**
4.1 Time-reversible systems 154
4.2 Approximating curves by trajectories 158

References **168**

1 Controllability and Lie bracket

Controllability properties of a control system are properties related to the following questions. (Q1) Can the system be steered from a given initial state x_0 to a given final state x_1? (Q2) Can this be done for any pair of initial and final states? (Q3) How large is the set of points to which the system can be steered from a given initial state x_0? (Q4) Which trajectories of the system are realizable and how do we find controls realizing them?

Such questions can be motivated by practical problems and they are basic for any qualitative study of control systems. Our aim in these lectures will be to develop tools which will enable us to answer such questions and to understand qualitative properties of nonlinear control systems. We will see that for a large class of problems a control system can be represented by a family of vector fields (dynamical systems). The qualitative properties of the control system depend on the properties of the vector fields (dynamical systems) and interactions between them. The basic tool which will enable us to understand the interactions between different vector fields will be the Lie bracket.

1.1 Control systems and controllability problems

By a control system we shall mean a system of the form

$$\Sigma: \qquad \dot{x} = f(x, u),$$

where x, called *state* of Σ, takes values in an open subset X of $\mathbb{R}^n$ (or in a differentiable manifold X of dimension n) and u, called *control*, takes values in a set U. We call X the *state space* of the system and U the *control set*. When the control u is fixed the system equation $\dot{x} = f(x, u)$ defines a single dynamical system. Thus, the control system Σ can be viewed as a collection of dynamical systems parametrized by the control as parameter. We will see later that this interpretation is fruitful.

Example 1.1 *Boat on a lake.* Consider a motor boat on a lake. We can choose some coordinate system in which the lake is identified with a subset X of $\mathbb{R}^2$ and the state of the boat with a point $x = (x_1, x_2) \in X$. The simplest mathematical model of the motion of the boat is the following control system

$$\dot{x} = u$$

where the control $u = (u_1, u_2)$ is the velocity vector which belongs to the set $U = \{u \in \mathbb{R}^2 : \|u\| \leq m\}$, where $\|u\| = \sqrt{u_1^2 + u_2^2}$ is the norm of u and m is the maximal possible velocity of the boat.

A different version of the problem is obtained if we consider a motor boat (or a rowing boat) on a river. Then the set of velocities of the boat $F(x)$ depends on the current of the river at this point. This means that in our model we have to change the equation $\dot{x} = u$ for

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

where the control u is in the set $U = \{u \in \mathbb{R}^2 : \|u\| \leq m\}$ and $f(x)$ denotes the velocity vector of the current of the river at the point x. We could also keep the equation $\dot{x} = u$ and choose the control set $\tilde{U}(x) = f(x) + U$ depending on x (we will usually try to avoid the latter possibility as more complicated). Clearly, if the set of available velocities $F(x) = f(x) + U$ contains 0 in its interior then the boat can be steered from any initial position to any final position if we use enough time.

Example 1.2 *Sailing boat.* A more interesting system is obtained when the boat is a sailing boat. Assuming that the wind is stable (of constant direction and force) we can model the motion of the boat on a lake by the equation

$$\dot{x} = v(\theta),$$

where θ is the angle of the axis of the boat with respect to the wind. The angle θ is treated as control and takes values in the set $U = (\alpha, 2\pi - \alpha)$, where α is the minimal angle with which the boat can sail against the wind. The velocity v, as a function of θ, depends on the characteristics of the boat related to the wind and it usually looks like in Figure 1 (a). An interesting problem for a sailor appears when the target is placed in the "dead cone" of the boat, when we look at it from the starting point. In that case sailing consists of a series of tacks chosen in such a way that the target is reached even if it is placed in the dead cone. In fact, sailing against the wind can be restricted to using only two values of the control $\theta = \pm\theta_{opt}$, where θ_{opt} maximizes the parallel to the wind component of $v(\theta)$ (directed against the wind). In this case the system reduces to two dynamical systems with two available velocities $v^+ = v(\theta_{opt})$ and $v^- = v(-\theta_{opt})$. By changing the tacks (Figure 1 (b)) with the time spent for each (left and right) tack proportional, respectively, to constants λ_+ and λ_- (where $\lambda_+ + \lambda_- = 1$) the

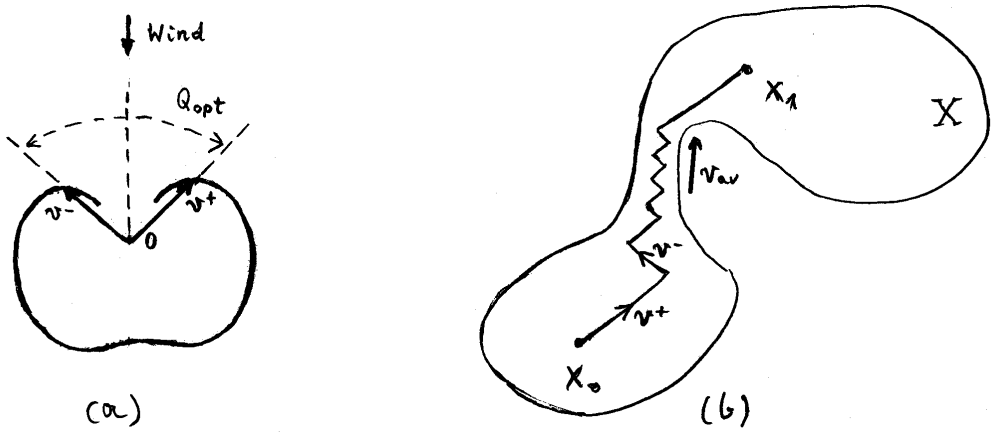


Figure 1

sailing boat changes its position as it was sailing with the average velocity $v_{av} = \lambda_+ v(\theta_{opt}) + \lambda_- v(-\theta_{opt})$.

The observation of the above example can be generalized to the following informal (but intuitively plausible)

Conclusion (*principle of convexification*). In analysing controllability properties of systems Σ we can replace the set of available velocities $F(x) = \{f(x,u) : u \in U\}$ by its convex hull, the trajectories of the convexified system can be approximated (in C^0 topology) by the trajectories of the original system. In particular, if

$$0 \in \operatorname{int} \operatorname{co} F(x)$$

for all $x \in X$, then the system is completely controllable (any state can be reached from any other state).

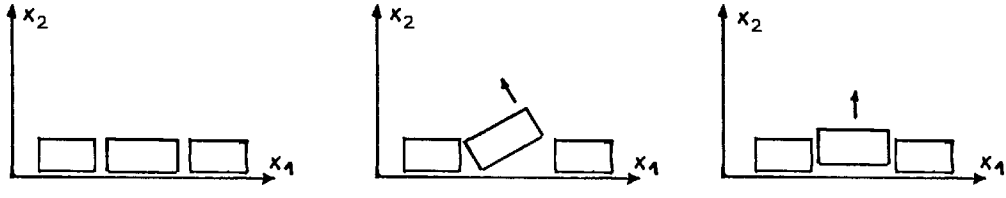
Example 1.3 *Car parking I.* Suppose we would like to unpark our car blocked by two other cars parked on the side of the street (Figure 2 (a)). The simplest but not always applicable strategy is to use a series of moves that gradually turn the car until it points to the free part of the street (Figure 2 (b)).

We use the following mathematical model of our problem. We let x_1 and x_2 denote the Euclidean coordinates of the geometric center of the back axle of the car and ϕ will denote the angle between the axis of the car

and the x_1-axis. We assume that the street is parallel to the x_1-axis. It is enough to consider movements with two extreme positions of the steering wheel. If we assume that the car moves with a constant angular velocity $\pm b$ then the velocity of the center of the rear axle moves along a circle (at each position of the steering wheel). The kinematic movements of the car in coordinates $x = (x_1, x_1, \phi)$ can be described by the following two vector fields on $\mathbb{R}^2 \times (-\pi, \pi) \subset \mathbb{R}^3$

$$f = (r\cos\phi, r\sin\phi, b)^T, \quad g = (r\cos\phi, r\sin\phi, -b)^T,$$

where r is a constant. Our strategy is to use a series of short moves (with equal length) where we interchange moving forwards with the leftmost position of the steering wheel (the vector field f) and moving backwards with the rightmost position of the steering wheel (the vector field $-g$). Intuitively, the overall movement should be approximately described by the vector which is a linear combination of the vectors f and $-g$. We have $(1/2)f - (1/2)g = (0, 0, b)$ which suggests that our series of movements can be approximated by a pure turn.



(a) (b) (c)

Figure 2

We shall later show that our approximation is justified by a suitable mathematical result (Proposition 1.8). The above strategy cannot be used if the cars are approximately rectangular and the blocking cars are parked very close to our car (then their geometry will not allow for the turn of our car). In this case we have to use a more sophisticated strategy (Example 1.10) based on the notion of Lie bracket of vector fields. This strategy allows, approximately, to drive our car almost parallel in the direction perpendicular to the street (Figure 2 (c)).

In fact, we shall be able to show later the following much stronger controllability property of the car. "Given $\epsilon > 0$ and any compact curve in the

state space $X = \{(x_1, x_2, \phi) \in \mathbb{R}^2 \times S^1\}$, there exist admissible moves of the car which approximately follow the curve. More precisely, they bring it from the initial position of the curve to the final position of the curve and the car is never at a distance (in the state space) larger than ϵ from the curve."

1.2 Vector fields and flows

Let X denote an open subset of $\mathbb{R}^n$, possibly equal to $\mathbb{R}^n$ (the reader familiar with the theory of differentiable manifolds may assume from the beginning that X is a manifold). We denote by T_pX the space of tangent vectors to X at the point p. In the case where X is an open subset of $\mathbb{R}^n$ one can identify T_pX with $\mathbb{R}^n$ (this identification depends on the coordinate system).

A *vector field* on X is a mapping

$$X \ni p \longrightarrow f(p) \in T_pX$$

which assigns a tangent vector at p to any point p in X (Figure 3). An analogous mapping defined on an open subset of X, only, will be called *partial vector field*. In a given system of coordinates f can be expressed as a column vector

$$f = (f_1, \ldots, f_n)^T,$$

where "T" stands for transposition. We say that f is of class C^k if its components are of class C^k.

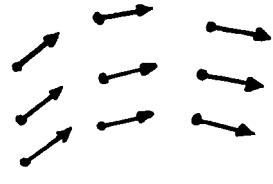


Figure 3

We shall usually assume that the vector fields considered here are of class C^∞. The space of such vector fields forms a linear space (with natural, pointwise operations of summation and multiplication by numbers) denoted by $V(X)$.

For any vector field (or partial vector field) f we can write the differential equation

$$\dot{x} = f(x).$$

From theorems on existence of solutions of ordinary differential equations it follows that, if f is of class C^k and $k \geq 1$, then for any initial point p in the domain of f there is an open interval I containing zero and a differentiable curve $t \to x(t) = \gamma_t(p), t \in I$, which satisfies the above equation and $x(0) = \gamma_0(p) = p$. If f is of class C^∞, then from elementary properties of differential equations it follows that the map

$$(t, p) \longrightarrow \gamma_t(p)$$

is also of class C^∞ and is well defined on a maximal open subset of $\mathbb{R} \times X$. The resulted family γ_t of local maps of X (Figure 4), called the *local flow* or simply the *flow* of the vector field f, has the following group type properties ("$\circ$" denotes composition of maps)

$$\gamma_{t_1} \circ \gamma_{t_2} = \gamma_{t_1+t_2}, \quad \gamma_{-t} = (\gamma_t)^{-1}, \quad \gamma_0 = \mathrm{id}. \tag{1}$$

If the solution $\gamma_t(p)$ is well defined for all $t \in \mathbb{R}$ and $p \in X$, then the

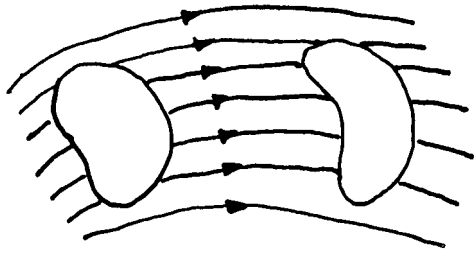


Figure 4

vector field f is called *complete* and its flow forms a one parameter group of (global) diffeomorphisms of X. Any one parameter family of maps which satisfies conditions (1) defines a unique vector field through the formula

$$f(p) = \left.\frac{\partial}{\partial t}\right|_{t=0} \gamma_t(p),$$

and the flow of this vector field coincides with γ_t.

We shall denote the local flow of a vector field f by γ_t^f or by $\exp(tf)$. A reason for the latter notation will become clear later.

Example 1.4 The linear vector field $f(x) = Ax$ is complete and the corresponding flow is the one-parameter group of linear transformations $p \longrightarrow e^{At}p$, i.e.

$$\gamma_t = e^{At},$$

where $e^{At} = \sum_{i \geq 0} A^i \frac{t^i}{i!}$.

1.3 Lie bracket and its properties

A nonlinear control system can be considered as a collection of dynamical systems (vector fields) parametrized by a parameter called control. It is natural to expect that basic properties of such a system depend on interconnections between the different dynamical systems corresponding to different controls. We represent our dynamical systems by vector fields as this allows us to perform algebraic operations on them such as taking linear combinations and a taking a product called Lie bracket. It is the Lie product which allows studying interconnections between different dynamical systems in a coordinate independent way.

The Lie bracket of two vector fields is another vector field which, roughly speaking, measures noncommutativeness of the flows of both vector fields. Noncommutativeness means here dependence of the result of applying the flows on the order of applying these flows. This remark, as well as the definition of Lie bracket is made precise below.

There are three equivalent definitions of Lie bracket and each of them will be useful to us later. We start with the easiest (but coordinate dependent) definition in $\mathbb{R}^n$. Let $X \subset \mathbb{R}^n$, and let f and g be vector fields on X. The *Lie bracket* of f and g is another vector field on X defined as follows

$$[f,g](x) = \frac{\partial g}{\partial x}(x)f(x) - \frac{\partial f}{\partial x}(x)g(x), \tag{2}$$

where $\partial f/\partial x$ and $\partial g/\partial x$ denote the Jacobi matrices of f and g. We will call this *the Jacobian definition of Lie bracket.*

Example 1.5 For the vector fields $f = (1,0)^T$ and $g = (0,x_1)^T$ on $\mathbb{R}^n$ one easily finds that $[f,g] = (0,1)^T$. Note that the Lie bracket of f and g adds a new direction to the space spanned by f and g at the origin.

Let $f = b$ be a constant vector field and $g = Ax$ be a linear vector field. Then $[f,g] = [b,Ax] = Ab - 0 = Ab$. Similar trivial calculations show that the following holds.

Proposition 1.6 *The Lie bracket of two constant vector fields is zero. The Lie bracket of a constant vector field with a linear vector field is a constant vector field. Finally, the Lie bracket of two linear vector fields is a linear vector field.*

The basic geometric properties of Lie bracket are stated in the following propositions. The first one says that vanishing of Lie bracket $[f,g]$ is equivalent to the fact that starting from a point p and going along trajectory of f for time t and then along trajectory of g for time s gives always the same result as with the order of taking f and g reversed (Figure 5).

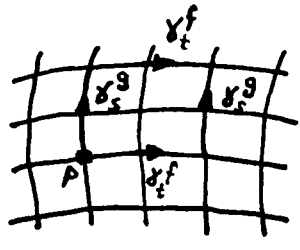


Figure 5

Proposition 1.7 *The Lie bracket of vector fields f and g is equal identically to zero if and only if their flows commute, i.e.*

$$[f,g] \equiv 0 \quad \Longleftrightarrow \quad \gamma_t^f \circ \gamma_s^g(p) = \gamma_s^g \circ \gamma_t^f(p) \quad \forall s,t \in \mathbb{R}, \forall p \in X,$$

where the equality on the right should be satisfied for those s,t and p for which both sides are well defined.

Proof. To prove the implication "$\Longleftarrow$" it is enough to note that by computing the partial derivatives $(\partial/\partial t)(\partial/\partial s)$ at $t = s = 0$ of the left side of the equality $\gamma_t^f \circ \gamma_s^g(p) = \gamma_s^g \circ \gamma_t^f(p)$ and the same partial derivatives (but in reverse order) of the right side gives the equality $(\partial f/\partial x)g = (\partial g/\partial x)f$. The converse implication will be shown after Proposition 1.13. ∎

Two vector fields having the property of Proposition 1.7 will be called *commuting*.

Proposition 1.8 *Let us fix a $p \in X$ and consider the curve* (*Figure* 6)

$$\alpha(t) = \gamma_{-t}^g \circ \gamma_{-t}^f \circ \gamma_t^g \circ \gamma_t^f(p).$$

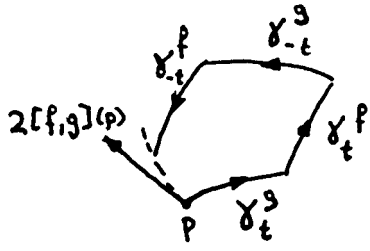


Figure 6

Then we have that its first derivative at zero vanishes, $\alpha'(0) = 0$ *and the second derivative is given by the Lie bracket:*

$$\alpha''(0) = 2[f, g](p).$$

The above means that, after a reparametrization, the tangent vector at zero to the curve $t \to \alpha(t)$ is equal to $2[f, g](p)$ (see Figure 6). This implies that the points attainable from p by means of the vector fields f and g lie not only in the "directions" $f(p)$ and $g(p)$, but also in the "direction" of the Lie bracket $[f, g](p)$. This fact will be of basic importance for studying controllability properties of nonlinear control systems.

The proof of the above proposition is omitted and follows from a more general fact proved in Section 4 (see also Spivak [Sp], page 224). Note that the formula in Proposition 1.8 can be used for defining the Lie bracket $[f, g]$.

Proposition 1.9 *Suppose we are given two vector fields* f *and* g *on* X *and a point* $p \in X$ *and let* λ_1, λ_2 *be real constants. Define the following (local) diffeomorphisms of* X

$$\phi_t = \gamma^f_{\lambda_1 t} \circ \gamma^g_{\lambda_2 t}, \quad \psi_t = \gamma^g_{-t} \circ \gamma^f_{-t} \circ \gamma^g_t \circ \gamma^f_t.$$

Then the families of curves (Figure 7)

$$\alpha_k(t) = \phi_{t/k} \circ \cdots \circ \phi_{t/k}(p), \quad k\text{-times}$$
$$\beta_k(t) = \psi_{t/k} \circ \cdots \circ \psi_{t/k}(p), \quad k^2\text{-times}$$

converge to the trajectories of the vector fields $\lambda_1 f + \lambda_2 g$ *and* $[f, g]$*, respectively. More precisely, we have the convergence*

$$\alpha_k(t) \longrightarrow \gamma^{\lambda_1 f + \lambda_2 g}_t(p), \quad \textit{and} \quad \beta_k(t) \longrightarrow \gamma^{[f,g]}_{t^2}(p) \quad \textit{as } k \longrightarrow \infty.$$

Figure 7

We will not prove this proposition here, sending the reader to Section 4. However, the reader should find the first property about the convergence of α_k intuitively clear (compare the principle of convexification from Section 1.1). Namely, the movement which jumps sufficiently often between trajectories of two vector fields (and the time spent for these vector fields is proportional to some weights) follows, approximately, a trajectory of the linear combination of these vector fields (with the same weights). This property is used, for example, by sailors passing through narrow rivers or canals. A sailing boat can go against the wind only with certain minimal positive or negative angle (Example 1.2). But, even if the direction of the canal is in the "dead" cone and the boat cannot go straight in this direction, the sailor tacks sufficiently often spending suitable amount of time for the left and the right tacks to reach the desired direction.

The property of convergence of β_k can be illustrated by the following example.

Example 1.10 *Car parking II.* Suppose the strategy of turning the car in Example 1.3 is inadmissible because the blocking cars are too close. There is a better strategy for unparking which works in any situation. Namely, we use repeatedly the following series of 4 moves: LF, RF, LB, RB, where "L" and "R" stand for the leftmost and rightmost positions of the steering wheel while "F" and "B" stand for forward and backward motions. This means that our strategy is precisely the zig-zaging strategy described by $\beta_k(t)$ in Proposition 1.9. Therefore, the resulting movement follows approximately the Lie bracket of the vector fields

$$f = (r\cos\phi, r\sin\phi, b)^T, \quad g = (r\cos\phi, r\sin\phi, -b)^T.$$

We compute

$$\frac{\partial f}{\partial x} = \begin{pmatrix} 0 & 0 & -r\sin\phi \\ 0 & 0 & r\cos\phi \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \frac{\partial g}{\partial x} = \begin{pmatrix} 0 & 0 & -r\sin\phi \\ 0 & 0 & r\cos\phi \\ 0 & 0 & 0 \end{pmatrix}$$

and the Lie bracket of f and g equals to

$$[f,g] = 2br(-\sin\phi, \cos\phi, 0)^T.$$

In particular, at $\phi = 0$ we have that

$$[f,g] = (0, 2br, 0)^T.$$

The zig-zaging strategy produces movement approximating the trajectory of the Lie bracket $[f,g]$, that is the movement keeping the axis of the car approximately constant ($\phi = 0$) and changing its x_2-coordinate only (Figure 2 (c)). This means that we should be able to unpark the car no matter how close the other cars are.

1.4 Coordinate changes and Lie bracket

To study what happens with vector fields and flows under coordinate changes let us consider a global diffeomorphism $\Phi : X \longrightarrow X$ (or a partial diffeomorphism i.e. a diffeomorphism between two open subsets of X). As tangent

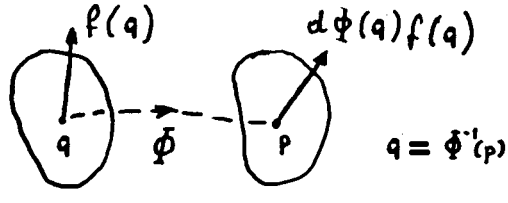


Figure 8

vectors are transformed through the Jacobian map of a diffeomorphism, our diffeomorphism defines the following transformation of a vector field f (see Figure 8)

$$\mathrm{Ad}_\Phi(f)(p) = D\Phi(q)\, f(q), \qquad q = \Phi^{-1}(p),$$

where $D\Phi$ denotes the tangent map of Φ (Jacobian mapping of Φ represented, in coordinates, by the Jacobi matrix $\partial\Phi/\partial x$). Another commonly used notation for the linear operator on $V(X)$ corresponding to the change of coordinates Φ is

$$\Phi_* f = \mathrm{Ad}_\Phi.$$

Note that the coordinate change $p = \Phi(q)$ transforms the differential equation $\dot{p} = f(p)$ into the equation $\dot{q} = \tilde{f}(q)$ where $\tilde{f} = \mathrm{Ad}_\Phi f$.

If Φ is a global diffeomorphism of X, then the operation Ad_Φ is a linear operator on the space of vector fields on X, i.e. $\mathrm{Ad}_\Phi(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 \mathrm{Ad}_\Phi(f_1) + \lambda_2 \mathrm{Ad}_\Phi(f_2)$. Additionally, if Ψ is another global diffeomorphism of X, then

$$\mathrm{Ad}_{\Phi\circ\Psi}(f) = \mathrm{Ad}_\Phi \mathrm{Ad}_\Psi(f),$$

where "$\circ$" denotes composition of maps.

For further reference we state the following fact.

Proposition 1.11 *Consider the vector field* $\mathrm{Ad}_\Phi(f)$. *The local flow of this vector field is given by*

$$\sigma_t = \Phi \circ \gamma_t \circ \Phi^{-1}.$$

Proof. It is easy to see that σ_t satisfies the group conditions (1) and we have

$$\left.\frac{\partial}{\partial t}\right|_{t=0} \Phi \circ \gamma_t \circ \Phi^{-1}(p) = D\Phi(\Phi^{-1}(p))\, f(\Phi^{-1}(p)) = (\mathrm{Ad}_\phi f)(p).$$

■

It is not immediately clear from the definition of Lie bracket in Section 1.3 that so defined $[f, g]$ is a vector field, that is, it is transformed with coordinate changes like a vector field. There are also other disadvantages of this definition which are not shared by the following *geometric definition of Lie bracket*. We define the Lie bracket of f and g as the derivative with respect to t, at $t = 0$, of the vector field g transformed by the flow of the field f. More precisely, we define (Figure 9)

$$[f, g](p) = \frac{\partial}{\partial t} D\gamma^f_{-t}(\gamma^f_t(p))\, g(\gamma^f_t(p)) = \frac{\partial}{\partial t}(\mathrm{Ad}_{\gamma^f_{-t}} g)(p). \tag{3}$$

Let us check that this definition coincides with the Jacobian definition from Section 1.3. By taking the partial derivative $\partial/\partial t$ at $t = 0$ and taking

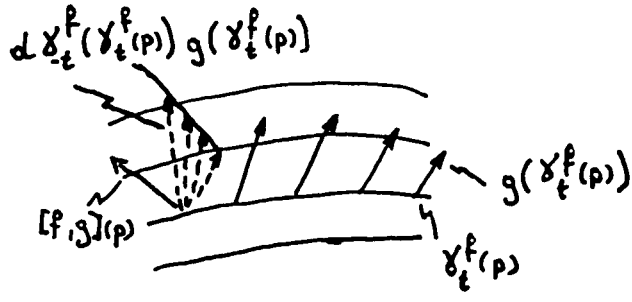


Figure 9

into account that $\gamma_0^f = \mathrm{id}$ and $\gamma_0^f(p) = p$ we find that the above definition, where t appears three times, gives

$$[f,g](p) = \Big(D\frac{\partial}{\partial t}\Big|_{t=0}\gamma_{-t}\Big)(p)g(p) + \frac{\partial}{\partial t}\Big|_{t=0} D(\mathrm{id})(\gamma_t(p))g(p) + \mathrm{id}\,\frac{\partial}{\partial t}\Big|_{t=0} g(\gamma_t(p)),$$

where we interchanged the order of taking the tangent map "D" (which is a matrix of partial derivatives with respect to the coordinates) and the partial derivative $\partial/\partial t$ in the first expression. The first term gives $-Df(p)g(p)$, the second is equal to zero, and the third equals to $Dg(p)f(p)$, which means that this definition coincides with the previous one.

It follows from the second definition of Lie bracket that $[f,g]$ transforms with coordinate changes like a vector field, that is via the Jacobi matrix of the coordinate change. Namely, we have the following basic property of equivariance of Lie bracket with coordinate changes.

Proposition 1.12 *If Φ is a (partial or global) diffeomorphism of X then*

$$[\mathrm{Ad}_\Phi f, \mathrm{Ad}_\Phi g] = \mathrm{Ad}_\Phi [f,g].$$

Proof. As we have established earlier, the flow of the vector field $\mathrm{Ad}_\Phi f$ is equal to $\sigma_t = \Phi \circ \gamma_t^f \circ \Phi^{-1}$. Thus, applying the geometric definition of Lie bracket gives

$$[\mathrm{Ad}_\Phi f, \mathrm{Ad}_\Phi g](p) = \frac{\partial}{\partial t}\Big|_{t=0} (\mathrm{Ad}_{\Phi\circ\gamma_{-t}^f\circ\Phi^{-1}}\mathrm{Ad}_\Phi g)(p)$$

$$= \frac{\partial}{\partial t}\Big|_{t=0} (\mathrm{Ad}_\Phi \mathrm{Ad}_{\gamma_{-t}^f}\mathrm{Ad}_{\Phi^{-1}}\mathrm{Ad}_\Phi g)(p) = \frac{\partial}{\partial t}\Big|_{t=0} (\mathrm{Ad}_\Phi \mathrm{Ad}_{\gamma_{-t}^f} g)(p) = \mathrm{Ad}_\Phi [f,g].$$

From the geometric definition of Lie bracket we deduce the following relation. Note that $\mathrm{Ad}_{\gamma_t^f} f = f$.

Proposition 1.13 *We have*

$$\frac{\partial}{\partial t}\mathrm{Ad}_{\gamma_t^f} g = -[f, \mathrm{Ad}_{\gamma_t^f} g] = -\mathrm{Ad}_{\gamma_t^f}([f,g]).$$

Proof. To show the first equality it is enough to note that

$$\frac{\partial}{\partial t}\mathrm{Ad}_{\gamma_t^f} g = \left.\frac{\partial}{\partial h}\right|_{h=0} \mathrm{Ad}_{\gamma_h^f}\mathrm{Ad}_{\gamma_t^f} g$$

and apply the geometric definition of Lie bracket to the vector fields $-f$ and $\mathrm{Ad}_{\gamma_t^f} g$. The second equality follows analogously from $\left.\frac{\partial}{\partial t}\right|_{t=0}\mathrm{Ad}_{\gamma_t^f} g = \left.\frac{\partial}{\partial h}\right|_{h=0}\mathrm{Ad}_{\gamma_t^f}\mathrm{Ad}_{\gamma_h^f} g$. ∎

Proof of Proposition 1.7. To show the converse implication note that from $[f,g] \equiv 0$ and the equalities in Proposition 1.13 it follows that $\mathrm{Ad}_{\gamma_t^f} g$ is independent of t, i.e. $\mathrm{Ad}_{\gamma_t^f} g = \mathrm{Ad}_{\gamma_0^f} g = g$. Therefore, the flow of g is equal to the flow of the vector field $\mathrm{Ad}_{\gamma_t^f} g$, i.e. $\gamma_t^f \circ \gamma_s^g \circ \gamma_{-t}^f = \gamma_s^g$, by Proposition 1.11. This implies that $\gamma_t^f \circ \gamma_s^g = \gamma_s^g \circ \gamma_t^f$ and the proposition is proved. ∎

Below and in the following sections we shall use the following notation. We denote $\mathrm{ad}_f g = [f,g]$. Thus, ad_f is a linear operator in the space of vector fields $V(X)$. We also consider its iterations

$$\mathrm{ad}_f^0 g = g \quad \text{and} \quad \mathrm{ad}_f^i g = \mathrm{ad}_f \cdots \mathrm{ad}_f g \quad i\text{-times}.$$

The following dependence between the operations Ad and ad follows from the formula in Proposition 1.13

$$\frac{\partial}{\partial t}(\mathrm{Ad}_{\gamma_t^f} g)(p) = -(\mathrm{ad}_f(\mathrm{Ad}_{\gamma_t^f} g))(p). \tag{4}$$

In the analytic case we also have an expansion formula which follows from this relation.

Proposition 1.14 *If the vector fields f and g are real analytic, then we have the following expansion formula for the vector field g transformed by*

the flow of the vector field f:

$$(\mathrm{Ad}_{\gamma_t^f} g)(p) = \sum_{i=0}^{\infty} \frac{(-t)^i}{i!} (\mathrm{ad}_f^i g)(p),$$

where the series converges absolutely for t in a neighborhood of zero (more precisely, each of n components of this series converges absolutely).

Proof. Applying iteratively the formula (4) and taking into account that $\gamma_0^f = \mathrm{id}$ we find that

$$\Big(\frac{\partial}{\partial t}\Big)^i (\mathrm{Ad}_{\gamma_t^f} g)(p)|_{t=0} = (-1)^i \mathrm{ad}_f^i g(p).$$

Therefore, our equality is simply the Maclaurin series of the left-hand side.

■

1.5 Vector fields as differential operators

A smooth vector field f on X defines a linear operator L_f on the space of smooth functions $C^\infty(X)$ in the following way

$$(L_f \phi)(p) = \frac{\partial}{\partial t}\Big|_{t=0} \phi(\gamma_t^f(p)) = \sum_{i=1}^{n} f_i(p) \frac{\partial}{\partial x_i} \phi(p). \tag{5}$$

This operator is called *directional derivative along f* or *Lie derivative along f* and it is a differential operator of order one.

Conversely, any differential operator of order one with no zero order term can be written as

$$L = \sum_{i=1}^{n} a_i(x) \frac{\partial}{\partial x_i}$$

and it defines a unique vector field given in coordinates as $f = (a_1, \ldots, a_n)^T$. (We can easily check that the coordinate vector $(a_1, \ldots, a_n)$ of the operator L transforms with a coordinate change Φ by the Jacobi matrix $\partial\Phi/\partial x$. Thus so defined f is a vector field on X.) This means that there is a unique correspondence

$$f \to L_f$$

between vector fields and differential operators of order one (with no zero order term).

Because of the above correspondence mathematicians often identify vector fields f with the corresponding differential operators L_f and write

$$L_f = f = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i}.$$

We will rather try to distinguish between these two objects.

We shall close this subsection with a third definition of Lie bracket and some useful corollaries to it. Let f, g be vector fields and L_f, L_g the corresponding differential operators. Consider the commutator of these operators defined by

$$[L_f, L_g] := L_f L_g - L_g L_f.$$

Proposition 1.15 *The commutator $[L_f, L_g]$ is a differential operator of order one which corresponds to the Lie bracket $[f, g]$, i.e.,*

$$[L_f, L_g] = L_{[f,g]}.$$

Proof. Given any smooth function ϕ, we compute the composed differential operator on ϕ

$$L_f L_g \phi = \sum_i f_i \frac{\partial}{\partial x_i}\Big(\sum_j g_j \frac{\partial}{\partial x_j}\phi\Big) = \sum_{ij} f_i g_j \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\phi + \sum_{ij} f_i \frac{\partial g_j}{\partial x_i}\frac{\partial \phi}{\partial x_j}.$$

The analogous expression for $L_g L_f \phi$ has the same first summand, due to commutativeness of partial derivatives with respect to x_i and x_j, thus we have

$$[L_f, L_g]\phi = L_f L_g \phi - L_g L_f \phi = \sum_{ij} f_i \frac{\partial g_j}{\partial x_i}\frac{\partial \phi}{\partial x_j} - \sum_{ij} g_i \frac{\partial f_j}{\partial x_i}\frac{\partial \phi}{\partial x_j}.$$

We see that $[L_f, L_g]$ is a differential operator of order one. Using the Jacobian definition of Lie bracket from Section 1.3 we see that $L_{[f,g]}\phi$ gives the same expression

$$L_{[f,g]}\phi = \sum_j \Big(\sum_i f_i \frac{\partial g_j}{\partial x_i} - g_i \frac{\partial f_j}{\partial x_i}\Big)\frac{\partial \phi}{\partial x_j},$$

which means that $[L_f, L_g] = L_{[f,g]}$. ∎

If we identify vector fields f with the corresponding differential operators L_f, i.e. write $f = L_f = \sum_i f_i \partial/x_i$, then Proposition 1.15 suggests that we can equivalently define the *Lie bracket* as the *commutator*

$$[f,g] = fg - gf = \sum_j \Big(\sum_i \frac{\partial g_j}{\partial x_i} f_i - \frac{\partial f_j}{\partial x_i} g_i \Big) \frac{\partial}{\partial x_j},$$

where $g = \sum_j g_j \partial/\partial x_j$. We shall call this the *algebraic definition of Lie bracket.* Clearly, this definition coincides in a given coordinate system with the Jacobian definition, if we use the identifications $f = L_f$, $g = L_g$.

Commutator of linear operators is antisymmetric and satisfies the Jacobi identity $[A,[B,C]] + [B,[C,A]] + [C,[A,B]] = 0$ (verify this using the definition $[A,B] = AB - BA$ of commutator). Therefore, we have the following properties of Lie bracket

$$\begin{aligned} [f,g] &= -[g,f] \qquad \text{(antisymmetry)}, \\ [f,[g,h]] + [g,[h,f]] + [h,[f,g]] &= 0 \quad \text{(Jacobi identity)}, \end{aligned}$$

for any vector fields f, g, in $V(X)$. The former property also follows easily from the first definition of Lie bracket. Because of the above properties the linear space $V(X)$ of smooth vector fields on X, with the Lie bracket as product, is called the *Lie algebra of vector fields on* X.

Further material concerning the basic geometric notions used in this and the following chapters can be found in any textbook on differential geometry, we refer especially to [1] and [6].

Appendix 1: Lie Algebras

A *Lie algebra* is a linear space L with a bilinear map $[\cdot,\cdot] : L \times L \longrightarrow L$ which satisfies the following properties

$$\begin{aligned} [f,g] &= -[g,f] \qquad \text{(antisymmetry)}, \\ [f,[g,h]] + [g,[h,f]] + [h,[f,g]] &= 0 \qquad \text{(Jacobi condition)}. \end{aligned}$$

The Jacobi condition can be equivalently written as the following Leibniz-Jacobi condition

$$[f,[g,h]] = [[f,g],h] + [g,[f,h]],$$

or equivalently

$$\mathrm{ad}_f[g,h] = [\mathrm{ad}_f g, h] + [g, \mathrm{ad}_f h], \qquad \text{(LJ1)}$$

where ad_f denotes the linear operator in L defined by the formula

$$\mathrm{ad}_f g = [f, g].$$

The Leibniz-Jacobi condition has also the following equivalent form

$$\mathrm{ad}_{[g,h]} f = \mathrm{ad}_g \mathrm{ad}_h f - \mathrm{ad}_h \mathrm{ad}_g f = [\mathrm{ad}_g, \mathrm{ad}_h] f, \tag{LJ2}$$

where the square bracket on the right denotes the commutator of linear operators in L: $[\mathrm{ad}_g, \mathrm{ad}_h] = \mathrm{ad}_g \mathrm{ad}_h - \mathrm{ad}_h \mathrm{ad}_g$.

A linear subspace K of L which is closed under the product $[\cdot, \cdot] : L \times L \longrightarrow L$ is called a *Lie subalgebra of* L. A *Lie subalgebra generated by a subset* or simply *Lie algebra generated by a subset* $S \subset L$ is the smallest Lie subalgebra of L which contains S. A *Lie ideal of* L is a linear subspace $I \subset L$ such that $[f, g] \in I$, whenever $f \in L$ and $g \in I$.

Example 1.16 The space $gl(n)$ of all square $n \times n$ matrices with the commutator

$$[A, B] = AB - BA$$

forms a Lie algebra. There are various Lie subalgebras of this algebra which are interesting and important for mathematics and physics. For example, skew symmetric matrices form a Lie subalgebra of this Lie algebra.

Example 1.17 The space $V(X)$ of smooth vector fields on a smooth manifold X (or simply on $X = \mathbb{R}^n$) forms a Lie algebra with Lie bracket as product. When the vector fields are treated as differential operators of order one, then the Lie bracket becomes the commutator of operators, as in the above case of square matrices (treated as linear operators). There is no surprise about this, namely, there is a Lie subalgebra of the algebra of vector fields which is formed by the space of linear vector fields: $f = Ax$, or in the operator form

$$f = \sum_{i,j} a_{ij} x_j \frac{\partial}{\partial x_i}.$$

Here, the Lie bracket corresponds to taking commutators of the corresponding matrices $[Ax, Bx] = (BA - AB)x = [B, A]x$.

Example 1.18 In the Lie algebra of linear vector fields as defined above there is an ideal which consists of all constant vector fields.

An iterative application of the Leibniz-Jacobi identity (LJ2) and of antisymmetry of Lie bracket leads to the following general property. Let $f_1,\ldots,f_k$ be elements of a Lie algebra L. We shall call an iterated Lie bracket of these elements any element of L obtained from these elements by applying iteratively the operation of Lie bracket in any possible order, e.g. $[[f_1,f_4],[f_3,f_1]]$. Left iterated Lie brackets will be brackets of the form $[f_{i_1},\ldots,[f_{i_{k-1}},f_{i_k}]\ldots]$.

Proposition 1.19 *Any iterated Lie bracket of $f_1,\ldots,f_k$ is a linear combination of left iterated Lie brackets of $f_1,\ldots,f_k$.*

For example

$$[[f_1,f_4],[f_3,f_1]] = [\mathrm{ad}_{f_1},\mathrm{ad}_{f_4}]\,[f_3,f_1] = [f_1,[f_4,[f_3,f_1]]] - [f_4,[f_1,[f_3,f_1]]].$$

Exercise Prove the above proposition (you may use induction with respect to the order of Lie bracket).

Appendix 2: Equivalence of families of vector fields

To close this chapter we shall show that the Lie brackets taken at a point of an analytic family of vector fields form a complete set of its invariants. As a control system can be represented by a family of vector fields, this will have direct applications to control systems. In another version of this result we will define a family of functions which forms a set of complete invariants for state equivalence.

Consider two general families of analytic vector fields on X and $\tilde{X}$, respectively, parametrized by the same parameter $u \in U$

$$F = \{f_u\}_{u\in U}, \qquad \tilde{F} = \{\tilde{f}_u\}_{u\in U}.$$

We shall call these families *locally equivalent* at the points p and $\tilde{p}$, respectively, if there is a local analytic diffeomorphism $\Phi : X \longrightarrow \tilde{X}$, $\Phi(p) = \tilde{p}$ which transforms the vector fields f_u into $\tilde{f}_u$ locally, i.e.

$$\mathrm{Ad}_\Phi f_u = \tilde{f}_u, \quad \text{for} \quad u \in U$$

locally around $\tilde{p}$.

Denote by $\mathcal{L}$ and $\tilde{\mathcal{L}}$ the Lie algebras of vector fields generated by the families F and $\tilde{F}$. Recall that a family of vector fields is called transitive at a point if its Lie algebra is of full rank at this point, i.e. the vector fields in this Lie algebra span the whole tangent space at this point.

We shall use the following notation for left iterated Lie brackets

$$f_{[u_1 u_2 \cdots u_k]} = [f_{u_1}, [f_{u_2}, \ldots, [f_{u_{k-1}}, f_{u_k}] \ldots]]$$

and analogous for the tilded family. In particular, $f_{[u_1]} = f_{u_1}$.

Theorem 1.20 *If the families F and $\tilde{F}$ are transitive at the points p and $\tilde{p}$, respectively, then they are locally equivalent at these points if and only if there exists a linear map between the tangent spaces $L : T_pX \longrightarrow T_{\tilde{p}}\tilde{X}$ such that*

$$L\ f_{[u_1 u_2 \cdots u_k]}(p) = \tilde{f}_{[u_1 u_2 \cdots u_k]}(\tilde{p}) \tag{6}$$

for any $k \geq 1$ and any $u_1, \ldots, u_k \in U$.

Proof. Necessity. If $\tilde{f}_u = \mathrm{Ad}_\Phi f_u$, then $\tilde{f}_u(\tilde{p}) = L\ f_u(p)$ where $L = d\Phi(p)$. To prove condition (6) in general it is enough to use iteratively the property of Lie bracket

$$[\mathrm{Ad}_\Phi f, \mathrm{Ad}_\Phi g] = \mathrm{Ad}_\Phi [f, g]$$

from which we get $\tilde{f}_{[u_1 \cdots u_k]} = \mathrm{Ad}_\Phi f_{[u_1 \ldots u_k]}$ and so the condition (6). ■

The proof of sufficiency is more involved an will be presented in the next section together with other versions of the above result.

2 Orbits, distributions, and foliations

2.1 Distributions and local Frobenius theorem

In this chapter we introduce notions and results which play a basic role in analysis and understanding the structure of nonlinear control systems. They are directly related to controllability properties of such systems. We denote by X an open subset of $\mathbb{R}^n$ or a diferentiable manifold of dimension n.

Definition 2.1 A *distribution on X* is, by definition, a map Δ which assigns to each point in X a subspace of the tangent space at this point, i.e.

$$X \ni p \longrightarrow \Delta(p) \subset T_pX.$$

The distribution Δ is called of class C^∞ if, locally around each point in X, there is a family of vector fields $\{f_\alpha\}$ (called local generators of Δ) which spans Δ, i.e. $\Delta(p) = \mathrm{span}_\alpha f_\alpha(p)$. Δ is called *locally finitely generated* if the

above family of vector fields is finite. Finally, the distribution Δ is called *of dimension* k if $\dim \Delta(p) = k$ for all points p in X, and of *constant dimension* if it is of dimension k, for some k.

We will tacitly assume that our distributions are of class C^∞.

Definition 2.2 We say that a vector field f belongs to a distribution Δ and write $f \in \Delta$ if $f(p) \in \Delta(p)$ for all p in X. A distribution Δ is called *involutive* if for any vector fields $f, g \in \Delta$ the Lie bracket is also in Δ; $[f, g] \in \Delta$. If the distribution has, locally, a finite number of generators $f_1, \ldots, f_m$ then involutivity of Δ means that

$$[f_i, f_j](p) = \sum_{k=1}^{m} \phi_{ij}^k(p) f_k(p), \qquad i, j = 1, \ldots, m,$$

where ϕ_{ij}^k are C^∞ functions.

Involutivity plays a fundamental role in the following Frobenius theorem.

Theorem 2.3 *If Δ is an involutive distribution of class C^∞ and of dimension k on X then, locally around any point in X, there exists a smooth change of coordinates which transforms the distribution Δ to the following constant distribution*

$$\operatorname{span}\{e_1, \ldots, e_k\},$$

where $e_1, \ldots, e_k$ are the constant versors $e_i = (0, \ldots, 1, \ldots, 0)^T$, with 1 at i-th place.

Proof. The proof will consist of two steps.

Step 1. We shall first show that the distribution Δ is locally generated by k pairwise commuting vector fields. Let us fix a point p in X and let $f_1, \ldots, f_k$ be any vector fields which generate the distribution Δ in a neighborhood of p. Treating f_i as column vectors, we form the $n \times k$ matrix $F = (f_1, \ldots, f_k)$. Note that multiplying F from the right by an invertible $k \times k$ matrix of smooth functions does not change the distribution spanned by the columns of F (it changes its generators, only). By a possible permutation of variables we achieve that the upper $k \times k$ submatrix of the matrix F is nonsingular. Multiplying F from the right by a suitable invertible matrix we obtain that this submatrix is equal to the identity, i.e. the new matrix

F takes the form

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ * & * & \dots & * \\ \vdots & \vdots & & \vdots \\ * & * & \dots & * \end{pmatrix},$$

where "*" denote unknown coefficients. The new vector fields formed by the columns of this matrix commute. In fact, since their first k coefficients are constant, the first k coefficients of any Lie bracket $[f_i, f_j]$ vanish. On the other hand, from involutivity it follows that this Lie bracket is a linear combination of the columns of F. Both these facts can only hold when the coefficients of this linear combination are equal to zero. This shows that the new vector fields commute.

Step 2. Assume that the vector fields $f_1, \dots, f_k$ generate the distribution Δ, locally around p, and they commute. We can choose other $n-k$ vector fields $f_{k+1}, \dots, f_n$ so that $f_1, \dots, f_n$ are linearly independent at p. Define a map Φ by

$$(t_1, \dots, t_n) \longrightarrow \exp(t_1 f_1) \circ \exp(t_2 f_2) \circ \dots \circ \exp(t_n f_n)(p).$$

As the flows of the vector fields $f_1, \dots, f_k$ commute, we see that the order of taking these flows in the above definition can be changed. Therefore, an integral curve of a vector field $e_i = (0, \dots, 1, \dots, 0)^T$, $1 \leq i \leq k$ is transformed to an integral curve of the vector field f_i (as we may place the flow of f_i to the most left place). It follows that the map Φ sends the vector fields $e_1, \dots, e_k$ to the vector fields $f_1, \dots, f_k$ and conversely does the inverse map Φ^{-1}. This inverse map is the desired map which transforms the distribution Δ spanned by $f_1, \dots, f_k$ to the constant distribution spanned by $e_1, \dots, e_k$. ∎

In order to state a global version of this theorem as well as other theorems related to transitivity of families of vector fields and integrability of distributions we need more definitions.

2.2 Submanifolds and foliations

Definition 2.4 A subset $S \subset X$ is called a *regular submanifold* of X of dimension k if for any $x \in S$ there exists a neighborhood U of x and a diffeomorphism $\Phi : U \longrightarrow V \subset \mathbb{R}^n$ onto an open subset V such that

$$\Phi(U \cap S) = \{x = (x_1, \ldots, x_n) \in V \mid x_{k+1} = 0, \ldots, x_n = 0\}$$

(see Figure 10). The regularity class of this submanifold is by definition the regularity class of the diffeomorphism Φ (we shall assume that this regularity is C^∞ or C^ω).

In other words, a regular submanifolds of dimension k is a subset which locally looks like a piece of subspace of dimension k, up to a change of coordinates. A slightly weaker notion of a submanifold is introduced in the following definition.

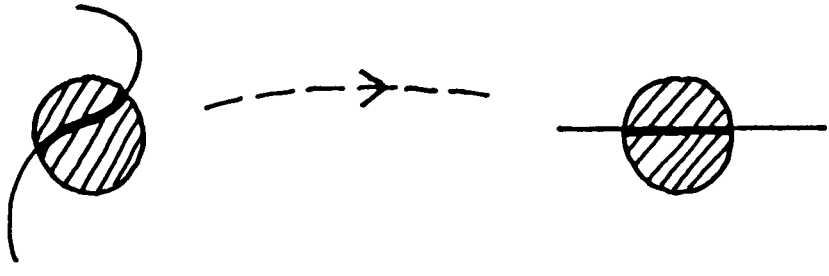


Figure 10

Definition 2.5 We call a subset $S \subset X$ an *immersed submanifold* of X of dimension k if

$$S = \bigcup_{i=1}^{\infty} S_i, \qquad \text{where} \qquad S_1 \subset S_2 \subset S_3 \subset \cdots \subset S$$

and S_i are regular submanifolds of X of dimension k.

In the case when S itself is a regular submanifold we can take $S_i = S$ and so S is also an immersed submanifold.

Example 2.6 In Figure 11 (a) and (b) are regular submanifolds of $\mathbb{R}^2$ while (c) and (d) are only immersed submanifolds.

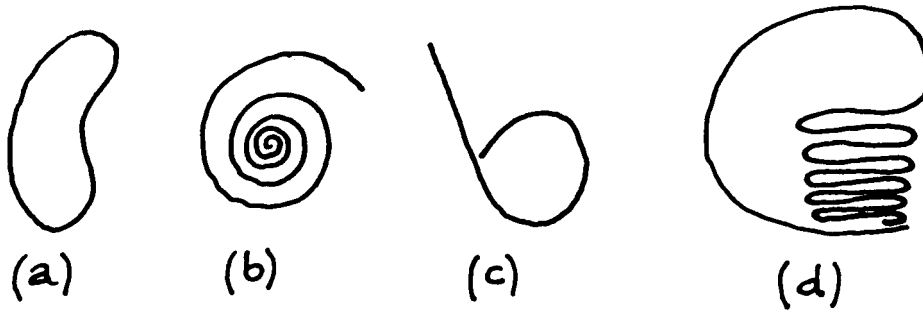


Figure 11

We shall later need two geometric properties of Lie bracket.

Property 1 If two vector fields f, g are tangent to an (immersed) submanifold S then also their Lie bracket $[f, g]$ is tangent to this submanifold.

This follows from the geometric definition of Lie bracket. In fact, if f is tangent to S, then its flow transforms points of S into points of S when the time is sufficiently small. Therefore, the tangent map to the flow $D\gamma_t^f$ transforms the tangent subspaces of S into tangent subspaces of S, in particular, it transforms the tangent vectors $g(p)$ into vectors tangent to S. Moreover, the vectors $v(t) = (\mathrm{Ad}_{\gamma_{-t}^f} g)(p)$ are all in the tangent space T_pS. Taking derivative with respect to t of this expression, which appears in the geometric definition of $[f, g]$, gives a tangent vector to S.

Definition 2.7 A *foliation* $\{S_\alpha\}_{\alpha\in A}$ of X of dimension k is a partition

$$X = \bigcup_{\alpha\in A} S_\alpha$$

of X into disjoint connected (immersed) submanifolds S_α, called *leaves*, which has the following property. For any $x \in X$ there exists a neighborhood U of x and a diffeomorphism $\Phi : U \longrightarrow V \subset \mathbb{R}^n$ onto an open subset V such that

$$\Phi((U \cap S_\alpha)_{cc}) = \{x = (x_1, \ldots, x_n) \in V \mid x_{k+1} = c_\alpha^{k+1}, \ldots, x_n = c_\alpha^n\},$$

where P_{cc} denotes a connected component of the set P and the above property should hold for any such connected component, with the constants c_α^i

depending on the leaf and the choice of the connected component (Figure 12). Similarly as for submanifolds, the regularity of the foliation is defined by the regularity of the diffeomorphism Φ.

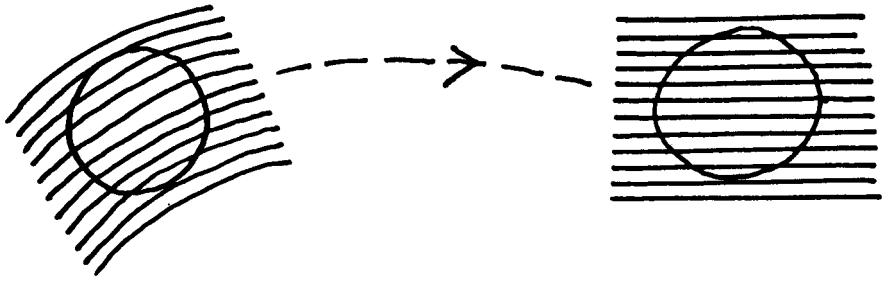


Figure 12

Examples of foliations on subsets of $\mathbb{R}^2$ are presented in Figure 13. A general example of a foliation of dimension $k = n - r$ is given by the following equations for leaves

$$S_\alpha = \{x \in X \mid h_1(x) = c_\alpha^1, \ldots, h_r(x) = c_\alpha^r\},$$

where c_α^i are arbitrary constants and $h = (h_1, \ldots, h_r)$ is a smooth map of constant rank r (i.e. its Jacobi map is of rank r).

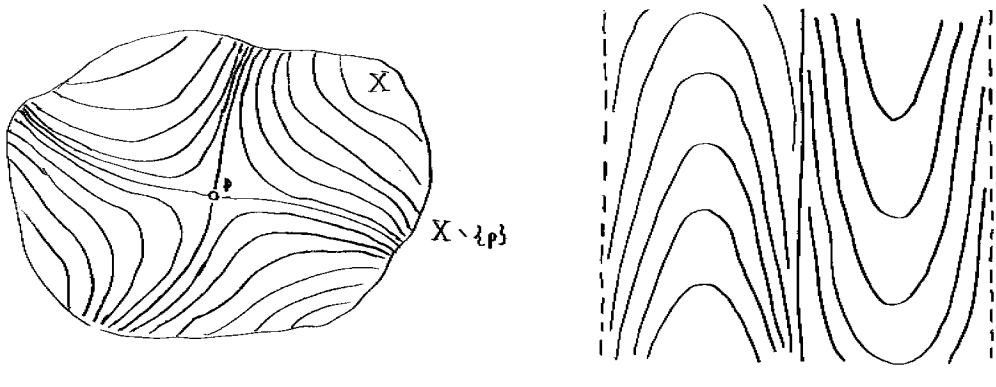


Figure 13

Property 2 Assume that a vector field g is tangent to a foliation $\{S_\alpha\}_{\alpha\in A}$, that is, it is tangent to its leaves. Then, if the flow of another vector field f locally preserves this foliation, the Lie bracket $[f,g]$ is tangent to this foliation.

Here by saying that the flow of f locally preserves the foliation $\{S_\alpha\}_{\alpha\in A}$ we mean that for any point $p\in S_\alpha$ there is a neighborhood U of p such that the image of a piece of a leaf $\gamma_t^f(S_\alpha\cap U)$ is contained in a leaf of the foliation (dependent on t), for any t sufficiently small.

To prove this property let us choose coordinates as in the definition of the foliation and assume that γ_t^f locally preserves $\{S_\alpha\}_{\alpha\in A}$. It follows that the tangent map to γ_t^f maps tangent spaces to leaves into tangent spaces to leaves. Therefore the vector $D\gamma_t^f(p)g(p)$ is tangent to leaves and, in particular, its last $n-k$ components are zero (here we use our special coordinates). Differentiating with respect to t at $t=0$ gives a vector with the last $n-k$ components equal to zero (and so tangent to a leaf), which by the geometric definition of Lie bracket is equal to $[f,g](p)$.

2.3 Orbits of families of vector fields

Consider a family of (global or partial) vector fields $\mathcal{F}=\{f_u\}_{u\in U}$ on X.

Definition 2.8 We define the *orbit of a point* $p\in X$ of this family as the set of points of X reachable from p piecewise by trajectories of vector fields in the family, i.e.

$$\mathrm{Orb}(p)=\{\gamma_{t_k}^{u_k}\circ\cdots\circ\gamma_{t_1}^{u_1}\mid k\geq 1,\ u_1,\ldots,u_k\in U,\ t_1,\ldots,t_k\in\mathbb{R}\},$$

where γ_t^u denotes the flow of the vector field f_u. Of course, if some of our vector fields are not complete then we consider only such $t_1,\ldots,t_k$ for which the above expression has sense.

The relation: "q belongs to the orbit of p" is an equivalence relation on the space X. In fact, a point q belongs to the orbit $\mathrm{Orb}(p)$ if and only if it is reachable from p piecewise by trajectories of the vector fields in the family $\mathcal{F}$. It is evident that q is reachable from p if and only if p is reachable from q (symmetry). Also, if q is reachable from p and r is reachable from q, then r is reachable from p (transitivity).

It follows then that the space X is a disjoint union of orbits (equivalence classes).

Definition 2.9 Let Γ be the smallest distribution on X which contains the vector fields in the family $\mathcal{F}$ (i.e. $f_u(p) \in \Gamma(p)$ for all $u \in U$) and is invariant under any flow γ_t^u, $u \in U$, that is

$$D\gamma_t^u(p)\Gamma(p) \subset \Gamma(\gamma_t^u(p)).$$

for all $p \in X$, $u \in U$ and t for which the above expression is well defined.

Equivalently, we can write the invariance property (using partial vector fields) in the form:

$$g \in \Gamma \Longrightarrow \mathrm{Ad}_{\gamma_t^u} g \in \Gamma, \quad \text{for any } u \in U \text{ and } t \in \mathbb{R}.$$

The following theorem was proved independently by H.J. Sussmann and P. Stefan. We state it here without proof.

Theorem 2.10 (Orbit Theorem) *Each orbit $S = \mathrm{Orb}(p)$ of a family of vector fields $\mathcal{F} = \{f_u\}_{u \in U}$ is an immersed submanifold (of class C^k if the vector fields f_u are of class C^k). Moreover, the tangent space to this submanifold is given by the distribution Γ,*

$$T_p S = \Gamma(p), \quad \textit{for all} \quad p \in X.$$

Corollary 2.11 *If the vector fields f_u are analytic, then the tangent space to the orbit can be computed as*

$$T_p X = L(p) = \{g(p) \mid g \in \mathrm{Lie}\{f_u\}_{u \in U}\},$$

where $\mathrm{Lie}\{f_u\}_{u \in U}$ denotes smallest family of (partial) vector fields which contains the family $\mathcal{F}$ and is closed under taking linear combinations and Lie bracket (this is the Lie algebra of vector fields generated by the family $\mathcal{F} = \{f_u\}_{u \in U}$ in the case when f_u are global vector fields). In the smooth case the following inclusion holds

$$L(p) \subset \Gamma(p).$$

Proof. We shall first prove the inclusion. Using the second form of the invariance property of Γ and the geometric definition of Lie bracket we obtain the following implication

$$g \in \Gamma \Longrightarrow [f_u, g] \in \Gamma.$$

Applying this implication iteratively, we deduce that the left iterated Lie brackets

$$[f_{u_k}, \cdots, [f_{u_2}, f_{u_1}] \cdots]$$

are in Γ. As all iterated Lie brackets are linear combinations of left iterated Lie brackets, it follows that $L(p) \subset \Gamma(p)$ for $p \in X$.

To prove the equality in the analytic case it is enough to use the formula

$$D\gamma_t^u(q) f_v(q) = \sum_{i \geq 0} \frac{(-t)^i}{i!} \mathrm{ad}_{f_u}^i f_v(p), \quad p = \gamma_t^u(q),$$

which shows that transformations of vectors under the tangent maps to flows of f_u can be expressed by taking (infinite) linear combinations of Lie brackets. This implies that $\Gamma(p) \subset L(p)$. ■

Example 2.12 The following system in the plane

$$\begin{aligned} \dot{x}_1 &= u_1 x_1, \quad |u_1| \leq 1, \\ \dot{x}_2 &= u_2 x_2, \quad |u_2| \leq 1, \end{aligned}$$

represented by the family of vector fields

$$f_u = (u_1 x_1, u_2 x_2)^T$$

has four 2-dimensional orbits (the open octants), four 1-dimensional orbits (open half-axes) and one zero dimensional orbit which is the origin.

Example 2.13 The family of three vector fields which represent rotations around the three axes

$$f_1 = (0, x_3, -x_2)^T, \quad f_2 = (x_3, 0, -x_1)^T, \quad f_3 = (x_2, -x_1, 0)^T$$

has a continuum of 2-dimensional orbits which are spheres with the center at the origin and one zero dimensional orbit which is the origin itself. Note that the orbits form a 2-dimensional foliation on the set $X = \mathbb{R}^3 \setminus \{0\}$.

The following example shows that in the nonanalytic case the equality $\Gamma(p) = L(p)$ may not hold.

Example 2.14 Consider the family of the following two C^∞ vector fields in the plane

$$f_1 = (1, 0)^T, \quad f_2 = (0, \phi(x_1))^T,$$

where $\phi(y)$ is a smooth function on $\mathbb{R}$ positive for $y < 0$ (for example $\phi(y) = \exp(1/y)$) and equal to zero for $y \geq 0$. Then the orbit of any point is equal to the whole $\mathbb{R}^2$ and from the orbit theorem it follows that $\dim \Gamma(p) = 2$ for any p. On the other hand, we have that $L(p)$ is spanned by the first vector field only, when $x_1 \geq 0$, so $\dim L(p) = 1$.

Corollary 2.15 (Chow and Rashevskii) *If* $\dim L(p) = n$ *for any* $p \in X$, *then any point of* X *is reachable from any other point piecewise by trajectories of* $\mathcal{F} = \{f_u\}_{u \in U}$ (*allowing positive and negative times*), *i.e.* $\mathrm{Orb}(p) = X$ *for any* p.

Proof. It follows from our assumption and the above corollary that $\Gamma(p)$ is equal to the whole tangent space T_pX for any p. From the orbit theorem it follows then that the orbit of any point is of full dimension, so it is an open subset of X. We conclude that X is a union of disjoint open subsets and, as X is connected, only one of them can be nonempty. Therefore, X consists of a single orbit and any point is reachable from any other point piecewise by trajectories of our family of vector fields. ■

2.4 Integrability of distributions and foliations

The above results, especially the orbit theorem, allow us to give criteria for integrability of distributions and prove some classical theorems.

Definition 2.16 We say that a distribution of constant dimension $p \longrightarrow \Delta(p)$ on X is *integrable* if there exists a foliation $\{S_\alpha\}_{\alpha \in A}$ on X such that for any $p \in X$

$$T_pS = \Delta(p),$$

where S is the leaf passing through p.

Finding the foliation which satisfies the condition of the above definition is usually called integrating this distribution, while the foliation and its leaves are called integral foliation and integral (sub)manifolds of the distribution.

Theorem 2.17 (Global Frobenius theorem) *A smooth distribution of constant dimension* Δ *is integrable if and only if it is involutive. The integral foliation of* Δ *is the partition of* X *into orbits of the family of* (*partial*) *vector fields* $\{g \mid g \in \Delta\}$.

Proof. Assume that our distribution is integrable and choose two vector fields $f, g \in \Delta$ and any point $p \in X$. Then f and g are tangent to the leaf S passing through p, therefore their Lie bracket $[f, g]$ is also tangent to this leaf by Property 1. As this happens for any p, it follows that $[f, g](p) \in T_p S = \Delta(p)$ for all p and so $[f, g] \in \Delta$.

Assume now that our distribution is involutive. Consider the family of partial vector fields $\mathcal{F} = \{f \mid f \in \Delta\}$. We shall prove that the partition of X into orbits of this family gives the desired foliation.

Let $f_1, \ldots, f_k, \in \Delta$ span this distribution in a neighborhood of p. We shall show that Δ is invariant under the flows of the vector fields $f \in \Delta$, that is the distribution Γ in the orbit theorem coincides with Δ. We have to prove that

$$D\gamma_t^f(p)\Delta(p) = \Delta(q), \quad q = \gamma_t^f(p),$$

for $f \in \Delta$. The left-hand side subspace is spanned by the vector fields

$$g_t^i = \mathrm{Ad}_{\gamma_t^f} f_i, \quad i = 1, \ldots, k.$$

From the involutiveness assumption we have that $[f, f_i] = \sum_j \phi_{ij} f_j$. Denote the functions $a_t^{ij} = -\phi_{ij} \circ \gamma_{-t}^f$. From Proposition 1.13 it follows that the spanning vector fields satisfy pointwise the following system of linear differential equations

$$\frac{d}{dt} g_t^i = -\mathrm{Ad}_{\gamma_t^f}[f, f_i] = \sum_j g_t^j a_t^{ij}.$$

As the solution of a linear differential equation depends linearly on its initial conditions, it follows that

$$g_t^i = \sum_j \psi_t^{ij} g_0^j = \sum_j \psi_t^{ij} f_j,$$

where ψ_t^{ij} are functions. Therefore, the subspace $D\gamma_t^f(p)\Delta(p)$ is spanned by the vectors $f_1(p), \ldots, f_k(p)$ and so it is equal to $\Delta(p)$.

It follows from the orbit theorem that Δ gives the tangent space to the orbits and completes the proof.

To complete the proof it is enough to show that the orbits indeed form a foliation of X. This follows immediately from the local version of the Frobenius theorem. In fact, our distribution is constant in appropriate coordinates and so the connected components of intersections of leaves look like in the definition of a foliation. ∎

In order to define integrability of distributions which are not of constant dimension we have to weaken the notion of foliation. We will do this in such a way that partitions of the space X into orbits of a family of vector fields form foliations in this weaker sense.

Definition 2.18 A *foliation with singularities* is a partition

$$X = \bigcup_{\alpha \in A} S_\alpha$$

of X into immersed submanifolds such that, locally, there is a family of vector fields $\{g_\beta\}_{\beta \in B}$ such that $T_p S_\alpha = \text{span}\,\{g_\beta(p) \mid \beta \in B\}$ for all p and α.

A distribution on X is called *integrable* if there exists a foliation with singularities $\{S_\alpha\}_{\alpha \in A}$ which satisfies $T_p S = \Delta(p)$ for any p and S denoting the leaf which passes through p.

Theorem 2.19 (Nagano) *Any analytic involutive distribution Δ is integrable.*

Proof. We take the partition of X into orbits of the family of vector fields $\{f \mid f \in \Delta\}$ as a candidate for the integral foliation. From the orbit theorem and the corollary to it follows that the tangent space to the leaf passing through p is equal to $\Gamma(p) = \mathcal{L}(p) = \Delta(p)$ (involutivity implies that the space of vector fields $\{f | f \in \Delta\} =: \mathcal{F}$ is closed under the Lie bracket and so coincides with $\text{Lie}\{\mathcal{F}\} = \mathcal{L}$. This means that the partition into orbits is the integral foliation of Δ indeed. ∎

Appendix: Global equivalence of families of vector fields

We close this chapter with a proof of sufficiency of the theorem about equivalence of families of vector fields and a global version of this result. The theorem of Nagano will be helpful in this proof.

Proof of Theorem 1.20. *Sufficiency.* In the proof we shall use the method of graph of Cartan and the theorem of Nagano. The method of graph consists of considering the product space $Z = X \times \tilde{X}$ and constructing the graph of the desired diffeomorphism $\Phi : X \longrightarrow \tilde{X}$ as an integral manifold of a distribution of vector fields on Z.

Define the product vector fields on Z by $h_u = f_u \times \tilde{f}_u$, $u \in U$, where, in $\mathbb{R}^n$,

$$f_u \times \tilde{f}_u = (f_u^1, \ldots, f_u^n, \tilde{f}_u^1, \ldots, \tilde{f}_u^n)^T.$$

Consider the distribution spanned by the Lie algebra $\mathrm{Lie}\{H\}$ generated by the family $H = \{h_u\}_{u\in U}$ of these vector fields. Nagano's theorem says that the distribution $Z \ni z \longrightarrow \mathrm{Lie}\{H\}(z)$ is integrable, i.e. for each point $z \in Z$ there is an integral manifold of $\mathrm{Lie}\{H\}$ passing through this point.

Take the point $z_0 = (p, \tilde{p}) \in Z$. We claim that the integral submanifold S passing through z_0 is of dimension n and it is the graph of a local diffeomorphism between X and $\tilde{X}$. Since S is the integral manifold, its dimension is equal to the dimension of the distribution $\mathrm{Lie}\{H\}$ at z_0. But the vectors defining $\mathrm{Lie}\{H\}(z_0)$ are of the form

$$h_{[u_1\cdots u_k]} = (f_{[u_1\cdots u_k]}, \tilde{f}_{[u_1\cdots u_k]}) = (f_{[u_1\cdots u_k]}, Lf_{[u_1\cdots u_k]}),$$

the latter equality following from the assumption. From transitivity of $\mathrm{Lie}\{\mathcal{F}\}$ at p and the above form of the vector fields $h_{[u_1\cdots u_k]}$ it follows that the dimension of $\mathrm{Lie}\{H\}(z_0)$ is at least n. On the other hand, since the second component of these vector fields depends on the first through the same linear map L, it follows that this dimension is precisely equal to n.

It follows that the integral submanifold S is of dimension n. To show that it defines a graph of a local diffeomorphism between X and $\tilde{X}$ we should check that the projections of the tangent space to S onto the tangent spaces of X and $\tilde{X}$ are "onto". From continuity, it is enough to show this at the point z_0. But $T_{z_0}S = \mathrm{Lie}\{H\}(z_0)$ and the "ontoness" follows immediately from the above form of vectors $h_{[u_1\cdots u_k]}$ and the transitivity of $\mathcal{F}$ and $\tilde{\mathcal{F}}$.

Let Φ be the local diffeomorphism from X to $\tilde{X}$ defined in a neighborhood of p via the submanifold S, $\Phi(p) = \tilde{p}$. Since among the vectors tangent to S there are vectors $h_u = (f_u, \tilde{f}_u)$, and S is the graph of Φ, it follows that there is the following relation between the domain component f_u of h_u and its codomain component $\tilde{f}_u$:

$$\tilde{f}_u(\Phi(x)) = D\Phi(x)f_u(x), \qquad \text{or} \qquad \tilde{f}_u(\tilde{x}) = D\Phi(x)f_u(x), \quad x = \Phi^{-1}(\tilde{x}).$$

The latter equality means that $\tilde{f}_u = \mathrm{Ad}_\Phi f_u$, $u \in U$. The proof of sufficiency is complete. ∎

Theorem 2.20 (Sussmann) *Assume that $\mathcal{F}$ and $\tilde{\mathcal{F}}$ are analytic transitive families of vector fields on compact, simply connected, analytic manifolds X and $\tilde{X}$ and the relation between the Lie brackets as in the local theorem holds. Then there exists a global diffeomorphism $\Phi : X \longrightarrow \tilde{X}$ such that $\mathrm{Ad}_\Phi f_u = \tilde{f}_u$, $u \in U$.*

Proof. The proof of this theorem is an extension of the above proof. Namely, it is enough to prove that under our assumptions the map Φ defined above is a global diffeomorphism.

We shall first show that the projection maps from S to X and $\tilde{X}$ are onto (and so they are coverings of X and of $\tilde{X}$, respectively). It is enough to show this for X. Take any point q on X. From the theorem of Chow and Rashevskii it follows that this point is reachable from p piecewise by the trajectories of the vector fields in F. Consider the point z_1 on S which corresponds to q and is reachable from z_0 piecewise by the lifted trajectories of the corresponding vector fields in H. It is easy to see that the projection of z_1 onto X is equal to q. Therefore, S is a covering of X.

As X is simply connected, it follows that this covering is a single covering, i.e. a diffeomorphism of S and X. In a similar way we show that the projection of S onto $\tilde{X}$ is a diffeomorphism. We conclude that the families F and $\tilde{F}$ are diffeomorphic. ∎

3 Controllability and accessibility

3.1 Basic definitions

We shall be dealing with two classes of control systems, the general nonlinear systems

$$\Sigma: \quad \dot{x} = f(x,u),$$

where $x(t) \in X$ and $u(t) \in U$, and the control-affine systems

$$\Sigma_{\text{aff}}: \quad \dot{x} = f(x) + \sum_{i=1}^{m} u_i g_i(x),$$

where $x(t) \in X$ and $u(t) = (u_1(t), \ldots, u_m(t)) \in U$. The state space X is assumed to be an open subset of $\mathbb{R}^n$ or a smooth differential manifold of dimension n. The control set U is an arbitrary set (with at least two elements), in the case of system Σ, and a subset of $\mathbb{R}^m$, in the case of Σ_{aff}. The vector fields $f_u = f(\cdot, u)$, defined by Σ, are assumed to be smooth (of class C^∞). Similarly, we assume that the vector fields $f, g_1, \ldots, g_m$ defined by Σ_{aff} are smooth. We will not need regularity of $f(x,u)$ in Σ with respect to u when we will use piecewise constant controls. Otherwise, we will assume that $f(x,u)$ together with the first partial derivatives with respect to u are smooth as functions of x and continuous with respect to (x,u).

We begin with the formal definition of reachable sets.

Definition 3.1 We shall call the set of points reachable from $p \in X$ for system Σ its *reachable set from* p and denote it by $\mathcal{R}(p)$. For the class of piecewise constant controls this is the set of points

$$\gamma_{t_k}^{u_k} \circ \cdots \circ \gamma_{t_1}^{u_1}(p), \qquad k \geq 1, \quad u_1, \ldots, u_k \in U, \quad t_1, \ldots, t_k \geq 0.$$

Similarly, the set of above points with $t_1 + \cdots + t_k = t$ will be called the *reachable set at time* t *from* p and denoted by $\mathcal{R}_t(p)$, and the set of such points with $t_1 + \cdots + t_k \leq t$ will be referred to as the *reachable set up to time* t *from* p and denoted by $\mathcal{R}_{\leq t}(p)$.

It is unreasonable to expect that the reachable set of a nonlinear control system will have a simple structure, in general. Almost never it will be a linear subspace, even if $X = \mathbb{R}^n$ and $U = \mathbb{R}^m$. For example, for the system in the plane

$$\dot{x}_1 = u_1^2, \quad \dot{x}_2 = u_2^2$$

with $U = \mathbb{R}^2$ the reachable set from the origin is the positive ortant.

Therefore, our aim will be to establish qualitative properties of the reachable sets. One of such basic properties is the following.

Definition 3.2 We shall say that the system Σ is *accessible from* p if its reachable set $\mathcal{R}(p)$ has a nonempty interior. Similarly, we will call this system *strongly accessible from* p if the reachable set $\mathcal{R}_t(p)$ has a nonempty interior for any $t > 0$.

3.2 Taylor linearization

We begin with a presentation of a rough sufficient condition for strong accessibility. Suppose that the set of admissible controls consists of piecewise continuous controls with values in $U \subset \mathbb{R}^m$.

Let (x_0, u_0) be an equilibrium point of our system Σ, i.e. $f(x_0, u_0) = 0$. Assume that f is of class C^1 with respect to (x, u). Denote

$$A(x,u) = \frac{\partial f}{\partial x}(x,u), \quad B(x,u) = \frac{\partial f}{\partial u}(x,u),$$

and let $A_0 = A(x_0, u_0)$, $B_0 = B(x_0, u_0)$.

Theorem 3.3 *If $u_0 \in \operatorname{int} U$ and the pair (A_0, B_0) satisfies the controllability rank condition, then the system is strongly accessible from x_0.*

A corresponding result outside an equilibrium can be stated as follows. Let $u^*(\cdot)$ be an admissible control and let $x^*(\cdot)$ be the corresponding trajectory of system Σ. Denote

$$A(t) = \frac{\partial f}{\partial x}(x^*(t), u^*(t)), \quad B(t) = \frac{\partial f}{\partial u}(x^*(t), u^*(t)).$$

Theorem 3.4 *If $u^*(t) \in \operatorname{int} U$ and the linear system $\dot{x} = A(t)x + B(t)u$, $x(0) = 0$ without constraints on the control is controllable on the interval $[0, T]$, then the reachable set $\mathcal{R}_T(x_0)$ of system Σ has a nonempty interior. In particular, if the Grammian rank condition* $\operatorname{rank} G(0,t) = n$ *is satisfied for our linear system for some $t > 0$ (equivalently, for any $t > 0$), then system Σ is strongly accessible.*

The Grammian matrix used above is defined by

$$G(0,t) = \int_0^t S(\tau)B(\tau)B^T(\tau)S^T(\tau)d\tau,$$

where $S(t)$ is the fundamental solution of $\dot{S}(t) = A(t)S(t)$, $S(0) = I$.

For the proof we shall need the following lemma of the theory of ordinary differential equations, which will be stated without proof.

Let $\bar{u}$ be a measurable, essentially bounded control and consider an admissible control in the form of the following variation

$$u_\epsilon = u^* + \epsilon\bar{u}.$$

Denote by x_ϵ the trajectory of system Σ, $x_\epsilon(0) = x_0$, corresponding to the control u_ϵ. Introduce the variation of the trajectory by

$$\bar{x}(t) = \left.\frac{\partial}{\partial\epsilon}\right|_{\epsilon=0} x_\epsilon(t).$$

Lemma 3.5 *If $f = f(x,u)$ is of class C^1, then the variation of the trajectory satisfies the following equation, called variational equation*

$$\dot{\bar{x}} = A(x^*(t), u^*(t))\bar{x} + B(x^*(t), u^*(t))\bar{u}, \quad \bar{x}(0) = 0.$$

Both above theorems follow from the criteria on controllability of linear systems without constraints (see the contribution of Zabczyk in this volume) and from the following lemma.

Lemma 3.6 *If the variational system (treated as a linear system without constraints on the control) is controllable, then the original system is strongly accessible.*

Proof. Denote the matrices $A(t)$ and $B(t)$ as above. As the variational system is controllable, there exist (bounded) controls v^i which steer this system from 0 to $e_i = (0, \dots, 1, \dots, 0)^T$ (with 1 at i-th place) at time T, $i = 1, \dots, n$. Take the control

$$u = u(\lambda_1, \dots, \lambda_n) = \lambda_1 v^1 + \cdots + \lambda_n v^n.$$

When applied to the original system with the initial condition $x(0) = x_0$ it gives a final state $x(T)$ dependent on the parameters $\lambda = (\lambda_1, \dots, \lambda_n)$ in a differentiable way. In particular, the variation

$$\left.\frac{\partial x(T)}{\partial \lambda_i}\right|_{\lambda=0} = \bar{x}^i$$

satisfies the variational equation with the control $\bar{u} = v^i$. As $\bar{x}^i(T) = e_i$, it follows that the Jacobi map of the nonlinear mapping

$$(\lambda_1, \dots, \lambda_n) \longrightarrow x(T)$$

is of full rank. Therefore, it follows from the inverse function theorem that this mapping maps a neighborhood of the origin onto a neighborhood of the origin. As $u(\lambda_1, \dots, \lambda_n)$ form admissible controls, for λ_i small, it follows that the reachable set $\mathcal{R}_T(x_0)$ contains a neighborhood of the point $x^*(T)$. ∎

3.3 Lie algebras of control system

We shall be using the following families of vector fields associated to the system Σ. Denote

$$f_u = f(\cdot, u),$$

and define the following families of vector fields

$$\mathcal{F} = \{f_u\}_{u \in U}$$

and

$$\mathcal{G} = \{f_u - f_v \mid u, v \in U\}.$$

We define the *Lie algebra of system* Σ as the smallest linear space $\mathcal{L}$ of vector fields on X which contains the family $\mathcal{F}$ and is closed under Lie bracket:

$$f_1,\ f_2 \in \mathcal{L} \implies [f_1, f_2] \in \mathcal{L},$$

or equivalently

$$f_1 \in \mathcal{F},\ f_2 \in \mathcal{L} \implies [f_1, f_2] \in \mathcal{L}.$$

Remark 3.7 Equivalence of both conditions follows by iterative application of the Jacobi identity written in the form

$$\mathrm{ad}_{[g,h]} f = \mathrm{ad}_g \mathrm{ad}_h f - \mathrm{ad}_h \mathrm{ad}_g f$$

and from bilinearity of Lie bracket (cf. Appendix 1, Section 1).

We also define the *Lie ideal of system* Σ as the smallest linear space $\mathcal{L}_0$ of vector fields on X which contains the family $\mathcal{G}$ and is closed under Lie bracket:

$$f_1 \in \mathcal{L},\quad f_2 \in \mathcal{L}_0 \quad \implies \quad [f_1, f_2] \in \mathcal{L},$$

or equivalently

$$f_1 \in \mathcal{F},\quad f_2 \in \mathcal{L}_0 \quad \implies \quad [f_1, f_2] \in \mathcal{L}_0.$$

$\mathcal{L}_0$ is closed under Lie bracket and so is a Lie algebra in the usual sense (cf. Appendix 1, Section 1).

One can see from both definitions that $\mathcal{L}$ and $\mathcal{L}_0$ can be equivalently defined through the iterative Lie brackets as follows

$$\mathcal{L} = \mathrm{span}\,\{[f_{u_1}, \cdots, [f_{u_{k-1}}, f_{u_k}] \cdots] \mid k \geq 1,\ u_1, \cdots, u_k \in U\},$$

$$\mathcal{L}_0 = \mathrm{span}\,\{[f_{u_1}, \cdots, [f_{u_{k-1}}, f_{u_k} - f_{u_{k+1}}] \cdots] \mid k \geq 1,\ u_1, \cdots, u_{k+1} \in U\}.$$

It follows then that

$$\mathcal{L} = \mathrm{span}\,\{f_{u^*},\ \mathcal{L}_0\},$$

where u^* is any fixed element of U. In fact, directly from the definitions we obtain that $\mathcal{L}_0 \subset \mathcal{L}$, and also $f_{u^*} \in \mathcal{L}$. The converse inclusion $\mathcal{L} \subset \mathrm{span}\,\{f_{u^*}, \mathcal{L}_0\}$ follows from the equalities

$$f_{u_1} = f_{u^*} + f_{u_1} - f_{u_2}, \quad u_2 = u^*,$$

$$[f_{u_1}, \cdots, [f_{u_{k-1}}, f_{u_k}] \cdots] = [f_{u_1}, \cdots, [f_{u_{k-1}}, f_{u_k} - f_{u_{k+1}}] \cdots],$$

where $u_{k+1} = u_{k-1}$.

For the control-affine system Σ_{aff} the corresponding Lie algebras can be expressed as

$$\begin{aligned}\mathcal{L} &= \operatorname{Lie}\{f, g_1, \ldots, g_m\} \\ &= \operatorname{span}\{[g_{i_1}, \cdots, [g_{i_{k-1}}, g_{i_k}] \cdots] \mid k \geq 1,\ 0 \leq i_1, \ldots, i_k \leq m\}, \\ \mathcal{L}_0 &= \operatorname{span}\{[g_{i_1}, \cdots, [g_{i_{k-1}}, g_{i_k}] \cdots] \mid k \geq 1,\ 0 \leq i_1, \ldots, i_k \leq m,\ i_k \neq 0\},\end{aligned}$$

where $g_0 = f$.

Remark 3.8 $\mathcal{L}_0$ is a Lie ideal in the Lie algebra $\mathcal{L}$.

Example 3.9 For illustration and also for further use we shall compute the Lie algebra and the Lie ideal of the linear system

$$\dot{x} = Ax + Bu = Ax + \sum_{i=1}^{m} u_i b_i,$$

where b_i are constant vector fields being columns of the matrix B. Taking into account that $g_1 = b_1, \ldots, g_m = b_m$, $f = g_0 = Ax$, and that Lie bracket of constant vector fields is zero, we find that in the above formula for $\mathcal{L}_0$ the only nonzero iterated Lie brackets are

$$[Ax, b_i] = -Ab_i, \quad [Ax, [Ax, b_i]] = [Ax, -Ab_i] = A^2 b_i, \ldots,$$

$$\operatorname{ad}_{Ax} \cdots \operatorname{ad}_{Ax} b_i = \operatorname{ad}_{Ax}^j b_i = (-1)^j A^j b_i.$$

Therefore, the Lie ideal $\mathcal{L}_0$ consists of constant vector fields only,

$$\mathcal{L}_0 = \operatorname{span}\{A^j b_i \mid j \geq 0,\ 1 \leq i \leq m\} = \operatorname{span}\{A^j b_i \mid 0 \leq j \leq n-1,\ 1 \leq i \leq m\},$$

and $\mathcal{L} = \operatorname{span}\{Ax, \mathcal{L}_0\}$, where in the second equality we use the Cayley-Hamilton theorem.

3.4 Accessibility criteria

Given a family of vector fields $\mathcal{H}$, we shall use the notation

$$\mathcal{H}(x) = \operatorname{span}\{h(x) \mid h \in \mathcal{H}\}.$$

In particular, $\mathcal{L}(x)$ and $\mathcal{L}_0(x)$ will denote the space of tangent vectors at x defined by the Lie algebra and the Lie ideal of system Σ. The following result was first proved by Sussmann and Jurdjevic [8].

Theorem 3.10

(a) *If for a state smooth system* Σ *the Lie algebra is of full rank at* x_0, $\dim \mathcal{L}(x_0) = n$, *then the attainable set up to time* t *from* x_0 *has the nonempty interior and so the system is accessible from* x_0.

(b) *If the system is state analytic and* $\dim \mathcal{L}(x_0) < n$, *then the system is not accessible from* x_0.

We present a proof of the first statement (due to A. Krener) which is very simple and gives insight to the problem of accessibility.

Proof of (a). It follows from the assumption $\dim \mathcal{L}(x_0) = n$ that $\dim \mathcal{L}(x) = n$ for x in a neighborhood of x_0 (the full rank is realized by n vector fields which are linearly independent in a neighborhood of x_0). It also follows from the same assumption that there is a $u_1 \in U$ such that $f_{u_1}(x_0) \neq 0$. Otherwise, it would follow from the Jacobian definition of Lie bracket that all the vector fields in $\mathcal{L}$ vanished at x_0 and so $\dim \mathcal{L}(x_0) = 0$. The trajectory $\gamma_{t_1}^{u_1} x_0$, $t \in V_1 = (0, \epsilon_1)$, $\epsilon_1 > 0$, forms a one dimensional submanifold of X which we denote by S_1.

We now claim that there is a $u_2 \in U$ such that the vector fields f_{u_1} and f_{u_2} are linearly independent at a point $x_1 \in S_1$. Otherwise, all the vector fields in $\mathcal{F}$ would be tangent to the submanifold S_1. As taking linear combinations and Lie bracket of vector fields tangent to a submanifold gives vector fields tangent to this submanifold, we would have that all the vector fields in $\mathcal{L}$ were tangent to S_1 which would contradict $\dim \mathcal{L}(x_0) = n$ (if $n > 1$).

Let f_{u_1} and f_{u_2} be linearly independent at $x_1 = \gamma_{t_1}^{u_1} x_0 \in S_1$, $0 < t_1 < \epsilon_1$. Define the map

$$V_2 \ni (t_1, t_2) \longrightarrow x = \gamma_{t_2}^{u_2} \circ \gamma_{t_1}^{u_1}(x_0),$$

where V_2 is an open subset of $\mathbb{R}^2$: $V_2 = (0, \epsilon_1) \times (0, \epsilon_2)$, $\epsilon_2 > 0$. For ϵ_2 sufficiently small the image of this map contains a submanifold of X of dimension 2 (this follows from linear independence of f_{u_1} and f_{u_2}) which we denote by S_2.

By an argument analogous to the above there exists a $u_3 \in U$ and a point $x_2 \in S_2$ such that the vector field f_{u_3} is not tangent to S_2 at x_2. Thus the image of the map

$$V_3 \ni (t_1, t_2, t_3) \longrightarrow x = \gamma_{t_3}^{u_3} \circ \gamma_{t_2}^{u_2} \circ \gamma_{t_1}^{u_1}(x_0)$$

(where $V_3 = (0, \epsilon_1) \times (0, \epsilon_2) \times (0, \epsilon_3)$) contains a submanifold S_3 of X of dimension 3. Of course, S_i, $i = 1, 2, 3$ are subsets of the reachable set.

After n steps of such a construction we obtain a submanifold S_n of X of dimension n, i.e. an open subset of X, which is contained in the reachable set $\mathcal{R}(x_0)$ and, more precisely, in the reachable set $\mathcal{R}_{\leq t}(x_0)$, where $t = \epsilon_1 + \cdots + \epsilon_n$. Since $\epsilon_1, \ldots, \epsilon_n$ could have been taken arbitrarily small, it follows that any attainable set $\mathcal{R}_t$, $t > 0$ has the nonempty interior.

Proof of (b). From the corollary to the orbit theorem it follows that the tangent space to the orbit from x_0 is equal to $\mathcal{L}(x_0)$. When $\dim \mathcal{L}(x_0) < n$, it follows that this orbit is a submanifold of dimension smaller than n. Thus, its interior is empty. As the reachable set is a subset of the orbit, its interior is empty also. ∎

The analyticity assumption in statement (b) cannot be dropped. This can be seen in the example presented after the orbit theorem in Section 2 (showing that in the smooth case we can have $\Gamma(x) \neq \mathcal{L}(x)$) by taking an initial point with positive second coordinate.

If the dimension of the Lie algebra of the system is not full at some point, still we have the following positive result.

Corollary 3.11 *If the system* Π *is state analytic, then the interior in the orbit* $\mathrm{Orb}(x_0)$ *of the reachable set* $\mathcal{R}(x_0)$ *is nonempty.*

Proof. If $\dim \mathcal{L}(x_0) = n$, then this is simply statement (a) of our theorem. When this dimension is smaller we can restrict our system to the orbit passing through the initial point. The corollary to the orbit theorem says that $\dim \mathcal{L}(x_0)$ is equal to the dimension of the orbit. Thus, our system reduced to the orbit satisfies the assumptions of statement (a) of our theorem and our result follows. ∎

Example 3.12 Consider the system with the scalar control $u \in U = \mathbb{R}$

$$\dot{x}_1 = u, \quad \dot{x}_2 = x_1^k, \qquad k \geq 2.$$

It is easy to check that the Taylor linearization of this system, at the equilibrium $x_0 = 0$ and $u_0 = 0$, is not controllable. Our system is control-affine with $f = (0, x_1^k)^T$ and $g = (1, 0)^T$. Then

$$[g, f] = (0, kx_1^{k-1})^T, \quad [g, [g, f]] = (0, k(k-1)x_1^{k-2})^T, \quad \mathrm{ad}_g^k f = (0, k!)^T,$$

and so $\dim \mathcal{L}_0(x) = \dim \mathcal{L}(x) = 2$ for all x, in particular the system is strongly accessible from the origin.

There is an analogous relation between the Lie ideal $\mathcal{L}_0$ and the attainable set at time t which is established by the following theorem.

Theorem 3.13

(a) *If the system is state smooth and* $\dim \mathcal{L}_0(x_0) = n$, *then the attainable set* $\mathcal{R}_t(x_0)$ *has a nonempty interior for any* $t > 0$.

(b) *If* $\dim \mathcal{L}_0(x_0) < n$, *then* $\operatorname{int} \mathcal{R}_t(x_0) = \emptyset$ *for any* $t > 0$.

Example 3.14 Consider the system on $\mathbb{R}^2$

$$\dot{x}_1 = 1, \quad \dot{x}_2 = u\, x_1^2,$$

and take $x_0 = (0,0)$, and $U = \mathbb{R}$. We have

$$\mathcal{F} = \{(1, u\, x_1^2)^T \mid u \in \mathbb{R}\}, \quad \mathcal{G} = \operatorname{span}\{(0, x_1^2)^T\}.$$

The Lie algebra $\mathcal{L}$ contains the vector fields

$$f_1 = (1,0)^T, \quad f_2 = (1, x_1^2)^T, \quad f_3 = [f_1, f_2] = (0, 2x_1)^T, \quad [f_1, f_3] = (0,2)^T.$$

Therefore, $\dim \mathcal{L}(x_0) = 2$ and so the system is accessible from x_0. (Note that one gets the same result if the set U is restricted to two values $U = \{0,1\}$). On the other hand $\mathcal{L}_0(x_0) = \operatorname{span}\{(0,1)^T\}$ and so the interior of the attainable set at time t, $t > 0$, is empty. In fact, it can be proved that the attainable set $\mathcal{R}(x_0)$ is equal to the open right half plain including the origin and the set $\mathcal{R}_t(x_0)$ is equal to the set $x_1 = t$, $x_2 \in \mathbb{R}$.

Example 3.15 *Accessibility of linear systems without constraints.*
As we computed earlier, for autonomous linear system Λ with unconstrained control we have $\mathcal{L}(x) = \operatorname{Im}[B, AB, \ldots, A^{n-1}B]$ and $\mathcal{L}(x) = \operatorname{span}\{Ax, \mathcal{L}(x)\}$. Thus, such a system is strongly accessible from x if and only if the controllability matrix

$$[B, AB, \ldots, A^{n-1}B]$$

is of rank n (such linear systems are called controllable).

Noncontrollable linear system may be accessible from x. This happens when $\dim \mathcal{L}(x) = n$ but $Ax \notin \operatorname{Im}[B, AB, \ldots, A^{n-1}B]$. Then the system is accessible from those x at which Ax is not in the image of the controllability matrix. The system is not accessible from the linear subspace of points at which Ax is in this image (this subspace is the counterimage under A of the image of the controllability matrix).

Exercise Analyse the orbits of the linear system without constrains on the control. Show that this system may have one orbit, three orbits, or a continuum of orbits. (The Kalman decomposition theorem for linear systems is helpful here.)

Example 3.16 *Accessibility of linear systems with constraints.*
Consider now a linear autonomous system with the constraints $u(t) \in U \subset \mathbb{R}^m$. If the interior of U is nonempty, then the controllability rank condition implies that the system is strongly accessible, as we have already established in the section about linear systems. When U has the empty interior then the situation is more complicated. One possibility is to use the principle of convexification. We will use the above theorems on accessibility and strong accessibility. It is more convenient to consider the system in the form

$$\dot{x} = Ax + v, \quad v \in V,$$

where V is the image of U under the linear map $B : \mathbb{R}^m \longrightarrow \mathbb{R}^n$. Let us introduce the set

$$W = \{v' - v'' \mid v', v'' \in V\}.$$

The one can easily compute that the Lie algebra of our system contains the vector fields $Ax + v' - (Ax + v'') = v' - v'' \in W$, i.e. all constant vector fields $f = w$, $w \in W$. Thus, it contains also the Lie brackets $[w, Ax + v] = Aw$, $w \in W$, and by induction it contains all the constant vector fields $A^i w$, $i \geq 0$, $w \in W$. It follows then from the Cayley-Hamilton theorem that the linear system with constraints is strongly accessible from x_0 if and only if

$$\dim \operatorname{span}\{A^i w \mid 0 \leq i \leq n-1, w \in W\} = n.$$

It is accessible from x_0 if and only if the same collection of vectors together with any fixed vector $Ax_0 + v$, $v \in V$, span the whole space.

Example 3.17 *Space-craft with two jets.* Consider a spacecraft with two pairs of jets placed so that they angular momenta are parallel to principal axes of the spacecraft. Then, the equations of motion for the angular velocities take the form

$$\begin{aligned}
\dot{\omega}_1 &= a_1 \omega_2 \omega_3 + u_1, \\
\dot{\omega}_2 &= a_2 \omega_3 \omega_1 + u_2, \\
\dot{\omega}_3 &= a_3 \omega_1 \omega_2.
\end{aligned}$$

Here the constants are given by the principal momenta of inertia: $a_1 = (I_2 - I_3)/I_1$, $a_2 = (I_3 - I_1)/I_2$, and $a_3 = (I_1 - I_2)/I_3$. Our system is control-affine with

$$f = (a_1\omega_2\omega_3, a_2\omega_3\omega_1, a_3\omega_1\omega_2)^T, \quad g_1 = (1,0,0)^T, \quad g_2 = (0,1,0)^T.$$

We compute

$$[f, g_1] = -(0, a_2\omega_3, a_3\omega_2)^T, \quad [f, g_2] = -(a_1\omega_3, 0, a_3\omega_1)^T,$$

$$[g_1, [g_2, f]] = (0, 0, a_3)^T.$$

It follows easily that

$$\dim \mathcal{L}_0(x) = 3 \iff \dim \mathcal{L}(x) = 3 \iff a_3 \neq 0,$$

for any $x = (\omega_1, \omega_2, \omega_3)$. It follows then that the above system is accessible (equivalently, strongly accessible) if and only if the momenta of inertia of the space-craft along the axes with two pairs of jets are different.

Example 3.18 *Space-craft with one jet.* The analysis of the space-craft with one jet, with the equations

$$\begin{aligned} \dot{\omega}_1 &= a_1\omega_2\omega_3 + u, \\ \dot{\omega}_2 &= a_2\omega_3\omega_1, \\ \dot{\omega}_3 &= a_3\omega_1\omega_2, \end{aligned}$$

gives a different result. We have that

$$f = (a_1\omega_2\omega_3, a_2\omega_3\omega_1, a_3\omega_1\omega_2)^T, \quad g = (1,0,0)^T,$$

$$[f, g] = -(0, a_2\omega_3, a_3\omega_2)^T = -(\omega_1)^{-1} f + (\omega_1)^{-1} a_1\omega_2\omega_3 g.$$

Computing the higher order Lie brackets does not give anything new:

$$[g, [f, g]] = 0, \quad [f, [f, g]] = (*, 0, 0)^T = \phi g,$$

where ϕ is a function. It follows that

$$\mathcal{L}(x) = \text{span}\,\{g(x), [f, g](x)\}$$

and these two vector fields span an involutive distribution. From the form of g and $[f, g]$ it follows that the orbits of the system consist of the Cartesian

product of lines along the first coordinate and the trajectories of the vector field $(a_2\omega_3, a_3\omega_2)^T$ along the last two coordinates. In particular, if $a_2 \neq 0 \neq a_3$, then there is one 1-dimensional orbit of the system (the first coordinate axis) corresponding to the equilibrium of the vector field $(a_2\omega_3, a_3\omega_2)^T$, and continuum of 2-dimensional orbits. Our system is not accessible from any $x = (\omega_1, \omega_2, \omega_3)$. We conclude that if there is only one pair of jets which gives the angular momentum parallel to one of the principal axes of inertia of the space-craft then, contrary to the case of two pairs of jets, the system is never accessible.

4 Controllability and path approximation

4.1 Time-reversible systems

In general, the reachable set is a proper subset of the orbit. It is reasonable to ask for which systems the reachable set coincides with the orbit. One class of such systems is called time-reversible systems.

Below we will also consider piecewise continuous controls, as admissible controls. By definition these will be functions $u : [0,T] \to U$ defined on finite intervals $[0,T]$ and continuous, except at a finite number of points, having left and right limits at such points (the set U will be assumed a subset of $\mathbb{R}^m$ or a metric space). We shall say that a function $g(x,u)$ is of class $C^{k,0}$ (respectively, of class $C^{k,\emptyset}$) if U is a metric space and g is continuous as a function of (x,u) together with all partial derivatives with respect to x of order not exceeding k (respectively, U is any set and $g(x,u)$ is of class C^k with respect to x, for any fixed $u \in U$).

Definition 4.1 We will call a system $\Sigma : \dot{x} = f(x,u)$ *time-reversible* if there are a function $U \ni u \to v(u) \in U$ and a positive valued function $\lambda(x,u)$ of class $C^{1,\emptyset}$ such that

$$f(x,u) = -\lambda(x,u) f(x, v(u)) \quad \text{for any} \quad (x,u) \in X \times U.$$

Similarly, Σ is called *feedback time-reversible* if there are functions $(x,u) \to v(x,u) \in U$ and $\lambda(x,u)$ (the latter positive valued) of class $C^{1,0}$ such that

$$f(x,u) = -\lambda(x,u) f(x, v(x,u)) \quad \text{for all} \quad (x,u) \in X \times U.$$

Example 4.2 The system $\dot{x} = \sum_{i=1}^{m} u_i g_i(x)$ is time reversible if the set $U \subset \mathbb{R}^m$ is symmetric with respect to the origin (then we can take $\lambda \equiv 1$ and $v(u) = -u$) or U contains a neighborhood of the origin.

Proposition 4.3 *For any time-reversible system with $f(x,u)$ of class $C^{1,\emptyset}$ and piecewise constant controls (respectively, for any feedback time-reversible system Σ with f of class $C^{1,0}$ and piecewise continuous controls) we have*

$$\mathcal{R}(x_0) = \mathrm{Orb}(x_0).$$

Proof. In the definition of the reachable set it is not allowed to go backward in time along trajectories of the vector fields $f_u = f(\cdot, u)$, contrary to the case of the orbit. This means that, for piecewise constant controls, we have the inclusion $\mathcal{R}(x_0) \subset \mathrm{Orb}(x_0)$ but, possibly, no converse inclusion. For a time-reversible system going backward in time along a trajectory of f_u can be replaced (up to time scale defined by λ) by going forward with the control $v(u)$. Therefore, for a time-reversible system the points which are piecewise forward-backward reachable by trajectories of f_u, $u \in U$, (definition of the orbit) are also forward reachable by such trajectories and the inclusion $\mathcal{R}(x_0) \supset \mathrm{Orb}(x_0)$ follows. The same argument works for proving this inclusion in the case of feedback time-reversible systems, where we use the control $\tilde{u}(t) = v(x(t), u)$ in order to go backwards along the trajectory of f_u. In the case of piecewise continuous controls the inclusion $\mathcal{R}(x_0) \subset \mathrm{Orb}(x_0)$ follows from the following proposition. ∎

Consider the system $\dot{x} = ux$, with $x \in \mathbb{R}$ and $u \in \mathbb{R}$. It has three orbits: the half-lines $(-\infty, 0)$, $(0, \infty)$, and the point $\{0\}$. The (unbounded) control $u(t) = 1/(t-1)$ produces the trajectory $x(t) = t - 1$, $t \in [0, 2]$, starting at $t = 0$ from $x_0 = -1$ in the first orbit and ending up at the point $x(2) = 1$ in the second orbit. This phenomenon cannot occur for piecewise continuous controls (having left and right limits at points of discontinuity).

Namely, the following proposition says that, for piecewise continuous controls, trajectories of Σ starting from $x_0 \in X$ cannot leave the orbit $\mathrm{Orb}(x_0)$.

Proposition 4.4 *If U is a metric space and $f(x,u)$ is of class $C^{1,0}$ then, for piecewise continuous controls, we have the inclusion*

$$\mathcal{R}(x_0) \subset \mathrm{Orb}(x_0).$$

In other words, any trajectory of Σ *corresponding to a piecewise continuous control and starting from* x_0 *stays in* $\mathrm{Orb}(x_0)$ *for all times* t *for which it is defined in* X.

Proof. We shall first prove the following weaker statement. If x_0 is a point in X, $u : [0,T] \to U$ is a piecewise continuous control and $t_0 \in (0,T)$, then there exists a neighborhood I of t_0 in $[0,T]$ such that the trajectory $x(t)$ of Σ corresponding to $u(\cdot)$ satisfying $x(t_0) = x_0$ is well defined, for t in I, and $x(t) \in \mathrm{Orb}(x_0)$ for $t \in I$. To prove this statement notice that $S := \mathrm{Orb}(x_0)$ is a submanifold, by the orbit theorem, and $f(x,u) \in T_x S$ for any $x \in S$ and $u \in U$. This means that the system Σ can be restricted to $S =: \tilde{X}$ (in suitable local coordinates S can be locally identified with an open subset of R^k, where $k = \dim S$). The right-hand side $\tilde{f}(\tilde{x},u)$ of the restricted system is also of class $C^{1,0}$, so we have existence (in S) and uniqueness of solution of the equation $\dot{\tilde{x}} = \tilde{f}(\tilde{x},u)$, with the initial condition $\tilde{x}(t_0) = x_0$. This solution coincides with the unique solution of $\dot{x} = f(x,u)$. This implies our statement.

To prove the proposition assume that the converse holds, i.e., there exists a piecewise continuous control $u : [0,T) \to U$ such that the corresponding trajectory $x(t)$ leaves the orbit $S := \mathrm{Orb}(x_0)$ at time $t^* \in [0,T)$. This means that $[0,t^*)$ is the maximal right-open interval such that $x([0,t^*))$ is contained in the orbit S. Suppose that the point $p^* := x(t^*)$ is in S. Taking $p = p^*$ and $t_0 = t^*$ in the statement proved above we see that $x(t) \in S$ for $t > t^*$ sufficiently close to t^*. This contradicts the maximality of the interval $[0,t^*)$. Suppose thus that $p^* = x(t^*)$ is not in S. Then p^* is in another orbit, namely in $\mathrm{Orb}(p^*)$. Again, we choose p^* as initial point of the trajectory $x(t)$, $x(t^*) = p^*$, corresponding to the original control. Then, by the above statement, the trajectory stays in $\mathrm{Orb}(p^*)$, for some $t < t^*$. Since $S = \mathrm{Orb}(x_0)$ and $\mathrm{Orb}(p^*)$ are disjoint, this contradicts to the fact that $x(t)$ is in S for all $t < t^*$. The proof is complete. ∎

Proposition 4.5 *For any time-reversible system and piecewise constant controls (or feedback time-reversible system and piecewise continuous controls), with the vector fields* f_u *of class* C^∞, *we have*

$$\dim \mathcal{L}(x_0) = n \quad \Longrightarrow \quad x_0 \in \operatorname{int} \mathcal{R}(x_0).$$

Proof. From our theorem on accessibility of systems and the Lie algebra rank condition $\dim \mathcal{L}(x_0) = n$ it follows that the reachable set corresponding to

piecewise constant controls has a nonempty interior. Let x_1 be a point in this interior, contained together with its neighborhood W in the reachable set. Thus

$$x_1 = \gamma_{t_k}^{u_k} \circ \cdots \circ \gamma_{t_1}^{u_1}(x_0),$$

for some $k \geq 1$, $u_1, \ldots, u_k \in U$ and $t_1, \ldots, t_k \in (0, \infty)$, where γ_t^u denotes the flow of f_u. The forward time trajectories of $f_{u_1}, \ldots, f_{u_k}$ can be followed backward using the controls $v_1 = v(u_1), \ldots, v_k = v(u_k)$ (defined in our definition of time-reversible systems), choosing suitable positive times $\tau_1, \ldots, \tau_k$ so that the point

$$x_2 = \gamma_{\tau_1}^{v_1} \circ \cdots \circ \gamma_{\tau_k}^{v_k} \circ \gamma_{t_k}^{u_k} \circ \cdots \circ \gamma_{t_1}^{u_1}(x_0) = \gamma_{\tau_1}^{v_1} \circ \cdots \circ \gamma_{\tau_k}^{v_k}(x_1)$$

coincides with x_0. This point is also in the interior of the reachable set as the composition of flows $\gamma_{\tau_1}^{v_1} \circ \cdots \circ \gamma_{\tau_k}^{v_k}$ is a local diffeomorphisms and maps the neighborhood W of x_1 onto a neighborhood V of $x_2 = x_0$. Since W was contained in the reachable set $\mathcal{R}(x_0)$, V is also contained in $\mathcal{R}(x_0)$. It follows that x_0 lies in the interior of the reachable set from x_0.

In the case of feedback time-reversible systems the proof is similar. In this case the flows of autonomous vector fields f_u corresponding to constant controls have to be replaced by the flows of nonautonomous vector fields $f_{u(t)}$ corresponding to continuous controls $t \to u(t)$. ∎

As a corollary we obtain another proof of the Chow-Rashevskii theorem (this proof is independent of the orbit theorem).

Corollary 4.6 *If the system is time-reversible, X is connected, f_u are of class C^∞ and $\dim \mathcal{L}(x) = n$ for all $x \in X$, then any point of X is forward reachable from any other by piecewise constant controls, i.e. $\mathcal{R}(x) = X$ for any $x \in X$.*

Proof. From Proposition 4.3 it follows that the reachable set $\mathcal{R}(x)$ coincides with the orbit. Moreover, it follows from Proposition 4.5 that $\mathcal{R}(x)$ is open since, after reaching any point, we can also reach a neighborhood of this point. Thus reachable sets coincide with orbits and are open subsets of X. As X is a disjoint union of orbits, it is a disjoint union of open orbits. From connectedness of X it follows that X consists of one orbit. Thus, for any x_0 the orbit of x_0 is equal to X. As the reachable set of x_0 coincides with the orbit, it is also equal to X. ∎

Example 4.7 Our example of the motion of a car (Examples 1.3 and 1.10) gives a time-reversible system if, together with the forward motions given by the vector fields f and g we introduce also backward motions $-f$ and $-g$. It follows from the above result that the reachable set is the whole X, which means that we can reach any position of the car. In fact, a much stronger result can be proved. Namely, the car can "approximately follow" any continuous curve in its state space. This will follow from the main result of the following subsection.

4.2 Approximating curves by trajectories

In this section we will show other controllability properties of the system which are related to its orbits. In particular, we will show that for a time-reversible system any curve lying in a single orbit can be C^0 approximated (up to time reparametrization) by trajectories of the system. We assume that $X \subset \mathbb{R}^n$. The same can be done for X a differential manifold, if we replace the Euclidean distance used below by the distance defined by a Riemannian metric on M.

Consider a continuous curve

$$c : [0,1] \to X.$$

We denote by $\operatorname{Im} c = c([0,1])$ the image of the curve in X and $p_0 := c(0)$.

Definition 4.8 We say that $c(\cdot)$ *can be* C^0 *approximated by trajectories of* Σ if for any $\epsilon > 0$ there exist $T > 0$, an admissible control $u : [0,T] \to U$, and a strictly increasing continuous function $\tau(t)$, $\tau(0) = 0$, $\tau(T) = 1$, such that $x(T, p_0, u(\cdot)) = c(1)$ and

$$\|x(t, p_0, u(\cdot)) - c(\tau(t))\| < \epsilon$$

for all $t \in [0,T]$, where $x(t, p_0, u(\cdot))$ is the trajectory starting at $t = 0$ from p_0 and $\|\cdot\|$ denotes the Euclidean norm.

Relations between the following conditions will be discussed below.

(i) *The image* $\operatorname{Im} c$ *lies in a single orbit of* Σ.
(ii) *The curve* $c : [0,1] \to X$ *can be* C^0 *approximated by trajectories of* Σ.
(iii) *The image* $\operatorname{Im} c$ *lies in the closure in* X *of a single orbit of* Σ.

Theorem 4.9

(a) *For $f(x,u)$ of class $C^{1,0}$ and piecewise continuous controls we have* (ii) $\Rightarrow$ (iii), *for any continuous curve $c:[0,1]\to X$.*

(b) *If Σ is time-reversible, the vector fields $f_u = f(\cdot,u)$, $u\in U$, are analytic and the controls are piecewise constant then* (i) $\Rightarrow$ (ii), *for any absolutely continuous curve $c:[0,1]\to X$. The requirement $f_u\in C^\omega$ can be replaced by $f_u\in C^\infty$ and $T_p\mathrm{Orb}(p) = \mathcal{L}(p)$, for any $p\in \mathrm{Im}\, c$.*

Proof. (a) Assume that $\mathrm{Im}\, c$ does not lie in the closure of a single orbit. Then there exists $s^*\in[0,1]$ such that $p^* := c(s^*)\notin cl(S)$, where $S = \mathrm{Orb}(p_0)$ — the orbit of the point p_0. This means that $\mathrm{dist}(p^*,S) = \epsilon > 0$. However, this inequality implies that the curve c cannot be approximated with accuracy better than ϵ by trajectories starting from p_0 (since all such trajectories stay in S, by Proposition 4.4). This means that (ii) implies (iii).

(b) The implication (i) $\Rightarrow$ (ii) will follow from the Chow-Rashevskii theorem stated in the preceding section. Choose $\epsilon > 0$. We cover $\mathrm{Im}\, c$ with open, connected sets V_i in S, each contained in an ϵ-ball in X with center in $\mathrm{Im}\, c$, such that $V_i\cap S$ are connected. By compactness of $\mathrm{Im}\, c$ we can choose a finite number of such open sets $V_0,\dots,V_r$ ordered in such a way that $p_0 = c(0)\in V_0$, $p_{r+1} := c(1)\in V_r$, and $V_{i-1}\cap V_i\cap \mathrm{Im}\, c\neq\emptyset$, for $i = 1,\dots,r$ (this is possible by connectedness of $\mathrm{Im}\, c$ and openess of V_i). Let us choose some points $p_i = c(s_i)$ in $V_{i-1}\cap V_i\cap \mathrm{Im}\, c$, $i = 1,\dots,r$, so that $0 =: s_0 < s_1 < \cdots < s_r \le s_{r+1} := 1$. From the assumption $f_u\in C^\omega$ and the orbit theorem it follows that $T_pS = \mathcal{L}(p)$, for any $p\in S$. Therefore, the system Σ restricted to the open subset V_i of the orbit S satisfies the Lie algebra rank condition $\dim\mathcal{L}(p) = \dim\tilde{X}$, where $\tilde{X} = V_i$. It follows from Corollary 4.6 that the point $p_i\in V_i$ can be joined to $p_{i+1}\in V_i$, with a trajectory not leaving V_i. Each point on this trajectory is at a distance not larger than 2ϵ from any point of the piece $c([s_i,s_{i+1}])$ of $\mathrm{Im}\, c$ (since V_i has the diameter not larger than 2ϵ and $c([s_i,s_{i+1}])$ is contained in V_i by the assumptions that $p_i = c(s_i)\in V_i$, $p_{i+1} = c(s_{i+1})\in V_i$ and $V_i\cap\mathrm{Im}\, c$ being connected). Concatenating the consecutive trajectories joining $p_0 = c(0)$ to p_1 in V_0, then p_1 to p_2 in V_1 etc., and finally p_r to $p_{r+1} = c(1)$ in V_r we obtain a trajectory which approximates c with accuracy 2ϵ (we can define the reparametrization of c, which appears in the definition of C^0 approximation, as continuous piecewise linear function $[0,T]\ni t\to s(t)\in[0,1]$ which is linear on the intervals $[T_{i-1},T_i]$ corresponding to the trajectories joining p_{i-1} to p_i and satisfies $s(T_i) = s_i$, $i = 0,\dots,r+1$). As ϵ was chosen

arbitrarily, we see that (ii) holds.

Note that, instead of using analyticity of f_u in the above proof (which implies $T_pS = \mathcal{L}(p)$, for $p \in S$) it is enough to use the assumption that $T_pS = \mathcal{L}(p)$ for $p \in \operatorname{Im} c$. Namely, this assumption implies the equalities $T_{p_i}S = \mathcal{L}(p_i)$, for $i = 0, \ldots, p_{r+1}$, which in turn imply analogous equalities in neighborhoods of p_i, so that we can assume that $T_pS = \mathcal{L}(p)$ holds on V_i. (The equality $T_{p_i}S = \mathcal{L}(p_i)$ implies the analogous equality in a neighborhood in S of p_i. This follows from two facts: (a) we always have $\mathcal{L}(p) \subset T_pS$, for $p \in S$; (b) the equality $\dim \mathcal{L}(p_i) = \dim S = k$ extends to a neighborhood in S of p_i, since $\mathcal{L}(p_i)$ is spanned by some k linearly independent vector fields, which remain linearly independent in a neighborhood of p_i.) ■

Example 4.10 Condition (ii) does not imply (i). An example of such a system Σ is the system with the state space X equal to the 2-dimensional torus $T^2 = \mathbb{R}^2/Z^2$, with all vector fields f_u equal to the same "constant" vector field with the "slope" irrational. Then the orbit of any point $p_0 \in T^2$ is a one dimensional immersed submanifold which is dense in T^2. Its closure is the whole of T^2, however, any curve which is transversal to the orbit cannot be C^0 approximated by the trajectories of Σ.

The conditions (i) and (iii) in the above theorem may be difficult to check. However, given an absolutely continuous curve $c : [0,1] \to X$, the following sufficient condition for C^0 approximation by trajectories is checkable.

(iv) *There exists a neighborhood W of* $\operatorname{Im} c$ *in* X *such that* $\dim \mathcal{L}(p) =$ const, *for* $p \in W$, *and*

$$\frac{dc}{ds}(s) \in \mathcal{L}(c(s)) \qquad \text{for almost all } \ s \in [0,1].$$

Theorem 4.11

(a) *If f_u, $u \in U$, are of class C^∞ then* (iv) $\Rightarrow$ (i), *for any absolutely continuous curve $c : [0,1] \to X$.*

(b) *If Σ is time-reversible, the vector fields $f_u = f(\cdot, u)$, $u \in U$, are analytic and the controls are piecewise constant then* (iv) $\Rightarrow$ (ii), *for any absolutely continuous curve $c : [0,1] \to X$. The requirement $f_u \in C^\omega$ can be replaced by $f_u \in C^\infty$ and $T_p\operatorname{Orb}(p) = \mathcal{L}(p)$, for any $p \in \operatorname{Im} c$.*

Proof. (a) We have to show that $\operatorname{Im} c$ is contained in $S = \operatorname{Orb}(p_0)$. Assume that the contrary holds and let s^* be the infimum of $s \in [0,1]$ such that $c(s) \notin S$. Define $p^* = c(s^*)$. Since $c(s) \in S$ for all $s < s^*$, we have

$p^* \in cl(S)$, by continuity of the curve. Consider a neighborhood V of p^* in which $\dim \mathcal{L}(p) = \text{const} = k$. Then the distribution $p \to \mathcal{L}(p)$ is of constant dimension on V and involutive (since $\mathcal{L} = \text{Lie}\{f_u\}_{u \in U}$ is a Lie algebra). Applying the local version of the Frobenius theorem we can assume, after a change of coordinates, that in a neighborhood V_1 of p^* we have $\mathcal{L}(p) = \text{span}\{e_1, \ldots, e_k\}$, where e_i denotes the i-th coordinate versor. This implies that the k-submanifold of V_1 defined by $\{x \in V_1 : x_{k+1} = p^*_{k+1}, \ldots, x_n = p^*_n\}$, with p^*_i-coordinates of p^*, is contained in $\text{Orb}(p^*)$ (by the Chow-Rashevskii theorem applied to the system restricted to V_1). Since the vectors in $\mathcal{L}(p)$ have zero components along the last $n - k$ versors, from the assumption $(dc/ds)(s) \in \mathcal{L}(c(s))$ it follows that the last $n - k$ coordinates of the curve c are constant, equal to the coordinates of p^*, for s sufficiently close to s^*. This implies that $c(s)$ is in $S = \text{Orb}(p^*) = \text{Orb}(p_0)$, for $s \geq s^*$ close to s^*. But this contradicts our definition of s^*, thus (iv) implies (i).

(b) From statement (a) it follows that (iv) implies (i). Using statement (b) of Theorem 4.9 we see that (i) implies (ii). (Note that in the proof of the latter implication we have shown existence of controls giving approximating trajectories, but the controls were not constructed.) ■

The following result shows that, for any analytic, time reversible system, a point $p \in X$, and a given vector $v \in \mathcal{L}(p)$, there is a piecewise constant control producing an infinitesimal movement of the state "in the direction v" from p.

Theorem 4.12 *If Σ is time reversible and the vector fields f_u, $u \in U$, are smooth then, for any $p \in X$ and $v \in \mathcal{L}(p)$, there exists a 1-parameter family of piecewise constant controls $u_\epsilon(t)$ such that for the corresponding trajectory $x_\epsilon(t)$ starting from p we have*

$$x_\epsilon(T(\epsilon)) = p + \epsilon v + O(\epsilon^{1+1/N})$$

for $\epsilon > 0$, where $T(\epsilon)$ depends continuously on ϵ and $T(\epsilon) \to 0$, if $\epsilon \to 0$. The constant N is the smallest integer k such that v is spanned by the vector fields f_u, $u \in U$, and their Lie brackets up to order k, evaluated at p.

The following corollary is an immediate consequence of the theorem.

Corollary 4.13 *Under the assumptions of the above theorem the set of vectors at $p \in X$ tangent to the curves of ends of 1-parameter families of trajectories $[0, T(\epsilon)] \to X$ of Σ, starting from p, coincides with $\mathcal{L}(p)$.*

If we drop the assumption that the system is time reversible, the set $K(p)$ of vectors tangent to curves of ends of 1-parameter families of trajectories is, in general, a proper subset of $\mathcal{L}(p)$. Its explicit description is of basic importance. We state it as open problem.

A research problem

(a) *Prove that $K(p)$ is a cone (not difficult).*

(b) *Describe $K(p)$ explicitly, assuming that the vector fields f_u of the system are analytic. (Lie bracket should play a basic role.) Consider the case $X = \mathbb{R}^2$ or $\mathbb{R}^3$.*

(c) *Find conditions for path approximation analogous to conditions in Theorems 4.9 and 4.11.*

The proof of Theorem 4.12 will follow from the following two basic propositions. The proofs of these propositions should give some insight to the above problem.

Let $\Phi^1_\epsilon, \ldots, \Phi^k_\epsilon$ be families of diffeomorphisms of X which are of class C^r with respect to (x, ϵ), $r \geq 1$, and are defined for ϵ close to 0. Assume that

$$\Phi^1_0 = \mathrm{id}, \ \ldots, \ \Phi^k_0 = \mathrm{id}.$$

For example, we can take as Φ^i_ϵ the flow of a time dependent vector field, with ϵ playing the role of time. (In fact, it will be enough to assume less, namely that Φ^i_ϵ are partial diffeomorphisms of X, i.e., each Φ^i_ϵ is a diffeomorphism of an open subset of X onto an open subset of X and the set of (x, ϵ) for which Φ^i_ϵ is defined is open in $X \times \mathbb{R}$ and contains $X \times \{0\}$.) Note that composition of such families gives a family of diffeomorphisms satisfying the same conditions.

From $\Phi^i_0 = \mathrm{id}$ it follows that $\epsilon \to \Phi^i_\epsilon(p)$ is a C^r curve passing through p, for each i and $p \in X$. Thus we can define

$$f_1(p) := \frac{\partial \Phi^1_\epsilon}{\partial \epsilon}(p)\Big|_{\epsilon=0}, \ \ldots\ldots \ , f_k(p) := \frac{\partial \Phi^k_\epsilon}{\partial \epsilon}(p)\Big|_{\epsilon=0},$$

where $f_1, \ldots, f_k$ are vector fields on X of class C^{r-1}.

Proposition 4.14 *For any constants $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ we have*

$$\frac{\partial}{\partial \epsilon}\, \Phi^1_{\lambda_1 \epsilon} \circ \cdots \circ \Phi^k_{\lambda_k \epsilon}(p)\Big|_{\epsilon=0} = \lambda_1 f_1(p) + \cdots + \lambda_k f_k(p).$$

The above formula can be equivalently written in the form

$$\Phi^1_{\lambda_1\epsilon} \circ \cdots \circ \Phi^k_{\lambda_k\epsilon}(p) = p + \epsilon V(p) + O(\epsilon^2),$$

where

$$V(p) = \lambda_1 f_1(p) + \cdots + \lambda_k f_k(p).$$

This means that the composition of such diffeomorphisms gives infinitesimal movement along the vector $\lambda_1 f_1(p) + \cdots + \lambda_k f_k(p)$. In particular, if we start with vector fields $f_{u_1}, \ldots, f_{u_k}$ and define $\Phi^i_\epsilon = \exp(\epsilon f_{u_i})$ — the flows of f_{u_i}, then we get $f_i = f_{u_i}$.

Now we want to construct a family of diffeomorphisms which gives infinitesimal movement along the iterated Lie bracket of the vector fields $f_1, \ldots, f_k$. Given two diffeomorphisms Φ and Ψ of X, we define their commutator as the diffeomorphism

$$[\Phi, \Psi] = \Phi^{-1} \circ \Psi^{-1} \circ \Phi \circ \Psi.$$

If Θ is another diffeomorphism of X and we denote $\chi = [\Phi, \Psi]$, we define the third order commutator

$$\begin{aligned}[[\Phi, \Psi], \Theta] &= \chi^{-1} \circ \Theta^{-1} \circ \chi \circ \Theta \\ &= \Psi^{-1} \circ \Phi^{-1} \circ \Psi \circ \Phi \circ \Theta^{-1} \circ \Phi^{-1} \circ \Psi^{-1} \circ \Phi \circ \Psi \circ \Theta.\end{aligned}$$

Analogously we can define higher order commutators of diffeomorphisms and, in particular, commutators of our families of diffeomorphisms $\Phi^1_\epsilon, \ldots, \Phi^k_\epsilon$. The infinitesimal vector field corresponding to the iterated commutator of such families of diffeomorphisms appears to be equal to the iterated Lie bracket of the vector fields $f_1, \ldots, f_k$. Note that if one of the diffeomorphisms is equal to identity then the commutator is also equal to identity.

Proposition 4.15 *If Φ^i_ϵ are of class C^r with respect to (x, ϵ) and $r \geq k+1$, then*

$$\left(\frac{\partial}{\partial \epsilon}\right)^j [\cdots [\Phi^1_\epsilon, \Phi^2_\epsilon], \cdots, \Phi^k_\epsilon](p)\Big|_{\epsilon=0} = 0, \qquad \text{for } 1 \leq j < k,$$

and

$$\left(\frac{\partial}{\partial \epsilon}\right)^k [\cdots [\Phi^1_\epsilon, \Phi^2_\epsilon], \cdots, \Phi^k_\epsilon](p)\Big|_{\epsilon=0} = k!\,[f_k, \cdots, [f_2, f_1]\cdots](p),$$

where the commutator of vector fields $f_1, \ldots, f_k$ is of class C^{r-k}.

The above equalities are equivalent to the Taylor formula

$$[\cdots[\Phi^1_\epsilon,\Phi^2_\epsilon],\cdots,\Phi^k_\epsilon](p)=p+\epsilon^k V(p)+O(\epsilon^{k+1}),$$

and, after reparametrization,

$$[\cdots[\Phi^1_{\epsilon^{1/k}},\Phi^2_{\epsilon^{1/k}}],\cdots,\Phi^k_{\epsilon^{1/k}}](p)=p+\epsilon V(p)+O(\epsilon^{1+1/k}),$$

where

$$V(p)=[f_k,\cdots,[f_2,f_1]\cdots](p).$$

This means that the commutator of such diffeomorphisms gives infinitesimal movement along the vector $[f_k,\cdots,[f_2,f_1]\cdots](p)$. (If we define $\Phi^i_\epsilon=\exp(\epsilon f_{u_i})$ — the flows of f_{u_i}, then we get $f_i=f_{u_i}$ and this infinitesimal movement is along the iterated Lie bracket of the vectors fields $f_{u_1},\ldots,f_{u_k}$ defined by a control system Σ.)

Proof of Theorem 4.12. Since $v\in\mathcal{L}(p)$, where $\mathcal{L}=\operatorname{Lie}\{f_u\}_{u\in U}$, we can write

$$v=\lambda_1 v_1+\cdots+\lambda_r v_r,$$

where λ_i are real constants, $v_i=V_i(p)$, and V_i are some of the vector fields f_u, $u\in U$, and their iterated Lie brackets. Since any iterated Lie bracket is equal to a linear combination of left iterated Lie brackets (see Proposition 1.19), after possibly rearranging the above sum we can assume that

$$V_i=[f_{u^i_{k(i)}},\cdots,[f_{u^i_2},f_{u^i_1}]\cdots]$$

for $i=1,\ldots,r$. We define the families of diffeomorphisms as iterated commutators of the flows,

$$\Phi^i_\epsilon:=[\cdots[\exp(\epsilon f_{u^i_1}),\exp(\epsilon f_{u^i_2})],\cdots,\exp(\epsilon f_{u^i_{k(i)}})].$$

Define

$$y(\epsilon)=\Phi^1_{\lambda_1\epsilon^{1/k(1)}}\circ\cdots\circ\Phi^r_{\lambda_1\epsilon^{1/k(r)}}(p).$$

Taking the derivative of $dy/d\epsilon$ at $\epsilon=0$ and using Propositions 4.14 and 4.15 and the definitions of Φ^i_ϵ we obtain the formula

$$\frac{dy}{d\epsilon}(0)=\lambda_1 V_1(p)+\cdots+\lambda_r V_r(p).$$

It remains to show that in the above construction of the 1-parameter family of points $y(\epsilon)$ we can use true forward-time trajectories of system Σ.

Notice that the coefficients $\lambda_1, \ldots, \lambda_r$ in the linear combination which gives v can be taken positive. In fact, if λ_i is negative then we can change the order of the vector fields $f_{u_1^i}$ and $f_{u_2^i}$ and then V_i and λ_i change signs. In the definition of the diffeomorphisms Φ_ϵ^i we use commutators of flows, where we apply the flows $\exp(-\epsilon f_u)$, with $\epsilon > 0$. We can replace such transformations by "time-forward movements" by using the control $v = v(u)$ given by the definition of time reversible system (if the function $\lambda(x, u)$ in this definition is not constant, the portion of time needed for obtaining the equivalent of the transformation $\exp(-\epsilon f_u)$ varies with trajectories). In this way we can replace all time-backward steps by time-forward steps. This shows the main formula in the theorem. The other statements are easy to see by our construction. ■

Proof of Proposition 4.14. Define the function $h(\epsilon) = f(s_1(\epsilon), \ldots, s_k(\epsilon))$ where

$$f(s_1, \ldots, s_k) = \Phi_{s_1}^1 \circ \cdots \circ \Phi_{s_k}^k(p)$$

and $s_i(\epsilon) = \lambda_i \epsilon$. Then the equality $f(0, \ldots, s_i, \ldots, 0) = \Phi_{s_i}^i(p)$ and the chain rule give

$$\frac{d}{d\epsilon} h(0) = \sum_{i=1}^k \frac{\partial f}{\partial s_i}(0, \ldots, 0)\lambda_i = \sum_{i=1}^k \lambda_i \frac{\partial \Phi_{s_i}^i}{\partial s_i}(p)|_{s_i=0} = \lambda_1 f_1(p) + \cdots + \lambda_k f_k(p)$$

which proves the proposition. ■

Proof of Proposition 4.15. Consider the function $h(\epsilon) = f(s_1(\epsilon), \ldots, s_k(\epsilon))$ where

$$f(s_1, \ldots, s_k) = [\cdots[\Phi_{s_1}^1, \Phi_{s_2}^2], \cdots, \Phi_{s_k}^k](p)$$

and $s_i(\epsilon) = \epsilon$. Then the chain rule gives

$$\left(\frac{d}{d\epsilon}\right)^j h(0) = \sum_{j_1 + \cdots + j_k = j} \frac{j!}{j_1! \cdots j_k!} \left(\frac{\partial}{\partial s_1}\right)^{j_1} \cdots \left(\frac{\partial}{\partial s_k}\right)^{j_k} f(0, \ldots, 0).$$

Since the iterated commutator is equal to identity, if one of the diffeomorphisms $\Phi_{s_i}^i$ is identity, any term in the above sum is equal to zero if $j_i = 0$, for some i. This implies that the derivative is equal to 0 if $j < k$, which shows the first equality in the proposition.

If $j = k$ then only one term, with all $j_i \neq 0$, can be nonzero and we get

$$\begin{aligned} \frac{d^k}{d\epsilon^k} h(0) &= k! \frac{\partial}{\partial s_1} \cdots \frac{\partial}{\partial s_k} f(0, \ldots, 0) \\ &= k! \frac{\partial}{\partial s_1} \cdots \frac{\partial}{\partial s_k} [\cdots[\Phi_{s_1}^1, \Phi_{s_2}^2], \cdots, \Phi_{s_k}^k](p)\Big|_{s_1 = \cdots = s_k = 0}. \end{aligned}$$

The above expression gives the second equality in Proposition 4.15 by the following formula (we denote $\Phi_* f = \mathrm{Ad}_\Phi f$, see Section 1.4).

Proposition 4.16

$$\frac{\partial}{\partial s_k}\cdots\frac{\partial}{\partial s_1}\,[\cdots[\Phi^1_{s_1},\Phi^2_{s_2}],\cdots,\Phi^k_{s_k}](p)\Big|_{s_1=\cdots=s_k=0}$$
$$=\frac{\partial}{\partial s_k}\cdots\frac{\partial}{\partial s_2}\,((\Phi^k_{s_k})^{-1}_*\cdots(\Phi^2_{s_2})^{-1}_* f_1)(p)\Big|_{s_2=\cdots=s_k=0}=[f_k,\cdots,[f_2,f_1]\cdots](p).$$

Proof. The second equality follows by induction from the formula

$$\frac{\partial}{\partial s_m}\,((\Phi^k_{s_k})^{-1}_*\cdots(\Phi^m_{s_m})^{-1}_*[f_{m-1},\cdots,[f_2,f_1]\cdots])(p)\Big|_{s_m=0}$$
$$=((\Phi^k_{s_k})^{-1}_*\cdots(\Phi^{m+1}_{s_{m+1}})^{-1}_*[f_m,\cdots,[f_2,f_1]\cdots])(p),$$

which is a consequence of $(\partial/\partial s)(\Phi^m_s)^{-1}_* g=[f_m,g]$. The latter formula can be verified directly in the same way as formula (3) in Section 1.4.

In order to check the first equality we use induction with respect to k. Denote $\Psi_{\bar s}:=[\cdots[\Phi^1_{s_1},\Phi^2_{s_2}],\cdots,\Phi^{k-1}_{s_{k-1}}]$, where $\bar s=(s_1,\ldots,s_{k-1})$. Then, at $s_1=\cdots=s_k=0$,

$$\frac{\partial}{\partial s_k}\cdots\frac{\partial}{\partial s_1}[\Psi_{\bar s},\Phi^k_{s_k}](p)=\frac{\partial}{\partial s_k}\cdots\frac{\partial}{\partial s_1}(\Psi_{\bar s})^{-1}\circ(\Phi^k_{s_k})^{-1}\circ\Psi_{\bar s}\circ\Phi^k_{s_k}(p)$$
$$=\frac{\partial}{\partial s_k}\cdots\frac{\partial}{\partial s_1}(\Phi^k_{s_k})^{-1}\circ\Psi_{\bar s}\circ\Phi^k_{s_k}(p)=\frac{\partial}{\partial s_k}\cdots\frac{\partial}{\partial s_2}((\Phi^k_{s_k})^{-1}_*\frac{\partial}{\partial s_1}\Psi_{\bar s})(p),$$

where the middle equality follows from the fact that $\Psi_{\bar s}|_{s_1=0}=\mathrm{id}$, so differentiating with respect to s_1 appearing in $(\Psi_{\bar s})^{-1}$ gives a term independent of s_k and its derivative $\partial/\partial s_k$ vanishing. The derivatives $\partial/\partial s_i$, $i=2,\ldots,k-1$, commute with $(\Phi^k_{s_k})^{-1}_*$, thus the required equality follows from the inductive formula, at $s_1=\cdots=s_{k-1}=0$, $\partial/\partial s_{k-1}\cdots\partial/\partial s_1\Psi_{\bar s}(p)=\partial/\partial s_{k-1}\cdots\partial/\partial s_2((\Phi^{k-1}_{s_{k-1}})^{-1}_*\cdots(\Phi^2_{s_2})^{-1}_* f_1)(p)$. ∎

Exercises

Exercise 1 For vector fields

$$f_1=x_3\frac{\partial}{\partial x_2}-x_2\frac{\partial}{\partial x_3},\quad f_2=x_1\frac{\partial}{\partial x_3}-x_3\frac{\partial}{\partial x_1},\quad f_3=x_2\frac{\partial}{\partial x_1}-x_1\frac{\partial}{\partial x_2},$$

on $\mathbb{R}^3$, with $x=(x_1,x_2,x_3)$, show that:

(a) the flow of f_1 is the rotation around the first axis:

$$\gamma_t^{f_1}(x) = (x_1, x_2 \cos t + x_3 \sin t, -x_2 \sin t + x_3 \cos t)^T;$$

(b) the Lie bracket $[f_1, f_2]$ is equal to f_3.

Exercise 2 Using a result stated in lecture notes justify the property: the motion along the vector field f_3 in Problem 1 can be approximated by a composition of motions along the vector fields f_1 and f_2. How this composition should be chosen? (This property can be stated as follows: "the sequence ..... of small rotations along the x_1-axis and x_2-axis produces, approximately, a rotation along the x_3-axis".)

Exercise 3 Show that the following system is accessible, but not strongly accessible, at any point $p \in \mathbb{R}^2$ different from the origin:

$$\dot{x}_1 = x_2 + ux_1, \qquad \dot{x}_2 = -x_1 + ux_2,$$

where the control set $U = \mathbb{R}$. Find the reachable sets $\mathcal{R}(p)$ and $\mathcal{R}_t(p)$.

Exercise 4 Consider three vector fields on $\mathbb{R}^3$ given in coordinates by

$$f = (x_2, -x_1, 0)^T, \quad g = (0, x_3, x_2)^T, \quad h = (x_3, 0, x_1)^T.$$

(a) Compute the Lie brackets $[f, g]$, $[f, h]$, $[g, h]$ and show that the Lie algebra $Lie\{f, g, h\}$ generated by f, g, h is a linear subspace of dimension 3, in the linear space of all smooth vector fields on $\mathbb{R}^3$. Show that it spans a subspace $\mathcal{L}(x)$ of dimension 2 of tangent vectors at any point $x \neq 0$.

(b) Show that the orbits of this family of 3 vector fields are hiperboloids $x_1^2 + x_2^2 - x_3^2 = \text{const}$ (or cones of revolution) and they are all of dimension 2, except of one orbit (which one?). Show that the partition of $X = \mathbb{R}^3 \setminus \{0\}$ into orbits forms a foliation of X of dimension 2.

Exercise 5 Solve the research problem given in the text of Lecture 4.

For further reading the reader is referred to the contributions of Kawski, Respondek and Agrachev in this volume. The reader may also find useful the textbooks [3] and [5], as well as the collections of expository papers [7] and [4]. A brief account of problems and results in nonlinear geometric theory is given in [2].

References

[1] William M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry* (Academic Press, New York, 1975).

[2] R. W. Brockett, "Nonlinear control theory and differential geometry", in *Proc. Int. Congress of Mathematicians, Warszawa 1983*, Vol. II, eds. Z. Ciesielski and C. Olech (Polish Scientific Publishers, Warszawa, 1984), pp. 1357–1368.

[3] Alberto Isidori, *Nonlinear Control Systems* (Springer-Verlag, London, 1995).

[4] B. Jakubczyk and W. Respondek eds., *Geometry of Feedback and Optimal Control* (Marcel Dekker, New York, 1998).

[5] Eduardo D. Sontag, *Mathematical Control Theory* (Springer-Verlag, New York, 1998).

[6] Michael Spivak, *Differential Geometry*, Vol. 1 (Publish and Perish Inc., Berkeley, 1997).

[7] H.J. Sussmann ed., *Nonlinear Controllability and Optimal Control* (Marcel Dekker, New York, 1990).

[8] H.J. Sussmann and V. Jurdjevic, "Controllability of nonlinear systems", J. Diff. Equations **12**, 95 (1972).

Introduction to Geometric Nonlinear Control; Linearization, Observability, Decoupling

Witold Respondek*

Laboratoire de Mathématiques, INSA de Rouen, France

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028004

*wresp@insa-rouen.fr

Abstract

These notes are devoted to the problems of linearization, observability, and decoupling of nonlinear control systems. Together with notes of Bronislaw Jakubczyk in the same volume, they form an introduction to geometric methods in nonlinear control theory. In the first part we discuss equivalence of control systems. We consider various aspects of the problem: state-space and feedback equivalence, local and global equivalence, equivalence to linear and partially linear systems. In the second part we present the notion of observability and give a geometric rank condition for local observability and an algebraic characterization of local observability. We discuss uniform observability, decompositions of nonobservable systems, and properties of generic observable systems. In the third part we introduce the notion of invariant distributions and discuss disturbance decoupling and input-output decoupling. Many concepts and results are illustrated with examples.

Contents

1 Introduction **173**

2 Feedback linearization: an introduction **173**

3 Equivalence of control systems **181**
3.1 State space equivalence 181
3.2 Feedback equivalence 184

4 Feedback linearization **186**
4.1 Static feedback linearization 186
4.2 Restricted feedback linearization 192
4.3 Partial linearization 194

5 Observability **196**
5.1 Nonlinear observability 196
5.2 Local decompositions 203
5.3 Uniform observability 205
5.4 Local observability: a necessary and sufficient condition . . . 206
5.5 Generic observability properties 207

6 Decoupling **210**
6.1 Invariant distributions 211
6.2 Disturbance decoupling 212
6.3 Input-output decoupling 214

References **219**

1 Introduction

These notes, together with notes of Jakubczyk [26] of the same volume, form an introduction to geometric nonlinear control. Section 2 has an elementary and introductory character. We formulate the problem of feedback linearization, show why the concept of Lie bracket appears naturally, and give necessary and sufficient conditions for feedback linearization in the single-input case. In Section 3 we introduce two concepts of equivalence of control systems: state space equivalence and feedback equivalence. We also state a result that any nonlinear control system is (locally) determined by iterated Lie brackets of vector fields corresponding to constant controls. In Section 4 we discuss various aspects of the feedback linearization problem. In particular, we consider the multi-input as well as non control-affine systems and the problems of global feedback linearization, restricted feedback linearization, and partial linearization. Section 5 is concerned with the concept of observability. We introduce observability rank condition and then discuss Kalman-like decomposition of nonlinear non observable systems, uniform observability, and generic properties of observable systems. Finally, in Section 6 we introduce te concept of invariant and controlled invariant distributions and, based on it, discuss solutions to the disturbance decoupling and input-output decoupling problems.

We do not provide proofs of the presented results and send the reader to the literature on geometric control theory (see the list of references) and, in particular, to monographs [18], [23], [29], [37]. As a small "recompense", we illustrate many notions, concepts, and results by simple, mainly mechanical, examples.

2 Feedback linearization: an introduction

The aim of this preliminary section is to introduce the concept of feedback linearization and a fundamental geometric tool of nonlinear control theory, which is the Lie bracket. Feedback linearization is a procedure of transforming a nonlinear system into the simplest possible form, that is, into a linear system. Necessary and sufficient conditions for this to be possible will be expressed using the notion of Lie bracket, which is omnipresent in very many nonlinear control problems.

The problem of feedback linearization is to transform the nonlinear con-

trol system

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

into a linear system of the form

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u}$$

via a diffeomorhism

$$(\tilde{x}, \tilde{u}) = (\Phi(x), \Psi(x, u)),$$

called feedback transformation. We will start with an introductory example.

Example 2.1 Consider a nonlinear pendulum (rigid one-link manipulator) consisting of a mass m with control torque u.

The evolution of the pendulum is described by the Euler-Lagrange equation with external force

$$ml^2\ddot{\theta} + mgl\sin\theta = u \ .$$

We rewrite it as

$$\begin{aligned} \dot{\theta} &= \omega \\ \dot{\omega} &= -\tfrac{g}{l}\sin\theta + \tfrac{u}{ml^2}. \end{aligned}$$

Denote $x_1 = \theta$ and $x_2 = \omega$ and consider the evolution of the pendulum on the state space $\mathbb{R}^2$, that is $x = (x_1, x_2)^T \in \mathbb{R}^2$. We get the system Σ

$$\Sigma: \quad \begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\tfrac{g}{l}\sin x_1 + \tfrac{u}{ml^2}. \end{aligned}$$

Replace the control u by

$$u = ml^2\tilde{u} + mlg\sin x_1,$$

which can be interpreted as a transformation in the control space U depending on the state $x \in X$. We get the linear control system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \tilde{u}. \end{aligned}$$

Using a simple transformation in the control space we thus brought the system into the simplest possible form: a linear one. Notice that the families of all trajectories of both systems coincide although they are parametrized (with respect to the control parameters u and $\tilde{u}$, respectively) in two different ways.

Now fix an angle θ_0. The goal is to stabilize the system around $x_0 = (x_{10}, x_{20})^T$, where $x_{10} = \theta_0$ and $x_{20} = 0$. Introduce new coordinates

$$\begin{aligned} \tilde{x}_1 &= x_1 - x_{10} \\ \tilde{x}_2 &= x_2. \end{aligned}$$

and apply the control

$$\tilde{u} = k_1 \tilde{x}_1 + k_2 \tilde{x}_2,$$

where k_1, k_2 are real parameters to be chosen. We get a *closed loop system* described by the system of linear differential equations

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{x}_2 \\ \dot{\tilde{x}}_2 &= k_1 \tilde{x}_1 + k_2 \tilde{x}_2, \end{aligned}$$

whose characteristic polynomial is given by

$$p(\lambda) = \lambda^2 - \lambda k_2 - k_1.$$

Let $\lambda_1, \lambda_2 \in \mathbb{C}$ be any pair of conjugated complex numbers. Take

$$\begin{aligned} k_1 &= -\lambda_1 \lambda_2 \\ k_2 &= \lambda_1 + \lambda_2, \end{aligned}$$

then the eigenvalues of the closed loop system are λ_1 and λ_2. In particular, by choosing λ_1 and λ_2 in the left half plane we stabilize exponentially the pendulum around an arbitrary angle θ_0 and a stabilizing control can be taken as

$$u = k_1 ml^2 (x_1 - x_{10}) + k_2 ml^2 x_2 + mgl \sin x_1.$$

Now fix for the system Σ an initial point $x_0 = (x_{10}, x_{20})^T \in \mathbb{R}^2$ and a terminal point $x_T = (x_{1T}, x_{2T})^T \in \mathbb{R}^2$ and consider the problem of finding a control $u(t)$, $0 \leq t \leq T$, which generates a trajectory $x(t)$, $0 \leq t \leq T$, such that $x(0) = x_0$ and $x(T) = x_T$. This is the *controllability problem*, called also *motion planning problem*. Due to the above described linearization, we get the following simple solution of the problem. Choose an arbitrary C^2-function $\varphi(t)$, $0 \leq t \leq T$, such that

$$\begin{aligned} \varphi(0) &= x_{10} \\ \varphi'(0) &= x_{20} \\ \varphi(T) &= x_{1T} \\ \varphi'(T) &= x_{2T}. \end{aligned}$$

and apply to the system the control

$$\tilde{u}(t) = \varphi''(t)$$

or, equivalently,

$$u(t) = ml^2\varphi''(t) + mlg \sin x_1(t).$$

Clearly, the proposed control solves the motion planning problem producing a trajectory that joins x_0 and x_T. □

Now consider a single-input linear control system of the form

$$\Lambda : \ \dot{x} = Ax + bu,$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}$ and assume that Λ is controllable, that is

$$\text{rank } \big(b, Ab, \ldots, A^{n-1}b\big) = n.$$

Choose a linear function $h = cx$, where c is a row vector, such that

$$cb = cAb = \cdots = cA^{n-2}b = 0$$

and

$$cA^{n-1}b = d \neq 0,$$

whose existence follows immediately from the controllability assumption. Introduce linear coordinates

$$\begin{array}{rcl} \tilde{x}_1 &=& cx \\ \tilde{x}_2 &=& cAx \\ & \vdots & \\ \tilde{x}_n &=& = cA^{n-1}x. \end{array}$$

We have

$$\begin{array}{rcccccl} \dot{\tilde{x}}_1 &=& c\dot{x} &=& cAx + cbu &=& \tilde{x}_2 \\ \dot{\tilde{x}}_2 &=& cA\dot{x} &=& cA^2x + cAbu &=& \tilde{x}_3 \\ & \vdots & & & & & \\ \dot{\tilde{x}}_{n-1} &=& cA^{n-2}\dot{x} &=& cA^{n-1}x + cA^{n-2}bu &=& \tilde{x}_n \\ \dot{\tilde{x}}_n &=& cA^{n-1}\dot{x} &=& cA^nx + cA^{n-1}bu &=& \sum_{i=1}^n a_i\tilde{x}_i + du, \end{array}$$

for some $a_i \in \mathbb{R}$, for $1 \leq i \leq n$. By introducing a new control variable

$$\bar{u} = \sum_{i=1}^n a_i\tilde{x}_i + du,$$

which can be viewed at as a state depending transformation in the control space U, we bring any single-input controllable linear system into the n-fold integrator

$$\dot{\tilde{x}}_1 = \tilde{x}_2,\ \dot{\tilde{x}}_2 = \tilde{x}_3,\ \ldots,\ \dot{\tilde{x}}_{n-1} = \tilde{x}_n,\ \dot{\tilde{x}}_n = \tilde{u}.$$

We will consider the problem of whether and when such a transformation is possible in the nonlinear case. Consider a single-input control affine system of the form

$$\Sigma:\ \dot{x} = f(x) + g(x)u,$$

where $x \in X$, an open subset of $\mathbb{R}^n$, and f and g are C^∞-smooth vector fields on X.

Recall that $L_v\varphi$ denotes the derivative of a function φ with respect to a vector field v, that is

$$L_v\varphi(x) = \sum_{i=1}^{n} \frac{\partial \varphi}{\partial x_i}(x) v_i(x).$$

Fix a point $x_0 \in X$ and assume that there exist a C^∞-smooth function φ on X such that (compare the linear case)

$$L_g\varphi = L_g L_f \varphi = \cdots = L_g L_f^{n-2}\varphi = 0$$

and

$$L_g L_f^{n-1}\varphi(x) = d(x),$$

where $d(x)$ is a smooth function such that $d(x_0) \neq 0$. If around the point x_0, the functions $\varphi, L_f\varphi, \ldots, L_f^{n-1}\varphi$ are independent (in the sense that their differentials are linearly independent around x_0), then in a neighborhood V of x_0 the map

$$\begin{aligned} \tilde{x}_1 &= \varphi \\ \tilde{x}_2 &= L_f\varphi \\ &\vdots \\ \tilde{x}_n &= L_f^{n-1}\varphi \end{aligned}$$

defines a local diffeomorphism, or, in other words, a local coordinate system. In the local coordinates $(\tilde{x}_1, \ldots, \tilde{x}_n)^T$ we have

$$\begin{array}{rclcll} \dot{\tilde{x}}_1 & = & < d\varphi, \dot{x} > & = & L_f\varphi + uL_g\varphi & = \tilde{x}_2 \\ & \vdots & & \vdots & & \vdots \\ \dot{\tilde{x}}_{n-1} & = & < dL_f^{n-2}\varphi, \dot{x} > & = & L_f^{n-1}\varphi + uL_gL_f^{n-2}\varphi & = \tilde{x}_n \\ \dot{\tilde{x}}_n & = & < dL_f^{n-1}\varphi, \dot{x} > & = & L_f^{n}\varphi + uL_gL_f^{n-1}\varphi & = L_f^n\varphi + ud(x). \end{array}$$

By introducing a new control variable

$$\tilde{u} = L_f^n \varphi + u L_g L_f^{n-1} \varphi,$$

which can be viewed at as a transformation in the control space U, depending nonlinearly on the state x, we bring our single-input nonlinear system into the n-fold integrator

$$\dot{\tilde{x}}_1 = \tilde{x}_2,\ \dot{\tilde{x}}_2 = \tilde{x}_3,\ \ldots,\ \dot{\tilde{x}}_{n-1} = \tilde{x}_n,\ \dot{\tilde{x}}_n = \tilde{u}.$$

The proposed method works under two assumptions. Firstly, we assumed the existence of a function φ such that $L_g\varphi = L_g L_f \varphi = \cdots = L_g L_f^{n-2}\varphi = 0$. Secondly, we assumed that the functions $\varphi, L_f\varphi, \ldots, L_f^{n-1}\varphi$ are independent in a neighborhood of x_0. The former is a system of $n-1$ first order partial differential equations. In order to see it, let us consider the two first equations $L_g\varphi = 0$ and $L_g L_f \varphi = 0$, which imply that

$$L_f L_g \varphi - L_g L_f \varphi = 0.$$

Although the expression on the left hand side involves a priori partial derivatives of order two, it depends on partial derivatives of φ of order one only and a direct calculation shows that we can represent it as

$$L_f L_g \varphi - L_g L_f \varphi = L_{[f,g]} \varphi,$$

where the vector field $[f,g]$ is given by

$$[f,g](x) = \mathrm{D}g(x) f(x) - \mathrm{D}f(x) g(x),$$

where $\mathrm{D}g(x)$ (resp. $\mathrm{D}f(x)$) stands for the derivative at x, that is, the Jacobi matrix of the map $g : X \longrightarrow \mathbb{R}^n$ (resp. $f : X \longrightarrow \mathbb{R}^n$). We will call $[f,g]$ the *Lie bracket* of the vector fields f and g. We would like to emphasize two important aspects of the nature of Lie bracket. Firstly, it is a vector field, because if we change coordinates then the Lie bracket is multiplied on the left by the Jacobi matrix of the derivative of the coordinate change. This shows its vector, i.e., contravariant, nature. Secondly, a Lie bracket $[f,g]$ acts on a function φ by the formula $L_{[f,g]}\varphi$, that is, acts as a first order differential operator. Notice that, as we have already said, the expression $L_f L_g \varphi - L_g L_f \varphi$ involves, a priori, second order derivatives of φ but all of them are mixed partials that mutually cancel due to Schwarz lemma.

Introduce the notation

$$\mathrm{ad}_f g = [f, g]$$

and, inductively,

$$\mathrm{ad}_f^{j+1} g = [f, \mathrm{ad}_f^j g],$$

for any integer $j \geq 1$. Put $\mathrm{ad}_f^0 g = g$. It can be shown by an induction argument that the existence of a function φ such that $L_g\varphi = L_g L_f \varphi = \cdots = L_g L_f^{n-2}\varphi = 0$ is equivalent to the solvability of the following system of first order partial differential equations

$$\left\{ \begin{array}{rcl} L_g\varphi & = & 0 \\ L_{\mathrm{ad}_f g}\varphi & = & 0 \\ & \vdots & \\ L_{\mathrm{ad}_f^{n-2} g}\varphi & = & 0, \end{array} \right. \tag{2.1}$$

which in coordinates is expressed as

$$\sum_{i=1}^{n} \frac{\partial \varphi}{\partial x_i} (\mathrm{ad}_f^j g)_i = 0, \quad \text{for } 0 \leq j \leq n-2,$$

where $(\mathrm{ad}_f^j g)_i$ denotes the i-th component, in the coordinates $(x_1, \ldots, x_n)^T$, of the vector field $\mathrm{ad}_f^j g$.

It can be shown that the requirement that the differentials $dL_f^j\varphi$, for $0 \leq j \leq n-1$, where φ is a nontrivial solution of the system (2.1), are linearly independent at x_0 is equivalent to the linear independence of $\mathrm{ad}_f^j g$ at x_0, for $0 \leq j \leq n-1$.

We will show that a necessary condition for the above system of first order PDE's to admit a nontrivial solution is that for any $0 \leq i, j \leq n-2$ the Lie bracket $[\mathrm{ad}_f^i g, \mathrm{ad}_f^j g](x)$ belongs to the linear space generated by $\{\mathrm{ad}_f^q g(x),\ 0 \leq q \leq n-2\}$. In view of the linear independence of the $\mathrm{ad}_f^q g$'s, this is equivalent to the existence of smooth functions α_q^{ij} such that

$$[\mathrm{ad}_f^i g, \mathrm{ad}_f^j g] = \sum_{q=0}^{n-2} \alpha_q^{ij} \mathrm{ad}_f^q g.$$

To prove it, assume that there exists a vector field v of the form $v = [\mathrm{ad}_f^i g, \mathrm{ad}_f^j g]$, for some $0 \leq i, j \leq n-2$, and a point $x \in X$, such that $v(x) \notin \mathrm{span}\,\{\mathrm{ad}_f^q g(x),\ 0 \leq q \leq n-2\}$. We have

$$L_v\varphi = L_{[\mathrm{ad}_f^i g, \mathrm{ad}_f^j g]}\varphi = L_{\mathrm{ad}_f^i g} L_{\mathrm{ad}_f^j g}\varphi - L_{\mathrm{ad}_f^j g} L_{\mathrm{ad}_f^i g}\varphi = 0.$$

The n vector fields v and $\mathrm{ad}_f^q g$, for $0 \leq q \leq n-2$, are linearly independent in a neighborhood of $x \in X$ and therefore the only solutions of the system of n first order PDE's

$$\begin{cases} L_v \varphi &= 0 \\ L_{\mathrm{ad}_f^j g} \varphi &= 0, \quad \text{for } 0 \leq j \leq n-2, \end{cases}$$

are $\varphi =$constant.

It turns out that the two above necessary conditions are also sufficient for the solvability of the problem. Indeed, we have the following result.

Theorem 2.2 *There exist a local change of coordinates $\tilde{x} = \phi(x)$ and a feedback of the form $u = \alpha(x) + \beta(x)\tilde{u}$, where $\beta(x) \neq 0$, transforming, locally around $x_0 \in X$, the nonlinear system*

$$\Sigma : \ \dot{x} = f(x) + g(x)u$$

into a linear controllable system of the form

$$\Lambda : \ \dot{\tilde{x}} = A\tilde{x} + b\tilde{u}$$

if and only if the system Σ satisfies in a neighborhood of x_0:

(C1) *$g(x), \mathrm{ad}_f g(x), \ldots, \mathrm{ad}_f^{n-1} g(x)$ are linearly independent;*

(C2) *for any $0 \leq i, j \leq n-2$, there exist smooth functions α_q^{ij} such that*

$$[\mathrm{ad}_f^i g, \mathrm{ad}_f^j g] = \sum_{q=0}^{n-2} \alpha_q^{ij} \mathrm{ad}_f^q g.$$

The condition (C2), called *involutivity*, is discussed in the general context in the section devoted to Frobenius theorem of [26] in this volume and in the context of feedback linearization in Section 4. It has a clear geometric interpretation. If the above defined system of PDE's $L_g \varphi = \cdots = L_{\mathrm{ad}_f^{n-2} g} \varphi = 0$ admits a nontrivial solution then for any constant $c \in \mathbb{R}$ the equation $\varphi = c$ defines a hypersurface in X. The vectors $g(x), \mathrm{ad}_f g(x), \ldots, \mathrm{ad}_f^{n-2} g(x)$ form at any $x \in \{\varphi(x) = c\}$ the tangent space to that hypersurface. In general, such a hypersurface need not exist; the involutivity condition (C2) guarantees its existence.

Especially simple is the planar case, that is, $n = 2$, in which the involutivity follows automatically from the linear independence condition.

Corollary 2.3 *A control-affine planar system*

$$\dot{x} = f(x) + g(x)u,$$

where $x \in \mathbb{R}^2$*, is locally feedback linearizable at* x_0 *if and only if* g *and* $\mathrm{ad}_f g$ *are independent at* x_0*.*

Example 2.4 (Example 2.1 cont.) We have $f = x_2 \frac{\partial}{\partial x_1}$ and $g = \frac{1}{ml^2}\frac{\partial}{\partial x_2}$. Thus the vector fields g and $\mathrm{ad}_f g = -\frac{1}{ml^2}\frac{\partial}{\partial x_1}$ are independent and hence, by Corollary 2.3, we can conclude feedback linearization of the pendulum, a property which we have established by a direct calculation in Example 2.1. □

3 Equivalence of control systems

The question of feedback linearization discussed in Section 2 is a subproblem of a more general problem of feedback equivalence. In this section we study equivalence of control systems. We start with state space equivalence in Section 3.1 and then we define feedback equivalence in Section 3.2. Various aspects of the problem of feedback linearization will be discussed in Section 4.

3.1 State space equivalence

Two systems are state-space equivalent if they are related by a diffeomorphism (and then also their trajectories, corresponding to the same controls, are related by that diffeomorphism). A question of particular interest is that of when a nonlinear system is equivalent to a linear one. If this is the case the nonlinearities of the considered system are not intrinsic, they appear because of a "wrong" choice of coordinates, and the nonlinear system shares all properties of its linear equivalent.

Consider a smooth nonlinear control system of the form

$$\Sigma \; : \quad \dot{x} = f(x, u),$$

where $x \in X$, an open subset of $\mathbb{R}^n$ (or an n-dimensional manifold) and $u \in U$, an open subset of $\mathbb{R}^m$ (or an m-dimensional manifold). The class of admissible controls $\mathcal{U}$ is fixed and $\mathcal{PC} \subset \mathcal{U} \subset \mathcal{M}$, where $\mathcal{PC}$ denotes the class of piece-wise constant controls with values in U and $\mathcal{M}$ the class of measurable controls with values in U.

Consider another control system of the same form with the same control space U and the same class of admissible controls $\mathcal{U}$

$$\tilde{\Sigma} \; : \quad \dot{\tilde{x}} = \tilde{f}(\tilde{x}, u),$$

where $\tilde{x} \in \tilde{X}$, an open subset of $\mathbb{R}^n$ (or an n-dimensional manifold) and $u \in U$. Analogously to the transformation $\Phi_* g$ of a vector field $g(\cdot)$ by a diffeomorphism Φ, we define the transformation of $f(\cdot, u)$ by Φ. Put

$$(\Phi_* f)(\tilde{p}, u) = \mathrm{D}\Phi(\Phi^{-1}(\tilde{p})) \cdot f(\Phi^{-1}(\tilde{p}), u).$$

We say that control systems Σ and $\tilde{\Sigma}$ are *state space equivalent* (respectively, *locally state space equivalent at points p and $\tilde{p}$*) if there exists a diffeomorphism $\Phi : X \to \tilde{X}$ (respectively, a local diffeomorphism $\Phi : X_0 \to \tilde{X}$, $\Phi(p) = \tilde{p}$, where X_0 is a neighborhood of p) such that

$$\Phi_* f = \tilde{f} \; .$$

Put

$$\mathcal{F} = \{f_u \mid u \in U\} \quad \text{and} \quad \tilde{\mathcal{F}} = \{\tilde{f}_u \mid u \in U\} \; ,$$

where $f_u = f(\cdot, u)$ and $\tilde{f}_u = \tilde{f}(\cdot, u)$, that is, $\mathcal{F}$ (resp. $\tilde{\mathcal{F}}$) stands for the family of all vector fields corresponding to constant controls of Σ (resp. of $\tilde{\Sigma}$). (Local) state space equivalence of Σ and $\tilde{\Sigma}$ means simply that

$$\Phi_* f_u = \tilde{f}_u \quad \text{for any } u \in U,$$

i.e., that Φ establishes a correspondence between vector fields defined by constant controls.

Recall the notion of the Lie algebra $\mathcal{L}$ of the system, see the section on controllability and accessibility of [26] in this volume. Assume $\dim \mathcal{L}(p) = \dim \tilde{\mathcal{L}}(\tilde{p}) = n$, which implies that Σ and $\tilde{\Sigma}$ are accessible at p and $\tilde{p}$, respectively.

The following observation shows that (local) state space equivalence is very natural.

Proposition 3.1 *Σ and $\tilde{\Sigma}$ are (locally) state space equivalent if and only if there exists a (local) diffeomorphism Φ which (locally, in neighborhoods of p and $\tilde{p}$) preserves trajectories corresponding to the same controls $u(\cdot) \in \mathcal{U}$, i.e.,*

$$\Phi(\gamma_t^u(p)) = \tilde{\gamma}_t^u(\tilde{p})$$

for any $u(\cdot) \in \mathcal{U}$ and any t for which both sides exist, where $\gamma_t^u(p)$ (resp. $\tilde{\gamma}_t^u(\tilde{p})$) denotes the trajectory of Σ (resp. $\tilde{\Sigma}$) corresponding to the control function $u(\cdot) \in \mathcal{U}$ and passing by p (resp. by $\tilde{p}$) for $t = 0$.

Introduce the following notation for left iterated Lie brackets

$$f_{[u_1 u_2 \ldots u_k]} = [f_{u_1}, [f_{u_2}, \ldots, [f_{u_{k-1}}, f_{u_k}] \ldots]]$$

and analogous for the tilded family. In particular $f_{[u_1]} = f_{u_1}$.

The following result was established by Krener [32] (see also Sussmann [42]).

Theorem 3.2 *Assume that the systems Σ and $\tilde{\Sigma}$ are analytic and that* $\dim \mathcal{L}(p) = n$ *and* $\dim \tilde{\mathcal{L}}(\tilde{p}) = \tilde{n}$.

(i) *Σ and $\tilde{\Sigma}$ are locally equivalent at p and $\tilde{p}$ if and only if there exists a linear isomorphism of the tangent spaces $F : T_pX \to T_{\tilde{p}}\tilde{X}$ such that*

$$F f_{[u_1 u_2 \ldots u_k]}(p) = \tilde{f}_{[u_1 u_2 \ldots u_k]}(\tilde{p}), \tag{3.1}$$

for any $k \geq 1$ and any $u_1, \ldots u_k \in U$.

(ii) *Assume, moreover, that X and $\tilde{X}$ are simply connected and that the Lie algebras $\mathcal{L}$ and $\tilde{\mathcal{L}}$ of Σ and $\tilde{\Sigma}$, respectively, consist of complete vector fields and satisfy Lie rank condition everywhere. If there exist points $p \in X$ and $\tilde{p} \in \tilde{X}$ and a linear isomorphism $F : T_pX \to T_{\tilde{p}}\tilde{X}$ satisfying* (3.1) *then Σ and $\tilde{\Sigma}$ are state space equivalent.*

This theorem shows that all information concerning (local) behavior is contained in the values at the initial condition of Lie brackets from $\mathcal{L}$. In a sense (iterative) Lie brackets form invariant (higher order) derivatives of the dynamics of the system and in the analytic case they completely determine its local properties as (higher order) derivatives do for analytic functions.

Consider a control-affine system of the form

$$\Sigma_{\text{aff}} : \quad \dot{x} = f(x) + \sum_{i=1}^{m} g_i(x) u_i.$$

Denote $g_0 = f$. Using the above theorem we obtain the following linearization result (compare [38], [42]).

Proposition 3.3 *Consider a control-affine analytic system Σ_{aff}.*

(i) *The system* $\Sigma_{\rm aff}$ *is locally state space equivalent at* $p \in X$ *to a linear controllable system of the form*

$$\Lambda_{\rm c}: \quad \dot{x} = Ax + c + Bu = Ax + c + \sum_{i=1}^{m} b_i u_i \ , \quad x \in \mathbb{R}^n \ , \quad u \in \mathbb{R}^m,$$

at $x_0 \in \mathbb{R}^n$ *if and only if*
(E1) $[g_{i_1}, [g_{i_2}, \ldots [g_{i_{k-1}}, g_{i_k}] \ldots]](p) = 0$
for any $k \geq 2$ *and any* $0 \leq i_j \leq m$, $1 \leq j \leq k$, *provided that at least two* i_j*'s are different from zero and*
(E2) $\dim \operatorname{span} \{\mathrm{ad}_f^j g_i(p) \mid 1 \leq i \leq m \ , \ 0 \leq j \leq n-1\}(p) = n.$

(ii) *The system* $\Sigma_{\rm aff}$ *is locally state space equivalent at* $p \in X$ *to a linear controllable system of the form*

$$\Lambda: \quad \dot{x} = Ax + Bu = Ax + \sum_{i=1}^{m} b_i u_i \ , \quad x \in \mathbb{R}^n \ , \quad u \in \mathbb{R}^m,$$

at $0 \in \mathbb{R}^n$ *if and only if* Σ *satisfies* (E1), (E2) *and* $f(p) = 0$.

(iii) *The system* $\Sigma_{\rm aff}$ *is globally state space equivalent to a controllable linear system* Λ *on* $\mathbb{R}^n$ *if and only if it satisfies* (E1), (E2), *there exists* $p \in X$ *such that* $f(p) = 0$, *the state space* X *is simply connected and, moreover,*
(E3) *the vector fields* f *and* $g_1, \ldots, g_m$ *are complete.*

Recall that a vector field f is complete if its flow $\gamma_t^f(p)$ is defined for any $(t, p) \in \mathbb{R} \times X$.

3.2 Feedback equivalence

The role of the concept of feedback in control cannot be overestimated and is very well understood, both in the linear and nonlinear cases. We would like to consider it as a way of transforming nonlinear systems in order to achieve desired properties. When considering state-space equivalence the controls remain unchanged. The idea of feedback equivalence is to enlarge state-space transformations by allowing to transform controls as well and to transform them in a way which depends on the state: thus *feeding* the state back to the system.

Consider two general control systems Σ and $\tilde{\Sigma}$ given respectively by $\dot{x} = f(x, u)$, $x \in X$, $u \in U$ and $\dot{\tilde{x}} = \tilde{f}(\tilde{x}, \tilde{u})$, $\tilde{x} \in \tilde{X}$, $\tilde{u} \in \tilde{U}$. Assume that U and

$\tilde{U}$ are open subsets of $\mathbb{R}^m$. We say that Σ and $\tilde{\Sigma}$ are *feedback equivalent* if there exists a diffeomorphism $\chi : X \times U \to \tilde{X} \times \tilde{U}$ of the form

$$(\tilde{x}, \tilde{u}) = \chi(x, u) = (\Phi(x), \Psi(x, u))$$

which transforms the first system into the second, i.e.,

$$\mathrm{D}\Phi(x) f(x, u) = \tilde{f}(\Phi(x), \Psi(x, u)).$$

Observe that Φ plays the role of a coordinate change in X and Ψ, called feedback transformation, changes coordinates in the control space in a way which is state dependent.

When studying dynamical control systems with parameters and their bifurcations, the situation is opposite: coordinate changes in the parameters space are state-independent, while coordinate changes in the state space may depend on the parameters.

For the control-affine case, i.e., for systems of the form

$$\Sigma_{\mathrm{aff}} : \quad \dot{x} = f(x) + \sum_{i=1}^{m} g_i(x) u_i = f(x) + g(x)u,$$

where $g = (g_1, \ldots, g_m)$ and $u = (u_1, \ldots, u_m)^T$, in order to preserve the control affine form of the system, we will restrict feedback transformations to control affine ones

$$\tilde{u} = \Psi(x, u) = \tilde{\alpha}(x) + \tilde{\beta}(x)u,$$

where $\tilde{\beta}(x)$ is an invertible $m \times m$ matrix and $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_m)^T$. Denote the inverse feedback transformation by $u = \alpha(x) + \beta(x)\tilde{u}$. Then feedback equivalence means that

$$\tilde{f} = \Phi_*(f + g\alpha) \quad \text{and} \quad \tilde{g} = \Phi_*(g\beta),$$

where $\tilde{g} = (\tilde{g}_1, \ldots, \tilde{g}_m)$.

For control linear systems of the form $\dot{x} = g(x)u = \sum_{i=1}^{m} g_i(x)u_i$, (local) feedback equivalence coincides with (local) equivalence of distributions $\mathcal{G}$ spanned by the vector fields g_i's.

4 Feedback linearization

Since feedback transformations change dynamical behavior of a system they are used to achieve some required properties of the system. In Sections 6.2 and 6.3 we will show how feedback transformations are used to synthesize controls with decoupling properties. In this Section we will study the problem of when a nonlinear system can be transformed to a linear form via feedback. A particular case of feedback linearization of single-input control affine systems has been discussed in Section 2. The interest in feedback linearization is two-fold. Firstly, if one is able to compensate nonlinearities by feedback then the modified system possesses all control properties of its linear equivalent and linear control theory can be used in order to study it and/or to achieve the desired control properties. This shows possible engineering applications of feedback linearization, compare Example 2.1. From mathematical (or system theory) viewpoint, if we would like to classify nonlinear systems under feedback transformations (which define a group action on the space of all systems) then one of the most natural problems is to characterize those nonlinear systems which are feedback equivalent to linear ones. In Section 4.1 we will study feedback linearization of multi-input and general nonlinear systems. In Section 4.2 we will consider linearization using feedback which changes the drift vector field only. Finally, in Section 4.3 we will study the problem of finding the largest possible linearizable subsystem of the given system.

4.1 Static feedback linearization

A general nonlinear control system

$$\Sigma : \ \dot{x} = f(x, u),$$

is *(locally at (x_0, u_0)) feedback linearizable* if it is (locally at (x_0, u_0)) feedback equivalent to a controllable linear system Λ_c of the form

$$\Lambda_c : \ \dot{\tilde{x}} = A\tilde{x} + c + B\tilde{u}.$$

Recall the notation

$$\mathcal{F} = \{f_u \ \mid \ u \in U\}.$$

For any $u \in U$, define the following distributions on X which will play the

fundamental role in solving the feedback linearization problem.

$$\begin{aligned}\Delta_1(x,u) &= \operatorname{Im}\frac{\partial f}{\partial u}(x,u)\\ \Delta_2(x,u) &= \Delta_1(x,u) + \operatorname{span}\left[\mathcal{F},\Delta_1\right](x,u)\\ &= \Delta_1(x,u) + \operatorname{span}\left\{[f_u,g](x,u) \mid f_u\in\mathcal{F},\ g\in\Delta_1\right\}\end{aligned}$$

and, inductively,

$$\begin{aligned}\Delta_j(x,u) &= \Delta_{j-1}(x,u) + \operatorname{span}\left[\mathcal{F},\Delta_{j-1}\right](x,u)\\ &= \Delta_{j-1}(x,u) + \operatorname{span}\left\{[f_u,g](x,u) \mid f_u\in\mathcal{F},\ g\in\Delta_{j-1}\right\}.\end{aligned}$$

Remark 4.1 For the linear system Λ_c we have

$$\Delta_1 = \operatorname{Im} B \qquad \Delta_j = \operatorname{Im}\left(B,\ldots,A^{j-1}B\right),\ j\geq 0.$$

In the control-affine case, the feedback linearization problem was solved by Jakubczyk and Respondek [27], and independently by Hunt and Su [22] (see Theorem 4.5 below). In the general case we have the follwing result.

Theorem 4.2 *Σ is locally feedback linearizable at (x_0,u_0) if and only if it satisfies in a neighborhood of (x_0,u_0) the following conditions*

(A0) *Δ_1 does not depend on u,*

(A1) $\dim\ \Delta_j(x,u) =$*const*, $j=1,...,n$,

(A2) *Δ_j are involutive*, $j=1,...,n$,

(A3) $\dim\Delta_n(x_0,u_0)=n$.

Remark 4.3 One can show that if Δ_1 is involutive, of constant rank, and does not depend on u then the successive distributions Δ_j, for $j\geq 2$, do not depend on u either. Thus we can check the involutivity condition (A2) for them for a single value u only (for example for u_0).

In applications, one is often interested in points of equilibria. Denote by Λ a linear system of the form

$$\Lambda\ :\ \dot{\tilde{x}} = A\tilde{x} + B\tilde{u},$$

that is, the system Λ_c with $c=0$.

Corollary 4.4 *Σ_{aff} is locally feedback equivalent at (x_0, u_0) to a controllable linear system Λ at $(0,0)$ if and only if it satisfies the conditions* (A0)-(A3) *and moreover $f(x_0, u_0) \in \Delta_1(x_0, u_0)$.*

Consider feedback equivalence of linear controllable multi-input systems Λ of the form $\dot{x} = Ax + Bu$ (in this case the diffeomorphism $\Phi(x)$ and feedback $\Psi(x,u)$ are taken to be linear with respect to the state and control). As shown by Brunovský [5], complete feedback invariants are the dimensions m_j of $\operatorname{Im} M^j$, where the map $M^j : \mathbb{R}^{mj} \longrightarrow \mathbb{R}^n$ is defined as $[B, AB, \ldots, A^{j-1}B]$. Put $n_0 = 0$ and $n_j = m_j - m_{j-1}$, for $1 \leq j \leq n$. Define

$$\kappa_j = \max\{n_i \mid n_i \geq j\}. \tag{4.1}$$

Observe that $\kappa_1 \geq \cdots \geq \kappa_m$ and $\sum_{i=1}^{m} \kappa_i = n$. The integers κ_i, called controllability (or Brunovský) indices, form another set of complete invariants of feedback equivalence of linear controllable systems.

Every controllable system Λ with indices $\kappa_1 \geq \cdots \geq \kappa_m$ is feedback equivalent to the system

$$\begin{array}{rcll} \dot{x}_{i,j} & = & x_{i,j+1}\ , & \text{for } 1 \leq j \leq \kappa_i - 1, \\ \dot{x}_{i,\kappa_i} & = & u_i\ , & \end{array} \tag{4.2}$$

where $1 \leq i \leq m$, called *Brunovský canonical form*, which consists of m independent series of κ_i integrators.

Very often we deal with control-affine systems Σ_{aff}. To state a feedback linearization result for Σ_{aff}, we define the following distributions

$$\begin{array}{rcl} \mathcal{D}^1(x) & = & \operatorname{span}\{g_i(x),\ 1 \leq i \leq m\} \\ \mathcal{D}^j(x) & = & \operatorname{span}\{\operatorname{ad}_f^{q-1} g_i(x),\ 1 \leq q \leq j,\ 1 \leq i \leq m\}, \end{array}$$

for $j \geq 2$. If the dimensions $d_j(x)$ of $\mathcal{D}^j(x)$ are constant (see (A1)' and (B1) below) we denote them by d_j and we define indices ρ_j as follows. Define $d_0 = 0$ and put $r_j = d_j - d_{j-1}$ for $1 \leq j \leq n$. Then (compare (4.1))

$$\rho_j = \max\{r_i \mid r_i \geq j\}. \tag{4.3}$$

If the distributions Δ_j defined earlier in this section are involutive, then they are feedback invariant. If, moreover, the system is affine with respect to controls then, clearly, $\Delta_j = \mathcal{D}^j$, for $j \geq 1$ and, in particular, $\rho_1 \geq \cdots \geq \rho_m$ are feedback invariant. In this case the indices ρ_j coincide with κ_j, the controllability indices of the linear equivalent of Σ.

The following result (see [27] and [22]) describes linearizable control-affine systems.

Theorem 4.5 *The following conditions are equivalent.*

(i) Σ *is locally feedback linearizable at* $x_0 \in \mathbb{R}^n$.

(ii) Σ *satisfies in a neighborhood of* x_0

(A1)' $\dim \mathcal{D}^j(x) = const$, for $1 \le j \le n$,

(A2)' $\mathcal{D}^j$ *are involutive,* for $1 \le j \le n$,

(A3)' $\dim \mathcal{D}^n(x_0) = n$.

(iii) Σ *satisfies in a neighborhood of* x_0

(B1) $\dim \mathcal{D}^j(x) = const$, for $1 \le j \le n$,

(B2) $\mathcal{D}^{\rho_j - 1}$ *are involutive,* for $1 \le j \le m$,

(B3) $\dim \mathcal{D}^{\rho_1}(x_0) = n$, *where* ρ_1 *is the largest controllability index.*

In the single-input case $m = 1$, the condition (A3) (or, equivalently, (B3)) states that $g(x_0), \ldots, \mathrm{ad}_f^{n-1} g(x_0)$ are independent, which implies that all distributions $\mathcal{D}^j$, for $1 \le j \le n$, are of constant rank. In the following Corollary of Theorem 4.5 we thus rediscover Theorem 2.2.

Corollary 4.6 *A scalar input system* Σ *is feedback linearizable if and only if it satisfies*

(C1) $g(x_0), \ldots, \mathrm{ad}_f^{n-1} g(x_0)$ *are independent,*

(C2) $\mathcal{D}^{n-1}$ *is involutive.*

Example 4.7 Consider the following rigid two-link robot manipulator (*double pendulum*); compare, e.g., [6] or [37].

$$\begin{aligned} \dot{x}^1 &= x^2 \\ \dot{x}^2 &= -M^{-1}(x^1)(C(x^1, x^2) + k(x^1)) + M^{-1}(x^1)u \ , \end{aligned}$$

where θ_1 and θ_2 represent the angles (between the horizontal and the first arm and between the arms) and $x^1 = (\theta_1, \theta_2)$, $x^2 = (\dot{\theta}_1, \dot{\theta}_2)$. The control torques applied to the joints are $u = (u_1, u_2)$ and the positive definite symmetric matrix $M(x^1)$ is given by

$$\begin{pmatrix} m_1 l_1^2 + m_2 l_1^2 + m_2 l_2^2 + 2 m_2 l_1 l_2 \cos\theta_2 & m_2 l_2^2 + m_2 l_1 l_2 \cos\theta_2 \\ m_2 l_2^2 + m_2 l_1 l_2 \cos\theta_2 & m_2 l_2^2 \end{pmatrix}.$$

The term $k(\theta)$ represents the gravitational force and the term $C(\theta, \dot{\theta})$ reflects the centripetal and Coriolis force.

We have that $\mathcal{D}^1 = \text{span}\{\frac{\partial}{\partial x^2}\} = \text{span}\{\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}\}$ is involutive and $\dim \mathcal{D}^2(x^1, x^2) = 4$. Hence the double pendulum is feedback linearizable. A linearizing feedback is given, e.g., by $u = C(x^1, x^2) + k(x^1) + M(x^1)\tilde{u}$. □

Example 4.8 Consider the following model of a permanent magnet stepper motor [46]

$$\begin{aligned}
\dot{x}_1 &= -K_1 x_1 + K_2 x_3 sin(K_5 x_4) + u_1 \\
\dot{x}_2 &= -K_1 x_2 + K_2 x_3 cos(K_5 x_4) + u_2 \\
\dot{x}_3 &= -K_3 x_1 sin(K_5 x_4) + K_3 x_2 cos(K_5 x_4) - K_4 x_3 + K_6 sin(4K_5 x_4) - \tau_L/J \\
\dot{x}_4 &= x_3 \ ,
\end{aligned}$$

where x_1, x_2 denote currents, x_3 denotes the rotor speed and x_4 its position, J is the rotor inertia, and τ_L is the load torque. We see the distributions $\mathcal{D}^1 = \text{span}\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\}$ and $\mathcal{D}^2 = \text{span}\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\}$ are involutive and that $\dim \mathcal{D}^3(x) = 4$ and thus the system is locally (and even globally!) feedback linearizable. □

Example 4.9 The goal of this example is to show how to solve nonlinear problems by transforming the system to an equivalent linear system and solving the linear version of the problem for the linear system. Consider the following system

$$\begin{aligned}
\dot{x} &= y + yz \\
\dot{y} &= z \\
\dot{z} &= u + \sin x,
\end{aligned}$$

where $(x, y, z) \in \mathbb{R}^3$. We want to stabilize it exponentially globally on $\mathbb{R}^3$. Firstly, we show that the system is feedback linearizable. To simplify calculations, replace f by $\tilde{f} = f + \alpha g = f - (\sin x)g = (y + yz, z, 0)^T$. We have $g = (0, 0, 1)^T$, $\text{ad}_{\tilde{f}} g = -(y, 1, 0)^T$, and $[g, \text{ad}_{\tilde{f}} g] = 0$. Thus the distributions $\mathcal{D}^1 = \text{span}\{\frac{\partial}{\partial z}\}$ and $\mathcal{D}^2 = \text{span}\{y\frac{\partial}{\partial x} + \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\}$ are involutive. We seek for a function φ whose differential anihilates $\mathcal{D}^2$ which means to find a solution of the following system of 1-st order partial differential equations (compare Section 2)

$$\begin{cases} \frac{\partial \varphi}{\partial z} = 0 \\ y\frac{\partial \varphi}{\partial x} + \frac{\partial \varphi}{\partial y} = 0. \end{cases}$$

We conclude that φ can be an arbitrary function of $x - \frac{y^2}{2}$ and we choose $\varphi = x - \frac{y^2}{2}$. Therefore we put, see Section 2, $\tilde{x} = x - \frac{y^2}{2}$, $\tilde{y} = L_f\varphi = y$, $\tilde{z} = L_f^2\varphi = z$ and, finally, $u = \tilde{u} - \sin x$. This yields the following linear system

$$\begin{aligned} \dot{\tilde{x}} &= \tilde{y} \\ \dot{\tilde{y}} &= \tilde{z} \\ \dot{\tilde{z}} &= \tilde{u}, \end{aligned}$$

which we stabilize on $\mathbb{R}^3$ globally and exponentially via a linear feedback of the form $\tilde{u} = k\tilde{x} + l\tilde{y} + m\tilde{z}$, where the matrix

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ k & l & m \end{pmatrix}$$

is Hurwitz. Therefore the nonlinear feedback

$$u = k(x - \frac{y^2}{2}) + ly + mz - \sin x$$

stabilizes globally and asymptotically on $\mathbb{R}^3$ the original system. □

Example 4.10 Consider the following model of the rigid body whose gas jets control the rotations around the two first principal axes.

$$\begin{aligned} \dot{\omega}_1 &= a_1\omega_2\omega_3 + u_1 \\ \dot{\omega}_2 &= a_2\omega_1\omega_3 + u_2 \\ \dot{\omega}_3 &= a_3\omega_1\omega_2. \end{aligned}$$

We have $f = (a_1\omega_2\omega_3, a_2\omega_1\omega_3, a_3\omega_1\omega_2)^T$, $g_1 = (1,0,0)^T$ and $g_2 = (0,1,0)^T$. We calculate $\mathrm{ad}_f g_1 = -(0,0,a_3\omega_2)^T$ and $\mathrm{ad}_f g_2 = -(0,0,a_3\omega_1)^T$. We thus see that the distribution $\mathcal{D}^1 = \mathrm{span}\,\{g_1, g_2\} = \mathrm{span}\,\{\frac{\partial}{\partial\omega_1}, \frac{\partial}{\partial\omega_2}\}$ is always involutive and of rank two everywhere while $\mathcal{D}^2 = \mathrm{span}\,\{g_1, g_2, ad_f g_1, ad_f g_2\}$ is of rank three if and only if $a_3 \neq 0$ and either $\omega_1 \neq 0$ or $\omega_2 \neq 0$. In the first case we put $\tilde{\omega}_1 = a_3\omega_1\omega_2$, $u_1 = \frac{1}{a_3\omega_2}(-a_1a_3\omega_2^2\omega_3 - \alpha_2\alpha_3\omega_1^2\omega_3 + a_3\omega_1\tilde{u}_2)$, and $u_2 = -a_2\omega_1\omega_3 + \tilde{u}_2$ and we get the linear system

$$\begin{aligned} \dot{\tilde{\omega}}_1 &= \tilde{u}_1 \\ \dot{\omega}_2 &= \tilde{u}_2 \\ \dot{\omega}_3 &= \tilde{\omega}_1. \end{aligned}$$

In the second case we we put $\tilde{\omega}_2 = a_3\omega_1\omega_2$ and define $\tilde{u}_1$ and $\tilde{u}_2$ in an anologous way. □

Example 4.11 Consider the following model of unicycle

$$\begin{aligned} \dot{x}_1 &= u_1 \cos\theta \\ \dot{x}_2 &= u_2 \sin\theta \\ \dot{\theta} &= u_2 \ , \end{aligned}$$

where $(x_1, x_2, \theta) \in \mathbb{R}^2 \times S^1$. We have

$$g_1 = (\cos\theta, \sin\theta, 0)^T, \ g_2 = (0, 0, 1)^T,$$

thus $[g_1, g_2] = (\sin\theta, -\cos\theta, 0)^T$ and hence $\mathcal{D}^1$ is not involutive: the unicycle is *not* static feedback linearizable. □

4.2 Restricted feedback linearization

Consider a control-affine system Σ_{aff} and a feedback transformation $u = \alpha(x) + \beta(x)\tilde{u}$ which can be interpreted as an (affine) change of coordinates, depending on the state, in the input space. The term β allows to choose generators of the distribution $\mathcal{D}^1 = \text{span}\{g_1, \ldots, g_m\}$ whereas the term α changes the drift f. Restricted feedback allows to transform the drift f only and keeps the g_i's unchanged. More precisely, two control affine systems Σ_{aff} and $\tilde{\Sigma}_{\text{aff}}$ are *restricted feedback equivalent* if there exist a diffeomorphism Φ between their state spaces and a restricted feedback of the form $u = \alpha(x) + \tilde{u}$ such that

$$\tilde{f} = \Phi_*(f + g\alpha) \quad \text{and} \quad \tilde{g}_i = \Phi_* g_i, \tag{4.4}$$

for $1 \leq i \leq m$.

We will be interested in equivalence to liner systems under such feedback and we will call it *restricted feedback linearization.*

The three main reasons to discuss restricted feedback linearization are as follows. Firstly, it was Brockett's restricted feedback linearization result [3] which begun an increasing interest in various kinds of feedback linearization problems for nonlinear systems. Secondly, there is a nice stochastic interpretation of the restricted feedback linearization [3]. Thirdly, it is relatively easy, as we will show it, to proceed from local results to global ones.

Consider single-input systems of the form

$$\Sigma_{\text{aff}} \ : \ \dot{x} = f(x) + g(x)u \ , \ x \in X \ , \ u \in \mathbb{R},$$

and study their equivalence to linear single-input systems of the form

$$\Lambda_c \ : \quad \dot{\tilde{x}} = A\tilde{x} + c + b\tilde{u} \ , \quad \tilde{x} \in \mathbb{R}^n \quad \tilde{u} \in \mathbb{R}.$$

We have the following result [1].

Theorem 4.12 *Σ_{aff} is locally restricted feedback linearizable at x_0 if and only if it satisfies in a neighborhood of x_0 the following conditions*

(RC1) *$g(x_0), \ldots, \text{ad}_f^{n-1} g(x_0)$ are independent.*

(RC1) *$[\text{ad}_f^q g, \text{ad}_f^r g] \subset \mathcal{D}^{n-2}$ for any $0 \leq q, r \leq n-1$,*

Remark 4.13 Like in the case of feedback linearization (compare Corollary 4.4), Σ_{aff} is restricted feedback equivalent at x_0 to Λ_c, with $c = 0$, at 0 if and only if $f(x_0) \in \mathcal{D}^1(x_0)$.

In the single-input case all linearizable systems are equivalent to the Brunovský canonical form (compare (4.2))

$$\begin{array}{rcl} \dot{x}_i & = & x_{i+1} \ , \text{ for } 1 \leq i \leq n-1 \ , \\ \dot{x}_n & = & u \ . \end{array} \tag{4.5}$$

If Σ_{aff} is (locally) feedback linearizable, then there are many pairs (α, β) and many (local) diffeomorphisms which transform Σ_{aff} into its Brunovský canonical form. However, if we allow for restricted feedback only, then α transforming Σ_{aff} into the canonical form is unique and is given by

$$\alpha = (-1)^{n-1} L_f^{n-1} \gamma_n \ , \tag{4.6}$$

where L_f stands for the Lie derivative along f and the smooth function γ_n is uniquely defined by

$$f = \sum_{i=1}^{n} \gamma_i \text{ad}_f^{i-1} g.$$

This observation is crucial for establishing the following result on restricted feedback linearization [9], [39].

Theorem 4.14 *Σ is restricted feedback globally linearizable, that is, globally equivalent via a restricted feedback to a linear system on $\mathbb{R}^n$, if and only if it satisfies the conditions* (RC1),(RC2) *and, moreover,*

(RC3) *the vector fields $\tilde{f}$ and g are complete, where $\tilde{f} = f + g\alpha$ and α is defined by* (4.6),

(RC4) *the state space X is simply connected.*

Example 4.15 (Continuation of Examples 2.1 and 2.4). We have $f = \omega\frac{\partial}{\partial\theta} - \frac{g}{l}\sin\theta\frac{\partial}{\partial\omega}$ and $g = \frac{1}{ml^2}\frac{\partial}{\partial\omega}$. Therefore $[\mathrm{ad}_f g, g] = 0$ and since g and $\mathrm{ad}_f g$ are independent everywhere, the system is restricted feedback linearizable. Indeed, it is immediate to see that the feedback $u = mgl\sin\theta + \tilde{u}$ brings the system to a linear form (no action of diffeomorphism is needed). □

The nonlinear pendulum defined on $S^1 \times \mathbb{R}^1$ is globally equivalent to a linear system evolving on $S^1 \times \mathbb{R}^1$. If we enlarge the class of linear systems to include systems of the form $\dot{x} = Ax + Bu$, where each component x_i of x is either a global coordinate on $\mathbb{R}^1$ or a global coordinate (angle) on S^1, then Theorem 4.14 remains true if we drop the assumption (RC4). This includes many mechanical control systems.

4.3 Partial linearization

The linearizability conditions are restrictive (except for the scalar input affine systems on the plane, compare Corollary 2.3). Given a nonlinearizable system it is therefore natural to ask what is its largest linearizable subsystem. Consider a partially linear system Λ_{part} of the form

$$\Lambda_{\mathrm{part}}: \quad \begin{array}{rcl} \dot{\tilde{x}}^1 & = & A\tilde{x}^1 + c + \sum_{i=1}^m b_i\tilde{u}_i \\ \dot{\tilde{x}}^2 & = & \tilde{f}^2(\tilde{x}^1, \tilde{x}^2) + \sum_{i=1}^m \tilde{g}_i^2(\tilde{x}^1, \tilde{x}^2)\tilde{u}_i, \end{array}$$

with $\tilde{x}^1$, $\tilde{x}^2$ being possibly vectors. Recall the notion of the Lie ideal $\mathcal{L}_0$ of the system (see [26] of this volume), which is defined as the Lie ideal generated by $g_1, \ldots, g_m$ in $\mathcal{L}$, or, in other words, $\mathcal{L}_0 = \mathrm{Lie}\{\mathrm{ad}_f^q g_i \mid 1 \le i \le m,\ q \ge 0\}$. With the help of $\mathcal{L}_0$ we define another Lie ideal by putting

$$\mathcal{L}^2 = [\mathcal{L}_0, \mathcal{L}_0] = \{[f_1, f_2] \ \mid \ f_1, f_2 \in \mathcal{L}_0\}.$$

It is $\mathcal{L}^2$ which contains all intrinsic nonlinearities not removable by the action of diffeomorphisms as the following result [40] shows.

Theorem 4.16 *Consider a control affine system* Σ_{aff}.

(i) *If* Σ_{aff} *is locally state space equivalent at* x_0 *to a partially linear system* Λ_{part} *then* $\dim \mathcal{L}^2(x) < n$ *in a neighborhood of* x_0.

(ii) *Assume that* Σ_{aff} *satisfies* $\dim \mathcal{L}_0(x_0) = n$ *and that* $\dim \mathcal{L}^2(x) = \sigma$*=const. in a neighborhood of* x_0. *Then* Σ_{aff} *is locally state space equivalent to a partially linear system* Λ_{part}, *such that the dimension*

of the linear subsystem is $\dim \tilde{x}^1 = n - \sigma$ *and, moreover, the linear subsystem is controllable.*

Corollary 4.17 *Let an analytic system* Σ_{aff} *satisfies* $\dim \mathcal{L}_0(x_0) = n$. *It is locally state space equivalent at* x_0 *to a partially linear system* Λ_{part} *if and only if*

$$\dim \mathcal{L}^2(x_0) < n.$$

Moreover, there exists a system Λ_{part}, *with* $n-\sigma$*-dimensional linear controllable subsystem, where* $\sigma = \dim \mathcal{L}^2(x_0)$, *which is state space equivalent to* Σ_{aff}.

Now we consider the problem of transforming a nonlinear system to a partially linear one via feedback. This problem has been studied and solved in the scalar-input case in [33] and in the multi-input case in [34] and [40]. Recall that for a smooth distribution $\mathcal{D}$ we denote by $\overline{\mathcal{D}}$ its involutive closure, that is, the smallest distribution containing $\mathcal{D}$ and closed under the Lie bracket.

Theorem 4.18 *Consider a single-input system* Σ_{aff}.

(i) *If* Σ_{aff} *is locally feedback equivalent at* x_0 *to a partially linear* Λ_{part} *with* ρ*-dimensional linear controllable subsystem then* Σ_{aff} *satisfies the following conditions:*

(PC1) $g(x_0), \ldots, \mathrm{ad}_f^{\rho-1}(x_0)$ *are independent,*

(PC2) $\dim \overline{\mathcal{D}}^{\rho-1}(x) < n$ *in a neighborhood of* x_0,

(PC3) $\mathrm{ad}_f^{\rho-1} g(x) \notin \overline{\mathcal{D}}^{\rho-1}(x)$.

(ii) *Assume that there exists an integer* ρ *such that* $\dim \overline{\mathcal{D}}^{\rho-1}(x)$=*const. and that* (PC1), (PC2), (PC3) *are satisfied. Then* Σ_{aff} *is locally feedback equivalent to a partially linear system* Λ_{part} *with* ρ*-dimensional linear controllable subsystem, that is* $\dim \tilde{x}^1 = \rho$. *Moreover, the largest* ρ *satisfying the above conditions gives the largest dimension of linear subsystem among all possible partial linearizations.*

Example 4.19 Consider a symmetric rigid body (two inertia momenta are equal) with one pair of jets

$$\begin{aligned} \dot{\omega}_1 &= a\omega_2\omega_3 + e_1 u \\ \dot{\omega}_2 &= -a\omega_1\omega_3 + e_2 u \\ \dot{\omega}_3 &= e_3 u \quad . \end{aligned}$$

Compute $g = (e_1, e_2, e_3)^T$, $\mathrm{ad}_f g = a(e_2\omega_3 + e_3\omega_2, -e_1\omega_3 - e_3\omega_1, 0)^T$. Hence for $\mathcal{D}^2$ to be involutive, that is, $[g, \mathrm{ad}_f g] = 2ae_3(-e_2, e_1, 0) \in \mathcal{D}^2 = \mathrm{span}\{g, \mathrm{ad}_f g\}$, we need either $e_3 = 0$ or $e_1 = e_2 = 0$. In the former case, ω_3 remains constant, in the latter, the symmetric spacecraft is controlled in a symmetric way: the angular momentum of the jet is parallel to the third principal axis. Notice that for all values of the control vector $e = (e_1, e_2, e_3)^T$, the system is not feedback linearizable. Indeed, either $\mathcal{D}^2$ is not involutive or the system is not accessible. On the other hand, for all values of the control vector field $e \neq 0$, the system contains a 2-dimensional linear subsystem for an open and dense set of initial conditions. □

5 Observability

In this chapter we consider briefly the concept of nonlinear observability. We start with geometric approach to the observability problem and in Section 5.1 we state a sufficient condition, called observability rank condition, based on successive Lie derivatives of the output along the dynamics. In Section 5.2 we discuss (local) decompositions into observable and completely unobservable parts which generalize the classical Kalman decomposition. Then in Section 5.3 we consider the problem of uniform observability, which means that we can observe the system for any input. In Section 5.4 we give a necessary and sufficient condition for local observability. Finally, in Section 5.5 we discuss generic properties: we give normal forms for generic systems and recall results concerning genericity of observability.

5.1 Nonlinear observability

Consider the class of nonlinear systems with outputs (measurements) of the form

$$\Sigma \; : \quad \begin{array}{rcl} \dot{x} & = & f(x,u)\,, \\ y & = & h(x)\,, \end{array}$$

where $x \in X$, $u \in U$, $y \in Y$. Here X, U, and Y are open subsets of $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^p$, respectively (or differentiable manifolds of dimensions n, m, and p, respectively)[1]. The map $h : X \to Y$ represents the vector of p measurements (observations), where $h_i \in C^\infty(X)$, for $1 \leq i \leq p$, and $h = (h_1, \ldots, h_p)^T$.

[1] Except for the second part of Section 5.5, where we assume U to be J^m, with J being a compact subinterval of $\mathbb{R}$.

Throughout this section, Σ will denote the above described nonlinear system with output.

The class of admissible controls $\mathcal{U}$ is fixed and $\mathcal{PC} \subset \mathcal{U} \subset \mathcal{M}$, where $\mathcal{PC}$ denotes the class of piece-wise constant controls with values in U and $\mathcal{M}$ the class of measurable controls with values in U.

Let $\mathcal{Y}$ denote the space of absolutely continuous functions on X with values in Y. For the system Σ we define the *response map*, called also *input-output map*,

$$R_\Sigma : X \times \mathcal{U} \longrightarrow \mathcal{Y},$$

which to any initial condition $q \in X$ and any admissible control $u(\cdot) \in \mathcal{U}$ attaches the output of the system

$$y_{q,u}(t) = y(t, q, u(\cdot)) = h(x(t, q, u(\cdot))),$$

where $x(t, q, u(\cdot))$ denotes the solution of $\dot{x} = f(x, u)$, for $u(\cdot) \in \mathcal{U}$, passing through q, that is $x(0, q, u(\cdot)) = q$. The control $u(\cdot)$ being defined on an interval $I_u \subset \mathbb{R}$, such that $0 \in I_u$, we consider the output $y(\cdot)$ on the maximal interval $I_y \subset I_u \subset \mathbb{R}$ on which it exists.

Roughly speaking, the problem of observability is that of the injectivity, with respect to the initial condition, of the response map.

We say that two states $q_1, q_2 \in X$ are *indistinguishable*, and we write $q_1 I q_2$, if

$$y_{q_1,u}(t) = y_{q_2,u}(t),$$

for any $u(\cdot) \in \mathcal{U}$ and any t for which both sides exist.

Definition 5.1 We call the system Σ *observable* if for any two states $q_1, q_2 \in X$ we have

$$q_1 I q_2 \Longrightarrow q_1 = q_2,$$

that is, if there exists an admissible control $u(\cdot) \in \mathcal{U}$ and a time $t \geq 0$ such that

$$y_{q_1,u}(t) \neq y_{q_2,u}(t).$$

meaning that the states q_1 and q_2 are *distinguishable*.

Definition 5.2 Σ is called *locally observable* at $q \in X$ if there is a neighborhood V of q such that for any $\tilde{q} \in V$, the states q and $\tilde{q}$ are distinguishable.

Given a system Σ and an open set $V \subset X$, by the restriction $\Sigma_{|V}$ we will mean a control system with the state space V, defined by the restrictions of f and h to $V \times U$ and V, respectively.

Definition 5.3 Σ is called *strongly locally observable* at a point $q \in X$ if there exists a neighborhood V of q such that the restricted system $\Sigma_{|V}$ is observable.

We would like to emphasize some features of the introduced concepts of observability. Strong local observability is a local concept in two aspects. Firstly, strong local observability means that, in general, we are able to distinguish neighboring points only. Secondly, we are able to do so considering trajectories which stay close to the initial condition. Of course, observability implies strong local observability at any point (we can take $V = X$), which, in turn, implies local observability at any point (for each point we take the neighborhood V existing due to the strong local observability). In general, the reversed implications do not hold, see Examples 5.6 and 5.7 below.

Example 5.4 Consider a mechanical system evolving according to Newton's law

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = u,$$

where x_1 denotes the position, x_2 the velocity, and u is the control force. We observe the position

$$y = x_1.$$

This system is clearly observable. Indeed, let $y_{q,u}(t)$ and $\tilde{y}_{\tilde{q},u}(t)$ be the outputs of the system initialized, respectively, at $q = (x_{10}, x_{20})^T$ and at $\tilde{q} = (\tilde{x}_{10}, \tilde{x}_{20})^T$ and governed by a control $u(\cdot)$. Assume that $y_{q,u}(t) = \tilde{y}_{\tilde{q},u}(t)$ for any t. Then comparing at $t = 0$ both sides of the above equality as well as derivatives at $t = 0$ of both sides, we get $x_{10} = \tilde{x}_{10}$ and $x_{20} = \tilde{x}_{20}$, which proves the observability.

Now assume that, for the same control system, we observe the velocity

$$y = x_2.$$

The system is not observable. Indeed, the initial conditions $q = (x_{10}, x_{20})^T$ and $\tilde{q} = (\tilde{x}_{10}, \tilde{x}_{20})^T$, such that $x_{10} \neq \tilde{x}_{10}$ but $x_{20} = \tilde{x}_{20}$, produce the same output $y(t) = \int_0^t u(s)ds + x_{20}$. Mechanically, this is obvious: we cannot estimate the position if we observe the velocity only. □

Example 5.5 Consider the linear oscillator (linear pendulum) given by

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1, \end{aligned}$$

where x_1 denotes the position and x_2 the velocity. Assume that we observe the position

$$y = x_1.$$

Then the system is observable and, given the output function $y(\cdot)$, we can deduce the initial condition $(x_{10}, x_{20})^T$ by taking, like in Example 5.4, the output and its first derivative with respect to time at $t = 0$.

Now assume that we observe the velocity

$$y = x_2.$$

This system is also observable and once again we can deduce the initial condition by looking at the values at $t = 0$ of the output and its first time derivative. The reason for which observing the velocity renders the system observable is that the evolution of the velocity x_2 depends on the position x_1, which is not the case of the system of Example 5.4. □

Example 5.6 Consider the unicycle

$$\begin{aligned} \dot{x}_1 &= u_1 \cos\theta, & y_1 &= x_1 \\ \dot{x}_2 &= u_1 \sin\theta, & y_2 &= x_2 \\ \dot{\theta} &= u_2, \end{aligned}$$

where $(x_1, x_2)^T \in \mathbb{R} \times \mathbb{R}$ is the position of the center of the mass of the unicycle and $\theta \in S^1$ is the angle between the horizontal and the axis of the unicycle. We observe the position of the center of the mass.

The unicycle is observable. To see it, consider the outputs $y_{q,u}(t)$ and $\tilde{y}_{\tilde{q},u}(t)$ of the system controlled by $u(t) = (u_1(t), u_2(t))^T$, such that $u_1(t) \equiv 1$, passing for $t = 0$ by $q = (x_{10}, x_{20}, \theta_0)^T$ and $\tilde{q} = (\tilde{x}_{10}, \tilde{x}_{20}, \tilde{\theta}_0)^T$, respectively. Assume that $y_{q,u}(t) = \tilde{y}_{\tilde{q},u}(t)$. Thus $x_{10} = \tilde{x}_{10}$, $x_{20} = \tilde{x}_{20}$, and moreover $\sin\theta(t) = \sin\tilde{\theta}(t)$ and $\cos\theta(t) = \cos\tilde{\theta}(t)$. Hence we conclude that $\theta_0 = \tilde{\theta}_0$, where θ_0, $\tilde{\theta}_0 \in S^1$.

Now consider the unicycle, with the same observations $y_1 = x_1$ and $y_2 = x_2$, evolving on $\mathbb{R}^3$, that is, we consider $\theta \in \mathbb{R}$. It turns out that the system is not observable. To see this, we will show that the outputs $y_{q,u}(t)$ and $\tilde{y}_{\tilde{q},u}(t)$ of the system coincide for $q = (x_{10}, x_{20}, \theta_0)^T$ and $\tilde{q} =$

$(\tilde{x}_{10}, \tilde{x}_{20}, \tilde{\theta}_0)^T$ such that $x_{10} = \tilde{x}_{10}$, $x_{20} = \tilde{x}_{20}$, and $\tilde{\theta}_0 = \theta_0 + 2k\pi$. We have $\theta(t) = \int_0^t u_2(s)ds + \theta_0$ and $\tilde{\theta}(t) = \int_0^t u_2(s)ds + \tilde{\theta}_0$ and hence $\tilde{\theta}(t) = \theta(t) + 2k\pi$. Thus $\sin\tilde{\theta}(t) = \sin\theta(t)$ and $\cos\tilde{\theta}(t) = \cos\theta(t)$ implying that $y_{q,u}(t) = \tilde{y}_{\tilde{q},u}(t)$, for any control $u(\cdot) \in \mathcal{U}$ and the initial conditions as above. In other words, the points $q = (x_{10}, x_{20}, \theta_0)^T$ and $\tilde{q} = (\tilde{x}_{10}, \tilde{x}_{20}, \tilde{\theta}_0)^T$ such that $x_{10} = \tilde{x}_{10}$, $x_{20} = \tilde{x}_{20}$, and $\tilde{\theta}_0 = \theta_0 + 2k\pi$ are indistinguishable. Of course, the system is strongly locally observable at any $q \in \mathbb{R}^3$. □

Example 5.7 Consider the system

$$\dot{x} = 0 \ , \quad y = x^2 \ ,$$

where $x \in \mathbb{R}$ and $y \in \mathbb{R}$. The system is not observable because the initial conditions x_0 and $-x_0$ give the same output trajectories. This system is strongly locally observable at any $x_0 \neq 0$. Notice that it is locally observable at any point, in particular at $0 \in \mathbb{R}^2$, although in any neighborhood of 0 there are indistinguishable states. This shows that local observability is indeed a weaker property than strong local observability. □

We will give now a sufficient condition for strong local observability. To this end, we will introduce the following concepts.

Definition 5.8 The *observation space* of Σ is defined as

$$H = \mathrm{span}_{\mathbb{R}} \{ L_{f_{u_k}} \cdots L_{f_{u_1}} h_i \ \mid \ 1 \leq i \leq p, \ k \geq 0, \ u_1, \ldots, u_k \in U \},$$

where $f_{u_j}(\cdot) = f(\cdot, u_j)$ and $L_g\varphi$ stands for the Lie derivative of a smooth function φ with respect to a smooth vector field g, i.e.,

$$L_g\varphi(x) = d\varphi(x) \cdot g(x).$$

Observe that H is the smallest linear subspace of $C^\infty(X)$ containing the observations $h_1, \ldots, h_p$ and closed with respect to Lie differentiation by all elements of $\mathcal{F} = \{f(\cdot, u), u \in U\}$, i.e., all vector fields corresponding to constant controls. Using functions from H we define the following codistribution

$$\mathcal{H} = \mathrm{span} \ \{ d\phi \ : \ \phi \in H \} .$$

Notice that, in general, $\mathcal{H}$ is not of constant rank.

In the case of control affine systems of the form

$$\Sigma_{\text{aff}} \; : \qquad \begin{array}{rcl} \dot{x} & = & f(x) + \sum_{i=1}^{m} g_i(x) u_i \,, \\ y & = & h(x) \,, \end{array}$$

we have

$$\mathcal{H} = \text{span} \left\{ dL_{g_{j_k}} \cdots L_{g_{j_1}} h_i \; : \; 1 \leq i \leq p, \; 0 \leq j_l \leq m \right\},$$

where $g_0 = f$.

The following result of Hermann and Krener [21] gives a fundamental criterion for nonlinear observability.

Theorem 5.9 *Assume that the system Σ satisfies*

$$\dim \mathcal{H}(q) = n. \tag{5.1}$$

Then Σ is strongly locally observable at q.

The condition (5.1) will be called *observability rank condition*. It can be considered as a counterpart of the accessibility and strong accessibility rank conditions (see, e.g., the survey [26] of this volume), although the duality is not perfect, as we will see in the next example.

Example 5.10 The converse of Theorem 5.9 does not hold (even in the analytic case) as the following simple example shows. Consider

$$\dot{x} = 0 \; , \quad y = x^3 \; ,$$

where $x \in \mathbb{R}$, $y \in \mathbb{R}$. Of course, the system is strongly locally observable at any $x \in \mathbb{R}$ (even observable on $\mathbb{R}$) but it does not satisfy the rank condition at $0 \in \mathbb{R}$, since we have $\mathcal{H} = \text{span}\{ x^2 dx \}$. This shows also that the rank of $\mathcal{H}$ need not be constant. This is to be compared with the accessibility rank condition, which, in the analytic case, is necessary and sufficient for accessibility. □

Example 5.11 Consider a linear control system with outputs of the form

$$\Lambda \; : \qquad \begin{array}{rcl} \dot{x} & = & Ax + Bu \,, \\ y & = & Cx \; , \end{array}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$. We have $f = Ax$, $g_k = b_k$, for $1 \leq k \leq m$, and $h_i = C_i x$, for $1 \leq i \leq p$, where C_i denotes the i-th row of the matrix C. We calculate

$$L_f^j h_i = C_i A^j x \quad \text{and} \quad L_{g_k} L_f^j h_i = C_i A^j b_k.$$

Thus $\mathcal{H}(q) = \operatorname{span}\{C_i A^j \mid 1 \leq i \leq p,\ 0 \leq j \leq n-1\}$ and $\dim \mathcal{H}(q) = \operatorname{rank} O$, where O is the Kalman observability matrix

$$O = \begin{pmatrix} C \\ CA \\ \cdots \\ CA^{n-1} \end{pmatrix}.$$

Therefore a linear system satisfies the observability rank condition if and only if it satisfies Kalman observability condition $\operatorname{rank} O = n$. In this case, as it follows from Theorem 5.9, the system is strongly locally observable. Moreover, we know from the linear control theory, see e.g. [30], that the system is observable. Indeed, the response map R_Λ of the linear system Λ given by

$$y(t) = Cx(t) = Ce^{At}x_0 + \int_0^t Ce^{A(t-s)}Bu(s)ds$$

and associating to an initial condition x_0 the output trajectory, is affine with respect to the initial condition x_0 and thus local injectivity implies global injectivity. Notice that observability properties of a linear system do not depend on the chosen control; indeed, they depend only on the injectivity of the map

$$x_0 \longmapsto Ce^{At}x_0.$$

In the next example we will show that this no longer true in the nonlinear case. □

Example 5.12 The aim of this example is to show that, contrary to the linear case, controls play an important role in the nonlinear observability. In general, there may exist controls which do not distinguish points nevertheless the system can be observable if other controls distinguish. To illustrate that phenomenon, consider the bilinear system

$$\begin{aligned} \dot{x}_1 &= x_2 - x_2 u\ , \quad y = x_1 \\ \dot{x}_2 &= 0\ , \end{aligned}$$

where $(x_1, x_2)^T \in \mathbb{R}^2$. This system is observable, because if we put $u(t) \equiv 0$ we get an observable linear system. Notice, however, that the constant control $u(t) \equiv 1$ does not distinguish x_0 and $\tilde{x}_0$ such that $x_{10} = \tilde{x}_{10}$ and $x_{20} \neq \tilde{x}_{20}$ (we will come back to this phenomenon in Section 5.3). Of course, we can deduce strong local observability at any point from the rank condition. Indeed, we have $h = x_1$, $f = x_2 \frac{\partial}{\partial x_1}$, and $L_f h = x_2$. Hence $\mathcal{H} = \text{span}\{dx_1, dx_2\}$. □

Example 5.13 Consider the unicycle, see Example, 5.6, for which we observe $y_1 = x_1$ and $y_2 = x_2$. We have $h_1 = x_1$, $h_2 = x_2$, $g_1 = \cos\theta \frac{\partial}{\partial x_1} + \sin\theta \frac{\partial}{\partial x_2}$, and $g_2 = \frac{\partial}{\partial \theta}$. Hence $L_{g_1} h_1 = \cos\theta$ and $L_{g_1} h_2 = \sin\theta$. Thus $\mathcal{H} = \text{span}\{dx_1, dx_2, d\sin\theta, d\cos\theta\}$ implying that $\dim \mathcal{H}(q) = 3$, for any $q \in \mathbb{R}^2 \times S^1$. Therefore the unicycle satisfies the observability rank condition at any point of its configuration space. □

5.2 Local decompositions

Let us start with linear systems of the form

$$\Lambda \; : \quad \begin{array}{rcl} \dot{x} & = & Ax + Bu, \\ y & = & Cx \; , \end{array}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$. Denote by W, the kernel of the linear map defined by the Kalman observability matrix O (see Example 5.11). If Λ is not observable then we can find new coordinates (x^1, x^2), with x^1 and x^2 being possibly vectors and $\dim x^1 = k$, where $\dim W = n - k$, such that $x \in W$ if and only if $x = (0, x^2)$. Then Λ reads

$$\begin{array}{rcllll} \dot{x}^1 & = & A^1 x^1 + & + & B^1 u \; , & y = C^1 x^1 \; , \\ \dot{x}^2 & = & A^{21} x^1 + A^{22} x^2 & + & B^2 u \; , & \end{array}$$

where the pair (C^1, A^1) is observable.

As a consequence, any two initial states whose difference is not in W are distinguishable from each other, in particular, by means of the output produced by the zero input. Contrary, if their difference is in W, then they are indistinguishable. The factor system $\Lambda_{/I}$, where I is the indistinguishability equivalence relation, is observable and is given by

$$\Lambda^1 \; : \; \dot{x}^1 = A^1 x^1 + B^1 u \; , \; y = C^1 x^1 .$$

Geometrically, Λ^1 is obtained by factoring the system through the subspace W and the factor system is well defined since W is invariant under A. A

natural question is whether we can proceed similarly for the nonlinear system Σ?

Theorem 5.14 *Consider the nonlinear system Σ. Assume that the distribution $\mathcal{H}$ is of constant rank equal to k locally around q. Then we have.*

(i) *The codistribution $\mathcal{H}$ is integrable and there exist local coordinates $(x^1, x^2)^T$ defined in a neighborhood V of q, with x^1, x^2 possibly being vectors, such that $\mathcal{H} = \text{span}\{dx_1^1, \ldots, x_k^1\}$.*

(ii) *In the local coordinates $(x^1, x^2)^T$, the system Σ takes the form*

$$\begin{aligned} \dot{x}^1 &= f^1(x^1, u) , \qquad y = h^1(x^1) , \\ \dot{x}^2 &= f^2(x^1, x^2, u). \end{aligned}$$

(iii) *By taking V sufficiently small, two pints $q, \tilde{q} \in V$ are indistinguishable for $\Sigma_{|V}$ if and only if $\tilde{q} \in S_q$, where S_q is the integral leaf, passing through q, of the codistribution $\mathcal{H}$ restricted to V.*

(iv) *In V, factoring the system through the foliation of the integrable codistribution $\mathcal{H}$, produces the strongly locally observable system Σ^1 which, in $(x^1, x^2)^T$- coordinates, is given by*

$$\Sigma^1 \; : \; \dot{x}^1 = f^1(x^1, u) , \qquad y = h^1(x^1).$$

This result says that locally and under the constant rank assumption, the leaves of the foliation of the integrable codistribution $\mathcal{H}$ consist of indistinguishable points and that, on the other hand, we can distinguish the leaves.

From Theorem 5.14 we immediately get the two following corollaries.

Corollary 5.15 *If $\mathcal{H}$ is of constant rank in a neighborhood of q then the following conditions are equivalent.*

(i) *Σ is locally observable at q.*

(ii) *Σ is strongly locally observable at q.*

(iii) $\dim \mathcal{H}(q) = n$.

Corollary 5.16 *If Σ is locally observable at any point of X then* $\dim \mathcal{H}(q) = n$, *for $q \in X'$, an open and dense subset of X.*

An important case when the observability rank is constant is given by the following.

Proposition 5.17 *Assume that an analytic control system Σ satisfies the accessibility Lie rank condition everywhere on X. Then $\mathcal{H}$ is of constant rank on X. In particular, the system is locally observable at q (or, equivalently, strongly locally observable at q) if and only if* $\dim \mathcal{H}(q) = n$.

To illustrate the decomposition result of this section we consider the following example.

Example 5.18 Consider the unicycle, see Example 5.6, for which we measure the angle θ only, that is $h = \theta$. We have $L_{g_1}h = L_{g_2}h = 0$. Thus $\mathcal{H} = \operatorname{span}\{d\theta\}$ defines the foliation

$$\{\theta = \text{const.}\},$$

whose leaves consist of indistinguishable points. Indeed, if $\theta_0 = \tilde{\theta}_0$ then the points $q = (x_{10}, x_{20}, \theta_0)^T$ and $\tilde{q} = (\tilde{x}_{10}, \tilde{x}_{20}, \tilde{\theta}_0)^T$ are indistinguishable. The obvious reason for this is that the evolution of the observed variable $y(t) = \theta(t)$ is independent of that of $x_1(t)$ and $x_2(t)$. □

5.3 Uniform observability

In Example 5.12 we pointed out that for observable nonlinear systems there may exist controls that render the system unobservable. In this section we describe a class of systems, for which all controls distinguish points.

Definition 5.19 The system Σ is called *uniformly observable*, with respect to the inputs, if for any two states $q_1, q_2 \in X$, such that $q_1 \neq q_2$, and any control $u(\cdot) \in \mathcal{U}$

$$y_{q_1,u}(t) \not\equiv y_{q_2,u}(t).$$

Σ is *uniformly locally observable* at $q \in X$, if there exists a neighborhood V of q, such that Σ restricted to V is uniformly observable.

Example 5.20 Example 5.12 illustrates the existence of nonlinear systems that are not uniformly observable. Another example is the unicycle, see Example 5.6, for which we observe $y_1 = x_2$ and $y_2 = x_2$. The system is observable, nevertheless, for the control $u_1(t) \equiv 0$, any two points $q = (x_{10}, x_{20}, \theta_0)^T$ and $\tilde{q} = (\tilde{x}_{10}, \tilde{x}_{20}, \tilde{\theta}_0)^T$, such that $x_{10} = \tilde{x}_{10}$ and $x_{20} = \tilde{x}_{20}$ are indistinguishable. □

Of course, linear observable systems are uniformly observable. We will describe now a class of nonlinear uniformly observable systems. Consider a single-input single-output control-affine system of the form

$$\Sigma_{\text{aff}}: \quad \begin{aligned} \dot{x} &= f(x) + g(x)u \\ y &= h(x), \end{aligned}$$

where $x \in X$, $u \in \mathbb{R}$, $y \in \mathbb{R}$ and f, g are smooth vector fields on X.

The following result is due to Gauthier and Bornard [14].

Theorem 5.21 *For the system Σ_{aff} we have:*

(i) *If Σ_{aff} is uniformly locally observable at any $q \in X$, then around any point of on an open and dense submanifold X' of X there exist local coordinates $(x_1, \ldots, x_n)^T$ in which the system takes the following normal form*

$$\text{(UO)} \qquad \begin{aligned} \dot{x}_1 &= x_2 + ug_1(x_1), & y = x_1 \\ \dot{x}_2 &= x_3 + ug_2(x_1, x_2) \\ &\vdots \quad\quad \vdots \\ \dot{x}_{n-1} &= x_n + ug_{n-1}(x_1, \ldots, x_{n-1}) \\ \dot{x}_n &= f_n(x_1, \ldots, x_n) + ug_n(x_1, \ldots, x_n). \end{aligned}$$

(ii) *If Σ_{aff} admits, locally at q, the form* (UO) *then it is uniformly locally observable at q.*

(iii) *A necessary and sufficient condition for Σ_{aff} to admit locally at q the normal form* (UO) *is that* $\dim \operatorname{span}\{dh, \ldots, d\mathrm{L}_f^{n-1}h\}(q) = n$ *and that in a neighborhood of q*

$$[D_j, g] \subset D_j \ ,$$

for any $1 \leq j \leq n$, where $D_j = \ker\{dh, \ldots, d\mathrm{L}_f^{j-1}h\}$.

5.4 Local observability: a necessary and sufficient condition

Recall that the Hermann-Krener observability rank condition gives only a sufficient condition for (strong) local observability (compare Example 5.10). Following Bartosiewicz [1] we will provide in this section a necessary and sufficient condition for local observability.

Consider a nonlinear system Σ and assume that it is analytic, that is, X is an analytic manifold, the vector fields f_u are analytic and h is an analytic map.

We start with the following simple observation.

Proposition 5.22 *The points q_1 and q_2 are indistinguishable if and only if for any $\phi \in \mathcal{H}$ we have $\phi(q_1) = \phi(q_2)$.*

Introduce now the *observation algebra* of Σ. It is the smallest subalgebra over $\mathbb{R}$ of $C^\omega(X)$, the algebra of analytic functions on X, which contains h_i and is closed under Lie derivatives with respect to f_u, $u \in U$. We denote it by H_A. Observe that H_A consists of all elements of H and of all constant functions.

For $x \in X$, by $\mathcal{O}_x$ we denote the algebra over $\mathbb{R}$ of germs of analytic functions at x. Denote by m_x the unique maximal ideal of $\mathcal{O}_x$. It consists of all germs that vanish at x. For $x \in X$ we define I_x to be the ideal in $\mathcal{O}_x$ generated by germs of those functions from H_A which vanish at x. Of course, $I_x \subset m_x$. The *real radical* of an ideal I in a commutative ring R is

$$\sqrt[\mathbb{R}]{I} = \{a \in R \mid a^{2m} + b_1^2 + \cdots b_k^2 \in I \text{ for some } m > 0,\ k \geq 0,\ b_1, \ldots, b_k \in R\}.$$

Clearly, the real radical is an ideal.

Theorem 5.23 *The system Σ is locally observable at x if and only if*

$$\sqrt[\mathbb{R}]{I_x} = m_x.$$

Example 5.24 We can easily see that for the system $\dot{x} = 0$, $y = x^3$ (compare Example 5.10), which is clearly locally observable at any $x \in \mathbb{R}$, we have $\sqrt[\mathbb{R}]{I_x} = m_x$ for any $x \in \mathbb{R}$, in particular, for $x = 0$. □

5.5 Generic observability properties

In this section we discuss the problem of what observability properties are shared by generic control systems. We consider C^∞-Whitney topology for smooth systems. Recall that a sequence of smooth function φ_n on a manifold X converges in C^∞-Whitney topology to a smooth function φ if there exists a compact subset $C \subset X$ such that the all derivatives $\varphi_n^{(i)}$, for $i \geq 0$, converge uniformly on C to the corresponding $\varphi^{(i)}$ and $\varphi_n \equiv \varphi$ on $X \setminus C$, for all n sufficiently large. In the case of a compact state space X, it is just the topology of C^∞ uniform convergence on X. We start by presenting results of Jakubczyk and Tchoń [28] who classified uncontrolled observed dynamics

of the form

$$\Sigma \; : \; \begin{array}{rcl} \dot{x} & = & f(x) \; , \\ y & = & h(x) \; , \end{array}$$

where $x \in X$ and $y \in \mathbb{R}$, f is a smooth vector field and h is a smooth $\mathbb{R}$-valued function. Let Ξ denote the family of all systems Σ of the above form equipped with the C^∞-Whitney topology.

Theorem 5.25 *There exists an open and dense subset $\Xi_0 \subset \Xi$ such that any $\Sigma \in \Xi_0$ is locally equivalent at any $q \in X$ to one of the following normal forms.*

(i) *If $f(q) \neq 0$ then Σ is equivalent to*

$$h(x) = x_1^{r+1} + x_2 x_1^{r-1} + \cdots + x_r x_1 + \eta(x_2, \ldots, x_n), \tag{5.2}$$

$$f(x) = f_1(x_1, \ldots, x_n)\frac{\partial}{\partial x_1}, \tag{5.3}$$

where $0 \leq r \leq n$, and f_1 and η are C^∞-functions of the indicated arguments such that $f_1(0) > 0$.

(ii) *If $f(q) = 0$ then Σ is locally equivalent to*

$$h(x) = x_1 + c \; , \tag{5.4}$$

$$f(x) = x_2\frac{\partial}{\partial x_1} + \cdots + x_n\frac{\partial}{\partial x_{n-1}} + f_n(x_1, \ldots, x_n)\frac{\partial}{\partial x_n}, \tag{5.5}$$

where c is a constant and f_n is a C^∞-function such that $f_n(0) = 0$.

In the item (i) above, if $r = 1$ we can always take $h = x_1^2 + \eta$ while for $r = 0$ we take $h = x_1 + c$

Observe that in the case (i), for the "time-rescaled" system $\frac{dx}{d\tau} = \frac{1}{f_1(x)}f(x)$, where $d\tau = f_1(x(t))dt$, we have $x_i(\tau)$=constant, for $2 \leq i \leq n$, and thus in the new time scale

$$y(\tau) = h(x(\tau)) = \tau^{r+1} + x_2\tau^{r-1} + \cdots x_r\tau + c,$$

where $x_2 = c_2, \ldots, x_n = c_n$ and c are constants. It follows that, firstly, responses are polynomial with respect to the new time τ, with at most r different local extreme points. Secondly, there are always initial conditions, close to q, producing $y(\tau)$ with r different local extrema.

For systems which are not generic but satisfy the observability rank condition, an analogous normal form can be established.

Theorem 5.26 *If Σ satisfies the observability rank condition at q then it is locally equivalent either to the form* (5.4)-(5.5) *if $f(q) = 0$ or, otherwise, to one of the following normal forms*

$$h(x) = x_1^{r+1} + \phi_{r-1}x_1^{r-1} + \cdots + \phi_1 x_1 + \phi_0 + c, \tag{5.6}$$

$$f(x) = f_1(x_1, \ldots, x_n)\frac{\partial}{\partial x_1}, \tag{5.7}$$

where $r \geq 0$, and ϕ_i are C^∞-functions of $x_2, \ldots, x_n$, for $0 \leq i \leq r-1$, satisfying $\phi_i(0) = 0$, and f_1 is a C^∞-function such that $f_1 > 0$.

If $r = 1$ we can always take $h = x_1^2 + \phi_2$, while for $r = 0$ we take $h = x_1 + c$.

We end up this chapter by stating some results of Gauthier and Kupka devoted to the genericity of uniform observability. Consider an observed smooth control system of the form

$$\Sigma \; : \quad \begin{aligned} \dot{x} &= f(x,u), \\ y &= h(x,u) \end{aligned}$$

where $x \in X$, $u \in U$, and $y \in Y$. Notice that we assume the output $y = h(x,u)$ to depend explicitly on the control u.

Recall that for the system Σ we define the response map, called also input-output map

$$R_\Sigma : X \times \mathcal{U} \longrightarrow \mathcal{Y},$$

which to any initial condition $x_0 \in X$ and any admissible control $u(\cdot) \in \mathcal{U}$ attaches the output of the system

$$y(t, x_0, u(\cdot)) = h(x(t, x_0, u(\cdot)), u(t)).$$

The control $u(\cdot)$ being defined on an interval $0 \in I_u \subset \mathbb{R}$, we consider the output $y(\cdot)$ on the maximal interval $I_y \subset I_u \subset \mathbb{R}$ on which it exists.

In the remaining part of this section, we will assume that the state space X is a compact manifold and $U = J^m$, where J is some compact interval of $\mathbb{R}$. We denote by Ξ the class of such systems equipped with the topology of C^∞ uniform convergence on $X \times I^m$.

For any C^k-function $w(t)$ of time we will denote $\bar{w}^k(t) = (w(t), w'(t), \ldots, w^{(k)}(t))$. For the system Σ, for any integer k and for a C^k-differentiable input $u(t)$, we define the k-prolongation of the response map as

$$R_\Sigma^k(x_0, \bar{u}^k(t)) = (y(t), y'(t), \ldots, y^{(k)}(t)) = \bar{y}^k(t),$$

that is, as the vector formed by the output and its first k derivatives with respect to time t.

For an open subset W of $\mathbb{R}^q$ (or a differential manifold), and for a C^k-differentiable function w of $I \subset \mathbb{R}$ into W, such that $0 \in I$, we denote by $j^k w$ the k-jet at $0 \in \mathbb{R}$ of w. We will denote by $J^k W$ the space of k-jets at $0 \in \mathbb{R}$ of maps from I into W. Now we consider the k-jet

$$j^k R_\Sigma : X \times J^k U \longrightarrow J^k Y$$

of the map R_Σ defined by

$$j^k R_\Sigma(x_0, j^k u) = j^k y,$$

where $j^k y = \bar{y}^k(0)$, $\bar{y}^k(t) = R^k_\Sigma(x_0, \bar{u}^k(t))$, and $u(t)$ is any C^k-control such that $\bar{u}^k(0) = j^k u$.

The following fundamental result has been proved by Gauthier and Kupka [16], [17], [18].

Theorem 5.27 *Assume $p > m$, that is, the number of outputs is greater than that of inputs. Fix a sufficiently large positive integer k.*

(i) *The set of systems Σ such that $j^k R_\Sigma(\cdot, j^k u)$ is an immersion of X into $\mathbb{R}^{p(k+1)}$, for all $j^k u \in J^k U$, contains an open dense subset of Ξ.*

(ii) *The set of systems Σ such that $j^k R_\Sigma(\cdot, j^k u)$ is an embedding of X into $\mathbb{R}^{p(k+1)}$, for all $j^k u \in J^k U$, is a residual subset of Ξ.*

(iii) *For any compact subset C of $J^k U$, the set of systems Σ such that $j^k R^k_\Sigma(\cdot, j^k u)$ is an embedding, for all $j^k u \in C$, is open dense in Ξ.*

The above result implies that, in the case $p > m$, the set of systems that are observable for all C^k inputs is residual, that is, it is a countable intersection of open dense sets. If a bound on the derivatives of the controls is given a-priori, that is $\mid u^{(i)}(t) \mid \leq M$, for some constant M and any $0 \leq i \leq k$, then this set is open dense. If the number of outputs is not greater than that of controls all statements of the above theorem are *false*.

6 Decoupling

In this section we show how static feedback allows to transform the dynamics of a nonlinear system in order to achieve desired decoupling properties. In Section 6.1 we will introduce a crucial concept of invariant distributions. In Section 6.2 we consider disturbance decoupling while in Section 6.3 we deal with input-output decoupling.

6.1 Invariant distributions

Consider a smooth nonlinear control system of the form

$$\Sigma: \quad \dot{x} = g_0(x) + \sum_{i=1}^{m} g_i(x)u_i = g_0(x) + g(x)u,$$

where $x \in X$, $u \in \mathbb{R}^m$, $g = (g_1, \ldots, g_m)$ and $u = (u_1, \ldots, u_m)^T$. Notice that for simplicity we denote the drift of the system by $f = g_0$.

A distribution $\mathcal{D}$ is called *invariant* for Σ if

$$[g_i, \mathcal{D}] \subset \mathcal{D}, \quad \text{for } 0 \leq i \leq m.$$

If a distribution is not invariant for Σ it may become invariant under a suitable feedback modification. A distribution $\mathcal{D}$ is called *controlled invariant* if there exists an invertible feedback of the form

$$u = \alpha(x) + \beta(x)\tilde{u}, \quad \beta(\cdot) - \text{invertible},$$

such that $\mathcal{D}$ is invariant under the feedback modified dynamics

$$\dot{x} = \tilde{g}_0(x) + \sum_{i=1}^{m} \tilde{g}_i(x)\tilde{u}_i,$$

that is,

$$[\tilde{g}_i, \mathcal{D}] \subset \mathcal{D}\ ,$$

for $0 \leq i \leq m$, where

$$\tilde{g}_o = g_o + g\alpha, \qquad \tilde{g} = g\beta.$$

Example 6.1 In the case of a linear system of the form

$$\Lambda: \quad \dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m,$$

a subspace $V \subset \mathbb{R}^n$ is said to be invariant for Λ if $AV \subset V$. We say that V is controlled invariant (or (A, B)-invariant) if there exists a linear feedback of the form $u = Fx + G\tilde{u}$ such that

$$(A + BF)V \subset V.$$

Observe that in the linear case (A, B)-invariance does not depend on G so one can take $G = \mathrm{Id}$ or G-noninvertible.

One can check by a direct calculation that (A, B)-invariance is equivalent to

$$AV \subset V + \mathrm{Im}B. \tag{6.1}$$

We refer to [45] for an extensive treatment of the concept of invariance in the linear case.

Put $\mathcal{G} = \mathrm{span}\,\{g_1, \ldots, g_m\}$. In the nonlinear case, controlled invariance of a distribution $\mathcal{D}$ implies the following property of *local controlled invariance* (compare (6.1))

$$[g_i, \mathcal{D}] \subset \mathcal{D} + \mathcal{G}, \quad \text{for } 0 \leq i \leq m.$$

For involutive distributions the converse holds locally under regularity assumptions, see [20],[25],[35].

Proposition 6.2 *Assume that the distributions $\mathcal{D}$, $\mathcal{G}$, and $\mathcal{D} \cap \mathcal{G}$ are of constant rank. If $\mathcal{D}$ is involutive and locally controlled invariant then it is controlled invariant, locally at any point $x \in X$.*

6.2 Disturbance decoupling

In this Section we apply the concept of controlled invariant distributions to solve the nonlinear disturbance decoupling problem. Consider the following nonlinear system with output affected by disturbances $d = (d_1, \ldots, d_k)^T$, which are assumed to be bounded measurable $\mathbb{R}^k$-valued functions of time.

$$\Sigma_{\mathrm{dist}}: \quad \begin{aligned} \dot{x} &= g_0(x) + \sum_{i=1}^{m} g_i(x)u_i + \sum_{i=1}^{k} q_i(x)d_i = g_0(x) + g(x)u + q(x)d \\ y &= h(x)\ , \end{aligned}$$

where $x \in X$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, and $d \in \mathbb{R}^k$. All data are smooth, i.e., $f, g_1, \ldots, g_m \in V^\infty(X)$, $h_i \in C^\infty(X)$, where $h = (h_1, \ldots, h_p)^T$, and $q_1, \ldots, q_k \in V^\infty(X)$. We denote $q = (q_1, \ldots, q_k)$ and call them disturbance vector fields.

We say that the *disturbance decoupling problem*, shortly *DDP*, is solvable, if there exists an invertible feedback of the form $u = \alpha(x) + \beta(x)\tilde{u}$ such that the output $y(t) = h(x(t))$ of the feedback modified system

$$\dot{x} = \tilde{g}_0(x) + \sum_{i=1}^{m} \tilde{g}_i(x)\tilde{u}_i + \sum_{i=1}^{k} q_i(x)d_i$$

does not depend on the disturbances $d(t)$. By the latter we mean that

$$y(t, q, \tilde{u}(\cdot), d(\cdot)) \equiv y(t, q, \tilde{u}(\cdot), \tilde{d}(\cdot)),$$

for any initial condition $q \in X$, any control $\tilde{u}(\cdot) \in \mathcal{U}$, and any disturbances $d(\cdot)$ and $\tilde{d}(\cdot)$.

Put $\mathcal{Q} = \operatorname{span}\{q_1, \ldots, q_k\}$. The following result has been proved by Isidori et al [25].

Theorem 6.3 *If DDP is solvable then there exists an involutive controlled invariant distribution $\mathcal{V}$ such that*

$$\mathcal{Q} \subset \mathcal{V} \subset \ker dh.$$

This result suggests the following approach to *DDP*. Look for the maximal controlled invariant distribution in $\ker dh$ and check whether it contains $\mathcal{Q}$. In general, however, such a maximal distribution may not exist. Moreover, even if it exists and contains the disturbance vector fields it is not necessarily true that *DDP* is solvable. On the other hand there always exists $\mathcal{V}^*$, the *maximal locally controlled invariant distribution* in $\ker dh$, which leads to the following solution of *DDP* (see [20] and [25]).

Theorem 6.4 *Assume that the distributions $\mathcal{V}^*, \mathcal{V}^* \cap \mathcal{G}$, and $\mathcal{G}$ are of constant rank. If*

$$\mathcal{Q} \subset \mathcal{V}^*,$$

then DDP is solvable, locally, around any point of X.

The structure of the decoupled system can be described as follows. Let (α, β) be an invertible feedback which locally renders the distribution $\mathcal{V}^*$ invariant (it always exists under the regularity assumptions of Theorem 6.4, see Proposition 6.2). Let $x = (x^1, x^2)$ be local coordinates, with x^1, x^2 being possibly vectors, such that $\mathcal{V}^* = \operatorname{span}\{\frac{\partial}{\partial x^2}\}$. Then the feedback modified system reads as

$$\Sigma_{\text{dist}}: \quad \begin{aligned} \dot{x}^1 &= \tilde{g}_0^1(x^1) + \tilde{g}^1(x^1)\tilde{u} \\ \dot{x}^2 &= \tilde{g}_0^2(x^1, x^2) + \tilde{g}^2(x^1, x^2)\tilde{u} + q^2(x^1, x^2)d \\ y &= h^1(x^1)\ , \end{aligned}$$

where $\tilde{f} = f + g\alpha$ and $\tilde{g} = g\beta$. Now it is clear, compare Section 5.2, that the output $y(t)$ of the system does not depend on $d(t)$ since the latter affects the x^2-part of the system only which, in turn, is not observed by the output y.

Example 6.5 Consider the linear system with disturbances

$$\Lambda_{\text{dist}}: \quad \begin{array}{rcl} \dot{x} & = & Ax + Bu + Ed \\ y & = & Cx \ , \end{array}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $d \in \mathbb{R}^k$, and d denotes the disturbances. DDP is solvable if and only if $\operatorname{Im} E \subset V^*$, where V^* is the largest controlled invariant subspace in $\ker C$ (compare [45]). □

Example 6.6 Consider a particle of unit mass moving on the surface of a cylinder according to a potential force given by the potential function V (see [37])

$$\begin{array}{rclcrcl} \dot{q}_1 & = & p_1 & & \dot{q}_2 & = & p_2 \\ \dot{p}_1 & = & -\frac{\partial V}{\partial q_1}(q_1, q_2) + u & & \dot{p}_2 & = & -\frac{\partial V}{\partial q_2}(q_1, q_2) + d \ , \end{array}$$

where $(q_1, q_2, p_1, p_2) \in T(S^1 \times \mathbb{R})$. Let the output be given as $y = q_1$. We can see that $\mathcal{V}^* = \operatorname{span}\{\frac{\partial}{\partial q_2}, \frac{\partial}{\partial p_2}\}$. Moreover, the disturbance vector field $\frac{\partial}{\partial p_2} \in \mathcal{V}^*$ and hence DDP is solvable by the feedback $u = \frac{\partial V}{\partial q_1}(q_1, q_2) + \tilde{u}$. □

6.3 Input-output decoupling

Consider a smooth nonlinear control affine system with outputs of the form

$$\Sigma_{\text{aff}}: \quad \begin{array}{rcl} \dot{x} & = & f(x) + \sum_{i=1}^m u_i g_i(x) \\ y & = & h(x) \ , \end{array}$$

where $x \in X$, $u \in \mathbb{R}^m$, and $y \in R^p$.

We say that the *input-output decoupling problem* (called also *I-O decoupling problem* or *noninteracting problem*) is solvable for Σ if there exists an invertible feedback of the form $u = \alpha(x) + \beta(x)\tilde{u}$ such that the feedback modified system $\dot{x} = \tilde{f}(x) + \sum_{i=1}^m \tilde{u}_i \tilde{g}_i(x)$ with $y = h(x)$, where $\tilde{f} = f + g\alpha$, $\tilde{g} = g\beta$, satisfies

$$y_i^{(k_i)} = \tilde{u}_i, \text{ for } 1 \leq i \leq p, \tag{6.2}$$

for suitable nonnegative integers k_i. Observe that we assume that the input-output map of the modified system is linear. Therefore there is no lose of generality in assuming the form (6.2) because if the transfer matrix of the input-output response is diagonal (which is the usual definition of noninteracting) we can always achieve (6.2) by applying a suitable linear feedback.

Fix an initial condition $x_0 \in X$. For each output channel we define its *relative degree* ρ_i, called also *characteristic number*, to be the smallest integer such that for any neighborhood V_{x_0} of x_0

$$L_{g_j} L_f^{\rho_i - 1} h_i(x) \neq 0,$$

for some $1 \leq j \leq m$ and for some $x \in V_{x_0}$. By $L_f^{\rho} h$ we will mean the vector of p smooth functions whose i-entry is $L_f^{\rho_i} h_i$.

Define the $(p \times m)$ decoupling matrix $D(x)$, denoted also by $L_g L_f^{\rho} h$, whose (i,j)-entry is

$$L_{g_j} L_f^{\rho_i - 1} h_i(x).$$

Theorem 6.7 *Consider a control affine system* Σ_{aff}.

(i) *The system* Σ_{aff} *is input-output decouplable at* x_0 *via an invertible feedback of the form* $u = \alpha(x) + \beta(x)\tilde{u}$ *if and only if*

$$\operatorname{rank} D(x_0) = p.$$

(ii) *Moreover, for the square system, i.e.,* $m = p$, *the feedback*

$$u = -(L_g L_f^{\rho - 1} h)^{-1} L_f^{\rho} h + (\mathrm{L}_g L_f^{\rho - 1} h)^{-1} \tilde{u} \tag{6.3}$$

yields $y_i^{(k_i)} = \tilde{u}_i$, *where* $k_i = \rho_i$, *for* $1 \leq i \leq p$.

Remark 6.8 Inverting formula (6.3), we get the following expression for the new controls

$$\tilde{u}_i = L_f^{\rho_i} h_i + \sum_{j=1}^{m} u_j L_{g_j} L_f^{\rho_i - 1} h_i,$$

for $1 \leq i \leq m = p$. An analogous formula holds also in the non-square case. Indeed, if the system satisfies the decoupling condition $\operatorname{rank} D(x_0) = p$, then we can assume after a permutation of controls, if necessary, that the first p columns of the matrix $D(x_0)$ are independent. Then a decoupling feedback can be taken as

$$\begin{array}{rcll} \tilde{u}_i & = & L_f^{\rho_i} h_i + \sum_{j=1}^{m} u_j L_{g_j} L_f^{\rho_i - 1} h_i, & \text{for } 1 \leq i \leq p, \\ \tilde{u}_i & = & u_i, & \text{for } p+1 \leq i \leq m. \end{array} \tag{6.4}$$

Example 6.9 Consider the following rigid two-link robot manipulator or, in other words, double pendulum, (see [37], compare also Example 4.7)

$$\begin{aligned} \dot{x}^1 &= x^2 \\ \dot{x}^2 &= -M(x^1)^{-1}(C(x^1, x^2) + k(x^1)) + M(x^1)^{-1}u \ , \end{aligned}$$

where $x^1 = \theta = (\theta_1, \theta_2)^T$, $x^2 = \dot{\theta} = (\dot{\theta}_1, \dot{\theta}_2)^T$, $u = (u_1, u_2)^T$. The term $k(\theta)$ represents the gravitational force and the term $C(\theta, \dot{\theta})$ reflects the centripetal and Coriolis forces, and the positive definite symmetric matrix $M(x^1)$ is given by

$$\begin{pmatrix} m_1 l_1^2 + m_2 l_1^2 + m_2 l_2^2 + 2m_2 l_1 l_2 \cos\theta_2 & m_2 l_2^2 + m_2 l_1 l_2 \cos\theta_2 \\ m_2 l_2^2 + m_2 l_1 l_2 \cos\theta_2 & m_2 l_2^2 \end{pmatrix}.$$

As the outputs we take the cartesian coordinates of the endpoint

$$\begin{aligned} y_1 &= h_1(\theta_1, \theta_2) = l_1 \sin\theta_1 + l_2 \sin(\theta_1 + \theta_2) \\ y_2 &= h_2(\theta_1, \theta_2) = l_1 \cos\theta_1 + l_2 \cos(\theta_1 + \theta_2). \end{aligned}$$

By a direct computation we get $\rho_1 = \rho_2 = 2$ and $\operatorname{rank} D(x) = 2$ if and only if $l_1 l_2 \sin\theta_2 \neq 0$. Thus the system is input-output decouplable if $\theta_2 \neq k\pi$, that is, we have to exclude configurations at which the two robot arms are parallel. □

Example 6.10 Consider the unicycle (compare Examples 4.11 and 5.6) and assume that we observe the x_1-cartesian coordinate and the angle θ

$$\begin{aligned} \dot{x}_1 &= u_1 \cos\theta \qquad y_1 = x_1 \\ \dot{x}_2 &= u_1 \sin\theta \\ \dot{\theta} &= u_2 \qquad\quad y_2 = \theta. \end{aligned}$$

The control u_2 has a direct impact on the second component y_2 of the output as well as, through $\cos\theta$, on the first component. Thus the system is not input-output decoupled but it can be decoupled via a static feedback. We obviously have $\rho_1 = \rho_2 = 1$ and the decoupling matrix is

$$D = \begin{pmatrix} \cos\theta & 0 \\ 0 & 1 \end{pmatrix}.$$

Therefore the system is input-output decouplable at all points such that $\theta \neq \frac{\pi}{2} + k\pi$. □

Example 6.11 Consider the same dynamics of the unicycle and suppose that this time we observe x_1 and x_2

$$\begin{array}{rclrcl} \dot{x}_1 &=& u_1 \cos\theta & y_1 &=& x_1 \\ \dot{x}_2 &=& u_1 \sin\theta & y_2 &=& x_2 \\ \dot{\theta} &=& u_2. & & & \end{array}$$

Obviously, we have $\rho_1 = \rho_2 = 1$ but this time the decoupling matrix

$$D = \begin{pmatrix} \cos\theta & 0 \\ \sin\theta & 0 \end{pmatrix}$$

is of rank one everywhere and thus the system is not *I-O* decouplable. □

If the system is *I-O* decouplable then it is straightforward to calculate $\mathcal{V}^*$, the maximal locally controlled invariant distribution in $\ker dh$ (compare Section 6.2 and, consequently, solve the *DDP* problem.

Proposition 6.12 *Consider the system* Σ_{dist}. *Assume that the undisturbed system, that is, when* $d_i = 0$, *for* $1 \leq i \leq k$, *is input-output decouplable. Then*

(i) $\mathcal{V}^* = \mathcal{P}^\perp$, *where* $\mathcal{P} = \operatorname{span}\{dL_f^j h_i,\ 1 \leq i \leq p,\ 0 \leq j \leq \rho_i - 1\}$.

(ii) *If, moreover,* $\mathcal{Q} \subset \mathcal{P}^\perp$ *then the DDP problem is solvable and the feedback* (6.4) *simultaneously decouples the disturbances and renders the system input-output decoupled and input-output linear.*

Example 6.13 To illustrate this result let us consider the following model of the unicycle. We suppose that the dynamics is affected by a disturbing rotation (of un unknown varying strength $d(t)$) and that we measure the angle and the square of the distance from the origin:

$$\begin{array}{rclrcl} \dot{x}_1 &=& u_1 \cos\theta + x_2 d & y_1 &=& x_1^2 + x_2^2 \\ \dot{x}_2 &=& u_1 \sin\theta - x_1 d & & & \\ \dot{\theta} &=& u_2 & y_2 &=& \theta. \end{array}$$

The decoupling matrix is

$$D = \begin{pmatrix} 2x_1 \cos\theta + 2x_2 \sin\theta & 0 \\ 0 & 1 \end{pmatrix}$$

and is of rank two at any point away from $N = \{(x_1, x_2, \theta) \mid x_1 \cos\theta + x_2 \sin\theta = 0\}$. Notice that N consists of points where the direction of the

unicycle is perpendicular to the ray from the origin passing through the center of the unicycle. At points of $(\mathbb{R}^2 \times S^1) \setminus N$, the system is input-output decouplable and, moreover, $\mathcal{P} = \operatorname{span}\{x_1 dx_1 + x_2 dx_2, d\theta\}$. The vector field $q = x_2 \frac{\partial}{\partial x_1} - x_1 \frac{\partial}{\partial x_2}$ is annihilated by $\mathcal{P}$ and thus the feedback $u_1 = (2x_1 \cos\theta + 2x_2 \sin\theta)^{-1} \tilde{u}_1$ and $u_2 = \tilde{u}_2$ decouples the disturbances from the output yielding an I-O decoupled and I-O linear system expressed, in (R, φ, θ)-coordinates, where $R = r^2 = x_1^2 + x_2^2$, $x_1 = r\cos\varphi$, $x_2 = r\sin\varphi$, by

$$\begin{array}{rclrcl} \dot{R} &=& \tilde{u}_1 & y_1 &=& R \\ \dot{\varphi} &=& \frac{1}{2r^2}\tilde{u}_1 \tan(\theta - \varphi) - d & & & \\ \dot{\theta} &=& \tilde{u}_2 & y_2 &=& \theta. \end{array}$$

The disturbance d does not affect the output $(y_1, y_2) = (R, \theta)$. □

References

[1] Z. Bartosiewicz, Local observability of nonlinear systems, *Syst. Contr. Lett.* **25** (1995), 295-298.

[2] Battilotti, *Noninteracting Control with Stability for Nonlinear Systems*, Springer-Verlag, London, 1994.

[3] R. W. Brockett, Feedback invariants for nonlinear systems, *IFAC Congress* **6**, Helsinki, 1978, 1115-1120,.

[4] R. W. Brockett, Asymptotic stability and feedback stabilization, in *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman, and H. J. Sussmann (eds.), Birkhäuser, Boston, 1983, 181-191.

[5] P. Brunovský, A classification of linear controllable systems, *Kybernetica* **6** (1970), 173-188.

[6] J. J. Craig, *Introduction to Robotics, Mechanics and Control*, Addison Wesley, Reading, 1986.

[7] P. E. Crouch and B. Bonnard, An appraisal of linear analytic system theory with applications to attitude control, *ESA ESTEC Contract reprint* (1980).

[8] P. E. Crouch, Spacecraft attitude control and stabilization: applications of geometric control theory to rigid body models, *IEEE Tr. Aut. Contr.* **AC-29** (1984), 321-331.

[9] W. P. Dayawansa, W. M. Boothby, and D. Elliot, Global state and feedback equivalence of nonlinear systems, *Syst. Contr. Lett.* **6** (1985), 229-234.

[10] J. Descusse and C. H. Moog, Decoupling with dynamic compensation for strong invertible affine nonlinear systems, *Int. J. Control* **43** (1985), 1387-1398.

[11] S. Diop and M. Fliess, On nonlinear observability, in *Proc. of the* 1^{st} *ECC*, Hermès, Paris, 1991, 152-157.

[12] S. Diop and M. Fliess, Nonlinear observability, identifiability, and persistent trajectories, *Proc. of the 30th CDC*, 1991, 749-719.

[13] S. Diop and Y. Wang, Equivalence of algebraic and local generic observability, *Proc. of the 30th CDC*, 1993, 2864-2865.

[14] J. P. Gauthier and G. Bornard, Observability for any u(t) of a class of nonlinear systems, *IEEE Tr. Automat. Contr.* **26** (1981), 922-926.

[15] J. P. Gauthier, H. Hammouri, and S. Othman, A simple observer for nonlinear systems, applications to bioreactors, *IEEE Tr. Automat. Contr.* **37** (1992), 875-880.

[16] J. P. Gauthier and I. Kupka, Genericity of observability and the existence of asymptotic observers, in *Geometry in Nonlinear Control and Differential Inclusions*, B. Jakubczyk, W. Respondek, and T. Rzezuchowski (eds.), Banach Center Publications vol. 32, Warsaw, 1995, 227-244.

[17] J. P. Gauthier and I. Kupka, Observability for systems with more outputs than inputs and asymptotic observers, *Matematische Zeitschriften* **223** (1996), 47-78.

[18] J. P. Gauthier and I. Kupka, *Deterministic Observation Theory and Applications*, Cambridge University Press, N.Y., 2001.

[19] I. J. Ha and E. G. Gilbert, A complete characterization of decoupling control laws for a general class of nonlinear systems, *IEEE Tr.. Aut. Contr.* **AC-31** (1986), 823-830.

[20] R. M. Hirschorn, (A,B)-invariant distributions and disturbance decoupling of nonlinear systems, *SIAM J. Contr. Optimiz.* **19** (1981), 1-19.

[21] R. Hermann and A. J. Krener, Nonlinear controllability and observability, *IEEE Trans. Aut. Contr.* **AC-22** (1977), 728-740.

[22] L. R. Hunt and R. Su, Linear equivalents of nonlinear time varying systems, *Proc. of the MTNS*, Santa Monica, 1981, 119-123.

[23] A. Isidori, *Nonlinear Control Systems*, 3rd edition, Springer-Verlag, London, 1995.

[24] A. Isidori and J. W. Grizzle, Fixed modes and nonlinear noninteracting control with stability, *IEEE Tr. Aut. Contr.* **AC-33** (1988), 907-914.

[25] A. Isidori, A. J. Krener, C. Gori Giorgi, and S. Monaco, Nonlinear decoupling via feedback: a differential geometric approach, *IEEE Trans. Aut. Contr.* **AC-26** (1981), 331-345.

[26] B. Jakubczyk, Introduction to geometric nonlinear control; Controllability and Lie bracket, this volume.

[27] B. Jakubczyk and W. Respondek, On linearization of control systems, *Bull. Acad. Polonaise Sci., Ser. Sci. Math.*, **28** (1980), 517-522.

[28] B. Jakubczyk and K. Tchoń, Singularities and normal forms of observed dynamics, *Math. Control Signal Systems* **2** (1989), 19-31.

[29] V. Jurdjevic, *Geometric Control Theory*, Studies in Advanced Math., 52, Cambridge University Press, N.Y., 1997.

[30] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[31] A. J. Krener, On the equivalence of control systems and linearization of nonlinear systems, *SIAM J. Contr.* **11** (1973), 670-676.

[32] A. J. Krener, A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems, *SIAM J. Contr.* **12** (1974), 43-52.

[33] A. J. Krener, A. Isidori, and W. Respondek, Partial and robust linearization by feedback, *Proc. 22nd CDC, San Antonio*, (1983), 126-130.

[34] R. Marino, On the largest feedback linearizable subsystem, *Syst. Contr. Lett.* **7**, (1986), 345-351.

[35] H. Nijmeijer, Controlled invariance for affine control systems, *Int. J. Contr.* **34** (1981), 824-833.

[36] H. Nijmeijer and W. Respondek Dynamic input-output decoupling of nonlinear control systems, *IEEE Trans. Aut. Contr.* **AC-33**, 1988, 1065-1070.

[37] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[38] W. Respondek, Geometric methods in linearization of control systems, in *Mathematical Control Theory*, Cz. Olech, B. Jakubczyk, and J. Zabczyk (eds.), Banach Center Publications, vol. 14, PWN-Polish Scientific Publishers, Warsaw, 1985, 453-467.

[39] W. Respondek, Global aspects of linearization, equivalence to polynomial forms and decomposition of nonlinear control systems, in *Algebraic and Geometric Methods in Nonlinear Control Theory*, M. Fliess and M. Hazewinkel (eds.), Reidel, Dordrecht, 1986, 257-284.

[40] W. Respondek, Partial linearization, decompositions and fibre linear systems, in *Theory and Applications of Nonlinear Control Systems*, C.I. Byrnes and A. Lindquist (eds.), North-Holland, Amsterdam, 1986, 137-154.

[41] H. J. Sussmann, Orbits of families of vector fields and integrability of distributions, *Trans. Am. Math. Soc.* **180** (1973), 171-180.

[42] H. J. Sussmann, Lie brackets, real analyticity, and geometric control, in *Differential Geometric Control Theory*, R. W. Brockett, R. S. Millman, and H. J. Sussmann (eds.), Birkhäuser, Boston, 1983, 1-116.

[43] H. J. Sussmann and V. Jurdjevic, Controllability of nonlinear systems, *J. Diff. Eqs.* **12** (1972), 95-116.

[44] A. J. van der Schaft, Linearization and input-output decoupling for general nonlinear systems, *Syst. Contr. Lett.* **5** (1984), 27-33.

[45] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, Berlin, 1979.

[46] M. Zribi and J. Chiasson, Exact linearization control of a PM stepper motor, *Proc. American Control Conference*, Pittsburgh, 1989.

The Combinatorics of Nonlinear Controllability and Noncommuting Flows

Matthias Kawski*

Department of Mathematics, Arizona State University, Tempe, Arizona, USA

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028005

*kawski@asu.edu

Abstract

These notes accompany four lectures, giving an introduction to new developments in, and tools for problems in nonlinear control. Roughly speaking, after the successful development, starting in the 1960s, of methods from linear algebra, complex analysis and functional analysis for solving linear control problems, the 1970s and 1980s saw the emergence of differential geometric tools that were to mimic that success for nonlinear systems. In the past 30 years this theory has matured, and now connects with many other branches of mathematics.

The focus of these notes is the role of algebraic combinatorics for both illuminating structures and providing computational tools for nonlinear systems. On the control side, we focus on problems connected with controllability, although the combinatorial tools obviously have just as much use for other control problems, including e.g. path-planning, realization theory, and observability.

The lectures are meant to be an introduction, sketching the road from the comparatively naive, bare-handed constructions used in the early years, to the elegant and powerful insights from recent years. One of the main targets is to develop an explicit, continuous analogue of the classical Campbell-Baker-Hausdorff formula, and of a related exponential product expansion. The purpose of such formulae is to separate the time-dependent and control-dependent parts of solution curves from the invariant underlying geometrical structure inherent in each control system.

The key theme is that effective tools (including effective notation) from algebraic combinatorics are essential, for both theoretical analysis and for practical computation (beyond some miniscule academic examples). On a practical level we want the reader to take home the message to never write out complicated iterated integrals, as it is both a waste of paper and time, as it obscures the underlying structure. On the theoretical level, the key object is the chronological algebra isomorphism from the free chronological algebra to an algebra of iterated integral functionals, denoted by Υ in our exposition.

Reiterating, these notes are meant to be an introduction. As such, they provide many examples and exercises, and they emphasize as much getting a hands-on experience and intuitive understanding of various structural terms, as they are meant to establish the need for, and appreciation of tools from algebraic combinatorics. We leave a formal treatment of the abstract structures and isomorphism to future lectures, and until then refer the reader to pertinent recent literature.

Contents

0 Organization and objectives **227**

1 Nonlinear controllability **228**
1.1 Introductory examples 228
1.2 Controllability 232
1.3 Piecewise constant controls and the CBH formula 236
1.4 Approximating cones and conditions for STLC 242

2 Series expansion, nilpotent approximating systems **247**
2.1 Introduction to the Chen Fliess series 247
2.2 Families of dilations 250
2.3 Nilpotent approximating systems 254

3 Combinatorics of words and free Lie algebras **259**
3.1 Intro: Trying to partially factor the Chen Fliess series 259
3.2 Combinatorics and various algebras of words 265
3.3 Hall Viennot bases for free Lie algebras 274

4 A primer on exponential product expansions **282**
4.1 Ree's theorem and exponential Lie series 282
4.2 From infinite series to infinite products 286
4.3 Sussmann's exponential product expansion 292
4.4 Free nilpotent systems 300

References **306**

0 Organization and objectives

These notes contain the background information and the contents (in roughly the same order) of four 75 minute lectures given during the 2001 summer school on mathematical control. They shall provide an introduction to nonlinear controllability and the algebraic-combinatorial tools used to study it. An effort is made to keep the level elementary, assuming familiarity primarily with the theory of differential equations and knowledge from selected preceding lectures in this summer school that addressed geometric methods in control, and an introduction to nonlinear control systems. Consequently, in several places a comparatively *"pedestrian approach"* is taken which may not be the cleanest or most elegant formulation, as the latter may typically presume more advanced ways of thinking in differential geometry or algebraic combinatorics. However, in most such places comments point to places in the literature where more advanced approaches may be found.
Similarly, proofs are given or sketched where they are illuminating and of reasonable length when using tools at the level of this course. In other cases comments refer to the literature where detailed, or more efficient proofs may be found.

Several examples are provided, and revisited frequently, both in order to provide motivation, and to provide the hands-on experience that is so important for making sense of otherwise abstract recipes, and to provide the ground for further developments. In this sense, the exercises imbedded in the notes are an essential component and the reader is urged to get her/his hands *dirty* by working out the details.

Aside from providing an introductory survey of some aspects of modern differential geometric control theory, the overarching objective is to develop a sense of necessity, and an appreciation of the algebraic and combinatorial tools, which provide as much an elegant algebraization of the theory as they provide the essential means that allow one to carry out real calculations that without these tools would be practically almost impossible.

1 Nonlinear controllability

1.1 Introductory examples

The problem of *parallel parking a car* provides one of the most intuitive introductions to many aspects of nonlinear control, especially controllability, and it may be analyzed at many different levels. Here we introduce a simplified version of the problem, and use it to motivate questions which naturally beg for generalization. The example will be revisited in later sections as a model case on which to try out newly developed tools and algorithms.

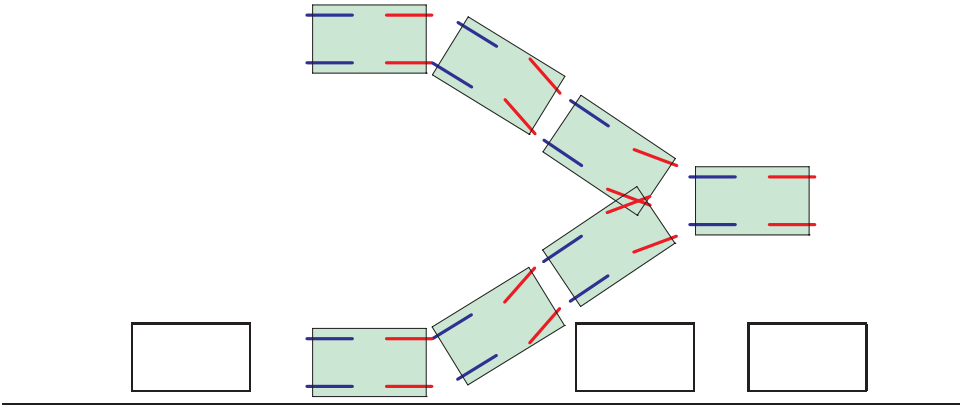


Figure 1. Parallel parking a car (exaggerated parallel displacement)

Example 1.1

Think about driving a real car, and the experience of parallel parking a car in an empty spot along the edge of the road. If the open gap is large, this is very easy – but it becomes more challenging when the length of the gap is just barely larger than the length of the car. For the sake of definiteness, suppose the initial position and orientation of the car as indicated in the diagram (with much exaggerated parallel displacement, and an exaggerated length of the gap), with steering wheels in the direction of the road.
Everyday experience says that, while it is impossible to directly move the car sideways, it is possible to do so indirectly via a series of careful maneuvers

that involve going back and fourth with suitably matching motions of the steering wheels.
One may consider different choices as possible controls. In this case let us use the forward acceleration of the rear wheels as one control, and the steering angle as a second control.

Exercise 1.1 *Develop different possible series of maneuvers that result in a car that is in the same location, with zero speed, but rotated by* $\frac{\pi}{2}$ *of by* π. *Describe the maneuvers verbally, and sketch the states as functions of time*

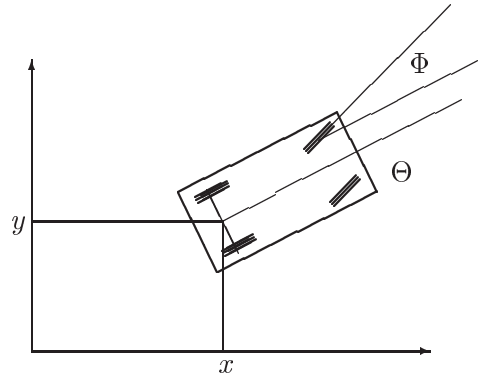


Figure 2. Defining the *states* of the system

To obtain a mathematical model consider the simpler (less controversial case as it does not require *differentials* to adjust for the different speeds of *inside* and *outside* wheels) of a bicycle! In particular, let $(x, y) \in \mathbb{R}^2$ denote the point of contact of the rear wheel with the plane (center of rear axle in the case of a car). Let $\theta \in S^1$ be the angle of the bicycle with the x_1-axis, and by $\phi \in S^1$ the angle of the front wheel(s) with the direction of the bicycle. An algebraic constraint captures that the distance between front and rear wheel is constant, equal to the length L. Thus the position of the front wheel (point of contact with plane) is $(x + L\cos\theta, y + L\sin\theta)$. The conditions that the wheels can slip neither forward nor sideways, each can only roll in the direction of the wheel is captured in

$$\left\{ \begin{array}{rcl} 0 & = & \cos\theta\, dy - \sin\theta\, dx \\ 0 & = & \sin(\theta+\phi)\, d(x + L\cos\theta) - \cos(\theta+\phi)\, d(y + L\sin\theta) \end{array} \right. \tag{1}$$

Introducing the speed $v = \|\dot{x}^2 + \dot{x}^2\|$ of the rear wheel (or of the center of the rear axle), write $\dot{x} = v\cos\theta$ and $\dot{y} = v\sin\theta$.

Exercise 1.2 *Discuss what happens in this model when the forward speed of the rear wheel is zero and the angle of the steering wheel is* $\phi = \pi/2$. *Can the bicycle move?*
Develop an alternative front-wheel drive *model, i.e. with controlled speed* v *of front wheel. Continue working that model in parallel to the one discussed here in the notes.*

Using the first constraint, solve the second constraint for

$$d\theta = \frac{v\,dt}{L} \cdot \frac{\cos\theta \cdot \tan(\theta+\phi) - \sin\theta}{\cos\theta + \tan(\theta+\phi)\sin\theta} = \frac{v}{L} \cdot \tan\phi\, dt \tag{2}$$

(The last step is immediate from basic trigonometric identities after multiplying numerator and denominator by $\cos(\theta+\phi)$.) Write the model as a system of controlled ordinary differential equations (for simplicity we choose units such that $L = 1$)

$$\begin{cases} \dot\phi &= u_1 \\ \dot v &= u_2 \\ \dot x &= v\cos\theta \\ \dot\theta &= v\tan\phi \\ \dot y &= v\sin\theta \end{cases} \tag{3}$$

Exercise 1.3 *Using your practical driving experience, suggest specific control functions* u_1, u_2 *(e.g. piecewise constant or sinusoidal, with switching times as parameters to be determined) such that the corresponding solution* steers *the system from* $(\phi, v, x, \theta, y)(0) = (0,0,0,0,0)$ *to* $(\phi, v, x, \theta, y)(T) = (0,0,0,0,H)$ *for some* $T > 0$ *and* $H \neq 0$.
Sketch the graphs of the states as functions of time (compare figure 3).

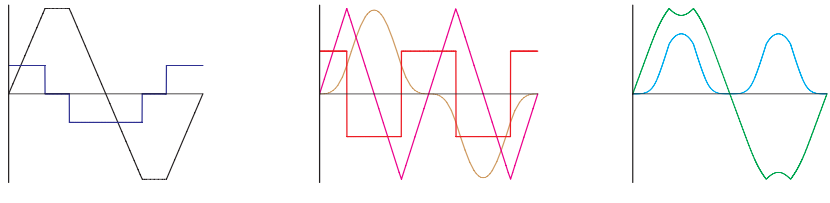


Figure 3. One possible, very symmetric, parallel parking maneuver

Exercise 1.4 *In figure 3, identify which curve represents which state or control.*

Another well-studied [3] introductory example is that of a *rolling penny* in he plane.

Example 1.2

Consider a disk of radius a and negligible thickness *standing on its edge* that may roll without slipping in the plane, and which may rotate about its vertical axis. Denoting by $(x_1, x_2) \in \mathbb{R}^2$ its point of contact with the plane, by $\theta \in S^1$ its angle with the x_1-axis, and by $\phi \in S^1$ its *rolling* angle from a fixed reference angle, the non-slip constraints may be written as:

$$\begin{cases} \cos\theta\, dx_1 + \sin\theta\, dx_2 &= a\, d\phi \\ \sin\theta\, dx_1 - \cos\theta\, dx_2 &= 0 \end{cases} \tag{4}$$

Equivalently, considering the angular velocities as controls the system is written as

$$\begin{cases} \dot{\phi} &= u_1 \\ \dot{\theta} &= u_2 \\ \dot{x}_1 &= a u_1 \cos\theta \\ \dot{x}_2 &= a u_1 \sin\Theta \end{cases} \tag{5}$$

Alternatively, considering the accelerations, or rather the torques as controls (suitably scaled), the system is described by

$$\begin{cases} \dot{\omega}_1 &= u_1 \\ \dot{\omega}_2 &= u_2 \\ \dot{\phi} &= \omega_1 \\ \dot{\theta} &= \omega_2 \\ \dot{x}_1 &= a u_1 \cos\theta \\ \dot{x}_2 &= a u_1 \sin\theta \end{cases} \tag{6}$$

One of the more intriguing question is whether it is possible to roll, and turn the penny in such a way that at the end it is back at its original location with original orientation but rotated about a desired angle about its horizontal axis.
Moreover, one may ask if it always possible to achieve such a reorientation without moving far from the starting state. Alternatively, one may ask whether one can in any arbitrarily small time interval achieve at least a small reorientation.

Exercise 1.5 *Develop an intuitive strategy that results in such a reorientation. I.e. describe the maneuver in words, and sketch the general shapes of the states as functions of time.*

Exercise 1.6 *Develop an intuitive strategy that results in such a reorientation. I.e. describe the maneuver in words, and sketch the general shapes of the states as functions of time.*

Exercise 1.7 *Find an analytic solution using piecewise constant controls defined on an arbitrary short time-interval* $[0, T]$ *that rotates the penny by a given angle* $\varepsilon \in \mathbb{R}$.

Exercise 1.8 *Repeat the previous exercise using controls that are piecewise trigonometric functions of time, or that are trigonometric polynomials.*

With such mechanical examples as there is no question about the model and we concentrate on the analysis and geometry. But the methodology developed in sequel is just applicable to controlled dynamical systems that arise in electric and communication networks, in biological and bio-medical systems, in macro-economic and financial systems etc.

1.2 Controllability

For a given control $u(t)$, a control system $\dot{x} = f(x, u)$ with initial value $x(0)$ is simply a dynamical system, which is straightforward to analyze and *solve* using basic techniques from differential equations. What makes control so much more intellectually challenging is the inverse nature of most questions – e.g. given a target $x(T)$, *find* a control u that steers from $x(0)$ to $x(T)$. The first step, before one may start any construction or optimization, is to ask whether there exists any solution in the first place. This is the question about controllability.

Exercise 1.9 *Review the examples and exercises in the previous section, and relate the notion of controllability to the questions raised in that section.*

One may well say that the study of controllability is analogous, and just as fundamental as the questions of existence and uniqueness of solutions of differential equations. In further analogy, the study of controllability actually leads one to algorithmic constructions of more advanced problems such as path planning, much in the same way as proofs for existence and uniqueness

of solutions of differential equations yield e.g. recipes for obtaining infinite series and numerical solutions.

Recall the case of linear systems $\dot{x} = Ax + Bu$ (with state and control vectors x and u and matrices A and B of appropriate sizes). Using variation of parameters one quickly obtains a formula for the solution curve

$$x(t) = x(0)e^{tA} + \int_0^t e^{(t-s)A} Bu(s)\, ds \tag{7}$$

It is readily apparent that the set of points that can be reached from $x(0) = 0$ (via piecewise constant, measurable controls or any similar sufficiently rich class) is always a subspace of the state space. Moreover, scaling of the control $u \mapsto cu$ immediately carries over to the solution curve $x(t, cu) = cx(t, u)$ (assuming $x(0) = 0$). Consequently, the size of the control is no major factor in the discussion of linear controllability, and neither is the time T that is available. The scaling invariance implies that most local properties and global properties are the same. Finally, the solution formula (7) also quickly yields (e.g. via Taylor expansions and the Cayley-Hamilton theorem) a simple algebraic criterion for linear controllability:

Theorem 1.1 (Kalman rank condition) *The linear system $\dot{x} = Ax + Bu$ with $x \in \mathbb{R}^n$ is controllable (for* any *reasonable technical definition of controllable) iff the block-matrix $(B, AB, A^2B, \ldots A^(n-1)B)$ has full rank.*

In the case of nonlinear systems almost *everything* is different: There are many, many equally reasonable notions of controllability which are not equivalent to each other. Local and global notions are generally very different. The class of admissible controls has to be very carefully stated – e.g. bounds on the control size can make all the difference. Assumptions about regularity properties of both the vector fields (e.g. smooth versus analytic) and the controls (e.g. measurable versus piecewise constant) are critically important. Similarly, a system may be controllable (in a reasonable) sense given sufficiently much time, but may be uncontrollable for small positive times.

In these lectures we shall concentrate on one of the best studied notions, and which is of significant importance for a variety of further theories (e.g. a sufficient condition for some notions of feedback stabilizability, and duality to optimality). Thus from now on, unless otherwise stated the following blanket assumptions shall generally apply: We consider affine, analytic systems that are of the form

$$\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^{m} u_i(t)\, f_i(x(t)) \text{ usually initialized at } x(0) = 0 \tag{8}$$

where f_i are (real) analytic vector fields, and the controls u are assumed to be measurable with respect to time, and assumed to take values in a compact subset $U \subset \mathbb{R}^m$, often taken as $[-1,1]^m$. The vector field f_0 is called the drift vector field, while f_i for $i \geq 1$ are called the control (or controlled) vector fields. In the case that $f_0 \equiv 0$ (i.e. is absent) the system (8) is called *"without drift"*.

Much different techniques are needed when allowing more general dependence of the dynamics on the control $\dot{x} = f(x,u)$, compare the lectures by Jacubczyk in this series. One may also demand less regularity, e.g. only Lipschitz-continuity of the vector fields associated to fixed values of the controls. A mature theoretical framework for that case provided by *differential inclusions*, compare the lectures by Frankowska in this series.

Revisit the parking example of the first section and introduce standard, uniform notation by defining $x = (x_1, x_2, x_3, x_4, x_5) \stackrel{\text{def}}{=} (\phi, v, x, \theta, y)$, and write the system (3) in the form (8)

$$f_0(x) = \begin{pmatrix} 0 \\ 0 \\ x_2 \cos x_4 \\ x_2 \tan x_1 \\ x_2 \sin x_4 \end{pmatrix}, \qquad f_1(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad f_2(x) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{9}$$

This is a system with drift – which corresponds to the *unforced* dynamics of the car, and with two controlled vector fields which correspond to forward acceleration/deceleration, and to changing the steering angle.

Exercise 1.10 *Revisit the second example, the rolling penny, from the first section. Again write the states as $x = (x_1, x_2, \ldots)$ and write the systems* (5) *and* (6) *in the form* (8) *(i.e. identify the controlled vector field(s), and the drift vector field.*

Explain in practical terms how the choice of controls as acceleration or as velocities effects the presence of a drift. (Note, these are models for the kinetic versus dynamic behaviours).

Definition 1.1

The reachable set $\mathcal{R}_p(T)$ of system (8) *at time $T \geq 0$, subject to the initial condition $x(0) = p$ is the set*

$$\mathcal{R}_p(T) = \{x(T,u) \colon x(0) = p \ \ \textit{and} \ \ u \colon [0,T] \mapsto U \ \textit{measurable} \ \} \tag{10}$$

Definition 1.2 *The system* (8) *is accessible from* $x(0) = p$*, if the reachable sets* $\mathcal{R}_p(t)$ *have non-empty interior for all* $t > 0$.
The system (8) *is small-time locally controllable* (STLC) *about* $x(0) = p$*, if* $x(0) = p$ *is contained in the interior or the reachable sets* $\mathcal{R}_p(t)$ *for all* $t > 0$.
Consider

Accessibility and controllability (STLC) are generally different from each other.

$$\left\{ \begin{array}{rclcrcl} \dot{x}_1 & = & u & & x(0) & = & 0 \\ \dot{x}_2 & = & x_1^k & & \|u(\cdot)| & \leq & 1 \end{array} \right. \tag{11}$$

with measurable controls $u(\cdot)$ bounded by $\|u(\cdot)| \leq 1$ and $k \in \mathbf{Z}^+$ fixed. Using e.g. piecewise constant controls with a single switching one easily that the reachable sets $\mathcal{R}_0(t)$ have two dimensional interior for all $t > 0$ while $x(0) \notin \mathcal{R}_0(t)$ for all $t \geq 0$ if k is even.

Exercise 1.11 *Using methods from optimal control, one may show that the boundaries of the reachable sets at time* $T \geq 0$ *of the system* (11) *are contained in the set of endpoints of trajectories resulting from* bang-bang controls *with at most one switching, i.e. controls of the form* $u_{+-,t_1}(t) = 1$ *if* $0 \leq t \leq t_1$ *and* $u_{+-,t_1}(t) = -1$ *if* $t < t_1 \leq T$*, or* $u_{-+,t_1}(t) = 1$ *if* $0 \leq t \leq t_1$ *and* $u_{-+,t_1}(t) = -1$ *if* $t < t_1 \leq T$*. Calculate these curves of endpoints (as curves parameterized by* t_1*). Rewrite these as (unions of) graphs of functions* $x_2 = f(x_1)$*, and sketch the reachable sets.*

Exercise 1.12 *Continuing the previous exercise in the case of* k *an even integer, identify all pairs of switching times* t_1*,* t_2 *such that* $x(1; u_{+-,t_1}(t) = x(1; u_{-+,t_2})(t)$.

A few further remarks about controllability: Clearly, controllability is a geometric notion, independent of any choice of local coordinates. While for calculations it often is convenient to choose and fix a set of specific coordinates, it is desirable to obtain conditions for controllability that are geometric, too (compare the Kalman rank condition which involves the *geometric* property of the rank). In these notes we are concerned only with local properties. Consequently, we generally may assume that the underlying manifold is $\mathbb{R}^n$. In particular, when working with approximating vectors we shall conveniently identify the tangent spaces $T_p\mathbb{R}^n$ to $\mathbb{R}^n$with $\mathbb{R}^n$. Nonetheless, occasionally we may phrase our observations and results so as to emphasize that they really also apply to general manifolds.

In the linear setting there is a very distinctive duality between controllability and observability. In the nonlinear case this moves more to the background. However, STLC is *dual* to optimality: Controllability means that one can reach a neighborhood, whereas optimality means that a trajectory lies on the boundary of the *funnel* of all reachable sets. Consequently, necessary conditions for STLC translate immediately into necessary conditions for optimality, and vice versa.
Also the implication that controllable systems are feedback stabilizable carries over to the nonlinear framework when using the notion of STLC and *time-periodic static feedback* as shown by Coron [7]. Note, that a simple reversal of time, i.e. replacing each f_i by $(-f_i)$ makes the obvious conceptual transition from *controllable from* an equilibrium p, to *stabilizable to* an equilibrium p.

1.3 Piecewise constant controls and the CBH formula

A natural first approach to studying controllability is to start with the analysis of trajectories corresponding to piecewise constant controls. As illustrated in the explorations in the parallel parking example in the first section, it is the lack of commutativity that is the key to obtaining *new directions* by conjugation of flows corresponding to different (constant) control values. This section further explores piecewise constant controls and their connection to Lie brackets, which measure the lack of commutativity.
Consider a collection of switching times $0 = t_0 \leq t_1 < t_2 < \ldots < t_{s-1} < t_s = T$ and fixed control values $c_1, c_2, \ldots c_s \in U$, and define the control $u = u_{t_1,t_2,\ldots,t_s;c_1,c_2,\ldots c_s} \colon [0,T] \mapsto U$ by $u(t) = c_i$ if $t_{i-1} < t \leq t_i$ (and e.g. $u(0) = 0$). As a piecewise constant control, u is measurable and thus admissible. The endpoint $x(T,u)$ of the trajectory starting at $x(0)$ is obtained by concatenating the solutions of s differential equations $\dot{x} = f_0(x) + \sum_{j=1}^m c_{ij} f_j(x)$. In other words, $x(T,u)$ is obtained from $x(0)$ by composing the flows for times $(t_i - t_{i-1})$ of the vector fields $F_i = f_0 + \sum_{j=1}^m c_{ij} f_j$, $i = 1 \ldots s$ and evaluating the composition at $x(0)$. It is customary to write this as a *product of exponentials*

$$x(T,u) = e^{(t_s - t_s - 1)F_s} \cdots e^{(t_3 - t_2)F_1} e^{(t_2 - t_1)F_1} e^{t_1 F_1} x(0) \tag{12}$$

Here the exponential is just a convenient shorthand notation for the flow $(t,p) \mapsto e^{tX}p$ for the flow of the vector field X, i.e. defined by $e^{0X}p = p$ and $\frac{d}{dt} e^{tX} p$ equals the value of the vector field X at the point pe^{tX} for every t (in the domain of the flow).

A word of caution:

While practices vary around the world, and change with time, it is customary in geometric control theory to adopt the convention of writing xf for the value of a function f at a point x (replacing the traditional $f(x)$. In particular, one writes $\frac{d}{dt}pe^{tX} = pe^{tX}X$ for the value of the vector field X at the point pe^{tX}.
In more generality, in an expression $pe^X Ye^Z\phi$ it is understood that p is a point (on a manifold M), e^X the flow of the vector field X at time 1, Y is a vector field on M, e^Z is the tangent map of the flow of the vector field Z at time 1, and ϕ is a function on M. Particularly nice is that there is no need for parentheses, or a need to write additional *"stars"* for the tangent maps (see below). E.g. $pe^X Y$ is a tangent vector at the point pe^X, while e.g. $Y\phi$ is a function on M, $p\phi$ and $pY\phi$ are numbers.
It is important to remember at all times that these exponentials denote flows, and thus they are manipulated exactly as flows are manipulated. In particular, in general $e^{tX}e^{sY} \neq e^{sY}e^{tX}$. However, with careful attention to the legal rules of operation, this proves to be very effective notation for many calculations. For some impressive examples of substantial calculations see [35]. For an extension of this symbolism to *time varying vector fields* see Agrachev [1, 2]. We note on the side, that in the differentiation rules $\frac{d}{ds}pe^{tX}e^{sY} = pe^{tX}e^{sY}Y$ and $\frac{d}{dt}pe^{tX}e^{sY} = pe^{tX}Xe^{sY}$ exponentials to the right of a tangent vector (like pXe^{tY}) stand for the tangent maps of the flows, which in classical differential geometry is often denoted by a lower star: If $\Phi\colon M \mapsto N$ is a map between differentiable manifolds, and $p \in M$ then $\Phi_{*p}\colon T_pM \mapsto T_{\Phi(p)}N$ and $\Phi_*\colon TM \mapsto TN$ denote the tangent map in classical notation. In our case, the position of the exponentials will always make it clear which map it stands for, i.e. there is no need to write stars.

In these introductory notes we shall **not** follow this convention. There are just too many examples and calculations from areas other than geometric control where a consistent application of these rules would look very awkward (just think of $x\sqrt{\ }$ sin). However, we note that the reversal of the order in which certain expressions are to be interpreted will cause the (dis)appearance of sign correction factors $(-)^k$ in many places, i.e. one has to be very careful when combining formulas from different sources.

One major advantage of the exponential notation is that it not only matches the symbols used in the study of Lie groups and the symbolism used in formal power series, but that the properties and rules for manipulating them are often identical, making it very easy to mentally move back and fourth.

Rather than directly constructing a control that steers to any given point in a neighborhood of $x(0)$, the first simplification results from using the implicit or inverse function theorem. The basic idea is to construct a comparatively simple control, parameterized e.g. by a finite number of switching times (and/or control values) that *returns* the state to the starting point. If these data are interior (i.e. not extreme values of U), one can to conclude STLC if the Jacobian matrix of this endpoint map has full rank. This basic construction is applicable to much more general settings, compare e.g. the discussion of controllability of partial differential equations in the lectures by Coron. (However, for daily computations in finite dimensional systems we now know simpler tests that will be discussed in the sequel).

Example 1.3 (Stefani [59], 1985)

$$\left\{\begin{array}{lcll} \dot{x}_1 &=& u & x(0) = 0 \\ \dot{x}_2 &=& x_1 & \|u(\cdot)| \leq 1 \\ \dot{x}_3 &=& x_1^3 x_2 & \end{array}\right. \tag{13}$$

Consider the piecewise constant controls that take values $+1, -1, +1, -1, 0$ on the intervals $[0, a]$, $(a, a+b]$, $(a+b, a+b+c]$, $(a+b+c, a+b+c+d]$, and $(a+b+c+d, T]$, and calculate the endpoint (using a computer algebra system)

$$\begin{aligned}
x_1(T, u_{+-+-;a,b,c,d}) &= a + c - b - d, \\
x_2(T, u_{+-+-;a,b,c,d}) &= \tfrac{1}{2}a^2 + ab - \tfrac{1}{2}b^2 - bc + ac + \tfrac{1}{2}c^2 + ad + cd - bd - \tfrac{1}{2}d^2, \\
x_3(T, u_{+-+-;a,b,c,d}) &= -2bcd^4 + \tfrac{1}{12}a^6 + \tfrac{1}{2}a^5 b + 2ab^3 cd - 4a^3 bcd - 9a^2 bc^2 d \\
&\quad -3a^2 bcd^2 - 8abc^3 d + 6ab^2 c^2 d + 6ab^2 cd^2 + 6abcd^3 - \tfrac{1}{4}a^4 b^2 - \tfrac{2}{3}a^3 b^3 \\
&\quad +a^2 b^4 - \tfrac{1}{2}ab^5 + b^2 c^4 + \tfrac{1}{2}b^5 c + \tfrac{1}{12}c^6 + \tfrac{1}{12}d^6 + 4a^2 b^3 c - 2a^3 b^2 c - \tfrac{5}{2}ab^4 c \\
&\quad +ab^3 c^2 + 2ab^2 c^3 - 2bac^4 - \tfrac{1}{2}ba^4 c - 2ba^3 c^2 - 3ba^2 c^3 + \tfrac{1}{12}b^6 + 2acd^4 \\
&\quad -2ac^2 d^3 + \tfrac{5}{2}ac^4 d - ac^3 d^2 - \tfrac{5}{2}abd^4 - 5ab^2 d^3 - \tfrac{5}{2}ab^4 d - 5ab^3 d^2 \\
&\quad +2bc^2 d^3 - \tfrac{5}{2}bc^4 d + bc^3 d^2 - \tfrac{1}{4}b^4 c^2 - \tfrac{2}{3}b^3 c^3 - \tfrac{1}{2}bc^5 + \tfrac{5}{4}a^2 c^4 + \tfrac{1}{2}a^5 c \\
&\quad +\tfrac{5}{4}a^4 c^2 + \tfrac{5}{3}a^3 c^3 + \tfrac{1}{2}ac^5 - \tfrac{1}{2}ad^5 + a^2 d^4 + \tfrac{1}{2}a^5 d - \tfrac{1}{4}a^4 d^2 - \tfrac{2}{3}a^3 d^3 \\
&\quad +\tfrac{1}{2}bd^5 + \tfrac{5}{4}b^2 d^4 + \tfrac{5}{3}b^3 d^3 + \tfrac{1}{2}b^5 d + \tfrac{5}{4}b^4 d^2 - \tfrac{1}{2}a^4 bd + \tfrac{5}{2}a^4 cd - 2a^3 b^2 d \\
&\quad -2a^3 bd^2 + 5a^3 c^2 d - a^3 cd^2 + 4a^2 bd^3 + 4a^2 b^3 d + 6a^2 b^2 d^2 - 2a^2 cd^3 \\
&\quad +5a^2 c^3 d - 3/2a^2 c^2 d^2 + 4b^2 c^3 d - \tfrac{1}{2}b^4 cd - 2b^3 c^2 d - 2b^3 cd^2 - 3b^2 cd^3 \\
&\quad -\tfrac{1}{2}cd^5 + c^2 d^4 - \tfrac{2}{3}c^3 d^3 + \tfrac{1}{2}c^5 d - \tfrac{1}{4}c^4 d^2.
\end{aligned} \tag{14}$$

It is easy to check that $x(10, u_{+-+-;1,1+\sqrt{2},1+\sqrt{2},1}) = (0,0,0)$, and that

$$\text{rank } \left.\frac{\partial x(10, u_{+-+-;a,b,c,d})}{\partial(a,b,c,d)}\right|_{(1,1+\sqrt{2},1+\sqrt{2},1)} = 3 \tag{15}$$

Thus, by the implicit function theorem, there exists some open neighborhood W of $x(0) = x(10, u_{+-+-;(1,1+\sqrt{2},1+\sqrt{2},1)} = 0$ such that for every $p \in W$, there exists some values (a,b,c,d) near $(1, 1+\sqrt{2}, 1+\sqrt{2}, 1)$ such that $x(10, u_{+-+-;a,b,c,d)} = p$. Thus the system is locally controllable about 0, and via some simple arguments using homogeneity (see the next sections), also STLC about 0.

Challenge exercise 1.13 *(use CAS!). Consider the slightly modified system*

$$\left\{\begin{array}{lcllcl} \dot{x}_1 &=& u & x(0) &=& 0 \\ \dot{x}_2 &=& x_1^3 & \|u(\cdot)| &\leq& 1 \\ \dot{x}_3 &=& x_1 x_2 & & & \end{array}\right. \tag{16}$$

Find a piecewise constant, bang-bang control $u\colon [0,T] \mapsto \{-1,+1\}$ *(for some* $T > 0$*) such that corresponding trajectory of* (16) *returns to* 0*, and such that the Jacobian matrix of partial derivatives of the endpoint* $x(T,u)$ *with respect to the switching times has rank* 3 *at your choice of switching times. (Use a computer algebra system.)*

Challenge exercise 1.14 *(use CAS!). Repeat the previous exercise, but now with the values of the piecewise constant control considered as variables, while the switching times are considered fixed. I.e. find a piecewise constant control* $u\colon [0,T] \mapsto (-1,1)$ *(for some* $T > 0$*)* $u_i(t) = c_i$ *if* $t_{i-1} \leq t \leq t_i$ *such that corresponding trajectory of* (16) *returns to* 0*, and such that the Jacobian matrix of partial derivatives of the endpoint* $x(T,u)$ *with respect to the values* c_i *of the control has rank* 3 *at your choice of control values. Use a computer algebra system.*

In terms of compositions of flows or products of exponentials the previous example employed

$$x(10,u) = e^{(10-a-b-c-d)f_0} e^{d(f_0-f_1)} e^{c(f_0+f_1)} e^{b(f_0-f_1)} e^{a(f_0+f_1)}(0) \tag{17}$$

and found that in particular

$$x(10,u_*) = e^{(8-2\sqrt{2})f_0} e^{(f_0-f_1)} e^{(1+\sqrt{2})(f_0+f_1)} e^{(1+\sqrt{2})(f_0-f_1)} e^{(f_0+f_1)}(0) = 0 \tag{18}$$

Differentiation of (17) with respect to the times a, b, c, d then was used to establish controllability. (Compare exercise 4.12 for manual symbolic manipulations.) For many systems this approach is impractical as e.g. exact switching times which return the system to the starting point may be difficult or practically impossibleto find. Thus it is natural to look for alternative methods. In particular, the key is to study the lack of commutativity.

Recall that the Lie bracket $[F_1, F_2]$ of two smooth vector fields F_1 and F_2 on a manifold M is algebraically defined as the vector field $[F_1, F_2]: C^\infty(M) \mapsto C^\infty(M)$ via $[F_1, F_2]\phi = F_1(F_2\phi) - F_1(F_2\phi)$.
In coordinates, with vector fields written as column vectors, and denoting the Jacobian matrix by D, one calculates the Lie bracket as $[F_1, F_2] = (DF_2)F_1 - (DF_1)F_2$.

Example 1.4

Consider $f_0(x) = x_1 \frac{\partial}{\partial x_2}$ and $f_1(x) = \frac{\partial}{\partial x_1}$. Then

$$[f_0, f_1](x) = \left(x_1 \frac{\partial}{\partial x_2}\right) \circ \left(\frac{\partial}{\partial x_1}\right) - \left(\frac{\partial}{\partial x_1}\right) \circ \left(x_1 \frac{\partial}{\partial x_2}\right)$$

$$= x_1 \left(\frac{\partial^2}{\partial x_1 \partial x_2} - \frac{\partial^2}{\partial x_2 \partial x_1}\right) - \frac{\partial x_1}{\partial x_1}\frac{\partial}{\partial x_2} = -\frac{\partial}{\partial x_2} \qquad (19)$$

In matrix / column vector notation the same calculation reads

$$[\begin{pmatrix} 0 \\ x_1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}] = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ x_1 \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}. \qquad (20)$$

Exercise 1.15 *For the vector fields $f_0(x) = x_1^4 \frac{\partial}{\partial x_2}$ and $f_1(x) = \frac{\partial}{\partial x_1}$ calculate the iterated Lie brackets $[f_0, f_1]$, $[f_0, [f_0, f_1]]$, and $[[f_0, f_1], f_1]$.*
Find an iterated Lie bracket f_π (of higher order) such that $f_\pi(0) = \frac{\partial}{\partial x_2}$.

Geometrically,the Lie bracket is defined by the limit (for $\phi \in C^\infty(M)$)

$$[f_1, f_2]\phi(p) = \lim_{t \longrightarrow 0} \frac{1}{t^2} \left(\phi(e^{-tf_2} e^{-tf_1} e^{tf_2} e^{tf_1} p) - \phi(p)\right) \qquad (21)$$

i.e. as the infinitesimal measure of the lack of commutativity of the flows of f_1 and f_2 at the point p. It is very instructive to calculate these flows in a simple explicit example, and see how the limit gives rise to a new direction.

Back to example 1.4. Staring at $p = (p_1, p_2)$, calculate

$$\begin{array}{rcl} p' = e^{tf_0}(p) & = & (p_1, p_2 + tp_1) \\ p'' = e^{tf_1}(p') & = & (p_1 + t, p_2 + tp_1) \\ p''' = e^{-tf_0}(p'') & = & (p_1 + t, p_2 - t^2) \\ p'''' = e^{-tf_1}(p''') & = & (p_1, p_2 - t^2) \end{array} \tag{22}$$

and thus

$$[f_0, f_1]\phi(p) = \lim_{t\to 0} \frac{1}{t^2}(\phi(p'''') - \phi(p)) \lim_{t\to 0} \frac{1}{t^2}(\phi(p_1, p_2 - t^2) - \phi(p_1, p_2)) = -\frac{\partial \phi}{\partial x_2}(p) \tag{23}$$

which is in agreement with the earlier algebraic calculation that yielded $[f_0, f_1] = -\frac{\partial}{\partial x_2}$.

Exercise 1.16 *Check the calculations in* (22). (Note, a common source of confusion is the very questionable use of the same symbol t for both the length of each interval and also as integration variable along each piece.). *Sketch the curves, and repeat all with the order of f_1 and f_2 reversed. Work out the special case of $p = 0$ – this is a useful case to remember as it helps to recover the sign-convention used in any locality.*

One of the major goals of these lectures is to develop methods and tools that allow one to more easily work with the compositions of noncommuting flows. One of the oldest such tools is the classical Campbell Baker Hausdorff formula which asserts that

$$e^X \cdot e^Y = e^{\log(e^X \cdot e^Y)} = e^{X+Y+\frac{1}{2}[X,Y]+\frac{1}{12}[X,[X,Y]]-\frac{1}{12}[Y,[X,Y]+\ldots}} \tag{24}$$

One of the nice features of this formula is that it is just as correct in the sense of formal power series in *noncommuting indeterminates* X and Y, as it is correct for analytic vector fields X and Y (as long as all flows are defined). This is no accident! It is easy to informally verify this identity by simply using the standard Taylor expansion for exponentials, formally expanding both sides and recursively using the definition $[V, W] = VW - WV$. A rigorous justification that one can indeed go easily back and fourth between geometric/analytic and algebraic/combinatorial interpretations can be made in many ways – indeed, manipulations of analytic objects are by nature often purely algebraic. Arguably one of the more elegant ones starts with the classical identification of points on a manifold with multiplicative functionals on the algebra of smooth functions on a manifold, and then proceeds with identifying flows with formal partial differential operators of infinite order, compare e.g. [19, 35].

Exercise 1.17 *Repeatedly use the CBH-formula to write the end point of 4 flows corresponding to* bang-bang *controls as a single* exponential*:*

$$\begin{aligned} x(T, u_{t_1,t_2,t_3}) &= 0 \cdot e^{t_1(f_0+f_1)} \cdot e^{(t_2-t_1)(f_0-f_1)} \cdot e^{(t_3-t_2)(f_0+f_1)} \cdot e^{(T-t_3)(f_0-f_1)} \\ &\stackrel{!}{=} e^{p_0(t)f_0+p_1(t)f_1+p_{01}(t)[f_0,f_1]+p_{011}(t)[f_0,[f_0,f_1]]+p_{101}(t)[f_1,[f_0,f_1]]+\cdots}(0) \end{aligned} \tag{25}$$

Find explicit formulas for polynomial expressions $p_I(t) = p_I(t_1, t_2, t_3, t_4)$ *(in the switching times) for* $I = 0, 1, 01, 011, 110$.

The following lectures aim at obtaining similar formulas that are easier to use, and that also allow for controls that are not necessarily piecewise constant! The starting point will be the Chen Fliess series expansion.

1.4 Approximating cones and conditions for STLC

Instead of constructing controls that steer exactly to a specific point, which generally is very hard, *analysis* is about building arguments that use approximate directional information obtained from derivatives. This discussion also establishes a close link between STLC and optimal control.
The key idea is to develop a tangent, or derivative object for the reachable sets that is easy to constuct/compute, that has reasonably nice convexity properties, and that nicely approximates the reachable sets. While prior efforts in control largely focused on constructing specific kinds of control variations and then created arguments why these control variations can be combined especially, Frankowska [16, 17] pioneered a different approach that provides a very general open mapping principle under very general hypotheses, which then may be applied to many special cases after simply checking that the specific tangent vectors satisfy some mild conditions. Its general theory applies in Banach space settings, and requires only minimal smoothness (thus the name *nonsmooth analysis*. In these notes we generally follow the original definitions and constructions of [16, 17]. The open mapping theorem stated below is a very special case of the general results in [16, 17], and we use its simple language as it is meets exactly the needs of our systems.
The following is one of the most simple possible notions of tangent vectors, yet in the special case of affine smooth systems these vectors are automatically in the *contingent cone* and thus Frankowska's general open mapping principle [17] applies.

Definition 1.3 ([17, 29]) *Consider systems of form* (8) *on* $\mathbb{R}^n$ *with* $f_0(0) = 0$ *and* $0 \in \text{int}U$. *A vector* $\xi \in R^n$ *is called a k-th order tangent vector to*

the family $\{\mathcal{R}_t(0)\}_{t\geq 0}$ *at* 0 *if there exists a parameterized family of control variations* $u_s\colon [0,s] \mapsto U$, $s \geq 0$, *such that*

$$x(s, u_s) = 0 + s^k \xi + o(s^k). \tag{26}$$

The set of all k*-th order tangent vectors* (to $\{\mathcal{R}_t(0)\}_{t\geq 0}$ at zero) *is denoted by* C^k, *while* $\overline{C^k} = \bigcup_{\lambda>0} \lambda C^k$ *is the set of* tangent rays *to* $\{\mathcal{R}_t(0)\}_{t\geq 0}$ *at zero.*

The parameterization $s \mapsto u_s$ is not required to be smooth. Indeed, it suffices to require sequences $s_k \searrow 0$.

Exercise 1.18 *Find 6 families of control variations* $u_s^{\pm i}\colon [0,s] \mapsto [-1,1]$ *that generate the tangent vectors* $\pm \frac{\partial}{\partial x_i}\Big|_0$, $i = 1,2,3$, *for the system* (13).

Exercise 1.19
Repeat the previous exercise using the control sizes, as in exercise 1.14, *as parameter* s.

The following properties are easy to establish:

Proposition 1.2 ([16, 17, 29])
(a) *If* $\lambda^k \in [0,1]$, *then* $\lambda^k C^k \subseteq C^k$.
(b) *If* $k \leq \ell$ *then* $C^k \subseteq C^\ell$.
(c) *If* $v_1, v_2 \in C^k$ *and* $\lambda^k \in [0,1]$ *then* $\lambda^k v_1 + (1-\lambda)^k v_2 \in C^k$.

Thus the sets C^k form an increasing sequence of truncated convex cones. The *approximation* property of these cones is established by:

Theorem 1.3 ([16, 17, 18, 29])
If $\overline{C^l}$ *is a closed convex cone (with vertex* $0 \in \mathbf{R}^n$*) such that* $\overline{C^l} \setminus \{0\} \subseteq \mathrm{int}\overline{C^k}$ *for some* $k < \infty$, *then there exist* $c > 0$, $T > 0$ *such that* $\overline{C^l} \cap B(0, ct^k) \subseteq \mathcal{A}(t)$ *for all* $0 \leq t \leq T$.

In [29] an explicit constructive proof is given for the theorem in the special case for systems of form (8), which has subsequently used as a starting point for feedback stabilization. But analogous results hold in much more generality, see. Frankowska [17, 18] for infinite dimensional versions only requiring minimal regularity.

Corollary 1.4 *If* $\overline{C^k} = \mathbf{R}^n$ *then there are constants* $c > 0$, $T > 0$ *such that* $B(0, ct^k) \subseteq \mathcal{A}(t)$ *for all* $0 \leq t \leq T$.

These approximating cones are just as useful for obtaining high-order versions of the maximum principle of optimal control as basically the dichotomy is the same: Does the reference trajectory (here $x \equiv 0$ lie in the interior or on the boundary of the reachable sets?

The preceding discussions, examples, and exercises suggest that there should be universal families of control variations that generate specific Lie brackets as tangent vectors to the reachable sets. This is indeed the case – and much research in the 1980 focused on developing the following conditions, which really emanate from arguments why certain families of control variations generate some Lie brackets. First introduce the following notation:

For smooth vector fields f and g define recursively define $(ad^0 f, g) = g$ and $(\mathrm{ad}^{k+1} f, g) = [f, (\mathrm{ad}^k(f, g)]$. For smooth vector fields $f_0, f_1, \dots f_m$ let $L(f_0, f_1, \dots f_m)$ denote the Lie algebra spanned by all iterated brackets of the vector fields f_i.
For any multi-index $r = (r_0, r_1, \dots r_m) \in \mathbf{Z}^{+(m+1)}$ let $L^r(f_0, f_1, \dots f_m)$ be the subspace spanned by all iterated brackets with r_i factors f_i, $i = 0, 1, \dots m$.
Also write $\mathcal{S}^k(f_0, f_1, \dots f_m)$ for the subspace spanned by all iterated brackets exactly k factors from $f_1, , \dots f_m$ and any numbers of factors f_0. For a set S of vector fields and a point p, we write $S(p)$ for the set $\{v(p) \colon v \in S\}$.
Note that it is possible to have e.g. $0 \neq [f_1, [f_1, f_0]] \in L^{(0,1)}(f_0, f_1)$ as $[f_1, [f_1, f_0]] = f_1$ is possible, e.g. if $f_1 = \frac{\partial}{\partial x_1}$ and $f_2 = \frac{1}{2} x_1^2 \frac{\partial}{\partial x_1}$. See the next chapter for more elaborate language that carefully distinguishes *binary labeled trees* (or *formal brackets* from Lie brackets. Since all our considerations are local, we identify the tangent space $T_0 \mathbb{R}^n$ with $\mathbb{R}^n$.

Recall that for nonlinear systems in general accessibility and controllability are not equivalent. For analytic systems, accessibility is comparatively easy to decide.

Theorem 1.5
The system (8) *initialized at* $x(0) = 0$ *is accessible if and only if*
$\dim L(f_0, f_1, \dots f_m)(0) = n$.

If the system is reversible, e.g. if U is symmetric about 0 and the system has no drift ($f_0 \equiv 0$), then accessibility implies STLC. (Consequently, the controlled kinematics are comparatively easy to deal with as opposed to the full dynamic model. Compare (5) versus (6).)
The closest analogue of the Kalman rank condition (theorem 1.1) for linear systems is the following condition, which basically says that if the Taylor

linearization is linearly controllable, then the original system is controllable (STLC) in the sense of nonlinear systems.

Theorem 1.6 (Linear Test) *If* $\mathcal{S}^1(0) = \mathbf{R}^n$ *then the system* (8) *is STLC.*

A complementary necessary condition for single-input system is a special case of the Clebsch-Legendre condition of optimal control:

Theorem 1.7
If $m = 1$ *and the system* (8) *is STLC then* $[f_1, [f_0, f_1]](0) \in \mathcal{S}^1(f_0, f_1)(0)$.

The exercises in the preceding section and above, aimed at generating tangent vectors via families of control variations should have suggested that for certain brackets their negatives are generated by the negatives of the controls. On the other hand system (11) shows that at least some *even* powers may be *obstructions* to STLC. The correctness of this intuition is formally established in:

Theorem 1.8 (Hermes [22], Sussmann [62]) *If* $m = 1$ *and* (8) *is accessible, and* $\mathcal{S}^{2k}(f_0, f_1)(0) \subseteq \mathcal{S}^{2k-1}(f_0, f_1)(0)$ *for all* $k \in \mathbf{Z}^+$ *then* (8) *is STLC.*

A complementary necessary condition is:

Theorem 1.9 (Stefani [60]) *If* $m = 1$ *and the system* (8) *is STLC then* $(ad^{2k} f_0, f_1)(0) \in \mathcal{S}^{2k-1}(f_0, f_1)(0)$ *for all* $k \in \mathbf{Z}^+$.

Finally, the most general sufficient condition known today allows one to weight the drift and the controlled fields differently when counting the order of a bracket.

Theorem 1.10 (Sussmann [65]) *If the system the system* (8) *is accessible and there exists a weight* $\theta \in (0, 1]$ *such that for all odd* k *and even* $\ell_1, \ldots \ell_m$

$$L^{(k,\ell_1,\ldots\ell_m)}(f_0, \ldots f_m)(0) \subseteq \sum_{(k^i,\ell^i)} L^{(k^i,\ell_1^i,\ldots\ell_m^i)}(f_0, \ldots f_m)(0) \tag{27}$$

where the sum extends over all (k^i, ℓ^i) *such that*

$$\theta k^i + \ell_1^i + \ldots + \ell_m^i < \theta k + \ell_1 + \ldots + \ell_m \tag{28}$$

then the system (8) *is STLC.*

Loosely phrased, this theorem singles out brackets of type (odd,even,... even) as *potential obstructions* to STLC, and it describes a way how these may be *neutralized* so that the system is STLC. The commonly used term *bad brackets*[??] for the potential obstructions is unfortunate since the obstructions should be identified with *supporting hyperplanes* of the the approximating cones, and thus are elements of the *dual space* of a free Lie algebra, compare [39]. (It is also possible to also use different weights for the controlled fields. The best notation uses weight 1 for f_0 and weights $\sigma_i = \frac{1}{\theta_i} \in [1, \infty)$ for the fields f_i.) Several small extensions, and specific examples that explore the region between the necessary and sufficient conditions can be found in the literature, see e.g. [32] for an overview.

Example (1.3) revisited. Extract the vector fields f_0 andf_1 from (13) and calculate iterated Lie brackets.

$$f_0(x)=\begin{pmatrix}0\\ x_1\\ x_1^3x_2\end{pmatrix},\ f_1(x)=\begin{pmatrix}1\\ 0\\ 0\end{pmatrix},\ [f_0,f_1](x)=\begin{pmatrix}0\\ -1\\ 3x_1^2x_2\end{pmatrix},\ [f_0,[f_0,f_1]](x)=\begin{pmatrix}0\\ 0\\ 6x_1^3\end{pmatrix}$$

$$[f_1,[f_1,f_0]](x)=\begin{pmatrix}0\\ 0\\ -6x_1x_2\end{pmatrix},\ (\mathrm{ad}^3f_1,f_0)(x)=\begin{pmatrix}0\\ 0\\ -6x_2\end{pmatrix},\ (\mathrm{ad}^4f_1,f_0)(x)=0 \quad (29)$$

$$[f_0,[f_1,[f_1,f_0]]](x)=[f_1,[f_0,[f_1,f_0]]](x)=\begin{pmatrix}0\\ 0\\ 12x_1^3\end{pmatrix},\ [f_1,f_0],(\mathrm{ad}^3f_1,f_0)](x)=\begin{pmatrix}0\\ 0\\ -6\end{pmatrix}$$

Note that $(\mathrm{ad}^k f_0, f_1)(x) = 0$ when $k > 2$. Since $\dim L(f_0, f_1)(0) = 3$ the system is accessible, but is *not linearly controllable* due to $\dim \mathcal{S}^1(f_0, f_1)(0) = 2 < 3$. Since $[f_1, [f_1, f_0]](0) = 0 \in \mathcal{S}^1(f_0, f_1)(0)$ and similarly $(\mathrm{ad}^4 f_1, f_0)(0) = 0 \in \mathcal{S}^3(f_0, f_1)(0)$, Stefani's necessary conditions are satisfied. The only brackets which give the $\frac{\partial}{\partial x_3}$ direction have 4 factors of f_1, and thus the Hermes' condition does not apply. However, since all brackets with an even number of factors f_1 and an odd number of factors f_0 vanish at 0, Sussmann's condition affirms STLC – something which we proved earlier by a brute-force construction.

Historical note: This example by Stefani was first shown to be STLC using the method we exhibited in the previous section. It clearly shows that the Hermes condition is far from necessary, and it served as a substantial motivation for the eventual sharpened version of Sussmann's general theorem that was proven in [65].

With a computer algebra system such calculations are very easy and quickly executed – but the big question is: Which brackets does one have to calculate? For very short brackets it is quite obvious that e.g. only one of $\{[f_0,[f_0,f_1]],\ [f_0,[f_1,f_0]],\ [f_1,[f_0,f_0]],\ [[f_0,f_0],f_1],\ [[f_0,f_1],f_0],\ [[f_1,f_0],f_0]\}$ needs to be computed (due to anticommutativity $[X,Y]=-[Y,X]$ and the Jacobi identity $[X,[Y,Z]]+[Z,[X,Y]]+[Y,[Z,X]]=0$ for all X,Y,Z in a Lie algebra). But as the length increases, the number of a-priori possible brackets sky-rockets, yet it is apparent that there will be lots of duplication. The subsequent lectures on combinatorics and algebra will provide nice answers by providing bases that are very easily constructed. The question "when can one stop?" is also answered in the next lecture (for nilpotent systems).

Exercise 1.20 *Determine whether the car model* (9) *from example* (1.1) *is STLC.*

Exercise 1.21 *Determine whether the models* (5) *and* (6) *for the kinematics and the dynamics of the rolling penny example* (1.2) *are STLC.*

2 Series expansion, nilpotent approximating systems

2.1 Introduction to the Chen Fliess series

Much classical work investigated the whether the sets of points reachable by piecewise constant controls agree with those reachable by means of arbitrary measurable controls, see e.g. [15] Grasse (late 1980s). But one may expect that in general one may need very large numbers of *pieces* in order to well approximate measurable controls. The subsequent very large number of repeated applications of the CBH-formula is even less attractive. Thus one is led to look for expansions that do not rely on piecewise constancy, and which allow one to combine a large number of *pieces* in one step.
One of the most basic formulas is obtained by simple Picard iteration. For an analytic *output function* $\phi\colon \mathbb{R}^n \mapsto \mathbb{R}$ (especially, for $\phi = x_i$ a coordinate function) first rewrite the system of differential equations with initial condition

$$\frac{d}{dt}\phi(x(t)) = \sum_{i=0}^{n} u_i(t)(f_i\phi)(x(t)), \quad \phi(x(0)) = \phi(p) \tag{30}$$

as an equivalent integral equation, and then iterate the analogous rewriting for the

subsequently appearing Lie derivatives $(f_{i_s} \dots f_{i_2} f_{i_1} \phi)$ of ϕ

$$\begin{aligned}
\phi(x(t)) &= \phi(p) + \int_0^t \sum_{i_1=0}^m u_{i_1}(t_1)(f_{i_1}\phi)(x(t_1))\, dt_1 \\
&= \phi(p) + \int_0^t \sum_{i_1=0}^m u_{i_1}(t_1) \Big((f_{i_1}\phi)(p) + \int_0^{t_1} \sum_{i_2=0}^m u_{i_2}(t_2)(f_{i_2} f_{i_1}\phi)(x(t_2))\, dt_2 \Big)\, dt_1 \\
&= \phi(p) + \int_0^t \sum_{i_1=0}^m u_{i_1}(t_1) \Big((f_{i_1}\phi)(p) + \int_0^{t_1} \sum_{i_2=0}^m u_{i_2}(t_2) \Big((f_{i_2} f_{i_1}\phi)(p) \\
&\qquad + \int_0^{t_2} \sum_{i_3=0}^m u_{i_3}(t_3)(f_{i_3} f_{i_2} f_{i_1}\phi)(x(t_3))\, dt_3 \Big)\, dt_2 \Big)\, dt_1
\end{aligned} \tag{31}$$

and so on. Note that each of these is an *exact* equation, where the last term to be considered an *error term* in a finite series approximation. The usefulness of this expansion is that it separates the time- and control dependence of the solution from the nonvarying geometry of the system which is captured by the vector fields f_i (or in the iterated Lie derivatives $(f_{i_s} \dots f_{i_2} f_{i_1} \phi)$ which may be computed off-line, and only once). For compatibility with later formulas, we reverse the names of the integration variables and indices used, e.g. rename i_1 to become i_3 and vice versa, and expand the sums

$$\begin{aligned}
\phi(x(t)) &= \phi(p) + \sum_{i_1=0}^m \left(\int_0^t u_{i_1}(t_1) dt_1 \right) \cdot (f_{i_1}\phi)(p) \\
&+ \sum_{i_2=0}^m \sum_{i_1=0}^m \left(\int_0^t \int_0^{t_1} u_{i_2}(t_2) u_{i_1}(t_1) dt_1 dt_2 \right) (f_{i_1} f_{i_2}\phi)(p) \\
&+ \sum_{i_3=0}^m \sum_{i_2=0}^m \sum_{i_1=0}^m \left(\int_0^t \int_0^{t_3} \int_0^{t_2} u_{i_3}(t_3) u_{i_2}(t_2) u_{i_1}(t_1) dt_1 dt_2 dt_3 \right) (f_{i_1} f_{i_2} f_{i_3}\phi)(x(t_3))
\end{aligned} \tag{32}$$

Note that the indices in the partial derivatives and in the integrals are in *opposite order*. The pattern emerges clearly, and this procedure may be iterated ad infinitum, yielding a formal infinite series. (In a later lecture we shall repeat this derivation using the very compact notation of chronological products, without writing any integrals.)

Definition 2.1 (Chen-Fliess series) *For any measurable control* $u\colon [0,T] \mapsto \mathbb{R}^{m+1}$ *and a set of* $(n+1)$ *indeterminates* X_0, $X_1, \dots X_m$ *define the formal*

series

$$S_{CF}(T,u)=\sum_I \underbrace{\int_0^T\int_0^{t_{p-1}}\cdots\int_0^{t_3}\int_0^{t_2} u^{i_p}(t_p)\dots u^{i_1}(t_1)\,dt_1\dots dt_p}_{\Upsilon^I(u)(T)}\underbrace{X_{i_1}\dots X_{i_p}}_{X_I} \tag{33}$$

where the sum ranges over all multi-indices $I=(i_1,\dots i_s)$, $s\geq 0$ *with each* $i_j\in\{0,1,\dots m\}$.

This series originates in K. T. Chen's study [6] in the 1950s of geometric invariants of curves in $\mathbb{R}^n$. In the early 1970s Fliess recognized its utility for the analysis of control systems. Using careful analytic estimates one may prove (compare [62]) that this so far only formal series actually converges

Theorem 2.1 *Suppose* f_i *are analytic vector fields on* $\mathbb{R}^n$, $\phi\colon\mathbb{R}^n\mapsto\mathbb{R}$ *is analytic and* $U\subset\mathbb{R}^{m+1}$ *is compact. Then for every compact set* $K\subseteq\mathbb{R}^n$, *there exists* $T>0$ *such that the series (with* Υ^I *and the range of the sum as above)*

$$S_{CF,f}(T,u)(\phi)=\sum_I \Upsilon^I(u)(T)\cdot(f_I\phi)(p) \tag{34}$$

converges uniformly to the solution $x(t,u)$ *of* (30) *with* $x(0)=p$ *for* $p\in K$ *and* $u\colon[0,T]\mapsto U$.

This series solution is not just good for piecewise constant controls, but for all measurable controls. To get a better feeling for the terms, revisit example 13 with

$$f_0=x_1\frac{\partial}{\partial x_2}+x_1^3x_2\frac{\partial}{\partial x_3}\text{ and }\quad f_1=\frac{\partial}{\partial x_1} \tag{35}$$

and consider the Chen-Fliess series for the coordinate functions $\phi=x_i$ about $p=0$. As usual we use $u_0\equiv 1$ and write $u_1=u$. Obviously, for $\phi=x_1$, the series collapses to a single term, yielding $x_1(T,u)(u)(T)=\Upsilon^1(u)(T)=\int_0^T u(t)dt$. For $\phi=x_2$, the series collapses to the single term corresponding to the multi-index $(1,0)$(or *"word"* 10)

$$x_2(T,u)=\Upsilon^{10}(u)(T)=\textstyle\int_0^T\int_0^{t_2}u(t_2)u_0(t_1)\,dt_1\,dt_2=\int_0^T\int_0^{t_2}u(t_1)\,dt_1\,dt_2 \tag{36}$$

As expected, the series just returns the integral form for the linear, double integrators part of the system 13.

For $\phi=x_3$ note that $f_1x_3\equiv 0$ and $f_0x_3=x_1^3x_2$. (Meticulous attention to the *two slots* of differential operators and careful notation are advised: A

differential operator X acts on a function Φ and is evaluated at a point p – the usual identification of points with their coordinates causes the appearance of the same symbol x in both slots!) Next, e.g. $f_1 f_0 x_3 = 3x_1^2 x_2$, while $f_0 f_0 x_3 = x_1^4$. We leave further calculations to

Exercise 2.1 (Important!) *Continuing this example, find all partial derivatives $f_I x_3$ which are not identically zero. What is the highest order non-zero derivative (length of the word I)? How could you have found that length* by inspection, *without calculating any partial derivatives? Find all words I for which $(f_I x_3)(0) \neq 0$ and calculate the values of these derivatives at $p = 0$. Write out the corresponding iterated integrals and explicitly write out the Chen-Fliess series expansion for $x_3(T, u)$.*

The previous exercise, and the following challenge are excellent motivation for all later work. It really helps to first get one's hands dirty with comparatively naive and messy hand-calculations. This way the later elegant combinatorial and algebraic simplifications will be much more appreciated!

Exercise 2.2 (Important!) *Compare the resulting expression of the previous exercise with the* obvious *integral formula*

$$x_3(T,u) = \int_0^T \left(\left(\int_0^{t_3} u(t_2)\, dt_2 \right) \right)^3 \cdot \left(\int_0^{t_3} \int_0^{t_2} u(t_1)\, dt_1\, dt_2 \right) dt_3. \qquad (37)$$

Reconcile these expressions via repeated integration by parts and suitably combining terms.

The example considered above is apparently very special, yielding finite, *polynomial* series expansions in terms of *iterated integrals*. This property is easily traced to the *triangular nature* of the (Jacobian matrices of the) vector fields f_i *together* with their polynomial entries. Such very desirable structure is indeed the objective of nilpotent approximations, to be discussed in after we introduce some technical tools in the next section.

2.2 Families of dilations

For (nonconstant) polynomials of one variable, each derivative lowers the degree by one – something similar clearly is happening in the iterated Lie derivatives in the examples considered in previous sections. This is complemented by the degree with respect to the switching times in the *responses*

$x(T,u)$ in the explicit constructions of the previous chapter. The formal definitions of *families of dilations* are useful to capture the apparent patterns, and for e.g. to identify *leading terms* allowing one to construct approximating systems. Working with fixed coordinates $(x_1, x_2, \dots x_n)$ it is convenient to make the following definition which is a special case of the general geometric, coordinate free, notion of homogeneity of [33]:

Definition 2.2 *Consider* $\mathbb{R}^n$ *with fixed coordinates* $(x_1, x_2, \dots x_n)$ *and* r_1, $r_2, \dots r_n \geq 1$. *A* one-parameter family of dilations *is a map* $\Delta\colon \mathbb{R}^+ \times \mathbb{R}^n$ *defined by*

$$\Delta_s(x) = (s^{r_1}x_1, s^{r_2}x_2, \dots, s^{r_n}x_n). \tag{38}$$

A smooth function $\phi\colon \mathbb{R}^n \mapsto \mathbb{R}$ *and a smooth vector field* F *on* $\mathbb{R}^n$ *are homogeneous of degrees* m *and* k *(with respect to* Δ*), written* $\phi \in H_m$ *and* $F \in \underline{n}_k$, *respectively, if*

$$\phi \circ \Delta_s = s^m \phi \quad \textit{and } Fx_k \in \mathcal{H}_{m+r_k} \quad \textit{for } k = 1, 2, \dots, n. \tag{39}$$

The Euler vector field *of this dilation is the vector field*

$$\nu(x) = r_1 x_1 \frac{\partial}{\partial x_1} + r_2 x_2 \frac{\partial}{\partial x_2} + \dots + r_n x_n \frac{\partial}{\partial x_n}. \tag{40}$$

For example consider $n = 3$, $r = (1, 2, 6)$. The practical meaning of the *exponents* r_i are as *weights* of the coordinate functions, i.e. $x_1 \in \mathcal{H}_1$, $x_2 \in \mathcal{H}_2$ and $x_3 \in \mathcal{H}_6$. With these weights e.g. $\phi(x) = x_1x_3 - x_1^7 + x_1x_2^3 \in \mathcal{H}_7$ is homogeneous of degree 7.
Similarly, the coordinate vector fields are homogeneous of degrees $\frac{\partial}{\partial x_1} \in \underline{n}_{-1}$, $\frac{\partial}{\partial x_2} \in \underline{n}_{-2}$, and $\frac{\partial}{\partial x_3} \in \underline{n}_{-6}$. The Lie derivatives of the homogeneous polynomial ϕ in the directions of the coordinate fields are again homogeneous $\frac{\partial}{\partial x_1}\phi \in \mathcal{H}_6$, $\frac{\partial}{\partial x_1}\phi \in \mathcal{H}_5$, and $\frac{\partial}{\partial x_1}\phi \in \mathcal{H}_6$.

The following properties hold also for more general, geometric dilations as defined in [33].

Proposition 2.2
Let Δ *be a one-parameter family of dilations on* $\mathbb{R}^n$ *with coordinates* $(x_1, \dots x_n)$.
If $\phi \in \mathcal{H}_m$, *and* $\psi \in \mathcal{H}_k$, *then* $\phi\psi \in \mathcal{H}_{m+k}$.
If $F \in \underline{n}_m$, *and* $G \in \underline{n}_k$ *then* $[F, G] \in \underline{n}_{m+k}$.
If $\phi \in \mathcal{H}_m$, *and* $F \in \underline{n}_k$, *then* $F\phi \in \mathcal{H}_{m+k}$.
If $m < -r_n$ *then* $\underline{n}_m = \{0\}$.

Together with the obvious properties for sums, these properties provide the algebras of polynomials and of polynomial vector fields with graded structures: E.g. every polynomial can be uniquely written as a sum of homogeneous polynomials, and every polynomial vector field can be uniquely decomposed into a sum of homogeneous vector fields. This is used e.g. in nilpotent approximating systems and for high-order analogues of linear stability (if the *leading term* of a dynamical system is asymptotically stable, then the system is locally asymptotically stable, compare e.g. [23, 56]

Exercise 2.3 *Prove the assertions made in proposition* (2.2).

The Euler vector field ν is (up to rescaling by a logarithm) the *infinitesimal generator* of the dilation group Δ, and it allows for particularly elegant characterizations of homogeneity.

Proposition 2.3 [33]
Let Δ be a one-parameter family of dilations on $\mathbb{R}^n$ with coordinates $(x_1, \dots x_n)$.
A smooth function ϕ on $\mathbb{R}^n$ is homogeneous $\phi \in \mathcal{H}_m$ iff $\nu\phi = m\phi$.
A smooth vector field on $\mathbb{R}^n$ is homogeneous $F \in \underline{n}_m$, and $G \in\in \underline{n}_k$ iff $[\nu, F] = mF$.

Actually, one may start with a vector field ν such that $\dot{x} = -\nu(x)$ is globally asymptotically stable, and then define a dilation Δ associated to ν via the properties in proposition 2.3 [33].

Exercise 2.4 *Prove the assertions made in proposition* (2.3).

A typical use of the last property in proposition (2.2) is to justify stopping to compute Lie brackets of vector fields after reaching a certain maximal length. E.g. suppose that the vector fields f_0 and f_1 have polynomial components and for a some choice of exponents $(r_1, \dots r_n)$ they are sums of homogeneous vector fields all of which have negative degrees. Then every bracket of length larger than $-r_n$ is identically zero.
Reconsider the vector fields $f_0 = (0, x_1, x_1^3 x_2)^T$ and $f_1 = (1, 0, 0)^T$ from example (1.3). These are homogeneous of degrees $f_0 \in \underline{n}_0$ and $f_1 \in \underline{n}_{-1}$ with respect to the dilation defined by $r = (1, 1, 4)$, while they are homogeneous of degrees $f_0, f_1 \in \underline{n}_{-1}$ with respect to the dilation defined by $r = (1, 2, 7)$. Using the second dilation we conclude that any Lie bracket involving more than 7 factors f_0 or f_1, in any order, with any bracketing is identically zero.
Recall:

Definition 2.3 *A Lie algebra L is called* nilpotent *if there exists a number s such that every iterated Lie bracket of elements of L of length greater than s is zero.*

Thus in the example, we conclude that $L(f_0, f_1)$ is nilpotent. It can be shown [31] that if the Lie algebra $L(f_0, f_1, \dots f_n)$ is nilpotent, then the control system (8) can be brought into a strictly lower triangular form with polynomial vector fields (with well-defined maximal degrees) via a change of local coordinates: In the new coordinates each component $f_i x_j$ is a polynomial in $x_1, x_2, \dots x_{j-1}$ only! Consequently, solution curves corresponding to any control $u(t)$ can be found by simple integrations of functions of a single variable, no nontrivial differential equations need to be integrated! This makes nilpotent systems a very attractive class to work with, and predestined to serve as a class of approximating systems – to be discussed in the next section.

The examples, and especially exercises 1.18 and 1.19, using piecewise constant controls also illustrated that, at least in the case of *homogeneous* systems, the length of each Lie bracket corresponds to the degree of the the polynomial expression in the data (switching times, control values). This is made precise using the notion of homogeneity.
Fix a control $u\colon [0,T] \mapsto U$. For $\varepsilon, \delta \in [0,1]$ define the families of rescaled controls

$$u_{\varepsilon,\delta}\colon [0, \delta T] \mapsto \varepsilon U \subseteq U \quad \text{by} \quad u_{\varepsilon,\delta}(\delta t) = \varepsilon u(t) \tag{41}$$

For the scaling by *amplitude*, using ε, to make sense, assume that the set U is star-shaped with respect to zero, i.e. $[0,1]U \subseteq U$ (meaning $\lambda c \in U$ for all $c \in U$ and all $0 \leq \lambda \leq 1$).

Proposition 2.4 *Suppose $\Delta\colon (s,x) \mapsto \Delta_s(x) = (s^{r_1} x_1, \dots s^{r_n} x_n)$ is a family of dilations on $\mathbb{R}^n$. If the system is homogeneous such that $f_1 \in \underline{n}_{-1}$ and $f_0 \in \underline{n}_{-\theta}$ for some $\theta \in [0,1]$ then*

$$x(s^\theta T, u_{s^{1-\theta}, s^\theta}) = \Delta_s(x(T,u)) \quad \textit{for all} \quad s \in [0,1]. \tag{42}$$

Of particular importance are the special cases $\theta = 0$ and $\theta = 1$ which yield, respectively:

$$\text{if } f_0 \in \underline{n}_0,\ f_1 \in \underline{n}_{-1} \quad \text{then} \quad x(T, u_{\varepsilon,1}) = \Delta_\varepsilon(x(T,u)) \quad \text{for all} \quad \varepsilon \in [0,1] \tag{43}$$

$$\text{if } f_0, f_1 \in \underline{n}_{-1} \quad \text{then} \quad x(\delta T, u_{1,\delta}) = \Delta_\delta(x(T,u)) \quad \text{for all} \quad \delta \in [0,1] \tag{44}$$

Many control systems, especially *free nilpotent* systems, admit several different dilations so that with respect to one dilation one may e.g. have $f_0 \in \underline{n}_0$ while w.r.t. another dilation one has $f_0 \in \underline{n}_{-1}$. In such case one may directly use separate scalings for time and size:

$$x(\delta T, u_{\varepsilon,\delta}) = \Delta_{\varepsilon}^{(1)}(\Delta_{\delta}^{(2)}(x(T,u))) \text{ for all } \varepsilon, \delta \in [0,1]. \tag{45}$$

A simple proof uses uniqueness of solutions of initial value problems, showing that both the right and left hand side of (45) are solutions of the same dynamical system, see e.g. [32].
This proposition is at the heart of many classical sufficient conditions for STLC as it basically allows one to construct control variations that will generate a specific tangent vector to the reachable sets, and which in some sense singles out the lowest order term or bracket according to some weighting scheme. The classical needle variations are built around arguments involving basically the dilations $\Delta_{1,\delta}$ (i.e. $\theta = 1$), while a Taylor expansion in the control sizes, and Hermes' sufficient condition is built around the dilation $\Delta_{\varepsilon,1}$ (i.e. $\theta = 0$). Sussmann's general sufficient condition allows a trade-off between the time-scale and amplitude.

Exercise 2.5 *If possible find a one-parameter family of dilations so that the following system, considered by Jakubczyk in the 1970s, is homogeneous. Find all values of* $\frac{r_2 - r_1}{r_1}$*, or of "*θ*" for which the term* x_1^3 *is of lower order than the definite term* x_2^2 *(which appears a potential obstruction to STLC) (compare theorem* 1.10*). Also, compute all nonzero Lie brackets of the vector fields* f_0 *and* f_1 *defining this system.*

$$\left\{ \begin{array}{lcll} \dot{x}_1 & = & u & |u(\cdot)| \leq c \\ \dot{x}_2 & = & x_1 & x(0) = 0 \\ \dot{x}_3 & = & x_2^2 + x_1^3 & \end{array} \right. \tag{46}$$

2.3 Nilpotent approximating systems

When a nonlinear control system of form (8) is controllable by virtue of the linear condition (theorem (1.6)), then it makes sense for many applications (that involve only/primarily the local behaviour near the equilibrium) to approximate the system (8) by the linear system $\dot{x} = Ax + Bu$ where $A = (Df_0)(0)$ equals the Jacobian matrix of partial derivatives if the drift f_0, and where the i-th column of B equals the value of $f_i(0)$, $i = 1, \ldots, m$. (Of course, this can be (but rarely is) formulated in a coordinate-free geometric way that does not mix up the state space and its tangent spaces.)

Exercise 2.6 *Calculate the standard linearized systems for the models* (9) of a car/bicycle (example (1.1)) and for the models (5) and (6) for the dynamics of a rolling penny (example (1.2)). Discuss the (linear) controllability properties of the linearized systems, and contrast these with the earlier findings from the first sections.

The exercises make it clear that for some nonlinear systems that are reasonably *"realistic"* the standard linearization causes a dramatic loss of information. Thus one asks for alternatives: Reasonable demands are that the approximating systems are elements of a reasonably rich class of systems that allows for the preservation of controllability or stabilizability properties, that systems in this class are amenable to reasonable analysis and computation, and that the approximation is algorithmic and allows for explicit computation. At this time no such ideal approximating scheme is known – the main culprit being the lack of conditions for STLC that are both necessary and sufficient. However, a very good solution is known that preserves STLC for virtually all systems that are known to be STLC by virtue of Sussmann's general sufficiency condition, theorem 1.10. However, there exist STLC systems for which the standard algorithms yield a nilpotent approximating system that is not STLC. But such systems are considered to be quite exotic – the most simple case is the system (54) considered at the end of this section.

We give a crude outline of an algorithm attributed to Hermes (compare the review [23]) (very similar constructions were employed at almost the same time by Stefani and others), omitting some technical steps that are not central and not essential here. See the review [23], or the original references for more details, especially Stefani [59] for details about adapted charts.) Assuming that the original systems of form (8) is STLC by virtue of theorem 1.10, the objective of this procedure is to construct a nilpotent approximating systems, on the *same* state space $\mathbb{R}^n$, of the same form

$$\dot{y}(t) = g_0(y) + \sum_{i=1}^{m} u_i(t)\ g_i(y) \tag{47}$$

(together with coordinates $y_1, \ldots, y_n$) such that not only $L(g_0, g_1, \ldots, g_n)$ is nilpotent, but so that in addition the vector fields g_j are polynomial and (their Jacobian matrices of partial derivatives w.r.t. y_j are) strictly lower triangular. Recall, that for any such system the solution curves for any given function $u(t)$ are obtained explicitly via simple quadratures (no solution of

nonlinear differential equations is needed). Thus, one considers *nilpotent approximations* as the natural nonlinear analogue of linearizations for systems that exhibit truly nonlinear behaviour, i.e. are more than just nonlinear perturbations of linearly controllable systems.

Start with calculating iterated Lie brackets of the vector fields of increasing length until their values at $x_0 = 0$ span the tangent space $T_0\mathbb{R}^n$. If necessary, continue further until brackets are found that *neutralize possible obstructions to STLC* as defined in theorem 1.10 for a suitable weight $\theta \in (0, 1]$. (It may happen that one can choose among different weights, and thus construct many different nilpotent approximating systems.) It is always possible to choose all weights to be rational. Determine the Lie brackets f_{π_i} such that

$$\text{span}\{f_{\pi_1}(0), f_{\pi_2}(0), \ldots, f_{\pi_n}(0)\} = T_0\mathbb{R}^n \tag{48}$$

and they are of lowest possible weight, defined as the weighted sum of θ times the number of factors f_0 plus the number of factors of the controlled fields f_i, $i \geq 1$ in f_{π_i}. (This is very sloppy, see the discussion of *formal brackets* in the next section.) Define the exponents r_i to equal these weighted sums. If necessary, perform a linear coordinate change such that $f_{\pi_1}(0) = \frac{\partial}{\partial x_i}$ for $i = 1, 2, \ldots n$. Commonly one thrives to have $1 \leq r_1 \leq r_2 \leq \ldots, r_n$. (If homogeneity of the new vector fields is needed, e.g. as for feedback stabilization techniques, a strictly triangular polynomial coordinate change may have to be performed, see [59] for *"adapted charts"*.) Using the new coordinates, again called $(x_1, \ldots, x_n)$, define a group of dilations by $\Delta_s(x) = (s^{r_1}x_1, \ldots, s^{r_n}x_n)$.

Expand each component $f_i x_j$ in a Taylor series in the new coordinates, and truncate each expansion keeping only polynomials $p_{ij}(x)$ of order less or equal to $r_j - 1$ for $i \geq 1$, and $r_j - \theta$ for $i = 0$. Define the vector fields $g_j = \sum_{j=1}^{n} p_{ij}(x)\frac{\partial}{\partial x_j}$. These are easily checked to be (sums of) homogeneous vector fields of negative degree of homogeneity and thus, they generate a nilpotent Lie algebra. The preservation of STLC properties follows from the observation that if g_σ is an iterated Lie bracket of the g_i, and f_σ is the corresponding bracket of the f_i, then their components $g_i x_j$ and $f_i x_j$ agree up to a well-defined degree, and in particular,

$$f_{\pi_i}(0) = g_{\pi_i}(0) \quad \text{for all} \quad i = 1, \ldots n. \tag{49}$$

Note that this is only a rough outline of the procedure as a precise description requires a few more technical details and symbols. See the original references of the survey [23] for details.

For illustration consider the model (9) of a car/bicycle (example 1.1). Recall:

$$f_0(x) = \begin{pmatrix} 0 \\ 0 \\ x_2 \cos x_4 \\ x_2 \tan x_1 \\ x_2 \sin x_4 \end{pmatrix}, \qquad f_1(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \qquad f_2(x) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{50}$$

One readily computes $[f_1, f_2] \equiv 0$. Selected other brackets are:

$$[f_1, f_0] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ x_2 \sec^2 x_1 \\ 0 \end{pmatrix}, \; [f_2, f_0] = \begin{pmatrix} 0 \\ 0 \\ \cos x_4 \\ \tan x_1 \\ \sin x_4 \end{pmatrix}, \; [[f_0, f_1], f_2] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \sec^2 x_1 \\ 0 \end{pmatrix},$$

$$\text{and finally} \qquad [[f_0, f_2], [[f_0, f_1], f_2]] = \begin{pmatrix} 0 \\ 0 \\ -\sec^2 x_1 \sin x_4 \\ 0 \\ \sec^2 x_1 \cos x_4 \end{pmatrix}. \tag{51}$$

In principle there is a large number of other brackets that should be calculated, too. However, advanced knowledge from the next lectures (Hall bases) allow one to calculate only a minimal number of brackets. And once Sussmann's theorem 1.10 applies one always can stop. Note that at the origin these vector fields have the values:

$$f_1(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; f_2(0) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; [f_2, f_0](0) = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

$$f_{\pi_4}(0) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \; f_{\pi_5}(0) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{52}$$

where $f_{\pi_4} = [[f_0, f_1], f_2]$ and $f_{\pi_5} = [[f_0, f_2], [[f_0, f_1], f_2]]$. These iterated brackets span the tangent space at the origin, thereby guaranteeing accessibility. Clearly the system is not linearly controllable (it does not satisfy the conditions in theorem 1.6).

Exercise 2.7 *Explain why no matter how many brackets one uses that contain any number of factors f_0, but only a single factor f_1 or f_2, their values at 0 will never span $T_0\mathbf{R}^5$.*

While technically one needs to verify that indeed no *lower order possible obstructions* are nonzero at 0, it is quite apparent that no surprises can happen. (For a rigorous argument, use Hall bases from the next section, and check *ALL* brackets of length at most 5 that appear in such a basis.) Define $f_{\pi_1} = f_1$, $f_{\pi_2} = f_2$, and $f_{\pi_3} = [f_0, f_2]$.
As no *potential obstructions to STLC had to be neutralized*, we are free to choose any weight $\theta \in (0, 1]$, e.g. $\theta = 1$. Thus the weight of each of the five selected brackets agrees with its length (see next chapter for more precise language), and we obtain $r = (1, 1, 2, 3, 5)$. There is no need to perform any linear coordinate change as already $f_{\pi_i}(0) = \frac{\partial}{\partial x_i}\Big|_0$ for $i = 1, 2, 3, 4, 5$.
Expanding the components of $f_i x_j$ into Taylor series and keeping in each component $f_j x_i$ only the terms Δ-lowest term of degree no larger than $r_i - 1$ (and for f_0, in general, no larger than $r_i - \theta$) one obtains the approximating fields

$$g_0(x) = \begin{pmatrix} 0 \\ 0 \\ x_2 \\ x_1 x_2 \\ x_2 x_4 \end{pmatrix}, \; g_1(x) = f_1(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{ and } g_2(x) = f_2(x) = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{53}$$

Exercise 2.8 *Verify directly, i.e. using the theorem* 1.10 *that this nilpotent approximating system* (53) is indeed STLC about 0. Moreover verify that the corresponding brackets f_{π_j} and g_{π_j} have the same values at 0.

Exercise 2.9 *Give a (counter)example of a system that illustrates that the choice of the weight $\theta = 0$ may yield an approximating system that is not necessarily nilpotent. (Remark: However, the Lie algebra will be solvable, and thus still allow for a choice of coordinates in which the approximating vector fields are polynomial and triangular, thus allowing for still comparatively simple calculations of trajectories, compare Crouch* [8]*).*

Exercise 2.10 *Calculate an STLC nilpotent approximating systems for the models* (5) and (6) for the dynamics of a rolling penny (example (1.2)).

Finally consider the following system

$$\left\{\begin{array}{lcll} \dot{x_1} &=& u & x(0) = 0 \\ \dot{x_2} &=& x_1 & |u(\cdot)| \leq \varepsilon_0 \\ \dot{x_3} &=& x_1^3 & \\ \dot{x_4} &=& x_3^2 + x_2^7 & \end{array}\right. \tag{54}$$

which has been shown to be STLC in [30]. However, for every weighting $\theta \in [0,1]$ of $\varepsilon = s^{1-\theta}$ and $\delta = s^\theta$, the definite term $\int_0^{\delta T} x_3^2(t, u_{\varepsilon,\delta})dt = \varepsilon^6\delta^9 \int_0^T x_3^2(t, u_{1,1})dt$ is of lower order in s than the term $\int_0^{\delta T} x_2^7(t, u_{\varepsilon,\delta})dt = \varepsilon^7\delta^{15} \int_0^T x_2^7(t, u_{1,1})dt$ which provides controllability! As a consequence, none of the traditional control variations can be used to generate $-\frac{\partial}{\partial x_4}\Big|_0$ as a tangent vector to the reachable sets in order to conclude STLC, and different kinds of families of control variations were invented [30].

Exercise 2.11 *Calculate* all *iterated Lie brackets for the fields in system* (54) *that are nonzero at* 0 *and* recover *the scaling exponents* $(6, 9)$ *and* $(7, 15)$. *Verify that for no choice of* $\theta \in (0, 1]$ *the system* (54) *satisfies Sussmann's sufficient conditions in theorem* 1.10.

3 Combinatorics of words and free Lie algebras

3.1 Intro: Trying to partially factor the Chen Fliess series

This section shall serve as the final motivation to get rid of all excessive symbols, such as iterated integrals, when facing either large computations or deeper theoretical analysis. While the sample calculations may appear rather simple and naive, past experience shows that for many a reader of the subsequent abstract material, they are an essential guide that connects the combinatorial structures with control.

Consider a single input system of form (8), i.e. with $m = 1$ and $u_0 \equiv 1$, $u = u$. Write out the first few terms in the Chen Fliess series (34)

$$\begin{aligned} S_{CF,f}(T,u)(\phi) = 1\cdot\phi(0) &+ \int_0^T 1dt\cdot(f_0\phi)(0) + \int_0^T u(t)dt\cdot(f_1\phi)(0) \\ &+ \int_0^T\int_0^{t_2} 1\cdot 1 dt_1dt_2\cdot(f_0f_0\phi)(0) + \int_0^T\int_0^{t_2} u(t_2)u(t_1)dt_1dt_2\cdot(f_1f_1\phi)(0) \\ &+ \int_0^T\int_0^{t_2} u(t_2)1\, dt_1dt_2\cdot(f_0f_1\phi)(0) + \int_0^T\int_0^{t_2} 1\cdot u(t_1)dt_1dt_2\cdot(f_1f_0\phi)(0) \end{aligned} \tag{55}$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} 1\cdot 1\cdot 1 dt_1 dt_2 dt_3\cdot(f_0 f_0 f_0\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} u(t_3)\cdot 1\cdot 1 dt_1 dt_2 dt_3\cdot(f_0 f_0 f_1\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} 1\cdot u(t_2)\cdot 1 dt_1 dt_2 dt_3\cdot(f_0 f_1 f_0\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} 1\cdot 1\cdot u(t_1) dt_1 dt_2 dt_3\cdot(f_1 f_0 f_0\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} u(t_3)\cdot u(t_2)\cdot 1 dt_1 dt_2 dt_3\cdot(f_0 f_1 f_1\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} u(t_3)\cdot 1\cdot u(t_1) dt_1 dt_2 dt_3\cdot(f_1 f_0 f_1\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} 1\cdot u(t_2)u(t_1) dt_1 dt_2 dt_3\cdot(f_1 f_1 f_0\phi)(0)$$

$$+\int_0^T\int_0^{t_3}\int_0^{t_2} u(t_3)u(t_2)u(t_1) dt_1 dt_2 dt_3\cdot(f_1 f_1 f_1\phi)(0)$$

$$+ \text{ higher order terms}$$

This is just the beginning, and one never should manually manipulate such a huge expression. Indeed, each of the summands is identified by a simple *word* such as 101 or 10 (to be read as finite sequence, like $(1,0,1)$ or $(1,0)$. The *identification* is captured in form of the two maps

$$\mathcal{F}\colon w = a_1 a_2 \dots a_s \mapsto \left(\phi \mapsto (f_w\phi)(0) = (f_{a_1}\circ f_{a_2}\circ\cdots f_{a_s}\phi)(0)\right), \quad \text{and} \tag{56}$$

$$\Upsilon\!\!\!\Upsilon\colon w = a_1 a_2 \dots a_s \mapsto \left(u \mapsto \int_0^T u_{a_s}(t_s)\int_0^{t_s}\dots\int_0^{t_2} u_{a_1}(t_1) dt_1 \dots dt_{s-1} dt_s\right) \tag{57}$$

These two maps take the advanced point of view that each image is itself an operator: In the first case the image is a partial differential operator on (output) functions on the state space. In the second case, the image is an *iterated integral functional* on the space of admissible controls on an interval

$[0,T]$. For later convenience we already define a companion map Υ in terms of the primitives $U(t) = \int_0^t u(s)ds$ of the usual controls.

$$\Upsilon\colon w = a_1 a_2 \dots a_s \mapsto \left(U \mapsto \int_0^T U'_{a_s}(t_s) \int_0^{t_s} \dots \int_0^{t_2} U'_{a_1}(t_1) dt_1 \dots dt_{s-1} dt_s \right) \tag{58}$$

It is well known that there are many ways to rewrite the huge expression of the Chen Fliess series, ways which are better in the sense of both providing much more insight for theoretical analysis and for being much more amenable for calculation and design (such as path planning). Such alternative forms may be obtained through direct simultaneous manipulation of the analytical objects on right hand sides of (56) and (57), or alternatively through purely algebraic and combinatorial manipulation of the combinatorial objects on the left hand side of (56) and (57).

For illustration, we shall perform some of the analytic operations for a typical objective on some of the low order terms written out above. Then we will repeat the same working only with the indices w. This hopefully will lead even the last skeptics to look positively on combinatorics, and it will motivate the *chronological algebra* structure which makes Υ a *chronological algebra homomorphism.*

One reasonable question to ask in view of this series, and in view of the ubiquitous presence of iterated Lie brackets (and their important geometric roles) in nonlinear control, as exhibited in the previous section, is: *"Where are the Lie brackets in the Chen Fliess series"* (or in above big expression (3.1)). The previous chapters analyzed systems using almost exclusively vector fields which are first order derivatives (all Lie brackets are vector fields!), whereas above formula contains primarily partial differential operators of arbitrarily high order!

Let us consider the terms containing one f_0 and one f_1, followed by looking at the terms containing one f_0 and two f_1. In particular, noting that $[f_1, f_0]\phi = f_1 f_0 \phi - f_0 f_1 \phi$, we add and subtract the following term (which does not appear in the series!) (alternative choices are possible)

$$\int_0^T \int_0^{t_2} 1 \cdot u(t_1) dt_1 dt_2 \cdot (f_0 f_1 \phi)(0)$$

then combine the results appropriately (alternatively integrate by parts)

$$\int_0^T \int_0^{t_2} u(t_2) \cdot 1 \, dt_1 dt_2 \cdot (f_0 f_1 \phi)(0) + \int_0^T \int_0^{t_2} 1 \cdot u(t_1) \, dt_1 dt_2 \cdot (f_1 f_0 \phi)(0) =$$

$$
\begin{aligned}
&= \left(\int_0^T u(t_2) \int_0^{t_2} 1 \, dt_1 dt_2 + \int_0^T 1 \cdot \int_0^{t_2} u(t_1) \, dt_1 dt_2 \right) \cdot (f_0 f_1 \phi)(0) \\
&\qquad + \int_0^T \int_0^{t_2} 1 \cdot u(t_1) dt_1 dt_2 \cdot ((f_1 f_0 - f_0 f_1)\phi)(0) \\
&= \left(\int_0^T u(t) dt \right) \cdot \left(\int_0^T 1 \, dt \right) \cdot (f_0 f_1 \phi)(0) \\
&\qquad + \left(\int_0^T 1 \cdot \left(\int_0^{t_2} u(t_1) \, dt_1 \right) dt_2 \right) \cdot ([f_1, f_0]\phi)(0)
\end{aligned}
\tag{59}
$$

An important observation is that above sum of two second order partial derivatives with iterated integral coefficients is now expressed as a sum of one first order derivative with an iterated integral coefficient and a second order partial derivative with a product of integrals as coefficient.

For comparison let us write down the bare essentials to code all the terms in above calculation.

$$01 \otimes 01 + 10 \otimes 10 = (01 + 10) \otimes 01 + 10 \otimes (10 - 01) = (0 \sqcup\!\sqcup 1) \otimes 01 + 10 \otimes [10]$$

Barely one line, and already providing a preview of a product on *words* that will encode the pointwise multiplication of functions of a single variable, or of iterated integral functionals. This *shuffle product* shall be studied formally in subsequent sections.

Now consider the third order terms that contain exactly two factors of f_1 and one f_0. This time strategically integrate by parts repeatedly, instead of judiciously adding and subtracting terms. This has the same effect, and illustrates the duality. (The first approach, which focused on the vector fields as opposed to the integrals, though, appears to be closer to the technique of *rewriting systems* of algebraic combinatorics, compare [48] and [54]. (Caveat: The following might be done a little faster, but in the end one should always use the algebra, instead of trying to improve the lengthy integrations by parts.) Start with integrating by parts the inside integral in the first term

$$
\begin{aligned}
&\int_0^T u(t_3) \int_0^{t_3} u(t_2) \int_0^{t_2} 1 \, dt_1 dt_2 dt_3 \cdot (f_0 f_1 f_1 \phi)(0) \\
&\qquad + \int_0^T u(t_3) \int_0^{t_3} 1 \int_0^{t_2} u(t_1) dt_1 dt_2 dt_3 \cdot (f_1 f_0 f_1 \phi)(0)
\end{aligned}
$$

$$+\int_0^T 1\int_0^{t_3} u(t_2)\int_0^{t_2} u(t_1)dt_1dt_2dt_3\cdot(f_1f_1f_0\phi)(0)$$

$$=\left(\int_0^T u(t_3)\left(\cdot\left(\int_0^{t_3}u(t_2)dt_2\right)\cdot\left(\int_0^{t_3}1dt_2\right)-\int_0^{t_3}1\int_0^{t_2}u(t_1)dt_1dt_2\right)dt_3\right)\cdot(f_0f_1f_1\phi)(0)$$

$$+\int_0^T u(t_3)\int_0^{t_3}1\int_0^{t_2}u(t_1)\,dt_1dt_2dt_3\cdot(f_1f_0f_1\phi)(0)$$

$$+\int_0^T 1\int_0^{t_3}u(t_2)\int_0^{t_2}u(t_1)\,dt_1dt_2dt_3\cdot(f_1f_1f_0\phi)(0)$$

After suitably regrouping the first term again integrate by parts, and combine the second and third term, recognizing that $f_0f_1f_1-f_0f_1f_1=[f_1,f_0]f_1$

$$=\left(\left(\int_0^T 1\,dt\right)\cdot\left(\int_0^T u(t_3)\int_0^{t_3}u(t_2)dt_2dt_3\right)-\int_0^T 1\int_0^{t_3}u(t_2)\int_0^{t_2}u(t_1)dt_1dt_2dt_3\right)\cdot(f_0f_1f_1\phi)(0)$$

$$+\left(\int_0^T\left(u(t_3)\int_0^{t_3}1\int_0^{t_2}u(t_1)\,dt_1dt_2\right)dt_3\right)\cdot([f_1,f_0]f_1\phi)(0)$$

$$+\int_0^T 1\int_0^{t_3}u(t_2)\int_0^{t_2}u(t_1)dt_1dt_2dt_3\cdot(f_1f_1f_0\phi)(0)$$

Combine the second and fourth terms, and integrate the third term by parts (outer integral). Also write the first term as a product of three integrals.

$$=\tfrac{1}{2}\cdot\left(\int_0^T 1\,dt\right)\cdot\left(\int_0^T u(t)\,dt\right)^2\cdot(f_0f_1f_1\phi)(0)$$

$$+\int_0^T 1\int_0^{t_3}u(t_2)\int_0^{t_2}u(t_1)\,dt_1dt_2dt_3\cdot(f_1f_1f_0-f_0f_1f_1\phi)(0)$$

$$+\left(\left(\int_0^T u(t)dt\right)\left(\int_0^T 1\int_0^{t_3}u(t_2)dt_2dt_3\right)-\int_0^T 1\cdot\left(\int_0^{t_3}u(t_2)dt_2\right)^2dt_3\right)\cdot([f_1,f_0]f_1\phi)(0)$$

Finally combine the second and fourth terms, recognizing that the integral in the fourth term is twice the integral in the second term.

$$=\tfrac{1}{2}\cdot\left(\int_0^T 1dt\right)\cdot\left(\int_0^T u(t)dt\right)^2\cdot(f_0f_1f_1\phi)(0)$$

$$+\left(\int_0^T u(t)dt\right)\cdot\left(\int_0^T 1\int_0^{t_3}u(t_2)dt_2dt_3\right)\cdot([f_1,f_0]f_1\phi)(0)$$

$$+\left(\int_0^T 1\int_0^{t_3}u(t_2)\int_0^{t_2}u(t_1)dt_1dt_2dt_3\right)\cdot[f_1,[f_1,f_0]]\phi)(0)\ .$$

The last step used that

$$f_1f_1f_0 - f_0f_1f_1 - 2[f_1,f_0]f_1 = f_1f_1f_0 - 2f_1f_0f_1 + f_0f_1f_1 = [f_1,[f_1,f_0]]. \quad (60)$$

What matters, aside from experiencing the painful book-keeping, is that again the three third-order partial derivatives with iterated integral coefficients of the original series can be written as a sum of a first, a second order and third order partial derivative, with corresponding products of iterated integrals as coefficients. The emerging pattern is very suggestive. However, this naive approach of repeatedly integrating by parts is no way to deal with the infinite series.

To illustrate the usefulness of this expression, suppose that $f(0) = 0$ and ϕ is a function such that $f_1\phi \equiv 0$ and $[f_1,f_0]\phi(0) = 0$ (this is very similar to the examples discussed in the first chapter). In this case the *leading term* in the rewritten Chen Fliess series (assuming that similar calculations to above have been carried out with analogous results for the other *homogeneous components*) is the last term in the result of our previous calculation. If the iterated integrals corresponding to the *words* 1 and 10 both vanish (by, say, a judicious choice of a piecewise constant control), then again the lowest order nonvanishing term is the last term in our result. Note, in the first argument we used the *product structure* of the partial differential operators (e.g. if $f_1\phi \equiv 0$ then $f_\pi f_1\phi \equiv 0$ for *every* partial differential operator f_π). In the second argument we used the product structure of the rewritten iterated integrals that appear as coefficients of the non-first order operators. Clearly, there are lots of opportunities to combine these arguments, and indeed this is a route towards obtaining conditions for STLC and for optimality!

It turns out that the *expected* result is true, and even more: The entire series can be written as a product of nice flows (of constant vector fields!), or as the exponential of a single field. A partial factorization was used for obtaining a new necessary condition for STLC in [28], but it was clear that this is not the way to go. In [64] Sussmann managed to factor the entire series using differential equations techniques, compare section 4.3. An elegant alternative is to do away with all integrals and such, and proceed purely combinatorially, which allows one to focus on the underlying algebraic structure.

We conclude this last motivation for *combinatorics of words* with the combinatorial analogue of above calculation:

$$011 \otimes 011 + 101 \otimes 101 + 110 \otimes 110 =$$

$$= ((011 + \overbrace{101) - 101}) \otimes 011 + 101 \otimes 101 + 110 \otimes 110$$

$$= ((011+101+\overbrace{110) - 110}) \otimes 011 + 101 \otimes (101-011) + 110 \otimes 110$$

$$= (011+101+110) \otimes 011 + ((101+\overbrace{2*110) - 2*110}) \otimes (101-011)$$
$$+ \ 110 \ \otimes \ (110-011)$$

$$= (011+101+110)\otimes 011 + (101+2*110)\otimes(101-011)$$
$$+ \ 110 \otimes ((110-011) - 2 * (101 - 011))$$

$$= \tfrac{1}{2}(0 \sqcup\!\sqcup 1 \sqcup\!\sqcup 1) \otimes 011 + (10 \sqcup\!\sqcup 1) \otimes [10]\,1 + 110 \otimes [1[10]]$$

with the last line containing the abbreviated form involving *shuffle products*, see below. At this time the combinatorial *rewriting rules* used here may still look unfamiliar, but they simply code *integration by parts*. The following lectures shall give an introduction into this world of a different algebra. We shall aim first for a formal definition of the product $\sqcup\!\sqcup$ on words that encodes products of iterated integral functionals is needed. Together with a *systematic* choice of bases, it should reduce the above calculations to simply inverting the matrix corresponding to a change of basis in some vector space. Being able to use simple linear algebra, it will turn out to be rather easy to compute a powerful continuous analogue of the Campbell Baker Hausdorff formula [39]

3.2 Combinatorics and various algebras of words

This section provides a very basic introduction to the terminology commonly used in an area of combinatorial algebra commonly known as *combinatorics of words*. For a comprehensive introduction accessible to the non-specialist we refer the reader to consult the book *Combinatorics on words* by *"Lothaire"* [46] with the same title. For a more advanced treatment of many of the objects with applications to nonlinear control, we refer to the book *Free Lie algebras* by Reutenauer [54].

The basic idea from the control-perspective is to directly manipulate the *multi-indices* that appeared in the preceding calculations, rather than carry around the bulky overhead of iterated integrals, control functions and vector fields, when carrying out what effectively are purely algebraic or combinatorial manipulations. Moreover, as indicated previously, there is a need to work with formal brackets as opposed to brackets of vector fields (which are just vector fields, and thus have no numbers of factors etc.). Finally, there

are many algebraic theorems and constructions available, starting with constructions of bases and formulas for their dual bases, that are very useful on control.

Start with a set Z whose elements are in one-to-one correspondence with the vector fields $f_0, f_1, \dots f_m$ and with the controls $u_0 \equiv 1, u_1, \dots u_m$. Occasionally it is convenient to simply use the indices $Z = \{0, 1, 2, \dots, m\} \subseteq \mathbf{Z}_0^+$ considered as formal symbols (not as integers). In general $Z = \{X_0, X_1, \dots X_m\}$ is just a set of *indeterminates* X_i. In the sequel we shall refer to this set as an *alphabet*, and to its elements as *letters*. In principle, this set can be infinite, but for most of our purposes finite sets suffice (Lazard elimination in chapter 4 is an exception). A *word* is a finite sequence $(a_1, \dots a_s)$ with $a_i \in Z$ and $s \in \mathbf{Z}_0^+$. It is customary to write $a_1 a_2 a_3 \dots a_s$ for the sequence $(a_1, \dots a_s)$, to use $a, b, c, \dots$ for letters in Z while u, v, w, z for words, to write e or 1 for the *empty word* defined by $we = ew = w$ for all words w. Write Z^+ for the set of all nonempty words and $Z^* = Z^+ \cup \{e\}$ for the set of all words. The set Z^* of all words forms a free monoid (semigroup) (associative, but noncommutative) under the concatenation product

$$(a_1 a_2 \dots a_s \ , \ b_1 b_2 \dots b_r) \mapsto a_1 a_2 \dots a_s b_1 b_2 \dots b_r \tag{61}$$

From the control perspective, on the side of the vector fields, this concatenation product clearly just corresponds (via the map $\mathcal{F}$ in (56)) to compositions of partial differential operators. But on the control and iterated integrals side, via the map Υ from (58) (or $\mathit{\Upsilon\!\Upsilon}$ from (57)) it is much more interesting as a product $\Upsilon(w)\Upsilon(z)$ of iterated integrals in the form special form as they arose in the derivation (31) of the Chen Fliess series is *not* an iterated integral of the same form – although, conceivably, through laborious repeated integration by parts, it can be written as a linear combination of iterated integrals in that special form. One of the purposes of this section is to take care of that kind of manipulation once for all!

As linear combinations are clearly needed, we consider the *free associative algebra* $A(Z) = A_{\mathbf{R}}(Z)$ of all finite linear combinations (with real coefficients) of words in Z^*, and linearly extending the concatenation product in the obvious way. (This algebra is also known as the *algebra of polynomials in noncommuting variables.*) Write $\hat{A}(Z) = \hat{A}_{\mathbf{R}}(Z)$ for the algebra of formal power series over Z (with the same concatenation product).

Define the Lie bracket as the bilinear map $[\cdot, \cdot]\colon A(Z) \times A(Z) \mapsto A(Z)$ that satisfies $[w, z] = wz - zw$ for words $w, z \in Z^*$.

Exercise 3.1 *Verify that is* $(A, \circ)$ *is any associative algebra, then the commutator* $[\cdot,\cdot]\colon A \times A \mapsto A$ *defined by* $[x, y] = x \circ y - y \circ x$ *satisfies the Jacobi identity.*

An element of $A(Z)$ is called a Lie polynomial if it lies in the smallest subspace of $A(Z)$ that contains Z and that is closed under the Lie bracket. It is nontrivial, requiirng one consequence of the Poincaré-Birkhoff-Witttheorem (4.6) (compare [54]) to show that this subspace of Lie polynomials (with the Lie bracket as above) is *the free Lie algebra* over Z, denoted $L(Z)$.
The next section will address the quest for bases of the free Lie algebra $L(Z)$. Under the natural extension of the map $\mathcal{F}$ in (56) to $A(Z)$ and thus to $L(Z)$, any such basis maps to a spanning set of $L(f_0, f_1, \ldots f_m)$, i.e. provides a *minimal* set of Lie brackets to calculated / considered in control applications.

Next we try to distill the essence of the algebraic structure of the iterated integrals in (31). After little reflection it is clear that the construction of iterating the integral form of the differential equation invariably leads one to the noncommutative product

$$* : (U(t),\ V(t)) \mapsto (U * V)(t) = \int_0^t U(s)V'(s)ds \qquad \left(\ = \int_0^t V'(s)U(s)ds \right). \tag{62}$$

(or to its mirror image). In (31) consider for example

$$U(t) = \int_0^t u_{i_3}(t_3) \int_0^{t_3} u_{i_2}(t_2) \int_0^{t_2} u_{i_1}(t_1)dt_1dt_2dt_3 \ \text{ and } \ V(t) = \int_0^t u_{i_4}(s)ds \tag{63}$$

and their *chronological product*

$$(U * V)(t) = \int_0^t u_{i_4}(t_4) \int_0^{t_4} u_{i_3}(t_3) \int_0^{t_3} u_{i_2}(t_2) \int_0^{t_2} u_{i_1}(t_1)dt_1dt_2dt_3dt_4 \tag{64}$$

We shall quickly identify the *defining identity* satisfied by this product, and then equip the free associative algebra $A(Z)$ with an analogous product, so that the map Υ, linearly extended to $A(Z)$ will be an homomorphism for the resulting algebra structure.
Looking for a three term identity (analogous to associativity or the Jacobi identity) that might possibly characterize this algebra structure, consider the products (of say, absolutely continuous functions) $f, g, h\colon \mathbb{R} \mapsto \mathbb{R}$ taken in different orders:

$$(f * (g * h))(t) = \int_0^t f(s) \cdot g(s) \cdot h'(s)\, ds \quad \text{and} \tag{65}$$

$$(f * g) * h)(t) = \int_0^t \left(\int_0^s f(\sigma) \cdot g'(\sigma) d\sigma \right) h'(s)\, ds \tag{66}$$

This reminds (with good reason) of the laborious integrations by parts in the previous chapter. Indeed, it is almost immediately apparent that this product satisfies, for all (absolutely continuous) functions $f, g, h \colon \mathbb{R} \mapsto \mathbb{R}$ (that vanish at 0) the three term *right chronological identity*

$$f * (g * h) = (f * g) * h + (g * f) * h \tag{67}$$

Definition 3.1 *A (right) chronological algebra is a vector space C with a bilinear product $* \colon C \times C \mapsto C$ that satisfies the* right chronological identity (67) *for all $fg, h \in C$.*

The naturalness and usefulness of this algebra structure for nonlinear control, as well as its natural appearance as the *(Koszul-)dual structure* to that of *Leibniz algebras* which recently have received much attention by algebraists [12, 13, 20, 42, 43, 44, 45, 55], suggest that one study this algebra structure in its own right, just like associative, commutative, and Lie algebras. Refer to [35] for some more abstract investigations. In these notes we shall basically just *use* this product.

Exercise 3.2 *Let $V = L^1_{\mathrm{loc}}(\mathbb{R})$ be the space of locally integrable functions on $\mathbb{R}$ and define Verify that $(V, \star)$ is a chronological algebra with the product $\star; : V \times V \mapsto V$ defined by*

$$(f \star g)(t) = \left(\int_0^t f(s) ds \right) \cdot g(t) \tag{68}$$

There are many interesting chronological subalgebras of the algebras $AC_0(\mathbb{R})$ of absolutely continuous functions that vanish at 0, and of the algebra L^1_{loc} locally integrable functions that deserve attention in their own right. The following exercise gives further examples that open new doors for constructing further chronological algebras.

Exercise 3.3 *Verify directly that each of the products on polynomials and exponentials defined below is a right chronological product.*

$$\begin{array}{rclcrcl} X^k \star X^\ell &=& \frac{1}{k+1} X^{k+\ell+1} & & e^{ikt} \star e^{i\ell t} &=& \frac{(-i)}{k} e^{i(k+\ell)t} \\ X^k * X^\ell &=& \frac{\ell}{k+\ell} X^{k+\ell} & & e^{ikt} * e^{i\ell t} &=& \frac{\ell}{k+\ell} e^{i(k+\ell)t} \end{array} \tag{69}$$

It is not surprising that one can make sense of a *free chronological algebra* $C(Z)$ and even construct it from the free associative algebra $A(Z)$ by defining a *chronological product of words* in terms of the concatenation product. For any letter $a \in Z$, and words *nonempty* $w, z \in Z^+$ define inductively (on the length of the second word)

$$w * a \stackrel{\text{def}}{=} wa \qquad \text{and} \qquad w * (za) = (w * z + z * w)a \tag{70}$$

and extend bi-linearly to the subspace $A^+(Z)$ of $A(Z)$ that is spanned by all nonempty words. Note, it is impossible to extend the definition to all of $A(Z) \times A(Z)$ without loosing some key properties. (However, for some purposes it will be convenient to allow *one* factor to be the empty word e and set $w * e = 0$ and $e * w = w$ if $w \in Z^* \setminus \{e\}$.) With these definitions it is apparent that the following holds:

Theorem 3.1 *The map Υ from $C(Z)$ to a chronological algebra of iterated integral functionals $\mathcal{III}(\mathcal{U})$ is a chronological algebra homomorphism, i.e. for any $w, z \in C(Z)$*

$$\Upsilon(w * z) = \Upsilon(w) * \Upsilon(z) \tag{71}$$

For details on suitable domains $\mathcal{U}$ of the iterated integral functionals see [35]. Here we only note that any such set of functionals immediately inherits the chronological algebra structure from its domain, e.g. the set of absolutely continuous functions. One can show that the map Υ is actually is a chronological algebra *isomorphism* provided the class of admissible inputs is *sufficiently rich*, compare [35]. We sketch the key *inductive step* for nonempty words $w, z \in Z^*$, a letter $a \in Z$, $U \in AC_0([0,T]$ and $T \geq 0$, written out in mini-steps:

$$
\begin{aligned}
\Upsilon_{w*(za)}(U)(T) &= \Upsilon_{(w*z+z*w)a}(U)(T) \\
&= \int_0^T U_a'(t) \cdot \Upsilon_{w*z+z*w}(U)(t) dt \\
&= \int_0^T U_a(t) \cdot \left(\Upsilon_{w*z}(U)(t) + \Upsilon_{z*w}(U)(t)\right) dt \qquad (72) \\
&= \int_0^T U_a'(t) \cdot \left(\int_0^t (\Upsilon_z(U))' \cdot \Upsilon_w(U)(s) ds + \int_0^t (\Upsilon_w(U))' \cdot \Upsilon_z(U)(s) ds \right) dt \\
&= \int_0^T U_a'(t) \Upsilon_z(U)(t) \cdot \Upsilon_w(U)(t) dt \\
&= \int_0^T \frac{d}{dt} \left(\int_0^t U_a'(s) \Upsilon_z(U)(s) ds \right) \cdot \Upsilon_w(U)(t) dt \\
&= (\Upsilon_w(U) * \Upsilon_{za}(U))(T)
\end{aligned}
$$

The first and last step use the definition (58). In between, aside from using the linearity of Υ and the induction hypothesis, the key steps are integration by parts followed by suitable regrouping – exactly the steps from the section 2.1 that we wanted to combinatorially encode.

The symmetrization of the chronological product of functions (or iterated integral functionals) yields the pointwise multiplication, which is both commutative and associative. Since this product is also routinely used in control, it makes sense to formally define and name the corresponding product on the free associative algebra $C(Z)$, or its extension to the free associative algebra $A(Z)$.

Definition 3.2 *The shuffle product is the bilinear map* $\sqcup\!\sqcup : A(Z) \times A(Z) \mapsto A(Z)$ *that satisfies* $w \sqcup\!\sqcup e = e \sqcup\!\sqcup w = w$ *for all* $w \in A(Z)$ *and*

$$
w \amalg z = w * z + z * w \qquad \text{for all } w, z \in C(Z) \qquad (73)
$$

Corollary 3.2 *The map* Υ *from* $A(Z, \sqcup\!\sqcup)$ *to a associative algebra of iterated integral functionals* $\mathcal{III}(\mathcal{U})$ *(with pointwise mutliplication) is a associative algebra homomorphism.*

$$
\Upsilon(w \amalg z) = \Upsilon(w) \cdot \Upsilon(z) \qquad (74)
$$

Using the recursive definition (70) of the chronological product one immediately obtains a recursive formula for the shuffle product. For letters $a, b \in Z$

and words $w, z \in Z^*$

$$(wa) \sqcup\!\sqcup (zb) = (w \sqcup\!\sqcup (zb))a + ((wa) \sqcup\!\sqcup z)b \tag{75}$$

Exercise 3.4 *Verify by direct computation, using* (67)*, that the shuffle product is associative.*

Exercise 3.5 *Calculate at least a handful of shuffle products to get a feeling for it. E.g. calculate the following (but feel free to make your own choices)* $a \sqcup\!\sqcup b$, $a \sqcup\!\sqcup a$, $a \sqcup\!\sqcup a \sqcup\!\sqcup a$, $a \sqcup\!\sqcup a \sqcup\!\sqcup b$, $a \sqcup\!\sqcup b \sqcup\!\sqcup c$, $(ab) \sqcup\!\sqcup c$, $(ab) \sqcup\!\sqcup b$, $(ab) \sqcup\!\sqcup a$, $(ab) \sqcup\!\sqcup (cd)$, $(ab) \sqcup\!\sqcup (ab)$, ...

Our definition of the shuffle product as the symmetrized chronological product makes sense from the point of view of nonlinear control – but this does not do it justice as its algebraic characterization shows that it is a fundamental map. First introduce yet another product, actually a *"coproduct"*.

Definition 3.3
Define the coproduct $\Delta\colon A(Z) \mapsto A(Z) \otimes A(Z)$ *as the linear algebra homomorphism defined on generators* $a \in Z$ *by (using* 1 *for the* empty word *(previously denoted by* e*).*

$$\Delta\colon a \mapsto a \otimes 1 + 1 \otimes a \tag{76}$$

Use [54] as a gentle introduction, and source for more advanced references for the realm of coproducts, co-gebras, bi-grebas, and ultimately Hopf-algebras. (They appear to play a quite useful, though still largely unrecognized role in nonlinear control.)
It is instructive to work a few examples. Suppose $a, b \in Z$. Then

$$\begin{aligned} \Delta(ab) &= \Delta(a)\Delta(b) \\ &= (a \otimes 1 + 1 \otimes a)(b \otimes 1 + 1 \otimes b) \\ &= ab \otimes 1 + a \otimes b + b \otimes a + 1 \otimes ab \\ &\neq ab \otimes 1 + 1 \otimes ab \end{aligned} \tag{77}$$

By symmetry, it is clear that

$$\Delta([a, b]) = [a, b] \otimes 1 + 1 \otimes 1, b] \tag{78}$$

The previous calculation not only holds for $a, b \in Z$, but in much more generality, it is true for any $p, q \in A(Z)$ provided $\Delta(p) = p \otimes 1 + 1 \otimes p$ and $\Delta(q) = q \otimes 1 + 1 \otimes q$. Thus the set of $p, q \in A(Z)$ for which this holds is, maybe not surprisingly:

Theorem 3.3 *A polynomial $p \in A(Z)$ is a Lie polynomial if and only if*

$$\Delta(p) = p \otimes 1 + 1 \otimes p \tag{79}$$

Exercise 3.6 *Prove theorem* 3.3 *using the characterization that the set of Lie polynomials is the smallest subspace of $A(Z)$ that contains Z and is closed under the Lie bracket.*

Before returning to the shuffle product, we note that one may define an inner product on $A(Z)$ by demanding that the *basis* Z^* of $A(Z)$ is an orthonormal basis, i.e. define $<\cdot,\cdot>: A(Z) \times A(Z) \mapsto \mathbb{R}$ for $P_w, Q_w \in \mathbb{R}$ by

$$<P, Q> = \sum_{w \in Z^*} P_w Q_w \quad \text{if } P = \sum_{w \in Z^*} P_w w, Q = \sum_{w \in Z^*} Q_w w \in A(Z), \tag{80}$$

This map extends immediately to a map $<\cdot,\cdot>: \hat{A}(Z) \times A(Z) \mapsto \mathbb{R}$ (or $<\cdot,\cdot>: A(Z) \times \hat{A}(Z) \mapsto \mathbb{R}$), which then is considered as a *bilinear pairing* upon noting that the algebra $\hat{A}(Z)$ of noncommuting formal power series is the *algebraic dual* (space of all linear functionals) of the algebra $\times A(Z)$of noncommuting polynomials. In turn, with the usual topology (compare [35, 54]), $A(Z)$ is the *topological dual* (space of *continuous* linear functionals of $\hat{A}(Z)$. (Note, that if $P = \sum_{w \in Z^*} P_w w$ then $P_w = 0$ for all except a finite number of $w \in Z^*$, and thus the sum in (80) is finite.

Remark 3.4 In the case of a finite alphabet Z, the topology can be characterized by the metric $\|P, Q\| = 2^{-k}$ where k is the *length* of the shortest word $w \in Z^*$ such that $<w, Q - P> \neq 0$. Alternatively, a sequence $P^{(n)} \in A(Z)$ converges to $Q \in A(Z)$ if for every $M > 0$ there exists $N < \infty$ such that $<w, Q - P^{(n)}> = 0$ for all words of length at most N. In the case of an infinite one uses a similar topology, albeit its characterization in terms of neighborhood bases is slightly more technical [54].

Algebraically, the shuffle product is defined in elementary terms using the coproduct:

Definition 3.4 *(Alternate definition) The shuffle product $ш: \hat{A}(Z) \times \hat{A}(Z) \mapsto \hat{A}(Z)$ is the transpose of the coproduct $\Delta: A(Z) \mapsto A(Z) \otimes A(Z)$:*

$$<u ш v, w> = <u \otimes v, \Delta(w)> \quad \textit{for all } u, v, w \in A(Z) \tag{81}$$

Exercise 3.7 *Proof by induction on the lengths of the words that this algebraic definition is equivalent to the recursive combinatorial definition in equation* (75) *or alternatively equations* (70) *and* (73).

While there are many formulas that mix shuffle and concatenation product, they naturally *live* on spaces that are dual to each other, the algebra $A(Z)$ of noncommutative polynomials and the algebra $\hat{A}(Z)$ of noncommutative formal power series. But the latter naturally contains the former, and one may equip each with both products, giving rise to two *Hopf algebra* structures on $A(Z)$ see [54] for details. In these notes we shall not go deeper into this, for details see e.g. [54, 35].
This algebraic characterization makes it easy to establish that the set of Lie polynomials is *orthogonal* to nontrivial shuffles:

Proposition 3.5 *If $p \in L(Z) \subseteq A(Z)$ is a Lie polynomial and and $u, v \in Z^* \setminus \{1\}$ are nonempty words, then*

$$< u \sqcup\!\sqcup v, p > = 0. \tag{82}$$

The proof is a short calculation, using the natural pairing of $\hat{A}(Z) \otimes \hat{A}(Z)$ with $A(Z) \otimes A(Z)$. For a Lie polynomial p and nonempty words u, v calculate

$$\begin{array}{rcl} < u \sqcup\!\sqcup v, p > & = & < u \otimes v, \Delta(p) > \\ & = & < u \otimes v, p \otimes 1 + p \otimes 1 > \\ & = & < u, p > \cdot < v, 1 > \; + \; < u, 1 > \cdot < v, p > \\ & = & 0. \end{array} \tag{83}$$

Finally consider the action and *anti-action* [?] of $A(Z)$ on $A(Z)$ by right and translations, i.e. for $a \in Z$ and $w \in Z^*$ define

$$\varrho_a, \lambda_a \colon A(Z) \mapsto A(Z), \qquad \varrho_a(w) = wa \quad \text{and} \quad \lambda_a(w) = aw \tag{84}$$

Note that $\varrho_a \varrho_b(w) = wba$ reverses the order, while $\lambda_a \lambda_b(w) = abw$ preserves order. It is easy to extend ϱ_w and λ_w to $w \in A(Z)$ but we shall have no need for this. Instead, we are interested in the transposes $\varrho_a^\dagger, \lambda_a^\dagger \colon \hat{A}(Z) \mapsto \hat{A}(Z)$ of these translations which are defined on words $w, z \in Z^*$ by

$$< \varrho_a^\dagger w, z > = < w, za > \qquad \text{and} \qquad < \lambda_a^\dagger w, z > = < w, az > \tag{85}$$

Clearly, if $a, b \in Z$ then $\rho_a^\dagger b = \lambda_a^\dagger b = 1$ if $a = b$ and 0 else. If $b \in Z$ and $w \in Z^* \setminus Z$ are words that are not letters, then

$$\varrho_a^\dagger(wb) = \left\{ \begin{array}{ll} w & \text{if } a = b \\ 0 & \text{else} \end{array} \right\} = \lambda_a^\dagger(bw) \tag{86}$$

Recall that a linear map $D \colon A \mapsto A$ on an algebra A with product $\cdot$ is called a derivation if $D(f \cdot g) = (D(f)) \cdot g + f \cdot (D(g))$ for all $f, g \in A$.

Exercise 3.8 *Show that the composition* $D_2D_1 \colon f \mapsto D_2(D_1(f))$ *of two derivations* D_1 *and* D_2 *need not be a derivation, but the commutator* $D_2D_1 - D_1D_2$ *is always a derivation. Show that the set of derivations on an algebra always has a Lie algebra structure with the commutator as product.*

The following observation appears to have very profound consequences, compare the next section and the next chapter.

Theorem 3.6 *Both transposes* $\varrho_a^\dagger$ *and* $\lambda_a^\dagger$ *are derivations on the shuffle algebra* $(\hat{A}(Z), \sqcup\!\sqcup)$*, but* only *the transpose* $\lambda_a^\dagger$ *of the left translation* λ_a *by a letter* a *is a derivation on the chronological algebra* $(C(Z), *)$*.*

Exercise 3.9 *Prove theorem* (3.6) *using the recursive combinatorial definitions of the shuffle and chronological products in equations* (75) *and* (70)*, and/or the algebraic definition in* (3.4)*.*

3.3 Hall Viennot bases for free Lie algebras

The goal of this section is to develop bases for free Lie algebras, and get some insight into their background, constructions and properties. We start with some remarks introducing binary labeled trees which will be necessary to later construct Hall bases as these critically depend on the notion of left and right factors. Returning to Lie algebras, we then consider the process of Lazard elimination which is the main procedure for generating bases of a free Lie algebra. (However, [54] provides an independent argument.) Finally we survey Hall-Viennot bases and their most important properties.

First recall some of the problems we encountered earlier. In every Lie algebra $[x,[y,[x,y]]] = [y,[x,[x,y]]]$ because $[[x,y],[x,y]] = 0$. This re-emphasizes that a Lie-bracket does not have well-defined left and right factors, and a need for a language of *formal brackets*. Such language also allows one to more precisely phrase the conditions for STLC of the previous chapter: Recall if $f_1 = \frac{\partial}{\partial x_1}$ and $f_0 = x_1^2 \frac{\partial}{\partial x_1}$ then strictly speaking the number of factors f_1 and f_0 in $[f_1,[f_1,f_0]$ is not well defined as e.g. $[f_1,[f_1,f_0] = 2\frac{\partial}{\partial x_1} = f_1$.
Formally, introduce the *"free magma"* $\mathcal{M}(Z)$, the set of rooted binary trees labeled in Z (also called *parenthesized words*). This set is recursively constructed by $M^1(Z) = \{Z\}$,

$$M^{n+1}(Z) = \cup_{k=1}^{n} M^k(Z) \times M^{n+1-k}(Z) \quad \text{and} \quad \mathcal{M}(Z) = \cup_{n=1}^{\infty} M^n(Z) \tag{87}$$

There are canonical maps $\varphi\colon \mathcal{M}(Z) \mapsto L(Z)$ and $\psi\colon \mathcal{M}(Z) \mapsto A(Z)$ defined for $a \in Z$,

$$\varphi(a) = \psi(a) = a \quad \text{and} \quad \varphi(t', t'') = [\varphi(t'), \varphi(t'')]a \quad \text{and} \quad \psi(t', t'') = \psi(t')\psi(t'') \tag{88}$$

Note that every tree $t \in \mathcal{M}(Z)$ is either a letter $t \in Z$ or it has well-defined left and right subtrees $t', t'' \in \mathcal{M}(Z)$, i.e. $t = (t', t'')$. Also, for each tree the numbers of times that each letter appears as a leaf are well-defined. Formally, for $a \in Z$ define $\|\cdot\|_a\colon \mathcal{M}(Z) \mapsto \mathbf{Z}_0^+$ by $\|a\|_a = 1$, $\|b\|_a = 0$ if $a \neq b \in Z$, and recursively for general trees $\|(t', t'')\|_a = \|t'\|_a = \|t''\|_a$. For example if $t = (a(ab))$ then $\|t\|_a = 2$ and $\|t\|_b = 1$. This map naturally carries over to Z^* to Lie monomials (images of single trees under φ. However, the notions of left and right subtree do *not* carry over to $L(Z)$, e.g. consider $Z = \{a, b\}$. Then $(a, b) \in \mathcal{M}(Z)$ maps to $\varphi((a, b)) = [a, b] = [-b, a] \in L(Z)$. Similarly, $(a, (b, (a, b))) \neq (a, (b, (a, b))) \in \mathcal{M}(Z)$, yet

$$\varphi((a, (b, (a, b))) = [a, [b, [a, b]] = [b, [a, [a, b]] = \varphi((a, (b, (a, b))) \tag{89}$$

due to anti-commutativity and the Jacobi-identity in $L(Z)$).

The language of formal brackets or trees is the one which allows for very precise statements of the theorems for controllability, such as Sussmann's theorem (1.10). However, one finds an abundance of rather sloppy statements of this and similar conditions in the recent literature, and our presentation in section 1.4 is just barely *border-line*. See the original article (1.10) for utmost precision.

It may take some time to get used to the precision needed to describe formal operations such as *expressing each iterated Lie bracket as a linear combination of a set of specified brackets*. E.g. working in $L(Z)$ it is correct to write $[[x, y], z] = [x, [y, z]] - [y, [z, x]]$, but for trees $((x, y), z) \neq (x, (y, z)) - (y, (z, x))$ (it is straightforward to introduce linear combinations of trees), visualized as

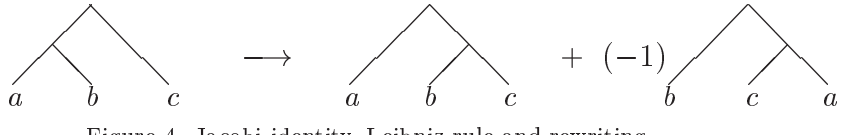


Figure 4. Jacobi identity, Leibniz rule and rewriting

This distinction is precisely what is needed in order to specify an algorithm how to *rewrite* formal brackets (and trees), and then map to identities of Lie brackets in $L(Z)$.

A tree $t \in \mathcal{M}(Z)$ is called a *Dynkin tree*, written $Dyn(Z)$, if either $t \in Z$ or $t = (a, t'')$ with $a \in Z$ and $t'' \in Dyn(Z)$. Note that these trees correspond exactly to the iterated Lie derivatives that appear in control when repeatedly differentiating e.g. output functions along solution curves of the system. The next exercise shows that the image $\varphi(Dyn(Z)) \subseteq L(Z)$ spans $L(Z)$. But since e.g. $(a,(b,(a,b))), (a,(b,(a,b))) \in Dyn(Z)$, yet $[a,[b,[a,b]]] = [b,[a,[a,b]] \in L(Z)$ it is clear that (even after removing obviously trivial trees like (a,a)) that $\varphi(Dyn(Z))$ is not a basis for $L(Z)$. (Nonetheless, this set is often called a *Dynkin basis*, clearly a bad misnomer.) Note that a Lie bracket may be a the image of a Dynkin tree even if it looks quite different, e.g. $[[y,[[x,y],x]],x] =\in \varphi(Dyn(Z))$ because $[[y,[[x,y],x]],x] = [x,[y,[x,[x,y]]]] = \varphi(\,(x,(y,(x,(x,y))))\,) \in \varphi(Dyn(Z))$.

Exercise 3.10 *Show that* $\varphi(Dyn(Z)) \subseteq L(Z)$ *spans* $L(Z)$. Work with trees, and adapt the *"rewriting rule"* $((x,y),z) \longrightarrow \{(x,(y,z)), \quad (y,(x,z))\}$ to recursively reduce any tree to a subset of $Dyn(Z)$. For precise, technical notions of *rewriting systems* see [54].)

The basic construction of bases for $L(Z)$, as well as e.g. the construction of Sussmann's infinite exponential product expansion of theorem (4.10) are fundamentally based on the concept of Lazard elimination. This rests on the following basic theorem, simple proofs (but using technical language beyond the scope of these notes) of which may be found in [4, 54]. We shall be satisfied with a simple illustration of the elimination process.

Theorem 3.7 (Lazard elimination). *Suppose* Z *is a (not necessarily finite) set and* $a \in Z$. *Then the free Lie algebra* $L(Z)$ *is the direct sum of the one-dimensional subspace* $\{a\}_{\mathbf{R}} = \{\lambda a \colon \lambda \in \mathbb{R}\}$, *spanned by* a, *and a Lie algebra that is freely generated (as a Lie algebra) by the set* $\{\mathrm{ad}^{\mathrm{k}}\mathrm{a},\mathrm{b}) \colon \mathrm{k} \geq 0,\ \mathrm{b} \in \mathrm{Z} \setminus \{\mathrm{a}\}\ \}$.

Exercise 3.11 *Adapt the* rewriting *process from exercise* (3.10) *to show that* $L(\{a,b\}) \subseteq \{a\}_{\mathbf{R}} \ \oplus \ L(\{(\mathrm{ad}^{\mathrm{k}}\mathrm{a},\mathrm{b}) \colon \mathrm{k} \geq 0,\ \mathrm{b} \in \mathrm{Z} \setminus \{\mathrm{a}\}\ \})$. In view of the exercise (3.10), it suffices to show (by induction) that every Dynkin bracket

$[a_{i_r}, [a_{i_{r-1}}, [, \dots [a_{i_2}, a_{i_1}]] \dots]$ with $a_{i_j} \in \{a, b\}$ can be written as a linear combination of brackets of the form $[(\mathrm{ad}^{i_s} a, b), \dots [(\mathrm{ad}^{i_2} a, b), (\mathrm{ad}^{i_1} a, b)]] \dots]$ with $i_j \geq 0$. Work either on the level of trees or in $L(Z)$, but carefully reflect on your choices.

For illustration consider a two letter alphabet $Z = \{a, b\}$. Then

$$\begin{aligned} L(Z) &= \{a\}_{\mathbb{R}} \bigoplus \{b, [a, b], [a, [a, b]], \dots\} \\ &= \{a\}_{\mathbb{R}} \oplus \{b\}_{\mathbb{R}} \oplus \{[a, b], [a, [a, b]], \dots, [b, [a, b]], [b, [b, [a, b]], \dots\} \\ &= \{a\}_{\mathbb{R}} \oplus \{b\}_{\mathbb{R}} \oplus \{[a, b]\}_{\mathbb{R}} \oplus \{(\mathrm{ad}^k [a, b], (\mathrm{ad}^j b, (\mathrm{ad}^i a, b))) \colon i, j, k \geq 0\} \\ &= \{a\}_{\mathbb{R}} \oplus \{b\}_{\mathbb{R}} \oplus \{[a, b]\}_{\mathbb{R}} \oplus \{[a, [a, b]]\}_{\mathbb{R}} \\ &\qquad \oplus \{(\mathrm{ad}^\ell [a, [a, b]], (\mathrm{ad}^k [a, b], (\mathrm{ad}^j b, (\mathrm{ad}^i a, b)))) \colon i, j, k, \ell \geq 0\} \\ &= \{a\}_{\mathbb{R}} \oplus \{b\}_{\mathbb{R}} \oplus \{[a, b]\}_{\mathbb{R}} \oplus \{[a, [a, b]]\}_{\mathbb{R}} \oplus \{[b, [a, b]]\}_{\mathbb{R}} \oplus \dots \end{aligned} \tag{90}$$

Note that at every stage the infinite dimensional part is replaced by a new infinite dimensional part that is *generated by infinitely many times "more" generators* than the previous. Nonetheless, on can anticipate the *convergence* as all these generators become *"longer"* and longer (provided the eliminated brackets are chosen properly), compare remark 3.4. What is important to remember for applications and in the sequel, are the successive elimination of one bracket at a time, thereby conceivably constructing a basis, and the type of bracket that is common to all generators, and thus also to all eliminated elements. Note, the above elimination process should again be done on trees, and then mapped to $L(Z)$ – however, we presented it in the more traditional Lie algebra setting.

One defines *Lazard sets* as subsets of $\mathcal{M}(Z)$, that, roughly speaking, arise from infinite repetition of the illustrated Lazard elimination process – the technical statement is quite lengthy [54], and we omit it here. What matters is the following (again, for a proof see [54])

Theorem 3.8 *The image of a Lazard set $\mathcal{L} \subseteq \mathcal{M}(Z)$ under the map φ is a basis for $L(Z)$.*

Starting with Marshall Hall in the 1940s, whose work builds on Phillip Hall's studies of commutator groups in the 1930s, several bases for free Lie algebras

have been proposed. Aside from *Hall bases*, the best known names are *Lyndon bases* and *Sirsov bases*. The latter two were eventually found to basically be same. In the 1970s Viennot [68] slightly relaxed the conditions in the construction of Hall bases, and showed that with that relaxed notion Lyndon (and thus also Sirsov) bases are just special cases of the generalized Hall bases. When we want to emphasize the distinction, we will refer to the latter as *Hall Viennot bases*.

Definition 3.5 *A Hall set over a set Z is any strictly ordered subset $\tilde{\mathcal{H}} \subseteq \mathcal{M}(\mathbf{Z})$ that satisfies*

(i) $Z \subseteq \tilde{\mathcal{H}}$

(ii) *Suppose* $a \in Z$. *Then* $(t, a) \in \tilde{\mathcal{H}}$ *iff* $t' \in \tilde{\mathcal{H}}$, $t' < a$ *and* $a < (t', a)$.

(iii) *Suppose* $u, v, w, (u, v) \in \tilde{\mathcal{H}}$.
Then $(t', (t''', t'''')) \in \tilde{\mathcal{H}}$ *iff* $t''' \leq t' \leq (t''', t'''')$ *and* $t' < (t', (t''', t''''))$.

The definition given here twice reverses the convention of the one given in [54]: All trees have left and right factors swapped, and larger and less have been reversed. On the other hand, the definition given here is compatible with the conventions of [4].
The original Hall bases, as presented in Bourbaki [4], require that ordering be compatible with the length, i.e. if the length of the word $\psi(t)$ is less then the length of the word $\psi(t')$, then $t < t'$. Viennot replaced this condition (and minor other parts) by condition (ii) in definition (3.5).
For the sake of completeness, we state (even without having given a definition of Lazard sets)

Theorem 3.9 *Every Hall set is a Lazard set.*

Theorem 3.10 *The image of a Hall set under the map $\varphi\colon \mathcal{M}(Z) \mapsto L(Z)$ is an ordered basis for $L(Z)$.*

While this is an immediate corollary of theorems 3.9 and 3.8, its importance earns it the title of a theorem. Also Reutenauer [54] also gives a direct proof that is not based on Lazard sets.
It is very straightforward to inductively construct Hall sets, especially when choosing an order that is compatible with the length of the associated word. See the example given in figure 5. However, for many applications Lyndon bases seem to be even more efficient – and effective algorithms have been coded that factor e.g. every word into a product of Lyndon words etc.

Definition 3.6 *Order the alphabet Z. A word $w \in Z^+$ is a* Lyndon word *if it is strictly smaller than its cyclic rearrangements with respect to the lexicographical ordering on Z^*, i.e.*

$$\textit{If } w = uv \textit{ with } u, v \in Z^+ \textit{ then } uv < vu. \tag{91}$$

This ordering is compatible with the choices for Hall bases in [54]. To match our choices, we need to read the words backwards *and* replace *strictly smaller* by *strictly larger* (i.e. reverse the ordering of Z). Note that it is very easy to tell whether a *given* word is a Lyndon word, and it is also easy to factor:

b
$(((((ab)b)b)b)b)$
$((((ab)b)b)b)$
$(a((((ab)b)b)b))$
$(((ab)b)b)$
$((ab)(((ab)b)b))$
$(a(((ab)b)b))$
$(a(a(((ab)b)b)))$
$((ab)b)$
$((ab)((ab)b))$
$(a((ab)((ab)b)))$
$(a((ab)b))$
$\boxed{((ab)(a((ab)b)))}$
$(a(a((ab)b)))$
$(a(a(a((ab)b))))$
(ab)
$((a(ab))(ab))$
$(a((a(ab))(ab)))$
$(a(ab))$
$(a(a(ab)))$
$(a(a(a(ab))))$
$(a(a(a(a(ab)))))$
a

$((a(ab))(b(ab)))$
$((ab)(b(b(ab))))$
$((ab)(b(a(ab))))$
$((ab)(a(a(ab))))$
$(b(b(b(b(ab)))))$
$(b(b(b(a(ab)))))$
$(b(b(a(a(ab)))))$
$(b(a(a(a(ab)))))$
$((ab)(b(ab)))$
$((ab)(a(ab)))$
$(b(b(b(ab))))$
$(b(b(a(ab))))$
$(b(a(a(ab))))$
$(a(a(a(ab))))$
$(b(b(ab)))$
$(b(a(ab)))$
$(a(a(ab)))$
$(b(ab))$
$(a(ab))$
(ab)
b
a

Figure 5. Doubly reversed Lyndon trees and Ph. Hall trees

Note that for short words the reversed Lyndon words with $b < a$ are almost the same as the usual Lyndon words with $a < b$. The framed tree in figure 5 shows that this is just a coincidence for small trees.

Suppose $w \in Z^+$ is a Lyndon word. If $w \in Z$ is a letter there is nothing to be factored. If not, then there exists a pair $(u, v) \in Z^+ \times Z^+$ such that u is the longest *left factor* of w that is a Lyndon word. It can be shown that then v is also a Lyndon word. Repeating this factorization recursively (i.e. for the left and right factors) one obtains a map from the set of Lyndon words into the set $\mathcal{M}(Z)$ of trees which is the right inverse (on the set of Lyndon words) of the map $\psi\colon \mathcal{M}(Z) \mapsto Z^*$ which *forgets* the tree structure and maps each tree to its (ordered sequence of) leaves.

Exercise 3.12
Construct all Lyndon trees with at most 5 *leaves for a three letter alphabet* $Z = \{0, 1, 2\}$.

Exercise 3.13 *Consider the three letter alphabet* $Z = \{0, 1, 2\}$ *and construct an ordered subset of a set of Hall trees containing all trees with at most* 5 *leaves. Be aware of the freedom to choose the ordering of newly constructed trees (and the consequences of the choice upon later constructed trees).*

Exercise 3.14 *Verify that the restriction of the map* $\psi\colon \mathcal{M}(\{0, 1, 2\}) \mapsto Z^+$ *to the set of Hall trees given in figure 5 is one-to-one. Write down the image of this set in* Z^+*, and develop an algorithm that recovers the trees* $t \in \mathcal{M}(\{a, b\})$ *from the images* $\psi(t) \in Z^+$.

Exercise 3.15 *Verify that the restriction of the map* $\psi\colon \mathcal{M}(\{0, 1, 2\}) \mapsto Z^+$ *to the set constructed in exercise* 3.13 *is one-to-one. Write down the image of this set in* Z^+*, and develop an algorithm that recovers the trees* $t \in \mathcal{M}(\{0, 1, 2\})$ *from the images* $\psi(t) \in Z^+$.

The one-to-one-ness of the restrictions of the map $\psi\colon \mathcal{M}(\{0, 1, 2\}) \mapsto Z^+$ to Hall sets $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is one of the most important properties of Hall sets. As a practical consequence, it allows one to carry out most calculations using the words $\psi(t)$ (e.g. as indices) rather than the trees t themselves (which take much more effort to write without mistakes).

Definition 3.7 *Consider a fixed Hall set* $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$. *A word* $h \in Z^+$ *is called a* Hall word *if it is the image* $\psi(t)$ *of a Hall-tree* $t \in \tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$.

In many practical cases one may be quite sloppy, identifying a tree t with the Hall-word $h = \psi(t)$. However, it is very important to understand that one must specify which Hall set one is working with, as over every alphabet Z with at least two letters there exists many Hall sets, compare the next exercise.

Exercise 3.16 *Construct an example of different trees $t_1 \neq t_2$ that belong to different Hall sets $t_i \in \tilde{\mathcal{H}}_i \subseteq \mathcal{M}(Z)$ (over the same alphabet Z), but which have the same* foliage $\psi_1(t_1) = \psi_2(t_2)$ *under the* "forget" *maps $\psi_i : t_i \in \mathcal{H}_i \mapsto Z^+$.*

Directly related to this one-to-one-ness of the restriction of the maps ψ to Hall-Viennot sets, is the fact that every word $w \in Z^+$ has a unique factorization into a structured product, as made precise in the next theorem. This property has been characterized by Viennot as one of the fundamental building blocks of Hall-Viennot sets, and as *the* property that makes Hall-Viennot bases *optimal* [48, 49, 54, 68]. Compare also the Poincaré-Birkhoff-Witt theorem 4.6.

Theorem 3.11 *Suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall Viennot set and $\mathcal{H} = \psi(\tilde{\mathcal{H}}) \subseteq Z^+$ is the corresponding set of Hall words with the induced ordering. Then every word $w \in Z^+$* factors *uniquely into a nonincreasing product of Hall words, i.e. there exist unique $s \geq 0$, $h_j \in \mathcal{H}$, such that*

$$w = h_1 h_2 \dots h_s \quad \textit{and} \quad h_1 \geq h_2 \geq \dots \geq h_s \tag{92}$$

Exercise 3.17 *To get a feeling for this factorization, consider the sets given in figure 5, write down a list of increasingly longer* random *words (e.g. up to length* 20*) and factor according to theorem* (3.11).

Exercise 3.18
Repeat the preceding exercise for the sets constructed in exercises (3.12) *and* (3.13).

An important special case is the case that $w = h \in \mathcal{H}$ is itself a Hall word. Clearly h is its own unique factorization into a nonincreasing product of Hall words. However, if we truncate the last letter of h, e.g. if $h = za$ with $a \in Z$, then we may consider the unique factorization of z into a nonincreasing product $z = h_1 H - 2 \dots h_s$ of Hall words. A little reflection shows that necessarily also $h_s < a$, as illustrated in the figure 6. We will return to this diagram when discussing the structure formula for the dual Poincaré-Birkhoff-Witt bases in section 4.3, compare figure 7.

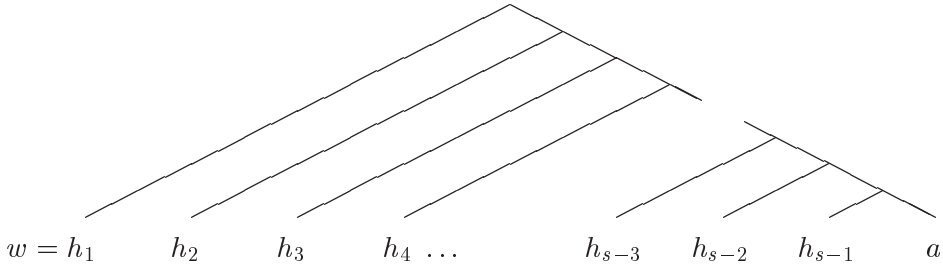


Figure 6. Structure and factorization of Hall trees

This tree very suggestively shows how Hall sets *grow out* of successive Lazard elimination. Note also the close correspondence to theorem 3.6, that while $\varrho_a^\dagger$ and $\lambda_a^\dagger$ are derivations on the shuffle algebra $(\hat{A}(Z), \sqcup\!\sqcup)$, *only* the transpose $\lambda_a^\dagger$ of the left translation λ_a by a letter a is a derivation on the chronological algebra $(C(Z), *)$.

While it is convenient to write a word $w \in Z^+$ where one really means the iterated Lie bracket $\varphi(\psi^{-1}(w)) \in L(Z)$, care needs to be taken not to confuse these two. This becomes even more important as the coordinate $< \varphi(\psi^{-1}(w)), w >$ of w in the Lie polynomial $\varphi(\psi^{-1}(w)) \in L(Z)$ (with respect to the basis Z^* of $A(Z)$) generally is not zero. Indeed, for some Hall bases (especially, Lyndon bases) the word w is always the smallest (largest) word that appears with nonzero coefficient in the Lie polynomial $\varphi(\psi^{-1}(w))$. For further details see [48, 49, 54].

4 A primer on exponential product expansions

4.1 Ree's theorem and exponential Lie series

Many of the constructions and properties of noncommuting polynomials and Lie polynomials from section 3.2 carry directly over to infinite series and infinite Lie series, although some extra caution needs to be taken when working with infinite alphabets. In the following we assume that Z is finite unless otherwise noted. We only summarize a few key notions, and concentrate on what is new and relevant for control. For a detailed description see [54].

Definition 4.1 *A formal series $f \in \hat{A}(Z)$ is a Lie series, written $f \in \hat{L}(Z)$ if for every $n \in \mathbf{Z}^+$ the* homogeneous component *f_n is a Lie polynomial, where*

$$f_n \stackrel{\text{def}}{=} \sum_{|w|=n} < f, w > w \quad \in L(Z) \tag{93}$$

For example, the characterization of Lie polynomials in theorem 3.3 carries directly over to Lie series.

Theorem 4.1 *A formal power series $f \in \hat{A}(Z)$ is a Lie series if and only if*

$$\Delta(f) = f \otimes 1 + 1 \otimes f \tag{94}$$

Exercise 4.1 *Prove theorem* 4.1 *using definition* 4.1 *and theorem* 3.3.

Just as in the case of Lie polynomials, see (82), one immediately obtains that a series $f \in \hat{A}(Z)$ is a Lie series if and only if both $< f, 1 > = 0$ and f is orthogonal to all nontrivial shuffles $< f, u \sqcup\!\sqcup v > = 0$ for all $u, v \in A(Z)$.

Two common ways in which series arise from polynomials (and from other series) are the exponential map and its inverse, both defined on suitable domains.

Definition 4.2
For any power series $s \in \hat{A}(Z)$ with zero constant term define the exponential by

$$e^s = \sum_{k=0}^{\infty} \frac{s^k}{k!} = 1 + s + \frac{s^2}{2} + \frac{s^3}{6} + \frac{s^4}{24} + \ldots \tag{95}$$

Definition 4.3
For any power series $s \in \hat{A}(Z)$ with constant term $< s, 1 > = 1$, define the logarithm by

$$\log s = \sum_{k=1}^{\infty} \frac{(-)^{k+1}}{k}(s-1)^k = (s-1) - \frac{(s-1)^2}{2} + \frac{(s-1)^3}{3} + \frac{(s-1)^4}{4} - \ldots \tag{96}$$

Exercise 4.2
Verify that exp: $s \mapsto e^s$ *and* log, *as defined above, are right, respectively left, inverses of each other. Carefully state the domains on which they are inverses. Check which of the* usual *identities for exponentials and logarithms hold for these maps on $\hat{A}(Z)$ defined as series.*

Theorem 4.2 (Friederich's criterion) *A power series $p \in \hat{A}(Z)$ with constant term $< p, 1 >= 1$, is an exponential Lie series (i.e. $\log p \in \hat{L}(Z)$), written $p \in \hat{G}(Z)$, if and only if*

$$\Delta(p) = p \otimes p. \tag{97}$$

Formally this is the result of a short calculation, shown below in one direction (assuming that $f \in \hat{L}(Z)$), using the continuity of the exponential, of the coproduct, and of the maps $f \mapsto 1 \otimes f$ and $f \mapsto f \otimes 1$. (For careful justifications of all steps see theorem 3.2 in [54].)

$$\Delta(e^f) = e^{\Delta(f)} = e^{f \otimes 1 + 1 \otimes f} = e^{f \otimes 1} e^{1 \otimes f} = \left(e^f \otimes 1\right) \cdot \left(1 \otimes e^f\right) = e^f \otimes e^f \tag{98}$$

Exercise 4.3 *Suppose $s \in \hat{A}(Z)$ is a formal power series with constant term $< s, 1 >= 1$. Verify that $s^{-1} \in \hat{A}(Z)$ and $s^{-1}s = ss^{-1} = 1$ where*

$$s^{-1} \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} (-)^k (s-1)^k = 1 - (s-1) + (s-1)^2 - (s-1)^3 + (s-1)^4 + \dots \tag{99}$$

Use your result to find all series $s \in \hat{A}(Z)$ which are invertible in this sense.

Exercise 4.4 *Show that $\hat{G}(Z)$ is a group under multiplication, i.e. verify that Δ extends* continuously *to an associative algebra homomorphism on $\hat{A}(Z)$, and then use theorem 4.2 to verify that if $p, q \in \hat{G}(Z)$, then also $p \cdot q \in \hat{G}(Z)$ and $p^{-1} \in \hat{G}(Z)$.*

This justifies the name *"group-like elements"* for power series in $\hat{G}(Z) \subseteq \hat{A}(Z)$, i.e. $\hat{G}(Z)$ is a formal Lie group whose algebra is $\hat{L}(Z)$. Translating back to control, the Lie series $\log p \in \hat{L}(Z)$ correspond to (infinite) linear combination of) vector fields, while the exponential $p \in \hat{G}(Z)$ corresponds to the flow of $\log p$ evaluated at time $t = 1$ (or the flow of $\frac{1}{c} \log p$ evaluated at time c). Thus $\hat{G}(Z)$ corresponds to a formal group of diffeomorphisms, and we think of $p \in \hat{G}(Z)$ as a *point*. Alternatively, these points may be characterized as *multiplicative linear functionals* on the associative algebra of *smooth functions on the state space of the system*, see [35] for more details. In the algebraic setting of this section, this corresponds to the following theorem (recall that Υ maps shuffle multiplication to pointwise multiplication of functions, see corollary 3.2):

Proposition 4.3 *The* points $p \in \hat{G}(Z)$ *are* multiplicative linear maps *on the algebra $A(Z, ш)$.*

The proof is a straightforward calculation using the algebraic characterization (81) of the shuffle product.

$$<p, \phi \sqcup\!\sqcup \psi> = <\Delta(p), \phi \otimes \psi> = <p \otimes p, \phi \otimes \psi> = <p, \phi> \cdot <p, \psi> \quad (100)$$

This may be thought of as a multi-dimensional, noncommutative analogue of the identification of the *point* $p \in (\mathbb{R}, +)$ with the translation $\tau_p : x \mapsto (x+p)$ (considered as a diffeomorphism of $\mathbb{R}$), and with the Taylor formula (for more advanced analysis see e.g. [1, 2, 19, 35])

$$\begin{array}{ccc} p & : & f \mapsto f(p). \\ \updownarrow & & \\ \tau_p & : & f(\cdot) \mapsto f(\cdot + p). \\ \updownarrow & & \end{array} \quad (101)$$

$$e^{p\frac{d}{dx}}\Big|_0 = \sum_{k=0}^{\infty} \frac{p^k}{k!} \left(\frac{d}{dx}\right)^n \Big|_0 \quad : \quad f \mapsto \sum_{k=0}^{\infty} \frac{(p-0)^k}{k!} f^{(n)}(0)$$

Continuing with making connections to control, we have the following theorem which follows almost immediately from re-reading (100). (For a short technical proof of the *"if"* direction see theorem 3.2 in [54].)

Theorem 4.4 (Ree's theorem*) A noncommutative power series*

$$p = 1 + \sum_{w \in Z^+} p_w w \in A(Z)$$

(*with constant term* $<p, 1>=0$) *is an exponential Lie series,* (*i.e.* $p \in \hat{G}(Z)$ *or equivalently* $\log p \in \hat{L}(Z)$) *if and only if its* coefficients satisfy the shuffle relations*, i.e. if the map*

$$w \mapsto p_w \stackrel{\text{def}}{=} <w, p> \quad (102)$$

is an (associative algebra) homomorphism from $A(Z, \sqcup\!\sqcup)$ *to* $\mathbb{R}$.

Note that in (102) any $w \in A(Z)$ is allowed, whereas previously p_w was defined only for $w \in Z^*$. In view of theorem 3.1 and corollary 3.2, that the map Υ is both a chronological and an associative algebra homomorphism, this yields right away the following fundamental fact. (We will return to this in the subsequent sections, e.g. in (115)).

Proposition 4.5 *The Chen Fliess series is the image of an exponential Lie series.*

4.2 From infinite series to infinite products

The main item of this section is the Poincaré-Birkhoff-Witt theorem which relates Lie algebras to certain associative algebras, specifically relating their bases. This naturally leads one to consider infinite exponential products, especially for the Chen Fliess series for which the coefficients will be determined in the next section. But we start with a brief review - using the more compact notation and terminology developed in recent chapters. Recall from section 2.1 how the Chen Fliess series as an *infinite series* arose from solving a universal control system by *iteration*. (In contrast, the infinite exponential product in the next section will arise from *variation of parameters*).

Definition 4.4 *For any finite alphabet Z the* universal control system *is the formal bilinear system on $\hat{A}(Z)$*

$$\dot{s} = s \cdot \sum_{a \in Z} u_a a \quad \textit{with initial condition} \quad s(0) = 1. \tag{103}$$

Here $u_a : t \mapsto u_a(t)$ are locally integrable scalar controls and $s : t \mapsto \hat{A}(Z)$ is the solution curve.

Exercise 4.5 *For the case of the three-letter alphabet $Z = \{0, 1, 2\}$ using the* basis Z^*, *write out the first few components of the system* (103), *i.e. $\dot{s}_e = \ldots$, $\dot{s}_a = \ldots$, $\dot{s}_{ab} = \ldots$, etc. (Here e is the empty word, and $a, b \in Z$). Write out the components $s_w(t)$ of the solution curve $s(t)$ using iterated integrals of the controls.*

Using the chronological product $(U * V)(t) = \int_0^t U(\tau) V'(\tau) d\tau$, and writing $U_a(t) = \int_0^t u_a(\tau)\, d\tau$ for the integrals of the controls, the integrated form of the *universal control system* (103)

$$s(t) = 1 + \int_0^t s(\tau) F'(\tau)\, d\tau \quad \text{with} \quad F = \sum_{a \in Z} U_a a, \tag{104}$$

is most compactly written as

$$s = 1 + s * F \tag{105}$$

Iteration yields the explicit series expansion

$$\begin{aligned} s &= 1 + (1 + s * F) * F \\ &= 1 + F + ((1 + s * F) * F) * F \\ &= 1 + F + (F * F) + (((1 + s * F) * F) * F) * F \\ &= 1 + F + (F * F) + ((F * F) * F) + ((((1 + s * F) * F) * F) * F) * F \\ &\vdots \\ &= 1 + F + (F * F) + ((F * F) * F) + (((F * F) * F) * F) \ldots \end{aligned}$$

Using intuitive notation for chronological powers (compare definition 4.8) this solution formula in the form of an infinite series is compactly written as

$$s = \sum_{n=0}^{\infty} F^{*n} \quad = 1 + F + F^{*2} + F^{*3} + F^{*4} + F^{*5} + F^{*6} + \ldots \qquad (106)$$

After this review of how *solving differential equations by iteration* yields infinite series expressions for the solution curves, we develop some abstract background that will lead to effective product expansions of the solution curves.

Every Lie algebra can be *imbedded* into an associative algebra: The universal enveloping algebra $\mathcal{U}$ of a Lie algebra $\mathcal{L}$ (with natural Lie algebra homomorphism $\imath$) is, by definition, the associative algebra (which exists, and is unique up to homomorphism) such that whenever A is an associative algebra and $\Phi: \mathcal{L} \mapsto A$ is a Lie algebra homomorphism, then there exists a map $\Psi: \mathcal{U} \mapsto A$ such that $\Phi = \Psi \circ \imath$. The fundamental theorem (following [54]) is

Theorem 4.6 (Poincaré-Birkhoff-Witt theorem) *Suppose $\mathcal{B} = \{b_\alpha : \alpha \in I\}$ is an ordered basis for a Lie algebra $\mathcal{L}$. Further suppose $\mathcal{U}$ is the universal enveloping algebra of $\mathcal{L}$ with inclusion map $\imath: \mathcal{L} \mapsto \mathcal{U}$. Then a basis for $\mathcal{U}$ is given by the set of decreasing products*

$$\{\imath(b_{\alpha_n})\imath(b_{\alpha_{n-1}}) \ldots \imath(b_{\alpha_2})\imath(b_{\alpha_1}) : \alpha_n \geq \alpha_{n-1} \geq \ldots \alpha_2 \geq \alpha_1, \quad \alpha_i \in I\} \qquad (107)$$

For a proof of the Poincaré-Birkhoff-Witt theorem see e.g. [46] or any textbook on Lie algebras. Note, we earlier used a consequence of this theorem to conclude that the Lie algebra $L(Z)$ of all Lie polynomials over Z is indeed the free Lie algebra over Z.

Of interest to us in the next section is the structure of the Poincaré-Birkhoff-Witt *basis* of the universal enveloping algebra, which in the case of $\mathcal{L} = L(Z)$

being the free Lie algebra agrees with the free associative algebra $A(Z) = \mathcal{U}$ over Z. More specifically, the set Z^* of all words over Z forms one basis of $A(Z)$, while every basis $\mathcal{B}$ of $L(Z)$ (in particular, each Hall-Viennot basis of the previous chapter) gives rise to a different Poincaré-Birkhoff-Witt *basis* of $A(Z)$.

Exercise 4.6 *Fix a Hall set $\tilde{\mathcal{H}}$ for the two element alphabet $Z = \{a, b\}$ (compare figure 5 in section 3.3) For each homogeneous component $A^{(k,\ell)}(Z)$ (i.e. the subspace spanned by all words $w \in Z^*$ with $\|w\|_a = k$ and $\|w\|_b = \ell$) with $k + \ell \leq 4$ write out the induced bases that arise from the Poincaré-Birkhoff-Wittbasis $\mathcal{P}$ built from $\tilde{\mathcal{H}}$, and find the transition matrix for the basis change from the standard basis Z^*.* (This requires the expansion of Lie polynomials. See also the next example.)

For illustration consider the subspace $A^{(1,2)}(\{0,1\})$ whose standard basis (coming from Z^*) is $\mathcal{B}^{(1,2)} = \{110, 101, 011\}$. Considering a Hall set starting with $\tilde{\mathcal{H}} = \{0, 1, (1,0), (1,(1,0)),\ (0,(1,0)), \ldots\}$, the subset of the induced PBW-basis $\mathcal{P}$ for this homogeneous component is (compare with the calculations in section 2.1).

$$\mathcal{P}^{(1,2)} = \{[1,[10]], [1,0]1, 011\} = \{110 - 2 \cdot 010 + 011, 101 - 011, 011\} \quad (108)$$

and the transition matrix between the two bases is

$$\begin{pmatrix} [1,[10]] \\ [1,0]1 \\ 011 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 110 \\ 010 \\ 011 \end{pmatrix} \quad (109)$$

The inverse of this matrix transforms the *dual bases* (see below) according to

$$\begin{pmatrix} \tilde{\xi}_{[1,[10]]} \\ \tilde{\xi}_{[1,0]} \sqcup\!\sqcup \tilde{\xi}_1 \\ \tilde{\xi}_0 \sqcup\!\sqcup \tilde{\xi}_1 \sqcup\!\sqcup \tilde{\xi}_1 \end{pmatrix} = \begin{pmatrix} \tilde{\xi}_{[1,[10]]} \\ \tilde{\xi}_{[1,0]1} \\ \tilde{\xi}_{011} \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 110 \\ 010 \\ 011 \end{pmatrix} \quad (110)$$

To see the importance of the dual bases for control, recall the illustrative manipulations of the Chen Fliess series in section 2.1. Working with the words, or multi-indices, we started with the expression

$$110 \otimes 110 + 101 \otimes 101 + 011 \otimes 011 \quad (111)$$

and using integration by parts and collecting Lie polynomials transformed it into an expression of the form

$$\tilde{\xi}_{[1,[10]]} \otimes [1,[1,0]] + \tilde{\xi}_{[1,0]1} \otimes [1,0]1 + \tilde{\xi}_{011} \otimes 011 \tag{112}$$

One way to think of this is as *resolving* the *identity map* (on a finite dimensional vector space) with respect to two different bases. E.g. suppose $\{e_1, \dots e_n\}$ and $\{v_1, \dots v_n\}$ are bases for V, and $\{e^1, \dots e^n\}$ and $\{v^1, \dots v^n\}$ are the corresponding dual bases for the dual space V^*, then upon the isomorphism $\mathrm{Hom}(V,V) \sim V^* \otimes V$ the identity map $\mathrm{id_v}\colon V \mapsto V$ may be identified with

$$\mathrm{id_v} \sim e^1 \otimes e_1 + e^2 \otimes e_2 + e^3 \otimes e_3 = v^1 \otimes v_1 + v^2 \otimes v_2 + v^3 \otimes v_3 \tag{113}$$

In case of the Chen Fliess series the vector space $V = \hat{A}(Z)$ is infinite dimensional. But due to its graded structure one may consider one homogeneous component at a time. In case of a finite alphabet Z each such component is finite dimensional. The case of $A^{(1,2)}(\{0,1\})$ is typical.
Stepping back, we have ordered bases (Hall-Viennot bases) for $L(Z)$, and they induce PBW-bases on its universal enveloping algebra which is $A(Z)$. Thus it is straightforward to write down the products on the *right hand side*, e.g. $\{[1,[1,0]],[1,0]1,011\}$ (corresponding to the partial differential operators $\{[f_1,[f_1,f_0]],[f_1,f_0]f_1, f_0f_1f_1\}$ in control). The interesting question is about the structure of the terms on the *left hand side*, algebraically the dual Poincaré-Birkhoff-Witt bases (in control the corresponding iterated integral functionals). This section is to give an elegant formula for these dual bases $\tilde{\xi}_p$.

Note the similarity of the *resolution of the identity* (113) with the Chen Fliess series: Indeed, the series is the image of the resolution of the identity map $\mathrm{id}\colon \hat{A}(Z) \mapsto \hat{A}(Z)$ with respect to the basis Z^* under the map

$$\Upsilon \otimes \mathcal{F}\colon A(Z) \otimes \hat{A}(Z) \mapsto \mathcal{II}\mathcal{F}(\mathcal{U}) \otimes \hat{A}(\{f_a\colon a \in Z\})\Big|_0 \tag{114}$$

(to a series of partial differential operators – evaluated at "zero" – with iterated integral functionals as coefficients), i.e.

$$\Upsilon \otimes \mathcal{F}\colon \sum_{w \in Z^*} w \otimes w \mapsto \sum_{w \in Z^*} \Upsilon(w) \otimes \mathcal{F}(w) \tag{115}$$

But as observed above, instead of using the standard basis Z^* for $A(Z)$, one may resolve the identity using any other basis. Of course, most useful will be

the Poincaré-Birkhoff-Witt bases built on Hall bases for the free Lie algebra $L(Z)$. More specifically, suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set. We introduce the following convenient notation:

Notation: If $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$, then write $[\,\cdot\,] = \varphi \circ \psi^{-1} \colon \psi(\tilde{\mathcal{H}}) \subseteq Z^* \mapsto L(Z)$ for the map that sends each Hall word to the corresponding Lie bracket.

The Poincaré-Birkhoff-Witt bases corresponding to the Hall set $\tilde{\mathcal{H}}$ is the set

$$\mathcal{P} = \{[h_n][h_{n-1}] \cdots [h_3][h_2][h_1] \colon h_n \geq h_{n-1} \geq \ldots h_2 \geq h_1,\ h_k \in \mathcal{H},\ n \geq 0\} \tag{116}$$

where the products of $[h_k]$ are taken in $A(Z)$ identified with the universal enveloping algebra $\mathcal{U}$ of $L(Z)$. In agreement with the prior usage in examples and in (110) formally define

Definition 4.5 *For a Poincaré-Birkhoff-Witt basis $\mathcal{P} \subseteq A(Z)$ denote by $\tilde{\xi}_v \in A(Z)$ the* dual basis elements *that are uniquely determined by*

$$< \tilde{\xi}_v, p > = \delta_{v,p} \quad \textit{for all} \quad p \in \mathcal{P} \quad \textit{(Kronecker delta)} \tag{117}$$

In analogy to (113), the preimage of the Chen-Fliess series (from (115)) may thus equally be resolved as

$$\mathrm{id}_{\mathrm{A(Z)}} \sim \sum_{w \in Z^*} w \otimes w = \sum_{v \in \mathcal{P}} \tilde{\xi}_v \otimes v \tag{118}$$

(using that Z^* is self-dual) where $\mathcal{P}$ is any Poincaré-Birkhoff-Witt basis for $A(Z)$.
Earlier manipulations, e.g. (110) and section 2.1 demonstrated that explicit formulas for the dual bases elements $\tilde{\xi}_v$ (for Poincaré-Birkhoff-Witt bases over Hall sets) critically encode the iterated integral functionals in effective solution formulas. Later in this chapter we will see that indeed it suffices to obtain formulas for $\tilde{\xi}_{[h]}$ for Hall words $h \in \mathcal{H}$.

From the previous sections it is known that the Chen Fliess series is an exponential Lie series, i.e. for any basis, especially Hall basis of $L(Z)$ there exist $\zeta_h \in A(Z)$ such that

$$\sum_{w \in Z^*} w \otimes w = e^{\sum_{h \in H} \zeta_h \otimes [h]} \tag{119}$$

Such expression may be considered the formal analogue (preimage under the map $\Upsilon \otimes \mathcal{F}$) of a *continuous* Campbell-Baker-Hausdorff formula. One

can obtain simple formulas for the entire exponent $\log\left(\sum_{w\in Z^*} w\otimes w\right)$ [54], compare also [64] for a cleaned up version – but such infinite linear combinations that do not use a basis obviously do not have uniquely determined coefficients. Using a Hall basis, explicit formulae for the case of a two-letter alphabet (alas a single input system with drift, or two input system without drift, in control) have been calculated for all terms up to fifth order [39]. Most recently [55] has used purely algebraic means to develop a reasonably simple, general formula for the exponent that, while not using a basis, uses a *comparatively small* spanning set for $L(Z)$. From that formula, one can obtain formulae for the ζ_h using a Hall basis, but these are again less attractive.

Indeed, the search for such simple expressions for ζ_h as h ranges over a basis of $L(Z)$ is still subject of ongoing research, with much evidence pointing to the need for a completely different construction of bases for $L(Z)$ (as Hall Viennot bases together with the Lazard elimination process are *inextricably* linked to the exponential product expansions discussed in the sequel.

Exercise 4.7 *Use the definitions* (95) *and* (96) *and the identity* (118) *to calculate explicit formulae for ζ_h for short Hall words from the initial segment $\{0, 1, 10, 110, 001, 1110, 0110,\ 0010, \ldots\}$ of a Hall set.* (This is basically a linear algebra exercise.)

An alternative to writing the series (119) as the exponential of an infinite Lie series, is to write it as an *infinite directed product of exponentials* (where both the directed product and the exponential still need to be defined, see below)

$$\sum_{w\in Z^*} w\otimes w = \overrightarrow{\prod}_{h\in H} e^{\xi_h\otimes[h]} \tag{120}$$

Exercise 4.8 *Referring to the formal definitions of directed products, use the definition* (95) *and structure of the the Poincaré-Birkhoff-Witt basis* (107) *to infer that the coefficients ξ_h in* (120) *indeed agree for $[h]$ in a Hall basis with the definition of the dual Poincaré-Birkhoff-Witt basis $\tilde{\xi}_{[h]}$ in* (4.5).

Notation: Since the key information is contained in the formulas $\tilde{\xi}_{[h]} = \xi_h$ for Hall elements h, and the map from Hall-trees to Hall words is injective, it is convenient to use the (deparenthesized) Hall words $h \in Z^*$ as indices rather than the coresponding Lie polynomials $[h] = \varphi(\psi^{-1}(h)) \in A(Z)$.

It remains to formally define directed products, and to consider the convergence properties of infinite products, compare remark 3.4.

Definition 4.6 *For a sequence* $\{s_k : k \in \mathbf{Z}_0^+\} \subseteq \hat{A}(Z)$ *of formal series inductively define the directed products via*

$$\overrightarrow{\prod}_{\emptyset} s_k = \overleftarrow{\prod}_{\emptyset} s_k = 1 \text{ and}$$

$$\overrightarrow{\prod}_{k=1}^{n+1} s_k = \left(\overrightarrow{\prod}_{k=1}^{n} s_k\right) \cdot s_{n+1} \quad \text{and} \quad \overleftarrow{\prod}_{k=1}^{n+1} s_k = s_{n+1} \cdot \left(\overleftarrow{\prod}_{k=1}^{n} s_k\right) \quad (121)$$

Proposition 4.7 *Suppose* $\{s_k : k \in \mathbf{Z}^+\} \subseteq\in \hat{A}(Z)$ *is a sequence of formal series such that for every* $N < \infty$ *there exists* $k_N < \infty$ *such that* $< s_k, w > = 0$ *for all* $k > k_N$ *and for all words* $w \in Z^*$ *with length* $\|w\| < N$. *Then the infinite directed products* $\overrightarrow{\prod}_{k=1}^{\infty} s_k$ *and* $\overleftarrow{\prod}_{k=1}^{\infty} s_k$ *are well defined.*

Exercise 4.9 *Prove proposition* (4.7) *and using remark* 3.4

Corollary 4.8 *Suppose* $\{f_k : k \in \mathbf{Z}^+\} \subseteq\in \hat{L}(Z)$ *is a sequence of Lie series such that for every* $N < \infty$ *there exists* $k_N < \infty$ *such that* $< f_k, w > = 0$ *for all* $k > k_N$ *and for all words* $w \in Z^*$ *with length* $\|w\| < N$. *Then the infinite directed products* $\overrightarrow{\prod}_{k=1}^{\infty} e^{f_k}$ *and* $\overleftarrow{\prod}_{k=1}^{\infty} e^{f_k}$ *are well defined.*

Exercise 4.10 *Prove corollary* (4.8) *assuming proposition* (4.7) *and using remark 3.4*

4.3 Sussmann's exponential product expansion

This section demonstrates how to write the Chen Fliess series as an infinite exponential product. The approach follows the construction originally given by Sussmann [64] (but utilizing terminology form prior lectures in this series). Alternative constructions have been given using repeated differentiation and analysis of the derivatives [21], and by using entirely combinatorial and algebraic methods [48, 49, 54, 58]. The approach relies on repeatedly employing the method of *variation of parameters* from differential equations to develop a formula for the solution of the *universal control system* (103) The key strategy is to methodically match the recursive design with the Lazard elimination process, compare theorem 3.7. To improve the readability, we concentrate on the differential equations formulation (e.g. work directly with iterated integrals) and only state the analogous combinatorial formulas (in terms of formulas in $\hat{A}(Z) \otimes A(Z)$).

We begin with a review of some technical manipulations that are essential for the variation of parameters approach. The following formulas is one of the most useful and most often used elementary formulas in control. It is instructive to compare the algebraic and geometric/analytic proofs and definitions, and the quite different appearance. The following again just reiterates that in the *analytic setting* many apparently analytic arguments are really purely algebraic.

Proposition 4.9 *If $x \in \hat{L}(Z)$ is a Lie series and $y \in \hat{A}(Z)$ then*

$$e^x y e^{-x} = e^{\mathrm{ad}x} y = \sum_{k=0}^{\infty} \tfrac{1}{k!}(\mathrm{ad}^k x, y) \ = y + [x,y] + \tfrac{1}{2}[x,[x,y]] + \tfrac{1}{6}[x,[x,[x,y]] + \dots \tag{122}$$

We will give a *formal differential equations* argument below. However, it is instructive to write out the first few terms by hand, to get a feeling how the words *combine into Lie polynomials.*

Exercise 4.11 *By formally expanding each exponential into its series (using the definition of the exponential), formally derive at least the first few terms of the formula (which is the form in which the previous formula is used most often in control)*

$$e^f g = (e^f g e^{-f}) e^f \ = (g + [f,g] + \tfrac{1}{2}[f,[f,g]] + \tfrac{1}{6}[f,[f,[f,g]]] + \dots) e^f \tag{123}$$

This simple formula is very useful in control as differentiation of products of flows typically yields expressions like $e^{t_3 f_3} e^{t_2 f_2} f_2 e^{t_1 f_1} p$ (corresponding to a *variation* at $e^{t_1 f_1} p$ *transported* along the flows of first f_2, and then f_3 to the same terminal point $e^{t_3 f_3} e^{t_2 f_2} e^{t_1 f_1} p$ where different such tangent vectors are combined in the usual arguments of approximating cones yielding conditions for optimality and controllability). This means finding a formula for $\tilde{f}$ such that

$$e^{t_3 f_3} e^{t_2 f_2} f_2 e^{t_1 f_1} p = \tilde{f} e^{t_3 f_3} e^{t_2 f_2} e^{t_1 f_1} p \tag{124}$$

Clearly, f_2 commutes with $e^{t_2 f_2}$, but the tangent map of the third flow has an effect which is quantified by the above formula.

Exercise 4.12 (Control application) *Differentiate* (17) *with respect to each of the switching times a, b, c and d, and then use formula* (123) *to move each of the vector fields to the left of all exponentials, i.e. (geometrically) transport each vector back to the same point $x(10, u)$.*

Suppose $z \in \hat{L}(Z)$ is a Lie series with constant term $< z, 1 > = 0$. Consider the *curve* $\gamma\colon \mathbb{R} \mapsto \hat{A}(Z)$ defined by $\gamma\colon t \mapsto e^{tZ}$. Define the *derivative* $\gamma'\colon \mathbb{R} \mapsto \hat{A}(Z)$ of γ at t as

$$\gamma'(t) = \tfrac{d}{dt} e^{tz} \stackrel{\text{def}}{=} e^{tz} z. \tag{125}$$

This allows an elegant formal derivation of (122), connecting algebra and geometry in a bootstrapping argument. Suppose $x \in \hat{L}(Z)$ is a Lie series and $y \in \hat{A}(Z)$. Consider the curve $\gamma\colon \mathbb{R} \mapsto \hat{A}(Z)$ defined by $\gamma\colon t \mapsto e^{tx} y e^{-tx}$ and differentiate.

$$\gamma'(t) = e^{tx} x y e^{-tx} + e^{tx} y e^{-tx}(-x) = e^{tx}(xy - yx)e^{-tx} = e^{tx}(\mathrm{ad}x, y)e^{-tx}. \tag{126}$$

using that $e^{tz} z = z e^{tz}$ for all $z \in \hat{L}(Z)$. Recursively, replacing y in above calculation by $(\mathrm{ad}^k x, y)$ one obtains

$$\left(\tfrac{d}{dt}\right)^k \Big|_{t=0} e^{tx} y e^{-tx} = e^{tx}(\mathrm{ad}^k x, y)e^{-tx}\Big|_{t=0} = (\mathrm{ad}^k x, y) \tag{127}$$

It is helpful to recall that in traditional notation in differential geometry, e.g. Spivak [57], the expression $e^{tf} g e^{-tf}$ is written as $\Phi_{t*} g$ where $(t, q) \mapsto \Phi_t(q)$ denotes the flow of the vector field f. The second exponential corresponds to the ever-present *inverse* in the *push-forward* (of a vector field), or in the *tangent map* (of the diffeomorphism) Φ_t, e.g. written as

$$(\Phi_{t*} g)(p) = \Phi_{t*\Phi_{-t}(p)} \left(g(\Phi_{-t}(p))\right) \tag{128}$$

says that the value of the vector field g pushed forward by the tangent map Φ_{t*} (bundle to bundle) is the same as the value of the vector field g at the *preimage* $\Phi_{-t}(p)$, i.e. the tangent vector $g(\Phi_{-t}(p))$, mapped forward by the tangent map $\Phi_{t*\Phi_{-t}(p)}$ from the fibre $T_{\Phi_{-t}(p)}M$ to the fibre $T_p M$. To complete the side-trip, recall the definition of the Lie derivative of a vector field g in the direction of a vector field f at a point p in terms of the flow Φ of f:

$$(L_f g)(p) = \lim_{t \to 0} \tfrac{1}{t} \Big(g(p) - (\Phi_{t*} g)(p)\Big) \tag{129}$$

which resounds well with the differential equations argument given above (127) (read backwards in the case of $k = 1$), written as $[f, g] = (\mathrm{ad}f, g) = \lim_{t\to 0} \frac{1}{t} \left(e^{tf} g e^{-tf}\right)$.

We are now ready to apply this formula in the variation of parameters approach that leads to Sussmann's exponential product expansion of the

Chen Fliess series. For illustration consider a two input system, i.e. a two-letter alphabet $Z = \{a, b\}$ and the solution curve $y(\cdot)$ taking values in $\hat{G}(\{a,b\}) \subseteq \hat{A}(\{a,b\})$. The *controls* $U'_a, U'_b : [0,t] \mapsto \mathbb{R}$ are assumed to be integrable.

$$y'(t) = y(t) \cdot (U'_a(t) \cdot a + U'_b(t) \cdot b) \tag{130}$$

Make the *Ansatz*

$$y(t) = y_1(t) \cdot e^{U_a(t)a} \text{ for some } y_1(\cdot) \in \hat{A}(Z) \tag{131}$$

Here $e^{U_a(t)a} \in \hat{A}(Z)$ may be thought of as the solution of the initial value problem $y'_1 = za$ with $y_1(0) = 1$ evaluated at *time* $U_a(t) = \int_0^t U'_a(s)ds$. Differentiate (131) and use (130) to obtain a differential equation for $y_1(\cdot)$

$$\Big(y_1(t) \cdot e^{U_a(t)a}\Big) \cdot (U'_a(t) \cdot a + U'_b(t) \cdot b) = y'_1(t) \cdot e^{U_a(t)a} + y_1(t) \cdot e^{U_a(t)a} \cdot (U'_a(t) \cdot a) \tag{132}$$

i.e. after collecting like terms

$$\begin{aligned} y'_1(t) &= y_1(t) \cdot \Big(e^{U_a(t)a} \cdot (U'_a(t) \cdot a + U'_b(t) \cdot b) - e^{U_a(t)a} \cdot (U'_a(t) \cdot a)\Big) \cdot e^{-U_a(t)a} \\ &= y_1(t) \cdot e^{U_a(t)a} \cdot (U'_b(t) \cdot b)) \cdot e^{-U_a(t)a} \\ &= y_1(t) \cdot \Big(\sum_{k=0}^{\infty} \tfrac{1}{k!} U_a^k(t) U'_b(t) \cdot (\mathrm{ad}^k a, b)\Big) \end{aligned} \tag{133}$$

Note that the resulting differential equation for $y_1(t)$ is of the same form as the original one (130) for $y(t)$, albeit now with an infinite linear combination of *control vector fields* $(\mathrm{ad}^k a, b)$. Important is that these are all elements of a Hall-basis for $L(\{a,b\})$, and

$$U'_{a^k b}(t) \stackrel{\mathrm{def}}{=} \tfrac{1}{k!} (U_a(t))^k \cdot U'_b(t) \tag{134}$$

plays a role as a *virtual* control associated to the *vector field* $(\mathrm{ad}^k a, b) \in L(Z)$. Iterating this process, we make the *Ansatz*

$$y_1(t) = y_2(t) \cdot e^{U_b(t)b} \text{ for some } y_2(\cdot) \in \hat{A}(Z) \tag{135}$$

Differentiate (135) and use (133) to obtain a differential equation for $y_2(\cdot)$

$$\Big(y_2(t) \cdot e^{U_b(t)b}\Big) \cdot \sum_{k=0}^{\infty} U'_{a^k b}(t) \cdot (\mathrm{ad}^k a, b) = y'_2(t) \cdot e^{U_b(t)b} + y_2(t) \cdot e^{U_b(t)b} \cdot (U'_b(t) \cdot b) \tag{136}$$

which after collecting like terms becomes

$$\begin{aligned} y_2'(t) &= y_2(t)\cdot\left(e^{U_b(t)b}\cdot(\sum_{k=0}^{\infty}U'_{a^kb}(t)\cdot(\mathrm{ad}^k a,b)) - e^{U_b(t)b}\cdot(U_b'(t)\cdot b)\right)\cdot e^{-U_b(t)b} \\ &= y_2(t)\cdot e^{U_b(t)b}\cdot\left(\sum_{k=1}^{\infty}U'_{a^kb}(t)\cdot(\mathrm{ad}^k a,b))\right)\cdot e^{-U_b(t)b} \\ &= y_2(t)\cdot\left(\sum_{\ell=0}^{\infty}\sum_{k=1}^{\infty}\tfrac{1}{\ell!}\tfrac{1}{k!}U_b^{\ell}(t)U'_{a^kb}(t)\cdot(\mathrm{ad}^{\ell}b,(\mathrm{ad}^k a,b)\right) \end{aligned} \tag{137}$$

The resulting differential equation for $y_2(t)$ is again of the same form, now with a doubly-infinite linear combination of *control vector fields* $(\mathrm{ad}^{\ell}b,(\mathrm{ad}^k a,b))$. Again, these are all elements of a Hall-basis for $L(\{a,b\})$, and

$$U'_{b^{\ell}a^kb}(t) = \tfrac{1}{\ell!}U_b^{\ell}(t)\cdot U_{a^kb}(t) \tag{138}$$

plays the role of a *virtual* control associated to $(\mathrm{ad}^{\ell}b(\mathrm{ad}^k a,b))\in L(Z)$.

This process may be iterated infinitely many times. The critical choice at each step is the selection of the *smallest* Hall element among the *control fields*

$$(\mathrm{ad}^{k_n}h_n,(\mathrm{ad}^{k_{n-1}}h_{n-1},\ldots(\mathrm{ad}^{k_3}h_3,(\mathrm{ad}^{k_2}h_2,h_1))...)) \tag{139}$$

in the previous stage. This will assure that in the next step again all *control vector fields* will be of the same form, again with $h_{j+1}>h_j$ for all j. By virtue of the Lazard elimination process 3.7 they will again be Hall elements. Combining the *Ansätze* of all steps yields

$$y(t)=y_1(t)e^{U_a(t)a}=y_2(t)e^{U_b(t)b}e^{U_a(t)a}=y_3(t)e^{U_{ab}(t)ab}e^{U_b(t)b}e^{U_a(t)a}=\ldots \tag{140}$$

For Hall sets (over finite alphabets) whose ordering is compatible with the lengths of the words (i.e. $\|h\|<\|h'\|$ implies $h\prec h'$, such as in the definition of Philipp Hall bases given in Bourbaki [4]) it is easily seen that this iterative process converges in the topology considered here (compare remark 3.4). However, a little reflection shows that such formula actually also holds for infinite alphabets, and also for Hall sets whose ordering is not necessarily compatible with the length (e.g. recall that Lyndon bases have *infinite segments*).

We refer the reader to the original references [64], also [35], for detailed proofs, and only state the main result as it applies to analytic control systems on finite dimensional manifolds such as those encountered in the examples of the first chapter:

Theorem 4.10 (Sussmann [64]) *Suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set, f_a, $a \in Z$, are analytic vector fields on a manifold M, and u_a, $a \in Z$, are measurable bounded controls defined on $[0, \infty]$. Then for every compact set $K \subseteq M$ there exists $T > 0$ such that for all initial conditions $x_0 \in K$ the infinite directed exponential product*

$$s(t) = \overrightarrow{\prod}_{h \in \mathcal{H}} e^{U_h(t)\, f_{[h]}} \tag{141}$$

converges for $0 \leq t \leq T$ uniformly on K to the solution of the control system

$$\dot{x} = \sum_{a \in Z} u_a f_a(x), \quad x(0) = x_0 \tag{142}$$

Here $f_{[h]} = \mathcal{F}([h])$ is the vector field on M that is obtained from the commutator $[h] \in L(Z)$ by substituting *the control vector fields f_a for the letters $a \in Z$ in $h \in \mathcal{H}$. The iterated integrals U_h for Hall words $h \in \mathcal{H} \setminus Z$ satisfy*

$$U_h(t) = \tfrac{1}{k!} \cdot \int_0^t U_{h_1}^k(\tau) U'_{h_2}(\tau)\, d\tau \tag{143}$$

if $h = h_1^k h_2 \in \mathcal{H}$, $h_1 > h_2$ and either $h_2 \in Z$ is a letter, or the left factor *of h_2 is strictly smaller than h_1.*

Exercise 4.13 *Consider the nilpotent control system $\dot{x_1} = u_1$, $\dot{x_2} = u_2$, and $\dot{x_3} = x_2 u_1$ on $\mathbb{R}^3$. Identify the system vector fields f_1 and f_2, and explicitly write out the images $f_{[h]}$ of a suitable Hall-basis for $L(\{1, 2\})$, the associated iterated integrals $U_h(t)$ and the* flows *in the product* (141). *Verify that the solutions $x(t, u)$ agree with the products of the flows as in the theorem.*

Exercise 4.14 *Repeat the previous exercise for the nilpotent control system $\dot{x_1} = u$, $\dot{x_2} = x_1^p$ on $\mathbb{R}^2$, for $p = 2, 3$ or any other integer.*

Exercise 4.15 *Returning to the previous exercise in the case of $p = 2$ investigate how the choice of $0 < 1$ or $1 < 0$ in the Hall set over $Z = \{0, 1\}$ affects the structure of the iterated integrals U_h.*

The essence in the combinatorial analogue of the theorem 4.10 is captured in the formula for the elements ξ_v of the dual basis for $A(Z)$ indexed by elements $v \in \mathcal{P}$ of a Poincaré-Birkhoff-Witt basis $\mathcal{P}$ for $A(Z)$ that is built on a Hall set, or Hall basis for $L(Z)$. We first give a technical definition.

Definition 4.7 *Suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set. Define the function $\mu: Z^* \mapsto \{\frac{1}{n}: n \in \mathbf{Z}^+\}$ by*

$$\mu(wa) = \mu(w) = \frac{1}{m_1! m_2! \cdot m_{s-1}! m_s!} \tag{144}$$

if $a \in Z$ and $h = wa \in \mathcal{H}$ factors uniquely into $h = h_1^{m_1} h_2^{m_2} \cdots h_{s-1}^{m_{n-1}} h_n^{m_n} a$ with $a \in Z$, $m_i \in \mathbf{Z}^+$, $h_i \in \mathcal{H}$ and $h_1 > h_2 > \ldots > h_{s-1} > h_s < a$ (compare theorem (3.11)).

Theorem 4.11 *Suppose $\hat{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set and the Hall word $h \in \mathcal{H} \subseteq Z^*$ factors uniquely into Hall words $h = h_1 h_2 \cdots h_{n-1} h_n a$ with $a \in Z$ and $h_i \in \mathcal{H}$ and $h_1 \geq h_2 \geq \ldots \geq h_n < a$. Then*

(145)

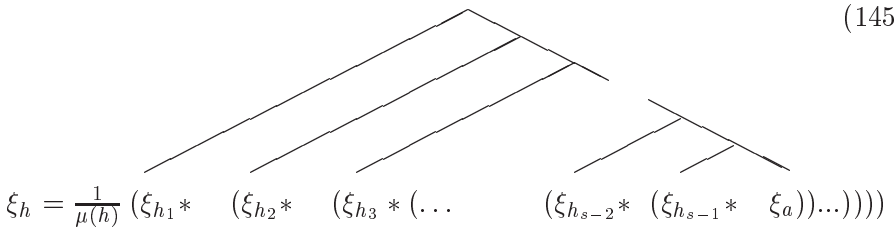


Figure 7. Structure of the dual Poincaré-Birkhoff-Wittbases for Hall trees

Compare the unique factorization theorem 3.11 for Hall Viennot words and the similar figure 6. Note also the close correspondence to theorem 3.6, that while $\varrho_a^\dagger$ and $\lambda_a^\dagger$ are derivations on the shuffle algebra $(\hat{A}(Z), ш)$, *only* the transpose $\lambda_a^\dagger$ of the left translation λ_a by a letter a is a derivation on the chronological algebra $(C(Z), *)$.

One way to establish theorem 4.11 as a consequence of theorem 4.10 is to use that the map Υ is a chronological algebra isomorphism from the free chronological algebra $C(Z)$ onto the space of $\mathcal{IIF}(\mathcal{U})$ of iterated integral functionals – provided the space $\mathcal{U}$ of admissible controls is sufficiently large (compare [35]). An alternative proof of a purely combinatorial nature was given by Melancon and Reutenauer [48, 49]. Essentially the same formula may also be found in Schützenberger [58] and Grayson and Grossman [21].

Introducing left and right chronological powers not only facilitates the writing, but in some cases it may *make factorials disappear*. More specifically, rewriting formulas with another product may result in the disappearance of

factorials, e.g. via the Taylor expansions of $\frac{1}{1-x}$ with one product, becomes the Taylor expansion e^x with respect to another product.

Definition 4.8 *For $w \in A^+(Z) = A(Z^+)$ define $w^{*1} = \lambda^1(w) = w^{\sqcup\!\sqcup 1} = w$, and inductively for $n \geq 1$* (sometimes it is convenient also allow $w^{*0} = 1 = w^{\sqcup\!\sqcup 0}$ and $\lambda^0(w) = 0$)

$$\begin{array}{rcl} \lambda^{n+1}(w) & = & w * \lambda^n(w) \\ w^{*(n+1)} & = & w^{*n} * w \\ w^{\sqcup\!\sqcup (n+1)} & = & w \sqcup\!\sqcup w^{\sqcup\!\sqcup n} = w^{\sqcup\!\sqcup n} \sqcup\!\sqcup w \end{array}$$

Proposition 4.12 *For $w \in A^+(Z) = A(Z^+)$ and $n \in \mathbf{Z}^+$ the following identities hold:*

$$\begin{array}{rcll} w * w^{*(n-1)} & = & (n-1) \cdot w^{*n} & \\ \lambda^n(w) & = & (n-1)! \cdot w^{*n} & \\ w^{\sqcup\!\sqcup n} & = & n! \cdot w^{*n} & (= n\lambda^n(w)) \end{array} \tag{146}$$

Exercise 4.16 *Prove the identities in proposition* 4.12. (Take advantage of bilinearity and first prove the identities, by induction on n, for words $w \in Z^+$.)

For the sake of completeness we also note the structure of the complete dual Poincaré-Birkhoff-Witt bases built on Hall sets. A complete combinatorial proof is given in [48], also see [54]. Alternatively, use the chronological algebra isomorphism Υ to obtain this result from theorem 4.10.

Proposition 4.13 *Suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set and $\mathcal{P}$ the associated Poincaré-Birkhoff-Witt basis for $A(Z)$. If $w = [h_1]^{m_1}[h_2]^{m_2} \ldots [h_n]^{m_n} \in \mathcal{P}$ for Hall words $h_i \in \mathcal{H}$ with $h_i > h_{i+1}$, then the dual basis elements are*

$$\tilde{\xi}_v = \frac{1}{m_1! m_2! \cdots m_n!} \cdot \xi_{h_n}^{\sqcup\!\sqcup m_n} \sqcup\!\sqcup \xi_{h_{n-1}}^{\sqcup\!\sqcup m_{n-1}} \sqcup\!\sqcup \ldots \sqcup\!\sqcup \xi_{h_2}^{\sqcup\!\sqcup m_2} \sqcup\!\sqcup \xi_{h_1}^{\sqcup\!\sqcup m_1} \tag{147}$$

We conclude the section with a series of challenge exercises that aim at bridging the combinatorial and analytical / differential equations arguments that culminate in theorems 4.10 and 4.11.

Exercise 4.17
Verify that the product $*: (A(Z) \otimes \hat{A}(Z)) \times (A(Z) \otimes \hat{A}(Z)) \mapsto (A(Z) \otimes \hat{A}(Z))$ defined by

$$(w \otimes f) * (z \otimes g) = (w * z) \otimes (fg) \tag{148}$$

is a chronological product, i.e. it satisfies the right chronological identity (67).

Exercise 4.18 *Rewrite the universal control system* (103) *as an equation on* $A(Z) \otimes \hat{A}(Z)$ *using the chronological product from the previous exercise.*

Exercise 4.19 *Capture the combinatorial essence of the* variation of parameters *technique using chronological products – first rewrite the differential equation, e.g.* (130) *as an equivalent integral equation, and then use chronological products.*

4.4 Free nilpotent systems

A simple way to state and remember, and a very useful application for the formula (145) in theorem 4.11 is as a normal form for free nilpotent systems. Recall that nilpotent control systems are systems of the form (8) for which the Lie algebra $L(f_0, f_1, \dots f_m)$ generated by the system vector fields is nilpotent. Via a local coordinate change they can always be brought into a form in which the vector fields are polynomial and have a *cascade* structure. Such systems are sufficiently rich that they have good approximation properties (controllability, stabilizability etc.), and they are very manageable: E.g. solution curves can be computed by simple quadratures, requiring no intractable solution of nonlinear differential equations.
A natural objective is to write down a canonical form for the most general such system (up to a certain order). However, any naive try such as the one starting with

$$\left\{\begin{array}{llllll} \dot{x}_1 = u & \dot{x}_5 = x_4 & \dot{x}_9 = x_1^3 \\ \dot{x}_2 = x_1 & \dot{x}_6 = x_5 & \dot{x}_{10} = x_9 \\ \dot{x}_3 = x_2 & \dot{x}_7 = x_2^2 & \dot{x}_{11} = x_1^2 x_2 \\ \dot{x}_4 = x_1^2 & \dot{x}_8 = x_1 x_3 & \dot{x}_{12} = x_1^4 \end{array}\right. \tag{149}$$

does not do the job as the system is not accessible due to redundant terms: Along every solution curve the function $\Phi(x) = x_7 + x_8 - x_2 x_3$ is constant.

Exercise 4.20 *Verify by direct calculation of the iterated Lie brackets of the system vector fields that the system* (149) *does not satisfy the Lie algebra rank condition for accessibility.*

To make precise what we mean by a (maximally) free nilpotent system, define:

Definition 4.9 *For any integer $r > 0$ define $L^{(r)}(Z)$ to be quotient of $L(Z)$ by the ideal $L(Z) \cap \cup_{k=r+1}^{\infty} A^{(k)}(Z)$.*

Definition 4.10 *Suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set and $H = \psi(\mathcal{H}) \subseteq Z^*$ is the set of corresponding Hall words. Define $\mathcal{H}^{(r)} \stackrel{\text{def}}{=} \{h \in \mathcal{H} : |h| \leq r\}$ to be the subset of Hall words of length at most r.*

Definition 4.11 *Suppose $\tilde{\mathcal{H}} \subseteq \mathcal{M}(Z)$ is a Hall set and $r > 0$. The* normal form of the free nilpotent system determined by $\mathcal{H}^{(r)}$ *is the control system*

$$\begin{array}{rcll} \dot{x}_a & = & u_a & \text{if } a \in Z \\ x_{wz} & = & x_w * x_z & \text{if } w,\, z,\, wz \in \mathcal{H}^{(r)} \subseteq Z^* \end{array}$$

Theorem 4.14 *The Lie algebra $L(\{f_a : a \in Z\})$ generated by the vector fields of the system* (4.11) *(written in the form $\dot{x} = \sum_{a \in Z} u_a f_a(x)$) is free nilpotent of step r.*

Example: A normal from for a free nilpotent system (of rank $r = 5$) using a typical Hall set on the alphabet $Z = \{0, 1\}$ is

$$\begin{array}{rcl|l|rcl} \dot{x}_0 & = & u_0 & & & & \\ \dot{x}_1 & = & u_1 & & & & \\ \dot{x}_{01} & = & x_0 \cdot \dot{x}_1 \;=\; x_0\, u_1 & & & & \\ \dot{x}_{001} & = & x_0 \cdot \dot{x}_{01} \;=\; x_0^2\, u_1 & \text{from} & \psi^{-1}(001) & = & (0(01)) \\ \dot{x}_{101} & = & x_1 \cdot \dot{x}_{01} \;=\; x_1 x_0\, u_1 & \text{from} & \psi^{-1}(101) & = & (1(01)) \\ \dot{x}_{0001} & = & x_0 \cdot \dot{x}_{001} \;=\; x_0^3\, u_1 & \text{from} & \psi^{-1}(0001) & = & (0(0(01))) \\ \dot{x}_{1001} & = & x_1 \cdot \dot{x}_{001} \;=\; x_1 x_0^2\, u_1 & \text{from} & \psi^{-1}(1001) & = & (1(0(01))) \\ \dot{x}_{1101} & = & x_1 \cdot \dot{x}_{101} \;=\; x_1^2 x_0\, u_1 & \text{from} & \psi^{-1}(1101) & = & (1(1(01))) \\ \dot{x}_{00001} & = & x_0 \cdot \dot{x}_{001} \;=\; x_0^4\, u_1 & \text{from} & \psi^{-1}(00001) & = & (0(0(0(01)))) \\ \dot{x}_{10001} & = & x_1 \cdot \dot{x}_{0001} \;=\; x_1 x_0^3\, u_1 & \text{from} & \psi^{-1}(10001) & = & (1(0(0(01)))) \\ \dot{x}_{11001} & = & x_1 \cdot \dot{x}_{1001} \;=\; x_1^2 x_0^2\, u_1 & \text{from} & \psi^{-1}(11001) & = & (1(1(0(01)))) \\ \dot{x}_{01001} & = & x_{01} \cdot \dot{x}_{001} \;=\; x_{01} x_0^3\, u_1 & \text{from} & \psi^{-1}(01001) & = & ((01)(0(01))) \\ \dot{x}_{01101} & = & x_{01} \cdot \dot{x}_{101} \;=\; x_{01} x_1^2 x_0\, u_1 & \text{from} & \psi^{-1}(01101) & = & ((01)(1(01))) \end{array} \tag{150}$$

Remark 4.15 *It is noteworthy, and almost essential for effective calculations that the coordinates x_h are indexed by Hall words, rather than by consecutive natural numbers!*
On the other hand, by virtue of the unique factorization theorem 3.11*, one may use the Hall words as indices, and* does not *to use trees or* parenthesized *words (which would make for very cumbersome notation).*

Exercise 4.21 *Write the system* (150) *in the form* $\dot{x} = u_0 f_0(x) + u_1 f_1(x)$, *and calculate iterated Lie brackets of* f_0 *and* f_1 *of length at most* 5 *(using the same Hall set). Verify that for each such iterated Lie bracket* $f_w = \mathcal{F}(w)$, *its value* $f_w(0)$ *at* $x = 0$ *is a multiple of the corresponding coordinate direction* $\frac{\partial}{\partial x_w}\Big|_0$

Exercise 4.22 *Use a different Hall-Viennot basis for* $Z = \{0, 1\}$ *(e.g. the Lyndon basis from figure* 5*), to construct a different representation of the free nilpotent system* (150).
Demonstrate that these systems are equivalent under a global, polynomial coordinate change.

Exercise 4.23 *Use a Hall-Viennot basis for* $Z = \{0, 1, 2\}$ *to construct an explicit coordinate representation similar to* (150) *for a free nilpotent (of order* 4*) two-input system with drift* $\dot{x} = f_0(x) + u_1 f_1(x) + u_2 f_2(x)$.

Exercise 4.24 *Prove that the Lie algebra* $L(\{f_a : a \in Z\})$ *of the vector fields of the system* (4.11) *is nilpotent.* (Introduce a suitable family of dilations so that all vector fields are homogeneous of strictly negative order.)

Exercise 4.25 *Prove that the Lie algebra* $L(\{f_a : a \in Z\})$ *of the vector fields of the system* (4.11) *is isomorphic to* $L^{(r)}(Z)$. (Demonstrate that the Lie algebra has maximal dimension by showing that $\mathcal{F}([h])(0) = c_h \frac{\partial}{\partial x_h}(0)$ for some $c_h \neq 0$.)

On the side we mention another useful number, the dimensions of the *homogeneous* components $L^{(\alpha)}(Z) = L(Z) \cap A^{(\alpha)}(Z)$ of the free Lie algebra $L(Z)$. Here $A^{(\alpha)}(Z)$ is the linear span of all words containing α_a times the letter a (for each $a \in Z$). First recall the Moebius function from enumerative combinatorics [4]:

Definition 4.12 *The Moebius function* $\mathrm{Moe}: \mathbf{Z}^+ \mapsto \{-1, 0, 1\}$ *is defined by* $\mathrm{Moe}(n) = 0$ *if* n *is divisible by the square of a prime, and else* $\mathrm{Moe}(n) = (-)^s$ *if* s *is the number of distinct prime factors of* n.

The Moebius function can also be characterized as the unique function from $\mathbf{Z}^+$ to $\{-1, 0, 1\}$ such that $\mathrm{Moe}(1) = 1$ and

$$\sum_{d|n} \mathrm{Moe}(d) = 0 \quad \text{for all } n \in \mathbf{Z}^+ \tag{151}$$

Proposition 4.16
Suppose $\alpha_a \geq 0$ for $a \in Z$ are nonnegative integers. Then the dimension of $L^{(\alpha)}(Z)$ is

$$\dim L^{(\alpha)}(Z) = \frac{1}{|\alpha|} \sum_{d|\alpha} \mathrm{Moe}(d) \frac{(|\alpha|/d)!}{(\alpha/d)!} = \frac{1}{\sum_{a\in Z} \alpha_a} \sum_{d|\alpha} \mathrm{Moe}(d) \frac{(\sum_{a\in Z} \alpha_a)/d)!}{\prod_{a\in Z} (\alpha_a/d)!} \tag{152}$$

Exercise 4.26 *Calculate, and tabulate, the dimensions of the homogeneous components $L^{(\alpha)}(\{a, b\})$ for $|\alpha| \leq 6$.*

Exercise 4.27 *Calculate, and tabulate, the dimensions of the homogeneous components $L^{(\alpha)}(\{a, b, c\})$ for $|\alpha| \leq 4$.*

We conclude with a few comments about the path planning problem, which given a system of form (8) and two points $p, q \in \mathbb{R}^n$ in the state space, asks for a control u, defined on some time interval $[0, T]$ which steers the system from $x(0) = p$ to $x(T, u) = q$. For a detailed discussion and a variety of results see e.g. [24, 25, 41, 51, 52, 66].
A reasonably tractable class consists of nilpotent systems (possibly used as approximating systems, yielding approximate path planning results). One of the most useful features of free nilpotent systems is that they can provide *universal* solutions to the path planning problems, as any specific nilpotent system *lifts* to a free system. In other words, the trajectories of the free system map, or project to the trajectories of the specific system. Thus the general solution of the problem for the free system yields also a (many) solutions(s) for the specific problem.
More specifically, suppose $\Sigma\colon \dot{x} = \sum_{i=1}^m u_i f_i(x)$ is a specific system such that $L(f_1, \ldots f_m)$ is nilpotent of order r. Then let $\Sigma^{(r)}\colon \dot{x} = \sum_{i=1}^m u_i F_i(x)$ be a free nilpotent system (of order r) on $\mathbb{R}^N$. Due to the *freeness* there exists a smooth map $\Phi\colon \mathbb{R}^N \mapsto \mathbb{R}^n$ that maps trajectories of $\Sigma^{(r)}$ to trajectories of Σ corresponding to the same controls. Thus in order to steer the system Σ from $p \in \mathbb{R}^n$ to $q \in \mathbb{R}^n$ one may use any control of the presumedly solved path planning problem steering $\Sigma^{(r)}$ from any $P \in \Phi^{-1}(p) \subseteq \mathbb{R}^N$ to any $Q \in \Phi^{-1}(q) \subseteq \mathbb{R}^N$.

This short discussion justifies that one take a closer look at the general path planning problem for free nilpotent systems. For systems with nonzero drift the possible lack of controllability remains a formidable obstacle to a

general solution. Thus here we only take a brief look at systems without drift (for which accessibility is the same as controllability). For practical solutions, a key step is to reduce the problem from the very large space $\mathcal{U}$ of all possible controls to smaller sets, typically finite dimensional subspaces. Typical examples include those spanned by trigonometric polynomials (with fixed base frequency and maximal order), polynomial controls, and piecewise constant or piecewise polynomial controls.

For illustration, in the following exercises, calculate the iterated integrals $U_h(T)$ for $h \in \mathcal{H}^{(r)}(\{0,1\})$ with $r = 3, 4$ or 5 (depending on available computer algebra system resources) for the specified parameterized families of controls u_α. Note that this is tantamount to calculating the trajectories of the system (150) for $p = 0$.
Then analyze the nature of the inverse problem of finding the control parameterized by α that yields a given target point $Q = (U_h(T))_{h\in\mathcal{H}^{(r)}(\{0,1\})} = (U_0(T), U_1(T), U_{10}(T), U_{110}(T), \ldots)$.

Exercise 4.28 *Consider polynomial controls (e.g.* $(u_0, u_1)(t) = (\alpha_1 + \alpha_2 t + \alpha_3 t^2, \alpha_4 + \alpha_5 t + \alpha_6 t^2)$*. – Use as many parameters as the dimension of the subsystem you are working with. Is the map from* α *to the endpoint (or sequence of iterated integrals) injective? invertible? What problems do you see as the dimension increases?*

Exercise 4.29 *Consider controls that are trigonometric polynomials (e.g.* $(u_0, u_1)(t) = (\alpha_1 \cos t + \alpha_2 \cos 2t + \alpha_3 \cos 3t, \alpha_4 \sin t + \alpha_5 \sin 2t + \alpha_6 \sin 3t)$*. Use as many parameters as the dimension of the subsystem you are working with. Is the map from* α *to the endpoint (or sequence of iterated integrals) injective? invertible? Contrast this map with the analogous map for linear systems! What problems do you see as the dimension increases?*

These exercises open many new question in an area that still leaves a lot to be explored. One suggestion for exploration is, instead of considering the full free nilpotent system, to restrict one's attention to some special class of system – e.g. one class of systems that has been popular in the 1990s is that of systems in *chain-form*, compare e.g. [67].

Acknowledgments

Supported in part by NSF-grant DMS 00-72369. The author greatly appreciates and the support and hospitality of the Abdus Salam Institute of

Theoretical Physics and of SISSA, who provided this fantastic opportunity to interact with this new generation of bright scientists, and who provided the locality for advancing my own research through continuous daily interactions with the best of my peers from around the world.
My special thanks go to the organizers of this summer school, A. Agrachev, B. Jakubczyk, and C. Lobry. I am most thankful for the continuous encouragement and support by all participants, especially the unusually diverse group of student-participants without whom these lectures would not have happened, without whom these notes would not exist, and who greatly contributed through excellent observations, conjectures and questioning.
With such a superb new generation of geometric control scientists becoming ready to lead the world, I am excited that this wonderful subject will not only stay alive, but prosper in the future, and contribute to making this an even better world.

References

[1] A. Agrachev and R. Gamkrelidze, *Exponential representation of flows and chronological calculus*, Math. USSR Sbornik (Russian) **107**, N4 (1978), 487–532. Math. USSR Sbornik, **35** (1978) 727–

[2] A. Agrachev and R. Gamkrelidze, *Chronological algebras and nonstationary vector fields*, Journal Soviet Math. **17**, no.1 (1979), 1650–1675.

[3] A. Bloch and H. McClamroch, "Controllability and stabilizability properties of nonholonomic control systems", Proc. of 29$^{\text{th}}$ IEEE CDC (1990) 1312–1317.

[4] N. Bourbaki, Lie Groups and Lie algebras, (1989) (Springer).

[5] R. Brockett, *Differential geometric methods in system theory*, in Proc. IEEE Conf.Dec.Cntrl. (1971) 176-180.

[6] K. T. Chen, *Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula*, Annals of Mathematics **65** (1957) 163–178.

[7] J.-M. Coron, *On the stabilization in finite time of locally controllable systems by means of continuous time-varying feedback law*, SIAM J. Cntrl. & Opt., **33** (1995) 804-833.

[8] P. Crouch, *Solvable Approximations to Control Systems*, SIAM J. Cntrl. & Opt., **22** (1984) 40–45.

[9] P. Crouch and F. Lamnabhi-Lagarrigue, *Algebraic and multiple integral identities*, Acta Appl. Math. **15** (1989) 235–274.

[10] P. Crouch and R. Grossman, *The explicit computation of integration algorithms and first integrals for ordinary differential equations with polynomial coefficients using trees*, Proc. Int. Symposium on Symbolic and Algebraic Computation, ACM Press (1992) 89-94.

[11] P. Crouch and F. Leite, *the dynamical interpolation problem: on Riemannian manifolds, Lie groups and symmetric spaces*, J. Dynam.Control Systems **1** (1995) 177 - 202.

[12] A. Dzhumadil'daev, *Non-associative algebras without unit*, Comm. Algebra (2002) (to appear).

[13] A. Dzhumadil'daev, *Trees, free right-symmetric algebras, free Novikov algebras and identities*, Homotopy, Homology and Applications (2001) (to appear).

[14] M. Fliess, *Fonctionelles causales nonlinéaires et indeterminées noncommutatives*, Bull. Soc. Math. France **109** (1981) 3–40.

[15] K. Grasse, *On the relation between small-time local controllability and normal self-reachability* Math. Control Signals and Systems, **5** (1992) 41 – 66.

[16] H. Frankowska, *An open mapping principle for set-valued maps*, J.Math.Anal.Appl. **127** (1987) 172–180.

[17] H. Frankowska, *Local controllability of control systems with feedback*, J.Opt.Theory Applics. **60** (1989) 277–296.

[18] H. Frankowska, *A conical open mapping principle for set-valued maps*, Bull.Australian Math.Soc **45** (1992) 53–60.

[19] I. Gelfand, D. Raikov, and G. Shilov, *Commutative normed rings*, **1991** (1964 Chelsea).

[20] V. Ginzburg and M. Kapranov, *Koszul duality for operads*, Duke Math.J. **76** (1994) 203–272.

[21] M. Grayson and R. Grossman, *Models for free nilpotent algebras*, J. of Algebra, **135** (1990) 177–191.

[22] H. Hermes, *Controlled Stability*, Annali di Matematica pura ed applicata IV, **CXIV** (1977) 103-119.

[23] H. Hermes, *Nilpotent and high-order approximations of vector field systems*, SIAM Review **33** (1991) 238–264.

[24] G. Jacob, *Lyndon discretization and exact motion planning*, Proc. Europ. Control Conf., Grenoble (1991), pp. 1507–1512.

[25] G. Jacob, *Motion planning by piecewise constant or polynomial inputs*, Proc. NOLCOS, Bordeaux (1992).

[26] V. Jurdjevic and H. Sussmann, *Controllability of nonlinear systems*, J. Diff. Eqns. **12** (1972) 95–116.

[27] V. Jurdjevic and H. Sussmann, *Control systems on Lie groups*, J. Diff. Eqns. **12** (1972) 313–329.

[28] M. Kawski, *A new necessary condition for local controllability*, AMS Contemporary Mathematics **68** (1987) 143–156.

[29] *An angular open mapping theorem*, in: Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lect. Notes in Control and Information Sciences **111** (1988) 361-371.

[30] M. Kawski, *Control variations with an increasing number of switchings*, Bull.Amer.Math.Soc. **18** (1988) 149-152.

[31] M. Kawski, *Nilpotent Lie algebras of vector fields*, J. Reine & Angewandte Mathematik, **188** (1988) pp. 1-17.

[32] M. Kawski, M. Kawski, *High-order small-time local controllability*, in: Nonlinear Controllability and Optimal Control, H. Sussmann, ed. (1990) 441-477 (Dekker)

[33] M. Kawski, *Geometric homogeneity and applications to stabilization* in: Nonlinear Control Systems Design, A.Krener and D.Mayne, eds., (1995), pp.147-152.

[34] M. Kawski, *Chronological algebras and nonlinear control*, Proc. Asian Conf. Control, Tokyo, 1994.

[35] M. Kawski and H. J. Sussmann *Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory*, in: Operators, Systems, and Linear Algebra, U. Helmke, D. Prätzel-Wolters and E. Zerz, eds., Teubner (1997), 111–128.

[36] M. Kawski, *Controllability via chronological calculus*, Proc. 38th IEEE Conf.Dec.Contr. (1999) 2920-2925.

[37] M. Kawski, *Calculating the logarithm of the Chen Fliess series*, submitted to Int. J.Cntrol (special issue Proc. MTNS 2000 Perpignan).

[38] M. Kawski, *Chronological algebras: combinatorics and control*, Itogi Nauki i Techniki, vol.68 (2000) pp.144-178. (translation in J. Math. Sciences).

[39] M. Kawski, *Controllability and coordinates of the first kind*, in: Contemporary Trends in Non-linear Geometric Control Theory and its Applications, eds. A. Anzaldo-Meneses, B. Bonnard, J.P. Gauthier, F. Monroy-Perez. (World Sci. Publ.) (to appear)

[40] H. Knobloch, *Higher Order Necessary Conditions in Optimal Control Theory*, Lect. Notes in Control and Information Sciences **34** (1981) (Springer)

[41] G. Laffarière and H. J. Sussmann, *Motion planning for controllable systems without drift*, IEEE Conf. Robotics and Automation, (1991), pp. 1148–1153.

[42] J.-L. Loday, *Une version non commutative des algèbres de Lie: les algèbres de Leibniz*, L'Enseignement Mathématique **39** (1993), pp. 269–293.

[43] J.-L. Loday and T. Pirashvili, *Universal enveloping algebras of Leibniz algebras and (co)homology*, Math. Annalen **196** (1993), pp. 139–158.

[44] J.-L. Loday and T. Pirashvili, *Leibniz representations of Lie algebras*, J. Algebra **181** (1996), pp. 414–425.

[45] J.-L. Loday, *Cup-product for Leibniz cohomology and dual Leibniz algebras*, Math. Scand. 77 (1995), no. 2, pp. 189–196.

[46] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading, Mass., (1983).

[47] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Copmm. Pure Appl. Math. **VII** (1954), 649–673.

[48] G. Melançon and C. Reutenauer, *Lyndon words, free algebras and shuffles*, Canadian J. Math. **XLI** (1989), 577–591.

[49] G. Melançon and C. Reutenauer C., *Combinatorics of Hall trees and Hall words*, J. Comb. Th. Ser. A **59** (1992), 285–299.

[50] H. Munthe-Kaas and B. Owren, *Computations in a Free Lie Algebra*, report no. 148, Dept. Informatics, Univ. Bergen, 1998.

[51] R. Murray and S. Sastry, *Nonholonomic path planning: steering with sinusoids*, IEEE T. Automatic Control, **38** (1993) 700–716.

[52] R. Murray, *Nilpotent bases for a class of nonintegrable distribution swith applications to trajectory generation for nonholonomic systems*, Math. Controls, Signals, & Systems, **7** (1994) 58–75.

[53] R. Ree, *Lie elements and an algebra associated with shuffles*, Annals of Mathematics **68** (1958) 210–220.

[54] C. Reutenauer, *Free Lie algebras*, (1993), Oxford (Clarendon Press).

[55] E. Rocha, *On computation of the logarithm of the Chen-Fliess series for nonlinear systems*, Proc. NCN Sheffield (2001).

[56] L. Rosier, *Homogeneous Lyapunov functions for homogeneous continuous vector fields*, Systems and Control Letters, **19** (1992) 467 – 473.

[57] M. Spivak, A comprehensive introduction to differential geometry (5 volumes). Publish or Perish.

[58] M. Schützenberger, *Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un probléme de mathématiques appliquées*, Séminaire P. Dubreil, Algèbres et Théorie des Nombres, Faculté des Sciences de Paris (1958/59).

[59] G. Stefani, *On the local controllability of a scalar-input system*, in "Theory and Applications of Nonlinear Control Systems," C. I. Byrnes and A. Lindquist eds., Elsevier, North-Holland (1986) 167–179.

[60] G. Stefani, *Polynomial approximations to control systems and local controllability* Proc. 25^{th} IEEE CDC (1985) 33–38

[61] H. J. Sussmann, *An extension of a theorem of Nagano on transitive Lie algebras*, Proc. Amer. Math. Soc., **45** (1974) 349–356.

[62] H. J. Sussmann, *Lie brackets and local controllability: a sufficient condition for scalar-input systems*, SIAM J. Cntrl. & Opt., **21** (1983) 686–713.

[63] H. J. Sussmann, *Lie brackets, Real Analyticity, and Geometric Control*, in Differential Geometric Control, (R. W. Brockett, R. S. Millman, H. J. Sussmann, eds.) (1983) 1–116.

[64] H. J. Sussmann, *A product expansion of the Chen series*, in "Theory and Applications of Nonlinear Control Systems," C. I. Byrnes and A. Lindquist eds., Elsevier, North-Holland (1986), 323–335.

[65] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control & Opt., **25** (1987) 158-194

[66] H. J. Sussmann, *New differential geometric methods in nonholonomic path finding*, Progress Systems and Control Theory, **12**(1992), pp. 365.

[67] D. Tilbury, R. Murray and S. Sastry, *Trajectory generation for the N-trailer problem using Goursat normal form*, IEEE Trans.Aut.Cntrl. **40** (1995) 802–819.

[68] G. Viennot, *Algèbres de Lie Libres et Monoïdes Libres*, Lecture Notes, Math., 692, Springer, Berlin, 1978.

Stability Analysis Based on Direct Liapunov Method

A. Bacciotti*

Dipartimento di Matematica del Politecnico, Torino, Italy.

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028006

* bacciotti@polito.it

Abstract

According to the present meaning of the term, Control theory is a very broad subject. In its rapid development, it incorporated many pre-existing theories in a more general context. This happened, for instance, to stability theory. Already present in the work of Lagrange, and formalized by the russian mathematician A.M. Liapunov more than 100 years ago, today it is recognized as a fundamental component of control theory. The aim of this short course is to introduce basic concepts and methodologies of stability theory from the classical point of view, and then to point out their relevance for applications to modern control theory.

Contents

1 Unforced systems **318**
1.1 Basic stability notions 318
1.2 Liapunov functions 320
1.3 Sufficient conditions 321
1.4 Converse theorems 322
1.4.1 Asymptotic stability 322
1.4.2 Stability 323
1.5 Time-dependent Liapunov functions 325

2 Stability and nonsmooth analysis **325**
2.1 Generalized derivatives 328
2.2 Criteria for stability 332
2.3 Converse theorems for asymptotic stability 334

3 Stabilization **335**
3.1 Jurdjevic-Quinn method 335
3.2 Optimality 337
3.2.1 The associated optimization problem 337
3.2.2 From stabilization to optimality 338
3.2.3 From optimality to stabilizability 338
3.2.4 Hamilton-Jacobi equation 339
3.3 Dissipation 339
3.4 The generality of damping control 343

4 Control Liapunov functions **344**
4.1 Smooth control Liapunov functions 345
4.2 Asymptotic controllability 347

References **351**

Introduction

We are interested in time-invariant control systems of the form

$$\dot{x} = f(x, u) \tag{1}$$

where $x \in \mathbb{R}^n$ represents the physical state of the system, and $u \in \mathbb{R}^m$ represents the input from the exterior world. In general, the input is decomposed as a sum $u = u_c + u_r + u_d + \dots$ (u_c =control, u_r =reference signal, u_d =disturbance, ...). The action of the control consists of finding u_c in such a way that the system evolves according to some prescribed goals. Usually two typical control actions can be performed

- *open loop control:* $u_c = c(t)$ (it may also depend on the initial state)
- *closed loop (automatic, feedback) control:* $u_c = k(x)$.

Just in order to fix the notation, assume that a notion of solution has been specified. Then, we denote by $\mathcal{S}_{x_0,u(\cdot)}$ the set of all solutions of (1) corresponding to a given initial state x_0 and a given input $u = u(t)$. When we want to emphasize the dependence of a particular solution $\varphi(t) \in \mathcal{S}_{x_0,u(\cdot)}$ on the initial state and the input, we may also write $x = \varphi(t; x_0, u(\cdot))$. When only the input is provided by a feedback $u = k(x)$, solutions of (1) are denoted by $x = \varphi_{k(\cdot)}(t; x_0)$.

Clearly, to every feedback $u = k(x)$ and every initial state there corresponds an open loop control $u = k(\varphi_{k(\cdot)}(t; x_0))$, but not vice versa.

Preliminary to control synthesis is system analysis; that is, the analysis of the way the solutions $x = \varphi(t; x_0, u(\cdot))$ are affected by the choice of the input $u = u(t)$. A first step in this direction is the investigation of the so-called *unforced system*

$$\dot{x} = f(x, 0) \ . \tag{2}$$

Since there is no energy supply, we expect that the initial energy is dissipated during the evolution, so that any solution converges to some equilibrium position. However, this is not necessarily the case because of possible unmodeled effects. The behavior could also be affected by undesired phenomena (resonance, multiple equilibrium positions, limit cycles, bifurcations, etc.). The stabilizability problem consists of finding a feedback $u = k(x)$ such that the closed loop system

$$\dot{x} = f(x, k(x)) \tag{3}$$

exhibits improved stability performances. As we shall see, stability of the unforced system is related to a better behavior of (1) with respect to external unpredictable inputs.

Prerequisites. We assume that the reader is familiar with the theory of linear systems of ordinary differential equations, and with basic facts about existence, uniqueness and continuous dependence of (classical) solutions of nonlinear ordinary differential equations.

1 Unforced systems

1.1 Basic stability notions

The mathematical formalization of stability concepts is due to A.M. Liapunov (1892). For convenience, we refer to a system of ordinary differential equation

$$\dot{x} = f(x) \qquad (x \in \mathbb{R}^n) \ . \tag{4}$$

For the moment, we assume that f is continuous on the whole of $\mathbb{R}^n$, so that for each measurable, locally bounded input and each initial condition a (classical) solution exists, but it is not necessarily unique. Solutions of (4) will be denoted by $x = \varphi(t; x_0)$; we shall also write $\mathcal{S}_{x_0}$ instead of $\mathcal{S}_{x_0,0}$.

Definition 1 *We say that (4) is* (Liapunov) stable *at the origin (or that the origin is stable for (4)) if for each $\varepsilon > 0$ there exists $\delta > 0$ such that for each x_0 with $||x_0|| < \delta$ and all the solutions $\varphi(\cdot) \in \mathcal{S}_{x_0}$ the following holds: $\varphi(\cdot)$ is right continuable for $t \in [0, +\infty)$ and*

$$||\varphi(t)|| < \varepsilon \quad \forall t \geq 0 \ .$$

Problem 1 *Prove that if the origin is stable, then it is an equilibrium position for (4) i.e., $f(0) = 0$.*

Definition 2 *We say that (4) is* Lagrange stable *(or that it has the property of uniform boundedness of solutions) if for each $R > 0$ there exists $S > 0$ such that for $||x_0|| < R$ and all the solutions $\varphi(\cdot) \in \mathcal{S}_{x_0}$ one has that $\varphi(\cdot)$ is right continuable for $t \in [0, +\infty)$ and*

$$\|\varphi(t)\| < S \ , \quad \forall t \geq 0 \ .$$

A very special (but very important for engineering applications) case arises when the system is linear i.e.,

$$\dot{x} = Ax \tag{5}$$

where A is a square matrix with constant entries.

Problem 2 *Prove that in the linear case Liapunov stability and Lagrange stability imply each other; give an example to prove that in general Liapunov stability and Lagrange stability are distinct properties.*

Definition 3 *We say that system (4) is* locally asymptotically stable *at the origin (or that the origin is locally asymptotically stable for (4)) if it is stable at the origin and, in addition, the following condition holds: there exists $\delta_0 > 0$ such that*

$$\lim_{t \to +\infty} \|\varphi(t)\| = 0$$

for each x_0 such that $\|x_0\| < \delta_0$, and all the solutions $\varphi(\cdot) \in \mathcal{S}_{x_0}$.

The origin is said to be globally asymptotically stable *if δ_0 can be taken as large as desired.*

Problem 3 *Prove that for linear systems, the Liapunov stability requirement can be dropped in the previous definition (in the sense that it is implied by the remaining conditions).*

Problem 4 *Find an example which shows that in general, the Liapunov stability requirement cannot be dropped in the previous definition (difficult: see [21], [63]).*

Problem 5 *Find an example of a system which is Liapunov stable but not asymptotically stable (easy: there are linear examples).*

Problem 6 *Prove that every linear system which is locally asymptotically stable is actually globally asymptotically stable.*

Remark 1 When dealing with systems without uniqueness, one should distinguish between weak and strong notions. The previous definitions are *strong* notions in the sense that the properties are required to hold for all the solutions, and not only for some of them (see also Remark 5, next section).

Remark 2 Definitions 1 and 3 can be referred to any equilibrium position, that is any point x_0 such that $f(x_0) = 0$. The choice $x_0 = 0$ implies no loss of generality. ■

1.2 Liapunov functions

Liapunov functions are energy-like functions which can be used to test stability. Actually, for each concept of stability there is a corresponding concept of Liapunov function.

Notation: $B_r = \{x \in \mathbb{R}^n : \|x\| < r\}$ and $B^r = \{x \in \mathbb{R}^n : \|x\| > r\}$.

Definition 4 *A* smooth weak Liapunov function in the small *is a real map $V(x)$ which is defined on B_r for some $r > 0$, and fulfills the following properties:*

(i) $V(0) = 0$

(ii) $V(x) > 0$ *for* $x \neq 0$

(iii) $V(x)$ *is of class* C^1 *on* B_r

(iv) $\nabla V(x) \cdot f(x) \leq 0$ for each $x \in B_r$.

When a real function $V(x)$ satisfies (ii), it is usual to say that it is *positive definite*. The function

$$\dot{V}(x) \underset{\text{def}}{=} \nabla V(x) \cdot f(x)$$

is called the *derivative of V with respect to (4)*. Condition (iv) means that $\dot{V}$ is *semi-definite negative*.

A real function $V(x)$ is said to be *radially unbounded* if it is defined on B^r for some $r > 0$, and

$$\lim_{\|x\| \to +\infty} V(x) = +\infty \ .$$

Problem 7 *Radial unboundedness is equivalent to say that the level sets $\{x \in \mathbb{R}^n : V(x) \leq a\}$ are bounded for each $a \in \mathbb{R}$.*

Definition 5 *A function $V(x)$ defined on B^r for some $r > 0$, which is radially unbounded and fulfills (iii) and (iv) of Definition 4 (with B_r replaced by B^r), will be called a* smooth weak Liapunov function in the large.

Definition 6 *A* smooth strict Liapunov function in the small *is a weak Liapunov function such that* $\dot{V}(x)$ *is* negative definite*; in other words, it satisfies, instead of (iv),*

(v) $\nabla V(x) \cdot f(x) < 0$ *for each* $x \in B_r$ $(x \neq 0)$.

A function $V(x)$ *defined for all* $x \in \mathbb{R}^n$*, which is radially unbounded and fulfills the properties (i), (ii), (iii), (v) with* B_r *replaced by* $\mathbb{R}^n$*, will be called a* smooth global strict Liapunov function.

Remark 3 As far as Liapunov functions are assumed to be of class (at least) C^1, condition (iv) is clearly equivalent to the following one:

(iv') for each solution $\varphi(\cdot)$ *of (4) defined on some interval* I *and lying in* B_r*, the composite map* $t \mapsto V(\varphi(t))$ *is non-increasing on* I.

Such a monotonicity condition can be considered as a "nonsmooth analogous" of properties (iii), (iv). Indeed, it can be stated without need of any differentiability (or even continuity) assumption about V.

■

Definition 7 *Let* $r > 0$. *A function* $V : B_r \to \mathbb{R}$ *is called a* generalized weak Liapunov function in the small *if it satisfies (i), (iv') and, in addition, the following two properties:*

(ii') *for some* $\eta < r$ *and for each* $\sigma \in (0, \eta)$ *there exists* $\lambda > 0$ *such that* $V(x) > \lambda$ *when* $\sigma \leq ||x|| \leq \eta$

(iii') $V(x)$ *is continuous at* $x = 0$.

1.3 Sufficient conditions

Theorem 1 *If there exists a smooth weak Liapunov function in the small, then (4) is stable at the origin.*

Theorem 2 *If there exists a smooth strict Liapunov function in the small, then (4) is locally asymptotically stable at the origin.*

If there exists a smooth global strict Liapunov function, then (4) is globally asymptotically stable at the origin.

These theorems are respectively called First and Second Liapunov Theorem. Next theorem is due to Yoshizawa.

Theorem 3 *If there exists a smooth weak Liapunov function in the large, then (4) is Lagrange stable.*

Problem 8 *Prove that if there exists a symmetric, positive definite real matrix P such that*

$$A^{\mathbf{t}}P + PA \leq 0$$

then $V(x) = x^{\mathbf{t}}Px$ is a weak Liapunov function for the linear system (5), so that the system is stable.

Problem 9 *Prove that if P and Q are symmetric, positive definite real matrices such that*

$$A^{\mathbf{t}}P + PA = -Q \tag{6}$$

then $V(x) = x^{\mathbf{t}}Px$ is a strict Liapunov function in the large for (5).

1.4 Converse theorems

From a mathematical point of view, the question whether Theorems 1, 2 and 3 are invertible is quite natural. Recently, it has been recognized to be an important question also for applications to control theory.

1.4.1 Asymptotic stability

Great contributions to studies about the invertibility of second Liapunov Theorem were due to Malkin, Barbashin and Massera, around 1950. In particular, in [95] Massera proved the converse under the assumption that the vector field f is locally Lipschitz. For such vector fields, he proved that asymptotic stability actually implies the existence of a Liapunov function of class C^∞. In 1956, Kurzweil ([89]) proved that the regularity assumption about f can be relaxed.

Theorem 4 *Let f be continuous. If (4) is locally asymptotically stable at the origin then there exists a C^∞ strict Liapunov function in the small.*

If the system is globally asymptotically stable at the origin, then there exists a C^∞ global strict Liapunov function.

It is worth noticing that Kurzweil's Theorem provides a Liapunov function of class C^∞ in spite of f being only continuous.

1.4.2 Stability

The invertibility of first Liapunov theorem is a more subtle question.

Problem 10 *Find an example in order to prove that a system with a stable equilibrium position may admit no continuous Liapunov functions (difficult: see [8], [85]).*

For one-dimensional systems with a stable equilibrium position it is proven in [16] there may be a variety of situations.

- continuous but not locally Lipschitz Liapunov functions
- locally Lipschitz but not C^1 Liapunov functions.

However, if there exists a C^1 Liapunov function then there are also C^∞ Liapunov functions. For two-dimensional systems the situation is still worse. We may have Liapunov functions of class C^r but not of class C^{r+1} $(0 \leq r \leq \omega)$. All this can be done with $f \in C^\infty$.

The following results concerns generalized Liapunov functions. For reader's convenience, we insert the simple proofs. Note that the sufficiency part of Theorem 5 is an extension of the original First Liapunov Theorem.

Theorem 5 *System (4) is Liapunov stable at the origin if and only if there exists a generalized weak Liapunov function in the small.*

Proof. Assume that $V(x)$ is a generalized Liapunov function defined for $||x|| < r$. Let us fix $\varepsilon < \eta$, where η is as in property (ii') of Definition 7, and let $\sigma < \varepsilon$. Then, there exists $\lambda > 0$ such that $V(x) > \lambda$ on the annulus $\sigma \leq ||x|| \leq \varepsilon$. Since $V(0) = 0$ and V is continuous at $x = 0$, there exists $\delta > 0$ such that $V(x) < \lambda$ for $||x|| < \delta$. It is clear that $\delta < \sigma$. We claim that trajectories issuing from the ball B_δ remain inside the ball B_ε for each $t \geq 0$, as required by the definition of stability. In the opposite case, we should have $||\varphi(t)|| \geq \varepsilon$ for some x_0 with $||x_0|| < \delta$, some $t \geq 0$ and some $\varphi \in \mathcal{S}_{x_0}$. This would imply $V(x_0) < \lambda < V(\varphi(t))$, which is impossible since V is non-increasing along the trajectories of the system.

Now we prove the converse statement. Let us assume that the origin is stable and define

$$V(x) = \sup\{\|\varphi(t)\|, t \geq 0, \varphi \in \mathcal{S}_x\} \ .$$

According to the stability assumption, there must exist $r > 0$ such that $V(x) < +\infty$ for $x \in B_r$. It is obvious that $V(x) \geq \|x\|$ if $x \neq 0$, and this

in turn implies (ii'). It is also clear that $V(0) = 0$ because of stability. The stability assumption is invoked also in order to prove that $V(x)$ is continuous at the origin. Indeed, $\forall \varepsilon > 0\ \exists \delta > 0$ such that

$$\|x\| < \delta,\ \varphi \in \mathcal{S}_x \Longrightarrow \|\varphi(t)\| < \varepsilon,\ t \geq 0$$

and hence $V(x) \leq \varepsilon$.

The monotonicity condition (iv') is a trivial consequence of the construction of V and the fact that new solutions can be obtained piecing together solutions defined on consecutive intervals.

∎

Theorem 6 *Assume that the right-hand side of (4) is locally Lipschitz continuous. Then, if (4) is Liapunov stable at the origin there exists a lower semi-continuous generalized weak Liapunov function in the small.*

Proof. Let $V(x)$ be defined as in the proof of Theorem 5. We prove that V is lower semicontinuous at any arbitrary point $x_0 \in X$. By the construction of V, for each $\varepsilon > 0$ there exists $\tau > 0$ such that

$$V(x_0) - \varepsilon/2 < \|\varphi(\tau)\|$$

where $\varphi(t)$ is the solution issuing from x_0. We now use the assumption about $f(x)$. Recall that differential equations with locally Lipschitz continuous right-hand side exhibit (uniqueness of solutions and) continuous dependence with respect to the initial data. Thus, there is $\delta > 0$ such that for all x with $\|x - x_0\| < \delta$ we have

$$\|\varphi(\tau)\| - \|\psi(\tau)\| \leq \|\varphi(\tau) - \psi(\tau)\| < \varepsilon/2$$

where $\psi(t)$ is the solution issuing from x. Hence,

$$V(x_0) - \varepsilon/2 < \|\varphi(\tau)\| \leq \|\psi(\tau)\| + \varepsilon/2\ .$$

Again, by the definition of V, we have $\|\psi(\tau)\| \leq V(x)$. In conclusion,

$$V(x_0) - \varepsilon < \|\psi(\tau)\| \leq V(x)$$

that was required to prove.

∎

Problem 11 *State and prove analogous results for Lagrange stability.*

1.5 Time-dependent Liapunov functions

Another possible approach to the invertibility of first Liapunov theorem is to seek time-dependent Liapunov functions. Recall that $a \in \mathcal{K}_0$ means that $a : [0, +\infty) \to \mathbb{R}$ is a continuous, strictly increasing function such that $a(0) = 0$. If in addition $\lim_{r \to +\infty} a(r) = +\infty$, then we write $a \in \mathcal{K}_0^\infty$.

Definition 8 *A* time-dependent weak Liapunov function in the small *for (4) is a real map $V(t, x)$ which is defined on $[0, +\infty) \times B_r$ for some $r > 0$, and fulfills the following properties:*
(i) there exist $a, b \in \mathcal{K}_0$ such that

$$a(||x||) \leq V(t, x) \leq b(||x||) \qquad \text{for } t \in [0, +\infty),\ x \in B_r$$

(ii) for each solution $\varphi(\cdot)$ of (4) and each interval $I \subseteq [0, +\infty)$ one has

$$t_1, t_2 \in I,\ t_1 < t_2 \Longrightarrow V(t_1, \varphi(t_1)) \geq V(t_2, \varphi(t_2))$$

provided that $\varphi(\cdot)$ is defined on I and $\varphi(t) \in B_r$ for $t \in I$.

From (i) it follows $V(t, 0) = 0$. The existence of a time-dependent weak Liapunov function is sufficient to prove stability of the origin for (4), as well. The following statement is a particular case of a theorem independently proved by Krasovski, Kurzweil and Yoshizawa around 1955.

Theorem 7 *Consider the system (4), and assume that $f(x)$ is locally Lipschitz continuous. If the origin is stable, then, there exists a weak Liapunov function in the small of class C^∞.*

Unfortunately, the conclusion fails if f is only continuous ([90]).

2 Stability and nonsmooth analysis

In control theory, one often needs to resort to discontinuous feedback. For this reason, we are interested in the extension of stability theory to systems

$$\dot{x} = f(x) \tag{7}$$

with discontinuous right-hand-side. More precisely, we assume that $f : \mathbb{R}^n \to \mathbb{R}^n$ is locally bounded and (Lebesgue) measurable. Under these assumptions,

the existence of *classical* (i.e., differentiable everywhere and satisfying (7) everywhere) is not guaranteed.

We say that $\varphi(t)$ is a *Carathéodory* solution if $\varphi \in AC$ and it satisfies (7) a.e..

We say that $\varphi(t)$ is a *Filippov* solution if $\varphi \in AC$ and it satisfies a.e. the differential inclusion

$$\dot{x} \in F(x)$$

where

$$F(x) = \mathbf{F}f(x) \underset{\text{def}}{=} \bigcap_{\delta>0} \bigcap_{\mu(N)=0} \overline{\text{co}}\,\{f(B_\delta(x)\backslash N)\} \tag{8}$$

where $\overline{\text{co}}$ denotes the convex closure of a set and μ is the usual Lebesgue measure of $\mathbb{R}^n$.

We say that $\varphi(t)$ is a *Krasowski* solution if $\varphi \in AC$ and it satisfies a.e. the differential inclusion

$$\dot{x} \in K(x)$$

where

$$K(x) = \mathbf{K}f(x) \underset{\text{def}}{=} \bigcap_{\delta>0} \overline{\text{co}}\,\{f(B_\delta(x))\}\ . \tag{9}$$

Problem 12 *Compute* $\mathbf{F}f$ *and* $\mathbf{K}f$ *in the following cases:*

$$f(x) = \operatorname{sgn} x\ , \quad f(x) = |\operatorname{sgn} x|\ , \quad f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}\ .$$

Every Filippov solution is a Krasowski solution but there may be Carathéodory solutions which are not Filippov solution (find an example).

It is not yet clear what type of solution is the best for control theory applications. Here, we focus on Filippov solutions. In particular, we want to give criteria for stability which apply to discontinuous systems and involves nonsmooth (say, locally Lipschitz continuous) Liapunov functions.

We recall that if $f(x)$ is measurable and locally bounded, then the multivalued map $F(x) = \mathbf{K}_x f(x)$ enjoys the following properties

$\mathbf{H}_1$) $F(x)$ is a nonempty, compact, convex subset of $\mathbb{R}^n$, for each $x \in \mathbb{R}^n$

$\mathbf{H}_2$) $F(x)$, as a multivalued map of x, is upper semi-continuous i.e.,

$$\forall x \ \forall \varepsilon \ \exists \delta \ : \quad \|\xi - x\| < \delta \Longrightarrow F(\xi) \subseteq F(x) + B_\varepsilon$$

$\mathbf{H}_3$) for each $R > 0$ there exists $M > 0$ such that

$$F(x) \subset \{v : \ ||v|| \leq M\}$$

for $0 \leq ||x|| \leq R$.

When $f(x)$ is locally bounded, there is also an equivalent (perhaps more intuitive) definition (see [100]). Indeed, it is possible to prove that there exists a set $N_0 \subset \mathbb{R}^n$ (depending on f) with $\mu(N_0) = 0$ such that, for each $N \subset \mathbb{R}^n$ with $\mu(N) = 0$, and for each $x \in \mathbb{R}^n$,

$$\mathbf{F}f(x) = \text{co}\{v : \ \exists\{x_i\} \text{ with } x_i \to x \text{such that } x_i \notin N_0 \cup N \text{ and } v = \lim f(x_i)\} \, . \tag{10}$$

In [100], the reader will find also some useful rules of calculus for the "operator" $\mathbf{F}$.

Remark 4 A second, important reason to consider differential inclusions is given by the fact that a system with free inputs can be actually reviewed as a differential inclusion of a particular type.

Consider a system with a continuous right-hand side $f(x, u)$. Let U be a given subset of $\mathbb{R}^m$, and assume that an input function $u(\cdot)$ is admissible only if it fulfills the constraint $u(t) \in U$ a.e. $t \geq 0$. Then, it is evident that every solution corresponding to an admissible input is a solution of a differential inclusion with the right-hand side defined by $f(x, U)$.

A celebrated theorem by Filippov states that also the converse is true, provided that $f(x, u)$ is continuous and U is a compact set. We recall that under the same assumptions on $f(x, u)$ and U, $f(x, U)$ turns out to be Hausdorff continuous[1]. On the other hand, if $f(x, u)$ is continuous and locally Lipschitz continuous with respect to x (uniformly with respect to u) then $f(x, U)$ is locally Lipschitz Hausdorff continuous with respect to x.

[1]Hausdorff continuity is continuity of set valued maps with respect to Hausdorff distance; the Hausdorff distance between nonempty, compact subsets of $\mathbb{R}^n$, usually denoted by h, is given by

$$h(A, B) = \max\{\sup_{a \in A} \text{dist}\,(a, B), \sup_{b \in B} \text{dist}\,(b, A)\}$$

where $\text{dist}\,(a, B) = \inf_{b \in B} ||a - b||$.

We can retain the following conclusion. From the point of view of control theory, it is interesting to consider differential inclusions

$$\dot{x} \in \mathcal{F}(x) \tag{11}$$

where either $\mathcal{F}$ satisfies assumptions $\mathbf{H}_1$, ..., $\mathbf{H}_3$ or $\mathcal{F}$ is locally Lipschitz Hausdorff continuous. The extension of Definitions 1, 2 and 3 to differential inclusions is straightforward, but the following remark is appropriate.

Remark 5 Let us recall that in the literature about differential inclusions, there are two possible way to interpret the classical notions of stability. The notions labelled "weak" are deduced by asking that the respective conditions are satisfied for at least one solution corresponding to prescribed initial data. Although they are not studied in the present work, these notions are not irrelevant from a control theory point of view: indeed, they are related to controllability problems, feedback stabilization, viability theory and so on.

On the contrary, the notions labelled "strong" imply that all the solutions corresponding to the prescribed initial data satisfy the respective conditions. From our point of view, this type of stability is the ideal one we can look for, when the inputs are interpreted as disturbances. Indeed, it is obviously desirable that the effect of a disturbance is quickly absorbed and that it does not affect too much the evolution of the system.

From now on, we shall not use explicitly the qualifiers "weak" and "strong" since we are interested only in the "strong" notions.

2.1 Generalized derivatives

Let $V(x) : \mathbb{R}^n \to \mathbb{R}$ be defined on an open subset Q of $\mathbb{R}^n$. For $x \in Q, v \in \mathbb{R}^n$ and $h \in \mathbb{R}$, we are interested in the difference quotient

$$\mathcal{R}(h, x, w) = \frac{V(x + hw) - V(x)}{h} .$$

Let finally $\bar{x} \in Q, \bar{w} \in \mathbb{R}^n$. The usual directional derivative at $\bar{x}$ with respect to $\bar{w}$ is defined as

$$DV(\bar{x}, \bar{w}) = \lim_{h \to 0} \mathcal{R}(h, \bar{x}, \bar{w})$$

provided that the limit exists and it is finite. When the existence of the limit is not guaranteed, certain notions of generalized derivatives may represent useful substitutes. The most classical type of generalized derivatives are *Dini*

derivatives. The idea is as follows. To V, $\bar{x}$ and $\bar{w}$ we associate four numbers $\overline{D^+}V(\bar{x},\bar{w})$, $\underline{D^+}V(\bar{x},\bar{w})$, $\overline{D^-}V(\bar{x},\bar{w})$, $\underline{D^-}V(\bar{x},\bar{w})$. The former is defined as

$$\limsup_{h\to 0^+} \mathcal{R}(h,\bar{x},\bar{w})$$

and the other are defined in similar way, taking the infimum instead of the supremum and the left limit instead of the right one, according to the notation. In this paper we shall make use of Dini derivatives, but we need also other types of generalized derivatives.

The *upper right contingent derivative* $\overline{D_K^+}(\bar{x},\bar{w})$ is defined as

$$\limsup_{\substack{h\to 0^+ \\ w\to \bar{w}}} \mathcal{R}(h,\bar{x},w) \ .$$

Analogously, one can define $\underline{D_K^+}V(\bar{x},\bar{w})$, $\overline{D_K^-}V(\bar{x},\bar{w})$, $\underline{D_K^-}V(\bar{x},\bar{w})$.

Problem 13 *Show that the following relations hold:*

$$\underline{D_K^+}V(\bar{x},\bar{w}) = \underline{D_K^-}(-V)(\bar{x},-\bar{w}) = -\overline{D_K^-}V(\bar{x},-\bar{w}) = -\overline{D_K^+}(-V)(\bar{x},\bar{w}) \ .$$

Contingent derivatives are in some way related to the so-called contingent cone, introduced by Bouligand in 1930. Note that if V is locally Lipschitz continuous, then any contingent derivative coincides with the corresponding Dini derivative and the same is true if $n = 1$ and $\bar{w} \neq 0$.

More recently, *upper Clarke directional derivative* $\overline{D_C}V(\bar{x},\bar{w})$ appears in the context of nonsmooth optimization theory ([33]). It is defined as

$$\limsup_{\substack{h\to 0 \\ x\to \bar{x}}} \mathcal{R}(h,x,\bar{w})$$

(in this case we do not distinguish between right and left limits, since they always coincide). Similarly, we can define $\underline{D_C}V(\bar{x},\bar{w})$. Note that $\underline{D_C}V(\bar{x},\bar{w}) = -\overline{D_C}V(\bar{x},-\bar{w})$.

It is not difficult to verify that the map

$$w \mapsto \overline{D^+}V(\bar{x},w)$$

from $\mathbb{R}^n$ to $\mathbb{R} \cup \{\pm\infty\}$ is positively homogeneous. The same is true for any other type of generalized (Dini, contingent or Clarke, upper or lower, left or right) derivative. In addition, $w \mapsto \overline{D_C}V(\bar{x}, w)$ is subadditive (and hence a convex function).

In general, $\overline{D^+}V(\bar{x}, \bar{w}) \leq \overline{D_C}V(\bar{x}, \bar{w})$ and $\overline{D^+}V(\bar{x}, \bar{w}) \leq \overline{D_K^+}V(\bar{x}, \bar{w})$.

Problem 14 *It may happen that for some $\bar{x}$ and $\bar{w}$*

$$\overline{D^+}V(\bar{x}, \bar{w}) < \overline{D_C}V(\bar{x}, \bar{w}) \quad , \quad \overline{D^+}V(\bar{x}, \bar{w}) < \overline{D_K^+}V(\bar{x}, \bar{w}) \ ,$$
$$\overline{D_C}V(\bar{x}, \bar{w}) > \overline{D_K^+}V(\bar{x}, \bar{w}) \quad , \quad \overline{D_C}V(\bar{x}, \bar{w}) < \overline{D_K^+}V(\bar{x}, \bar{w}) \ .$$

Give at least one example for each inequality.

Clarke gradient of V at x is given by

$$\partial_C V(x) = \{p \in \mathbb{R}^n : \ \forall w \in \mathbb{R}^n \text{ one has } \underline{D_C}V(x, w) \leq p \cdot w \leq \overline{D_C}V(x, w)\} \ .$$

The set $\partial_C V(x)$ is convex for each $x \in Q$. Moreover, if V is Lipschitz continuous, then $\partial_C V(x)$ turns out to be compact. The upper Clarke derivative can be recovered from Clarke gradient. Indeed,

$$\overline{D_C}V(x, w) = \sup_{p \in \partial_C V(x)} p \cdot w$$

(and, in a similar way, $\underline{D_C}V(x, w) = \inf_{p \in \partial_C V(x)} p \cdot w$).

If V is locally Lipschitz continuous, by Rademacher's Theorem its gradient $\nabla V(x)$ exists almost everywhere. Let S be the subset of $\mathbb{R}^n$ where the gradient does not exist. Then, it is possible to characterize Clarke generalized gradient as:

$$\partial_C V(x) = \text{co} \left\{ \lim_{i \to \infty} \nabla V(x_i), \ x_i \to x, \ x_i \notin S \cup S_1 \right\}$$

where S_1 is any subset of $\mathbb{R}^n$, with $\mu(S_1) = 0$. This suggests an analogy between Clarke gradient and Filippov's operator $\mathbf{F}$ (see [100]).

A map $V(x)$ is said to be *regular* if the usual one-side derivative

$$D^+V(\bar{x}, \bar{w}) = \lim_{h \to 0^+} \mathcal{R}(h, \bar{x}, \bar{w})$$

exists for each $\bar{x}$ and $\bar{w}$, and coincides with the upper Clarke derivative $\overline{D_C}V(\bar{x},\bar{w})$ (equivalently, $D^-V(\bar{x},\bar{w}) = \underline{D_C}V(\bar{x},\bar{w})$). Note that if V is regular,

$$\underline{D_C}V(\bar{x},\bar{w}) = -\overline{D_C}V(\bar{x},-\bar{w}) = -D^+V(\bar{x},-\bar{w}) = D^-V(\bar{x},\bar{w}) \ .$$

By analogy with Clarke's theory, we associate with the contingent derivatives the following two sets:

$$\underline{\partial}V(x) = \{p \in \mathbb{R}^n : \ \overline{D_K^-}V(x,w) \le p \cdot w \le \underline{D_K^+}V(x,w), \ \forall w \in \mathbb{R}^n\} \qquad (12)$$

and

$$\overline{\partial}V(x) = \{p \in \mathbb{R}^n : \ \overline{D_K^+}V(x,w) \le p \cdot w \le \underline{D_K^-}V(x,w), \ \forall w \in \mathbb{R}^n\} \ .$$

These sets are both convex and closed and may be empty. In addition, they are bounded provided that the contingent derivatives take finite values for each direction. If one of them contains two distinct elements, the other is necessarily empty.

Note that since the contingent derivatives are not convex functions, it is not possible in general to recover their values for arbitrary directions from $\overline{\partial}V(x)$ and $\underline{\partial}V(x)$.

It turns out (see [59]) that $\overline{\partial}V(x)$ and $\underline{\partial}V(x)$ coincide respectively with the so-called *generalized super* and *sub-differentials*. They can be defined in an independent way, by means of a suitable extension of the classical definition of Fréchet differential. More precisely, one has

$$\overline{\partial}V(x) = \{p \in \mathbb{R}^n : \ \limsup_{h\to 0} \frac{V(x+h) - V(x) - p\cdot h}{|h|} \le 0\}$$

and

$$\underline{\partial}V(x) = \{p \in \mathbb{R}^n : \ \liminf_{h\to 0} \frac{V(x+h) - V(x) - p\cdot h}{|h|} \ge 0\} \ .$$

Using this representation, it is not difficult to see that if $\underline{\partial}V(x)$ and $\overline{\partial}V(x)$ are both nonempty, then they coincide with the singleton $\{\nabla V(x)\}$ and V is differentiable at x in classical sense.

Clarke gradient and generalized differentials are related by $\overline{\partial}V(x) \cup \underline{\partial}V(x) \subseteq \partial_C V(x)$.

In the class of locally Lipschitz functions, regularity can be characterized in terms of generalized differentials.

Proposition 1 *Let V be locally Lipschitz continuous. Then, V is regular if and only if $\partial_C V(x) = \underline{\partial}V(x)$ for all x.*

We finally recall the definition of the proximal gradient. In analytic terms, the *proximal subgradient* of V at x is the set of all vectors p which enjoy the following property. There exists $\sigma \geq 0$ and $\delta \geq 0$ such that for each z with $|z - x| < \delta$,

$$V(z) - V(x) \geq p \cdot (z - x) - \sigma |z - x|^2 \ .$$

The proximal subgradient is denoted $\partial_P V(x)$. It is of course possible to define also the proximal supergradient $\partial^P V(x)$. For each x, $\partial_P V(x)$ is convex but not necessarily closed. In general, $\partial_P V(x) \subseteq \underline{\partial}V(x)$.

Relationship among these types of generalized derivatives, gradients and differentials, and comments on their possible geometric interpretation can be found in [36], [37].

2.2 Criteria for stability

The notion of generalized Liapunov function can be easily extended to the case of differential inclusions. More or less with the same proof as Theorem 5, we obtain the following necessary and sufficient condition.

Theorem 8 *Consider the differential inclusion (11) and assume that for each initial state x_0 there exists at least one solution. The origin is stable if and only if there exists a generalized weak Liapunov function in the small.*

There is also a version of Theorem 6 for differential inclusions with locally Lipschitz (Hausdorff) continuous right-hand side. Moreover, under such assumption the monotonicity condition can be expressed by means of the contingent derivative ([12], [36]).

Theorem 9 *Assume that $\mathcal{F}(x)$ is compact valued and locally Lipschitz (Hausdorff) continuous. The origin is stable if and only if there exists a positive definite, lower semi-continuous function $V(x)$ such that*

$$\sup_{v \in \mathcal{F}(x)} \underline{D_K^+} V(x, v) \leq 0$$

for all x in some neighborhood of the origin.

Next we focus on differential equations with discontinuous right-hand side.

Theorem 10 *Let us consider system (7), with f measurable and locally bounded. Let $V(x)$ be positive definite and locally Lipschitz continuous. Assume that*

$$\underline{D^+} V(x, v) \leq 0$$

for each $v \in F(x)$ and each $x \in \mathbb{R}^n$. Then, the origin is stable (with respect to Filippov solutions).

Since the upper Clarke's directional derivative majorizes the corresponding upper right Dini's derivative (and this in turn majorizes the lower one), it is clear that if

$$\overline{D_C} V(x, v) \leq 0$$

for each $x \in \mathbb{R}^n$ and $v \in F(x)$, then Theorem 10 applies. However, a criterion based on this inequality is too much conservative, since Clarke's gradient is a very large object and contains in general non-essential directions. On the other hand, Clarke's gradient possesses a rich amount of properties, so that its use could be advisable in view of certain applications. We obtain now a very sharp criterion which allows us to exploit the properties of Clarke's gradient: it avoids at the same time unnecessary verifications. The cost to be paid for this advantages is a new (mild) assumption on V.

Assume that V is a locally Lipschitz continuous and, in addition, a regular function. Let us define the *set valued derivative of V with respect to* (7)

$$\dot{\overline{V}}(x) = \{a \in \mathbb{R} : \ \exists v \in F(x) \text{ such that } \forall p \in \partial_C V(x) \text{ one has } v \cdot p = a\} \ .$$

It is easy to check that $\dot{\overline{V}}$ is closed, bounded and convex. Note that $\dot{\overline{V}}$ may be empty at some point.

Lemma 1 *Let V be locally Lipschitz continuous and regular, and let $\varphi : I \to \mathbb{R}^n$ be a Filippov solution. Let $N \subset I$ be the set of zero measure such that $N = N_0 \cup N_1 \cup N_2$ where:*
N_0 is the set where $\dot{\varphi}(t)$ does not exist
N_1 is the set where $\dot{\varphi}(t) \notin F(\varphi(t))$
N_2 is the set where $\frac{dV}{dt}(\varphi(t))$ does not exist.
Then, for $t \in I \setminus N$, we have $\frac{dV}{dt}(\varphi(t)) \in \dot{\overline{V}}(\varphi(t)))$.

This lemma provides a chain rule for nonsmooth functions: it is essentially due to [123] (see also [11]). As an immediate consequence we obtain new stability criteria.

Theorem 11 *Assume that V is locally Lipschitz continuous and regular. Assume further that*

$$\dot{\overline{V}}(x) \subset (-\infty, 0]$$

for each x in some neighborhood of the origin of $\mathbb{R}^n$. Then, system (7) is stable at the origin, with respect to Filippov solutions.

By means of $\dot{\overline{V}}(x)$ we can also give a sufficient condition for asymptotic stability.

Theorem 12 *Assume that V is radially unbounded, locally Lipschitz continuous and regular. Assume further that there exists a function $\omega \in \mathcal{K}_0$ such that*

$$\dot{\overline{V}}(x) \subset (-\infty, -\omega(\|x\|)]$$

for each $x \in \mathbb{R}^n$. Then, system (7) is globally asymptotically stable at the origin, with respect to Filippov solutions.

In fact, in the previous theorem it is sufficient to assume $\dot{\overline{V}}(x) \subset (-\infty, 0)$ for each $x \in \mathbb{R}^n$ (see [11]).

2.3 Converse theorems for asymptotic stability

For the purposes of this section, we find convenient to refer again to a general differential inclusion

$$\dot{x} \in \mathcal{F}(x) \ , \tag{13}$$

where $\mathcal{F}$ takes for each $x \in \mathbb{R}^n$ nonempty compact values. The first converse of second Liapunov theorem in this context has been given by Lin, Sontag and Wang in [92].

Theorem 13 *Assume that the origin is globally asymptotically stable for (13), where $\mathcal{F}$ is a locally Lipschitz continuous multivalued map, which takes nonempty compact values. Then there exists a C^∞ global strict Liapunov function V, which satisfies*

$$\langle \nabla V(x), v \rangle \leq -c(\|x\|) \qquad \forall x \in \mathbb{R}^n, \ \forall v \in \mathcal{F}(x),$$

for some function $c \in \mathcal{K}_0^\infty$.

Actually, Theorem 13 is stated in [92] in a somewhat different manner: (i) $\mathcal{F}$ takes in [92] the special form $\mathcal{F}(x) := \{f(x,d), d \in D\}$, where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a smooth function and D is a compact set in $\mathbb{R}^m$; (ii) Theorem 2.9 in [92] deals with the asymptotic stability with respect to any *compact invariant set*, instead of the origin.

Another converse Liapunov theorem has been obtained a few years later by Clarke, Ledyaev and Stern in [35] for another class of multivalued maps.

Theorem 14 *Assume that the origin is globally asymptotically stable for (13), where $\mathcal{F}$ is an upper semi-continuous multivalued map, which takes nonempty compact convex values. Then there exists a C^∞ global strict Liapunov function V, which satisfies $\langle \nabla V(x), v \rangle \leq -W(x)$ for each $x \in \mathbb{R}^n$ and each $v \in \mathcal{F}(x)$, for some definite positive continuous function W.*

3 Stabilization

3.1 Jurdjevic-Quinn method

One of the most popular approaches to the nonlinear stabilization problem (and probably the first that has been deeply studied from the mathematical viewpoint) is known as Jurdjevic-Quinn method in the western literature, and *speed gradient method* in the russian literature. In fact, it is not a general

method for stabilization, but rather a method for improving stability performances. It can be described as follows. Let a nonlinear (affine) system be given. Assume that when the input is disconnected, the system has a stable (but not asymptotically stable) equilibrium position. If a (weak) Liapunov function $V(x)$ for the (unforced) system is known and some other technical assumptions are fulfilled, the system can be asymptotically stabilized at the equilibrium by a feedback law whose construction involves $\nabla V(x)$.

The idea can be reviewed as an extension of certain classical stabilization procedures of practical engineering. For instance, let us consider a mechanical system representing a nonlinear elastic force $\ddot{x} = -f(x) + u$ (with $f(x)x > 0$ for $x \neq 0$). In order to study its stability, it is natural to take $V(x, \dot{x}) = \frac{(\dot{x})^2}{2} + \int f(x)\, dx$ as a Liapunov function. Now, asymptotic stabilization can be achieved by "proportional derivative" control, which actually amounts to add friction to the system. It is not difficult to see that this is actually a particular case of feedback depending on the gradient of $V(x, \dot{x})$. For this reason, the method is sometimes also called *damping control.*

From now on, we restrict our attention to affine systems

$$\dot{x} = f(x) + \sum_{i=1}^{m} u_i g_i(x) = f(x) + G(x)u \tag{14}$$

where $x \in \mathbb{R}^n$, $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$. The vector fields $f, g_1, \ldots, g_m$ are required to be at least continuous, and $f(0) = 0$. Affine systems represent a natural generalization of the well-known linear systems

$$\dot{x} = Ax + Bu\ . \tag{15}$$

The basic assumption of the Jurdjevic-Quinn method is that the unforced system is stable at the origin, and that a smooth, weak Liapunov function $V(x)$ is known. Motivated by the previous discussion, we try the feedback

$$u = k(x) = -\frac{\gamma}{2}(\nabla V(x) G(x))^{\mathbf{t}} \tag{16}$$

where $\gamma > 0$ (the coefficient $1/2$ is due to technical reasons).

Problem 15 *Prove that the closed loop system is still Liapunov stable at the origin.*

The second typical assumption of the Jurdjevic-Quinn method is that the vector fields appearing in (31) are C^∞. Recall that the Lie bracket operator associates to an (ordered) pair f_0, f_1 of vector fields the vector field

$$[f_0, f_1] = Df_1 \cdot f_0 - Df_0 \cdot f_1$$

(here, Df_i denotes the jacobian matrix of f_i, $i = 0, 1$). The "*ad*" operator is iteratively defined by

$$ad^1_{f_0} f_1 = [f_0, f_1] \qquad ad^{k+1}_{f_0} f_1 = [f_0, ad^k_{f_0} f_1] \ .$$

Theorem 15 *(*JURDJEVIC-QUINN*) Assume that a weak Liapunov function $V(x)$ of class C^∞ for the unforced system associated to (31) is known. Assume further that for each $x \neq 0$ in some neighborhood of the origin we have $\nabla V(x) \neq 0$ and*

$$\dim \operatorname{span} \{ f(x), ad^k_f g_i(x), \ \ i = 1, \dots m, \ \ k = 0, 1, 2, \dots \ \} = n \ .$$

Then, for any $\gamma > 0$, the system is stabilized by the feedback (16).

The proof of this theorem relies on LaSalle's invariance principle.

3.2 Optimality

We consider again affine systems, but now we assume that the vector fields $f, g_1, \dots, g_m$ are locally Lipschitz continuous, so that uniqueness of solutions is guaranteed for any admissible input (but not under continuous feedback).

We need also to limit the class of admissible inputs. From now on, by an *admissible input* we mean any piecewise continuous, locally bounded function $u(t) : [0, +\infty) \to \mathbb{R}^m$. Without loss of generality, we always assume that any admissible input is right-continuous.

Assume that (14) can be asymptotically stabilized by a feedback law of the form (16). Then, an optimization problem can be associated to the stabilization problem. The solution of the optimization problem can be put in feedback form: it is exactly two times the feedback law (16). It follows some details.

3.2.1 The associated optimization problem

Let a continuous, positive definite and radially unbounded function $h(x)$ be given. We associate to (14) the following cost functional

$$J(x_0, u(\cdot)) = \frac{1}{2}\int_0^{+\infty} \left(h(\varphi(t)) + \frac{\|u(t)\|^2}{\gamma} \right) dt \tag{17}$$

where $\varphi(t) = \varphi(t; x_0, u(\cdot))$. For a given initial state x_0, we say that the minimization problem defined by (17) is solvable if there exists an admissible input, denoted by $u^*_{x_0}(t)$ such that

$$J(x_0, u^*_{x_0}(\cdot)) \leq J(x_0, u(\cdot))$$

for any other admissible input $u(t)$. The value function is defined by

$$V(x_0) = \inf_u J(x_0, u(\cdot)) \ .$$

$V(x_0)$ is actually a minimum if and only if the minimization problem is solvable for x_0.

3.2.2 From stabilization to optimality

Assume that there exist a radially unbounded, positive definite, C^1 function $V(x)$ and a positive number γ such that (14) is asymptotically stabilizable by means of the continuous feedback (16). Assume further that the closed-loop system admits $V(x)$ as a strict Liapunov function, with the additional requirement that the derivative of $V(x)$ with respect to the closed loop system is radially unbounded (this last assumption is not restrictive).

Set $h(x) = -2\nabla V(x) f(x) + \gamma \|\nabla V(x) G(x)\|^2$.

Then, the optimization problem has a solution for each x_0, the solution can be put in feedback form

$$u = k(x) = -\gamma (\nabla V(x) G(x))^{\mathbf{t}} \tag{18}$$

and the value function coincides with $V(x_0)$. Going from stabilization to optimality is called an inverse optimization problem in [122] (where the problem is treated with h positive semi-definite).

3.2.3 From optimality to stabilizability

Assume that there exist a continuous, positive definite, radially unbounded function $h(x)$ and a positive number γ such that the minimization problem (17) is solvable for each initial state x_0. Moreover, assume that the value

function $V(x_0)$ is radially unbounded and of class C^1. Then, system (14) is asymptotically stabilizable by means of the continuous feedback

$$u = k(x) = -\alpha(\nabla V(x)G(x))^{\mathbf{t}} \tag{19}$$

for any $\alpha \geq \frac{\gamma}{2}$. Moreover, the value function V represents a strict Liapunov function for the closed loop system.

3.2.4 Hamilton-Jacobi equation

Solvability of the optimization problem (17) is equivalent to the following statement.

The first order partial differential equation (of the Hamilton-Jacobi type)

$$\nabla U(x)f(x) - \frac{\gamma}{2}\|\nabla U(x)G(x)\|^2 = -\frac{h(x)}{2} \tag{20}$$

has a solution $U(x)$ which is radially unbounded, positive definite and of class C^1.

Problem 16 *Prove that if the system is linear, $h(x) = 2\|x\|^2$ and $\gamma = 1/2$, then the Hamilton Jacobi equation reduces to the matrix equation (the so-called Algebraic Riccati equation)*

$$PA + A^{\mathbf{t}}P - PBB^{\mathbf{t}}P = -I \tag{21}$$

where I is the identity matrix of $\mathbb{R}^n$ and the unknown P is symmetric and positive definite.

3.3 Dissipation

So far we were mainly concerned with internal stability properties. However, there are also relevant notions of "stability" which relate the behavior of the output (or the state evolution) to the size of the external input. The most popular is probably the notion of ISS, due to E. Sontag. We report here the original definition (but many variants are known). For the sake of generality, we state the definition for the general system

$$\dot{x} = f(x, u) \tag{22}$$

although many applications are limited to the relevant case of affine systems. Recall that $\beta \in \mathcal{KL}$ means that $\beta : [0, +\infty) \times [0, +\infty) \to \mathbb{R}$ is decreasing

to zero with respect to the first variable and of class $\mathcal{K}_0$ with respect to the second one.

Definition 9 *We say that (22) possesses the* input-to-state stability *(in short, ISS) property, or that it is an ISS system, if there exist maps $\beta \in \mathcal{KL}$, $\gamma \in \mathcal{K}_0$ such that, for each initial state x_0, each admissible input $u : [0, +\infty) \to \mathbb{R}^m$, each solution $\varphi(\cdot) \in \mathcal{S}_{x_0,u(\cdot)}$ and each $t \geq 0$*

$$\|\varphi(t)\| \leq \beta(t, \|x_0\|) + \gamma(\|u\|_\infty) \ .$$

Problem 17 *If the system is ISS and we set $u \equiv 0$, then we obtain a globally asymptotically stable system. Prove it.*

The following Liapunov-like characterization of the ISS property is very useful [137], [138]).

Theorem 16 *For the system (22) the following statements are equivalent:*

(i) the system possesses the ISS property

(ii) there exist a positive definite, radially unbounded C^∞ function $V : \mathbb{R}^n \to \mathbb{R}$ and two functions $\rho, \chi \in \mathcal{K}_0^\infty$ such that

$$\nabla V(x) \cdot f(x, u) < -\chi(||x||)$$

for all $x \in \mathbb{R}^n$ $(x \neq 0)$ and $u \in \mathbb{R}^m$, provided that $\|x\| \geq \rho(\|u\|)$

(iii) there exist a positive definite, radially unbounded C^∞ function $V : \mathbb{R}^n \to \mathbb{R}$ and two functions $\omega, \alpha \in \mathcal{K}_0^\infty$ such that

$$\nabla V(x) \cdot f(x, u) \leq \omega(\|u\|) - \alpha(\|x\|)$$

for all $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$.

A systems is said to be *IS-stabilizable* if the ISS property can be recovered by applying a suitable feedback law of the form $u = k(x) + \tilde{u}$. The following result concerns the affine system (14) ([126]).

Theorem 17 *Every globally asymptotically stable (or continuously globally asymptotically stabilizable) affine system of the form (31) is IS-stabilizable.*

In fact, ISS systems can be reviewed as special cases of dissipative systems. We proceed to introduce this notion. First of all, we complete the description of the system by associating with (22) an observation function $c(x) : \mathbb{R}^n \to \mathbb{R}^p$. In other words, we consider

$$\begin{cases} \dot{x} = f(x,u) \\ y = c(x) \ . \end{cases} \tag{23}$$

The variable y is called the *output*: it represents the available information about the evolution of the system. Let $w : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ be a given function, which will be called the supply rate, and consider the following three dissipation inequalities.

(D1) (intrinsic version) For each admissible input $u(\cdot)$, each $\varphi \in \mathcal{S}_{0,u(\cdot)}$ and for each $t \geq 0$

$$\int_0^t w(c(\varphi(s)), u(s))\, ds \geq 0$$

(note the initialization at $x_0 = 0$).

(D2) (integral version) There exists a positive semidefinite function $S(x)$ (called a *storage function* such that for each admissible input $u(\cdot)$, each initial state x_0, each $\varphi \in \mathcal{S}_{x_0,u(\cdot)}$ and for each $t \geq 0$

$$S(\varphi(t)) \leq S(x_0) + \int_0^t w(c(\varphi(s)), u(s))\, ds \ .$$

(D3) (differential version) There exists a positive semidefinite function $S(x) \in C^1$ such that for each $x \in \mathbb{R}^n, u \in \mathbb{R}^m$

$$\nabla S(x) f(x,u) \leq w(c(x), u) \ .$$

It is clear that **(D3)** $\Longrightarrow$ **(D2)** $\Longrightarrow$ **(D1)**. However, these conditions are not equivalent in general. The implication **(D1)** $\Longrightarrow$ **(D2)** requires a complete controllability assumption, while the implication **(D2)** $\Longrightarrow$ **(D3)** requires the existence of at least one storage function of class C^1.

In the literature, inequalities **(D1)**, **(D2)**, **(D3)** are alternatively used to define dissipative systems ([149], [72], [146], [130]). Moreover, several notions of "external" stability can be given by specializing the supply rate w. For instance we have

1) *passivity*, for $w = yu$

2) *finite L_2-gain*, for $w = k^2\|u\|^2 - \|y\|^2$, where k is some real constant.

To explain the name given to the second property, observe that it implies the estimation

$$\int_0^t \|y(s)\|^2\, ds \leq k^2 \int_0^t \|u(s)\|^2\, ds\ .$$

The ISS property can be interpreted as an extension of the finite L_2-gain property. Indeed, according to Theorem 16 (iii), ISS systems are dissipative in the sense of **(D3)**, with $c(x)$ =Identity and supply rate $w(x,u) = \omega(\|u\|) - \alpha(\|x\|))$. Hence, for zero initialization, the following estimation holds

$$\int_0^t \alpha(\|\varphi(s)\|)\, ds \leq k^2 \int_0^t \omega(\|u(s)\|)\, ds \tag{24}$$

(alternatively, we can set $c(x) = \sqrt{\alpha(\|x\|)}$, so that the integrand on the left-hand side becomes $\|y\|^2$).

In the remaining part of this section we focus on the finite L_2-gain property, which has been deeply studied in [146]. Moreover, we limit to affine systems or, more precisely, systems of the form (23) where (22) is replaced by (14).

It is well known that if (23) possesses the finite L_2-gain property and a suitable observability condition is fulfilled, then the unforced part of the system is asymptotically stable at the origin. In particular, the required observability condition is automatically satisfied when $c(x)$ is positive definite. Vice-versa, assume that (14) is smoothly stabilizable. Then, by using a possibly different feedback the system can be rendered ISS (Theorem 17). As a consequence, we have an estimation of the form (24), but in general we cannot predict the nature of the functions ω and α. As an application of the theory developed in the previous sections, we now give a more precise result. For notational consistency, we put $k^2 = 1/(2\gamma)$. The starting point is the following important result ([72], [146])[2].

Theorem 18 *Assume that there exists a positive semidefinite function $\Phi(x) \in C^1$ which solves the equation (of the Hamilton-Jacobi type)*

$$\nabla\Phi(x)f(x) + \frac{\gamma}{2}\|\nabla\Phi(x)G(x)\|^2 = -\|c(x)\|^2 \tag{25}$$

for each $x \in \mathbb{R}^n$. Then, the affine system (23) has a finite L_2-gain.

[2]The theorem is invertible under some restrictive assumptions, but here we need only the direct part

Theorem 19 *Associated with the affine system (14) we consider the optimization problem (17), where h is positive definite and continuous. Assume that the problem is solvable for each x_0, and that the value function $V(x)$ is C^1. Then, by applying the feedback*

$$u = k(x, \tilde{u}) = -\gamma(\nabla V(x)G(x))^{\mathbf{t}} + \tilde{u}$$

and choosing the observation function $c(x) = \sqrt{(h(x)/2)}$, the system (23) has a finite L_2-gain.

As a corollary, we see that if the affine system (14) is stabilizable by a damping feedback

$$u = k(x) = -\frac{\gamma}{2}(\nabla V(x)G(x))^{\mathbf{t}}$$

where $V(x)$ can be taken as a strict Liapunov function for the closed loop system, then the "doubled" feedback $u = 2k(x) + \tilde{u}$ gives rise to a system with finite L_2-gain.

3.4 The generality of damping control

It is well known that if a linear system is stabilizable by means of a continuous feedback, then it is also stabilizable by means of a linear feedback and in fact by a feedback in damping form ($u = -\alpha B^{\mathbf{t}} P x$, where $\alpha \geq 1/2$ and P is a solution of (21)). Surprisingly, this fact has an analogue for the nonlinear case.

The following result is basically due to [78].

Theorem 20 *Consider the affine system (14) and assume that*

$$\|f(x)\| \leq A\|x\|^2 + C \quad \text{and} \quad \|G(x)\| \leq D$$

for some positive constants A, C, D. Assume further that (14) admits a stabilizer $u = k(x)$ such that:

(i) $k(x)$ is of class C^1 and $k(0) = 0$,

(ii) $k(x)$ guarantees sufficiently fast decay: more precisely, we require that each solution of the closed loop system is square integrable i.e.,

$$\int_0^{+\infty} \|\varphi_{k(\cdot)}(t; x_0)\|^2 \, dt < +\infty \tag{26}$$

for each $x_0 \in \mathbb{R}^n$.

Then, there exists a map $V(x)$ *such that the feedback law (16) is a global stabilizer for our systems. In other words, the system also admits a damping control.*

4 Control Liapunov functions

Consider for the moment a general system of the form

$$\dot{x} = f(x, u) \tag{27}$$

where f is continuous and $f(0,0) = 0$. The non-existence of continuous stabilizers for (27) is related to certain obstructions of topological nature. The most famous one is pointed out by the following result, usually referred to as Brockett's test (see [27], [118], [155]).

Theorem 21 *Consider the system (27) and assume that* f *is continuous and that* $f(0,0) = 0$. *A necessary condition for the existence of a continuous stabilizer* $u = k(x)$ *with* $k(0) = 0$, *is that for each* $\varepsilon > 0$ *there exist* $\delta > 0$ *such that*

$$\forall y \in B_\delta \ \exists x \in B_\varepsilon\,, \ \exists u \in B_\varepsilon \quad \text{such that} \quad y = f(x, u)\ .$$

In other words, f must map any neighborhood of the origin in $\mathbb{R}^{n+m}$ onto some neighborhood of the origin in $\mathbb{R}^n$ (note that in the linear case, the condition of this theorem reduces to $\operatorname{rank}(A, B) = n$). There exist whole families of systems (typically, full rank nonholonomic systems with less inputs than states) which do not possess the property of Theorem 21. The most famous example of a system which does not satisfy Brockett's test is the so-called *nonholonomic integrator*

$$\begin{cases} \dot{x}_1 = u_1 \\ \dot{x}_2 = u_2 \\ \dot{x}_3 = x_1 u_2 - x_2 u_1\ . \end{cases} \tag{28}$$

The following interesting example is due to Z. Artstein. It passes Brockett's test. Nevertheless, it cannot be stabilized by a continuous feedback.

$$\begin{cases} \dot{x}_1 = u(x_1^2 - x_2^2) \\ \dot{x}_2 = 2u x_1 x_2 \end{cases} \tag{29}$$

(see [132] for a discussion).

4.1 Smooth control Liapunov functions

We need the following variant of the notion of Liapunov function (see [126], [124]).

Definition 10 *We say that (27) satisfies a* smooth global control Liapunov condition *(or that (27) has a smooth global control Liapunov function) if there exists a radially unbounded, positive definite, C^1 function $V(x)$ vanishing at the origin and enjoying the following property: for each $x \in \mathbb{R}^n$ there exists $u \in \mathbb{R}^m$ such that*

$$\nabla V(x) \cdot f(x,u) < 0 \ . \tag{30}$$

According to Kurzweil converse Theorem, it is clear that if there exists a continuous global stabilizer $u = k(x)$ for (27), then there exists also a smooth global control Liapunov function. The converse is false in general.

Problem 18 *Prove that the system*

$$\begin{cases} \dot{x}_1 = u_2 u_3 \\ \dot{x}_2 = u_1 u_3 \\ \dot{x}_3 = u_1 u_2 \end{cases}$$

possesses the control Liapunov function $V(x_1,x_2,x_3) = x_1^2 + x_2^2 + x_3^2$ but it does not pass Brockett's test.

However, it turns out to be true in the affine case

$$\dot{x} = f(x) + \sum_{i=1}^{m} u_i g_i(x) \tag{31}$$

where $f, g_1, \ldots, g_m$ are continuous vector fields of $\mathbb{R}^n$ (Z. Artstein [4]; but see also [127]). In order to state the theorem, we need to update the terminology. A feedback law $u = k(x)$ is said to be *almost continuous* if it is continuous at every $x \in \mathbb{R}^n \setminus \{0\}$. Moreover, we say that a control Liapunov function satisfies the *small control property*[3] if for each $\varepsilon > 0$ there exists $\delta > 0$ such that for each $x \in B_\delta$, (30) is fulfilled for some $u \in B_\varepsilon$.

[3] If the system admits a continuous stabilizer $u = k(x)$ such that $k(0) = 0$, then the small control property is automatically fulfilled.

Theorem 22 *If there exists a smooth global control Liapunov function for the affine system (31), then the system is globally stabilizable by an almost continuous feedback $u = k(x)$. If there exists a control Liapunov function which in addition satisfies the small control property, then it is possible to find a stabilizer $u = k(x)$ which is everywhere continuous.*

We do not report here the proof of this theorem, but some illustrative comments are appropriate. For sake of simplicity, we limit ourselves to the single input case ($m = 1$). If the vector fields f and g_1 are of class C^q ($0 \leq q \leq +\infty$) and a control Liapunov function of class C^r ($1 \leq r \leq +\infty$) is known, the stabilizing feedback whose existence is ensured by Theorem 22, can be explicitly constructed according to Sontag's "universal" formula

$$k(x) = \begin{cases} 0 & \text{if } b(x) = 0 \\ \frac{a(x) - \sqrt{a^2(x) + b^4(x)}}{b(x)} & \text{if } b(x) \neq 0 \end{cases} \tag{32}$$

where $a(x) = -\nabla V(x) \cdot f(x)$ and $b(x) = \nabla V(x) \cdot g_1(x)$ (see [127] for more details). We emphasize that such $k(x)$ is of class C^s (with $s = \min\{q, r-1\}$) on $\mathbb{R}^n \setminus \{0\}$. If the small control property is assumed, then the feedback law given by (32) turns out to be continuous also at the origin, but further regularity at the origin can be obtained only in very special situations.

It is worth noticing that the universal formula above has a powerful regularizing property. Indeed, if a continuous stabilizer for (31) is known, then Kurzweil's Converse Theorem applies. Hence, the existence of a C^∞ strict Liapunov function $V(x)$ for the closed loop system is guaranteed. It is not difficult to see that the same $V(x)$ is a control Liapunov function for (31). But then, the universal formula can be applied with this $V(x)$, and we obtain a new stabilizing feedback with the same order of differentiability as f and g_1 (at least for $x \neq 0$).

We have already noticed that Artstein's theorem is limited to affine systems. However, the following extension holds (see the remark after Lemma 2.1 in [46]; see also [107]).

Theorem 23 *Consider a system of the form (27), where f is continuous and $f(0,0) = 0$. The following statements are equivalent.*

(i) There exists a discontinuous feedback which stabilizes the system in Filippov's sense and which fulfills the additional condition

$$\lim_{\delta \to 0} \operatorname*{ess\,sup}_{\|x\| < \delta} \|u(x)\| = 0 \tag{33}$$

(ii) There exists a smooth control Liapunov function for which the small control property holds.

4.2 Asymptotic controllability

In this section we assume that $f(x,u)$ is continuous with respect to the pair $(x,u) \in \mathbb{R}^n \times \mathbb{R}^m$, and Lipschitz continuous with respect to x (uniformly with respect to u). We assume also that $f(0,0) = 0$.

Definition 11 *System (27) is said to be* globally asymptotically controllable *at the origin (see [34]) if there exist $C_0 > 0$, $C > 0$ such that:*

(a) for each $x_0 \in \mathbb{R}^n$ there exists an admissible input $u_{x_0}(t) : [0, +\infty) \to \mathbb{R}^m$ such that the unique solution $\varphi(t; x_0, u_{x_0}(\cdot))$ is defined for all $t \geq 0$ and satisfies

$$\lim_{t \to +\infty} \varphi(t; x_0, u_{x_0}(\cdot)) = 0 \tag{34}$$

(b) for each $\varepsilon > 0$ it is possible to find $\eta > 0$ such that if $\|x_0\| < \eta$ then there exists an admissible input $u_{x_0}(t)$ such that (34) holds, and in addition

$$\|\varphi(t; x_0, u_{x_0}(\cdot))\| < \varepsilon \quad \text{for all} \quad t \geq 0 \tag{35}$$

(c) if in (b) the state x_0 satisfies also $\|x_0\| < C_0$, then the input $u_{x_0}(t)$ can be chosen in such a way

$$\|u_{x_0}(t)\| \leq C$$

for a.e. $t \geq 0$.

If (34) is required to hold only for each x_0 in some neighborhood of the origin, then we say that the system is locally asymptotically controllable.

The meaning of this definition is that the system is asymptotically driven toward zero by means of an open loop, bounded control which depends on the initial state.

It is clear that if (27) is stabilizable by means of a continuous feedback, then it is asymptotically controllable. The converse is true if the system is linear[4], but not in general. The classical counter-example is given by the

[4]For linear systems asymptotic controllability, stabilizability by continuous feedback and stabilizability by linear feedback are all equivalent: see [67].

nonholonomic integrator: it is possible to prove that the system is asymptotically controllable: however, we know that it does not pass Brockett's test, so that it is not continuously stabilizable. In fact, because of Ryan's extension of Brockett's test [118], it follows that large classes of asymptotically controllable systems can be stabilized not even by discontinuous feedback, at least as far as the solutions are intended in Filippov's sense. Important progress toward the solution of this problem has been recently made ([34], [1], [105]) by exploiting suitable extensions of the notion of control Liapunov function and/or new notions of solutions for discontinuous ordinary differential equations.

In order to give an idea of such developments, we start by a simple remark. It is clear that if an affine system without drift (like the nonholonomic integrator and the Artstein example (29)) admits a smooth control Liapunov function, then the small control property is automatically fulfilled. It follows from this simple remark and Theorem 22, that there exist no smooth control Liapunov functions for (28) and (29). Nevertheless, both systems are asymptotically controllable. This suggests the possibility of characterizing asymptotic controllability by some weaker notion of control Liapunov function.

Note that if the differentiability assumption about V is relaxed, then the monotonicity condition can be no more expressed in the form (30). In [125] (see also [128]) E. Sontag proved that if f is locally Lipschitz continuous with respect to both x, u, then the global asymptotic controllability is equivalent to the existence of a continuous global control Liapunov function. The monotonicity condition is expressed in [125] by means of Dini derivatives along the solutions (see Chapter 2). In [136] (see also [34]) it is pointed out that the same condition can be also expressed by means of contingent directional derivatives.

With these motivations, we propose a general definition.

Definition 12 *Let $V : \mathbb{R}^n \to \mathbb{R}$ be continuous, positive definite and radially unbounded. Moreover, let $D(x)$ be a set valued map ($D(x)$ should be thought of as some generalized gradient of the map V). We say that V is a* (nonsmooth) global control Liapunov function (with respect to D) *if there exist two maps $W : \mathbb{R}^n \to \mathbb{R}$ and $\sigma : [0, +\infty) \to [0, +\infty)$ such that:*

1) W is continuous, positive definite and radially unbounded

2) σ is increasing

3) for each $x \in \mathbb{R}^n$ *and each* $p \in D(x)$ *there exists* $u_{x,p} \in \mathbb{R}^m$ *such that* $\|u_{x,p}\| \leq \sigma(\|x\|)$ *and*

$$p[f(x) + G(x)u_{x,p}] \leq -W(x) \ . \tag{36}$$

Problem 19 *There is another possible definition which does not make use of the map* σ*. There exists a continuous, positive definite and radially unbounded map* $W : \mathbb{R}^n \to \mathbb{R}$ *for which the following holds. For each compact set* $K \subset \mathbb{R}^n$ *there exists a compact set* $U \subset \mathbb{R}^m$ *such that for each* $x \in K$ *and each* $p \in D(x)$ *there exists* $u_{x,p} \in U$ *such that (36) holds. Prove that the two formulations are actually equivalent.*

Now, an improvement of the aforementioned Sontag's result can be stated in the following way ([105], [106], see also [132]).

Theorem 24 *Consider the system (27) and assume that* f *is locally Lipschitz continuous with respect to both* x, u*. Then, global asymptotic controllability is equivalent to the existence of a nonsmooth global control Liapunov function* $V(x)$ *(with respect to the proximal gradient* $\underline{\partial_P} V(x)$*). In addition,* $V(x)$ *can be taken locally Lipschitz continuous.*

Note that this result applies in particular to Artstein's example (29) (by the way, a locally Lipschitz continuous control Liapunov function for (29) is explicitly given in [132]).

We conclude this chapter by recalling the following stabilizability results.

Theorem 25 *([34]) Assume that the system (27) is globally asymptotically controllable. Then it can be stabilized by time-sampled discontinuous feedback*[5]*.*

Theorem 26 *([107]) Assume that the system (27) admits a locally Lipschitz continuous, nonsmooth global control Liapunov function* $V(x)$ *(with respect to Clarke gradient* $\partial_C V(x)$*). Then there exists also a* <u>smooth</u> *control Liapunov function, so that the system is actually stabilizable in Filippov sense.*

We emphasize that the tool used to express the monotonicity condition actually plays a crucial role.

[5] Roughly speaking, this means that the value of the feedback remains constant for a small interval of time

Acknowledgments

These notes have been prepared for a short summer course at ICTP (Trieste, Italy), September 2001. I wish to thank Professor A. Agrachev, Professor B. Jakubczyk and Professor C. Lobry for the invitation. The exposed material is partially obtained by re-organizing previous papers, reports and some chapters of a monograph. For this reason, I am indebted with some co-authors, and in particular with Lionel Rosier and Francesca Ceragioli. I also wish to thank the partecipants of the course for many comments and suggestions.

Apart from a few of additions, the list of references in taken from [17] and it is therefore overestimated for the purposes of these notes. Anyway, it can be useful for readers interested in developments.

References

[1] Ancona F. and Bressan A., *Patchy Vector Fields and Asymptotic Stabilization*, ESAIM: Control, Optimisation and Calculus of Variations, **4** (1999), pp. 445-472

[2] Andriano V., Bacciotti A. and Beccari G., *Global Stability and External Stability of Dynamical Systems*, Journal of Nonlinear Analysis, Theory, Methods and Applications, **28** (1997), pp. 1167-1185

[3] Arnold V.I., *Algebraic Unsolvability of the Problem of Lyapunov Stability and the Problem of the Topological Classification of the Singular Points of an Analytic System of Differential Equations*, Funct. Anal. Appl., pp. 173-180 (translated from Funktsional'nyi Analiz i Ego Prilozheniya, **4** (1970), pp. 1-9)

[4] Artstein Z., *Stabilization with Relaxed Controls*, Nonlinear Analysis, Theory, Methods and Applications, **7** (1983), pp. 1163-1173

[5] Arzarello E. and Bacciotti A., *On Stability and Boundedness for Lipschitzian Differential Inclusions: the Converse of Liapunov's Theorems*, Set Valued Analysis, **5** (1998), pp. 377-390

[6] Aubin J.P. and Cellina A., *Differential Inclusions*, Springer Verlag, Berlin, 1984

[7] Aubin J.P. and Frankowska H., *Set Valued Analysis*, Birkhäuser, 1990

[8] Auslander J. and Seibert P., *Prolongations and Stability in Dynamical Systems*, Annales Institut Fourier, Grenoble, **14** (1964), pp. 237-268

[9] Bacciotti A., *Local Stabilizability of Nonlinear Control Systems*, World Scientific, Singapore, 1992

[10] Bacciotti A. and Beccari G., *External Stabilizability by Discontinuous Feedback*, Proceedings of the second Portuguese Conference on Automatic Control, 1996, pp. 495-498

[11] Bacciotti A. and Ceragioli F., *Stability and Stabilization of Discontinuous Systems and Nonsmooth Lyapunov Functions*, ESAIM: Control, Optimisation and Calculus of Variations, **4** (1999), pp. 361-376

[12] Bacciotti A., Ceragioli F., and Mazzi L., *Differential Inclusions and Monotonicity Conditions for Nonsmooth Liapunov Functions*, Set Valued Analysis, **8** (2000), pp. 299-309

[13] Bacciotti A. and Mazzi L., *Some Remarks on k-Asymptotic Stability*, Bollettino U.M.I. (7) 8-A (1994), pp. 353-363

[14] Bacciotti A. and Mazzi L., *A Necessary and Sufficient Condition for Bounded Input Bounded State Stability of Nonlinear Systems*, SIAM Journal on Control and Optimization, to appear

[15] Bacciotti A. and Rosier L., *Liapunov and Lagrange Stability: Inverse Theorems for Discontinuous Systems*, Mathematics of Control, Signals and Systems, **11** (1998), pp. 101-128

[16] Bacciotti A. and Rosier L., *Regularity of Liapunov Functions for Stable Systems*, Systems and Control Letters, **41** (2000), pp. 265-270

[17] Bacciotti A. and Rosier L., *Liapunov Functions and Stability in Control Theory*, Lecture Notes in Control and Information Sciences 267, Springer Verlag, London, 2001

[18] Bernstein D.S., *Nonquadratic Cost and Nonlinear Feedback control*, International Journal of Robust and Nonlinear Control, **3** (1993), pp. 211-229

[19] Bhat S.P. and Bernstein D.S., *Continuous Finite-Time Stabilization of the Translational and Rotational Double Integrators*, IEEE Trans. Automat. Control, **43** (1998), pp. 678-682

[20] Bhat S.P. and Bernstein D.S., *Finite-Time Stability of Continuous Autonomous Systems*, SIAM Journal on Control and Optimization, **38** (2000), pp. 751-766.

[21] Bhatia N.P. and Szëgo G.P., *Stability Theory of Dynamical Systems*, Springer Verlag, Berlin, 1970

[22] Blagodatskikh V.I., *On the Differentiability of Solutions with respect to Initial Conditions*, Differential Equations, pp. 1640-1643 (translated from Differentsial'nye Uravneniya, **9** (1973), pp. 2136-2140)

[23] Blagodatskikh V.I. and Filippov A.F., *Differential Inclusions and Optimal Control*, In *Topology, Ordinary Differential Equations, Dynamical Systems*, Proceedings of Steklov Institute of Mathematics, 1986, pp. 199-259

[24] Bloch A. and Drakunov S., *Stabilization and Tracking in the Nonholonomic Integrator via Sliding Modes*, Systems and Control Letters, **29** (1996), pp. 91-99

[25] Bocharov A.V. et al., *Symmetries and Conservation Laws for Differential Equations of Mathematical Physics*, Translations of Mathematical Monographs **182**, American Mathematical Society, Providence, 1999

[26] Brezis H., *Analyse Fonctionnelle, Théorie et Applications*, Masson (1983).

[27] Brockett R., *Asymptotic Stability and Feedback Stabilization*, in *Differential Geometric Control Theory*, Ed.s Brockett R., Millman R., Sussmann H., Birkhäuser, Boston, 1983

[28] Byrnes C.I. and Isidori A., *New Results and Examples in Nonlinear feedback Stabilization*, Systems and Control Letters, **12** (1989), pp. 437-442

[29] Canudas de Witt C. and Sordalen O.J., *Examples of Piecewise Smooth Stabilization of Driftless NL Systems with less Input than States*, Proceedings of IFAC-NOLCOS 92, Ed. Fliess M., pp. 26-30

[30] Čelikowský S. and Nijmeijer H., *On the Relation Between Local Controllability and Stabilizability for a Class of Nonlinear Systems*, IEEE Transactions on Automatic Control, **42** (1997), pp. 90-94

[31] Ceragioli F., *Some Remarks on Stabilization by means of Discontinuous Feedback*, preprint

[32] Chugunov P.I., *Regular Solution of Differential Inclusions*, Differential Equations **17** (1981), pp. 449-455, translated from Differentsial'nye Uravneniya **17** (1981), pp. 660-668

[33] Clarke F.H., *Optimization and Nonsmooth Analysis*, Wiley and Sons, 1983

[34] Clarke F.H., Ledyaev Yu.S., Sontag E.D. and Subbotin A.I., *Asymptotic Controllability Implies Feedback Stabilization*, IEEE Trans. Automat. Control, **42** (1997) 1394-1407

[35] Clarke F.H., Ledyaev Yu.S. and Stern R.J., *Asymptotic Stability and Smooth Lyapunov Functions*, Journal of Differential Equations, **149** (1998), pp. 69-114

[36] Clarke F.H., Ledyaev Yu.S., Stern R.J. and Wolenski P.R., *Qualitative Properties of Trajectories of Control Systems: a Survey*, Journal of Dynamical and Control Systems, **1** (1995), pp. 1-48

[37] Clarke F.H., Ledyaev Yu.S., Stern R.J. and Wolenski P.R., *Nonsmooth Analysis and Control Theory*, Springer Verlag, New York, 1998

[38] Coddington E.A. and Levinson N., *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955

[39] Conti R., *Linear Differential Equations and Control*, Academic Press, London, 1976

[40] Coron J.M., *Links between Local Controllability and Local Continuous Stabilization*, IFAC Nonlinear Control Systems Design, Bordeaux France, 1992, pp. 165-171

[41] Coron J.M., *Global Asymptotic Stabilization for Controllable Systems without Drift*, Mathematics of Control, Signals, and Systems, **5** (1992) pp. 295-312

[42] Coron J.M., *Stabilizing Time-varying Feedback*, Proceedings of IFAC-NOLCOS 95, A. Krener and D. Mayne, eds., Tahoe

[43] Coron J.M., *Stabilization in Finite Time of Locally Controllable Systems by Means of Continuous Time-varying Feedback Law*, SIAM Journal on Control and Optimization, **33** (1995), pp. 804-833

[44] Coron J.M., *On the Stabilization of Some Nonlinear Control Systems: Results, Tools, and Applications* in *Nonlinear Analysis, Differential Equations and Control*, Ed.s Clarke F.H., Stern R.J., Kluwer, Dordrecht, 1999

[45] Coron J.M. and Praly L., *Adding an Integrator for the Stabilization Problem*, Systems and Control Letters **17** (1991), pp. 89-104.

[46] Coron J.M. and Rosier L., *A Relation Between Continuous Time-Varying and Discontinuous Feedback Stabilization*, Journal of Mathematical Systems, Estimation, and Control, **4** (1994), pp. 67-84

[47] Dayawansa W.P., *Recent Advances in the Stabilization Problem for Low-Dimensional Systems*, in *Proceedings of IFAC Nonlinear Control Systems Design Conference*, Bordeaux, 1992, M. Fliess (ed.), pp. 1-8

[48] Dayawansa W.P. and Martin C.F., *A Remark on a Theorem of Andreini, Bacciotti and Stefani*, Systems and Control Letters, **13** (1989), pp. 363-364

[49] Dayawansa W.P. and Martin C.F., *Asymptotic Stability of Nonlinear Systems with Holomorphic Structure*, Proc. 28th Conf. on Decision and Control, Tampa, FL (1989)

[50] Dayawansa W.P., Martin C.F. and Knowles G., *Asymptotic Stabilization of a Class of Smooth Two Dimensional Systems*, SIAM Journal on Control and Optimization, **28** (1990), pp. 1321-1349

[51] Deimling K., *Multivalued Differential Equations*, de Gruyter, 1992

[52] Doob J.L., *Measure Theory*, Springer Verlag, New York, 1994

[53] Filippov A.F., *Differential Equations with Discontinuous Right-hand Side*, Kluwer Academic Publisher, 1988

[54] Filippov A.F., *Differential Equations with Discontinuous Right-hand Side*, Translations of American Mathematical Society, **42** (1964), pp. 199-231

[55] Filippov A.F., *Classical Solutions of Differential Equations with Multivalued Right-Hand Side*, SIAM J. Control, **5** (1967), pp. 609-621

[56] Filippov A.F., *On Certain Questions in the Theory of Optimal Control*, SIAM Journal of Control, **1** (1962), pp. 76-84

[57] Fradkov A.L., *Speed-Gradient Scheme and its Applications in Adaptive Control*, Automation and Remote Control, **40** (1979), pp. 1333-1342

[58] Frankowska H., *Hamilton-Jacobi Equations: Viscosity Solutions and Generalized Gradients*, Journal of Mathematical Analysis and Applications, **141** (1989), pp. 21-26

[59] Frankowska H., *Optimal Trajectories Associated with a Solution of the Contingent Hamilton-Jacobi Equation*, Applied Mathematics and Optimization, **19** (1989), pp. 291-311

[60] Galeotti M. and Gori F., *Bifurcations and Limit Cycles in a Family of Planar Polynomial Dynamical Systems*, Rend. Sem. Mat. Univers. Politecn. Torino, **46** (1988), pp. 31-58

[61] Gauthier J.P. and Bonnard G., *Stabilisation des systèmes non lineaires* in *Outils et modèles mathématiques pour l'automatique* Vol. I, Ed. Landau, Editions CNRS, 1981, pp. 307-322

[62] Hahn W., *Theory and Applications of Liapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, 1963

[63] Hahn W., *Stability of Motions*, Springer Verlag, Berlin, 1967

[64] Haimo V.T., *Finite Time Controllers*, SIAM Journal on Control and Optimization, **24** (1986), pp. 760-770

[65] Hájek O., *Discontinuous Differential Equations, I*, Journal of Differential Equations, **32** (1979), pp. 149-170

[66] Hartman P., *Ordinary Differential Equations*, Birkhäuser, Boston, 1982

[67] Hautus M.L.J., *Stabilization, Controllability and Observability of Linear Autonomous Systems*, Indagationes Mathematicae, **32** (1970), pp. 448-455

[68] Hermes H., *The Generalized Differential Equation* $\dot{x} \in R(t,x)$, Advances in Mathematics **4** (1970), pp. 149-169

[69] Hermes H., *Homogeneous Coordinates and Continuous Asymptotically Stabilizing Feedback Controls*, in: *Differential Equations, Stability and Controls*, S. Elaydi, Ed., Lecture Notes in Applied Math. **109**, Marcel Dekker, New York (1991), pp. 249-260

[70] Hermes H., *Nilpotent and High-Order Approximations of Vector Field Systems*, SIAM Review, **33** (1991), pp. 238-264

[71] Hermes H., *Asymptotically Stabilizing Feedback Controls and the Nonlinear Regulator Problem*, SIAM Journal on Control and Optimization **29** (1991), pp. 185-196

[72] Hill D. and Moyland P., *The Stability of Nonlinear Dissipative Systems*, IEEE Transaction on Automatic Control **21** (1976), pp. 708-711

[73] Hong Y., Huang J. and Xu Y., *On an Output Feedback Finite-Time Stabilization Problem*, IEEE CDC, Phoenix, 1999, pp. 1302-1307

[74] Il'jašenko J.S., *Analytic Unsolvability of the Stability Problem and the Problem of Topological Classification of the Singular Points of Analytic Systems of Differential Equations*, Math. USSR Sbornik, **28** (1976), pp. 140-152

[75] Isidori A., *Nonlinear Control Systems*, Springer Verlag, 1989

[76] Jurdjevic V., *Geometric Control Theory*, Cambridge University Press, 1997

[77] Jurdjevic V. and Quinn J.P., *Controllability and Stability*, Journal of Differential Equations, **28**, 1978, 381-389

[78] Kang W., *Zubov Theorem and Domain of Attraction for Controlled Dynamical Systems*, *Proceedings of IFAC-NOLCOS Nonlinear Control Systems Design Conference*, Tahoe, 1995, pp. 160-163

[79] Kawski M., *Nilpotent Lie Algebras of Vectorfields*, J. Reine Angew. Math. **388** (1988), pp. 1-17

[80] Kawski M., *Stabilization and Nilpotent Approximations*, Proc. 27th IEEE Conference on Decision & Control, II, (1988), pp. 1244-1248

[81] Kawski M., *Stabilization of Nonlinear Systems in the Plane*, Systems and Control Letters, **12** (1989), pp. 169-175

[82] Kawski M., *Homogeneous Stabilizing Feedback Laws*, Control Theory and Advanced Technology (Tokyo), **6** (1990), pp. 497-516

[83] Kawski M., *Geometric Homogeneity and Applications to Stabilization*, Proceedings of IFAC-NOLCOS 95, A. Krener and D. Mayne, eds., Tahoe, pp. 147-152

[84] Krasowski N.N., *The Converse of the Theorem of K.P. Persidskij on Uniform Stability*, Prikladnaja Matematika I Mehanica, **19** (1955), pp. 273-278 (in russian)

[85] Krasowski N.N., *Stability of Motion*, Stanford University Press, Stanford, 1963

[86] Krener A.J., *Nonlinear Stabilizability and Detectability*, Proceedings of the Int. Symp. MTNS '93, ed.s U. Helmke, R. Mennicken, J. Saurer, Akademie Verlag, pp. 231-250

[87] Krikorian R., *Necessary Conditions for a Holomorphic Dynamical System to Admit the Origin as a Local Attractor*, Systems and Control Letters, **20** (1993), pp. 315-318

[88] Kurzweil J., *On the Invertibility of the First Theorem of Lyapunov Concerning the Stability of Motion* (in russian with english summary), Czechoslovak Mathematical Journal, **80** (1955), pp. 382-398

[89] Kurzweil J., *On the Inversion of Liapunov's Second Theorem on Stability of Motion*, Translations of American Mathematical Society, **24** (1963), pp. 19-77 (originally appeared on Czechoslovak Mathematical Journal, **81** (1956), pp. 217-259)

[90] Kurzweil J. and Vrkoč I., *The Converse Theorems of Lyapunov and Persidskij Concerning the Stability of Motion* (in russian with english summary), Czechoslovak Mathematical Journal, **82** (1957), pp. 254-272

[91] Ledyaev Y.S. and Sontag E.D., *A Lyapunov Characterization of Robust Stabilization*, Nonlinear Analysis, Theory, Methods and Applications, **37** (1999), pp. 813-840

[92] Lin Y., Sontag E.D. and Wang Y., *A Smooth Converse Lyapunov Theorem for Robust Stability*, SIAM Journal on Control and Optimization, **34** (1996), pp. 124-160

[93] Malgrange B., *Ideals of Differentiable Functions*, Oxford Univ. Press, 1966

[94] Massera J.L., *On Lyapounoff's Conditions of Stability*, Annals of Mathematics, **50** (1949), pp. 705-721

[95] Massera J.L., *Contributions to Stability Theory*, Annals of Mathematics, **64** (1956), pp. 182-206

[96] M'Closkey R.T. and Murray R.M., *Non-holonomic Systems and Exponential Convergence: Some Analysis Tools*, in *Proc. IEEE Conf. Decision Control,*, 1993, pp. 943-948

[97] M'Closkey R.T. and Murray R.M., *Exponential Stabilization of Driftless Nonlinear Control Systems Using Homogeneous Feedback*, IEEE Trans. Automat. Control, **42** (1997), pp. 614-628

[98] McShane E.J., *Integration*, Princeton University Press, 1947

[99] Morin P., Pomet J.B. and Samson C., *Design of Homogeneous Time-varying Stabilizing Control Laws for Driftless Controllable Systems via Oscillatory Approximation of Lie Brackets in Closed Loop*, SIAM Journal on Control and Optimization, **38** (1999), pp. 22-49

[100] Paden B.E. and Sastry S.S., *A Calculus for Computing Filippov's Differential Inclusions with Applications to the Variable Structure Control of Robot Manipulators*, IEEE Transactions on Circuits and Systems, **34** (1987), pp. 73-81

[101] Pomet J.B., *Explicit Design of Time-varying Stabilizing Control Laws for a Class of Controllable Systems without Drift*, Systems and Control Letters, **18** (1992), pp. 147-158

[102] Pomet J.B. and Samson C., *Time-Varying Exponential Stabilization of Nonholonomic Systems in Power Form*, Technical Report 2126, INRIA, (1993)

[103] Praly L., *Generalized Weighted Homogeneity and State Dependent Time Scale for Linear Controllable Systems*, Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, 1997

[104] Prieur C., *A Robust Globally Asymptotically Stabilizing Feedback: the Example of the Artstein's Circles*, in *"Nonlinear Control in the Year 2000"* Ed.s Isidori A., Lamnabhi-Lagarrigue F. and Respondek W., Springer Verlag, 2000 pp. 279-300

[105] Rifford L., *Stabilization des systèmes globalement asymptotiquement commandables*, Comptes Rendus de l'Académie des Sciences, Paris, Série I Mathématique, **330** (2000), pp. 211-216

[106] Rifford L., *Existence of Lipschitz and Semiconcave Control Lyapunov Functions*, SIAM Journal on Control and Optimization, to appear

[107] Rifford L., *Nonsmooth Control-Lyapunov Functions; Application to the Integrator Problem*, preprint

[108] Rockafellar R.T. and Wets R.B., *Variational Analysis*, Springer Verlag, Berlin, 1998

[109] Rosier L., *Homogeneous Lyapunov Function for Homogeneous Continuous Vector Field*, Systems and Control Letters, **19** (1992), pp. 467-473

[110] Rosier L., *Inverse of Lyapunov's Second Theorem for Measurable Functions*, Proceedings of IFAC-NOLCOS 92, Ed. Fliess M., pp. 655-660

[111] Rosier L., *Etude de quelques Problèmes de Stabilisation*, Ph. D. Thesis, Ecole Normale Supérieure de Cachan (France), 1993.

[112] Rosier L., *Smooth Lyapunov Functions for Discontinuous Stable Systems*, Set-Valued Analysis, **7** (1999), pp. 375-405

[113] Rosier L. and Sontag E.D., *Remarks Regarding the Gap between Continuous, Lipschitz, and Differentiable Storage Functions for Dissipation Inequalities Appearing in H_∞ Control*, Systems and Control Letters, **41** (2000), pp. 237-249

[114] Rothschild L.P. and Stein E.M., *Hypoelliptic Differential Operators and Nilpotent Groups*, Acta Math., **137** (1976), pp. 247-320

[115] Rouche N., Habets P. and Laloy M., *Stability Theory by Liapunov's Direct Method*, Springer Verlag, 1977

[116] Rudin W., *Real and Complex Analysis*, McGraw Hill, 1970

[117] Rudin W., *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1987.

[118] Ryan E.P., *On Brockett's Condition for Smooth Stabilizability and its Necessity in a Context of Nonsmooth Feedback*, SIAM Journal on Control and Optimization, **32** (1994), pp. 1597-1604

[119] Sansone C. and Conti R., *Nonlinear Differential Equations*, Pergamon, Oxford, 1964

[120] Sepulchre R. and Aeyels D., *Stabilizability does not Imply Homogeneous Stabilizability for Controllable Homogeneous Systems*, SIAM Journal on Control and Optimization, **34** (1996), pp. 1798-1813

[121] Sepulchre R. and Aeyels D., *Homogeneous Lyapunov Functions and Necessary Conditions for Stabilization*, Math. Control Signals Systems, **9** (1996), pp. 34-58

[122] Sepulchre R., Jankovic M. and Kokotovic P., *Constructive Nonlinear Control*, Springer Verlag, London, 1997

[123] Shevitz D. and Paden B., *Lyapunov Stability Theory of Nonsmooth Systems*, IEEE Transactions on Automatic Control, **39** (1994), pp. 1910-1914

[124] Sontag E.D., *Mathematical Control Theory*, Springer Verlag, New York, 1990

[125] Sontag E.D., *A Lyapunov-like Characterization of Asymptotic Controllability*, SIAM Journal on Control and Optimization, **21** (1983), pp. 462-471

[126] Sontag E.D., *Smooth Stabilization Implies Coprime Factorization*, IEEE Transactions on Automatic Control, **34** (1989), pp. 435-443

[127] Sontag E.D., *A "Universal" Construction of Artstein's Theorem on Nonlinear Stabilization*, Systems and Control Letters, **13** (1989), pp. 117-123

[128] Sontag E.D., *Feedback Stabilization of Nonlinear Systems*, in *Robust Control of Linear Systems and Nonlinear Control*, Ed.s Kaashoek M.A., van Schuppen J.H., Ran A.C.M., Birkhäuser 1990, pp. 61-81

[129] Sontag E.D., *On the Input-to-State Stability Property*, European Journal of Control, **1** (1995) pp. 24-36

[130] Sontag E.D., *Comments on Integral Variants of ISS*, IEEE-CDC Conference Proceedings, 1998, pp.

[131] Sontag E.D., *Nonlinear Feedback Stabilization Revisited*, in *Dynamical Systems, Control, Coding, computer Vision*, Ed.s Picci G., Gillian D.S., Birkhäuser, Basel, 1999, pp. 223-262

[132] Sontag E.D., *Stability and Stabilization: Discontinuities and the Effect of Disturbances*, in *Nonlinear Analysis, Differential Equations and Control*, Ed.s Clarke F.H., Stern R.J., Kluwer, Dordrecht, 1999

[133] Sontag E.D., *Clocks and Insensitivity to Small Measurement Errors*, ESAIM: Control, Optimisation and Calculus of Variations, **4** (1999), pp. 537-576

[134] Sontag E.D. and Sussmann H.J., *Remarks on Continuous Feedback*, IEEE-CDC Conference Proceedings, Albuquerque 1980, pp. 916-921

[135] Sontag E.D. and Sussmann H.J., *Further Comments on the Stabilizability of the Angular Velocity of a Rigid Body*, Systems and Control Letters, **12** (1989), pp. 213-217

[136] Sontag E.D. and Sussmann H.J., *Nonsmooth Control-Lyapunov Functions*, IEEE-CDC Conference Proceedings, New Orleans 1995, pp. 2799-2805

[137] Sontag E.D. and Wang Y., *On Characterizations of the Input-to-State Stability Property*, Systems and Control Letters, **24** (1995), pp. 351-359

[138] Sontag E.D. and Wang Y., *New Characterizations of Input-to-State Stability*, IEEE Transaction on Automatic Control, **41** (1996), pp. 1283-1294

[139] Sontag E.D. and Wang Y., *A Notion of Input to Output Stability*, Proceedings of European Control Conference, Bruxels 1997

[140] Sontag E.D. and Wang Y., *Notions of Input to Output Stability*, Systems and Control Letters, **38** (1999), pp. 235-248

[141] Sontag E.D. and Wang Y., *Lyapunov Characterizations of Input to Output Stability*, SIAM Journal on Control and Optimization, to appear

[142] Sussmann H.J., *A General Theorem on Local Controllability*, SIAM Journal on Control and Optimization, **25** (1987), pp. 158-194

[143] Teel R.A. and Praly L., *A Smooth Lyapunov Function from a class.$\mathcal{KL}$ Estimate Involving Two Positive Semidefinite Functions*, ESAIM: Control, Optimisation and Calculus of Variations, **5** (2000), pp. 313-367

[144] Tsinias J., *A Local Stabilization Theorem for Interconnected Systems*, Systems and Control Letters, **18** (1992), pp. 429-434

[145] Tsinias J., *Sufficient Lyapunov-Like Conditions for Stabilizability*, Mathematics of Control, Signals and Systems, **2** (1989), pp. 343-357

[146] van der Schaft A.J., *L_2-gain analysis of Nonlinear Systems and Nonlinear State Feedback H_∞ Control*, IEEE Transaction on Automatic Control **37** (1992), pp. 770-784

[147] Varaiya P.P. and Liu R., *Bounded-input Bounded-output Stability of Nonlinear Time-varying Differential Systems*, SIAM Journal on Control, **4** (1966), pp. 698-704

[148] Vidyasagar M., *Nonlinear Systems Analysis*, Prentice hall, 1993

[149] Willems J.C., *Dissipative Dynamical Systems Part I: General Theory*, Archive for Rational Mechanics and Analysis, **45** (1972), pp. 321-351

[150] Wonham W.M., *Linear Multivariable Control: a Geometric Approach*, Springer Verlag, New York, 1979

[151] Yorke J.A., *Differential Inequalities and Non-Lipschitz Scalar Functions*, Mathematical Systems Theory, **4** (1970), pp. 140-153

[152] Yoshizawa T., *On the Stability of Solutions of a System of Differential Equations*, Memoirs of the College of Sciences, University of Kyoto, Ser. A, **29** (1955), pp. 27-33

[153] Yoshizawa T., *Liapunov's Functions and Boundedness of Solutions*, Funkcialaj Ekvacioj, **2** (1957), pp. 95-142

[154] Yoshizawa T., *Stability Theory by Liapunov's Second Method*, Publications of the Mathematical Society of Japan No. 9, 1966

[155] Zabczyk J., *Some Comment on Stabilizability*, Applied Mathematics and Optimization, **19** (1989), pp. 1-9

[156] Zabczyk J., *Mathematical Control Theory: an Introduction*, Birkhäuser, Boston, 1992

[157] Zubov V.I., *The Methods of Liapunov and their Applications*, Leningrad, 1957

Introduction to Optimal Control Theory

Andrei A. Agrachev*

Steklov Mathematical Institute, Moscow, Russia
and
International School for Advanced Studies (SISSA), Trieste, Italy

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028008

*agrachev@sissa.it

Abstract

These are lecture notes of the introductory course in Optimal Control theory treated from the geometric point of view. Optimal Control Problem is reduced to the study of controls (and corresponding trajectories) leading to the boundary of attainable sets. We discuss Pontryagin Maximum Principle, basic existence results, and apply these tools to concrete simple optimal control problems. Special sections are devoted to the general theory of linear time-optimal problems and linear-quadratic problems.

Contents

1 Optimal control problem **455**
1.1 Problem statement 455
1.2 Reduction to study of attainable sets 456

2 Pontryagin Maximum Principle **458**
2.1 Geometric statement of PMP and discussion 458
2.2 Geometric statement of PMP for free time 462
2.3 PMP for optimal control problems 464

3 Existence of Optimal controls **468**
3.1 Compactness of attainable sets 468
3.2 Time-optimal problem 471
3.3 Relaxations 471

4 Examples of optimal control problems **473**
4.1 The fastest stop of a train at a station 473
4.2 Control of a linear oscillator 476
4.3 The cheapest stop of a train 479
4.4 Control of a linear oscillator with cost 481
4.5 Dubins car 482

5 Linear time-optimal problem **488**
5.1 Problem statement 488
5.2 Geometry of polytopes 489
5.3 Bang-bang theorem 489
5.4 Uniqueness of optimal controls and extremals 492
5.5 Switchings of optimal control 494

6 Linear-quadratic problem **500**
6.1 Problem statement and assumptions 500
6.2 Existence of optimal control 501
6.3 Extremals 504
6.4 Conjugate points 506

References **513**

1 Optimal control problem

1.1 Problem statement

Consider a control system of the form

$$\dot{q} = f_u(q), \qquad q \in M, \quad u \in U \subset \mathbb{R}^m, \tag{1}$$

where M is an open domain in $\mathbb{R}^n$ and U an arbitrary subset of $\mathbb{R}^m$. For the right-hand side of the control system, we suppose that:

$$q \mapsto f_u(q) \text{ is a smooth vector field on } M \text{ for any fixed } u \in U, \tag{2}$$

$$(q,u) \mapsto f_u(q) \text{ is a continuous mapping for } q \in M,\ u \in \overline{U},, \tag{3}$$

$$(q,u) \mapsto \frac{\partial f_u}{\partial q}(q) \text{ is a continuous mapping for } q \in M,\ u \in \overline{U}. \tag{4}$$

Admissible controls are vector-functions:

$$u : t \mapsto u(t) \in U, \quad t \in \mathbb{R}.$$

The set of all admissible controls is denoted by $\mathcal{U}$. In this lectures $\mathcal{U}$ is either the set of all piecewise smooth functions with values in U or the set of all bounded measurable functions with values in U. All results except those of Section 3 are valid for both classes of admissible controls. Substitute such a control $u = u(t)$ for control parameter into system (1), then we obtain a nonautonomous ODE $\dot{q} = f_u(q)$. By the classical Carathéodory's Theorem, for any point $q_0 \in M$, the Cauchy problem

$$\dot{q} = f_u(q), \qquad q(0) = q_0, \tag{5}$$

has a unique solution defined on an interval in $\mathbb{R}$. We will often fix the initial point q_0 and then denote the corresponding solution to problem (5) as $q_u(t)$.

In order to compare admissible controls one with another on a segment $[0, t_1]$, introduce a *cost functional*:

$$J(u) = \int_0^{t_1} \varphi(q_u(t), u(t))\, dt \tag{6}$$

with an integrand

$$\varphi : M \times U \to \mathbb{R}$$

satisfying the same regularity assumptions as the right-hand side f, see (2)–(4).

Take any pair of points $q_0, q_1 \in M$. We consider the following *optimal control problem*:

MINIMIZE THE FUNCTIONAL J AMONG ALL ADMISSIBLE CONTROLS $u = u(t)$, $t \in [0, t_1]$, FOR WHICH THE CORRESPONDING SOLUTION $q_u(t)$ OF CAUCHY PROBLEM (5) SATISFIES THE BOUNDARY CONDITION

$$q_u(t_1) = q_1. \tag{7}$$

We study two types of problems, with fixed t_1 and free t_1. A solution u of this problem is called an *optimal control*, and the corresponding curve $q_u(t)$ is the *optimal trajectory*.

So the optimal control problem is the minimization problem for $J(u)$ with constraints on u given by control system and the fixed endpoints conditions (5), (7). These constraints cannot usually be resolved w.r.t. u, thus solving optimal control problems requires special techniques.

1.2 Reduction to study of attainable sets

Fix an initial point $q_0 \in M$. *Attainable set* of control system (1) for time $t \geq 0$ from q_0 with measurable locally bounded controls is defined as follows:

$$\mathcal{A}_{q_0}(t) = \{q_u(t) \mid u \in \mathcal{U}\}.$$

Similarly, one can consider the attainable sets for time not greater than t:

$$\mathcal{A}^t_{q_0} = \bigcup_{0 \leq \tau \leq t} \mathcal{A}_{q_0}(\tau)$$

and for arbitrary nonnegative time:

$$\mathcal{A}_{q_0} = \bigcup_{0 \leq \tau < \infty} \mathcal{A}_{q_0}(\tau).$$

It turns out that optimal control problems on the state space M can be essentially reduced to the study of attainable sets of some auxiliary control systems on the extended state space

$$\widehat{M} = \mathbb{R} \times M = \{\widehat{q} = (y, q) \mid y \in \mathbb{R},\ q \in M\}.$$

Namely, consider the following extended control system on $\widehat{M}$:

$$\frac{d\,\widehat{q}}{d\,t} = \widehat{f}_u(\widehat{q}), \qquad \widehat{q} \in \widehat{M},\ u \in U, \tag{8}$$

with the right-hand side

$$\widehat{f}_u(\widehat{q}) = \begin{pmatrix} \varphi(q,u) \\ f_u(q) \end{pmatrix}, \qquad q \in M, \quad u \in U,$$

where φ is the integrand of the cost functional J, see (6). Then solutions $\widehat{q}_u(t)$ of the extended system (8) with the initial conditions

$$\widehat{q}_u(0) = \begin{pmatrix} y(0) \\ q(0) \end{pmatrix} = \begin{pmatrix} 0 \\ q_0 \end{pmatrix}$$

are expressed through solutions $q_u(t)$ of the original system (1) as

$$\widehat{q}_u(t) = \begin{pmatrix} J_t(u) \\ q_u(t) \end{pmatrix},$$

where

$$J_t(u) = \int_0^t \varphi(q_u(\tau), u(\tau))\, d\tau.$$

Thus attainable sets of the extended system (8) from the point $(0, q_0)$ have the form

$$\widehat{\mathcal{A}}_{(0,q_0)}(t) = \{(J_t(u), q_u(t)) \mid u \in \mathcal{U}\}.$$

Let $q(t)$, $t \in [0, t_1]$, be an optimal trajectory for the optimal control problem in M. Consider the corresponding trajectory

$$\widehat{q}(t) = \begin{pmatrix} J_t \\ q(t) \end{pmatrix}, \qquad t \in [0, t_1],$$

of the extended control system in $\widehat{M}$. The endpoint $\widehat{q}(t_1)$ must belong to the boundary of the attainable set $\widehat{\mathcal{A}}_{(0,q_0)}(t_1)$; moreover, this set should not intersect the ray

$$\{(y, q_1) \in \widehat{M} \mid y < J_{t_1}\}.$$

Indeed, if there exist points

$$(y, q_1) \in \widehat{\mathcal{A}}_{(0,q_0)}(t_1), \quad y < J_{t_1},$$

then the trajectory of the extended system

$$\widehat{q}'(t) = \begin{pmatrix} J'_t \\ q'(t) \end{pmatrix}$$

that steers $(0, q_0)$ to (y, q_1):

$$\widehat{q}'(0) = \begin{pmatrix} 0 \\ q_0 \end{pmatrix}, \qquad \widehat{q}'(t_1) = \begin{pmatrix} y \\ q_1 \end{pmatrix},$$

gives a trajectory $q'(t)$, $q'(0) = q_0$, $q'(t_1) = q_1$, with a smaller value of the cost functional:

$$J'_{t_1} = y < J_{t_1},$$

a contradiction with optimality of $q(\cdot)$.

So optimal trajectories (more precisely, their lift to the extended state space $\widehat{M}$) must come to the boundary of the attainable set $\widehat{\mathcal{A}}_{(0,q_0)}(t_1)$. In order to find optimal trajectories, we find those coming to the boundary of $\widehat{\mathcal{A}}_{(0,q_0)}(t_1)$, and then select optimal among them. The first step is much more important than the second one, so solving optimal control problems essentially reduces to the study of dynamics of boundary of attainable sets.

2 Pontryagin Maximum Principle

In this section we discuss the fundamental necessary condition of optimality for optimal control problems — Pontryagin Maximum Principle (PMP).

2.1 Geometric statement of PMP and discussion

Consider the optimal control problem stated in Sec. 1.1 for a control system

$$\dot{q} = f_u(q), \qquad q \in M, \quad u \in U \subset \mathbb{R}^m, \tag{9}$$

with the initial condition

$$q(0) = q_0. \tag{10}$$

Define the following family of *Hamiltonians*:

$$h_u(p, q) = \langle p, f_u(q) \rangle, \qquad p \in \mathbb{R}^n, \ q \in M, \ u \in U,$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product.

In Sec. 1.2 we reduced the optimal control problem to the study of boundary of attainable sets. Now we give a necessary optimality condition in this geometric setting.

Theorem 1 (PMP). *Let $\tilde{u}(t)$, $t \in [0, t_1]$, be an admissible control and $\tilde{q}(t) = q_{\tilde{u}}(t)$ the corresponding solution of* (9), (10). *If*

$$\tilde{q}(t_1) \in \partial \mathcal{A}_{q_0}(t_1),$$

then there exists a Lipschitz vector-function

$$p(t) \in \mathbb{R}^n, \qquad 0 \le t \le t_1,$$

such that

$$p(t) \neq 0, \tag{11}$$

$$\dot{p}(t) = -\frac{\partial h_{\tilde{u}(t)}}{\partial q}(p(t), \tilde{q}(t)), \tag{12}$$

$$h_{\tilde{u}(t)}(p(t), \tilde{q}(t)) = \max_{u \in U} h_u(p(t), \tilde{q}(t)) \tag{13}$$

for almost all $t \in [0, t_1]$.

If $u(t)$ is an admissible control, $q(t)$ the corresponding solution of (9), (10), and $p(t)$ a Lipschitz vector-function such that conditions (11)–(13) hold, then the triple $(u(t), p(t), q(t))$ is said to satisfy PMP. In this case $(p(t), (q(t))$ is often called an *extremal*, and $q(t)$ is called an *extremal trajectory*.

Remark. If $(u(t), p(t), q(t))$ satisfies PMP, then

$$h_{u(t)}(p(t), q(t)) = \text{const}, \qquad t \in [0, t_1]. \tag{14}$$

We skip a rather technical proof of the Pontryagin Maximum Principle but try to clarify its statement.

First we give an euristic explanation of the way the vector-function $p(t)$ appears naturally in the study of trajectories coming to boundary of the attainable set. Indeed, let

$$q_1 = \tilde{q}(t_1) \in \partial \mathcal{A}_{q_0}(t_1).$$

Consider a local convex approximation of the attainable set $\mathcal{A}_{q_0}(t_1)$ in the neighborhood of the point q_1, a convex cone with the vertix q_1. This convex cone has a hyperplane of support at q_1 determined by its normal vector $p(t_1)$ (the vector $p(t_1)$ is actually an analog of classical Lagrange multipliers).

In order to construct the whole curve $p(t)$, $t \in [0, t_1]$, consider the flow $P_\tau^{t_1} : M \to M$ generated by the control $\tilde{u}(\cdot)$:

$$P_\tau^{t_1} : q(\tau) \mapsto q(t_1), \qquad \tau \in [0, t_1],$$

where $\dot{q}(t) = f_{\tilde{u}(t)}(q(t))$, $t \in [\tau, t_1]$. It is easy to see that

$$P_\tau^{t_1}(\mathcal{A}_{q_0}(\tau)) \subset \mathcal{A}_{q_0}(t_1), \qquad \tau \in [0, t_1].$$

Indeed, if a point $q \in \mathcal{A}_{q_0}(\tau)$ is reachable from q_0 by a control $u(t)$, $t \in [0, \tau]$, then the point $P_\tau^{t_1}(q)$ is reachable from q_0 by the control

$$v(t) = \begin{cases} u(t), & t \in [0, \tau], \\ \tilde{u}(t), & t \in [\tau, t_1]. \end{cases}$$

Further, the flow $P_\tau^{t_1}$ satisfies the condition

$$P_\tau^{t_1}(\tilde{q}(\tau)) = \tilde{q}(t_1) = q_1, \qquad \tau \in [0, t_1].$$

Thus if $\tilde{q}(\tau) \in \operatorname{int} \mathcal{A}_{q_0}(\tau)$, then $q_1 \in \operatorname{int} \mathcal{A}_{q_0}(t_1)$. By contradiction, we obtain

$$\tilde{q}(\tau) \in \partial \mathcal{A}_{q_0}(\tau), \qquad \tau \in [0, t_1].$$

Consequently, we can find a hyperplane of support to the convex approximation of $\mathcal{A}_\tau(q_0)$ and the corresponding normal vector at any instant τ:

$$p(\tau) \in \mathbb{R}^n \setminus 0, \qquad \tau \in [0, t_1].$$

The normal vectors $p(t)$ are defined up to nonzero factors. They can be renormalized so that satisfy the equation $\dot{p}(t) = -\frac{\partial h_{\tilde{u}(t)}}{\partial q}(p(t), \tilde{q}(t))$

So the vector-function $p(t)$ in Pontryagin Maximum Principle appears naturally from hyperplanes of support to convex approximations of attainable sets.

Now we show the power of PMP by the following statement.

Proposition 1. *Assume that the maximized Hamiltonian of PMP*

$$H(p, q) = \max_{u \in U} h_u(p, q), \qquad p \in \mathbb{R}^n, \ q \in M,$$

is defined and C^2-smooth on $(\mathbb{R}^n \setminus 0) \times M$.

If $(\tilde{u}(t), p(t), q(t))$, $t \in [0, t_1]$, *satisfies PMP, then*

$$\begin{cases} \dot{p}(t) = -\frac{\partial H}{\partial q}(p(t), q(t) \\ \dot{q}(t) = \frac{\partial H}{\partial p}(p(t), q(t)) \end{cases} \qquad t \in [0, t_1]. \tag{15}$$

Conversely, if a Lipshitzian vector-function $(p(t), q(t)) \in (\mathbb{R}^n \setminus 0) \times M$ *is a solution to the Hamiltonian system* (15)*, then one can choose an admissible control* $\tilde{u}(t)$, $t \in [0, t_1]$, *such that* $(\tilde{u}(t), p(t), q(t))$ *satisfy PMP.*

That is, in the favorable case when the maximized Hamiltonian H is C^2-smooth, PMP reduces the problem to the study of solutions to just one Hamiltonian system (15). From the point of view of dimension, this reduction is the best one we can expect. Indeed, for a full-dimensional attainable set ($\dim \mathcal{A}_{q_0}(t_1) = n$) we have $\dim \partial \mathcal{A}_{q_0}(t_1) = n - 1$, i.e. we need an $(n-1)$-parameter family of curves to describe the boundary $\partial \mathcal{A}_{q_0}(t_1)$. On the other hand, the family of solutions to Hamiltonian system (15) with the initial condition $\pi(\lambda_0) = q_0$ is n-dimensional. Taking into account that the Hamiltonian H is homogeneous: $H(cp, q) = cH(p, q)$, $c > 0$; thus $(p(t), q(t))$ is a solution to Hamiltonian system (15) if and only if $(cp(t), q(t))$ is a solution to the same system and we obtain the required $(n-1)$-dimensional family of curves.

Now we prove Proposition 1.

Proof. Set $\lambda = (p, q)$, $\lambda_t = (p(t), q(t))$. We are going to show that if an admissible control $\tilde{u}(t)$ satisfies the maximality condition $h_{\tilde{u}(t)}(p(t), q(t)) = \max_{u \in U} h_u(p(t), q(t))$, then

$$\frac{\partial h_{\tilde{u}(t)}}{\partial \lambda}(\lambda_t) = \frac{dH}{d\lambda}(\lambda_t), \qquad t \in [0, t_1]. \tag{16}$$

In particular,

$$\frac{\partial H}{\partial p}(p(t), q(t)) = \frac{\partial h_{\tilde{u}(t)}}{\partial p}(p(t), q(t)) = f_{\tilde{u}(t)}(q(t)).$$

By definition of the maximized Hamiltonian H,

$$H(\lambda) - h_{\tilde{u}(t)}(\lambda) \geq 0 \qquad \lambda \in T^*M, \quad t \in [0, t_1].$$

On the other hand, by the maximality condition of PMP (13), along the extremal λ_t this inequality turns into equality:

$$H(\lambda_t) - h_{\tilde{u}(t)}(\lambda_t) = 0, \qquad t \in [0, t_1].$$

That is why

$$\frac{dH}{d\lambda}\lambda_t = \frac{\partial h_{\tilde{u}(t)}}{\partial \lambda}(\lambda_t), \qquad t \in [0, t_1].$$

But the right-hand side of the Hamiltonian system is obtained from differential of the Hamiltonian by a standard linear transformation, thus equality (16) follows.

Conversely, let $\lambda_t = (p(t), q(t))$, $p(t) \neq 0$, be a trajectory of the Hamiltonian system (14). One can show that it is possible to choose an admissible control $\tilde{u}(t)$ that realizes maximum along λ_t:

$$H(\lambda_t) = h_{\tilde{u}(t)}(\lambda_t) = \max_{u \in U} h_u(\lambda_t).$$

As we have shown above, then there holds equality (16). So $(\tilde{u}(t), \lambda_t)$ satisfies PMP. □

2.2 Geometric statement of PMP for free time

In the previous section we discussed Pontryagin Maximum Principle for the case of fixed terminal time t_1. Now we consider the case of free t_1.

Theorem 2. *Let $\tilde{u}(\cdot)$ be an admissible control for control system (9) and $\tilde{q}(t) = q_{\tilde{u}}(t)$ the corresponding solution of (9), (10). If*

$$\tilde{q}(t_1) \in \partial \left(\cup_{|t-t_1|<\varepsilon} \mathcal{A}_{q_0}(t) \right)$$

for some $t_1 > 0$ and $\varepsilon \in (0, t_1)$, then there exists a Lipschitz vector-function

$$\lambda_t = (p(t), \tilde{q}(t)) \in (\mathbb{R}^n \setminus 0) \times M, \quad \lambda_t \neq 0, \qquad 0 \leq t \leq t_1,$$

such that

$$\begin{aligned} &\dot{p}(t) = -\frac{\partial h_{\tilde{u}(t)}}{\partial p}(\lambda_t), \\ &h_{\tilde{u}(t)}(\lambda_t) = \max_{u \in U} h_u(\lambda_t), \\ &h_{\tilde{u}(t)}(\lambda_t) = 0 \end{aligned} \tag{17}$$

for almost all $t \in [0, t_1]$.

Remark. In problems with free time, there appears one more variable, the terminal time t_1. In order to eliminate it, we have one additional condition — equality (17). This condition is indeed scalar since the previous two equalities imply that $h_{\tilde{u}(t)}(\lambda_t) = \text{const}$, see remark after formulation of Theorem 1.

Proof. We reduce the case of free time to the case of fixed time by extension of the control system via substitution of time. Admissible trajectories of the extended system are reparametrized admissible trajectories of the initial system (the positive direction of time on trajectories is preserved).

Let a new time be a smooth function

$$\varphi : \mathbb{R} \to \mathbb{R}, \quad \dot{\varphi} > 0.$$

We find an ODE for a reparametrized trajectory:

$$\frac{d}{dt} q_u(\varphi(t)) = \dot{\varphi}(t) f_{u(\varphi(t))}(q_u(\varphi(t))),$$

so the required equation is

$$\dot{q} = \dot{\varphi}(t) f_{u(\varphi(t))}(q).$$

Now consider along with the initial control system

$$\dot{q} = f_u(q), \qquad u \in U,$$

an extended system of the form

$$\dot{q} = v f_u(q), \qquad u \in U, \ |v - 1| < \delta, \tag{18}$$

where $\delta = \varepsilon / t_1 \in (0, 1)$. Admissible controls of the new system are

$$w(t) = (v(t), u(t)),$$

and the reference control corresponding to the control $\tilde{u}(\cdot)$ of the initial system is

$$\tilde{w}(t) = (1, \tilde{u}(t)).$$

It is easy to see that since $\tilde{q}(t_1) \in \partial \left(\cup_{|t - t_1| < \varepsilon} \mathcal{A}_{q_0}(t) \right)$, then the trajectory of the new system through the point q_0 corresponding to the control $\tilde{w}(\cdot)$ comes at the moment t_1 to the boundary of the attainable set of the new system for time t_1. Thus $\tilde{w}(t)$ satisfies PMP with fixed time. We apply Theorem 1 to the new system (18). The Hamiltonian for the new system is $v h_u(\lambda)$. Then the maximality condition (13) reads

$$1 \cdot h_{\tilde{u}(t)}(\lambda_t) = \max_{u \in U, \ |v-1| < \delta} v h_u(\lambda_t).$$

We take $u = \tilde{u}(t)$ under the maximum and obtain

$$h_{\tilde{u}(t)}(\lambda_t) = 0,$$

then we restrict the maximum to the set $v = 1$ and come to

$$h_{\tilde{u}(t)}(\lambda_t) = \max_{u \in U} h_u(\lambda_t).$$

The Hamiltonian systems along $\tilde{w}(\cdot)$ and $\tilde{u}(\cdot)$ coincide one with another, thus the proposition follows. □

2.3 PMP for optimal control problems

Now we apply PMP in geometric form to optimal control problems, starting from problems with fixed time.

For a control system

$$\dot{q} = f_u(q), \qquad q \in M, \quad u \in U, \tag{19}$$

with the boundary conditions

$$q(0) = q_0, \quad q(t_1) = q_1, \quad q_0, q_1 \in M \text{ fixed}, \tag{20}$$

$$t_1 > 0 \text{ fixed}, \tag{21}$$

and the cost functional

$$J(u) = \int_0^{t_1} \varphi(q_u(t), u(t))\, dt \tag{22}$$

we consider the optimal control problem

$$J(u) \to \min. \tag{23}$$

We transform the problem as in Sec. 1.2. We extend the state space:

$$\widehat{q} = \begin{pmatrix} y \\ q \end{pmatrix} \in \mathbb{R} \times M,$$

define the extended vector field

$$\widehat{f}_u(q) = \begin{pmatrix} \varphi(q,u) \\ f_u(q) \end{pmatrix},$$

and come to the new control system:

$$\frac{d\,\widehat{q}}{d\,t} = \widehat{f}_u(q) \quad \Leftrightarrow \quad \begin{cases} \dot{y} = \varphi(q,u), \\ \dot{q} = f_u(q) \end{cases} \tag{24}$$

with the boundary conditions

$$\widehat{q}(0) = \widehat{q}_0 = \begin{pmatrix} 0 \\ q_0 \end{pmatrix}, \qquad \widehat{q}(t_1) = \begin{pmatrix} J(u) \\ q_1 \end{pmatrix}.$$

If a control $\tilde{u}(\cdot)$ is optimal for problem (19)–(23), then the trajectory $\widehat{q}_{\tilde{u}}(t)$ of the extended system (24) starting from $\widehat{q}_0$ satisfies the condition

$$\widehat{q}_{\tilde{u}}(t_1) \in \partial\widehat{\mathcal{A}}_{\widehat{q}_0}(t_1),$$

where $\widehat{\mathcal{A}}_{\widehat{q}_0}(t_1)$ is the attainable set of system (24) from the point $\widehat{q}_0$ for time t_1. So we can apply Theorem 1.

But the geometric form of PMP applied to the extended system (24) does not distinguish minimum and maximum of the cost $J(u)$. In order to have conditions valid only for minimum, we introduce a new control parameter v and consider a new system of the form

$$\begin{cases} \dot{y} = \varphi(q,u) + v, \\ \dot{q} = f_u(q), \end{cases} \qquad v \geq 0, \quad u \in U. \tag{25}$$

Now the trajectory of system (25) corresponding to the controls $\tilde{v}(t) \equiv 0$, $\tilde{u}(t)$, comes to the boundary of the attainable set of this system at time t_1. We apply Theorem 1 to system (25). The Hamiltonian function for system (25) has the form

$$\widehat{h}_{(v,u)}(\nu,p,q) = \langle p, f_u(q)\rangle + \nu(\varphi + v),$$

and the Hamiltonian system of PMP is

$$\begin{cases} \dot{\nu} = \frac{\partial\widehat{h}}{\partial\,y} = 0, \\ \dot{y} = \varphi(q,u) + v, \\ \dot{p} = -\frac{\partial h_{\tilde{u}(t)}}{\partial p}(\nu,\lambda_t), \\ \dot{q} = f_{\tilde{u}(t)}(q(t)), \end{cases} \tag{26}$$

where

$$h_u(\nu,p,q) = \langle p, f_u(q)\rangle + \nu\varphi(q,u).$$

The first of equations (26) means that

$$\nu = \text{const}$$

along the reference trajectory.

The maximality condition has the form

$$\langle p(t), f_{\tilde{u}(t)}(\tilde{q}(t))\rangle + \nu\varphi(\tilde{q}(t), \tilde{u}(t)) = \max_{u\in U,\ v\geq 0} \left(\langle p(t), f_u(\tilde{q}(t))\rangle + \nu\varphi(\tilde{q}(t), u) + \nu v\right).$$

Since the previous maximum is attained, we have

$$\nu \leq 0,$$

thus $v = 0$ and

$$\langle p(t), f_{\tilde{u}(t)}(\tilde{q}(t))\rangle + \nu\varphi(\tilde{q}(t), \tilde{u}(t)) = \max_{u\in U} \left(\langle p(t), f_u(\tilde{q}(t))\rangle + \nu\varphi(\tilde{q}(t), u)\right).$$

So we obtain the following result.

Theorem 3. *Let $\tilde{u}(t)$, $\tilde{q}(t)$, $t \in [0, t_1]$, be an optimal control and the corresponding trajectory for problem (19)–(23):*

$$J(\tilde{u}) = \min\{J(u) \mid q_u(t_1) = q_1\}.$$

Define a Hamiltonian function

$$h_u^\nu(p, q) = \langle p, f_u\rangle + \nu\varphi(q, u), \qquad (p, q) \in \mathbb{R}^n \times M, \quad u \in U, \quad \nu \in \mathbb{R}.$$

Then there exists a nontrivial pair:

$$(\nu, p(t)) \neq 0, \qquad \nu \in \mathbb{R}, \quad p(t) \in \mathbb{R}^n,$$

such that the following conditions hold:

$$\dot{p}(t) = -\frac{\partial h^\nu_{\tilde{u}(t)}}{\partial q}(p(t), \tilde{q}(t)),$$
$$h^\nu_{\tilde{u}(t)}(p(t), \tilde{q}(t)) = \max_{u\in U} h^\nu_u(p(t), \tilde{q}(t)) \quad \forall\ a.e.\ t \in [0, t_1],$$
$$\nu \leq 0.$$

Remarks. (1) If we have a maximization problem instead of minimization problem (23), then the preceding inequality for ν should be reversed:

$$\nu \geq 0.$$

(2) For the problem with free time t_1: (19), (20), (22), (23), necessary optimality conditions of PMP are the same as in Theorem 3 plus one additional scalar equality $h^\nu_{\tilde{u}(t)}(p(t), \tilde{q}(t)) \equiv 0$.

There are two distinct possibilities for the constant parameter ν in Theorem 3:

(a) if $\nu \neq 0$, then $\lambda_t = (p(t), \tilde{q}(t))$ is called a *normal extremal*. Since the pair (ν, λ_t) can be multiplied by any positive number, we can normalize $\nu < 0$ and assume that $\nu = -1$ in the normal case;

(b) if $\nu = 0$, then λ_t is an *abnormal extremal*.

So we can always assume that $\nu = -1$ or 0.

Now consider the time-optimal problem:

$$\begin{aligned} &\dot{q} = f_u(q), \qquad q \in M, \quad u \in U, \\ &q(0) = q_0, \quad q(t_1) = q_1, \quad q_0, q_1 \text{ fixed}, \\ &t_1 = \int_0^{t_1} 1\, dt \to \min. \end{aligned}$$

For the time-optimal problem, Pontryagin Maximum Principle takes the following form.

Corollary 1. *Let an admissible control $\tilde{u}(t)$, $t \in [0, t_1]$, be time-optimal. Define a Hamiltonian function*

$$h_u(p, q) = \langle p, f_u(q) \rangle, \qquad p \in \mathbb{R}^n, \quad u \in U.$$

Then there exists a Lipschitz vector-function

$$p(t) \in \mathbb{R}^n, \quad p(t) \neq 0, \qquad t \in [0, t_1],$$

such that the following conditions hold for almost all $t \in [0, t_1]$:

$$\begin{aligned} &\dot{p}(t) = -\frac{\partial h_{\tilde{u}(t)}}{\partial q}(p(t), \tilde{q}(t)), \\ &h_{\tilde{u}(t)}(p(t), \tilde{q}(t)) = \max_{u \in U} h_u(p(t), \tilde{q}(t)), \\ &h_{\tilde{u}(t)}(p(t), \tilde{q}(t)) \geq 0. \end{aligned} \tag{27}$$

Proof. Apply Theorem 3 and the second remark after it, taking $\varphi \equiv 1$. Then the Hamiltonian system and the maximality condition follow. Inequality (27) is equivalent to conditions $h_{\tilde{u}(t)}(p(t), \tilde{q}(t)) + \nu = 0$ and $\nu \leq 0$.

The inequality $p(t) \neq 0$ is obtained as follows: if $p(t) = 0$, then $h_{\tilde{u}(t)}(p(t), \tilde{q}(t)) = 0$, thus $\nu = 0$. But the pair $(\nu, p(t))$ must be nontrivial, consequently, $p(t) \neq 0$. □

In all previous problems, boundary conditions for a trajectory $q(t)$ were of the form $q(0) = q_0$, $q(t_1) = q_1$. Consider more general boundary conditions:

$$q(0) \in N_0, \qquad q_u(t_1) \in N_1,$$

where $N_0, N_1 \subset M$ are smooth submanifolds. It is easy to see that optimal solutions in the new problem are optimal for the problem with fixed $q(0)$, $q(t_1)$ as well. So all conditions of Pontryagin Maximum Principle should be satisfied. In addition to them, we need $(\dim N_1 + \dim N_2)$ extra conditions for the initial and terminal points. They are called *transversality conditions*: the adjoint covector $p(t)$ must be orthogonal to the submanifold N_0 at $q(0)$ and to the submanifold N_1 at $q(t_1)$ at the moments of time 0 and t_1 respectively:

$$\begin{aligned} p(0) \perp T_{q(0)}N_0 \quad &\Leftrightarrow \quad \langle p(0), T_{q(0)}N_0 \rangle = 0, \\ p(t_1) \perp T_{q(t_1)}N_1 \quad &\Leftrightarrow \quad \langle p(t_1), T_{q_1}N_1 \rangle = 0, \end{aligned}$$

where T_qN is the tangent space to the submanifold N at the point $q \in N$.

3 Existence of Optimal controls

In this section, $\mathcal{U}$ is the set all of measurable bounded vector-functions $t \mapsto u(t)$ with values in U.

3.1 Compactness of attainable sets

Due to the reduction of optimal control problems to the study of attainable sets, existence of optimal solutions to these problems is reduced to compactness of attainable sets.

For control system (1), sufficient conditions for compactness of the attainable sets $\mathcal{A}_{q_0}(t)$ for time t and $\mathcal{A}^t_{q_0}$ for time not greater than t are given in the following proposition.

Theorem 4 (Filippov). *Let the space of control parameters $U \Subset \mathbb{R}^m$ be compact. Let there exist a compact $K \Subset M$ such that $f_u(q) = 0$ for $q \notin K$, $u \in U$. Moreover, let the velocity sets*

$$f_U(q) = \{f_u(q) \mid u \in U\} \subset T_qM, \qquad q \in M,$$

be convex. Then the attainable sets $\mathcal{A}_{q_0}(t)$ and $\mathcal{A}^t_{q_0}$ are compact for all $q_0 \in M$, $t > 0$.

Remark. The condition of convexity of the velocity sets $f_U(q)$ is natural since trajectories of the ODE

$$\dot{q} = \alpha(t) f_{u_1}(q) + (1 - \alpha(t)) f_{u_2}(q), \qquad 0 \le \alpha(t) \le 1,$$

can be uniformly approximated by the "fast switching" trajectories of the systems of the form

$$\dot{q} = f_v(q), \quad \text{where} \quad v(t) \in \{u_1(t),\, u_2(t)\}.$$

Now we give a sketch of the proof of Theorem 4.

Proof. Notice first of all that all nonautonomous vector fields $f_u(q)$ with admissible controls u have a common compact support, thus are complete. Further, under hypotheses of the theorem, velocities $f_u(q)$, $q \in M$, $u \in U$, are uniformly bounded, thus all trajectories $q(t)$ of control system (1) starting at q_0 are Lipschitzian with the same Lipschitz constant. Thus the set of admissible trajectories is precompact in the topology of uniform convergence. For any sequence $q_n(t)$ of admissible trajectories:

$$\dot{q}_n(t) = f_{u_n}(q_n(t)), \qquad 0 \le t \le t_1, \quad q_n(0) = q_0,$$

there exists a uniformly converging subsequence, we denote it again by $q_n(t)$:

$$q_n(\cdot) \to q(\cdot) \text{ in } C[0, t_1] \text{ as } n \to \infty.$$

Now we show that $q(t)$ is an admissible trajectory of control system (1).

Fix a sufficiently small $\varepsilon > 0$. Then

$$\frac{1}{\varepsilon}(q_n(t+\varepsilon) - q_n(t)) = \frac{1}{\varepsilon}\int_t^{t+\varepsilon} f_{u_n}(q_n(\tau))\, d\tau$$
$$\in \operatorname{conv} \bigcup_{\tau \in [t, t+\varepsilon]} f_U(q_n(\tau)) \subset \operatorname{conv} \bigcup_{q \in O_{q(t)}(c\varepsilon)} f_U(q)$$

where c is the doubled Lipschitz constant of admissible trajectories. Then we pass to the limit $n \to \infty$ and obtain

$$\frac{1}{\varepsilon}(q(t+\varepsilon) - q(t)) \in \operatorname{conv} \bigcup_{q \in O_{q(t)}(c\varepsilon)} f_U(q).$$

Now let $\varepsilon \to 0$. If t is a point of differentiability of $q(t)$, then

$$\dot{q}(t) \in f_U(q)$$

since $f_U(q)$ is convex.

In order to show that $q(t)$ is an admissible trajectory of control system (1), we should find a measurable selection $u(t) \in U$ that generates $q(t)$. We do this via the lexicographic order on the set $U = \{(u_1, \dots, u_m)\} \subset \mathbb{R}^m$.

The set

$$V_t = \{v \in U \mid \dot{q}(t) = f_v(q(t))\}$$

is a compact subset of U, thus of $\mathbb{R}^m$. There exists a vector $v_{\min}(t) \in V_t$ minimal in the sense of lexicographic order: to find $v_{\min}(t)$, we minimize the first coordinate v_1 among all $v = (v_1, \dots, v_m) \in V_t$, then minimize the second coordinate v_2 on the compact set found at the first step, etc. The control $u(t) = v_{\min}(t)$ is measurable, thus $q(t)$ is an admissible solution of control system (1).

The proof of compactness of the attainable set $\mathcal{A}_{q_0}(t)$ is complete. Compactness of $\mathcal{A}^t_{q_0}$ is proved by a slightly modified argument. □

Remark. In Filippov's theorem, the hypothesis of common compact support of the vector fields in the right-hand side is essential to ensure the uniform boundedness of velocities and comple teness of vector fields. In the domain M, sufficient conditions for completeness of a vector field cannot be given in terms of boundedness of the vector field and its derivatives: a constant vector field is not complete in a bounded domain in $\mathbb{R}^n$. Nevertheless, one can prove compactness of attainable sets for many systems without the assumption of common compact support. If for such a system we have a priori bounds on solutions, then we can multiply its right-hand side by a cut-off function, and obtain a system with vector fields having compact support. We can apply Filippov's theorem to the new system. Since trajectories of the initial and new systems coincide in a domain of interest for us, we obtain a conclusion on compactness of attainable sets for the initial system.

For control systems on $M = \mathbb{R}^n$, there exist well-known sufficient conditions for completeness of vector fields: if the right-hand side grows at infinity not faster than a linear field, i.e.,

$$|f_u(x)| \leq C(1 + |x|), \qquad x \in \mathbb{R}^n, \quad u \in U, \tag{28}$$

for some constant C, then the nonautonomous vector fields $f_u(x)$ are complete (here $|x| = \sqrt{x_1^2 + \cdots + x_n^2}$ is the norm of a point $x = (x_1, \dots, x_n) \in \mathbb{R}^n$).

These conditions provide an a priori bound for solutions: any solution $x(t)$ of the control system

$$\dot{x} = f_u(x), \qquad x \in \mathbb{R}^n, \quad u \in U, \tag{29}$$

with the right-hand side satisfying (28) admits the bound

$$|x(t)| \le e^{2Ct}\,(|x(0)| + 1)\,, \qquad t \ge 0.$$

So Filippov's theorem plus the previous remark imply the following sufficient condition for compactness of attainable sets for systems in $\mathbb{R}^n$.

Corollary 2. *Let system* (29) *have a compact space of control parameters* $U \Subset \mathbb{R}^m$ *and convex velocity sets* $f_U(x)$, $x \in \mathbb{R}^n$. *Suppose moreover that the right-hand side of the system satisfies a bound of the form* (28). *Then the attainable sets* $\mathcal{A}_{x_0}(t)$ *and* $\mathbb{C}A^t_{x_0}$ *are compact for all* $x_0 \in \mathbb{R}^n$, $t > 0$.

3.2 Time-optimal problem

Given a pair of points $q_0 \in M$, $q_1 \in \mathcal{A}_{q_0}$, the *time-optimal problem* consists in minimizing the time of motion from q_0 to q_1 via admissible controls of control system (1):

$$\min_u \{t_1 \mid q_u(t_1) = q_1\}. \tag{30}$$

That is, we consider the optimal control problem described in Sec. 1.1 with the integrand $\varphi(q, u) \equiv 1$ and free terminal time t_1.

Reduction of optimal control problems to the study of attainable sets and Filippov's Theorem yield the following existence result.

Corollary 3. *Under hypotheses of Theorem* 4, *time-optimal problem* (1), (30) *has a solution for any points* $q_0 \in M$, $q_1 \in \mathcal{A}_{q_0}$.

3.3 Relaxations

Consider a control system of the form (1) with a compact set of control parameters U. There is a standard procedure called *relaxation* of control system (1), which extends the velocity set $f_U(q)$ of this system to its convex hull $\operatorname{conv} f_U(q)$.

Recall that the *convex hull* $\operatorname{conv} S$ of a subset S of a linear space is the minimal convex set that contains S. A constructive description of convex

hull is given by the following classical proposition: any point in the convex hull of a set S in the n-dimensional linear space is contained in the convex hull of some $n+1$ points in S.

Lemma 1 (Carathéodory). *For any subset $S \subset \mathbb{R}^n$, its convex hull has the form*

$$\operatorname{conv} S = \left\{ \sum_{i=0}^{n} \alpha_i x_i \mid x_i \in S,\ \alpha_i \geq 0,\ \sum_{i=0}^{n} \alpha_i = 1 \right\}.$$

Relaxation of control system (1) is constructed as follows. Let $n = \dim M$ be dimension of the state space. The set of control parameters of the relaxed system is

$$V = \Delta^n \times \underbrace{U \times \cdots \times U}_{n+1 \text{ times}},$$

where

$$\Delta^n = \left\{ (\alpha_0, \dots, \alpha_n) \mid \alpha_i \geq 0,\ \sum_{i=0}^{n} \alpha_i = 1 \right\} \subset \mathbb{R}^{n+1}$$

is the standard n-dimensional simplex. So the control parameter of the new system has the form

$$v = (\alpha, u_0, \dots, u_n) \in V, \qquad \alpha = (\alpha_0, \dots, \alpha_n) \in \Delta^n,\ u_i \in U.$$

If U is compact, then V is compact as well.

The *relaxed system* is

$$\dot{q} = g_v(q) = \sum_{i=0}^{n} \alpha_i f_{u_i}(q), \qquad v = (\alpha, u_0, \dots, u_n) \in V, \quad q \in M. \tag{31}$$

By Carathéodory's lemma, the velocity set $g_V(q)$ of system (31) is convex, moreover,

$$g_V(q) = \operatorname{conv} f_U(q).$$

If all vector fields in the right-hand side of (31) have a common compact support, we obtain by Filippov's theorem that attainable sets for the relaxed system are compact. Any trajectory of relaxed system (31) can be uniformly approximated by families of trajectories of initial system (1). Thus attainable sets of the relaxed system coincide with closure of attainable sets of the initial system.

4 Examples of optimal control problems

In this chapter we apply Pontryagin Maximum Principle to solve concrete optimal control problems.

4.1 The fastest stop of a train at a station

Consider a train moving on a railway. The problem is to drive the train to a station and stop it there in a minimal time.

Describe position of the train by a coordinate x_1 on the real line; the origin $0 \in \mathbb{R}$ corresponds to the station. Assume that the train moves without friction, and we can control acceleration of the train by applying a force bounded by absolute value. Using rescaling if necessary, we can assume that absolute value of acceleration is bounded by 1.

We obtain the control system

$$\ddot{x}_1 = u, \qquad x_1 \in \mathbb{R}, \quad |u| \leq 1,$$

or, in the standard form,

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = u, \end{cases} \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2, \quad |u| \leq 1.$$

The time-optimal control problem is

$$\begin{aligned} &x(0) = x^0, \quad x(t_1) = 0, \\ &t_1 \to \min. \end{aligned}$$

First we verify existence of optimal controls by Filippov's theorem. The set of control parameters $U = [-1, 1]$ is compact, the vector fields in the right-hand side

$$f(x, u) = \begin{pmatrix} x_2 \\ u \end{pmatrix}, \qquad |u| \leq 1,$$

are linear, and the set of admissible velocities at a point

$$f(x, U) = \{f(x, u) \mid |u| \leq 1\}$$

is convex. By Corollary 3, the time-optimal control problem has a solution if the origin $0 \in \mathbb{R}^2$ is attainable from the initial point x^0. We will show that any point $x \in \mathbb{R}^2$ can be connected with the origin by an extremal curve.

Now we apply Pontryagin Maximum Principle. Introduce canonical coordinates:

$$M = \mathbb{R}^2 = \left\{ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right\}.$$

So x is a specialization of the state variable q of previous sections. An adjoint vector (a specialization of the vector p of Sec. 2) is denoted by ξ and presented as a row: $\xi = (\xi_1, \xi_2)$. The control-dependent Hamiltonian function of PMP is

$$h_u(\xi, x) = (\xi_1, \xi_2) \begin{pmatrix} x_2 \\ u \end{pmatrix} = \xi_1 x_2 + \xi_2 u,$$

and the corresponding Hamiltonian system has the form

$$\begin{cases} \dot{x} = \dfrac{\partial h_u}{\partial \xi}, \\ \dot{\xi} = -\dfrac{\partial h_u}{\partial x}. \end{cases}$$

In coordinates this system splits into two independent subsystems:

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = u, \end{cases} \qquad \begin{cases} \dot{\xi}_1 = 0, \\ \dot{\xi}_2 = -\xi_1. \end{cases} \tag{32}$$

By PMP, if a control $\tilde{u}(\cdot)$ is time-optimal, then the Hamiltonian system has a nontrivial solution $(\xi(t), x(t))$, $\xi(t) \not\equiv 0$, such that

$$h_{\tilde{u}(t)}(\xi(t), x(t)) = \max_{|u| \le 1} h_u(\xi(t), x(t)) \ge 0.$$

From this maximality condition, if $\xi_2(t) \neq 0$, then $\tilde{u}(t) = \operatorname{sgn} \xi_2(t)$. Notice that the maximized Hamiltonian

$$\max_{|u| \le 1} h_u(\xi, x) = \xi_1 x_2 + |\xi_2|$$

is not smooth. So we cannot apply Proposition 1, but we can obtain description of optimal controls directly from Pontryagin Maximum Principle, without preliminary maximization of Hamiltonian.

Since

$$\ddot{\xi}_2 = 0,$$

then ξ_2 is linear:

$$\xi_2(t) = \alpha + \beta t, \qquad \alpha,\ \beta = \text{const},$$

hence the optimal control has the form

$$\tilde{u}(t) = \operatorname{sgn}(\alpha + \beta t).$$

So $\tilde{u}(t)$ is piecewise constant, takes only the extremal values ± 1, and has not more than one switching (discontinuity point).

New we find all trajectories $x(t)$ that correspond to such controls and come to the origin. For controls $u = \pm 1$, the first of subsystems (32) reads

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = \pm 1. \end{cases}$$

Trajectories of this system satisfy the equation

$$\frac{d\,x_1}{d\,x_2} = \pm x_2,$$

thus are parabolas of the form

$$x_1 = \pm \frac{x_2^2}{2} + C, \qquad C = \text{const}\,.$$

First we find trajectories from this family that come to the origin without switchings: these are two semiparabolas

$$x_1 = \frac{x_2^2}{2}, \qquad x_2 < 0, \quad \dot{x}_2 > 0, \tag{33}$$

and

$$x_1 = -\frac{x_2^2}{2}, \qquad x_2 > 0, \quad \dot{x}_2 < 0, \tag{34}$$

for $u = +1$ and -1 respectively.

Now we find all extremal trajectories with one switching. Let $(x_{1s}, x_{2s}) \in \mathbb{R}^2$ be a switching point for anyone of curves (33), (34). Then extremal trajectories with one switching coming to the origin have the form

$$x_1 = \begin{cases} -x_2^2/2 + x_{2s}^2/2 + x_{1s}, & x_2 > x_{2s}, \quad \dot{x}_2 < 0, \\ \\ x_2^2/2 & 0 > x_2 > x_{2s}, \quad \dot{x}_2 > 0, \end{cases} \tag{35}$$

and

$$x_1 = \begin{cases} x_2^2/2 - x_{2s}^2/2 + x_{1s}, & x_2 < x_{2s}, \quad \dot{x}_2 > 0, \\ \\ -x_2^2/2 & 0 < x_2 < x_{2s}, \quad \dot{x}_2 < 0. \end{cases} \tag{36}$$

It is easy to see that through any point (x_1, x_2) of the plane passes exactly one curve of the forms (33)–(36). So for any point of the plane there exists exactly one extremal trajectory steering this point to the origin. Since optimal trajectories exist, then the solutions found are optimal.

4.2 Control of a linear oscillator

Consider a linear oscillator whose motion can be controlled by force bounded in absolute value. The corresponding control system (after appropriate rescaling) is

$$\ddot{x}_1 + x_1 = u, \qquad |u| \leq 1, \quad x_1 \in \mathbb{R},$$

or, in the canonical form:

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u, \end{cases} \qquad |u| \leq 1, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2.$$

We consider the time-optimal problem for this system.

By Filippov's theorem, optimal control exists. Similarly to the previous problem, we apply Pontryagin Maximum Principle: the Hamiltonian function is

$$h_u(\xi, x) = \xi_1 x_2 - \xi_2 x_1 + \xi_2 u, \qquad (\xi, x) \in T^*\mathbb{R}^2 = \mathbb{R}^{2*} \times \mathbb{R}^2,$$

and the Hamiltonian system reads

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u, \end{cases} \qquad \begin{cases} \dot{\xi}_1 = \xi_2, \\ \dot{\xi}_2 = -\xi_1. \end{cases}$$

The maximality condition of PMP yields

$$\xi_2(t)\tilde{u}(t) = \max_{|u| \leq 1} \xi_2(t) u,$$

thus optimal controls satisfy the condition

$$\tilde{u}(t) = \operatorname{sgn} \xi_2(t) \qquad \text{if } \xi_2(t) \neq 0.$$

For the variable ξ_2 we have the ODE

$$\ddot{\xi}_2 = -\xi_2,$$

hence

$$\xi_2 = \alpha \sin(t+\beta), \qquad \alpha,\ \beta = \text{const}\,.$$

Notice that $\alpha \neq 0$: indeed, if $\xi_2 \equiv 0$, then $\xi_1 = -\dot{\xi}_2(t) \equiv 0$, thus $\xi(t) = (\xi_1(t), \xi_2(t)) \equiv 0$, which is impossible by PMP. Consequently,

$$\tilde{u}(t) = \operatorname{sgn}(\alpha \sin(t+\beta)).$$

This equality yields a complete description of possible structure of optimal control. The interval between successive switching points of $\tilde{u}(t)$ has the length π. Let $\tau \in [0,\pi)$ be the first switching point of $\tilde{u}(t)$. Then

$$\tilde{u}(t) = \begin{cases} \operatorname{sgn}\tilde{u}(0), & t \in [0,\tau) \cup [\tau+\pi, \tau+2\pi) \cup [\tau+3\pi, \tau+4\pi) \cup \dots \\ -\operatorname{sgn}\tilde{u}(0), & t \in [\tau, \tau+\pi) \cup [\tau+2\pi, \tau+3\pi) \cup \dots \end{cases}$$

That is, $\tilde{u}(t)$ is parametrized by two numbers: the first switching time $\tau \in [0,\pi)$ and the initial sign $\operatorname{sgn}\tilde{u}(0) \in \{\pm 1\}$.

Optimal control $\tilde{u}(t)$ takes only the extremal values ± 1. Thus optimal trajectories $(x_1(t), x_2(t))$ consist of pieces that satisfy the system

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 \pm 1, \end{cases} \tag{37}$$

i.e., arcs of the circles

$$(x_1 \pm 1)^2 + x_2^2 = C, \qquad C = \text{const},$$

passed clockwise.

Now we describe all optimal trajectories coming to the origin. Let γ be any such trajectory. If γ has no switchings, then it is an arc belonging to one of the semicircles

$$(x_1 - 1)^2 + x_2^2 = 1, \qquad x_2 \leq 0, \tag{38}$$

$$(x_1 + 1)^2 + x_2^2 = 1, \qquad x_2 \geq 0 \tag{39}$$

and containing the origin. If γ has switchings, then the last switching can occur at any point of these semicircles except the origin. Assume that γ has the last switching on semicircle (38). Then the part of γ before the last switching and after the next to last switching is a semicircle of the circle $(x_1+1)^2 + x_2^2 = C$ passing through the last switching point. The next to last switching of γ occurs on the curve obtained by rotation of semicircle (38)

around the point $(-1, 0)$ in the plane (x_1, x_2) by the angle π, i.e., on the semicircle

$$(x_1 + 3)^2 + x_2^2 = 1, \qquad x_2 \geq 0. \tag{40}$$

To obtain the geometric locus of the previous switching of γ, we have to rotate semicircle (40) around the point $(1, 0)$ by the angle π; we come to the semicircle

$$(x_1 - 5)^2 + x_2^2 = 1, \qquad x_2 \leq 0.$$

The previous switching of γ takes place on the semicircle

$$(x_1 + 7)^2 + x_2^2 = 1, \qquad x_2 \geq 0,$$

and so on.

The case when the last switching of γ occurs on semicircle (39) is obtained from the case just considered by the central symmetry of the plane (x_1, x_2) w.r.t. the origin: $(x_1, x_2) \mapsto (-x_1, -x_2)$. Then the successive switchings of γ (in the reverse order starting from the end) occur on the semicircles

$$\begin{aligned}
&(x_1 + 1)^2 + x_2^2 = 1, \qquad x_2 \geq 0,\\
&(x_1 - 3)^2 + x_2^2 = 1, \qquad x_2 \leq 0,\\
&(x_1 + 5)^2 + x_2^2 = 1, \qquad x_2 \geq 0,\\
&(x_1 - 7)^2 + x_2^2 = 1, \qquad x_2 \leq 0,
\end{aligned}$$

etc. We obtained the switching curve in the plane (x_1, x_2):

$$\begin{aligned}
&(x_1 - (2k-1))^2 + x_2^2 = 1, \qquad x_2 \leq 0, \quad k \in \mathbb{N},\\
&(x_1 + (2k-1))^2 + x_2^2 = 1, \qquad x_2 \geq 0, \quad k \in \mathbb{N}.
\end{aligned} \tag{41}$$

This switching curve divides the plane (x_1, x_2) into two parts. Any extremal trajectory $(x_1(t), x_2(t))$ in the upper part of the plane is a solution of ODE (37) with -1 in the second equation, and in the lower part it is a solution of (37) with $+1$. For any point of the plane (x_1, x_2) there exists exactly one curve of this family of extremal trajectories that comes to the origin (it has the form of a "spiral" with a finite number of switchings). Since optimal trajectories exist, the constructed extremal trajectories are optimal.

The time-optimal control problem is solved: in the part of the plane (x_1, x_2) over the switching curve (41) the optimal control is $\tilde{u} = -1$, and

below this curve $\tilde{u} = +1$. Through any point of the plane passes one optimal trajectory which corresponds to this optimal control rule. After finite number of switchings, any optimal trajectory comes to the origin.

Now we consider optimal control problems with the same dynamics as in the previous two sections, but with another cost functional.

4.3 The cheapest stop of a train

As in Section 4.1, we control motion of a train. Now the goal is to stop the train at a fixed instant of time with a minimum expenditure of energy, which is assumed proportional to the integral of squared acceleration.

So the optimal control problem is as follows:

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = u, \end{cases} \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2, \quad u \in \mathbb{R},$$

$$x(0) = x^0, \quad x(t_1) = 0, \quad t_1 \text{ fixed},$$

$$\frac{1}{2} \int_0^{t_1} u^2 \, dt \to \min .$$

Filippov's theorem cannot be applied directly since the right-hand side of the control system is not compact. Although, one can choose a new time $t \mapsto \frac{1}{2} \int_0^t u^2(\tau) \, d\tau + C$ and obtain a bounded right-hand side, then compactify it and apply Filippov's theorem. In such a way existence of optimal control can be proved. See also the general theory of linear quadratic problems below in Chapter 6.

To find optimal control, we apply PMP. The Hamiltonian function is

$$h_u^\nu(\xi, x) = \xi_1 x_2 + \xi_2 u + \frac{\nu}{2} u^2, \qquad (\xi, x) \in \mathbb{R}^{2*} \times \mathbb{R}^2 .$$

Along optimal trajectories

$$\nu \le 0, \quad \nu = \text{const} .$$

From the Hamiltonian system of PMP, we have

$$\begin{cases} \dot{\xi}_1 = 0, \\ \dot{\xi}_2 = -\xi_1 . \end{cases} \tag{42}$$

Consider first the case of abnormal extremals:

$$\nu = 0.$$

The triple (ξ_1, ξ_2, ν) must be nonzero, thus

$$\xi_2(t) \not\equiv 0.$$

But the maximality condition of PMP yields

$$\tilde{u}(t)\xi_2(t) = \max_{u \in \mathbb{R}} u\,\xi_2(t). \tag{43}$$

Since $\xi_2(t) \neq 0$, the maximum above does not exist. Consequently, there are no abnormal extremals.

Consider the normal case: $\nu \neq 0$, we can take $\nu = -1$. The normal Hamiltonian function is

$$h_u(\xi, x) = h_u^{-1}(\xi, x) = \xi_1 x_2 + \xi_2 u - \frac{1}{2}u^2.$$

Maximality condition of PMP is equivalent to $\frac{\partial\, h_u}{\partial\, u} = 0$, thus

$$\tilde{u}(t) = \xi_2(t)$$

along optimal trajectories. Taking into account system (42), we conclude that optimal control is linear:

$$\tilde{u}(t) = \alpha t + \beta, \qquad \alpha,\ \beta = \text{const}\,.$$

The maximized Hamiltonian function

$$H(\xi, x) = \max_u h_u(\xi, x) = \xi_1 x_2 + \frac{1}{2}\xi_2^2$$

is smooth. That is why optimal trajectories satisfy the Hamiltonian system

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = \xi_2, \\ \dot{\xi}_1 = 0, \\ \dot{\xi}_2 = -\xi_1. \end{cases}$$

For the variable x_1 we obtain the boundary value problem

$$\begin{aligned} &x_1^{(4)} = 0, \\ &x_1(0) = x_1^0, \quad \dot{x}_1(0) = x_2^0, \quad x_1(t_1) = 0, \quad \dot{x}_1(t_1) = 0. \end{aligned} \tag{44}$$

For any (x_1^0, x_2^0), there exists exactly one solution $x_1(t)$ of this problem — a cubic spline. The function $x_2(t)$ is found from the equation $x_2 = \dot{x}_1$.

So through any initial point $x^0 \in \mathbb{R}^2$ passes a unique extremal trajectory arriving at the origin. It is a curve $(x_1(t), x_2(t))$, $t \in [0, t_1]$, where $x_1(t)$ is a cubic polynomial that satisfies the boundary conditions (44), and $x_2(t) = \dot{x}_1(t)$. In view of existence, this is an optimal trajectory.

4.4 Control of a linear oscillator with cost

We control a linear oscillator, say a pendulum with a small amplitude, by an unbounded force u, but take into account expenditure of energy measured by the integral $\frac{1}{2}\int_0^{t_1} u^2(t)\,dt$. The optimal control problem reads

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u, \end{cases} \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2, \quad u \in \mathbb{R},$$

$$x(0) = x^0, \quad x(t_1) = 0, \quad t_1 \text{ fixed},$$

$$\frac{1}{2}\int_0^{t_1} u^2\,dt \to \min.$$

Existence of optimal control can be proved by the same argument as in the previous section.

The Hamiltonian function of PMP is

$$h_u^\nu(\xi, x) = \xi_1 x_2 - \xi_2 x_1 + \xi_2 u + \frac{\nu}{2}u^2.$$

The corresponding Hamiltonian system yields

$$\begin{cases} \dot{\xi}_1 = \xi_2, \\ \dot{\xi}_2 = -\xi_1. \end{cases}$$

In the same way as in the previous problem, we show that there are no abnormal extremals, thus we can assume $\nu = -1$. Then the maximality condition yields

$$\tilde{u}(t) = \xi_2(t).$$

In particular, optimal control is a harmonic:

$$\tilde{u}(t) = \alpha \sin(t + \beta), \qquad \alpha,\ \beta = \text{const}.$$

The system of ODEs for extremal trajectories

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + \alpha \sin(t + \beta) \end{cases}$$

is solved explicitly:

$$\begin{aligned} x_1(t) &= -\frac{\alpha}{2} t \cos(t + \beta) + a \sin(t + b), \\ x_2(t) &= \frac{\alpha}{2} t \sin(t + \beta) - \frac{\alpha}{2} \cos(t + \beta) + a \cos(t + b), \quad a,\ b \in \mathbb{R}. \end{aligned} \tag{45}$$

Exercise 1. Show that exactly one extremal trajectory of the form (45) satisfies the boundary conditions.

In view of existence, these extremal trajectories are optimal.

4.5 Dubins car

Consider a car moving in the plane. The car can move forward with a fixed linear velocity and simultaneously rotate with a bounded angular velocity. Given initial and terminal position and orientation of the car in the plane, the problem is to drive the car from the initial configuration to the terminal one for a minimal time.

Admissible paths of the car are curves with bounded curvature. Suppose that curves are parametrized by length, then our problem can be stated geometrically. Given two points in the plane and two unit velocity vectors attached respectively at these points, one has to find a curve in the plane that starts at the first point with the first velocity vector and comes to the second point with the second velocity vector, has curvature bounded by a given constant, and has the minimal length among all such curves.

Remark. If curvature is unbounded, then the problem, in general, has no solutions. Indeed, the infimum of lengths of all curves that satisfy the boundary conditions without bound on curvature is the distance between the initial and terminal points: the segment of the straight line through these points can be approximated by smooth curves with the required boundary conditions. But this infimum is not attained when the boundary velocity vectors do not lie on the line through the boundary points and are not collinear one to another.

After rescaling, we obtain a time-optimal problem for a nonlinear system:

$$\begin{cases} \dot{x}_1 = \cos\theta, \\ \dot{x}_2 = \sin\theta, \\ \dot{\theta} = u, \end{cases} \tag{46}$$
$$x = (x_1, x_2) \in \mathbb{R}^2, \quad \theta \in S^1, \quad |u| \le 1,$$
$$x(0),\ \theta(0),\ x(t_1),\ \theta(t_1) \text{ fixed},$$
$$t_1 \to \min.$$

Existence of solutions is guaranteed by Filippov's Theorem. We apply Pontryagin Maximum Principle.

We have $(x_1, x_2, \theta) \in M = \mathbb{R}^2_x \times S^1_\theta$, let (ξ_1, ξ_2, μ) be the corresponding coordinates of the adjoint vector. Then

$$\lambda = (x, \theta, \xi, \mu) \in T^*M,$$

and the control-dependent Hamiltonian is

$$h_u(\lambda) = \xi_1 \cos\theta + \xi_2 \sin\theta + \mu u.$$

The Hamiltonian system of PMP yields

$$\dot{\xi} = 0, \tag{47}$$
$$\dot{\mu} = \xi_1 \sin\theta - \xi_2 \cos\theta, \tag{48}$$

and the maximality condition reads

$$\mu(t)u(t) = \max_{|u|\le 1} \mu(t)u. \tag{49}$$

Equation (47) means that ξ is constant along optimal trajectories, thus the right-hand side of (48) can be rewritten as

$$\xi_1 \sin\theta - \xi_2 \cos\theta = \alpha \sin(\theta + \beta), \qquad \alpha,\ \beta = \text{const}, \quad \alpha = \sqrt{\xi_1^2 + \xi_2^2} \ge 0. \tag{50}$$

So the Hamiltonian system of PMP (46)–(48) yields the following system:

$$\begin{cases} \dot{\mu} = \alpha \sin(\theta + \beta), \\ \dot{\theta} = u. \end{cases}$$

Maximality condition (49) implies that

$$u(t) = \operatorname{sgn} \mu(t) \quad \text{if } \mu(t) \neq 0. \tag{51}$$

If $\alpha = 0$, then $(\xi_1, \xi_2) \equiv 0$ and $\mu = \text{const} \neq 0$, thus $u = \text{const} = \pm 1$. So the curve $x(t)$ is an arc of a circle of radius 1.

Let $\alpha \neq 0$, then in view of (50), we have $\alpha > 0$. Conditions (47), (48), (49) are preserved if the adjoint vector (ξ, μ) is multiplied by any positive constant. Thus we can choose (ξ, μ) such that $\alpha = \sqrt{\xi_1^2 + \xi_2^2} = 1$. That is why we suppose in the sequel that

$$\alpha = 1.$$

Condition (51) means that behavior of sign of the function $\mu(t)$ is crucial for the structure of optimal control. We consider several possibilities for $\mu(t)$.

(0) If the function $\mu(t)$ does not vanish on the segment $[0, t_1]$, then the optimal control is constant:

$$u(t) = \text{const} = \pm 1, \qquad t \in [0, t_1], \tag{52}$$

and the optimal trajectory $x(t)$, $t \in [0, t_1]$, is an arc of a circle. Notice that an optimal trajectory cannot contain a full circle: a circle can be eliminated so that the resulting trajectory satisfy the same boundary conditions and is shorter. Thus controls (52) can be optimal only if $t_1 < 2\pi$.

In the sequel we can assume that the set

$$N = \{\tau \in [0, t_1] \mid \mu(\tau) = 0\}$$

is nonempty. Since N is open, it is a union of open intervals in $[0, t_1]$, plus, may be, semiopen intervals of the form $[0, \tau_1)$, $(\tau_2, t_1]$.

(1) Suppose that the set N contains an interval of the form

$$(\tau_1, \tau_2) \subset [0, t_1], \qquad \tau_1 < \tau_2. \tag{53}$$

We can assume that the interval (τ_1, τ_2) is maximal w.r.t. inclusion:

$$\mu(\tau_1) = \mu(\tau_2) = 0, \qquad \mu|_{(\tau_1, \tau_2)} \neq 0.$$

From PMP we have the inequality

$$h_{u(t)}(\lambda(t)) = \cos(\theta(t) + \beta) + \mu(t)u(t) \geq 0.$$

Thus

$$\cos(\theta(\tau_1) + \beta) \geq 0.$$

This inequality means that the angle

$$\widehat{\theta} = \theta(\tau_1) + \beta$$

satisfies the inclusion

$$\widehat{\theta} \in \left[0, \frac{\pi}{2}\right] \cup \left[\frac{3\pi}{2}, 2\pi\right).$$

Consider first the case

$$\widehat{\theta} \in \left(0, \frac{\pi}{2}\right].$$

Then $\dot{\mu}(\tau_1) = \sin\widehat{\theta} > 0$, thus at τ_1 control switches from -1 to $+1$, so

$$\dot{\theta}(t) = u(t) \equiv 1, \qquad t \in (\tau_1, \tau_2).$$

We evaluate the distance $\tau_2 - \tau_1$. Since

$$\mu(\tau_2) = \int_{\tau_1}^{\tau_2} \sin(\widehat{\theta} + \tau - \tau_1)\, d\tau = 0,$$

then $\tau_2 - \tau_1 = 2(\pi - \widehat{\theta})$, thus

$$\tau_2 - \tau_1 \in [\pi, 2\pi). \tag{54}$$

In the case

$$\widehat{\theta} \in \left[\frac{3\pi}{2}, 2\pi\right)$$

inclusion (54) is proved similarly, and in the case $\widehat{\theta} = 0$ we obtain no optimal controls (the curve $x(t)$ contains a full circle, which can be eliminated).

Inclusion (54) means that successive roots τ_1, τ_2 of the function $\mu(t)$ cannot be arbitrarily close one to another. Moreover, the previous argument shows that at such instants τ_i optimal control switches from one extremal value to another, and along any optimal trajectory the distance between any successive switchings τ_i, τ_{i+1} is the same.

So in case (1) an optimal control can only have the form

$$u(t) = \begin{cases} \varepsilon, & t \in (\tau_{2k-1}, \tau_{2k}), \\ -\varepsilon, & t \in (\tau_{2k}, \tau_{2k+1}), \end{cases} \tag{55}$$

$$\varepsilon = \pm 1,$$

$$\tau_{i+1} - \tau_i = \text{const} \in [\pi, 2\pi), \qquad i = 1, \ldots, N-1, \tag{56}$$

$$\tau_1 \in (0, 2\pi),$$

here we do not indicate values of u in the intervals before the first switching, $t \in (0, \tau_1)$, and after the last switching, $t \in (\tau_N, t_1)$. For such trajectories, control takes only extremal values ± 1 and the number of switchings is finite on any compact time segment. Such a control is called *bang-bang*.

Controls $u(t)$ given by (55), (56) satisfy PMP for arbitrarily large t, but they are not optimal if the number of switchings is $N > 3$. Indeed, suppose that such a control has at least 4 switchings. Then the piece of trajectory $x(t)$, $t \in [\tau_1, \tau_4]$, is a concatenation of three arcs of circles corresponding to the segments of time $[\tau_1, \tau_2]$, $[\tau_2, \tau_3]$, $[\tau_3, \tau_4]$ with

$$\tau_4 - \tau_3 = \tau_3 - \tau_2 = \tau_2 - \tau_1 \in [\pi, 2\pi).$$

Draw the segment of line

$$\tilde{x}(t), \qquad t \in [(\tau_1 + \tau_2)/2, (\tau_3 + \tau_4)/2], \quad \left|\frac{d\,\tilde{x}}{d\,t}\right| \equiv 1,$$

the common tangent to the first and third circles through the points $x\,((\tau_1 + \tau_2)/2)$ and $x\,((\tau_3 + \tau_4)/2)$. Then the curve

$$y(t) = \begin{cases} x(t), & t \notin [(\tau_1 + \tau_2)/2, (\tau_3 + \tau_4)/2], \\ \tilde{x}(t), & t \in [(\tau_1 + \tau_2)/2, (\tau_3 + \tau_4)/2], \end{cases}$$

is an admissible trajectory and shorter than $x(t)$. We proved that optimal bang-bang control can have not more than 3 switchings.

(2) It remains to consider the case where the set N does not contain intervals of the form (53). Then N consists of at most two semiopen intervals:

$$N = [0, \tau_1) \cup (\tau_2, t_1], \qquad \tau_1 \le \tau_2,$$

where one or both intervals may be absent. If $\tau_1 = \tau_2$, then the function $\mu(t)$ has a unique root on the segment $[0, t_1]$, and the corresponding optimal control is determined by condition (51). Otherwise

$$\tau_1 < \tau_2,$$

and

$$\mu|_{[0,\tau_1)} \neq 0, \quad \mu|_{[\tau_1,\tau_2]} \equiv 0, \quad \mu|_{(\tau_2,t_1]} \neq 0. \tag{57}$$

In this case the maximality condition of PMP (51) does not determine optimal control $u(t)$ uniquely since the maximum is attained for more than one

value of control parameter u. Such a control is called *singular*. Nevertheless, singular controls in this problem can be determined from PMP. Indeed, the following identities hold on the interval (τ_1, τ_2):

$$\dot{\mu} = \sin(\theta + \beta) = 0 \quad \Rightarrow \quad \theta + \beta = \pi k \quad \Rightarrow \quad \theta = \text{const} \quad \Rightarrow \quad u = 0.$$

Consequently, if an optimal trajectory $x(t)$ has a singular piece, which is a line, then τ_1 and τ_2 are the only switching times of the optimal control. Then

$$u|_{(0,\tau_1)} = \text{const} = \pm 1, \quad u|_{(\tau_2,t_1)} = \text{const} = \pm 1,$$

and the whole trajectory $x(t)$, $t \in [0, t_1]$, is a concatenation of an arc of a circle of radius 1

$$x(t), \quad u(t) = \pm 1, \qquad t \in [0, \tau_1],$$

a line

$$x(t), \quad u(t) = 0, \qquad t \in [\tau_1, \tau_2],$$

and one more arc of a circle of radius 1

$$x(t), \quad u(t) = \pm 1, \qquad t \in [\tau_2, t_1].$$

So optimal trajectories in the problem have one of the following two types:

(1) concatenation of a bang-bang piece (arc of a circle, $u = \pm 1$), a singular piece (segment of a line, $u = 0$), and a bang-bang piece, or

(2) concatenation of bang-bang pieces with not more than 3 switchings, the arcs of circles between switchings having the same central angle $\in [\pi, 2\pi)$.

If boundary points $x(0)$, $x(t_1)$ are sufficiently far one from another, then they can be connected only by trajectories containing singular piece. For such boundary points, we obtain a simple algorithm for construction of an optimal trajectory. Through each of the points $x(0)$ and $x(t_1)$, construct a pair of circles of radius 1 tangent respectively to the velocity vectors $\dot{x}(0) = (\cos\theta(0), \sin\theta(0))$ and $\dot{x}(t_1) = (\cos\theta(t_1), \sin\theta(t_1))$. Then draw common tangents to the circles at $x(0)$ and $x(t_1)$ respectively, so that direction of motion along these tangents was compatible with direction of rotation along the circles determined by the boundary velocity vectors $\dot{x}(0)$ and $\dot{x}(t_1)$. Finally, choose the shortest curve among the candidates obtained. This curve is the optimal trajectory.

5 Linear time-optimal problem

5.1 Problem statement

In this chapter we study the following optimal control problem:

$$\begin{aligned} &\dot{x} = Ax + Bu, \qquad x \in \mathbb{R}^n, \quad u \in U \subset \mathbb{R}^m, \\ &x(0) = x_0, \quad x(t_1) = x_1, \quad x_0,\ x_1 \in \mathbb{R}^n \text{ fixed}, \\ &t_1 \to \min, \end{aligned} \tag{58}$$

where U is a compact convex polytope in $\mathbb{R}^m$, and A and B are constant matrices of order $n \times n$ and $n \times m$ respectively. Such problem is called *linear time-optimal problem*.

The polytope U is the convex hull of a finite number of points $a_1, \dots, a_k$ in $\mathbb{R}^m$:

$$U = \operatorname{conv}\{a_1, \dots, a_k\}.$$

We assume that the points a_i do not belong to the convex hull of all the rest points a_j, $j \neq i$, so that each a_i is a vertex of the polytope U.

In the sequel we assume the following *General Position Condition*:

For any edge $[a_i, a_j]$ of U, the vector $e_{ij} = a_j - a_i$ satisfies the equality

$$\operatorname{span}(Be_{ij}, ABe_{ij}, \dots, A^{n-1}Be_{ij}) = \mathbb{R}^n. \tag{59}$$

This condition means that no vector Be_{ij} belongs to a proper invariant subspace of the matrix A. This is equivalent to controllability of the linear system $\dot{x} = Ax + Bu$ with the set of control parameters $u \in \mathbb{R}e_{ij}$. Condition (59) can be achieved by a small perturbation of matrices A, B.

We already considered examples of linear time-optimal problems in Sections 4.1, 4.2. Here we study the structure of optimal control, prove its uniqueness, evaluate the number of switchings.

Existence of optimal control for any points x_0, x_1 such that $x_1 \in \mathcal{A}(x_0)$ is guaranteed by Filippov's theorem. Notice that for the analogous problem with an unbounded set of control parameters, optimal control may not exist: it is easy to show this using linearity of the system.

Before proceeding with the study of linear time-optimal problems, we recall some basic facts on polytopes.

5.2 Geometry of polytopes

The convex hull of a finite number of points $a_1, \dots, a_k \in \mathbb{R}^m$ is the set

$$U = \operatorname{conv}\{a_1, \dots, a_k\} \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{k} \alpha_i a_i \mid \alpha_i \geq 0,\ \sum_{i=1}^{k} \alpha_i = 1 \right\}.$$

An affine hyperplane in $\mathbb{R}^m$ is a set of the form

$$\Pi = \{u \in \mathbb{R}^m \mid \langle \xi, u \rangle = c\}, \qquad \xi \in \mathbb{R}^{m*} \setminus \{0\}, \quad c \in \mathbb{R}.$$

A supporting hyperplane to a polytope U is a hyperplane Π such that

$$\langle \xi, u \rangle \leq c \quad \forall u \in U$$

for the covector ξ and number c that define Π, and this inequality turns into equality at some point $u \in \partial U$, i.e., $\Pi \cap U \neq \emptyset$.

A polytope $U = \operatorname{conv}\{a_1, \dots, a_k\}$ intersects with any its supporting hyperplane $\Pi = \{u \mid \langle \xi, u \rangle = c\}$ by another polytope:

$$\begin{aligned} &U \cap \Pi = \operatorname{conv}\{a_{i1}, \dots, a_{il}\}, \\ &\langle \xi, a_{i1} \rangle = \dots = \langle \xi, a_{il} \rangle = c, \\ &\langle \xi, a_j \rangle < c, \quad j \notin \{i_1, \dots, i_l\}. \end{aligned}$$

Such polytopes $U \cap \Pi$ are called faces of the polytope U. Zero-dimensional and one-dimensional faces are called respectively vertices and edges. A polytope has a finite number of faces, each of which is the convex hull of a finite number of vertices. A face of a face is a face of the initial polytope. Boundary of a polytope is a union of all its faces. This is a straightforward corollary of the separation theorem for convex sets (or the Hahn-Banach Theorem).

5.3 Bang-bang theorem

Optimal control in the linear time-optimal problem is bang-bang, i.e., it is piecewise constant and takes values in vertices of the polytope U.

Theorem 5. *Let $u(t)$, $0 \leq t \leq t_1$, be an optimal control in the linear time-optimal control problem* (58)*. Then there exists a finite subset*

$$\mathcal{T} \subset [0, t_1], \qquad \#\mathcal{T} < \infty,$$

such that

$$u(t) \in \{a_1, \dots, a_k\}, \qquad t \in [0, t_1] \setminus \mathcal{T}, \tag{60}$$

and restriction $u(t)|_{t \in [0,t_1] \setminus \mathcal{T}}$ is locally constant.

Proof. Apply Pontryagin Maximum Principle to the linear time-optimal problem (58). State vector and adjoint vectors are

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n, \qquad \xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^{n*}.$$

The control-dependent Hamiltonian is

$$h_u(\xi, x) = \xi A x + \xi B u$$

(we multiply raws by columns). The Hamiltonian system and maximality condition of PMP take the form:

$$\begin{cases} \dot{x} = Ax + Bu, \\ \dot{\xi} = -\xi A, \end{cases}$$
$$\xi(t) \neq 0,$$
$$\xi(t)Bu(t) = \max_{u \in U} \xi(t)Bu. \tag{61}$$

The Hamiltonian system implies that adjoint vector

$$\xi(t) = \xi(0)e^{-tA}, \qquad \xi(0) \neq 0, \tag{62}$$

is analytic along the optimal trajectory.

Consider the set of indices corresponding to vertices where maximum (61) is attained:

$$J(t) = \left\{ 1 \leq j \leq k \mid \xi(t)Ba_j = \max_{u \in U} \xi(t)Bu = \max\{\xi(t)Ba_i \mid i = 1, \dots, k\} \right\}.$$

At each instant t the linear function $\xi(t)B$ attains maximum at vertices of the polytope U. We show that this maximum is attained at one vertex always except a finite number of moments.

Define the set

$$\mathcal{T} = \{t \in [0, t_1] \mid \#J(t) > 1\}.$$

By contradiction, suppose that $\mathcal{T}$ is infinite: there exists a sequence of distinct moments

$$\{\tau_1, \dots, \tau_n, \dots\} \subset \mathcal{T}.$$

Since there is a finite number of choices for the subset $J(\tau_n) \subset \{1, \dots, k\}$, we can assume, without loss of generality, that

$$J(\tau_1) = J(\tau_2) = \cdots = J(\tau_n) = \cdots .$$

Denote $J = J(\tau_i)$.

Further, since the convex hull

$$\operatorname{conv}\{a_j \mid j \in J\}$$

is a face of U, then there exist indices $j_1, j_2 \in J$ such that the segment $[a_{j_1}, a_{j_2}]$ is an edge of U. We have

$$\xi(\tau_i)Ba_{j_1} = \xi(\tau_i)Ba_{j_2}, \qquad i = 1, 2, \dots \ .$$

For the vector $e = a_{j_2} - a_{j_1}$ we obtain

$$\xi(\tau_i)Be = 0, \qquad i = 1, 2, \dots \ .$$

But $\xi(\tau_i) = \xi(0)e^{-\tau_i A}$ by (62), so the analytic function

$$t \mapsto \xi(0)e^{-tA}Be$$

has an infinite number of zeros on the segment $[0, t_1]$, thus it is identically zero:

$$\xi(0)e^{-tA}Be \equiv 0.$$

We differentiate this identity successively at $t = 0$ and obtain

$$\xi(0)Be = 0, \quad \xi(0)ABe = 0, \quad \dots, \quad \xi(0)A^{n-1}Be = 0.$$

By General Position Condition (59), we have $\xi(0) = 0$, a contradiction to (62). So the set $\mathcal{T}$ is finite.

Out of the set $\mathcal{T}$, the function $\xi(t)B$ attains maximum on U at one vertex $a_{j(t)}$, $\{j(t)\} = J(t)$, thus the optimal control $u(t)$ takes value in the vertex $a_{j(t)}$. Condition (60) follows. Further,

$$\xi(t)Ba_{j(t)} > \xi(t)Ba_i, \qquad i \neq j(t).$$

But all functions $t \mapsto \xi(t)Ba_i$ are continuous, so the preceding inequality preserves for instants close to t. The function $t \mapsto j(t)$ is locally constant on $[0, t_1] \setminus \mathcal{T}$, thus the optimal control $u(t)$ is also locally constant on $[0, t_1] \setminus \mathcal{T}$. □

In the sequel we will need the following statement proved in the preceding argument.

Corollary 4. *Let $\xi(t)$, $t \in [0, t_1]$, be a nonzero solution of the adjoint equation $\dot{\xi} = -\xi A$. Then everywhere in the segment $[0, t_1]$, except a finite number of points, there exists a unique control $u(t) \in U$ such that $\xi(t)Bu(t) = \max\limits_{u \in U} \xi(t)Bu$.*

5.4 Uniqueness of optimal controls and extremals

Theorem 6. *Let the terminal point x_1 be reachable from the initial point x_0:*

$$x_1 \in \mathcal{A}(x_0).$$

Then linear time-optimal control problem (58) has a unique solution.

Proof. As we already noticed, existence of an optimal control follows from Filippov's Theorem.

Suppose that there exist two optimal controls: $u_1(t)$, $u_2(t)$, $t \in [0, t_1]$. By Cauchy's formula:

$$x(t_1) = e^{t_1 A}\left(x_0 + \int_0^{t_1} e^{-tA} B u(t)\, dt\right),$$

we obtain

$$e^{t_1 A}\left(x_0 + \int_0^{t_1} e^{-tA} B u_1(t)\, dt\right) = e^{t_1 A}\left(x_0 + \int_0^{t_1} e^{-tA} B u_2(t)\, dt\right),$$

thus

$$\int_0^{t_1} e^{-tA} B u_1(t)\, dt = \int_0^{t_1} e^{-tA} B u_2(t)\, dt. \tag{63}$$

Let $\xi_1(t) = \xi_1(0) e^{-tA}$ be the adjoint vector corresponding by PMP to the control $u_1(t)$. Then equality (63) can be written in the form

$$\int_0^{t_1} \xi_1(t) B u_1(t)\, dt = \int_0^{t_1} \xi_1(t) B u_2(t)\, dt. \tag{64}$$

By the maximality condition of PMP

$$\xi_1(t) B u_1(t) = \max_{u \in U} \xi_1(t) B u,$$

thus

$$\xi_1(t) B u_1(t) \geq \xi_1(t) B u_2(t).$$

But this inequality together with equality (64) implies that almost everywhere on $[0, t_1]$

$$\xi_1(t) B u_1(t) = \xi_1(t) B u_2(t).$$

By Corollary 4,

$$u_1(t) \equiv u_2(t)$$

almost everywhere on $[0, t_1]$. □

So for linear time-optimal problem, optimal control is unique. The standard procedure to find the optimal control for a given pair of boundary points x_0, x_1 is to find all extremals $(\xi(t), x(t))$ steering x_0 to x_1 and then to seek for the best among them. In the examples considered in Sections 4.1, 4.2, there was one extremal for each pair x_0, x_1 with $x_1 = 0$. We prove now that this is a general property of linear time-optimal problems.

Theorem 7. *Let $x_1 = 0 \in \mathcal{A}(x_0)$ and $0 \in U \setminus \{a_1, \dots, a_k\}$. Then there exists a unique control $u(t)$ that steers x_0 to 0 and satisfies Pontryagin Maximum Principle.*

Proof. Assume that there exist two controls

$$u_1(t), \quad t \in [0, t_1], \qquad \text{and} \qquad u_2(t), \quad t \in [0, t_2],$$

that steer x_0 to 0 and satisfy PMP.

If $t_1 = t_2$, then the argument of the proof of preceding theorem shows that $u_1(t) \equiv u_2(t)$ a.e., so we can assume that

$$t_1 > t_2.$$

Cauchy's formula gives

$$e^{t_1 A}\left(x_0 + \int_0^{t_1} e^{-tA} B u_1(t)\, dt\right) = 0,$$
$$e^{t_2 A}\left(x_0 + \int_0^{t_2} e^{-tA} B u_2(t)\, dt\right) = 0,$$

thus

$$\int_0^{t_1} e^{-tA} B u_1(t)\, dt = \int_0^{t_2} e^{-tA} B u_2(t)\, dt. \tag{65}$$

According to PMP, there exists an adjoint vector $\xi_1(t)$, $t \in [0, t_1]$, such that

$$\xi_1(t) = \xi_1(0) e^{-tA}, \qquad \xi_1(0) \neq 0, \tag{66}$$
$$\xi_1(t) B u_1(t) = \max_{u \in U} \xi_1(t) B u. \tag{67}$$

Since $0 \in U$, then

$$\xi_1(t) B u_1(t) \geq 0, \qquad t \in [0, t_1]. \tag{68}$$

Equality (65) can be rewritten as

$$\int_0^{t_1} \xi_1(t)Bu_1(t)\,dt = \int_0^{t_2} \xi_1(t)Bu_2(t)\,dt. \tag{69}$$

Taking into account inequality (68), we obtain

$$\int_0^{t_2} \xi_1(t)Bu_1(t)\,dt \le \int_0^{t_2} \xi_1(t)Bu_2(t)\,dt. \tag{70}$$

But maximality condition (67) implies that

$$\xi_1(t)Bu_1(t) \ge \xi_1(t)Bu_2(t), \qquad t \in [0, t_2]. \tag{71}$$

Now inequalities (70) and (71) are compatible only if

$$\xi_1(t)Bu_1(t) = \xi_1(t)Bu_2(t), \qquad t \in [0, t_2],$$

thus inequality (70) should turn into equality. In view of (69), we have

$$\int_{t_1}^{t_2} \xi_1(t)Bu_1(t)\,dt = 0.$$

Since the integrand is nonnegative, see (68), then it vanishes identically:

$$\xi_1(t)Bu_1(t) \equiv 0, \qquad t \in [t_1, t_2].$$

By the argument of Theorem 5, the control $u_1(t)$ is bang-bang, so there exists an interval $I \subset [t_1, t_2]$ such that

$$u_1(t)|_I \equiv a_j \neq 0.$$

Thus

$$\xi_1(t)Ba_j \equiv 0, \qquad t \in I.$$

But $\xi_1(t)0 \equiv 0$, this is a contradiction with uniqueness of the control for which maximum in PMP is obtained, see Corollary 4. □

5.5 Switchings of optimal control

Now we evaluate the number of switchings of optimal control in linear time-optimal problems. In the examples of Sections 4.1, 4.2 we had respectively one switching and an arbitrarily large number of switchings, although finite

on any segment. It turns out that in general there are two cases: non-oscillating and oscillating, depending on whether the matrix A of the control system has real spectrum or not. Recall that in the example with one switching, Section 4.1, we had

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad \mathrm{Sp}(A) = \{0\} \subset \mathbb{R},$$

and in the example with arbitrarily large number of switchings, Section 4.2,

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathrm{Sp}(A) = \{\pm i\} \not\subset \mathbb{R}.$$

We consider systems with scalar control:

$$\dot{x} = Ax + ub, \qquad u \in U = [\alpha, \beta] \subset \mathbb{R}, \quad x \in \mathbb{R}^n,$$

under the General Position Condition

$$\mathrm{span}(b, Ab, \dots, A^{n-1}b) = \mathbb{R}^n.$$

Then attainable set of the system is full-dimensional for arbitrarily small times. We can evaluate the minimal number of switchings necessary to fill a full-dimensional domain. Optimal control is piecewise constant with values in $\{\alpha, \beta\}$. Assume that we start from the initial point x_0 with the control α. Without switchings we fill a piece of a 1-dimensional curve $e^{(Ax+\alpha b)t}x_0$, with 1 switching we fill a piece of a 2-dimensional surface $e^{(Ax+\beta b)t_2} \circ e^{(Ax+\alpha b)t_1}x_0$, with 2 switchings we can attain points in a 3-dimensional surface, etc. So the minimal number of switchings required to reach an n-dimensional domain is $n-1$.

We prove now that in the non-oscillating case we never need more than $n-1$ switchings of optimal control.

Theorem 8. *Assume that the matrix A has only real eigenvalues:*

$$\mathrm{Sp}(A) \subset \mathbb{R}.$$

Then any optimal control in linear time-optimal problem (58) *has no more than $n-1$ switchings.*

Proof. Let $u(t)$ be an optimal control and $\xi(t) = \xi(0)e^{-tA}$ the corresponding solution of the adjoint equation $\dot{\xi} = -\xi A$. The maximality condition of PMP reads

$$\xi(t)bu(t) = \max_{u\in[\alpha,\beta]} \xi(t)bu,$$

thus

$$u(t) = \begin{cases} \beta & \text{if } \xi(t)b > 0, \\ \alpha & \text{if } \xi(t)b < 0. \end{cases}$$

So the number of switchings of the control $u(t)$, $t \in [0, t_1]$, is equal to the number of changes of sign of the function

$$y(t) = \xi(t)b, \qquad t \in [0, t_1].$$

We show that $y(t)$ has not more than $n-1$ real roots.

Derivatives of the adjoint vector have the form

$$\xi^{(k)}(t) = \xi(0)e^{-tA}(-A)^k.$$

By Cayley Theorem, the matrix A satisfies its characteristic equation:

$$A^n + c_1A^{n-1} + \cdots + c_n \operatorname{Id} = 0,$$

where

$$\det(t \operatorname{Id} - A) = t^n + c_1t^{n-1} + \cdots + c_n,$$

thus

$$(-A)^n - c_1(-A)^{n-1} + \cdots + (-1)^n c_n \operatorname{Id} = 0.$$

Then the function $y(t)$ satisfies an n-th order ODE:

$$y^{(n)}(t) - c_1y^{(n-1)}(t) + \cdots + (-1)^n c_n y(t) = 0. \tag{72}$$

It is well known that any solution of this equation is a quasipolynomial:

$$y(t) = \sum_{i=1}^{k} e^{-\lambda_i t} P_i(t),$$

$P_i(t)$ a polynomial,

$\lambda_i \neq \lambda_j$ for $i \neq j$,

where λ_i are eigenvalues of the matrix A and degree of each polynomial P_i is less than multiplicity of the corresponding eigenvalue λ_i, thus

$$\sum_{i=1}^{k} \deg P_i \le n - k.$$

Now the statement of this theorem follows from the next general lemma. □

Lemma 2. *A quasipolynomial*

$$y(t) = \sum_{i=1}^{k} e^{\lambda_i t} P_i(t), \qquad \sum_{i=1}^{k} \deg P_i \le n - k, \tag{73}$$
$$\lambda_i \ne \lambda_j \text{ for } i \ne j,$$

has no more than $n-1$ real roots.

Proof. Apply induction on k.

If $k = 1$, then a quasipolynomial

$$y(t) = e^{\lambda t} P(t), \qquad \deg P \le n - 1,$$

has no more than $n - 1$ roots.

We prove the induction step for $k > 1$. Denote

$$n_i = \deg P_i, \qquad i = 1, \dots, k.$$

Suppose that the quasipolynomial $y(t)$ has n real roots. Rewrite the equation

$$y(t) = \sum_{i=1}^{k-1} e^{\lambda_i t} P_i(t) + e^{\lambda_k t} P_k(t) = 0$$

as follows:

$$\sum_{i=1}^{k-1} e^{(\lambda_i - \lambda_k)t} P_i(t) + P_k(t) = 0. \tag{74}$$

The quasipolynomial in the left-hand side has n roots. We differentiate this quasipolynomial successively $(n_k + 1)$ times so that the polynomial $P_k(t)$ disappear. After $(n_k + 1)$ differentiations we obtain a quasipolynomial

$$\sum_{i=1}^{k-1} e^{(\lambda_i - \lambda_k)t} Q_i(t), \qquad \deg Q_i \le \deg P_i,$$

which has $(n - n_k - 1)$ real roots by Rolle's Theorem. But by induction assumption the maximal possible number of real roots of this quasipolynomial is

$$\sum_{i=1}^{k-1} n_i + k - 2 < n - n_k - 1.$$

The contradiction finishes the proof of the lemma. □

So we completed the proof of Theorem 8: in the non-oscillating case an optimal control has no more than $n - 1$ switchings on the whole domain (recall that $n-1$ switchings are always necessary even on short time segments since the attainable sets $\mathcal{A}_{q_0}(t)$ are full-dimensional for all $t > 0$).

For an arbitrary matrix A, one can obtain the upper bound of $(n - 1)$ switchings for sufficiently short intervals of time.

Theorem 9. *Consider the characteristic polynomial of the matrix A:*

$$\det(t\,\mathrm{Id} - A) = t^n + c_1 t^{n-1} + \cdots + c_n,$$

and let

$$c = \max_{1 \le i \le n} |c_i|.$$

Then for any time-optimal control $u(t)$ and any $\bar{t} \in \mathbb{R}$, the real segment

$$\left[\bar{t}, \bar{t} + \ln\left(1 + \frac{1}{c}\right)\right]$$

contains not more than $(n - 1)$ switchings of an optimal control $u(t)$.

In the proof of this theorem we will require the following general proposition, which I learned from S. Yakovenko.

Lemma 3. *Consider an ODE*

$$y^{(n)} + c_1(t) y^{(n-1)} + \cdots + c_n(t) y = 0$$

with measurable and bounded coefficients:

$$c_i = \max_{t \in [\bar{t}, \bar{t}+\delta]} |c_i(t)|.$$

If

$$\sum_{k=1}^{n} c_k \frac{\delta^k}{k!} < 1, \tag{75}$$

then any nonzero solution $y(t)$ of the ODE has not more than $n - 1$ roots on the segment $t \in [\bar{t}, \bar{t} + \delta]$.

Proof. By contradiction, suppose that the function $y(t)$ has at least n roots on the segment $t \in [\bar{t}, \bar{t}+\delta]$. By Rolle's Theorem, derivative $\dot{y}(t)$ has not less than $n-1$ roots, etc. Then $y^{(n-1)}$ has a root $t_{n-1} \in [\bar{t}, \bar{t}+\delta]$. Thus

$$y^{(n-1)}(t) = \int_{t_{n-1}}^{t} y^{(n)}(\tau)\, d\tau.$$

Let $t_{n-2} \in [\bar{t}, \bar{t}+\delta]$ be a root of $y^{(n-2)}(t)$, then

$$y^{(n-2)}(t) = \int_{t_{n-2}}^{t} d\tau_1 \int_{t_{n-1}}^{\tau_1} y^{(n)}(\tau_2)\, d\tau_2.$$

We continue this procedure by integrating $y^{(n-i+1)}(t)$ from a root $t_{n-i} \in [\bar{t}, \bar{t}+\delta]$ of $y^{(n-i)}(t)$ and obtain

$$y^{(n-i)}(t) = \int_{t_{n-i}}^{t} d\tau_1 \int_{t_{n-i+1}}^{\tau_1} d\tau_2 \cdots \int_{t_{n-1}}^{\tau_{i-1}} y^{(n)}(\tau_i)\, d\tau_i, \qquad i = 1, \ldots, n.$$

There holds a bound:

$$\begin{aligned}
\left| y^{(n-i)}(t) \right| &\le \int_{t_{n-i}}^{t} d\tau_1 \int_{t_{n-i+1}}^{\tau_1} d\tau_2 \cdots \int_{t_n - 1}^{\tau_{i-1}} \left| y^{(n)}(\tau_i) \right| d\tau_i \\
&\le \int_{\bar{t}}^{\bar{t}+\delta} d\tau_1 \int_{\bar{t}}^{\tau_1} d\tau_2 \cdots \int_{\bar{t}}^{\tau_{i-1}} \left| y^{(n)}(\tau_i) \right| d\tau_i \le \frac{\delta^k}{k!} \sup_{t \in [\bar{t}, \bar{t}+\delta]} \left| y^{(n)}(t) \right|.
\end{aligned}$$

Then

$$\left| \sum_{i=1}^{n} c_i(t) y^{(n-i)}(t) \right| \le \sum_{i=1}^{n} |c_i(t)| \left| y^{(n-i)}(t) \right| \le \sum_{i=1}^{n} c_i \frac{\delta^k}{k!} \sup_{t \in [\bar{t}, \bar{t}+\delta]} \left| y^{(n)}(t) \right|,$$

i.e.,

$$\left| y^{(n)}(t) \right| \le \sum_{i=1}^{n} c_i \frac{\delta^k}{k!} \sup_{t \in [\bar{t}, \bar{t}+\delta]} \left| y^{(n)}(t) \right|,$$

a contradiction with (75). The lemma is proved. □

Now we prove Theorem 9.

Proof. As we showed in the proof of Theorem 8, the number of switchings of $u(t)$ is not more than the number of roots of the function $y(t) = \xi(t)b$, which satisfies ODE (72).

We have

$$\sum_{k=1}^{n} |c_k| \frac{\delta^k}{k!} < c(e^\delta - 1) \quad \forall \delta > 0.$$

By Lemma 3, if

$$c(e^\delta - 1) \leq 1, \tag{76}$$

then the function $y(t)$ has not more than $n-1$ real roots on any interval of length δ. But inequality (76) is equivalent to the following one:

$$\delta \leq \ln\left(1 + \frac{1}{c}\right),$$

so $y(t)$ has not more than $n-1$ roots on any interval of the length $\ln\left(1+\frac{1}{c}\right)$. □

6 Linear-quadratic problem

6.1 Problem statement and assumptions

In this chapter we study a class of optimal control problems very popular in applications, *linear-quadratic problems*. That is, we consider linear systems with quadratic cost functional:

$$\begin{aligned} &\dot{x} = Ax + Bu, \qquad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m, \\ &x(0) = x_0, \quad x(t_1) = x_1, \qquad x_0, x_1, t_1 \text{ fixed}, \\ &J(u) = \frac{1}{2} \int_0^{t_1} \langle Ru(t), u(t) \rangle + \langle Px(t), u(t) \rangle + \langle Qx(t), x(t) \rangle \, dt \to \min. \end{aligned} \tag{77}$$

Here A, B, R, P, Q are constant matrices of appropriate dimensions, R, Q are symmetric:

$$R^* = R, \qquad Q^* = Q,$$

and angle brackets $\langle \cdot, \cdot \rangle$ denote the standard inner product in $\mathbb{R}^m$ and $\mathbb{R}^n$.

One can show that the condition $R \geq 0$ is necessary for existence of optimal control. We do not touch here the case of degenerate R and assume that $R > 0$. The substitution of variables $u \mapsto v = R^{1/2}u$ transforms the functional $J(u)$ to a similar functional with the identity matrix instead of R. That is why we assume in the sequel that $R = \mathrm{Id}$. Another change of

variables kills the matrix P (exercise: find this change of variables). So we can write the cost functional as follows:

$$J(u) = \frac{1}{2}\int_0^{t_1} |u(t)|^2 + \langle Qx(t), x(t)\rangle \, dt.$$

For dynamics of the problem, we assume that the linear system is controllable:

$$\operatorname{rank}(B, AB, \dots, A^{n-1}B) = n. \tag{78}$$

6.2 Existence of optimal control

Since the set of control parameters $U = \mathbb{R}^m$ is noncompact, Filippov's Theorem does not apply, and existence of optimal controls in linear-quadratic problems is a nontrivial problem.

In this chapter we assume that admissible controls are square-integrable:

$$u \in L_2^m[0, t_1]$$

and use the L_2^m norm for controls:

$$\|u\| = \left(\int_0^{t_1} |u(t)|^2 \, dt\right)^{1/2} = \left(\int_0^{t_1} u_1^2(t) + \dots + u_m^2(t) \, dt\right)^{1/2}.$$

Consider the set of all admissible controls that steer the initial point to the terminal one:

$$U(x_0, x_1) = \{u \in L_2^m[0, t_1] \mid x(t_1, u, x_0) = x_1\}.$$

We denote by $x(t, u, x_0)$ the trajectory of system (77) corresponding to an admissible control $u \in L_2^m$ starting at a point $x_0 \in \mathbb{R}^n$. By Cauchy's formula, the endpoint mapping

$$u \mapsto x(t_1, u, x_0) = e^{t_1 A}x_0 + \int_0^{t_1} e^{(t_1-\tau)A}Bu(\tau) \, d\tau$$

is an affine mapping from $L_2^m[0, t_1]$ to $\mathbb{R}^n$. Controllability of the linear system (77) means that for any $x_0 \in \mathbb{R}^n$, $t_1 > 0$, the image of the endpoint mapping is the whole $\mathbb{R}^n$. Thus

$$U(x_0, x_1) \subset L_2^m[0, t_1]$$

is an affine subspace,

$$U(0,0) \subset L_2^m[0,t_1]$$

is a linear subspace, and

$$U(x_0,x_1) = u + U(0,0) \quad \text{for any } u \in U(x_0,x_1).$$

Thus it is natural that existence of optimal controls is closely related to behavior of the cost functional $J(u)$ on the linear subspace $U(0,0)$.

Proposition 2. (1) *If there exist points $x_0, x_1 \in \mathbb{R}^n$ such that*

$$\inf_{u\in U(x_0,x_1)} J(u) > -\infty, \tag{79}$$

then

$$J(u) \geq 0 \quad \forall u \in U(0,0).$$

(2) *Conversely, if*

$$J(u) > 0 \quad \forall u \in U(0,0) \setminus 0,$$

then the minimum is attained:

$$\exists \min_{u\in U(x_0,x_1)} J(u) \quad \forall x_0,\ x_1 \in \mathbb{R}^n.$$

Remark. That is, the inequality

$$J|_{U(0,0)} \geq 0$$

is necessary for existence of optimal controls, at least for one pair (x_0,x_1), and the strict inequality

$$J|_{U(0,0)\setminus 0} > 0$$

is sufficient for existence of optimal controls for all pairs (x_0,x_1).

In the proof of Proposition 2, we will need the following auxiliary proposition.

Lemma 4. *If $J(v) > 0$ for all $v \in U(0,0) \setminus 0$, then*

$$J(v) \geq \alpha\|v\|^2 \qquad \text{for some } \alpha > 0 \text{ and all } v \in U(0,0),$$

or, which is equivalent,

$$\inf\{J(v) \mid \|v\| = 1,\ v \in U(0,0)\} > 0.$$

Proof. Let v_n be a minimizing sequence of the functional $J(v)$ on the sphere $\{\|v\| = 1\} \cap U(0,0)$. Closed balls in Hilbert spaces are weakly compact, thus we can find a subsequence weakly converging in the unit ball and preserve the notation v_n for its terms, so that

$$\begin{aligned} &v_n \to \widehat{v} \text{ weakly as } n \to \infty, \qquad \|\widehat{v}\| \le 1, \quad \widehat{v} \in U(0,0), \\ &J(v_n) \to \inf\{J(v) \mid \|v\| = 1,\ v \in U(0,0)\}, \qquad n \to \infty. \end{aligned} \tag{80}$$

We have

$$J(v_n) = \frac{1}{2} + \frac{1}{2}\int_0^{t_1} \langle Qx_n(\tau), x_n(\tau)\rangle \, d\tau.$$

Since the controls converge weakly, then the corresponding trajectories converge strongly:

$$x_n(\cdot) \to x_{\widehat{v}}(\cdot), \qquad n \to \infty,$$

thus

$$J(v_n) \to \frac{1}{2} + \frac{1}{2}\int_0^{t_1} \langle Qx_{\widehat{v}}(\tau), x_{\widehat{v}}(\tau)\rangle \, d\tau, \qquad n \to \infty.$$

In view of (80), the infimum in question is equal to

$$\frac{1}{2} + \frac{1}{2}\int_0^{t_1} \langle Qx_{\widehat{v}}(\tau), x_{\widehat{v}}(\tau)\rangle \, d\tau = \frac{1}{2}\left(1 - \|\widehat{v}\|^2\right) + J(\widehat{v}) > 0.$$

□

Now we prove Proposition 2.

Proof. (1) By contradiction, suppose that there exists $v \in U(0,0)$ such that $J(v) < 0$. Take any $u \in U(x_0, x_1)$, then $u + sv \in U(x_0, x_1)$ for any $s \in \mathbb{R}$.

Let $y(t)$, $t \in [0, t_1]$, be the solution to the Cauchy problem

$$\dot{y} = Ay + Bv, \qquad y(0) = 0,$$

and

$$J(u, v) = \frac{1}{2}\int_0^{t_1} \langle u(\tau), v(\tau)\rangle + \langle Qx(\tau), y(\tau)\rangle \, d\tau.$$

Then the quadratic functional J on the family of controls $u + sv$, $s \in \mathbb{R}$, is computed as follows:

$$J(u + sv) = J(u) + 2sJ(u, v) + s^2 J(v).$$

Since $J(v) < 0$, then $J(u + sv) \to -\infty$ as $s \to \infty$. The contradiction with hypothesis (79) finishes the proof of item (1) of this proposition.

(2) We have

$$J(u) = \frac{1}{2}\|u\|^2 + \frac{1}{2}\int_0^{t_1} \langle Qx(\tau), x(\tau)\rangle \, d\tau.$$

The norm $\|u\|$ is lower semicontinuous in the weak topology on L_2^m, and the functional $\int_0^{t_1} \langle Qx(\tau), x(\tau)\rangle \, d\tau$ is weakly continuous on L_2^m. Thus $J(u)$ is weakly lower semicontinuous on L_2^m. Since balls are weakly compact in L_2^m and the affine subspace $U(x_0, x_1)$ is weakly compact, it is enough to prove that $J(u) \to \infty$ when $u \to \infty$, $u \in U(x_0, x_1)$.

Take any control $u \in U(x_0, x_1)$. Then for any $v \in U(0,0) \setminus 0$, the control $u + v$ belongs to $U(x_0, x_1)$ and

$$J(u + v) = J(u) + 2\|v\| J\left(u, \frac{v}{\|v\|}\right) + J(v).$$

Denote $J(u) = C_0$. Further, $\left|J\left(u, \frac{v}{\|v\|}\right)\right| \le C_1 = \text{const}$ for all $v \in U(0,0) \setminus 0$. Finally, by Lemma 4, $J(v) \ge \alpha\|v\|^2$, $\alpha > 0$, for all $v \in U(0,0) \setminus 0$. Consequently,

$$J(u + v) \ge C_0 - 2\|v\|C_1 + \alpha\|v\|^2 \to \infty, \qquad v \to \infty, \ v \in U(0,0).$$

Item (2) of this proposition follows. □

So we reduced the question of existence of optimal controls in linear-quadratic problems to the study of the restriction $J|_{U(0,0)}$. We will consider this restriction in detail later.

6.3 Extremals

Now we write PMP for linear-quadratic problems. The control-dependent Hamiltonian is

$$h_u(\xi, x) = \xi Ax + \xi Bu - \frac{\nu}{2}(\|u\|^2 + \langle Qx, x\rangle), \qquad x \in \mathbb{R}^n, \ \xi \in \mathbb{R}^{n*}.$$

Consider first the abnormal case:

$$\nu = 0.$$

By PMP, adjoint vector along an extremal satisfies the ODE $\dot{\xi} = -\xi A$, thus $\xi(t) = \xi(0)e^{-tA}$. The maximality condition

$$\xi(t)Bu(t) = \max_{u\in\mathbb{R}^n} \xi(t)Bu \tag{81}$$

implies that

$$0 \equiv \xi(t)B = \xi(0)e^{-tA}B.$$

We differentiate this identity $n-1$ times, take into account the controllability condition (78) and obtain $\xi(0) = 0$. This contradicts PMP, thus there are no abnormal extremals.

In the sequel we consider the normal case: $\nu \neq 0$, thus we can assume

$$\nu = 1.$$

Then the control-dependent Hamiltonian takes the form

$$h_u(\xi, x) = \xi Ax + \xi Bu - \frac{1}{2}(\|u\|^2 + \langle Qx, x\rangle), \qquad x \in \mathbb{R}^n,\ \xi \in \mathbb{R}^{n*}.$$

The term $\xi Bu - \frac{1}{2}\|u\|^2$ depending on u has a unique maximum in $u \in \mathbb{R}^m$ at the point where

$$\frac{\partial h_u}{\partial u} = \xi B - u^* = 0,$$

thus

$$u = B^*\xi^*.$$

So the maximized Hamiltonian is

$$\begin{aligned} H(\xi, x) &= \max_{u\in\mathbb{R}^m} h_u(\xi, x) = \xi Ax - \frac{1}{2}\langle Qx, x\rangle + \frac{1}{2}|B^*\xi^*|^2 \\ &= \xi Ax - \frac{1}{2}\langle Qx, x\rangle + \frac{1}{2}|B\xi|^2. \end{aligned}$$

The Hamiltonian function $H(\xi, x)$ is smooth, thus extremals are solutions of the corresponding Hamiltonian system

$$\begin{cases} \dot{x} = Ax + BB^*\xi^*, \\ \dot{\xi} = x^*Q - \xi A. \end{cases}$$

6.4 Conjugate points

Now we study conditions of existence and uniqueness of optimal controls depending upon the terminal time. So we write the cost functional to be minimized as follows:

$$J_t(u) = \frac{1}{2}\int_0^t |u(\tau)|^2 + \langle Qx(\tau), x(\tau)\rangle\, d\tau.$$

Denote

$$U_t(0,0) = \{u \in L_2^m[0,t] \mid x(t,u,x_0) = x_1\},$$
$$\mu(t) \stackrel{\text{def}}{=} \inf\{J_t(u) \mid u \in U_t(0,0),\ \|u\| = 1\}. \tag{82}$$

We showed in Proposition 2 that if $\mu(t) > 0$ then the problem has solution for any boundary conditions, and if $\mu(t) < 0$ then there are no solutions for any boundary conditions. The case $\mu(t) = 0$ is doubtful. Now we study properties of the function $\mu(t)$ in detail.

Proposition 3. (1) *The function $t \mapsto \mu(t)$ is monotone nonincreasing and continuous.*

(2)

$$1 \geq 2\mu(t) \geq 1 - \frac{t^2}{2} e^{2t\|A\|}\, \|B\|^2 \|Q\|. \tag{83}$$

(3) *If $1 > 2\mu(t)$, then the infimum in* (82) *is attained, i.e., it is minimum.*

Proof. (3) Denote

$$I_t(u) = \frac{1}{2}\int_0^t \langle Qx(\tau), x(\tau)\rangle\, d\tau,$$

the functional $I_t(u)$ is weakly continuous on L_2^m. Notice that

$$J_t(u) = \frac{1}{2} + I_t(u) \quad \text{on the sphere } \|u\| = 1.$$

Take a minimizing sequence of the functional $I_t(u)$ on the sphere $\{\|u\| = 1\} \cap U_t(0,0)$. Since the ball $\{\|u\| \leq 1\}$ is weakly compact, we can find a weakly converging subsequence:

$$u_n \to \widehat{u} \text{ weakly as } n \to \infty, \qquad \|\widehat{u}\| \leq 1, \quad \widehat{u} \in U_t(0,0),$$
$$I_t(u_n) \to I_t(\widehat{u}) = \inf\{I_t(u) \mid \|u\| = 1,\ u \in U_t(0,0)\}, \qquad n \to \infty.$$

If $\widehat{u} = 0$, then $I_t(\widehat{u}) = 0$, thus $\mu(t) = \frac{1}{2}$, which contradicts hypothesis of item (3).

So $\widehat{u} \neq 0$, $I_t(\widehat{u}) < 0$, and $I_t\left(\frac{\widehat{u}}{\|\widehat{u}\|}\right) \leq I_t(\widehat{u})$. Thus $\|\widehat{u}\| = 1$, and $J_t(u)$ attains minimum on the sphere $\{\|u\| = 1\} \cap U_t(0,0)$ at the point $\widehat{u}$.

(2) Let $\|u\| = 1$ and $x_0 = 0$. By Cauchy's formula,

$$x(t) = \int_0^t e^{(t-\tau)A} B u(\tau)\, d\tau,$$

thus

$$|x(t)| \leq \int_0^t e^{(t-\tau)\|A\|}\, \|B\| \cdot |u(\tau)|\, d\tau$$

by Cauchy-Schwartz inequality

$$\begin{aligned} &\leq \|u\| \left(\int_0^t e^{(t-\tau)2\|A\|}\, \|B\|^2\, d\tau \right)^{1/2} \\ &= \left(\int_0^t e^{(t-\tau)2\|A\|}\, \|B\|^2\, d\tau \right)^{1/2}. \end{aligned}$$

We substitute this estimate of $x(t)$ into J_t and obtain the second inequality in (83).

The first inequality in (83) is obtained by considering a weakly converging sequence $u_n \to 0$, $n \to \infty$, in the sphere $\|u_n\| = 1$, $u_n \in U_t(0,0)$.

(1) Monotonicity of $\mu(t)$. Take any $\hat{t} > t$. Then the space $U_t(0,0)$ is isometrically embedded into $U_{\hat{t}}(0,0)$ by extending controls $u \in U_t(0,0)$ by zero:

$$\begin{aligned} u \in U_t(0,0) \quad &\Rightarrow \quad \widehat{u} \in U_{\hat{t}}(0,0), \\ \widehat{u}(\tau) = &\begin{cases} u(\tau), & \tau \leq t, \\ 0, & \tau > t. \end{cases} \end{aligned}$$

Moreover,

$$J_{\hat{t}}(\widehat{u}) = J_t(u).$$

Thus

$$\begin{aligned} \mu(t) &= \inf\{J_t(u) \mid u \in U_t(0,0), \|u\| = 1\} \\ &\geq \inf\{J_{\hat{t}}(u) \mid u \in U_{\hat{t}}(0,0), \|u\| = 1\} = \mu(\hat{t}). \end{aligned}$$

Continuity of $\mu(t)$: we show separately continuity from the right and from the left.

Continuity from the right. Let $t_n \searrow t$. We can assume that $\mu(t_n) < \frac{1}{2}$ (otherwise $\mu(t_n) = \mu(t) = \frac{1}{2}$), thus minimum in (82) is attained:

$$\mu(t_n) = \frac{1}{2} + I_{t_n}(u_n), \qquad u_n \in U_{t_n}(0,0), \quad \|u_n\| = 1.$$

Extend the functions $u_n \in L_2^m[0,t_n]$ to the segment $[0,t]$ by zero. Choosing a weakly converging subsequence in the unit ball, we can assume that

$$u_n \to u \text{ weakly as } n \to \infty, \qquad u \in U_t(0,0), \quad \|u_n\| \le 1,$$

thus

$$I_{t_n}(u_n) \to I_t(u) \ge \inf\{I_t(v) \mid v \in U_t(0,0), \quad \|v\| = 1\}, \quad t_n \searrow t.$$

Then

$$\mu(t) \le \frac{1}{2} + \lim_{t_n \searrow t} I_{t_n}(u_n) = \lim_{t_n \searrow t} \mu(t_n).$$

By monotonicity of μ,

$$\mu(t) = \lim_{t_n \searrow t} \mu(t_n),$$

i.e., continuity from the right is proved.

Continuity from the left. We can assume that $\mu(t) < \frac{1}{2}$ (otherwise $\mu(\tau) = \mu(t) = \frac{1}{2}$ for $\tau < t$). Thus minimum in (82) is attained:

$$\mu(t) = \frac{1}{2} + I_t(\widehat{u}), \qquad \widehat{u} \in U_t(0,0), \quad \|\widehat{u}\| = 1.$$

For the trajectory

$$\widehat{x}(\tau) = x(\tau, \widehat{u}, 0),$$

we have

$$\widehat{x}(\tau) = \int_0^\tau e^{(\tau-\theta)A} B\widehat{u}(\theta)\, d\theta.$$

Denote

$$\alpha(\varepsilon) = \| \widehat{u}|_{[0,\varepsilon]} \|$$

and notice that

$$\alpha(\varepsilon) \to 0, \qquad \varepsilon \to 0.$$

Denote the ball

$$B_\delta = \{u \in L_2^m \mid \|u\| \le \delta, \ u \in U(0,0)\}.$$

Obviously,

$$x(\varepsilon, B_{\alpha(\varepsilon)}, 0) \ni \widehat{x}(\varepsilon).$$

The mapping $u \mapsto x(\varepsilon, u(\cdot), 0)$ from L_2^m to $\mathbb{R}^n$ is linear, and the system $\dot{x} = Ax + Bu$ is controllable, thus $x(\varepsilon, B_{\alpha(\varepsilon)}, 0)$ is a convex full-dimensional set in $\mathbb{R}^n$ such that the positive cone generated by this set is the whole $\mathbb{R}^n$. That is why

$$x(\varepsilon, 2B_{\alpha(\varepsilon)}, 0) = 2x(\varepsilon, B_{\alpha(\varepsilon)}, 0) \supset O_{x(\varepsilon, B_{\alpha(\varepsilon)}, 0)}$$

for some neighborhood $O_{x(\varepsilon, B_{\alpha(\varepsilon)}, 0)}$ of the set $x(\varepsilon, B_{\alpha(\varepsilon)}, 0)$. Further, there exists an instant $t_\varepsilon > \varepsilon$ such that

$$\widehat{x}(t_\varepsilon) \in x(\varepsilon, 2B_{\alpha(\varepsilon)}, 0),$$

consequently,

$$\widehat{x}(t_\varepsilon) = x(\varepsilon, v_\varepsilon, 0), \qquad \|v_\varepsilon\| \le 2\alpha(\varepsilon).$$

Consider the following family of controls that approximate $\widehat{u}$:

$$u_\varepsilon(\tau) = \begin{cases} v_\varepsilon(\tau), & 0 \le \tau \le t_\varepsilon, \\ \widehat{u}(\tau + t_\varepsilon - \varepsilon), & t_\varepsilon < \tau \le t + \varepsilon - t_\varepsilon. \end{cases}$$

We have

$$\begin{aligned} &u_\varepsilon \in U_{t+\varepsilon-t_\varepsilon}(0,0), \\ &\|\widehat{u} - u_\varepsilon\| \to 0, \qquad \varepsilon \to 0. \end{aligned}$$

But $t + \varepsilon - t_\varepsilon < t$ and μ is nonincreasing, thus it is continuous from the left.

Continuity from the right was already proved, hence μ is continuous. □

Now we prove that the function μ can have not more than one root.

Proposition 4. *If $\mu(t) = 0$ for some $t > 0$, then $\mu(\tau) < 0$ for all $\tau > t$.*

Proof. Let $\mu(t) = 0$, $t > 0$. By Proposition 3, infimum in (82) is attained at some control $\widehat{u} \in U_t(0,0)$, $\|\widehat{u}\| = 1$:

$$\begin{aligned} \mu(t) &= \min\{J_t(u) \mid u \in U_t(0,0),\ \|u\| = 1\} \\ &= J_t(\widehat{u}) = 0. \end{aligned}$$

Then

$$J_t(u) \ge J_t(\widehat{u}) = 0 \quad \forall u \in U_t(0,0),$$

i.e., the control $\widehat{u}$ is optimal, thus it satisfies PMP. There exists a solution $(\xi(\tau), x(\tau))$, $\tau \in [0,t]$, of the Hamiltonian system

$$\begin{cases} \dot{\xi} = x^* Q - \xi A, \\ \dot{x} = Ax + BB^* \xi, \end{cases}$$

with the boundary conditions

$$x(0) = x(t) = 0,$$

and

$$u(\tau) = B^* \xi^*(\tau), \qquad \tau \in [0,t].$$

We proved that for any root t of the function μ, any control $u \in U_t(0,0)$, $\|u\| = 1$, with $J_t(u) = 0$ satisfies PMP.

Now we prove that $\mu(\tau) < 0$ for all $\tau > t$. By contradiction, suppose that the function μ vanishes at some instant $t' > t$. Since μ is monotone, then

$$\mu|_{[t,t']} \equiv 0.$$

Consequently, the control

$$u'(\tau) = \begin{cases} \widehat{u}(\tau), & \tau \le t, \\ 0, & \tau \in [t,t'], \end{cases}$$

satisfies the conditions:

$$\begin{aligned} &u' \in U_{t'}(0,0), \quad \|u'\| = 1, \\ &J_{t'}(u') = 0. \end{aligned}$$

Thus u' satisfies PMP, i.e.,

$$u'(\tau) B^* \xi^{*'}(\tau), \qquad \tau \in [0,t'],$$

is an analytic function. But $u'|_{[t,t']} \equiv 0$, thus $u' \equiv 0$, a contradiction with $\|u'\| = 1$. □

It would be nice to have a way to solve the equation $\mu(t) = 0$ without performing the minimization procedure in (82). This can be done in terms of the following notion.

Definition 1. A point $t > 0$ is *conjugate* to 0 for the linear-quadratic problem in question if there exists a nontrivial solution $(\xi(\tau), x(\tau))$ of the Hamiltonian system

$$\begin{cases} \dot{\xi} = x^* Q - \xi A, \\ \dot{x} = Ax + BB^* \xi \end{cases}$$

such that $x(0) = x(t) = 0$.

Proposition 5. *The function μ vanishes at a point $t > 0$ if and only if t is the closest to 0 conjugate point.*

Proof. Let $\mu(t) = 0$, $t > 0$. First of all, t is conjugate to 0, we showed this in the proof of Proposition 4.

Suppose that $t' > 0$ is conjugate to 0. Compute the functional $J_{t'}$ on the corresponding control $u(\tau) = B^* \xi^*(\tau)$, $\tau \in [0, t']$:

$$\begin{aligned} J_{t'}(u) &= \frac{1}{2} \int_0^{t'} \langle B^* \xi^*(\tau), B^* \xi^*(\tau) \rangle + \langle Qx(\tau), x(\tau) \rangle \, d\tau \\ &= \frac{1}{2} \int_0^{t'} \langle BB^* \xi^*(\tau), \xi^*(\tau) \rangle + \langle Qx(\tau), x(\tau) \rangle \, d\tau \\ &= \frac{1}{2} \int_0^{t'} \xi(\tau)(\dot{x}(\tau) - Ax(\tau)) + x^*(\tau) Q x(\tau) \, d\tau \\ &= \frac{1}{2} \int_0^{t'} (\xi \dot{x} + \dot{\xi} x) \, d\tau \\ &= \frac{1}{2} (\xi(t') x(t') - \xi(0) x(0)) = 0. \end{aligned}$$

Thus $\mu(t') \le J_{t'} \left(\frac{u}{\|u\|} \right) = 0$. Now the result follows since μ is nonincreasing. □

The first (closest to zero) conjugate point determines existence and uniqueness properties of optimal control in linear-quadratic problems.

Before the first conjugate point, optimal control exists and is unique for any boundary conditions (if there are two optimal controls, then their difference gives rise to a conjugate point).

At the first conjugate point, there is existence and nonuniqueness for some boundary conditions, and nonexistence for other boundary conditions.

And after the first conjugate point, the problem has no optimal solutions for any boundary conditions.

Exercises

1. Optimal U-turn of the Dubins car.
   Consider the system

$$\begin{cases} \dot{x}^1 = \cos\theta \\ \dot{x}^2 = \sin\theta \\ \dot{\theta} = u \end{cases} \qquad |u| \leq 1.$$

   Find a time-optimal control and trajectory for the boundary conditions: $z(0) = (0,0,0)$, $z(t_1) = (0,0,\pi)$, where $z = (x^1, x^2, \theta)$.

2. Time-optimal stabilization of the oscillator with friction.
   Consider the system

$$\begin{cases} \dot{x}^1 = x^2 \\ \dot{x}^2 = -x^1 - kx^2 + u \end{cases} \qquad |u| \leq 1.$$

   Design a time-optimal synthesis with the target $(x^1, x^2) = (0,0)$ for any friction coefficient $k > 0$.

3. Conjugate points.
   Consider the following linear-quadratic problem:

$$\min \int_0^T (u^2(t) - x^2(t))\, dt, \quad \ddot{x} = u.$$

   Find an approximate value (up to 0.01) of the nearest to zero conjugate point.

References

[1] A. A. Agrachev, Yu. L. Sachkov, *Lectures on geometric control theory*, Trieste, SISSA preprint, 2001.

[2] V. I. Arnold, *Ordinary differential equations*, Springer-Verlag, 1992.

[3] J. Hale, *Ordinary differential equations*, Robert E. Krieger Publishing Company, 1980.

[4] V. Jurdjevic, *Geometric control theory*, Cambridge University Press, 1997.

[5] L. S. Pontryagin, V. G. Boltyanskij, R. V. Gamkrelidze, E. F. Mishchenko, *The mathematical theory of optimal processes*, Oxford, Pergamon Press, 1964.

[6] R. T. Rocafellar, *Convex analysis*, Princeton, Princeton University Press, 1972.

Value Function in Optimal Control

Hélène Frankowska*

CNRS, CREA, Ecole Polytechnique, Paris, France

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028009

*franko@poly.polytechnique.fr

Abstract

These lecture notes concern the "value functions" in optimal control. The value function was intensively investigated since the end of 1950's, when Bellman introduced it for studying optimal control problems and related it to solutions to Hamilton-Jacobi equations. In the framework of Calculus of Variations, it was already considered by Carathéodory since 1902 and of differential games by Isaacs at the RAND Corporation in the late 1940's/50's. A number of books were written on this subject in the 1960's/80's in the case when the value function is continuously differentiable. It became clear however that in most situations, the value function is non smooth and for constrained problems even discontinuous. Carathéodory already observed this phenomenon in his habilitation introducing the method of characteristics for the related Hamilton-Jacobi equation and examples of control theory later confirmed this fact. Roughly speaking, the value function becomes discontinuous even for control systems with analytic right-hand sides and analytic cost whenever there are multiple optimal solutions. The developments of convex analysis by J. Moreau and R.T. Rockafellar, of non-smooth analysis by R.T. Rockafellar and F.H. Clarke, of set-valued analysis by a large group of scientists, of theory of first order PDE by S.N. Kruzkov, M.G. Crandall and P.L. Lions, of viability theory by J.-P. Aubin gave a new trend to adapt many known classical results to non smooth situations typical for nonlinear control and state constrained problems.

These notes are mostly based on the author's publications in the 1980's/ 2000 either alone or together with P. Cannarsa, S. Plaskacz, F. Rampazzo and R.B. Vinter. Due to the lack of space, with a regret, the bibliographical comments and a more complete bibliography are omitted. They can be found in already published articles and books. My apologies and a call for comprehension to all those who may be deceived by this fact.

The first section is devoted to Set-Valued Analysis. Roughly speaking, set-valued analysis is an extension of analysis dealing with subsets and set-valued maps instead of elements and functions. The second section indicates some links between control systems and differential inclusions. Section 3 deals with the value function of Mayer's problem. It discusses its regularity and qualitative properties of optimal solutions related to those of the value function. Relations between differentiability of the value function and uniqueness of optimal solutions are indicated as well. In Section 4 we introduce discontinuous solutions to Hamilton-Jacobi-Bellman equation and in Section 5 we discuss the method of characteristics for the Bolza optimal control problem, shocks of characteristics and related properties of matrix Riccati equations. Section 6 is devoted to the Hamilton-Jacobi equation for problems under state constraints.

Finally, the last, but not the least, I would like to thank students and colleagues that followed the course and who by their reactions and questions helped me to complete these lecture notes. My thanks are also due to the organizers of this course A.Agrachev, B.Jakubczyk and C.Lobry for their initiative and to the staff of ICTP for its kind help during my stay.

Contents

1 Preliminaries: Set-Valued Analysis **519**

1.1 Preliminaries 520
1.1.1 Limits of Sets 520
1.1.2 Tangent and Normal Cones to a Subset 521
1.1.3 Generalized Differentials of Non Smooth Functions . . 522
1.1.4 Semiconcave Functions 524
1.1.5 Subnormal Cones to the Epigraph 528
1.2 Regularity of Set-Valued Maps 529
1.3 Differential Inclusions 536
1.3.1 Filippov's Theorem 538
1.3.2 Relaxation Theorems 540
1.3.3 Infinitesimal Generator of Reachable Map 543
1.3.4 Variational Inclusions 545
1.3.5 Viability Theorem 545
1.4 Parametrization of Set-Valued Maps 547

2 Control Systems and Differential Inclusions **548**

2.1 Nonlinear Control Systems 550
2.1.1 Reduction to Differential Inclusion 551
2.1.2 Linearization 553
2.2 State Dependent Control Systems 554
2.2.1 Reduction to Differential Inclusion 554
2.2.2 Linearization 555
2.3 Linear Implicit Control Systems 556
2.4 Nonlinear Implicit Control Systems 559
2.4.1 Reduction to Differential Inclusion 560
2.4.2 Linearization of Implicit Systems 561

3 Value Function of Mayer's Problem **562**

3.1 Value Function 565
3.1.1 Mayer and Bolza Problems 565
3.1.2 Lipschitz Continuity of the Value Function 567
3.1.3 Optimal Feedback 570
3.2 Maximum Principle for Free End Point Problems 571
3.2.1 Adjoint System 571
3.2.2 Maximum Principle 574
3.3 Necessary and Sufficient Conditions for Optimality 574

3.3.1 Sufficient Conditions 574
3.3.2 Necessary and Sufficient Conditions 575
3.3.3 Co-state and Superdifferentials of Value 578
3.3.4 Hamiltonian System 580
3.3.5 Uniqueness of Optimal Solution and Differentiability of Value Function 582
3.4 Semiconcavity of Value Function 584
3.4.1 Differentiability along Optimal Solutions 588
3.4.2 Regularity of Optimal Feedback 589

4 Hamilton-Jacobi-Bellman Equation 591
4.1 Solutions to Hamilton-Jacobi Equation 592
4.2 Lower Semicontinuous Solutions 595
4.2.1 Lower Semicontinuous & Contingent Solutions 595
4.2.2 Monotone Behavior of Contingent Solutions 598
4.2.3 Value Function & Contingent Solutions 602
4.2.4 Regularity of Value Function at Boundary Points . . . 603
4.3 Viscosity Solutions 604

5 Value Function of Bolza Problem and Riccati Equations 607
5.1 Matrix Riccati Equations and Shocks 610
5.2 Matrix Riccati Equations 617
5.2.1 Comparison Theorems 617
5.2.2 Existence of Solutions 620
5.3 Value Function of Bolza Problem 621
5.3.1 Maximum Principle 623
5.3.2 Differentiability of Value Function and Uniqueness of Optimal Solutions 625
5.3.3 Smoothness of the Value Function 627
5.3.4 Problems with Concave-Convex Hamiltonians 632

6 Hamilton-Jacobi-Bellman Equation for Problems under State-Constraints 633
6.1 Constrained Hamilton-Jacobi-Bellman Equation 634
6.2 A Neighboring Feasible Trajectories Theorem 637
6.3 Proof of the Main Theorem 643

References 649

1 Preliminaries: Set-Valued Analysis

This Section is concerned with the *differential inclusion* (*multivalued equation*):

$$x'(t) \in F(t, x(t)), \quad x(t_0) = x_0 \tag{1}$$

Its investigation was initiated in the thirties by the Polish and French mathematicians Zaremba in [53], [54] and Marchaud [42], [43].

Control theory motivated the renewal of the interest to the differential inclusion (1) in the earlier sixties. Filippov [19] and Ważewski [52] have shown that under very mild assumptions the control system

$$x' = f(t, x, u(t)), \quad u(t) \in U \text{ is measurable}, \quad x(t_0) = x_0 \tag{2}$$

can be reduced to differential inclusion (1). This placed control systems in the framework of ordinary differential "equations" with the difference that the right-hand side of these equations is multivalued.

However, very fortunately, the development of differential inclusions followed the same route that ODEs. There are existence results of *Peano* and *Cauchy-Lipschitz* type. When F is Lipschitz, then solutions depend on the initial condition in a Lipschitz way. We can as well differentiate solutions with respect to the initial condition (and to obtain *variational inclusions* instead of variational equations.) The only, but very important difference, is due to the fact that the solution to (1) is *a set* (of absolutely continuous functions $x(\cdot)$ starting at x_0 and satisfying $x'(t) \in F(t, x(t))$ almost everywhere.) For this reason the set-valued analysis arguments [5] have to be used in an essential way to investigate differential inclusions. We provide here only some of the proofs and indicate the source where the others can be found.

In Subsection 1 we recall Painlevé-Kuratowski limits, tangents to sets and generalized derivatives of functions and in Subsection 2 definitions concerning regularity and differentiation of set-valued maps that we shall use. We also gather some results on measurability and integration. The detailed study of these topics can be found for instance in [5] together with bibliographical comments.

Subsection 3 is devoted to differential inclusions. We start by the fundamental Filippov theorem and its applications. This is more than an existence theorem *à la Cauchy-Lipschitz*, but implies the same kind of consequences than the Gronwall inequality. In particular, we can compare solutions under perturbations of dynamics and/or initial conditions, and, in this respect, this

theorem is particularly useful. We also discuss there a result due to Filippov and Ważewski which states that solutions to (1) are dense in solutions to the relaxed differential inclusion

$$x'(t) \in \overline{co}\, F(t, x(t)), \quad x(t_0) = x_0$$

This allows to extend the concept of infinitesimal generator to set-valued semigroups (reachable maps) and also to derive variational inclusions by differentiating solutions with respect to initial conditions.

Finally we state the very useful viability theorem for problems under state constraints. See [4] for many results of this theory.

A natural question do arise:

Can differential inclusion (1) *be reduced to control system* (2)?

This is not true in general and examples of "nonconvex" differential inclusions justify their study in the nonparametrized form. However, the answer is positive when F has convex images.

We state in Subsection 4 some theorems concerning parametrization of set-valued maps. Most of the results of this subsection are provided without proofs.

1.1 Preliminaries

1.1.1 Limits of Sets

Let X be a metric space supplied with a distance d. When K is a subset of X, we denote by

$$d_K(x) := d(x, K) := \inf_{y \in K} d(x, y)$$

the *distance from* x *to* K, where we set $d(x, \emptyset) := +\infty$. Limits of sets have been introduced by Painlevé in 1902, as it is reported by his student Zoretti. They have been popularized by Kuratowski in his famous book TOPOLOGIE and thus, often called *Kuratowski lower and upper limits* of sequences of sets.

Definition 1.1 *Let* $(K_n)_{n \in \mathbf{N}}$ *be a sequence of subsets of a metric space* X. *We say that the subset*

$$\mathrm{Limsup}_{n \to \infty} K_n := \left\{ x \in X \mid \liminf_{n \to \infty} d(x, K_n) = 0 \right\}$$

is the upper limit *of the sequence* K_n *and that the subset*

$$\text{Liminf}_{n\to\infty} K_n \;:=\; \{x \in X \;\mid\; \lim_{n\to\infty} d(x, K_n) = 0\}$$

is its lower limit. *A subset* K *is said to be the* limit *or the* set limit *of the sequence* K_n *if*

$$K \;=\; \text{Liminf}_{n\to\infty} K_n \;=\; \text{Limsup}_{n\to\infty} K_n \;=:\; \text{Lim}_{n\to\infty} K_n$$

Lower and upper limits are obviously *closed*. We also see at once that

$$\text{Liminf}_{n\to\infty} K_n \;\subset\; \text{Limsup}_{n\to\infty} K_n$$

and that the upper limits and lower limits of the subsets K_n and of their closures $\overline{K}_n$ do coincide, since $d(x, K_n) = d(x, \overline{K}_n)$.

Naturally, we can replace $\mathbf{N}$ by a metric (or even, topological) space X, and sequences of subsets $n \rightsquigarrow K_n$ by set-valued maps $x \rightsquigarrow F(x)$ (which associates with a point x a subset $F(x)$) and adapt the definition of upper and lower limits to this case, called the *continuous case*.

1.1.2 Tangent and Normal Cones to a Subset

We begin with a presentation of the contingent cones:

Definition 1.2 (Contingent Cones) *Let* $K \subset X$ *be a subset of a normed vector space* X *and* $x \in \overline{K}$ *belong to the closure of* K. *The* contingent cone $T_K(x)$ *is defined by*

$$T_K(x) \;:=\; \{v \mid \liminf_{h\to 0+} d_K(x + hv)/h = 0\} = \text{Limsup}_{h\to 0+} \frac{K - x}{h}$$

It follows from the definition that $T_K(x)$ *is a closed cone.*

It is very convenient to have the following characterization of this cone in terms of sequences:

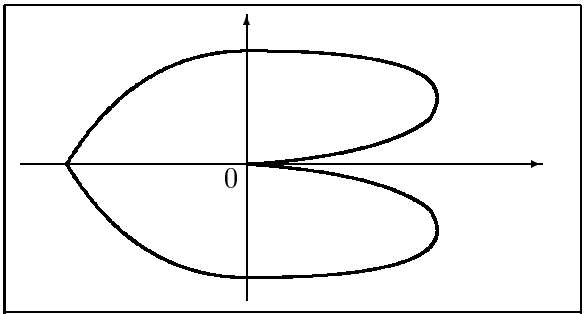
$$\begin{cases} v \;\in\; T_K(x) \;\text{ if and only if }\; \exists\, h_n \to 0+ \;\text{ and }\; \exists\, v_n \to v \\ \text{such that }\; \forall\, n, \;\; x + h_n v_n \in K \end{cases}$$

It implies that when K is convex, $T_K(x) = \overline{\bigcup_{\lambda\geq 0} \lambda(K - x)}$. We also observe that

$$\text{if} \;\; x \in \text{Int}(K), \;\; \text{then} \;\; T_K(x) \;=\; X$$

This situation may also happen when x does not belong to the interior of K (see Figure 1.)

Figure 1: Contingent Cone at a Boundary Point may be the Whole Space



Subset K such that $T_K(0) = X$

Theorem 1.3 *Let X be a finite dimensional vector-space and K be a closed subset of X. Then for every $x \in K$*

$$\text{Liminf}_{y \to_K x} T_K(y) \;=\; \text{Liminf}_{y \to_K x} \overline{co}(T_K(y)) \;\subset\; T_K(x)$$

See for instance [5] for the proof.

Definition 1.4 (Subnormal Cones) *Let $K \subset X$ be a subset of a normed vector space X and $x \in \overline{K}$ belong to the closure of K. The* subnormal cone *$N_K^0(x)$ is defined by*

$$N_K^0(x) \;:=\; \{p \in X^\star \mid \; < p, v > \leq 0 \;\; \forall\, v \in T_K(x)\}$$

1.1.3 Generalized Differentials of Non Smooth Functions

Definition 1.5 *Let X be a normed vector space, $\varphi : X \mapsto \mathbf{R} \cup \{\pm\infty\}$ be an extended function and $x_0 \in X$ be such that $\varphi(x_0) \neq \pm\infty$.*

The superdifferential *of φ at x_0 is the closed convex set defined by:*

$$\partial_+ \varphi(x_0) = \left\{ p \in \mathbf{R}^n \mid \limsup_{x \to x_0} \frac{\varphi(x) - \varphi(x_0) - < p, x - x_0 >}{\|x - x_0\|} \;\leq\; 0 \right\}$$

where $< \cdot, \cdot >$ denotes the scalar product.

The subdifferential *is defined in a similar way:*

$$\partial_- \varphi(x_0) = \left\{ p \in \mathbf{R}^n \mid \liminf_{x \to x_0} \frac{\varphi(x) - \varphi(x_0) - < p, x - x_0 >}{\|x - x_0\|} \;\geq\; 0 \right\}$$

We always have $\partial_+\varphi(x_0) = -\partial_-(-\varphi)(x_0)$.

The super and subdifferentials may also be characterized using *contingent epiderivatives*:

Definition 1.6 *Let X be a normed vector space, $\varphi : X \mapsto \mathbf{R} \cup \{\pm\infty\}$ be an extended function, $v \in X$ and $x_0 \in X$ be such that $\varphi(x_0) \neq \pm\infty$.*

The contingent epiderivative *of φ at x_0 in the direction v is given by*

$$D_\uparrow\varphi(x_0)(v) \;=\; \liminf_{h\to 0+,\; v'\to v} \frac{\varphi(x_0+hv') - \varphi(x_0)}{h}$$

and the contingent hypoderivative *of φ at x_0 in the direction v by*

$$D_\downarrow\varphi(x_0)(v) \;=\; \limsup_{h\to 0+,\; v'\to v} \frac{\varphi(x_0+hv') - \varphi(x_0)}{h}$$

Clearly

$$D_\uparrow\varphi(x_0) \;=\; -D_\downarrow(-\varphi)(x_0)$$

By a direct verification $D_\uparrow\varphi(x_0)$ is a lower semicontinuous map taking its values in $\mathbf{R} \cup \{\pm\infty\}$ whose epigraph is equal to the contingent cone to the epigraph of φ at $(x_0, \varphi(x_0))$.

When $\varphi : \mathbf{R}^n \mapsto \mathbf{R}$ is Lipschitz at x_0, then the contingent epi and hypoderivatives are reduced to the *Dini lower and upper derivatives*:

$$D_\uparrow\varphi(x_0)(v) \;=\; \liminf_{h\to 0+} \frac{\varphi(x_0+hv) - \varphi(x_0)}{h}$$

and

$$D_\downarrow\varphi(x_0)(v) \;=\; \limsup_{h\to 0+} \frac{\varphi(x_0+hv) - \varphi(x_0)}{h}$$

Proposition 1.7 [5] *Let $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$ be an extended function. Then*

$$\partial_-\varphi(x_0) \;=\; \{\; p \in \mathbf{R}^n \mid \quad \forall\, v \in \mathbf{R}^n,\;\; D_\uparrow\varphi(x_0)(v) \;\geq < p, v > \}$$

and

$$\partial_+\varphi(x_0) \;=\; \{\; p \in \mathbf{R}^n \mid \quad \forall\, v \;\in\; \mathbf{R}^n,\;\; D_\downarrow\varphi(x_0)(v) \;\leq < p, v > \}$$

It is not difficult to show that φ is Fréchet differentiable at x_0 if and only if both super and subdifferentials of φ at x_0 are nonempty. Moreover in this case

$$\partial_+\varphi(x_0) \;=\; \partial_-\varphi(x_0) \;=\; \{\; \nabla\varphi(x_0)\;\}$$

Definition 1.8 *Let $\varphi : \mathbf{R}^n \mapsto \mathbf{R}$ be Lipschitz at x_0. We denote by $\partial^\star \varphi(x_0)$ the set of all cluster points of gradients $\nabla\varphi(x_n)$, when x_n converge to x_0 and φ is differentiable at x_n, i.e.,*

$$\partial^\star \varphi(x_0) \ = \ \mathrm{Limsup}_{x \to x_0} \{ \nabla\varphi(x) \}$$

Proposition 1.9 (Clarke) *If $\partial^\star \varphi(x_0)$ is a singleton, then φ is differentiable at x_0.*

See [16, p.33] for the proof.

1.1.4 Semiconcave Functions

Definition 1.10 *Consider a convex subset K of $\mathbf{R}^n$. A function $\varphi : K \mapsto \mathbf{R}$ is called* semiconcave *if there exists $\omega : \mathbf{R}_+ \times \mathbf{R}_+ \mapsto \mathbf{R}_+$ such that*

$$\forall\, r \le R,\ \forall\, s \le S,\ \ \omega(r,s) \le \omega(R,S) \ \ \& \ \ \lim_{s \to 0+} \omega(R,s) \ = \ 0 \qquad (3)$$

and for every $R > 0,\ \lambda \in [0,1]$ and all $x, y \in K \cap RB$

$$\lambda\varphi(x) + (1-\lambda)\varphi(y) \le \varphi(\lambda x + (1-\lambda)y) + \lambda(1-\lambda)\|x-y\|\, \omega(R, \|x-y\|)$$

We say that φ is semiconcave at x_0 if there exists a neighborhood of x_0 in K such that the restriction of φ to it is semiconcave. We call the above function ω a modulus of semiconcavity of φ.

Observe that every concave function $\varphi : K \mapsto \mathbf{R}$ is semiconcave (with ω equal to zero.)

In general a Lipschitz function does not have directional derivatives. Our next result implies in particular that for a semi-concave function, the directional derivatives exist.

Theorem 1.11 *Let $K \subset \mathbf{R}^n$ be a convex set, $x_0 \in K$ and let a function $\varphi : K \mapsto \mathbf{R}$ be Lipschitz and semiconcave at x_0. Then for every $v \in T_K(x_0)$*

$$\liminf_{\substack{v' \to v,\, h \to 0+ \\ x' \to_K x_0,\ x' + hv' \in K}} \frac{\varphi(x' + hv') - \varphi(x')}{h}$$

$$= \lim_{\substack{v' \to v,\, h \to 0+ \\ x_0 + hv' \in K}} \frac{\varphi(x_0 + hv') - \varphi(x_0)}{h}$$

In particular, if $x_0 \in \mathrm{Int}(K)$, *then*

$$\partial_+\varphi(x_0) \;=\; co\,(\partial^\star\varphi(x_0)) \tag{4}$$

(Clarke's generalized gradient of φ *at* x_0*), where co states for the convex hull. Furthermore, setting* $\varphi = -\infty$ *outside of* K, *for all* $x_0 \in K$

$$\mathrm{Limsup}_{x\to_{\mathrm{Int}(K)}x_0}\,\partial_+\varphi(x) \;\subset\; \partial_+\varphi(x_0)$$

Proof — It is enough to consider the case $\|v\| < 1$. Fix such v and let $\delta > 0$ be so that φ is semiconcave on $K \cap B_{2\delta}(x_0)$ with semiconcavity modulus $\omega(\cdot) := \omega(2\delta, \cdot)$. Let $x \in K \cap B_\delta(x_0)$. Then for all $0 < h_1 \le h_2 \le \delta$ such that $x + h_2 v \in K$ we have

$$\varphi(x+h_1v) - \varphi(x) \;=\; \varphi\left(\tfrac{h_1}{h_2}(x+h_2v) \;+\; \left(1-\tfrac{h_1}{h_2}\right)x\right) \;-\; \varphi(x)$$

$$\ge\; \tfrac{h_1}{h_2}\varphi(x+h_2v) \;-\; \tfrac{h_1}{h_2}\varphi(x) - \;h_1\left(1-\tfrac{h_1}{h_2}\right)\|v\|\,\omega(h_2\,\|v\|)$$

Consequently,

$$\frac{\varphi(x+h_1v)-\varphi(x)}{h_1} \ge \frac{\varphi(x+h_2v)-\varphi(x)}{h_2} - \left(1-\frac{h_1}{h_2}\right)\omega(h_2\,\|v\|)$$

and we proved that for every $x \in K \cap B_\delta(x_0)$ and all $0 < h' \le h \le \delta$,

$$\frac{\varphi(x+h'v)-\varphi(x)}{h'} \;\ge\; \frac{\varphi(x+hv)-\varphi(x)}{h} - \omega(h\,\|v\|) \tag{5}$$

Thus for every $0 < h \le \delta$

$$\liminf_{\substack{h'\to 0+\\ v'\to v\\ x+h'v'\in K}} \frac{\varphi(x+h'v')-\varphi(x)}{h'} \ge \frac{\varphi(x+hv)-\varphi(x)}{h} - \omega(h\,\|v\|)$$

Taking lim sup in the right-hand side of the above inequality when $x = x_0$, we deduce that

$$\lim_{\substack{h\to 0+,\, v'\to v\\ x_0+hv'\in K}} \frac{\varphi(x_0+hv')-\varphi(x_0)}{h}$$

does exist. Fix $\varepsilon > 0$ and $0 < \lambda < \delta$. From the Lipschitz continuity of φ it follows that there exists $0 < \alpha < \delta$ such that for all $x \in K \cap B_\alpha(x_0)$ and $v' \in B_\alpha(v)$

$$\frac{\varphi(x_0 + \lambda v) - \varphi(x_0)}{\lambda} \leq \frac{\varphi(x + \lambda v') - \varphi(x)}{\lambda} + \varepsilon$$

where $x_0 + \lambda v \in K$, $x + \lambda v' \in K$. Thus, using (5), we obtain that for all sufficiently small $\alpha > 0$,

$$\frac{\varphi(x_0 + \lambda v) - \varphi(x_0)}{\lambda}$$

$$\leq \inf_{\substack{x \in K \cap B_\alpha(x_0) \\ h \in]0,\lambda], v' \in B_\alpha(v) \\ x + hv' \in K}} \frac{\varphi(x + hv') - \varphi(x)}{h} + \omega(\lambda\|v'\|) + \varepsilon$$

Letting ε, α and λ converge to zero we end the proof of the first statement. The second one results from the alternative definition of Clarke's generalized gradient, i.e. $p \in co\ (\partial^\star \varphi(x_0))$ if and only if for all v

$$\liminf_{\substack{v' \to v, h \to 0+ \\ x' \to x_0}} \frac{\varphi(x' + hv') - \varphi(x')}{h} \leq < p, v >$$

To prove the last statement we set $\varphi = -\infty$ outside of K. Consider a sequence $x_m \in \mathrm{Int}(K)$ converging to x_0 and a sequence $p_m \in \partial_+ \varphi(x_m)$ converging to some p. We have to show that $p \in \partial_+ \varphi(x_0)$.

From (4) and the Carathéodory theorem, we deduce that there exist $\lambda_i^m \geq 0$ and $x_i^m \in \mathrm{Int}(K)$ converging to x_0 when $m \to \infty$ such that φ is differentiable at x_i^m and for all i the sequence $\nabla\varphi(x_i^m)$ converges to some p_i when $m \to \infty$, and for every m, $\sum_{i=0}^n \lambda_i^m = 1$,

$$\lim_{m \to \infty} \left(\sum_{i=0}^n \lambda_i^m \nabla \varphi(x_i^m) \right) = p$$

Taking a subsequence and keeping the same notations, we may assume that $(\lambda_0^m, ..., \lambda_n^m)$ converge to some $(\lambda_0, ..., \lambda_n)$. Thus $p = \sum_{i=0}^n \lambda_i p_i$. Since $\partial_+ \varphi(x_0)$ is convex, the above yields that it is enough to prove our statement

only in the case when φ is differentiable at x_m. Fix $v \in T_K(x_0)$ and consider $h_m \to 0+$ such that $x_m + h_m v \in K$ and

$$\frac{\varphi(x_m + h_m v) - \varphi(x_m)}{h_m} \;\leq\; \langle \nabla\varphi(x_m),\; v\rangle \;+\; \frac{1}{m}$$

This and the first claim imply that

$$\limsup_{v' \to v,\; h \to 0+} \frac{\varphi(x_0 + hv') - \varphi(x_0)}{h} \;\leq\; \langle p, v\rangle$$

Hence from Proposition 1.7 we deduce that $p \in \partial_+\varphi(x_0)$. $\diamond$

Proposition 1.12 *Let $\varphi : \mathbf{R}^n \mapsto \mathbf{R}$ be Lipschitz and semiconcave at x_0. If $\partial_+\varphi(x_0)$ is a singleton, then φ is differentiable at x_0 and*

$$\partial^\star\varphi(x_0) \;=\; \{\; \nabla\varphi(x_0)\;\}$$

In particular, if $\partial_+\varphi(x)$ is a singleton for all x near x_0, then φ is continuously differentiable at x_0.

Proposition 1.13 *Let $\varphi : \mathbf{R}^n \mapsto \mathbf{R}$, $x_0 \in \mathbf{R}^n$. If φ is Lipschitz at x_0 and both φ and $-\varphi$ are semiconcave at x_0, then φ is continuously differentiable on a neighborhood of x_0.*

Proof — Since φ and $-\varphi$ are semiconcave at x_0, by Theorem 1.11, there exists a neighborhood $\mathcal{N}$ of x_0 such that for all $x \in \mathcal{N}$

$$\partial_+\varphi(x) = co(\partial^\star\varphi(x)), \quad \partial_-\varphi(x) = -\partial_+(-\varphi)(x) = -co(\partial^\star(-\varphi)(x))$$

Hence both $\partial_+\varphi(x)$ and $\partial_-\varphi(x)$ are nonempty. Therefore φ is differentiable on $\mathcal{N}$. The conclusion follows from Proposition 1.12. $\diamond$

We investigate next closedness of the level sets of regularized lower derivatives.

Proposition 1.14 *Let $K \subset \mathbf{R}^n$ and $\varphi : K \mapsto \mathbf{R}$ be locally Lipschitz. Define the set-valued map $Q : K \hookrightarrow \mathbf{R}^n$ by:*

for all $x \in K$, $Q(x)$ is equal to

$$\{v \mid \liminf_{\substack{v' \to v,\, h \to 0+ \\ x' \to_K x,\; x' + hv' \in K}} \frac{\varphi(x' + hv') - \varphi(x')}{h} \leq 0\}$$

Then Q has closed nonempty images and Graph(Q) *is closed.*

Proof — Clearly for every x, $0 \in Q(x)$. It remains to show that for every sequence $(x_n, v_n) \in K \times \mathbf{R}^n$ converging to some $(x, v) \in K \times \mathbf{R}^n$ and satisfying $v_n \in Q(x_n)$, we have $v \in Q(x)$. Fix such a sequence and let $\varepsilon_n \to 0+$. Then there exist $h_n \to 0+$, $x'_n \to_K x$, $v'_n \to v$ such that for every n, $x'_n + h_n v'_n \in K$ and

$$\frac{\varphi(x'_n + h_n v'_n) - \varphi(x'_n)}{h_n} \leq \varepsilon_n$$

Taking lim inf in the above inequality we end the proof. $\diamond$

1.1.5 Subnormal Cones to the Epigraph

Recall that

$$\mathcal{E}p(D_\uparrow \varphi(x_0)) = T_{\mathcal{E}p(\varphi)}(x_0, \varphi(x_0)) \tag{6}$$

where $\mathcal{E}p$ denotes the epigraph.

The subnormal cone to $\mathcal{E}p(\varphi)$ at $(x_0, \varphi(x_0))$ is given by

$$N^0_{\mathcal{E}p(\varphi)}(x_0, \varphi(x_0)) := \Big\{p \in \mathbf{R}^n \mid \forall\, v \in T_{\mathcal{E}p(\varphi)}(x_0, \varphi(x_0)),\ \langle p, v\rangle \leq 0\Big\}$$

Thus

Proposition 1.15 *Let $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$ and $x_0 \in \mathrm{Dom}(\varphi)$. Then the following statements are equivalent*

i) $p \in \partial_- \varphi(x_0)$

ii) $\forall\, u \in \mathbf{R}^n,\ < p, u > \leq D_\uparrow \varphi(x_0)(u)$

iii) $(p, -1) \in N^0_{\mathcal{E}p(\varphi)}(x_0, \varphi(x_0))$

We shall also need the following technical result.

Lemma 1.16 ([48]) *Consider an extended lower semicontinuous function $\varphi : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and $x_0 \in \mathrm{Dom}(\varphi)$. Let $p \in \mathbf{R}^n$ be such that*

$$(p, 0) \in N^0_{\mathcal{E}p(\varphi)}(x_0, \varphi(x_0)), \quad p \neq 0$$

Then for every $\varepsilon > 0$, there exist x_ε, p_ε in $\mathbf{R}^n$ and $q_\varepsilon < 0$ satisfying

$$\|x_\varepsilon - x_0\| \leq \varepsilon,\ \ \|p_\varepsilon - p\| \leq \varepsilon \quad \& \quad (p_\varepsilon, q_\varepsilon) \in N^0_{\mathcal{E}p(\varphi)}(x_\varepsilon, \varphi(x_\varepsilon))$$

1.2 Regularity of Set-Valued Maps

We recall next some definitions concerning set-valued maps. Let X, Y denote metric spaces and $F : X \hookrightarrow Y$ be a set-valued map. For every $x \in X$ the subset $F(x)$ is called the *image* of F at x. The *domain* of F is the subset

$$\mathrm{Dom}(F) := \{x \in X \mid F(x) \neq \emptyset\}$$

and its *graph*

$$\mathrm{Graph}(F) := \{(x, y) \in X \times Y \mid y \in F(x)\}$$

Definition 1.17 *The map F is called* upper semicontinuous *at x if and only if for any neighborhood $\mathcal{U}$ of $F(x)$,*

$$\exists\, \eta > 0 \quad \textit{such that } \forall\, x' \in B_\eta(x), \;\; F(x') \subset \mathcal{U}$$

It is said to be upper semicontinuous *on a subset $K \subset X$ if and only if it is upper semicontinuous at any point $x \in K$.*

The map F is called lower semicontinuous *at x if and only if for any open subset $\mathcal{U} \subset Y$ such that $\mathcal{U} \cap F(x) \neq \emptyset$,*

$$\exists\, \eta > 0 \text{ such that } \forall\, x' \in B_\eta(x), \;\; F(x') \cap \mathcal{U} \neq \emptyset$$

It is said to be lower semicontinuous *on a subset $K \subset X$ if for every $x \in K$ and for any open subset $\mathcal{U} \subset Y$ with $\mathcal{U} \cap F(x) \neq \emptyset$,*

$$\exists\, \eta > 0 \text{ such that } \forall\, x' \in B_\eta(x) \cap K, \;\; F(x') \cap \mathcal{U} \neq \emptyset$$

We shall say that F is continuous *at x if it is* both *upper and lower semicontinuous at x, and that it is* continuous *on a subset $K \subset X$ if and only if it is upper and lower semicontinuous on K.*

Notice that if F is upper semicontinuous on X, then its domain is closed.

When $F(x)$ is compact, F is upper semicontinuous at x if and only if

$$\forall\, \varepsilon > 0,\ \exists\, \eta > 0 \text{ such that } \quad \forall\, x' \in B_\eta(x), \;\; F(x') \subset \bigcup_{y \in F(x)} B_\varepsilon(y)$$

Proposition 1.18 [5] *The graph of an upper semicontinuous set-valued map $F : X \hookrightarrow Y$ with closed images is closed. The converse is true if we assume that Y* is compact.

Definition 1.19 *When (X, d_X) is a metric space and Y is a normed space, we shall say that $F : X \rightsquigarrow Y$ is* Lipschitz *(L−Lipschitz) on a subset $K \subset \mathrm{Dom}(F)$ if there exists $L \geq 0$ such that*

$$\forall\, x_1,\ x_2 \in K,\quad F(x_1)\ \subset\ F(x_2) + Ld_X(x_1, x_2)B$$

The set-valued map F is called locally Lipschitz around $x \in X$ if there exists a neighborhood $\mathcal{N}$ of x such that F is Lipschitz on $\mathcal{N}$.

We recall next definitions of derivatives of set-valued maps.

Definition 1.20 *Let X, Y be normed spaces, $F : X \rightsquigarrow Y$ be a set-valued map and $y \in F(x)$.*

The adjacent derivative *$dF(x, y)$ is the set-valued map from X to Y defined by*

$$\forall\, u\ \in\ X,\ \ v\ \in\ dF(x,y)(u)\ \iff\ \forall\, h_n \to 0+\quad \exists\, u_n \to u$$

$$\text{such that}\ \lim_{n\to\infty} \mathrm{dist}\left(v,\ \frac{F(x+h_nu_n)-y}{h_n}\right)\ =\ 0$$

If F is Lipschitz around x, then an equivalent definition is given by

$$\forall\, u\ \in\ X,\ \ dF(x,y)(u)\ =\ \mathrm{Liminf}_{h\to 0+}\ \tfrac{F(x+hu)-y}{h}\ =$$

$$\lim_{h\to 0+} \mathrm{dist}\left(v,\ \frac{F(x+hu)-y}{h}\right)\ =\ 0$$

We shall need the following proposition.

Proposition 1.21 [5] *Let us assume that the images of F are convex and that F is Lipschitz around x. Then for any $(x, y) \in \mathrm{Graph}(F)$ the images of the adjacent derivative $dF(x, y)$ are convex and*

$$dF(x,y)(0)\ =\ T_{F(x)}(y)$$

$$\forall u \in \mathrm{Dom}(dF(x,y)), D^\flat F(x,y)(u) + dF(x,y)(0) = dF(x,y)(u)$$

Proof — Let v_1 and v_2 belong to $dF(x, y)(u)$. Then, for any sequence $h_n > 0$ converging to 0, there exist sequences u_{1n} and u_{2n} converging to u and sequences v_{1n} and v_{2n} converging to v_1 and v_2 respectively such that

$$\forall\, n,\ \ y + h_nv_{in}\ \in\ F(x + h_nu_{in})\ \ (i = 1, 2)$$

Since F is Lipschitz around x, there exists $l > 0$ such that for all n large enough,

$$y + h_n v_{2n} \ \in \ F(x + h_n u_{1n}) + l h_n \|u_{2n} - u_{1n}\|$$

so that we can find another sequence v_{3n} converging to v_2 such that

$$y + h_n v_{3n} \ \subset \ F(x + h_n u_{1n})$$

Now, $F(x + h_n u_{1n})$ being convex, we deduce that for all $\lambda \in [0, 1]$,

$$y + h_n(\lambda v_{1n} + (1 - \lambda) v_{3n}) \ \in \ F(x + h_n u_{1n})$$

Since $\lambda v_{1n} + (1 - \lambda) v_{3n}$ converges to $\lambda v_1 + (1 - \lambda) v_2$, this element belongs to $dF(x, y)(u)$.

Notice that $v \in dF(x, y)(0)$ if and only if $d(v, (F(x) - y)/h)$ converges to 0. Since $F(x)$ is convex, it coincides with the tangent cone.

Since $0 \in dF(x, y)(0)$ we obtain that

$$\forall\ u,\ dF(x, y)(u) \ \subset \ dF(x, y)(u) \ + \ dF(x, y)(0)$$

To prove the opposite inclusion fix

$$v \ \in \ dF(x, y)(u) \quad \& \quad w \ \in \ dF(x, y)(0)$$

Let $h_n \to 0+,\ v_n \to v$ be such that

$$\forall\ n,\ \ y + h_n v_n \ \in \ F(x + h_n u)$$

By convexity of $F(x)$, there exist $w_n \to w$ such that for n large enough, $y + \sqrt{h_n} w_n \in F(x)$. Then, by the Lipschitz continuity of F, for all large n and for some w'_n, we have

$$y \ + \sqrt{h_n} w'_n \ \in \ F(x + h_n u)\ ;\ \|w'_n - w_n\| \ \leq \ l\sqrt{h_n}\, \|u\|$$

Thus

$$\left\{ \begin{array}{l} (1 - \sqrt{h_n})(y + h_n v_n) \ + \sqrt{h_n}(y + \sqrt{h_n} w'_n) \\ \\ = \ y + h_n(v_n + w'_n) \ - \sqrt{h_n}\, h_n v_n \ = \ y + h_n(v + w) + h_n \varepsilon(h_n) \\ \\ \in \ F(x + h_n u) \end{array} \right.$$

where $\varepsilon(h_n)$ converges to 0. Hence

$$\lim_{n\to\infty} \operatorname{dist}\left(v+w, \frac{F(x+h_n u)-y}{h_n}\right) \;=\; 0$$

This ends the proof. $\diamond$

Let X be a complete separable metric space, $t_0 < T$ be real numbers and $U : [t_0, T] \rightsquigarrow X$ be a set-valued map with closed, possibly empty images. It is called (Lebesgue) *measurable* if for every open subset $\mathcal{O} \subset X$, the set

$$\{\, t \in [t_0, T] \mid\; U(t) \cap \mathcal{O} \;\neq\; \emptyset \,\} \text{ is Lebesgue measurable}$$

or, equivalently, if for every closed subset $\mathcal{C} \subset X$, the set

$$\{\, t \in [t_0, T] \mid\; U(t) \cap \mathcal{C} \;\neq\; \emptyset \,\} \text{ is Lebesgue measurable}$$

A measurable single-valued map $u : [t_0, T] \mapsto X$ satisfying

$$\forall\, t \;\in\; [t_0, T], \quad u(t) \;\in\; U(t)$$

is called a *measurable selection* of $U(\cdot)$.

Measurable selections are dense:

Theorem 1.22 [5] *Let X be a complete separable metric space and $U : [t_0, T] \rightsquigarrow X$ be a set-valued map with closed nonempty images. Then the following two statements are equivalent:*

i) — U is measurable

ii) — There exist measurable selections $u_n(\cdot)$ of $U(\cdot)$, $n = 1, \ldots$ such that for every $t \in [t_0, T]$, $U(t) = \overline{\bigcup_{n\geq 1} u_n(t)}$.

Proposition 1.23 [5] *Let X be a complete separable metric space and $U_n : [t_0, T] \rightsquigarrow X$, $n = 1, \ldots$ be measurable set-valued maps with closed images. Then the set-valued maps*

$$t \;\rightsquigarrow\; \bigcap_{n\geq 1} U_n(t), \quad t \;\rightsquigarrow\; \overline{\bigcup_{n\geq 1} U_n(t)}$$

and

$$t \;\rightsquigarrow\; \operatorname{Liminf}_{n\to\infty} U_n(t), \quad t \rightsquigarrow \operatorname{Limsup}_{n\to\infty} U_n(t)$$

are measurable.

Corollary 1.24 *Let X, Y be complete separable metric spaces, $x : [t_0, T] \mapsto X$ be a measurable single-valued map and $F : [t_0, T] \times X \rightsquigarrow Y$ be a set-valued map with nonempty closed images satisfying the following assumptions:*

$$\begin{cases} i) & \forall\, x \in X \ \text{ the set-valued map } \ F(\cdot, x) \ \text{ is measurable} \\ \\ ii) & \text{For almost every } \ t \in [t_0, T], \ F(t, \cdot) \ \text{ is continuous at } \ x(t) \end{cases}$$

Then the map $t \rightsquigarrow F(t, x(t))$ is measurable.

Proof — Since $x(\cdot)$ is measurable, there exist measurable maps $x_n : [t_0, T] \mapsto X$ assuming only finite number of values such that for almost every $t \in [t_0, T]$, $\lim_{n\to\infty} x_n(t) = x(t)$. From the assumption $i)$ we deduce that the map $t \rightsquigarrow F(t, x_n(t))$ is measurable and from the assumption $ii)$, that for almost all $t \in [t_0, T]$

$$F(t, x(t)) \ = \ \text{Liminf}_{n\to\infty} \ F(t, x_n(t))$$

Proposition 1.23 completes the proof. $\diamond$

Let us denote by $B(x, \rho)$ the closed ball in X of center x and radius ρ. When $K \subset X$ and $y \in X$ we denote by $\Pi_K(y)$ the projection of y on K given by

$$\Pi_K(y) \ := \ \{ \ x \ \in \ K \mid \ d_X(x, y) \ = \ \text{dist}\,(y, K) \ \}$$

Of course it may happen that the set $\Pi_K(y)$ is empty. Denote by $\overline{co}$ the closed convex hull.

Proposition 1.25 [5] *Let X be a separable Banach space, $U : [t_0, T] \rightsquigarrow X$ be a measurable set-valued map with closed nonempty images and $g : [t_0, T] \mapsto X$, $k : [t_0, T] \mapsto \mathbf{R}_+$ be measurable single-valued maps. Then the maps*

$$t \ \rightsquigarrow \ \overline{co}\, U(t), \ \ t \ \rightsquigarrow \ B(g(t), k(t)), \ \ t \ \rightsquigarrow \ \Pi_{U(t)}(g(t))$$

and $t \mapsto \text{dist}(g(t), U(t))$ are measurable. Consequently, if

$$\{v \in U(t) \mid \|v - g(t)\| \ \leq \ k(t)\} \ \neq \ \emptyset \ \text{ almost everywhere in } \ [t_0, T]$$

then there exists a measurable selection $u(t) \in U(t)$ such that for almost all $t \in [t_0, T]$, $\|u(t) - g(t)\| \leq k(t)$.

Consider a metric space Y. We recall that a map $\varphi : [t_0, T] \times X \mapsto Y$ is called *Carathéodory*, if for every $x \in X$, $\varphi(\cdot, x)$ is measurable and for almost all $t \in [t_0, T]$, the map $\varphi(t, \cdot)$ is continuous.

Proposition 1.26 [5] *Consider complete separable metric spaces X, Y, a Carathéodory map $\varphi : [t_0, T] \times X \mapsto Y$ and a measurable set-valued map $U : [t_0, T] \rightsquigarrow X$ with closed nonempty images. Then for every measurable map $h : [t_0, T] \mapsto Y$ satisfying*

$$h(t) \in \varphi(t, U(t)) \text{ almost everywhere in } [t_0, T]$$

there exists a measurable selection $u(t) \in U(t)$ such that $h(t) = \varphi(t, u(t))$ for almost all $t \in [t_0, T]$.

Definition 1.27 *Consider metric spaces X, Y and a set-valued map $G : [t_0, T] \times X \rightsquigarrow Y$ with closed images. It is called a* Carathéodory set-valued map *if for every $x \in X$, the map $t \rightsquigarrow G(t, x)$ is measurable and for every $t \in [t_0, T]$, the map $x \rightsquigarrow G(t, x)$ is continuous.*

Theorem 1.28 (Direct Image [5]) *Let X be a complete separable metric space and $U : [t_0, T] \rightsquigarrow X$ a measurable set-valued map with closed images.*

Consider a Carathéodory set-valued map G from $[t_0, T] \times X$ to a complete separable metric space Y. Then, the map

$$[t_0, T] \ni t \rightsquigarrow \overline{G(t, U(t))}$$

is measurable.

Denote by $L^1(t_0, T; \mathbf{R}^n)$ the Banach space of (Lebesgue) integrable maps $u : [t_0, T] \mapsto \mathbf{R}^n$ with the norm

$$\|u\|_{L^1} = \int_{t_0}^{T} \|u(t)\| \, dt$$

Definition 1.29 *Consider a set-valued map $U : [t_0, T] \rightsquigarrow \mathbf{R}^n$ and denote by $\mathcal{U}$ the set of integrable selections of U, i.e.,*

$$\mathcal{U} := \{ u \in L^1(t_0, T; \mathbf{R}^n) \mid u(t) \in U(t) \text{ almost everywhere in } [t_0, T] \}$$

The integral of U on $[t_0, T]$ is defined by

$$\int_{t_0}^{T} U dt := \left\{ \int_{t_0}^{T} u(t) dt \mid u \in \mathcal{U} \right\}$$

We say that a set-valued map $U : [t_0, T] \hookrightarrow \mathbf{R}^n$ is *integrably bounded* if there exists an integrable function $\psi : [t_0, T] \mapsto \mathbf{R}_+$ such that $U(t) \subset \psi(t)B$ almost everywhere in $[t_0, T]$.

Let K be a nonempty subset of a vector space Y. A point $x \in K$ is called *extremal* if for all $y, z \in K$ and $0 < \lambda < 1$ satisfying $x = \lambda y + (1 - \lambda)z$, we have $x = y = z$.

Theorem 1.30 (Aumann) *Let $U : [t_0, T] \hookrightarrow \mathbf{R}^n$ be a measurable set-valued map with nonempty closed images. Then the integral $\int_{t_0}^T U dt$ is convex and extremal points of $\overline{co}\left(\int_{t_0}^T U dt\right)$ are contained in $\int_{t_0}^T U dt$. If in addition U is integrably bounded, then the integral of U is also compact and $\int_{t_0}^T U ds = \int_{t_0}^T \overline{co} U ds$.*

See for instance [5] for the proof.

Theorem 1.31 *Let $U : [t_0, T] \hookrightarrow \mathbf{R}^n$ be a measurable set-valued map with closed images having at least one integrable selection.*

Then for every $\varepsilon > 0$ and integrable selection $\overline{u}(t) \in \overline{co}\, U(t)$ there exists an integrable selection $u(t) \in U(t)$ such that

$$\sup_{t \in [t_0, T]} \left\| \int_{t_0}^t u(s) ds - \int_{t_0}^t \overline{u}(s) ds \right\| \leq \varepsilon$$

In particular this yields that

$$\overline{\int_{t_0}^T \overline{co}\, U dt} = \overline{\int_{t_0}^T U dt}$$

Proof — Fix $\varepsilon > 0$, an integrable selection $\overline{u}(t) \in \overline{co}\, U(t)$ and let $u_0(\cdot)$ be an integrable selection of $U(\cdot)$. Define measurable set-valued maps $U_n : [t_0, T] \hookrightarrow X$ with closed nonempty images by

$$\forall\, t \in [t_0, T], \quad U_n(t) = u_0(t) \cup (U(t) \cap nB)$$

and set $\varepsilon_n(t) := \operatorname{dist}(\overline{u}(t), \overline{co}(U_n(t)))$. By Proposition 1.25, $\varepsilon_n(\cdot)$ is measurable for each n. Furthermore, the sequence $\{\varepsilon_n(\cdot)\}_{n \geq 1}$ is integrably bounded and $\lim_{n \to \infty} \varepsilon_n(t) = 0$ for $t \in [t_0, T]$. Using again Proposition 1.25, we deduce that for every $n \geq 1$ there exists a measurable selection $u_n(t) \in \overline{co}\, U_n(t)$ such that

$$\|u_n(t) - \overline{u}(t)\| \leq \varepsilon_n(t) \text{ almost everywhere in } [t_0, T]$$

Therefore, by the Lebesgue dominated convergence theorem, the sequence u_n converges to $\overline{u}$ in $L^1(t_0,T;\mathbf{R}^n)$ and for all n large enough

$$\sup_{t\in[t_0,T]}\left\|\int_{t_0}^t u_n(s)ds-\int_{t_0}^t \overline{u}(s)ds\right\| \;\le\; \int_{t_0}^T \|u_n(s)-\overline{u}(s)\|\,ds \;\le\; \frac{\varepsilon}{2}$$

It remains to show that for every $n\ge 1$ there exists an integrable selection $u(t)\in U_n(t)\subset U(t)$ such that

$$\sup_{t\in[t_0,T]}\left\|\int_{t_0}^t u(s)ds-\int_{t_0}^t u_n(s)ds\right\| \;\le\; \frac{\varepsilon}{2}$$

Fix $n\ge 1$ and let $\psi:[t_0,T]\mapsto\mathbf{R}_+$ be an integrable function such that $U_n(t)\subset\psi(t)B$ for $t\in[t_0,T]$. Let $i\ge 1$ be so large, that for any measurable subset $I\subset[t_0,T]$ of the Lebesgue measure less than $(T-t_0)/i$ we have $\int_I\psi(s)ds\le\varepsilon/4$. We denote by I_j the interval

$$I_j \;=\; \left[\,t_0+\frac{j-1}{i}(T-t_0),\; t_0+\frac{j}{i}(T-t_0)\,\right],\quad j=1,...,i$$

By Theorem 1.30,

$$\forall\, j=1,...,i,\qquad \int_{I_j}\overline{co}\;U_n(s)ds \;=\; \int_{I_j}U_n(s)ds$$

This yields that for every $1\le j\le i$ there exists a measurable selection $f_j(t)\in U_n(t)$ such that

$$\int_{I_j}f_j(s)ds \;=\; \int_{I_j}u_n(s)ds$$

Let u be a selection of U_n equal to f_j on the interior of I_j for every $j=1,...,i$. Then for every $t\in[t_0,T]$, there exists j such that $t\in I_j$ and

$$\left\|\textstyle\int_{t_0}^t(u-u_n)(s)ds\right\|\le\left\|\sum_{r=1}^{j-1}\int_{I_r}(u-u_n)(s)ds\right\|+\int_{I_j}\|u-u_n\|(s)ds$$

$$\le\;\int_{I_j}\left(\|u(s)\|+\|u_n(s)\|\right)ds\;\le\;2\int_{I_j}\psi(s)ds\;\le\;\varepsilon/2\quad\diamond$$

1.3 Differential Inclusions

Consider $t_0<T$ and denote by $\mathcal{C}(t_0,T;\mathbf{R}^n)$ the Banach space of continuous maps from $[t_0,T]$ into $\mathbf{R}^n$ with the norm

$$\|x\|_{\mathcal{C}} \;=\; \sup_{t\in[t_0,T]}\|x(t)\|$$

We first define what we call a solution to differential inclusions.

In the case of differential equations, there is no ambiguity since the derivative $x'(\cdot)$ of a solution $x(\cdot)$ to a differential equation $x'(t) = f(t, x(t))$ inherits the properties of the map f and of the function $x(\cdot)$. It is continuous whenever f is continuous and measurable whenever f is continuous with respect to x and measurable with respect to t.

This is no longer the case with differential inclusions. The extension of Peano's Theorem to differential inclusions is due to Marchaud and Zaremba who proved independently in the thirties the existence of respectively *contingent* and *paratingent* solutions to differential inclusions (called *champs de demi-cônes* at the time). The generalization of the concept of derivative to the notion of contingent derivative is due to B. Bouligand, who wrote: "... Nous ferons tout d'abord observer ... que la notion de contingent éclaire celle de différentielle". Then Ważewski proposed at the beginning of the sixties to look for solutions among *absolutely continuous* functions. He wrote: "... I learned the results of Zaremba's dissertation before the second world war, since I was a referee of that paper. Then a few years ago I came across with some results on optimal control and I have noticed a close connection between the optimal control problem and the theory of Marchaud-Zaremba." (The author learned that this "coming across" happened during a seminar talk of C. Olech on a paper by LaSalle at Ważewski's seminar.)

Ważewski proved that one can replace the contingent or paratingent derivatives of functions by derivatives of absolutely continuous functions defined almost everywhere in the definition of a solution to a differential inclusion, that he called *orientor field*.

We recall that a function $x \in \mathcal{C}(t_0, T; \mathbf{R}^n)$ is called *absolutely continuous* if for almost all $t \in [t_0, T]$ the derivative $x'(t)$ exists, $x' \in L^1(t_0, T; \mathbf{R}^n)$ and

$$\forall\, t \;\in\; [t_0, T], \quad x(t) \;=\; x(t_0) \;+\; \int_{t_0}^{t} x'(s)ds$$

Let $W^{1,1}(t_0, T; \mathbf{R}^n)$ denote the Banach space of absolutely continuous functions from $[t_0, T]$ to $\mathbf{R}^n$ with the norm

$$\|x\|_{W^{1,1}} \;=\; \|x(t_0)\| + \int_{t_0}^{T} \|x'(t)\|dt$$

Consider a set-valued map F from $[t_0, T] \times \mathbf{R}^n$ into subsets of $\mathbf{R}^n$. We associate with it the differential inclusion

$$x' \;\in\; F(t, x) \tag{7}$$

An absolutely continuous function $x : [t_0, T] \mapsto \mathbf{R}^n$ is called a *solution* to (7) if

$$x'(t) \in F(t, x(t)) \text{ almost everywhere in } [t_0, T] \tag{8}$$

1.3.1 Filippov's Theorem

We investigate here some properties of solutions to differential inclusion (7) in the case when F is Lipschitz with respect to x.

We denote by $\mathcal{S}_{[t_0,T]}(x_0)$ the set of solutions to (7) starting at $x_0 \in \mathbf{R}^n$ and defined on the time interval $[t_0, T]$:

$$\mathcal{S}_{[t_0,T]}(x_0) = \{x \mid x \text{ is a solution to (7) on } [t_0, T],\ x(t_0) = x_0\}$$

and set $L^1(t_0, T) = L^1(t_0, T; \mathbf{R}_+)$ (the set of nonnegative integrable functions.)

Let $y \in W^{1,1}(t_0, T; \mathbf{R}^n)$ be an absolutely continuous function. Filippov's theorem provides an estimate of the distance from y to the set $\mathcal{S}_{[t_0,T]}(x_0) \subset W^{1,1}(t_0, T; \mathbf{R}^n)$ under the following assumptions on F:

$$\begin{cases} i) & \forall\, (t, x) \in [t_0, T] \times \mathbf{R}^n, \quad F(t, x) \text{ is closed} \\ ii) & \forall\, x \in \mathbf{R}^n \text{ the set-valued map } F(\cdot, x) \text{ is measurable} \\ iii) & \exists\, \beta > 0,\ k \in L^1(t_0, T) \text{ such that for almost all} \\ & t \in [t_0, T],\ F(t, x) \text{ is nonempty for } x \in y(t) + \beta B \\ & \text{the map } F(t, \cdot) \text{ is } k(t) - \text{Lipschitz on } y(t) + \beta B \end{cases} \tag{9}$$

Theorem 1.32 *Consider a set-valued map $F : [t_0, T] \times \mathbf{R}^n \hookrightarrow \mathbf{R}^n$ and an absolutely continuous function $y \in W^{1,1}(t_0, T; \mathbf{R}^n)$. Assume that (9) holds true and that the function*

$$t \mapsto \gamma(t) := \operatorname{dist}(y'(t), F(t, y(t)))$$

is integrable. Let $\delta \geq 0$ and set

$$\eta(t) = e^{\int_{t_0}^t k(\tau)d\tau}\delta + \int_{t_0}^t \gamma(s)\, e^{\int_s^t k(\tau)d\tau} ds$$

If $\eta(T) \leq \beta$, then for all $x_0 \in \mathbf{R}^n$ with $\|x_0 - y(t_0)\| \leq \delta$, there exists $x \in \mathcal{S}_{[t_0,T]}(x_0)$ such that

$$\forall\, t \in [t_0,\ T], \quad \|x(t) - y(t)\| \leq \eta(t)$$

and

$$\|x'(t) - y'(t)\| \leq k(t)\eta(t) + \gamma(t) \text{ a.e. in } [t_0, T]$$

Remark — From Corollary 1.24 and Proposition 1.25 follows that under assumptions (9) the function $t \mapsto \text{dist}(y'(t), F(t, y(t)))$ is always measurable. ◇

The proof can be found in [19], [1]. The above result can be extended to the whole half line:

Theorem 1.33 *Consider a set-valued map $F : \mathbf{R}_+ \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and an absolutely continuous function $y \in W^{1,1}(0, \infty; \mathbf{R}^n)$. Assume that (9) holds true with the time interval $[t_0, T]$ replaced by $\mathbf{R}_+$ and that the function $t \mapsto \gamma(t) := \text{dist}(y'(t), F(t, y(t)))$ is integrable on $[0, \infty[$. Let $\delta \geq 0$ and set*

$$\eta(t) = e^{\int_0^t k(\tau)d\tau}\delta + \int_0^t \gamma(s)\, e^{\int_s^t k(\tau)d\tau} ds$$

If $\limsup_{t\to\infty} \eta(t) \leq \beta$, then for all $x_0 \in \mathbf{R}^n$ with $\|x_0 - y(0)\| \leq \delta$, there exists $x \in \mathcal{S}_{[0,\infty[}(x_0)$ such that

$$\forall\, t \geq 0, \quad \|x(t) - y(t)\| \leq \eta(t)$$

and

$$\|x'(t) - y'(t)\| \leq k(t)\eta(t) + \gamma(t) \text{ a.e. in } [0, \infty[$$

Proof — Theorem 1.32 yields an estimate on the finite interval $[0, 1]$. Hence there exists a solution $x(\cdot) \in \mathcal{S}_{[0,1]}(x_0)$ satisfying the required estimates on the interval $[0, 1]$ and in particular

$$\|x(1) - y(1)\| \leq e^{\int_0^1 k(\tau)d\tau}\delta + \int_0^1 \gamma(s) e^{\int_s^1 k(\tau)d\tau} ds$$

This and Theorem 1.32 imply that there exists a solution $z(\cdot) \in \mathcal{S}_{[1,2]}(x(1))$ satisfying the required estimates on $[1, 2]$. Hence we can extend $x(\cdot)$ on the interval $[0, 2]$ by concatenating it with $z(\cdot)$ and we reiterate this process. ◇

The above theorems yield the following corollaries.

Corollary 1.34 *Consider a set-valued map $F : [t_0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and a point $x_0 \in \mathbf{R}^n$. We assume that F satisfies (9) with $y \equiv x_0$ and is lower semicontinuous at (t_0, x_0). Then for every $u \in F(t_0, x_0)$ there exist $t_1 > t_0$ and a solution $x(\cdot) \in \mathcal{S}_{[t_0,t_1]}(x_0)$ with $x'(t_0) = u$.*

Proof — Fix $u \in F(t_0, x_0)$. It is enough to consider the absolutely continuous function

$$\forall\, t \in [t_0, T], \quad y(t) = x_0 + (t - t_0)u$$

Then for every $t \in [t_0, T]$ such that $(t - t_0)\,\|u\| \leq \beta$ we have

$$\text{dist}\,(u, F(t, y(t))) \leq \text{dist}\,(u, F(t, x_0)) + k(t)\,\|y(t) - x_0\|$$

$$= \text{dist}(u, F(t, x_0)) + k(t)(t - t_0)\,\|u\|$$

By Theorem 1.32 there exist $t_1 > t_0$ and a solution $x \in \mathcal{S}_{[t_0,t_1]}(x_0)$ such that

$$\|x(t) - y(t)\| \leq \int_{t_0}^{t} (\text{dist}(u, F(s, x_0)) + k(s)(s - t_0)\,\|u\|)\, e^{\int_s^t k(\tau)d\tau} ds$$

$$\leq e^{\int_{t_0}^t k(s)ds} \left(\int_{t_0}^t \text{dist}(u, F(s, x_0))ds + (t - t_0)\,\|u\| \int_{t_0}^t k(s)ds \right)$$

for all $t \in [t_0, t_1]$. Thus

$$\forall\, t \in [t_0, t_1], \quad \|x(t) - x_0 - (t - t_0)u\| = o(t - t_0)$$

and the result follows. $\diamond$

Corollary 1.35 *Let $y_0 \in \mathbf{R}^n$, $y \in \mathcal{S}_{[t_0,T]}(y_0)$ and assume that F, y satisfy (9). Then there exists $\delta > 0$ depending only on $k(\cdot)$ such that for all $x_0 \in B(y_0, \delta)$ we have*

$$\inf_{x \in \mathcal{S}_{[t_0,T]}(x_0)} \|x - y\|_{\mathcal{C}} \leq e^{\int_{t_0}^T k(s)ds}\, \|x_0 - y_0\|$$

1.3.2 Relaxation Theorems

Let $x_0 \in \mathbf{R}^n$. In this section we compare solutions to the differential inclusion

$$\begin{cases} x'(t) \in F(t, x(t)) \text{ almost everywhere in } [t_0, T] \\ x(t_0) = x_0 \end{cases} \tag{10}$$

and of the convexified (relaxed) differential inclusion:

$$\begin{cases} x'(t) \in \overline{co}\, F(t, x(t)) \text{ almost everywhere in } [t_0, T] \\ x(t_0) = x_0 \end{cases} \tag{11}$$

Observe that if F satisfies (9), then so does the set-valued map $(t, x) \hookrightarrow \overline{co}(F(t, x))$.

Theorem 1.36 *Let $y : [t_0, T] \mapsto \mathbf{R}^n$ be a solution to the relaxed inclusion (11). Assume that F and y satisfy (9) and that the set-valued map $[t_0, T] \ni t \rightsquigarrow F(t, y(t))$ has at least one integrable selection (or, equivalently, that the map $t \mapsto dist(0, F(t, y(t)))$ is integrable.)*

Then for every $\varepsilon > 0$ there exists a solution x to (10) such that $\|x - y\|_C \leq \varepsilon$.

Proof — By Corollary 1.24 and assumptions (9) the set-valued map $t \rightsquigarrow F(t, y(t))$ is measurable and has closed images.

Fix $\varepsilon > 0$ so small that $\varepsilon < \beta - \varepsilon$. By Theorem 1.31 there exists an integrable selection $u(s) \in F(s, y(s))$ such that

$$\sup_{t \in [t_0, T]} \left\| \int_{t_0}^{t} (u - y')(s) ds \right\| \;\leq\; \varepsilon \, e^{-\int_{t_0}^{T} k(s) ds} \left(1 + \int_{t_0}^{T} k(s) ds \right)^{-1}$$

Define the absolutely continuous function $\overline{y} : [t_0, T] \mapsto \mathbf{R}^n$ by

$$\forall\, t \;\in\; [t_0, T], \;\; \overline{y}(t) \;=\; x_0 + \int_{t_0}^{t} u(s) ds$$

Then $\overline{y}(t_0) = x_0$ and

$$\forall\, t \;\in\; [t_0, T], \quad \|\overline{y}(t) - y(t)\| \;\leq\; \varepsilon$$

Thus $F(t, \cdot)$ is $k(t)-$Lipschitz on the ball $B(\overline{y}(t), \beta - \varepsilon)$. Furthermore, for almost all $t \in [t_0, T]$,

$$\mathrm{dist}(\overline{y}'(t), F(t, \overline{y}(t))) \;\leq\; k(t) \, \|\overline{y}(t) - y(t)\| \;=\; k(t) \left\| \int_{t_0}^{t} (u - y')(s) ds \right\|$$

Set

$$\eta(t) \;:=\; \int_{t_0}^{t} \mathrm{dist}(\overline{y}'(s), F(s, \overline{y}(s))) \, e^{\int_s^t k(\tau) d\tau} ds$$

Then, by the choice of u and ε, $\eta(T) \leq \varepsilon \leq \beta - \varepsilon$. Theorem 1.32 ends the proof. $\diamond$

Theorem 1.37 (Relaxation) *Let $F : [t_0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a set-valued map with closed nonempty images and $x_0 \in \mathbf{R}^n$. Assume that there exists $k \in L^1(t_0, T)$ such that for almost every $t \in [t_0, T]$, $F(t, \cdot)$ is $k(t)$-Lipschitz and that the map $t \rightsquigarrow F(t, 0)$ has at least one integrable selection.*

Then solutions to differential inclusion (10) are dense in solutions to the relaxed inclusion (11) in the metric of uniform convergence.

Proof — It is enough to observe that for every $y \in \mathcal{C}(t_0, T; \mathbf{R}^n)$ we have $F(t,0) \subset F(t, y(t)) + k(t)\,\|y(t)\|\,B$. Since $t \rightsquigarrow F(t,0)$ has an integrable selection, from Proposition 1.25 we infer that so does the set-valued map $t \rightsquigarrow F(t, y(t))$. Theorem 1.36 ends the proof. $\diamond$

Theorem 1.38 *Let $x_0 \in \mathbf{R}^n$ and $\mathcal{S}^{co}_{[t_0,T]}(x_0)$ denote the set of solutions to the relaxed inclusion (11). Under all assumptions of Theorem 1.37 suppose that the set-valued map $t \rightsquigarrow F(t,0)$ is integrably bounded.*

Then the closure of $\mathcal{S}_{[t_0,T]}(x_0)$ in the metric of uniform convergence is compact and is equal to $\mathcal{S}^{co}_{[t_0,T]}(x_0)$.

Proof — We first show that $\mathcal{S}_{[t_0,T]}(x_0)$ is relatively compact in $\mathcal{C}(t_0, T; \mathbf{R}^n)$ (i.e., its closure is compact.) Indeed consider a sequence $x_n(\cdot) \in \mathcal{S}_{[t_0,T]}(x_0)$ and let $\psi(\cdot) \in L^1(t_0, T)$ be such that $F(t,0) \subset \psi(t)B$ almost everywhere in $[t_0, T]$. Then for almost all $t \in [t_0, T]$ and for all $n \geq 1$ we have

$$\|x_n'(t)\| \;\leq\; \sup_{e \in F(t,0)} \|e\| \;+\; k(t)\,\|x_n(t)\| \;\leq\; \psi(t) + k(t)\,\|x_n(t)\|$$

Thus

$$\forall\, t \in [t_0, T], \;\; \|x_n(t)\| \;\leq\; \|x_0\| \;+\; \int_{t_0}^{t} \psi(s)ds \;+\; \int_{t_0}^{t} k(s)\,\|x_n(s)\|\,ds$$

This and Gronwall's lemma imply that there exists $M > 0$ such that

$$\forall\, t \;\in\; [t_0, T], \;\; \forall\, n \;\geq\; 1, \;\; \|x_n(t)\| \;\leq\; M$$

Thus the sequence $x_n'(\cdot)$ is integrably bounded and thereby the sequence $x_n(\cdot)$ is equicontinuous. By the Dunford-Pettis criterion a subsequence $\{x_{n_k}'\}$ converges weakly in $L^1(t_0, T; \mathbf{R}^n)$ to an integrable map $g : [t_0, T] \mapsto \mathbf{R}^n$.

Using Ascoli's theorem, taking a subsequence and keeping the same notations, we may also assume that $x_{n_k}(\cdot)$ converge uniformly to a continuous map $x : [t_0, T] \mapsto \mathbf{R}^n$. Since for every $n \geq 1$, $x_n(t) = x_0 + \int_{t_0}^{t} x_n'(s)ds$, taking the limit we obtain that

$$\forall\, t \;\in\; [t_0, T], \;\; x(t) \;=\; x_0 + \int_{t_0}^{t} g(s)ds$$

Thus $x(\cdot)$ is absolutely continuous and $x' = g$. Since

$$x_n'(t) \in \overline{co} F(t, x_n(t)) \subset \overline{co} F(t, x(t)) + k(t)\|x(t) - x_n(t)\|B$$

Mazur's theorem yields that $x(\cdot)$ is a solution to the differential inclusion (11). Theorem 1.37 ends the proof. $\diamond$

1.3.3 Infinitesimal Generator of Reachable Map

Consider $T > 0$, a set-valued map $F : [0,T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and let $x_0 \in \mathbf{R}^n$, $\rho > 0$ be given. In this subsection we assume that

$$\begin{cases} i) & \forall\ (t,x) \in [0,T] \times \mathbf{R}^n,\ \ F(t,x) \text{ is closed} \\ ii) & \forall\ x \in \mathbf{R}^n,\ \ F(\cdot,x) \text{ is measurable} \\ iii) & \forall\ (t,x) \in [0,T] \times B_\rho(x_0),\ \ F(t,x) \neq \emptyset \\ iv) & \exists\ L > 0 \text{ such that for every } t \in [0,T], \\ & \forall\ x,y \in B_\rho(x_0),\ \ F(t,x) \subset F(t,y) + L\,\|x-y\|\,B \end{cases} \tag{12}$$

For all $0 \leq t_0 \leq t_1 \leq T$ and $\xi \in \mathbf{R}^n$ set

$$R(t_1,t_0)\xi \ := \ \{\ x(t_1) \mid \ \ x \in \mathcal{S}_{[t_0,t_1]}(\xi)\ \}$$

This is the so-called *reachable set* of the inclusion

$$x' \ \in \ F(t,x) \tag{13}$$

from (t_0,ξ) at time t_1.

We first observe that the set-valued map R enjoys the following semigroup properties:

$$\begin{cases} \forall\ 0 \leq t_1 \leq t_2 \leq t_3 \leq T,\ \forall\ \xi \in \mathbf{R}^n,\ \ R(t_3,t_2)R(t_2,t_1)\xi = R(t_3,t_1)\xi \\ \forall\ 0 \leq t \leq T,\ \ \ \forall\ \xi \in \mathbf{R}^n,\ \ R(t,t)\xi \ = \ \xi \end{cases}$$

When F is sufficiently regular, the set-valued map $\overline{co}F(\cdot,\cdot)$ is the infinitesimal generator of the semigroup $R(\cdot,\cdot)$ in the sense that the difference quotients $(R(t+h,t)\xi - \xi)/h$ converge to $\overline{co}F(t,\xi)$:

Theorem 1.39 *Assume that (12) holds true and let* $t_0 \in [0,T[$.

If F is lower semicontinuous at (t_0,x_0), then

$$\overline{co}\ F(t_0,x_0) \ \subset \ \mathrm{Liminf}_{h\to 0+} \frac{R(t_0+h,t_0)x_0 - x_0}{h}$$

If F is upper semicontinuous at (t_0,x_0) and $F(t_0,x_0)$ is bounded, then

$$\mathrm{Limsup}_{h\to 0+} \frac{R(t_0+h,t_0)x_0 - x_0}{h} \ \subset \ \overline{co}\ F(t_0,x_0)$$

Consequently, if F is continuous at (t_0, x_0) and $F(t_0, x_0)$ is bounded, then

$$\mathrm{Lim}_{h\to 0+}\frac{R(t_0+h,t_0)x_0 - x_0}{h} \;=\; \overline{co}\; F(t_0,x_0)$$

Proof — The set-valued map $(t,x) \rightsquigarrow \overline{co}\; F(t,x)$ is lower semicontinuous at (t_0,x_0) if so is F. Fix $u \in \overline{co}F(t_0,x_0)$. By Corollary 1.34, there exist $t_1 > t_0$ and a solution $x(\cdot)$ to the relaxed inclusion (11) with T replaced by t_1 such that $x'(t_0) = u$. Using Theorem 1.36, we deduce that for every sufficiently small $h > 0$, there exists $x_h(\cdot) \in \mathcal{S}_{[t_0,t_0+h]}(x_0)$ such that $\|x_h(t_0+h) - x(t_0+h)\| \le h^2$. Hence

$$u \;\in\; \mathrm{Liminf}_{h\to 0+}\frac{R(t_0+h,t_0)x_0 - x_0}{h}$$

Since u is an arbitrary point in $\overline{co}F(t_0,x_0)$, the first statement follows.

To prove the second one we first observe that our assumptions imply that for some $\varepsilon > 0$, $M > 0$ and all $t \in [t_0, t_0+\varepsilon]$, $x \in B_\varepsilon(x_0)$ we have $F(t,x) \subset MB$. This yields that for some $t_1 > t_0$ and all $x \in \mathcal{S}_{[t_0,t_1]}(x_0)$

$$\forall\, t \;\in\; [t_0,t_1],\quad \|x(t) - x_0\| \;\le\; M(t-t_0)$$

Fix $v \in \mathrm{Limsup}_{h\to 0+}[R(t_0+h,t_0)x_0 - x_0]/h$ and consider a sequence $h_n > 0$ converging to zero and $x_n(\cdot) \in \mathcal{S}_{[t_0,t_0+h_n]}(x_0)$ such that

$$v \;=\; \lim_{n\to\infty}\frac{x_n(t_0+h_n) - x_n(t_0)}{h_n}$$

Since F is upper semicontinuous at (t_0,x_0), there exist $\varepsilon_n \to 0+$ such that

$$\forall\, t \;\in\; [t_0, t_0+h_n],\quad F(t,x_0) \;\subset\; F(t_0,x_0) \;+\; \varepsilon_n B$$

Since for all large n

$$x_n(t_0+h_n) - x_n(t_0) \;\in\; \textstyle\int_{t_0}^{t_0+h_n} F(t,x_n(t))dt$$

$$\subset\; \textstyle\int_{t_0}^{t_0+h_n} F(t,x_0)dt + \left(\int_{t_0}^{t_0+h_n} L\,\|x_n(t) - x_0\|\,dt\right)B$$

$$\subset\; \textstyle\int_{t_0}^{t_0+h_n} F(t_0,x_0)dt \;+\; \left(\int_{t_0}^{t_0+h_n}(\varepsilon_n + LM(t-t_0))dt\right)B$$

$$\subset h_n\overline{co}(F(t_0,x_0)) + (\varepsilon_n h_n + LMh_n^2)\,B$$

dividing by h_n and taking the limit we get $v \in \overline{co}(F(t_0,x_0))$.

1.3.4 Variational Inclusions

This subsection is devoted to differentiability of solutions to differential inclusion (7) with respect to the initial condition.

We denote by $d_x F(t, \overline{x}, \overline{y})$ the adjacent derivative of $F(t, \cdot, \cdot)$ (with respect to x) of the set-valued map $F(t, \cdot)$ at $(\overline{x}, \overline{y}) \in \mathrm{Graph}(F(t, \cdot))$.

Theorem 1.40 (Adjacent variational inclusion) [5] *Consider the solution map $\mathcal{S}_{[t_0,T]}(\cdot)$ as the set-valued map from $\mathbf{R}^n$ to $W^{1,1}(t_0, T; \mathbf{R}^n)$ and a solution $y(\cdot)$ to differential inclusion (10). Assume that (9) holds true, $u \in \mathbf{R}^n$ and let $w \in W^{1,1}(t_0, T; \mathbf{R}^n)$ be a solution to the linearized inclusion.*

$$\left\{ \begin{array}{lcl} w'(t) & \in & d_x F(t, y(t), y'(t))(w(t)) \ \text{ a.e. in } [t_0, T] \\ w(t_0) & = & u \end{array} \right. \tag{14}$$

Then for all $u_h \in \mathbf{R}^n$ converging to u when $h \to 0+$ and for all small $h > 0$, there exists $x_h \in \mathcal{S}_{[t_0,T]}(x_0 + h u_h)$ such that the difference quotients $(x_h - x)/h$ converge to w in $W^{1,1}(t_0, T; \mathbf{R}^n)$ when $h \to 0+$.

In particular, $w \in d\,\mathcal{S}(x_0, y(\cdot))(u)$.

The above result was proved in [5] in the case when $u_h = u$. Corollary 1.35 allows to extend it to an arbitrary sequence u_h.

Theorem 1.41 (Convex adjacent variational inclusion) *We consider the solution map $\mathcal{S}_{[t_0,T]}(\cdot)$ as the set-valued map from $\mathbf{R}^n$ to $\mathcal{C}(t_0, T; \mathbf{R}^n)$. Let y be a solution to the differential inclusion (10).*

Assume that (9) holds true, $u \in \mathbf{R}^n$ and let w be a solution to the inclusion

$$\left\{ \begin{array}{lcl} w'(t) & \in & d_x(\overline{co}\, F)(t, y(t), y'(t))(w(t)) \ \text{ a.e. in } [t_0, T] \\ w(t_0) & = & u \end{array} \right.$$

Then for all $u_h \in \mathbf{R}^n$ converging to u when $h \to 0+$ and for all small $h > 0$, there exists $x_h \in \mathcal{S}_{[t_0,T]}(x_0 + h u_h)$ such that the difference quotients $(x_h - x)/h$ converge to w in $\mathcal{C}(t_0, T; \mathbf{R}^n)$ when $h \to 0+$.

Proof — It is enough to apply Theorems 1.36 and 1.40. ◇

1.3.5 Viability Theorem

We recall here some definitions and the statement of Viability Theorem.

Let $F : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a set-valued map and $K \subset \mathrm{Dom}(F)$ be a nonempty subset.

The subset K enjoys the *viability property* for the differential inclusion

$$x' \in F(x) \tag{15}$$

if for any initial state $x_0 \in K$, there exists at least one solution $x(\cdot)$ to (15) starting at x_0 which is viable in K in the sense that $x(t) \in K$ for all $t \geq 0$. The viability property is said to be *local* if for any initial state $x_0 \in K$, there exist $T(x_0) > 0$ and a solution starting at x_0 which is viable in K on the interval $[0, T(x_0)]$ in the sense that for every $t \in [0, T(x_0)]$, $x(t) \in K$.

We say that K is a *viability domain* of F if

$$\forall\, x \in K, \quad R(x) := F(x) \cap T_K(x) \neq \emptyset$$

Theorem 1.42 (Viability Theorem) *If F is upper semicontinuous with nonempty compact convex images, then a locally compact set K enjoys the local viability property if and only if it is a viability domain of F. In this case, if for some $c > 0$, we have*

$$\forall\, x \in K, \quad \|R(x)\| := \inf_{u \in R(x)} \|u\| \leq c(\|x\| + 1)$$

and if K is closed, then K enjoys the viability property.

We refer to [4, Aubin] for the proof and many applications of viability theory.

The following result provides a very useful *duality* characterization of viability domains:

Proposition 1.43 (Ushakov, [38]) *Assume that the set-valued map $F : K \rightsquigarrow \mathbf{R}^n$ is upper semicontinuous with convex compact values. Then the following three statements are equivalent:*

$$\begin{aligned}
&i) \quad \forall\, x \in K, \ F(x) \cap T_K(x) \neq \emptyset \\
&ii) \quad \forall\, x \in K, \ F(x) \cap \overline{co}\,(T_K(x)) \neq \emptyset \\
&iii) \quad \forall\, x \in K, \ \forall\, p \in N_K^0(x), \ \sigma(F(x), -p) \geq 0
\end{aligned} \tag{16}$$

where $\sigma(F(x), \cdot)$ denotes the support function of $F(x)$.

(see for instance [5] for the proof).

1.4 Parametrization of Set-Valued Maps

We recall here few results concerning parametrization of set-valued maps. Their proofs can be found in [5, Chapter 9]. Theorems comparing solutions to differential inclusion and solutions to the corresponding parametrized system will be provided in the next Section.

Consider a metric space X, reals $t_0 < T$ and a set-valued map $F : [t_0, T] \times X \hookrightarrow \mathbf{R}^n$.

Definition 1.44 *Consider subsets $C(t) \subset X$, where $t \in [t_0, T]$. The set-valued map F is called* measurable/Lipschitz *on $\{C(t)\}_{t\in[t_0,T]}$ if for every $t \in [t_0, T]$, there exists $k(t) \geq 0$ such that*

$$\begin{cases} \forall\, x \in X,\ F(\cdot, x) \text{ is measurable} \\ \forall\, t \in [t_0, T],\ \forall\, x \in C(t),\ \ F(t, x) \neq \emptyset \text{ and is closed} \\ \forall\, t \in [t_0, T],\ \ F(t, \cdot) \text{ is } k(t) - \text{Lipschitz on } C(t) \end{cases}$$

Definition 1.45 *Let U be a metric space and $C(t) \subset X$, $t \in [t_0, T]$ be given nonempty subsets of X. We say that a single-valued map*

$$f : [t_0, T] \times X \times U \ \mapsto\ \mathbf{R}^n$$

is a measurable/Lipschitz parametrization *of F on $\{C(t)\}_{t\in[t_0,T]}$ with the constants $k(t)$, $t \in [t_0, T]$ if*

$$\begin{cases} i) & \forall\, (t, x) \in [t_0, T] \times X,\ \ F(t, x) = f(t, x, U) \\ ii) & \forall\, (x, u) \in X \times U,\ \ f(\cdot, x, u) \text{ is measurable} \\ iii) & \forall\, (t, u) \in [t_0, T] \times U,\ \ f(t, \cdot, u) \text{ is } k(t)\text{-Lipschitz on } C(t) \\ iv) & \forall\, (t, x) \in [t_0, T] \times X,\ \ f(t, x, \cdot) \text{ is continuous} \end{cases}$$

Theorem 1.46 (Parametrization of Unbounded Maps) *Consider a metric space X and a set-valued map $F : [t_0, T] \times X \hookrightarrow \mathbf{R}^n$ with closed convex images.*

Assume that F is measurable/Lipschitz on $\{C(t)\}_{t\in[t_0,T]}$ and let $k(t)$, $t \in [t_0, T]$ denote the corresponding Lipschitz constants.

Then there exists a measurable/Lipschitz parametrization f of F on $\{C(t)\}_{t\in[t_0,T]}$ with $U=\mathbf{R}^n$ such that:

$$\begin{cases} \forall\,(t,u)\ \in\ [t_0,T]\times\mathbf{R}^n,\ f(t,\cdot,u)\ \text{ is }\ ck(t)-\text{Lipschitz on }\ C(t) \\ \\ \forall\,(t,x)\ \in\ [t_0,T]\times X,\ f(t,x,\cdot)\ \text{ is }\ c-\text{Lipschitz on }\ \mathbf{R}^n \end{cases}$$

with c independent of F. Furthermore if F is continuous, so is f.

Theorem 1.47 (Parametrization of Bounded Maps) *Under the assumptions of Theorem 1.46 suppose that the images of F are compact.*

Then there exists a measurable/Lipschitz parametrization f of F on the family of sets $\{C(t)\}_{t\in[t_0,T]}$ with U equal to the closed unit ball B in $\mathbf{R}^n$ such that:

$$\begin{cases} i)\quad \forall\,(t,u)\in[t_0,T]\times B,\ f(t,\cdot,u)\ \text{ is }\ ck(t)-\text{Lipschitz on } C(t) \\ \\ ii)\quad \forall\,t\ \in\ [t_0,T],\ \ \forall\,x\ \in\ X,\ \ \forall\,u,\ v\ \in\ B \\ \\ \qquad \|f(t,x,u)-f(t,x,v)\|\ \le\ c\left(\max_{y\in F(t,x)}\|y\|\right)\|u-v\| \end{cases}$$

with c independent of F. Furthermore if F is continuous, so is f.

2 Control Systems and Differential Inclusions

In this Section we discuss several types of control systems and their relations to differential inclusions. Namely, we shall single out

- Explicit control systems
- State dependent control systems
- Implicit control systems

The explicit control system

$$x'\ =\ f(t,x,u(t)),\quad u(t)\ \in\ U(t)$$

is the most investigated in the literature. It is well adapted to the techniques of Ordinary Differential Equations and can be seen as a parametrized family

of ODE's. Indeed let us define the set of admissible controls $\mathcal{U}$ as the set of all measurable selections $u(t) \in U(t)$ and with every $u(\cdot) \in \mathcal{U}$, let us associate $\varphi_u(t,x) = f(t,x,u(t))$. Then the above control system may be replaced by ordinary differential equations

$$x' = \varphi_u(t,x), \quad u \in \mathcal{U}$$

So questions of existence, uniqueness and differentiability of solutions with respect to initial conditions may still be investigated using classical results. Another possible approach is to define the set-valued map F by $F(t,x) = f(t,x,U(t))$ and to consider the differential inclusion

$$x' \in F(t,x) \tag{17}$$

In Subsection 1 we show that under quite mild assumptions on the maps f and U, these two problems are equivalent. We apply this fact and variational inclusions from Section 1 to characterize variations of solutions. This will be used in Sections 3 and 5 to prove necessary conditions for optimality.

State dependent control systems

$$x' = f(t,x,u(t)), \quad u(t) \in U(t,x)$$

present additional difficulties: we can no longer choose controls independently of the state. A possible solution to this would be to pick first a selection $u(t,x) \in U(t,x)$ and then to consider the differential equation

$$x' = f(t,x,u(t,x))$$

However we have to use classical existence theorems to guarantee existence of a solution to such equation and, thereby, to assume at least continuity of u with respect to the state variable x. This would exclude a quite large number of solutions, because it is not possible to associate with every of them such regular selection u. This is why it is more natural in this case to use differential inclusion (17) with the set-valued map

$$F(t,x) = f(t,x,U(t,x)) = \{\, f(t,x,u) \mid u \in U(t,x) \,\}$$

In Subsection 2 we show that this new system has the same solution set and prove some results about variations of solutions.

Linear implicit system (*descriptor system*)

$$Ex' = Ax + Bu(t), \quad u(t) \in U$$

where E, A, B are possibly rectangular matrices, arises in models of electrical networks. When E, A are square, the above system is sometimes called *singular* because E may be noninvertible. Solutions to such system are usually understood in the distributional sense. Here we restrict our attention to *absolutely continuous solutions* only and prove in Subsection 3 that this implicit system may be reduced to the explicit one (in the sense that the sets of solutions are the same):

$$x' \ = \ Dx \ + \ v(t), \quad v(t) \ \in \ V, \ \ x \ \in \ Q$$

where $V \subset Q$ are subspaces obtained using E, A, B and D is a linear operator from Q into itself whose range is orthogonal to V.

Nonlinear implicit control systems

$$f(x, x', u(t)) \ = \ 0, \ \ u(t) \ \in \ U$$

appear often in different models. To investigate them, we shall use differential inclusion (17) with the set-valued map $F(t, x) = \{v \mid 0 \in f(x, v, U)\}$ for answering in Subsection 4 the same type of questions: comparison of solution sets and variations of solutions.

Although the nature of these systems appear to be different, the differential inclusion formulation allows to develop a unified approach to all of them. However one should always keep in mind that differential inclusions being rather an abstract representation, their investigation would remain unsatisfactory as long as the results are not translated in terms of the original systems. This is why we are also computing derivatives and variations of set-valued maps F defined in the above examples.

2.1 Nonlinear Control Systems

Consider a complete separable metric space $\mathcal{Z}$, real numbers $t_0 < T$ and a map (describing the dynamics)

$$f : [t_0, T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^n$$

Let $U : [t_0, T] \rightsquigarrow \mathcal{Z}$ be a set-valued map (of controls) with nonempty images. We associate with these data the control system

$$x' \ = \ f(t, x, u(t)), \ \ u(t) \ \in \ U(t), \ \ t \ \in \ [t_0, T] \tag{18}$$

An absolutely continuous function $x : [t_0, T] \mapsto \mathbf{R}^n$ is called a solution to (18) if there exists a measurable map $u : [t_0, T] \mapsto \mathcal{Z}$, called *admissible control*, such that

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U(t) \text{ almost everywhere in } [t_0, T]$$

2.1.1 Reduction to Differential Inclusion

Define the set-valued map from $[t_0, T] \times \mathbf{R}^n$ to $\mathbf{R}^n$ by

$$F(t, x) \;=\; f(t, x, U(t))$$

and consider the differential inclusion

$$x'(t) \;\in\; F(t, x(t)) \text{ almost everywhere in } [t_0, T] \tag{19}$$

Clearly every solution x to control system (18) satisfies (19). Hence x is also a solution to differential inclusion (19).

The natural question arises whether (19) has the same solutions than the control system (18)? The answer is positive for a quite large class of maps f.

We impose the following assumptions on f and U:

$$\begin{cases} \forall\, (x, u) \;\in\; \mathbf{R}^n \times \mathcal{Z}, \quad f(\cdot, x, u) \text{ is measurable} \\ \\ \forall\, t \;\in\; [t_0, T], \quad f(t, \cdot, \cdot) \text{ is continuous} \\ \\ U(\cdot) \text{ is measurable and has closed nonempty images} \end{cases} \tag{20}$$

Theorem 2.1 *Assume that (20) holds true. Then the set of solutions to control system (18) coincide with the set of solutions to differential inclusion (19).*

Proof — Fix a solution $x(\cdot)$ to differential inclusion (19). By Theorem 1.28 and our assumptions, the map $(t, u) \mapsto f(t, x(t), u)$ is Carathéodory. So the proof follows from Proposition 1.26. $\diamond$

The images of the set-valued map F defined above in general are not closed, while most Theorems of Section 1 deal only with closed valued maps. We provide next two results concerning "closure" of F.

Proposition 2.2 *Assume (20) and define the set-valued map clF by*

$$\forall\ (t,x)\ \in\ [t_0,T]\times\mathbf{R}^n,\quad clF(t,x)\ =\ \overline{f(t,x,U(t))}$$

Then $clF(\cdot,x)$ is measurable for every $x\in\mathbf{R}^n$. Furthermore if for some $x_0\in\mathbf{R}^n$, $\varepsilon>0$, $\overline{t}\in[t_0,T]$ and all $u\in U(\overline{t})$, $f(\overline{t},\cdot,u)$ is $k(\overline{t})$-Lipschitz on $B_\varepsilon(x_0)$, then so is $clF(\overline{t},\cdot)$. Finally, if $U(\cdot)$ has compact images, then so does F and, consequently, $clF=F$.

Proof — Measurability follows from Theorem 1.28. The proof of the last two statements is obvious. ◇

Theorem 2.3 *Assume (20) and let $\overline{x}(\cdot)$ be a solution to the differential inclusion*

$$x'(t)\ \in\ clF(t,x(t))\ \text{ almost everywhere in }\ [t_0,T] \tag{21}$$

Further assume that there exist $\rho>0$ and $k\in L^1(t_0,T)$ such that for almost every $t\in[t_0,T]$ and all $u\in U(t)$, the map $f(t,\cdot,u)$ is $k(t)$-Lipschitz on $B_\rho(\overline{x}(t))$.

Then for all $\varepsilon>0$ there exists a solution $x(\cdot)$ to (18) such that $x(t_0)=\overline{x}(t_0)$ and $\|x-\overline{x}\|_{W^{1,1}}\le\varepsilon$.

Proof — By Theorem 1.28 and (20) the map $(t,u)\mapsto f(t,\overline{x}(t),u)$ is Carathéodory. Fix $\varepsilon>0$, $N\ge1$. By Proposition 1.26 there exists a measurable selection $u(t)\in U(t)$ such that

$$\left\|\overline{x}'(t)-f(t,\overline{x}(t),u(t))\right\|\ \le\ \varepsilon/N$$

Consider the system

$$x'\ =\ f(t,x,u(t)),\quad x(t_0)\ =\ \overline{x}(t_0)$$

Choosing N large enough and using Filippov's Theorem 1.32 with $F(t,x)=f(t,x,u(t))$ and $y=\overline{x}$ we end the proof. ◇

Theorem 2.4 *Assume that (20) holds true and for some $\gamma\in L^1(t_0,T)$ and for almost all $t\in[t_0,T]$*

$$\forall\ x\ \in\ \mathbf{R}^n,\quad \sup_{u\in U(t)}\|f(t,x,u)\|\ \le\ \gamma(t)(1+\|x\|)$$

Further assume that for every $R > 0$ there exists $k_R \in L^1(t_0, T)$ such that for almost all $t \in [t_0, T]$ and for every $u \in U(t)$, $f(t, \cdot, u)$ is $k_R(t)$-Lipschitz on $B_R(0)$.

If the sets $f(t, x, U(t))$ are closed and convex, then the set of solutions to control system (18) starting at x_0 is compact in $\mathcal{C}(t_0, T; \mathbf{R}^n)$.

Proof — It is enough to apply Theorems 2.1 and 1.38. ◇

2.1.2 Linearization

Consider a solution z to control system (18) and let $\overline{u}$ be a corresponding control. We associate with it the following linearization of (18) along the solution-control pair $(z, \overline{u})$:

$$\begin{cases} w'(t) \ = \ \frac{\partial f}{\partial x}(t, z(t), \overline{u}(t))w(t) \ + \ v(t) \\ \\ v(t) \in V(t) := T_{\overline{co} f(t, z(t), U(t))}(f(t, z(t), \overline{u}(t))) \text{ a.e.} \end{cases} \tag{22}$$

where $T_{\overline{co} f(t,z(t),U(t))}(f(t, z(t), \overline{u}(t)))$ denotes the tangent cone to the convex set $\overline{co} f(t, z(t), U(t))$ at $f(t, z(t), \overline{u}(t))$.

We assume that

$$\begin{cases} \text{The derivative } \frac{\partial f}{\partial x}(t, z(t), \overline{u}(t)) \text{ exists a.e. in } [t_0, T] \\ \\ \text{For some } \varepsilon > 0, \ k \in L^1(t_0, T) \text{ and for a.e. } t \in [t_0, T] \\ \forall\, u \ \in \ U(t), \ \ f(t, \cdot, u) \ \text{ is } \ k(t) - \text{Lipschitz on } \ B_\varepsilon(z(t)) \end{cases} \tag{23}$$

Recall that the solution $w(\cdot)$ to (22) starting at w_0 and corresponding to an integrable selection $v(s) \in V(s)$ is given by

$$\forall\, t \ \in \ [t_0, T], \ \ w(t) \ = \ X(t)w_0 \ + \ \int_{t_0}^{t} X(t)X(s)^{-1}v(s)ds$$

where $X(\cdot)$ denotes the fundamental solution to the linear system

$$X'(t) \ = \ \frac{\partial f}{\partial x}(t, z(t), \overline{u}(t))X(t), \quad X(t_0) \ = \ Id \tag{24}$$

Theorem 2.5 *Assume that (20) and (23) hold true. Then for every solution $w(\cdot)$ to linearized system (22) and elements $\{w_h\}_{h>0}$ in $\mathbf{R}^n$ satisfying $\lim_{h \to 0+} w_h = w(t_0)$, there exist solutions $\{x_h\}_{h>0}$ to (18) such that*

$$x_h(t_0) \ = \ z(t_0) + hw_h \ \text{ for all } \ h > 0 \ \text{ small enough}$$

and the difference quotients $(x_h - z)/h$ converge uniformly to w when h goes to zero.

Proof — By Theorem 2.3 we may replace control system (18) by differential inclusion (21). Propositions 2.2, 1.21 allow to apply the variational inclusion (Theorem 1.41) and to deduce the result after observing that

$$\forall\, w,\quad \frac{\partial f}{\partial x}(t,z(t),\overline{u}(t))w \;\in\; d_xF(t,z(t),z'(t))(w) \quad \text{a.e. in } [t_0,T] \quad \diamond$$

2.2 State Dependent Control Systems

In the previous subsection we have considered the map of controls $U(\cdot)$ depending only on time. When it also depends on the states, then the control system is called a *state dependent* control system.

Let $\mathcal{Z}$ be a complete separable metric space and let

$$U : [t_0,T] \times \mathbf{R}^n \;\rightsquigarrow\; \mathcal{Z}$$

be a given set-valued map. Consider the control system

$$x' \;=\; f(t,x,u),\quad u \;\in\; U(t,x),\quad t \;\in\; [t_0,T] \tag{25}$$

An absolutely continuous function $x : [t_0,T] \mapsto \mathbf{R}^n$ is called a solution to (25) if for some measurable selection $u(t) \in U(t,x(t))$ we have

$$x'(t) \;=\; f(t,x(t),u(t)) \text{ almost everywhere in } [t_0,T]$$

2.2.1 Reduction to Differential Inclusion

We introduce the set-valued map $F : [t_0,T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ defined by

$$F(t,x) \;=\; f(t,x,U(t,x)) \;=\; \{f(t,x,v) \mid v \;\in\; U(t,x)\}$$

and replace (25) by the differential inclusion

$$x'(t) \;\in\; F(t,x(t)) \text{ almost everywhere in } [t_0,T] \tag{26}$$

We impose the following assumptions:

$$\begin{cases} \forall\,(x,u) \;\in\; \mathbf{R}^n \times \mathcal{Z},\quad f(\cdot,x,u) \text{ is measurable} \\ \forall\, t \;\in\; [t_0,T],\quad f(t,\cdot,\cdot) \text{ is continuous} \\ U \text{ is Carathéodory and has closed nonempty images} \end{cases} \tag{27}$$

Theorem 2.6 *If (27) holds true, then the sets of solutions to control system (25) and differential inclusion (26) do coincide.*

Proof — Clearly every solution to (25) solves also (26). Conversely, consider a solution x to differential inclusion (26). By Theorem 1.28 the set-valued map $t \rightsquigarrow U(t, x(t))$ is measurable and the map $(t, u) \mapsto f(t, x(t), u)$ is Carathéodory. Applying Proposition 1.26 we can find a measurable selection $u(t) \in U(t, x(t))$ such that $x'(t) = f(t, x(t), u(t))$ almost everywhere in $[t_0, T]$. $\diamond$

Hence we can rewrite dynamical system (25) in the differential inclusion formulation (26). In general F does not have closed values. However, using arguments comparable to those from the proof of Theorem 2.3 we get

Proposition 2.7 *Assume that (27) holds true. Then the set-valued map* $clF : [t_0, T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ *defined by*

$$clF(t, x) \;=\; \overline{f(t, x, U(t, x))}$$

is measurable with respect to t. *Furthermore if for some* $t \in [t_0, T]$, $k(t) \geq 0$, $l(t) \geq 0$, $x_0 \in \mathbf{R}^n$, $\rho > 0$, *the map* $f(t, \cdot, \cdot)$ *is* $k(t)-$*Lipschitz on* $B_\rho(x_0) \times \mathcal{Z}$ *and the set-valued map* $U(t, \cdot)$ *is* $l(t)-$*Lipschitz on* $B_\rho(x_0)$, *then* $clF(t, \cdot)$ *is* $k(t)(1 + l(t))-$*Lipschitz on* $B_\rho(x_0)$.

Let $\overline{x}(\cdot)$ *be a solution to the differential inclusion*

$$x'(t) \;\in\; clF(t, x(t)) \;\text{ almost everywhere in }\; [t_0, T] \tag{28}$$

Further assume that there exist $\rho > 0$ *and* $k \in L^1(t_0, T)$ *such that for almost every* $t \in [t_0, T]$ *and all* $u \in U(t, x)$, *the map* $f(t, \cdot, u)$ *is* $k(t)$*-Lipschitz on* $B_\rho(\overline{x}(t))$.

Then for all $\varepsilon > 0$ *there exists a solution* $x(\cdot)$ *to (25) such that* $x(t_0) = \overline{x}(t_0)$ *and* $\|x - \overline{x}\|_{W^{1,1}} \leq \varepsilon$.

2.2.2 Linearization

In this subsection we assume that $\mathcal{Z}$ is a separable Banach space. Consider a solution z to (25) and let $\overline{u}(t) \in U(t, z(t))$ be a corresponding control. We associate to it the following linearization of (25) along the pair $(z, \overline{u})$:

$$\begin{cases} w'(t) \;\in\; A(t)w(t) + B(t)d_x U(t, z(t), \overline{u}(t))(w(t)) + v(t) \\ \\ v(t) \;\in\; T_{\overline{co} f(t, z(t), U(t, z(t)))}(f(t, z(t), \overline{u}(t))) \;\text{ a.e. in }\; [t_0, T] \end{cases} \tag{29}$$

where

$$A(t) \;=\; \frac{\partial f}{\partial x}(t, z(t), \overline{u}(t)), \quad B(t) \;=\; \frac{\partial f}{\partial u}(t, z(t), \overline{u}(t))$$

and $d_x U$ denotes the (partial) adjacent derivative of U with respect to the state variable x.

We impose the following assumptions

$$\begin{cases} \text{The derivative } \frac{\partial f}{\partial (x,u)}(t, z(t), \overline{u}(t)) \text{ exists a.e. in } [t_0, T] \\ \\ \exists\, \varepsilon > 0 \text{ and functions } k,\ l : [t_0, T] \mapsto \mathbf{R}_+ \text{ such that} \\ f(t, \cdot, \cdot) \text{ is } k(t) - \text{Lipschitz on } B_\varepsilon(z(t)) \times \mathcal{Z} \text{ and} \\ U(t, \cdot) \text{ is } l(t) - \text{Lipschitz on } B_\varepsilon(z(t)) \text{ for a.e. } t \in [t_0, T] \end{cases} \tag{30}$$

Theorem 2.8 *Assume that $\mathcal{Z}$ is a separable Banach space, that (27), (30) hold true and the map $t \mapsto k(t)(1 + l(t))$ is integrable.*

If at least one of the following two conditions holds true:

$$\begin{cases} i) \quad \forall\, (t, x) \;\in\; [t_0, T] \times \mathbf{R}^n, \quad f(t, x, U(t, x)) \text{ is closed} \\ ii) \quad \forall\, (t, x) \;\in\; [t_0, T] \times \mathbf{R}^n, \quad U(t, x) \text{ is convex} \end{cases}$$

then for every solution $w(\cdot)$ to (29) and elements $\{w_h\}_{h>0}$ in $\mathbf{R}^n$ satisfying $\lim_{h \to 0+} w_h = w(t_0)$, there exists a family $\{x_h(\cdot)\}_{h>0}$ of solutions to (25) such that

$$x_h(t_0) \;=\; z(t_0) + h w_h \;\text{ for all small }\; h > 0$$

and the difference quotients $(x_h - z)/h$ converge uniformly to w when h goes to zero.

Proof — We apply the variational inclusion (Theorem 1.41) to deduce the result from the following relation:

$$\forall\, v \in \mathbf{R}^n,\ A(t)v + B(t) d_x U(t, z(t), \overline{u}(t))v \subset d_x F(t, z(t), z'(t))v$$

for almost all $t \in [t_0, T]$. $\diamond$

2.3 Linear Implicit Control Systems

Let $E, A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$ be linear operators from $\mathbf{R}^n$ into $\mathbf{R}^m$, U be a finite dimensional vector space and $B \in \mathcal{L}(U, \mathbf{R}^m)$. Consider the implicit control system

$$Ex' \;=\; Ax \;+\; Bu, \quad u \;\in\; U \tag{31}$$

When $n = m$ this system is sometimes called *singular*, because E may be non invertible.

An absolutely continuous function $x : [t_0, T] \mapsto \mathbf{R}^n$ is called a solution to (31) corresponding to a measurable control $u : [t_0, T] \mapsto U$ if

$$Ex'(t) \;=\; Ax(t) \;+\; Bu(t) \;\text{ almost everywhere in }\; [t_0, T]$$

Our aim is to reduce (31) to the explicit system

$$x' \;=\; Dx \;+\; v, \;\; v \;\in\; V, \;\; x \;\in\; Q \tag{32}$$

where $V \subset Q$ are subspaces of $\mathbf{R}^n$ and D is a linear operator from Q into Q.

Let us denote by $\mathcal{B}$ the range of B and for every $y \in \mathbf{R}^m$ set

$$E^{-1}(y) \;=\; \{\, x \in \mathbf{R}^n \mid \; Ex \;=\; y \,\}$$

We introduce a *decreasing family* of subspaces:

$$K_0 \;=\; \mathbf{R}^n, \;\ldots\;, \; K_{k+1} \;=\; A^{-1}(EK_k + \mathcal{B}), \;\; k \;\geq\; 0$$

Since they are subspaces of $\mathbf{R}^n$, we obtain

$$Q \;:=\; \bigcap_{k \geq 1} K_k \;=\; K_j$$

for some $j \leq n-1$. Furthermore $A^{-1}(EQ + \mathcal{B}) = Q$ and therefore the set-valued map $\mathcal{F} : Q \hookrightarrow Q$ given by

$$\forall\, x \;\in\; Q, \quad \mathcal{F}(x) \;:=\; E^{-1}(Ax + \mathcal{B}) \cap Q$$

has nonempty images. It is also clear that for every $x \in Q$, $\mathcal{F}(x)$ is an affine subspace of Q.

Theorem 2.9 *Every solution $x(\cdot)$ to (31) defined on the time interval $[t_0, T]$ satisfies $x(t) \in Q$ for all $t \in [t_0, T]$.*

Proof — Fix a solution $x : [t_0, T] \mapsto \mathbf{R}^n = K_0$. Assume that we already know that for some $0 \leq k < n-1$, $x(t) \in K_k$ for all t. Then $x'(t) \in K_k$ almost everywhere and, consequently,

$$x(t) \in A^{-1}(Ex'(t) + \mathcal{B}) \subset A^{-1}(EK_k + \mathcal{B}) = K_{k+1} \;\text{ for a.e. }\; t \in [t_0, T]$$

Continuity of $x(\cdot)$ yields that $x(t) \in K_{k+1}$ for all $t \in [t_0, T]$ and the proof ends by the induction argument. $\diamond$

Let the map $\mathcal{D} : Q \mapsto Q$ and the subspace $V \subset Q$ be defined by

$$\forall\, x \in Q, \;\; \mathcal{D}x \in \mathcal{F}(x), \;\; \|\mathcal{D}x\| = \min_{y \in \mathcal{F}(x)} \|y\|, \;\; V = E^{-1}(\mathcal{B}) \cap Q = \mathcal{F}(0)$$

Proposition 2.10 *The map $\mathcal{D}$ defined above is a linear operator from Q into itself. Furthermore for every $x \in Q$, $\mathcal{F}(x) = \mathcal{D}x + V$ and $\mathcal{D}x$ is orthogonal to V.*

Proof — Since $\mathrm{Graph}(\mathcal{F})$ is a subspace,

$$\mathcal{D}x + V \subset \mathcal{F}(x) + V = \mathcal{F}(x) + \mathcal{F}(0) \subset \mathcal{F}(x)$$

To prove the equality, consider $y \in \mathcal{F}(x) \subset Q$. Then $Ey \in Ax + \mathcal{B}$, $E\mathcal{D}x \in Ax + \mathcal{B}$ and therefore $y - \mathcal{D}x \in Q$ and $E(y - \mathcal{D}x) \in \mathcal{B}$. Hence $y - \mathcal{D}x \in V$ and $\mathcal{F}(x) = \mathcal{D}x + V$.

It remains to show that $\mathcal{D}$ is linear. The element $\mathcal{D}x$ being the orthogonal projection of zero onto $\mathcal{D}x + V$, we deduce that $\mathcal{D}(Q) \subset V^{\perp}$ (orthogonal to V in Q). Fix $x,\ y \in Q$. Then

$$-E\mathcal{D}x \in -Ax + \mathcal{B},\ -E\mathcal{D}y \in -Ay + \mathcal{B},\ E\mathcal{D}(x+y) \in Ax + Ay + \mathcal{B}$$

Adding these inclusions, we get $E(\mathcal{D}(x+y) - \mathcal{D}x - \mathcal{D}y) \in \mathcal{B}$. Thus

$$\mathcal{D}(x+y) - \mathcal{D}x - \mathcal{D}y \in V \cap V^{\perp} \quad \Longrightarrow \quad \mathcal{D}(x+y) = \mathcal{D}x + \mathcal{D}y$$

Finally $\mathcal{D}$ is homogeneous, because

$$\mathcal{F}(\lambda x) = E^{-1}(\lambda Ax + \mathcal{B}) \cap Q = \lambda \mathcal{F}(x) \ \diamond$$

Theorem 2.11 *Solutions to (31) and (32) do coincide. Denote by B^{+} the orthogonal right inverse of B. That is B^{+} is the linear operator from $\mathcal{B}$ into U with $B^{+}x$ equal to the orthogonal projection of zero onto $B^{-1}(x)$. Then the map*

$$u(x) = \begin{cases} B^{+}(E\mathcal{D}x - Ax) & \text{if } x \in Q \\ \emptyset & \text{if not} \end{cases}$$

is a regulation law for (31): for every $x_0 \in Q$ there exists a $C^{\infty}-$solution to the singular system

$$Ex' = Ax + Bu(x),\ x(0) = x_0 \tag{33}$$

defined on $[0, \infty[$. It is unique if and only if $\ker(E) \cap Q = \{0\}$.

Proof — By Theorem 2.9 and Proposition 2.10 every solution to (31) solves (32). Conversely every solution $x : [t_0, T] \mapsto \mathbf{R}^n$ to (32) satisfies

$Ex'(t) \in Ax(t) + \mathcal{B}$ almost everywhere. Hence, by Proposition 1.26, $x(\cdot)$ solves (31).

Fix $x_0 \in Q$ and consider the solution $x(\cdot) \in C^\infty$ to the linear system

$$x' \;=\; \mathcal{D}x, \quad x(0) \;=\; x_0$$

It is defined on $[0, +\infty[$ and is also a solution to (31). To prove the latter statement of our theorem, observe that (33) may be written as:

$$Ex' \;=\; E\mathcal{D}x, \quad x(0) \;=\; x_0 \tag{34}$$

So the solution to (33) is unique if and only if the solution to (34) is unique. But this happens whenever zero is the only solution to the differential inclusion

$$x' \;\in\; \mathcal{D}x + \ker(E) \cap Q, \quad x(0) \;=\; 0$$

Consequently uniqueness is equivalent to $\ker(E) \cap Q = \{0\}$. $\diamond$

Observe that the above results allow to study implicit system (31) even in the case when the solution corresponding to a given control and a given initial state is not unique. We investigate next necessary and sufficient conditions for uniqueness.

We say that (31) *enjoys uniqueness* if to every measurable control $\overline{u} : [0, T] \mapsto U$, $T > 0$ and every initial state $x_0 \in \mathbf{R}^n$ corresponds at most one solution to control system (31) starting at x_0. Set $(A^{-1}E)^0(\mathbf{R}^n) = \mathbf{R}^n$ and define recursively

$$\forall\, k \;\geq\; 0,\; (A^{-1}E)^{k+1}(\mathbf{R}^n) \;=\; A^{-1}E\left((A^{-1}E)^k(\mathbf{R}^n)\right)$$

Theorem 2.12 *Consider the subspace $P = (A^{-1}E)^{n-1}(\mathbf{R}^n)$. The following statements are equivalent :*

i) System (31) enjoys uniqueness

ii) $\ker E \cap P = \{\, 0 \,\}$

Proof — Observe that *i)* is equivalent to: $\overline{x}(\cdot) \equiv 0$ is the only solution to the linear system $Ex' = Ax$ starting at zero. Hence, by Theorem 2.11 applied with $B = 0$, (31) enjoys uniqueness if and only if *ii)* holds true. $\diamond$

2.4 Nonlinear Implicit Control Systems

Let $\mathcal{Z}$ be a complete separable metric space, $f : \mathbf{R}^n \times \mathbf{R}^n \times \mathcal{Z} \;\mapsto\; \mathbf{R}^m$ be a continuous map and $U \subset \mathcal{Z}$ be a given closed set. Consider the implicit

control system

$$f(x, x', u(t)) \ = \ 0, \quad u(t) \in U, \ \ t \ \in \ [t_0, T] \tag{35}$$

An absolutely continuous function $x : [t_0, T] \mapsto \mathbf{R}^n$ is called a solution to (35) if there exists a measurable map $u : [t_0, T] \mapsto U$ such that $f(x(t), x'(t), u(t)) = 0$ almost everywhere in $[t_0, T]$.

2.4.1 Reduction to Differential Inclusion

Define the set-valued map $F : \mathbf{R}^n \hookrightarrow \mathbf{R}^n$ by

$$F(x) \ = \ \{v \ \in \ \mathbf{R}^n \mid \ \exists \, u \ \in \ U \ \text{ with } \ f(x, v, u) \ = \ 0\}$$

and consider the differential inclusion

$$x'(t) \ \in \ F(x(t)) \ \text{ almost everywhere in } \ [t_0, T] \tag{36}$$

Clearly every solution to (35) solves (36). The following lemma implies the converse statement.

Lemma 2.13 *If f is continuous, then the solution sets of (36) and (35) do coincide.*

Proof — Fix a solution x to (36). The existence of a measurable selection $u(t) \in U$ with $f(x(t), x'(t), u(t)) = 0$ almost everywhere in $[t_0, T]$ follows from Proposition 1.26. $\diamond$

The introduced set-valued map F has a closed graph whenever f is continuous and U is compact. If moreover for all $\overline{x} \in \mathbf{R}^n$

$$\exists \, \varepsilon > 0 \ \text{ such that } \ \liminf_{\|v\| \to \infty} \ \min_{x \in B_\varepsilon(\overline{x}), \ u \in U} \|f(x, v, u)\| \ > \ 0 \tag{37}$$

then

$$\exists \, R \ > \ 0 \ \text{ such that } \ \forall \, x \ \in \ B_\varepsilon(\overline{x}), \ \ F(x) \ \subset B_R(0) \tag{38}$$

This and Proposition 1.18 yield that if (37) holds true and f is continuous, then F is upper semicontinuous on its domain of definition and has compact images.

Another sufficient condition for the upper semicontinuity of F is given by

Proposition 2.14 *Assume that $\mathcal{Z}$ is a finite dimensional vector space, f is continuous and for every $\overline{x} \in \mathbf{R}^n$*

$$\exists\, \varepsilon > 0 \text{ such that } \liminf_{\substack{\|v\| \to \infty \\ \|u\| \to \infty}} \inf_{x \in B_\varepsilon(\overline{x}),\, u \in U} \|f(x,v,u)\| > 0 \tag{39}$$

Then (38) holds true, F is upper semicontinuous on its domain of definition and has compact images.

In general the images of F are not convex and for this reason inclusion (36) is not easy to investigate even when F is upper semicontinuous.

Our next aim is to provide a sufficient condition for Lipschitz continuity of F. We assume that

$$\begin{cases} f(\cdot,\cdot,\cdot) \text{ is continuous} \\ \forall\, u \in U,\ \ f(\cdot,\cdot,u) \text{ is differentiable} \\ \dfrac{\partial f}{\partial v}(\cdot,\cdot,\cdot) \text{ is continuous} \end{cases} \tag{40}$$

Theorem 2.15 *Assume (40) and that for an open subset $\mathcal{N} \subset \mathbf{R}^n$ the following holds true*

$$\forall\, (x,v,u) \in f^{-1}(0) \text{ with } x \in \mathcal{N},\ u \in U,\ \ \frac{\partial f}{\partial v}(x,v,u) \text{ is surjective}$$

Further assume that at least one of the following two conditions is satisfied:

i) U is compact and for every $\overline{x} \in \mathcal{N}$ (37) is valid

ii) $\mathcal{Z}$ is finite dimensional and every $\overline{x} \in \mathcal{N}$ satisfies (39),

Then the set $\mathrm{Dom}(F) \cap \mathcal{N}$ *is open and F is locally Lipschitz on it. Furthermore for all $(x,v,u) \in f^{-1}(0)$ with $(x,u) \in \mathcal{N} \times U$, we have*

$$\ker\left(\frac{\partial f}{\partial (x,v)}(x,v,u)\right) \subset \mathrm{Graph}(dF(x,v))$$

The proof results from the inverse mapping theorems [30].

2.4.2 Linearization of Implicit Systems

Consider a solution z to (35) and let $\overline{u}$ be a corresponding control. We associate with it the following linear time dependent implicit system

$$\begin{cases} A(t)w(t) \ + \ B(t)(w'(t) - v(t)) \ = \ 0 \\ \\ v(t) \ \in \ T_{\overline{co}F(z(t))}(z'(t)) \ \text{ a.e. in } \ [t_0, T] \end{cases} \tag{41}$$

where

$$A(t) \;=\; \frac{\partial f}{\partial x}(z(t), z'(t), \overline{u}(t)), \quad B(t) \;=\; \frac{\partial f}{\partial v}(z(t), z'(t), \overline{u}(t))$$

Theorem 2.16 *Assume that all hypothesis of Theorem 2.15 hold true with* $\mathcal{N} = z([0,T]) + \rho B$, *where* $\rho > 0$. *Let* w *be a solution to (41).*

Then for all elements $\{w_h\}_{h>0}$ *in* $\mathbf{R}^n$ *satisfying* $\lim_{h\to 0+} w_h = w(t_0)$, *there exist solutions* x_h *to implicit system (35) such that*

$$x_h(t_0) \;=\; z(t_0) + hw_h \;\text{ for all small }\; h > 0$$

and the difference quotients $(x_h - z)/h$ *converge uniformly to* w *when* h *goes to zero.*

Proof — To prove this result we apply Lemma 2.13, Theorems 2.15, 1.21 and variational inclusion (Theorem 1.41). $\diamond$

3 Value Function of Mayer's Problem

In this Section we address the Mayer problem arising in control theory. We start with the free end point case:

$$\text{minimize } g(x(T))$$

over all solutions to the control system

$$x' \;=\; f(t, x, u(t)), \quad u(t) \;\in\; U(t) \tag{42}$$

satisfying the initial condition

$$x(0) \;=\; \xi_0 \tag{43}$$

By a simple change of variables the classical Bolza problem

$$\text{minimize} \left\{ g(x(T)) \;+\; \int_0^T L(t, x(t), u(t))dt \right\}$$

over all state-control solutions (x, u) of (42), (43) may be reduced to the Mayer problem.

The basic objective of optimal control theory is to find necessary and sufficient conditions for optimality and to construct optimal solutions. Necessary conditions are available in the form of Pontriagin's maximum principle. It implements the Fermat rule in the case of optimal control problems.

When applied to an abstract minimization problem: $\min_{x\in K}\varphi(x)$, where K is a subset of a normed space, Fermat rule states that if $\overline{x}\in K$ is a minimizer, then

$$\forall\, v\ \in\ T_K(\overline{x}),\ \ \langle\nabla\varphi(\overline{x}),\ v\rangle\ \geq\ 0$$

In the above $T_K(\overline{x})$ denotes the contingent cone to K at $\overline{x}$.

In the same way as the Euler-Lagrange equation is a consequence of the Fermat rule in Calculus of Variations, the maximum principle can be deduced from the above rule: If the state-control pair $(z,\overline{u})$ is optimal, then the solution $p(\cdot)$ (called the co-state) to the adjoint system

$$p'(t)\ =\ -\left(\frac{\partial f}{\partial x}(t,z(t),\overline{u}(t))\right)^{\star}p(t),\ \ p(T)\ =\ -\nabla g(z(T))$$

satisfies the transversality condition

$$p(t)\ \in\ N^0_{R(t)}(z(t))\ \text{ almost everywhere in }\ [0,T]$$

where $N^0_{R(t)}(z(t))$ denotes the subnormal cone to the reachable set $R(t)$ of (42) from ξ_0 at time t. This last inclusion implies the maximum principle (of Pontrjagin):

$$\langle p(t),\ f(t,z(t),\overline{u}(t))\rangle\ =\ \max_{u\in U(t)}\langle p(t),\ f(t,z(t),u)\rangle\ \ \text{a.e. in }\ [0,T]$$

In Subsection 3 we complete these conditions to obtain sufficient ones by using the value function

$$V(t_0,x_0)\ =\ \inf\{\ g(x(T))\mid x\text{ is a solution to (42)},\ x(t_0)=x_0\ \}$$

and the Hamiltonian H of the control system (42):

$$H(t,x,p)\ =\ \sup_{u\in U(t)}\langle p,f(t,x,u)\rangle$$

In general V and H are non smooth functions and we have to use notions of superdifferentials from Section 1.

The value function allows to single out optimal solutions. Indeed, it is non decreasing along solutions to (42) and is constant along optimal solutions.

Sufficient conditions that we prove are of the following type: for almost every $t \in [0,T]$, there exists $p(t)$ such that

$$(\langle p(t), z'(t)\rangle\,,\ -p(t)) \in \partial_+ V(t,z(t))$$

where $\partial_+ V$ denotes the superdifferential of V. We also show that the co-state of the maximum principle verifies the above relations.

To find the value function from its definition at first glance seems to be an impossible task, because it amounts to solving as many optimization problems as there are initial points (t_0,x_0). But very fortunately, under quite general assumptions, the value function is the *unique solution* to the Hamilton-Jacobi equation:

$$-\frac{\partial V}{\partial t}(t,x) \ + \ H\left(t,x,-\frac{\partial V}{\partial x}(t,x)\right) \ = \ 0, \quad V(T,\cdot) \ = \ g(\cdot)$$

However, since even in very regular situations the value function is merely Lipschitz, solutions of the above Hamilton-Jacobi equation have to be understood in a generalized sense, where derivatives are replaced by subdifferentials (see Section 4.)

We also investigate what are the regularity properties of the system which are inherited by the value function (Lipschitz continuity in Subsection 1, semiconcavity in Subsection 4 and lower semicontinuity in Section 4.) In Subsection 3 we show that differentiability of V is related to uniqueness of optima and is preserved along each optimal solution.

When the Hamiltonian H is smooth enough and the value function is differentiable at $(0,\xi_0)$, then the following necessary and sufficient condition for optimality holds true:

Let $x(\cdot),\ p(\cdot)$ solve the Hamiltonian system

$$\begin{cases} x'(t) \ = \ \frac{\partial H}{\partial p}(t,x(t),p(t)) \\ \\ p'(t) \ = \ -\frac{\partial H}{\partial x}(t,x(t),p(t)), \quad t \ \in \ [0,T] \end{cases}$$

Then x is optimal if and only if $x(0)=\xi_0,\ p(0)=-\frac{\partial V}{\partial x}(0,\xi_0)$.

The value function can be also used to construct the optimal feedback map:

$$G(t,x) \ = \ \left\{ v \ \in \ f(t,x,U(t)) \mid \ \frac{\partial V}{\partial(1,v)}(t,x) \ = \ 0 \right\}$$

Namely the following property holds true: a solution $\overline{x}$ to (42) is optimal for our minimization problem if and only if it is a solution to the differential inclusion

$$x' \in G(t,x), \quad x(0) = \xi_0 \tag{44}$$

To investigate the regularity of the set-valued map G, we show in Subsection 4 that for sufficiently smooth f and g, the value function is semiconcave. As a consequence, we obtain that the feedback map G is upper semicontinuous on $[0,T[\times X$ and has nonempty compact images. In particular whenever G is single-valued, it is continuous and optimal solutions are continuously differentiable.

If the data are convex, then the value function is convex, G has convex values and inclusion (44) fits the well investigated framework of upper semicontinuous convex valued maps (see [1].)

3.1 Value Function

3.1.1 Mayer and Bolza Problems

Consider $T > 0$, a complete separable metric space $\mathcal{Z}$, a set-valued map $U : [0,T] \hookrightarrow \mathcal{Z}$ and a map $f : [0,T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^n$. We associate with it the control system

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U(t) \tag{45}$$

Let an extended function $g : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and $\xi_0 \in \mathbf{R}^n$ be given. Consider the minimization problem, called *Mayer's optimal control problem*:

$$\min \{g(x(T)) \mid x \text{ is a solution to } (45),\ x(0) = \xi_0\} \tag{46}$$

The value function associated with this problem is defined by: for all $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$

$$V(t_0,x_0) = \inf\{g(x(T)) \mid x \text{ is a solution to (45)},\ x(t_0) = x_0\} \tag{47}$$

We impose the following assumptions

$$\begin{cases} \forall\, (x,u) \in \mathbf{R}^n \times \mathcal{Z}, \quad f(\cdot,x,u) \text{ is measurable} \\ \text{For a.e. } t \in [0,T], \quad f(t,\cdot,\cdot) \text{ is continuous} \\ U(\cdot) \text{ is measurable and has closed nonempty images} \end{cases} \tag{48}$$

and

$$\begin{cases} i) & \exists\, k \in L^1(0,T) \text{ such that for a.e. } t \in [0,T], \\ & \forall\, u \in U(t), f(t,\cdot,u) \text{ is } k(t) - \text{Lipschitz} \\ ii) & \exists\, \gamma \in L^1(0,T) \text{ such that for a.e. } t \in [0,T], \\ & \sup_{u\in U(t)} \|f(t,0,u)\| \le \gamma(t) \\ iii) & \text{g is locally Lipschitz} \end{cases} \tag{49}$$

These assumptions imply that V is actually equal to the value function of the relaxed problem in which system (45) is replaced by the differential inclusion

$$x'(t) \in \overline{co}\,(f(t,x(t),U(t))) \text{ almost everywhere} \tag{50}$$

Hence for all $(t_0,x_0) \in [0,T]\times \mathbf{R}^n$,

$$V(t_0,x_0) = \inf\{g(x(T)) \mid x \text{ solves (50)}, \ x(t_0)=x_0\} \tag{51}$$

The *Bolza problem* has the same nature, but its cost involves the integral functional: Consider in addition a function $L : [0,T]\times\mathbf{R}^n\times\mathcal{Z} \mapsto \mathbf{R}$ and the following minimization problem:

$$\text{minimize}\left\{ g(x(T)) + \int_0^T L(t,x(t),u(t))dt \right\} \tag{52}$$

over all solution-control pairs (x,u) to (45) with $x(0)=\xi_0$.

We denote by $\widehat{x}=(x^0,x)$ elements of $\mathbf{R}^{n+1}$ and we set

$$\forall\, t\in[0,T],\ \widehat{x}=(x^0,x),\ u\in\mathcal{Z},\ \widehat{f}(t,\widehat{x},u) := (L(t,x,u),f(t,x,u))$$

Then it is not difficult to realize that a solution-control pair $(z,\overline{u})$ of (45) is optimal for problem (52) if and only if the map

$$t \mapsto \widehat{z}(t) := \left(\int_0^t L(s,z(s),\overline{u}(s))ds,\ z(t)\right)$$

solves the problem

$$\text{minimize}\left(g(x(T)) + x^0(T)\right)$$

over all solutions to the control system

$$\widehat{x}'(t) = \widehat{f}(t,\widehat{x}(t),u(t)),\quad u(t)\in U(t),\quad \widehat{x}(0)=(0,\xi_0)$$

This new problem is of Mayer's type.

3.1.2 Lipschitz Continuity of the Value Function

More generally consider an extended function $g : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$, a set-valued map $F : [0,T] \times \mathbf{R}^n \hookrightarrow \mathbf{R}^n$, $\xi_0 \in \mathbf{R}^n$ and the differential inclusion

$$x'(t) \ \in \ F(t, x(t)) \ \text{ almost everywhere} \tag{53}$$

We investigate the minimization problem

$$\min \{g(x(T)) \mid x \text{ is a solution to (53)}, \ x(0) = \xi_0\}$$

The corresponding value function is given by:

For all $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$,

$$V(t_0, x_0) \ = \ \inf\{g(x(T)) \mid \ x \ \text{ solves } \ (53), \ x(t_0) = x_0\} \tag{54}$$

Let $\mathcal{S}_{[t_0,T]}(x_0)$ denote the set of solutions to (53) starting at x_0 at time t_0 and defined on the time interval $[t_0, T]$. The value function is non decreasing along solutions to (53):

$$\forall\, x \in \mathcal{S}_{[t_0,T]}(x_0), \ \ \forall\, t_0 \le t_1 \le t_2 \le T, \ \ V(t_1, x(t_1)) \le V(t_2, x(t_2))$$

and satisfies the following *dynamic programming principle*:

$$\forall\, t \in [t_0, T], \ \ V(t_0, x_0) \ = \ \inf \Big\{ V(t, x(t)) \mid \ x \in \mathcal{S}_{[t_0,T]}(x_0) \Big\} \tag{55}$$

Furthermore $x \in \mathcal{S}_{[t_0,T]}(x_0)$ is optimal for problem (54) if and only if $V(t, x(t)) \equiv g(x(T))$.

We impose the following assumptions on F and g

$$\left\{ \begin{array}{ll} i) & F \ \text{ has closed nonempty images} \\ \\ ii) & \forall\, x \in \mathbf{R}^n, \ \ F(\cdot, x) \ \text{ is measurable} \\ \\ iii) & \exists\, k \ \in \ L^1(0,T), \ \ \forall\, t \ \in \ [0,T], \ \ \forall\, x, y \ \in \ \mathbf{R}^n, \\ & F(t,x) \ \subset \ F(t,y) + k(t)\, \|x - y\|\, B \\ \\ iv) & \exists\, \gamma \in L^1(0,T), \ \forall\, t \in [0,T], \ \ F(t,0) \subset \gamma(t) B \\ \\ v) & g \ \text{ is locally Lipschitz} \end{array} \right. \tag{56}$$

and observe that if the map f from Subsection 3.1.1 satisfies (48) and (49), then, by Sections 1,2, assumptions (56) hold true for the set-valued map $F(t,x) := \overline{co}\,(f(t,x,U(t)))$. This and (51) yield that results of this subsection may be applied as well to the Mayer problem considered in Subsection 3.1.1.

We recall that the directional derivative of a function $\varphi : \mathbf{R}^m \mapsto \mathbf{R}$ at $x_0 \in \mathbf{R}^m$ in the direction $v \in \mathbf{R}^m$ (when it exists) is defined by

$$\frac{\partial \varphi}{\partial v}(x_0) \;=\; \lim_{h \to 0+} \frac{\varphi(x_0 + hv) - \varphi(x_0)}{h}$$

Theorem 3.1 *Assume (56). Then for every $R > 0$, there exists $L_R > 0$ such that*

i) For all $(t_0, x_0) \in [0,T] \times B_R(0)$ and every solution $x \in \mathcal{S}_{[t_0,T]}(x_0)$

$$\forall\, t \in [t_0, T], \quad \|x(t)\| \;\leq\; L_R$$

and the map $[t_0, T] \ni t \mapsto V(t, x(t))$ is absolutely continuous.

Furthermore for almost every $t \in [t_0, T]$, the directional derivative

$$\frac{\partial V}{\partial (1, x'(t))}(t, x(t))$$

does exist.

ii) For all $t \in [0,T]$, $V(t,\cdot)$ is L_R-Lipschitz on $B_R(0)$

Finally, if for all $R > 0$, there exists $c_R \geq 0$ such that

$$\text{For a.e. } t \in [0,T], \;\; \forall\, x \in B_R(0), \quad \sup_{y \in F(t,x)} \|y\| \;\leq\; c_R \tag{57}$$

then for every $R > 0$, there exists $C_R > 0$ such that

$$\forall\, x \;\in\; B_R(0), \quad V(\cdot, x) \;\text{ is }\; C_R - \text{Lipschitz}$$

Proof — Consider any solution $x \in \mathcal{S}_{[t_0,T]}(x_0)$ to differential inclusion (53). Then for almost all $t \in [t_0, T]$

$$x'(t) \;\in\; F(t, x(t)) \;\subset\; F(t,0) \;+\; k(t)\,\|x(t)\|\,B$$

Thus

$$\forall\, t \;\in\; [t_0, T], \;\; \|x(t)\| \;\leq\; \|x_0\| + \int_{t_0}^{t} \gamma(s)ds + \int_{t_0}^{t} k(s)\,\|x(s)\|\,ds$$

This and Gronwall's lemma yield the first statement. Since φ is locally Lipschitz we deduce *ii*) from Filippov's theorem.

Let $x_1 \in \mathcal{S}_{[t_0,T]}(x_0)$. We claim that the map $t \mapsto V(t, x_1(t))$ is absolutely continuous. Indeed fix $t_0 \le t_1 < t_2 \le T$. By (55), there exists $x_2 \in \mathcal{S}_{[t_1,T]}(x_1(t_1))$ such that

$$V(t_2, x_2(t_2)) \;\le\; V(t_1, x_1(t_1)) + |t_2 - t_1|$$

Then from *i*) we deduce that for $i = 1, 2$

$$\begin{aligned} \|x_i(t_2) - x_i(t_1)\| &\le \textstyle\int_{t_1}^{t_2} \gamma(s)ds + \int_{t_1}^{t_2} k(s)\,\|x_i(s)\|\,ds \\ &\le \textstyle\int_{t_1}^{t_2} \gamma(s)ds + L_{\|x_0\|}\int_{t_1}^{t_2} k(s)ds \end{aligned}$$

Thus, by *ii*), for a constant L depending only on $\|x_0\|$

$$\begin{aligned} 0 &\le V(t_2, x_1(t_2)) - V(t_1, x_1(t_1)) \\ &\le V(t_2, x_1(t_2)) - V(t_2, x_2(t_2)) + |t_2 - t_1| \\ &\le L\,\|x_1(t_2) - x_2(t_2)\| + |t_2 - t_1| \qquad (58) \\ &\le L(\|x_1(t_2) - x_1(t_1)\| + \|x_2(t_2) - x_1(t_1)\|) + |t_2 - t_1| \\ &\le 2L\textstyle\int_{t_1}^{t_2} \gamma(s)ds + 2L_{\|x_0\|}L\int_{t_1}^{t_2} k(s)ds + |t_2 - t_1| \end{aligned}$$

Recall the following characterization of absolutely continuous maps:

A function $f : [a, b] \mapsto \mathbf{R}$ is absolutely continuous if and only if

$$\begin{cases} i) \quad \exists\, v(f) > 0,\ \forall\, a = a_1 \le b_1 \le .... \le a_m \le b_m = b, \\ \qquad \textstyle\sum_{i=1}^m |f(b_i) - f(a_i)| \le v(f) \\ ii) \quad \forall\, \varepsilon > 0,\ \exists\, \delta > 0 \text{ such that } \forall\, a \le a_i < b_i \le b,\ i = 1, ..., m \\ \qquad \text{satisfying }]a_i, b_i[\cap]a_j, b_j[= \emptyset \text{ for } i \ne j,\ \textstyle\sum_{i=1}^m (b_i - a_i) \le \delta \\ \qquad \text{we have } \textstyle\sum_{i=1}^m |f(b_i) - f(a_i)| \le \varepsilon \end{cases}$$

Thus, by (58), the map $t \mapsto \varphi(t) := V(t, x_1(t))$ is absolutely continuous.

Fix $t \in [t_0, T]$ such that φ and x_1 are differentiable at t. Then from the local Lipschitz continuity of V with respect to the second variable

$$\lim_{h\to 0+} \frac{V(t+h, x_1(t)+hx_1'(t)) - V(t, x_1(t))}{h} = \lim_{h\to 0+} \frac{\varphi(t+h)-\varphi(t)}{h}$$

To prove the last statement of our theorem, observe that (57) and i) imply that for all $R > 0$, there exists l_R such that every $x \in \mathcal{S}_{[t_0,T]}(x_0)$ is l_R-Lipschitz whenever $x_0 \in B_R(0)$. Fix $0 \le t_0 < t_1 \le T$, $x_0 \in B_R(0)$. By (55) there exists $x \in \mathcal{S}_{[t_0,T]}(x_0)$ such that $V(t_1, x(t_1)) \le V(t_0, x_0) + |t_1 - t_0|$. Then

$$\begin{aligned} &|V(t_1, x_0) - V(t_0, x_0)| \\ &\le |V(t_1, x(t_1)) - V(t_0, x_0)| + |V(t_1, x(t_1)) - V(t_1, x_0)| \\ &\le |t_1 - t_0| + L_R \|x(t_1) - x_0\| \le (L_R l_R + 1)|t_1 - t_0| \quad \diamond \end{aligned}$$

3.1.3 Optimal Feedback

When the value function is directionally differentiable, it has many properties related to dynamics of the system.

Proposition 3.2 *Assume (56). If for some $(t_0, x_0) \in [0, T[\times \mathbf{R}^n$, F is lower semicontinuous at (t_0, x_0) and for some $v \in \overline{co}\,(F(t_0, x_0))$, the directional derivative of V at (t_0, x_0) in the direction $(1, v)$ exists, then this directional derivative is nonnegative.*

Proof — Consider a solution $x(\cdot)$ to differential inclusion (53) satisfying $x(t_0) = x_0$, $x'(t_0) = v$. Since $V(t, \cdot)$ is Lipschitz on a neighborhood of x_0 with the Lipschitz constant independent of t and since V is non decreasing along solutions to (53),

$$\begin{aligned} &\lim_{h\to 0+} \frac{V(t_0+h, x_0+hv) - V(t_0, x_0)}{h} \\ &= \lim_{h\to 0+} \frac{V(t_0+h, x(t_0+h)) - V(t_0, x_0)}{h} \ge 0 \quad \diamond \end{aligned}$$

To characterize optimal solutions, we introduce the following feedback map $G : [0, T] \times \mathbf{R}^n \hookrightarrow \mathbf{R}^n$ defined by

$$\forall\, (t, x) \in [0, T] \times \mathbf{R}^n, \quad G(t, x) = \left\{ v \in F(t, x) \mid \frac{\partial V}{\partial (1, v)}(t, x) = 0 \right\}$$

(notice that the sets $G(t,x)$ may be empty.)

Theorem 3.3 *Assume (56) and let $t_0 \in [0,T]$. Then the following two statements are equivalent:*

i) x is a solution to the differential inclusion

$$x'(t) \in G(t,x(t)) \text{ almost everywhere in } [t_0,T] \tag{59}$$

ii) x is a solution to differential inclusion (53) defined on the time-interval $[t_0,T]$ and for every $t \in [t_0,T]$, $V(t,x(t)) = g(x(T))$.

Proof — Fix a solution x to (53) defined on $[t_0,T]$ and set $\varphi(t) = V(t,x(t))$. By Theorem 3.1, φ is absolutely continuous and for almost all $t \in [t_0,T]$

$$\varphi'(t) = \frac{\partial V}{\partial(1,x'(t))}(t,x(t))$$

Assume that i) holds true. Hence, for almost every $t \in [t_0,T]$, the set $G(t,x(t))$ is nonempty and $\varphi'(t) = 0$ almost everywhere in $[t_0,T]$. Consequently $\varphi \equiv V(T,x(T)) = g(x(T))$.

Assume next that ii) is verified. Then, differentiating the map $t \mapsto \varphi(t)$, we obtain that for every $t_0 < t < T$, $\varphi'(t) = 0$. Therefore for almost all $t \in [t_0,T]$, $x'(t) \in G(t,x(t))$. $\diamond$

Corollary 3.4 *Assume (56). Then, a solution $x \in \mathcal{S}_{[t_0,T]}(x_0)$ is optimal for problem (54) if and only if it is a solution to differential inclusion (59) satisfying the initial condition $x(t_0) = x_0$.*

Theorem 3.5 *Assume (56) and that the images of F are convex. Then for every $t_0 \in [0,T]$ and $x_0 \in \mathbf{R}^n$, inclusion (59) has at least one solution $x \in \mathcal{S}_{[t_0,T]}(x_0)$.*

Proof — By Theorem 1.38, problem (54) has at least one optimal solution $\overline{x}$. Furthermore $V(t,\overline{x}(t)) \equiv g(\overline{x}(T))$. Theorem 3.3 ends the proof.

3.2 Maximum Principle for Free End Point Problems

3.2.1 Adjoint System

Consider a complete separable metric space $\mathcal{Z}$, real numbers $t_0 < T$ and

$$f : [t_0,T] \times \mathbf{R}^n \times \mathcal{Z} \mapsto \mathbf{R}^n$$

Let $U : [t_0, T] \rightsquigarrow \mathcal{Z}$ be a set-valued map and consider the control system

$$x' \ = \ f(t, x, u(t)), \quad u(t) \ \in \ U(t), \quad t \ \in \ [t_0, T] \tag{60}$$

Fix a state-control solution $(z, \overline{u})$ to control system (60). We assume that (48) holds true and

$$\begin{cases} \text{The derivative } \frac{\partial f}{\partial x}(t, z(t), \overline{u}(t)) \text{ exists a.e. in } [t_0, T] \\ \\ \text{For some } \varepsilon > 0, \ k \in L^1(t_0, T) \text{ and for a.e. } t \in [t_0, T] \\ \forall\, u \in U(t), \quad f(t, \cdot, u) \text{ is } k(t) - \text{Lipschitz on } B_\varepsilon(z(t)) \end{cases} \tag{61}$$

Denote by $X(\cdot)$ the fundamental solution to the linear system

$$X'(t) \ = \ \frac{\partial f}{\partial x}(t, z(t), \overline{u}(t))X(t), \quad X(t_0) \ = \ \mathrm{Id} \tag{62}$$

We recall the following property of the fundamental solution.

Proposition 3.6 *Let $X(t)^\star$ denote the transposed matrix. Then every solution p to the adjoint system*

$$-p' \ = \ \left(\frac{\partial f}{\partial x}(s, z(s), \overline{u}(s))\right)^\star p \tag{63}$$

verifies $p(t) = (X(t)^\star)^{-1} X(T)^\star p(T)$ for all $t \in [t_0, T]$.

Proof — Set $A(s) = \frac{\partial f}{\partial x}(s, z(s), \overline{u}(s))$. Then, differentiating the identity $X(t)X(t)^{-1} = \mathrm{Id}$, we obtain that for almost all $t \in [t_0, T]$

$$\begin{gathered} 0 \ = \ X'(t)X(t)^{-1} \ + \ X(t)(X^{-1})'(t) \\ = A(t)X(t)X(t)^{-1} + X(t)(X^{-1})'(t) = A(t) + X(t)(X^{-1})'(t) \end{gathered}$$

Hence $(X^{-1})'(t) = -X(t)^{-1}A(t)$ and therefore $(X(\cdot)^\star)^{-1}$ is the fundamental solution to

$$Y'(t) \ = \ -A(t)^\star Y(t), \quad Y(t_0) \ = \ \mathrm{Id}$$

Thus the solution $p(\cdot)$ to (63) verifies $p(t) = (X(t)^\star)^{-1} p(t_0)$ for all $t \in [t_0, T]$. So, $p(t_0) = X(T)^\star p(T)$ and the proof follows. $\diamond$

Let us associate with control system (60) the Hamiltonian $H : [0, T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ defined by

$$H(t, x, p) \ = \ \sup_{u \in U(t)} \ < p, f(t, x, u) >$$

Under assumptions (48), (49), H is measurable with respect to t, locally Lipschitz with respect to (x,p) and convex with respect to the third variable p.

Denote by $R(t)$ the reachable set of (60) at time t from $z(t_0)$:

$$R(t) \;=\; \{x(t) \mid \; x \;\text{ is a solution to }\; (60), \;\; x(t_0) = z(t_0)\}$$

by $T_{R(t)}(z(t))$ the contingent cone to $R(t)$ at $z(t)$ and by

$$N^0_{R(t)}(z(t)) \;:=\; \Big(T_{R(t)}(z(t))\Big)^-$$

the *subnormal cone* to $R(t)$ at $z(t)$ (negative polar cone of the contingent cone to $R(t)$ at $z(t)$.)

Lemma 3.7 *Assume that (48) and (61) hold true. If p is a solution to adjoint system* (63) *such that $p(T) \in N^0_{R(T)}(z(T))$, then*

$$\forall\, t \;\in\; [t_0,T], \quad p(t) \;\in\; N^0_{R(t)}(z(t)) \tag{64}$$

and the following maximum principle *holds true:*

$$\langle p(t), f(t,z(t),\overline{u}(t))\rangle \;=\; H(t,z(t),p(t)) \;\text{ a.e. in }\; [t_0,T] \tag{65}$$

Proof — From Theorem 1.40 we deduce that for all $v \in T_{R(t)}(z(t))$, $X(T)X(t)^{-1}v \in T_{R(T)}(z(T))$. Thus $\langle p(T), X(T)X(t)^{-1}v\rangle \le 0$ and, using Proposition 3.6, we obtain

$$< p(t), v > \;=\; \Big\langle (X(t)^\star)^{-1}\, X(T)^\star p(T),\; v\Big\rangle \;\le\; 0$$

If w solves the linearized system (22) and $t_0 = 0,\; w(0) = 0$, then $w(T) \in T_{R(T)}(z(T))$ and $w(T) = \int_0^T X(T)X(s)^{-1}v(s)ds$. Thus

$$0 \ge < p(T), w(T) > = < \int_0^T < (X(s)^\star)^{-1}X(T)^\star p(T), v(s) > ds$$

Thus $< \int_0^T < p(s), v(s)ds > \le 0$ for all integrable selection $v(s) \in T_{\overline{co}f(s,z(s),U(s))}(z'(s))$. Since

$$0 \in \overline{co}f(s,z(s),U(s)) - z'(s) \subset T_{\overline{co}f(s,z(s),U(s))}(f(s,z(s),\overline{u}(s)))$$

we deduce, using results on measurable set-valued maps from Section 1 that

$$0 \ge < \int_0^T \sup_{v \in \overline{co}f(s,z(s),U(s))-z'(s)} < p(s), v > ds \ge 0$$

which yields (65) $\diamond$

3.2.2 Maximum Principle

Let $g : \mathbf{R}^n \mapsto \mathbf{R}$ be a differentiable function and $x_0 \in \mathbf{R}^n$ be given. Consider the problem

$$\text{minimize } g(x(T))$$

over all solutions to control system (60) satisfying $x(t_0) = x_0$.

Theorem 3.8 *If a state-control solution $(z, \overline{u})$ solves the above problem and (48), (61) hold true, then the solution p to adjoint system* (63) *such that*

$$p(T) \;=\; -\nabla g(z(T))$$

satisfies (64) *and maximum principle* (65).

Remark — The map $p(\cdot)$ in the above theorem is called *co-state* or *adjoint variable* associated with $(z, \overline{u})$. $\diamond$

Proof — Since z is optimal, we have

$$\min_{y \in R(T)} g(y) \;=\; g(z(T))$$

Let $v \in T_{R(T)}(z(T))$ and $h_n \to 0+, v_n \to v$ be such that $z(T) + h_n v_n \in R(T)$. Then $g(z(T) + h_n v_n) - g(z(T)) \geq 0$. Dividing by h_n and taking the limit we obtain $< \nabla g(z(T)), v > \geq 0$. Hence $-\nabla g(z(T)) \in N^0_{R(T)}(z(T))$. Lemma 3.7 yields the conclusion. $\diamond$

3.3 Necessary and Sufficient Conditions for Optimality

We begin this subsection with a sufficient condition for optimality involving the superdifferential of the value function.

3.3.1 Sufficient Conditions

Theorem 3.9 *Assume that (48), (49) hold true and let $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$. Consider a solution $z : [t_0, T] \mapsto \mathbf{R}^n$ to control system (45) with $z(t_0) = x_0$ and let $\overline{u}$ be a corresponding control. If for almost every $t \in [t_0, T]$, there exists $p(t) \in \mathbf{R}^n$ such that*

$$(\langle p(t), z'(t) \rangle\,,\; -p(t)) \in \partial_+ V(t, z(t)) \tag{66}$$

then z is optimal for problem (47).

Proof — By Theorem 3.1 the map $\psi(t) := V(t, z(t))$ is absolutely continuous. Let $t \in [t_0, T]$ be such that the derivatives $\psi'(t)$ and $z'(t)$ do exist and (66) holds true. Then, using Theorem 3.1 and Proposition 1.7

$$0 \;=\; \langle (< p(t), z'(t) >, -p(t)), (1, z'(t)) \rangle \geq D_{\downarrow} V(t, z(t))(1, z'(t))$$

$$\geq \; \limsup_{h \to 0+} \frac{V(t+h, z(t+h)) - V(t, z(t))}{h} \;=\; \psi'(t)$$

This yields that ψ is non increasing. Since the value function is also non decreasing along solutions to control system (45), we deduce that the map $t \mapsto V(t, z(t))$ is constant. So z is optimal. $\diamond$

3.3.2 Necessary and Sufficient Conditions

Theorem 3.10 *Assume (48), (49), that f is differentiable with respect to x and g is differentiable. A state-control solution $(z, \overline{u})$ to control system (45) with $z(t_0) = x_0$ is optimal for problem (47) if and only if the solution $p : [t_0, T] \mapsto \mathbf{R}^n$ to the adjoint system*

$$-p'(t) \;=\; \left(\frac{\partial f}{\partial x}(t, z(t), \overline{u}(t)) \right)^{\star} p(t), \quad p(T) \;=\; -\nabla g(z(T)) \tag{67}$$

satisfies the maximum principle

$$< p(t), f(t, z(t), \overline{u}(t)) > = \; H(t, z(t), p(t)) \;\text{ a.e. in }\; [t_0, T] \tag{68}$$

and the generalized transversality conditions

$$(H(t, z(t), p(t)), -p(t)) \;\in\; \partial_+ V(t, z(t)) \;\text{ a.e. in }\; [t_0, T] \tag{69}$$

$$-p(t) \;\in\; \partial_+ V_x(t, z(t)) \;\text{ for every }\; t \in [t_0, T] \tag{70}$$

where $\partial_+ V_x(t, z(t))$ denotes the superdifferential of $V(t, \cdot)$ at $z(t)$.

Furthermore, if V is semiconcave and H is continuous, then (69) holds true everywhere in $[t_0, T]$.

Proof — Sufficiency is a straightforward consequence of Theorem 3.9 and (68), (69). The fact that (67) and (68) are necessary follows from Theorem 3.8.

Fix $t \in [t_0, T]$, $v \in \mathbf{R}^n$ and consider the solution $w(\cdot)$ to the linear system

$$\left\{ \begin{array}{rcl} w'(s) & = & \frac{\partial f}{\partial x}(s, z(s), \overline{u}(s)) w(s), \quad s \in [t, T] \\ \\ w(t) & = & v \end{array} \right.$$

Then $w(T) = X(T)X(t)^{-1}v$, where $X(\cdot)$ is the fundamental solution to (62). For every $h > 0$, let x_h be the solution to the differential equation

$$\begin{cases} x'(s) &= f(s, x(s), \overline{u}(s)), \quad s \in [t, T] \\ x(t) &= z(t) + hv \end{cases}$$

From the variational equation we know that the difference quotients $(x_h - z)/h$ converge uniformly to w.

To prove (70) it is enough to observe that

$$\langle -p(t), v\rangle = \langle (X(t)^\star)^{-1} X(T)^\star \nabla\varphi(z(T)), v\rangle = \langle \nabla\varphi(z(T)), w(T)\rangle$$

$$\geq \limsup_{h\to 0+} (V(t, z(t) + hv) - V(t, z(t)))/h$$

To prove the necessity of (69) fix $t \in [0, T[$ such that $z'(t) = f(t, z(t), \overline{u}(t))$ and equality (68) holds true, $v \in \mathbf{R}^n$, $\alpha \in \mathbf{R}$. Then from (68), using that $V(t, \cdot)$ is locally Lipschitz, that V is non decreasing along solutions to (88) and is constant along z, we deduce

$$\limsup_{h\to 0+} \left(V(t + \alpha h, z(t) + h(\alpha z'(t) + v)) - V(t, z(t))\right)/h$$

$$= \limsup_{h\to 0+} \left(V(t + \alpha h, z(t + \alpha h) + hw(t + \alpha h)) - V(t, z(t))\right)/h$$

$$= \limsup_{h\to 0+} \left(V(t + \alpha h, x_h(t + \alpha h)) \; - \; V(t, z(t))\right)/h$$

$$\leq \limsup_{h\to 0+} \left(\varphi(x_h(T)) - \varphi(z(T))\right)/h \; = \; \langle \nabla\varphi(z(T)), w(T)\rangle$$

$$= \langle \nabla\varphi(z(T)), X(T)X(t)^{-1}v\rangle = \left\langle (X(t)^\star)^{-1} X(T)^\star \nabla\varphi(z(T)), v\right\rangle$$

$$= \; \langle -p(t), v\rangle \; = \; \langle -p(t), -\alpha z'(t)\rangle \; + \; \langle -p(t), \alpha z'(t) + v\rangle$$

$$= \; \alpha H(t, z(t), p(t)) \; + \; \langle -p(t), \alpha z'(t) + v\rangle$$

Hence we deduce that for every $\alpha \in \mathbf{R}$ and $v_1 \in \mathbf{R}^n$

$$D_\downarrow V(t, z(t))(\alpha, v_1) \; \leq \; \alpha H(t, z(t), p(t)) \; + \; \langle -p(t), v_1\rangle$$

and (69) follows from Proposition 1.7.

When V is semiconcave, then, from Theorem 1.11, its superdifferential is upper semicontinuous. Thus the last statement results from (69) and continuity of $H(\cdot)$, $p(\cdot)$, $z(\cdot)$. $\diamond$

For autonomous systems, that is with f and U independent of t, the Hamiltonian is constant along any optimal state/co-state pair (z, p). Indeed, recalling (68), (67), for almost all $s \in [t_0, T]$ we obtain

$$\begin{cases} H(z(t), p(t)) - H(z(s), p(s)) \\ \geq \langle p(t), f(z(t), \overline{u}(s)) - f(z(s), \overline{u}(s)) \rangle + \langle p(t) - p(s), f(z(s), \overline{u}(s)) \rangle \\ = \ o(|t - s|) \end{cases}$$

for all $t \in [t_0, T]$. Since the above argument is symmetric,

$$|H(z(t), p(t)) - H(z(s), p(s))| \ \leq \ o(|t - s|)$$

Using that $t \mapsto H(z(t), p(t))$ is absolutely continuous, we deduce that $H(z(t), p(t))$ is constant.

When the Hamiltonian $H(t, \cdot, \cdot)$ is differentiable at $(z(t), p(t))$ for all $t \in [t_0, T]$, then z and the co-state p satisfy the *Hamiltonian system*

$$\begin{cases} z'(t) \ = \ \frac{\partial H}{\partial p}(t, z(t), p(t)) \\ \\ p'(t) \ = \ -\frac{\partial H}{\partial x}(t, z(t), p(t)) \ \text{ a.e. in } [t_0, T] \end{cases}$$

This follows from Theorem 3.10 and

Proposition 3.11 *Let $(t, z, \overline{p}) \in [t_0, T] \times \mathbf{R}^n \times \mathbf{R}^n$ and $\overline{u} \in U(t)$ be such that $\langle \overline{p}, f(t, z, \overline{u}) \rangle = H(t, z, \overline{p})$. Then*

i) If $H(t, \cdot, \overline{p})$ is differentiable at z, then

$$\frac{\partial H}{\partial x}(t, z, \overline{p}) \ = \ \left(\frac{\partial f}{\partial x}(t, z, \overline{u}) \right)^\star \overline{p}$$

ii) If $H(t, z, \cdot)$ is differentiable at $\overline{p}$, then

$$\frac{\partial H}{\partial p}(t, z, \overline{p}) \ = \ f(t, z, \overline{u})$$

Proof — It is enough to observe that for every $v \in \mathbf{R}^n$

$$\lim_{h \to 0+} \frac{H(t, z+hv, \overline{p}) - H(t, z, \overline{p})}{h} \ \geq \ \lim_{h \to 0+} \frac{\langle \overline{p}, \, f(t, z+hv, \overline{u}) - f(t, z, \overline{u}) \rangle}{h}$$

$$= \ \left\langle \overline{p}, \ \frac{\partial f}{\partial x}(t, z, \overline{u}) v \right\rangle \ = \ \left\langle \left(\frac{\partial f}{\partial x}(t, z, \overline{u}) \right)^\star \overline{p}, \ v \right\rangle$$

and

$$\lim_{h \to 0+} \frac{H(t, z, \overline{p}+hv) - H(t, z, \overline{p})}{h} \ \geq \ \lim_{h \to 0+} \frac{\langle \overline{p}+hv, \, f(t, z, \overline{u}) \rangle - \langle \overline{p}, \, f(t, z, \overline{u}) \rangle}{h}$$

$$= \ \langle v, f(t, z, \overline{u}) \rangle \quad \diamond$$

3.3.3 Co-state and Superdifferentials of Value

Proposition 3.12 *Assume (56) and let z be an optimal solution to problem (54). Then for almost every $t \in [t_0, T]$,*

$$\forall\, (p_t, p_x) \in \partial_+ V(t, z(t)), \quad -p_t \,-\, \langle p_x, z'(t)\rangle \,=\, 0 \tag{71}$$

Furthermore, if F is lower semicontinuous, then for almost every $t \in [t_0, T]$,

$$\forall\, (p_t, p_x) \in \partial_+ V(t, z(t)), \quad -p_t \,+\, H(t, z(t), -p_x) \,=\, 0 \tag{72}$$

If (57) holds true and F is continuous, then (72) is satisfied for all $t_0 < t < T$.

Proof — Let $t \in\,]t_0, T[$ be such that $z'(t) \in F(t, z(t))$. Then for all $(p_t, p_x) \in \partial_+ V(t, z(t))$

$$0 \geq \limsup_{s \to t+} \frac{V(s, z(s)) - V(t, z(t)) - p_t(s-t) - \langle p_x, z(s) - z(t)\rangle}{|s-t| + \|z(s) - z(t)\|}$$

Since $V(\cdot, z(\cdot))$ is constant, the above estimate yields

$$0 \;\geq\; \lim_{s \to t+} \frac{-p_t(s-t) - \langle p_x, z(s) - z(t)\rangle}{|s-t|} \;=\; -p_t \,-\, \langle p_x, z'(t)\rangle$$

By the same arguments, taking $s \to t-$ we derive $0 \geq p_t + \langle p_x, z'(t)\rangle$ and so (71) is proved. Assume next that F is lower semicontinuous and fix $v \in F(t_0, x_0), t_0 < T$. Consider $x \in \mathcal{S}_{[t,T]}(x_0)$ satisfying $x'(t_0) = v$. Since for all small $h > 0$, $V(t_0, x_0) \leq V(t_0 + h, x(t_0 + h))$, we deduce that $D_{\downarrow} V(t_0, x_0)(1, v) \geq 0$. Consequently, for all $(p_t, p_x) \in \partial_+ V(t_0, x_0)$, $p_t + < p_x, v > \geq 0$. But this yields $-p_t + H(t_0, x_0, -p_x) \leq 0$. From the last inequality and (71) we get (72).

If (57) holds true and F is continuous, then z is Lipschitz. Fix $t_0 < t < T$. Then for a sequence $h_n \to 0+$ and some $v \in \overline{co}\,(F(t, z(t)))$

$$\lim_{n \to \infty} \frac{z(t - h_n) - z(t)}{h_n} \;=\; -v$$

Fix $(p_t, p_x) \in \partial_+ V(t, z(t))$. Applying exactly the same arguments as before, we deduce that $p_t + \langle p_x, v\rangle \leq 0$ yielding equality (72) at t. $\diamond$

Theorem 3.13 *Assume (48), (49), that f is differentiable with respect to x and g is differentiable. Suppose further that $V(t_0,\cdot)$ is differentiable at x_0 and let $(z,\overline{u})$ be an optimal state-control solution to problem (47). Then the co-state $p:[t_0,T]\mapsto \mathbf{R}^n$ corresponding to $(z,\overline{u})$ and given by Theorem 3.10 verifies*

$$\{-p(t)\} \;=\; \partial_+ V_x(t,z(t)) \;\text{ for all }\; t\in[t_0,T]$$

Hence if $V(t,\cdot)$ is semiconcave, then $\frac{\partial V}{\partial x}(t,z(t)) = -p(t)$.

Remark — In Subsection 4 below, we show that under some additional regularity assumptions on f, V is semiconcave. ◇

Proof — We already know from Theorem 3.10 that

$$-p(t) \;\in\; \partial_+ V_x(t,z(t)) \;\text{ for all }\; t\in[t_0,T]$$

Thus $p(t_0) = -\frac{\partial V}{\partial x}(t_0,x_0)$.

Fix $v\in\mathbf{R}^n$ and let w, x_h have the same meaning as in the proof of Theorem 3.10 with t replaced by t_0. Then, since V is non decreasing along solutions to control system (88) and constant along z,

$$\begin{aligned}\langle -p(t_0),v\rangle \;&=\; \left\langle \frac{\partial V}{\partial x}(t_0,x_0),v\right\rangle \;=\; \lim_{h\to 0+}\frac{V(t_0,x_0+hv)-V(t_0,x_0)}{h}\\ &\le\; \limsup_{h\to 0+}\frac{V(t,x_h(t))-V(t,z(t))}{h}\\ &=\; \limsup_{h\to 0+}\frac{V(t,z(t)+hw(t))-V(t,z(t))}{h}\end{aligned}$$

for all $t\in[t_0,T]$. Hence for every $q\in\partial_+ V_x(t,z(t))$ we have

$$\langle -p(t_0),v\rangle \;\le\; \langle q,w(t)\rangle \;=\; \langle q,X(t)v\rangle \;=\; \langle X(t)^\star q,v\rangle$$

where X denotes the fundamental solution to (62).

Since $v\in\mathbf{R}^n$ is arbitrary, $p(t_0) = -X(t)^\star q$. On the other hand, $p(\cdot)$ being a solution to (67), we know that $p(t_0) = X(t)^\star p(t)$ and we deduce that $-p(t) = q$. This yields that $\partial_+ V_x(t,z(t))$ is a singleton and ends the proof.

◇

3.3.4 Hamiltonian System

Whenever H happens to be more regular we can prove the following theorem concerning optimal design. For every $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$ define

$$\partial^\star V_x(t_0, x_0) \;=\; \partial^\star W(x_0)$$

where W is given by $W(x) = V(t_0, x)$.

Theorem 3.14 *Assume (48), (49), that f is differentiable with respect to x, g is differentiable and that $H(t,\cdot,\cdot)$ is differentiable on $\mathbf{R}^n \times (\mathbf{R}^n \backslash \{0\})$ for almost every $t \in [t_0, T]$. It is well known that $H(t,x,\cdot)$ is not differentiable at zero when $f(t,x,U(t))$ is not a singleton. For this reason we exclude zero in our differentiability assumptions.*

Further assume that the sets $f(t,x,U(t))$ are convex and compact and for every $R > 0$, there exists a nonnegative integrable function $l_R \in L^1(0,T)$ such that for all $x, y \in RB$ and $p, q \in RB \backslash \frac{1}{R}B$

$$\begin{cases} \left\| \frac{\partial H}{\partial x}(t,x,p) - \frac{\partial H}{\partial x}(t,y,q) \right\| \;+\; \left\| \frac{\partial H}{\partial p}(t,x,p) - \frac{\partial H}{\partial p}(t,y,q) \right\| \\ \qquad\qquad \leq\; l_R(t)(\|x-y\| + \|p-q\|) \end{cases} \tag{73}$$

Let $(t_0, x_0) \in [t_0, T] \times \mathbf{R}^n$ and $p_0 \neq 0$ be such that $-p_0 \in \partial^\star V_x(t_0, x_0)$. Then the Hamiltonian system

$$\begin{cases} x'(t) \;=\; \frac{\partial H}{\partial p}(t, x(t), p(t)), & x(t_0) \;=\; x_0 \\ p'(t) \;=\; -\frac{\partial H}{\partial x}(t, x(t), p(t)), & p(t_0) \;=\; p_0 \\ p(t) \;\neq\; 0 \text{ for all } t \in [t_0, T] \end{cases} \tag{74}$$

has a unique solution $(z(\cdot), \overline{p}(\cdot))$ defined on $[t_0, T]$. Moreover $z(\cdot)$ is optimal for problem (47).

Consequently, problem (47) has at least as many optimal solutions as there are elements in the set $\partial^\star V_x(t_0, x_0) \backslash \{0\}$.

Furthermore, if $\nabla g(\cdot)$ is continuous at $z(T)$, then $\overline{p}(\cdot)$ is the co-state corresponding to $z(\cdot)$ given by Theorem 3.10.

Remark — A typical example of a nonlinear control system with closed convex images is the affine system:

$$x' \;=\; f(x) \;+\; \sum_{i=1}^{m} u_i g_i(x), \quad u_i \in [a_i, b_i]$$

where f and g_i are maps from $\mathbf{R}^n$ to itself and $a_i \leq b_i$ are given numbers.

$\diamond$

Proof — From the very definition of $\partial^\star V_x(t_0, x_0)$ it follows that there exists a sequence x_k converging to x_0 such that $V(t_0, \cdot)$ is differentiable at x_k and

$$-p_0 \;=\; \lim_{k\to\infty} \frac{\partial V}{\partial x}(t_0, x_k)$$

Let (z_k, u_k) be an optimal state-control solution for problem (47) with x_0 replaced by x_k. By Theorem 3.10 (applied with x_0 replaced by x_k), for every k there exists an absolutely continuous function $\overline{p}_k : [t_0, T] \mapsto \mathbf{R}^n$ such that

$$\left\{ \begin{array}{rcl} -\overline{p}_k'(t) & = & \left(\frac{\partial f}{\partial x}(t, z_k(t), u_k(t))\right)^\star \overline{p}_k(t), \quad \text{a.e. in } [t_0, T] \\ -\overline{p}_k(t_0) & = & \frac{\partial V}{\partial x}(t_0, x_k), \quad \overline{p}_k(T) \;=\; -\nabla g(z_k(T)) \end{array} \right. \tag{75}$$

Therefore, $p_k(t) \neq 0$ for all $t \in [t_0, T]$ and sufficiently large k. By Proposition 3.11, for every $t \in [t_0, T]$,

$$\left\{ \begin{array}{rcl} z_k(t) & = & x_0 \;+\; \displaystyle\int_{t_0}^{t} \frac{\partial H}{\partial p}(s, z_k(s), \overline{p}_k(s))ds \\ \\ \overline{p}_k(t) & = & p_0 \;-\; \displaystyle\int_{t_0}^{t} \frac{\partial H}{\partial x}(s, z_k(s), \overline{p}_k(s))ds \end{array} \right. \tag{76}$$

Recalling assumptions (49) and Theorem 3.1, we conclude that z_k, $k = 1, \ldots$ are equicontinuous and equibounded. Furthermore, from (49) and (75) it follows that $\overline{p}_k$ are also equicontinuous and equibounded, because the maps $t \mapsto \frac{\partial f}{\partial x}(t, z_k(t), u_k(t))$ are integrably bounded on $[t_0, T]$. So, taking a subsequence and keeping the same notation, we may assume that $(z_k, \overline{p}_k)$ converge uniformly to some $(z, \overline{p})$ and $\frac{\partial f}{\partial x}(\cdot, z_k(\cdot), u_k(\cdot))$ converge weakly in $L^1(t_0, T; \mathbf{R}^n \times \mathbf{R}^n)$ to some $A(\cdot)$. In particular $\overline{p}(t_0) = p_0 \neq 0$ and $\overline{p}$ solves the linear system

$$-\overline{p}'(t) \;=\; A(t)^\star \overline{p}(t), \quad \text{almost everywhere in } [t_0, T]$$

Thus $\overline{p}(t) \neq 0$ for all $t \in [t_0, T]$. Fix $R > 1$ so that

$$\forall\, s \in [t_0, T], \quad \frac{2}{R} \;\leq\; \|\overline{p}(s)\| \;\leq\; \frac{R}{2}$$

Then, for all sufficiently large k and all $s \in [t_0, T]$, we have

$$\frac{1}{R} \leq \|p_k(s)\| \leq R$$

So, using (73) and taking the limit in (76), we deduce $(z, \overline{p})$ is a solution to Hamiltonian system (74).

Since $\overline{p}$ never vanishes, assumption (73) implies that $(z, \overline{p})$ is the only solution to (74). On the other hand

$$V(t_0, x_0) = \lim_{k\to\infty} V(t_0, z_k(t_0)) = \lim_{k\to\infty} g(z_k(T)) = g(z(T))$$

and therefore z is optimal for problem (47).

If ∇g is continuous at $z(T)$, then from (75) it follows that $\overline{p}(T) = -\nabla g(z(T)) \neq 0$. Let p_1 be a co-state corresponding to the optimal solution z given by Theorem 3.10. Then $p_1(t) \neq 0$ for all $t \in [t_0, T]$ and, by Proposition 3.11 it solves the problem

$$\begin{cases} -p'(t) = \frac{\partial H}{\partial x}(t, z(t), p(t)), \quad \text{a.e. in } [t_0, T] \\ \\ p(T) = -\nabla g(z(T)) \end{cases}$$

Since $\overline{p}$ is also a solution to this system, $p_1 = \overline{p}$ by uniqueness. $\diamond$

3.3.5 Uniqueness of Optimal Solution and Differentiability of Value Function

Theorem 3.14 yields that if $\partial^\star V_x(t_0, x_0)\backslash\{0\}$ is not a singleton, then optimal solution to (47) is not unique. We prove a similar statement under less restrictive regularity assumptions on $H(t, x, \cdot)$.

Theorem 3.15 *Assume (48) (49), that g is continuously differentiable, f is differentiable with respect to x, $f(t, x, U(t))$ are convex and compact and that for every $t \in [0, T]$, $\frac{\partial H}{\partial x}(t, \cdot, \cdot)$ is continuous.*

Further assume that for every $R > 0$, there exists a nonnegative integrable function $l_R \in L^1(0, T)$ such that

$$\forall\, x, y, p \in RB, \quad \left\| \frac{\partial H}{\partial x}(t, x, p) - \frac{\partial H}{\partial x}(t, y, p) \right\| \leq l_R(t)\, \|x - y\|$$

If problem (47) has a unique optimal solution z, then for all $t \in [t_0, T]$, $\partial^\star V_x(t, z(t))$ is a singleton and, consequently, $V(t, \cdot)$ is differentiable at $z(t)$.

Proof — Observe that for every $t \in [t_0, T]$, problem (47) has a unique optimal solution with (t_0, x_0) replaced by $(t, z(t))$. For this reason we prove the result only for $V(t_0, \cdot)$.

By Proposition 1.9, it suffices to show that $\partial^\star V_x(t_0, x_0)$ is a singleton. Let $p_1, p_2 \in \partial^\star V_x(t_0, x_0)$ and consider sequences $\{x_k^1\}$ and $\{x_k^2\}$ converging to x_0, such that

$$\lim_{k \to +\infty} \frac{\partial V}{\partial x}(t_0, x_k^i) \;=\; p_i, \quad i = 1, 2$$

Let z_k^i be optimal solutions to problem (47) with x_0 replaced by x_k^i, $i = 1, 2$ and denote by p_k^i the corresponding co-states given by Theorem 3.10. Then, by Proposition 3.11,

$$\begin{cases} (p_k^i)'(t) \;=\; -\frac{\partial H}{\partial x}(t, z_k^i(t), p_k^i(t)) \;\text{ a.e. in } [t_0, T] \\ p_k^i(T) \;=\; -\nabla g(z_k^i(T)), \;\; p_k^i(t_0) \;=\; -\frac{\partial V}{\partial x}(t_0, x_k^i) \end{cases}$$

By Theorem 3.1, z_k^i are bounded, equicontinuous and $V(T, z_k^i(T)) = g(z_k^i(T))$. Since the solution to (47) is unique by our assumptions, we deduce that z_k^i converge uniformly to z for $i = 1, 2$. Taking subsequences and keeping the same notations, we may assume that p_k^i converge uniformly to the unique solution p to the system

$$p'(t) \;=\; -\frac{\partial H}{\partial x}(t, z(t), p(t)), \quad p(T) \;=\; -\nabla g(z(T))$$

Thus, $p_1 = p(t_0) = p_2$. $\diamond$

Theorem 3.16 *We posit all hypotheses of Theorem 3.14 and we assume that g is continuously differentiable. Then $V(t_0, \cdot)$ is differentiable at x_0 with the derivative different from zero if and only if there exists a unique optimal solution z to problem (47) satisfying $\nabla g(z(T)) \neq 0$.*

Proof — Assume that $\frac{\partial V}{\partial x}(t_0, x_0) \neq 0$. Let z be optimal for problem (47). By Theorem 3.10, $\nabla g(z(T)) \neq 0$. By Proposition 3.11, every optimal state/co-state pair solves Hamiltonian system (74) with $p_0 = -\frac{\partial V}{\partial x}(t_0, x_0)$. This and Theorem 3.14 yield uniqueness of optimal solution.

Conversely, assume that (47) has a unique optimal solution z and $\nabla g(z(T)) \neq 0$. By Theorem 3.15, $V(t_0, \cdot)$ is differentiable at x_0. Theorem 3.10 implies that $\frac{\partial V}{\partial x}(t_0, x_0) \neq 0$. $\diamond$

3.4 Semiconcavity of Value Function

We provide next a sufficient condition for semi-concavity of the value function on $[0,T]\times\mathbf{R}^n$. Throughout the whole subsection we assume the following

$$\left\{\begin{array}{l} \exists\ \omega:\mathbf{R}_+\times\mathbf{R}_+\mapsto\mathbf{R}_+ \text{ such that (3) holds true and} \\ \\ \forall\ \lambda\in[0,1],\ R>0,\ x_0,x_1\in B_R(0),\ t\in[0,T],\ u\in U(t) \\ \|\lambda f(t,x_0,u)+(1-\lambda)f(t,x_1,u)-f(t,x_\lambda,u)\| \\ \leq\ \lambda(1-\lambda)\,\|x_1-x_0\|\,\omega(R,\|x_1-x_0\|), \\ \text{where } x_\lambda=\lambda x_0+(1-\lambda)x_1 \\ \\ g:\mathbf{R}^n\mapsto\mathbf{R} \text{ is semiconcave} \end{array}\right. \tag{77}$$

Remark — Assumptions (77) hold true in particular when g is continuously differentiable and f is continuously differentiable with respect to x uniformly in (t,u). More precisely, if we assume that there exists a function $\omega:\mathbf{R}_+\times\mathbf{R}_+\mapsto\mathbf{R}_+$ satisfying (3) such that

$$\left\|\frac{\partial f}{\partial x}(t,x_1,u)-\frac{\partial f}{\partial x}(t,x_2,u)\right\|\ \leq\ \omega(R,\|x_1-x_2\|)$$

for all $t\in[0,T],\ u\in U(t)$ and $x_1,x_2\in B_R(0)$.

2) Vice versa, Proposition 1.13 implies that, if f satisfies (77), then f is continuously differentiable with respect to x. $\diamond$

Theorem 3.17 *Assume (48), (49) and (77). Then there exists* $\overline{\omega}:\mathbf{R}_+\times\mathbf{R}_+\mapsto\mathbf{R}_+$ *satisfying (3) such that for all* $t\in[0,T],\ \lambda\in[0,1],\ R>0$

$$\forall x_0,x_1\in B_R(0),\ \lambda V(t,x_1)+(1-\lambda)V(t,x_0)-V(t,\lambda x_1+(1-\lambda)x_0)$$

$$\leq\ \lambda(1-\lambda)\,\|x_1\ -\ x_0\|\,\overline{\omega}(R,\|x_1\ -\ x_0\|)$$

Consequently for every $t\in[0,T],\ V(t,\cdot)$ *is semiconcave.*

Proof — For every $t\in[0,T]$ and control $u(s)\in U(s)$ (admissible control), we denote by $y(\cdot;t,x,u)$ the solution to the system

$$\left\{\begin{array}{rcl} y'(s) & = & f(s,y(s),u(s)),\ \ s\ \in\ [t,T] \\ \\ y(t) & = & x \end{array}\right.$$

By Theorem 3.1 for every $R > 0$ there exists L_R such that

$$\forall\, x \in B_R(0),\ \ \forall\, s \in [t,T],\ \ \|y(s;t,x,u)\| \ \leq\ L_R \tag{78}$$

Moreover, by the Gronwall lemma, for all $t \in [0,T],\ s \in [t,T],\ x_0, x_1 \in \mathbf{R}^n$ and all admissible control $u(\cdot)$, we have

$$\|y(s;t,x_1,u)\ -\ y(s;t,x_0,u)\| \ \leq\ e^{\int_t^s k(\tau)d\tau}\,\|x_1\ -\ x_0\| \tag{79}$$

Step 1. We claim that there exists $\omega_1 : \mathbf{R}_+ \times \mathbf{R}_+ \mapsto \mathbf{R}_+$ satisfying (3) such that for all $0 \leq t \leq s \leq T,\ R > 0,\ x_0, x_1 \in B_R(0),\ \lambda \in [0,1]$ and admissible control $u(\cdot)$, we have

$$\|\lambda y(s;t,x_1,u) + (1-\lambda)y(s;t,x_0,u) - y(s;t,\lambda x_0 + (1-\lambda)x_1,u)\|$$

$$\leq\ \lambda(1-\lambda)\,\|x_1 - x_0\|\,\omega_1(R, \|x_1 - x_0\|)$$

Indeed set $x_\lambda = \lambda x_0 + (1-\lambda)x_1$ and define

$$y_\lambda(\tau) = \lambda y(\tau;t,x_1,u) + (1-\lambda)y(\tau;t,x_0,u) - y(\tau;t,x_\lambda,u)$$

Then $y_\lambda(t) = 0$ and

$$\begin{aligned} y'_\lambda(\tau) \ &=\ \lambda f(\tau, y(\tau;t,x_1,u), u(\tau)) \ + \\ &+ (1-\lambda) f(\tau, y(\tau;t,x_0,u), u(\tau)) - f(\tau, y(\tau;t,x_\lambda,u), u(\tau)) \end{aligned}$$

From assumptions (49), (77) we obtain

$$\begin{aligned} \|y'_\lambda(\tau)\| \ &\leq k(\tau)\,\|y_\lambda(\tau)\| + \lambda(1-\lambda)\,\|y(\tau;t,x_1,u) - y(\tau;t,x_0,u)\| \times \\ &\times\, \omega(L_R, \|y(\tau;t,x_1,u) - y(\tau;t,x_0,u)\|) \end{aligned}$$

Our claim results from (79) and the Gronwall lemma.

Step 2. Fix $\varepsilon > 0$ and a control u_ε such that

$$V(t,x_\lambda) \ >\ g(y(T;t,x_\lambda,u_\varepsilon)) - \varepsilon$$

Let ω_g denote a modulus of semiconcavity of g and C_R a Lipschitz constant

of g on the ball of radius L_R. Then from (79) and Step 1,

$$\lambda V(t,x_1) \ + \ (1-\lambda)V(t,x_0) \ - \ V(t,x_\lambda) \ <$$

$$\lambda g(y(T;t,x_1,u_\varepsilon)) + (1-\lambda)g(y(T;t,x_0,u_\varepsilon)) - g(y(T;t,x_\lambda,u_\varepsilon)) + \varepsilon$$

$$\leq \lambda(1-\lambda)\,\|y(T;t,x_1,u_\varepsilon) - y(T;t,x_0,u_\varepsilon)\| \times$$

$$\times\, \omega_g(L_R, \|y(T;t,x_1,u_\varepsilon) - y(T;t,x_0,u_\varepsilon)\|)$$

$$+ C_R\,\|\lambda y(T;t,x_1,u_\varepsilon) + (1-\lambda)y(T;t,x_0,u_\varepsilon) - y(T;t,x_\lambda,u_\varepsilon)\| + \varepsilon$$

$$\leq e^{\int_t^T k(s)ds}\lambda(1-\lambda)\,\|x_1-x_0\|\,\omega_g\left(L_R,\ e^{\int_t^T k(s)ds}\,\|x_1-x_0\|\right)$$

$$+ C_R\lambda(1-\lambda)\,\|x_1-x_0\|\ \omega_1(R,\|x_1-x_0\|) \ + \ \varepsilon$$

Since $\varepsilon > 0$ is arbitrary the proof follows. $\diamond$

If we impose some additional assumptions on f, then V is semiconcave also with respect to t.

Theorem 3.18 *Assume (48), (49) and (77), that k and U are* time independent *and for every $R > 0$, there exists $k_R > 0$ such that for all $x \in B_R(0)$ and $u \in U$, $f(\cdot,x,u)$ is k_R-Lipschitz.*

Further assume that for all $R > 0$, there exists $c_R \geq 0$ such that

$$\forall\,(t,u) \in [0,T]\times U,\ \ \forall\, x \in B_R(0),\ \ \|f(t,x,u)\| \ \leq \ c_R \tag{80}$$

Then the value function is semi-concave on $[0,T]\times \mathbf{R}^n$.

Proof — Consider $0 \leq t_1 < t_0 \leq T,\ R > 0$ and let $x_0, x_1 \in B_R(0),\ \lambda \in [0,1]$. Define

$$x_\lambda \ = \ \lambda x_1 + (1-\lambda)x_0, \quad t_\lambda \ = \ \lambda t_1 + (1-\lambda)t_0$$

and pick any $\varepsilon > 0$. By (55) there exists an admissible control u_ε such that

$$V(t_0, y(t_0;t_\lambda,x_\lambda,u_\varepsilon)) \ < \ V(t_\lambda,x_\lambda) \ + \ \varepsilon$$

Define

$$\tau(s) \ = \ \begin{cases} \lambda s \ + \ (1-\lambda)t_0, & \text{if } t_1 \ \leq \ s \ \leq \ t_0 \\ s & \text{otherwise} \end{cases} \tag{81}$$

Since the value function is non decreasing along solutions to our control system, we have

$$\lambda V(t_1, x_1) + (1-\lambda)V(t_0, x_0) - V(t_\lambda, x_\lambda) \;\le\; (1-\lambda)V(t_0, x_0) +$$

$$\lambda V(t_0, y(t_0; t_1, x_1, u_\varepsilon \circ \tau)) - V(t_0, y(t_0; t_\lambda, x_\lambda, u_\varepsilon)) + \varepsilon$$

Define L_R as in the proof of Theorem 3.17 and set

$$y_1(s) \;=\; y(s; t_1, x_1, u_\varepsilon \circ \tau), \quad y_\lambda(s) \;=\; y(s; t_\lambda, x_\lambda, u_\varepsilon)$$

Let K_R denote the Lipschitz constant of V on $[0,T] \times L_R B$ (which exists by Theorem 3.1.) From Theorem 3.17 and the latter inequality we obtain

$$\begin{aligned} &\lambda V(t_1, x_1) \;+\; (1-\lambda)V(t_0, x_0) \;-\; V(t_\lambda, x_\lambda) \\ &\le \lambda(1-\lambda)\,\|y_1(t_0) - x_0\|\,\overline{\omega}(L_R, \|y_1(t_0) - x_0\|) \qquad (82) \\ &+ K_R\,\|\lambda y_1(t_0) \;+\; (1-\lambda)x_0 \;-\; y_\lambda(t_0)\| + \varepsilon \end{aligned}$$

On the other hand from assumption (80) it follows that

$$\forall\, s \in [t_1, t_0], \quad \|y_1(s) - x_0\| \;\le\; \|x_1 - x_0\| + M_R(t_0 - t_1) \qquad (83)$$

where $M_R = c_{L_R}$. Set

$$z(s) \;=\; \lambda y_1\left(\tau^{-1}(s)\right) \;+\; (1-\lambda)x_0 \;-\; y_\lambda(s)$$

and notice that

$$z(t_\lambda) \;=\; 0, \quad z(t_0) \;=\; \lambda y_1(t_0) + (1-\lambda)x_0 - y_\lambda(t_0)$$

On the other hand, by assumptions of theorem there exists $L > 0$ such that for every $t \in [0,T]$ and $u \in U$, $f(t, \cdot, u)$ is L-Lipschitz on $\mathbf{R}^n$. Hence, using (77), we obtain the following estimates

$$\begin{aligned} &\|z'(s)\| \;=\; \left\|f\left(\tau^{-1}(s), y_1 \circ \tau^{-1}(s), u_\varepsilon(s)\right) \;-\; f(s, y_\lambda(s), u_\varepsilon(s))\right\| \\ &\le\; C_R|\tau^{-1}(s) - s| + L\left\|y_1 \circ \tau^{-1}(s) \;-\; y_\lambda(s)\right\| \\ &\le\; L\,\|z(s)\| \;+\; L(1-\lambda)\left\|y_1 \circ \tau^{-1}(s) \;-\; x_0\right\| \;+\; C_R\,\tfrac{1-\lambda}{\lambda}(t_0 - s) \end{aligned}$$

where $C_R := k_{L_R}$. Therefore from the Gronwall inequality and (83) we deduce that for some $c > 0$ depending only on L and c_R

$$\begin{cases} \|z(t_0)\| \leq c(1-\lambda)\int_{t_\lambda}^{t_0} (\|y_1 \circ \tau^{-1}(s) - x_0\| + \frac{t_0 - s}{\lambda})\, ds \\ \\ \leq c\lambda(1-\lambda)(t_0 - t_1)\left(\|x_1 - x_0\| + (\frac{1}{2} + M_R)(t_0 - t_1)\right) \end{cases} \tag{84}$$

Since $\varepsilon > 0$ is arbitrary, inequalities (82), (83), (84) imply the conclusion.

◇

3.4.1 Differentiability along Optimal Solutions

In this subsection we provide further results concerning differentiability of V along optimal solutions.

Theorem 3.19 *Under all assumptions of Theorems 3.15 and 3.18, suppose that problem (47) has a unique optimal solution z. Then V is differentiable at $(t, z(t))$ for all $t \in [t_0, T]$.*

The proof of this statement is left as an exercise.

Theorem 3.16 yields

Corollary 3.20 *Under hypotheses of Theorems 3.14 and 3.18, assume that g is continuously differentiable. Then $V(\cdot,\cdot)$ is differentiable at (t_0, x_0) with the partial derivative $\frac{\partial V}{\partial x}(t_0, x_0)$ different from zero if and only if there is a unique optimal solution z to problem (47) satisfying $\nabla g(z(T)) \neq 0$.*

Usually the value function is not everywhere differentiable. However this is always the case for "convex" problems, as we prove below.

Theorem 3.21 *Assume (48), (49), (77), that g is convex and*

$$\forall\, t \in [0,T], \quad \text{Graph}(f(t,\cdot,U(t))) \text{ is closed and convex} \tag{85}$$

Then $V(t,\cdot)$ is convex and continuously differentiable on $\mathbf{R}^n$.

Moreover, if all assumptions of Theorem 3.18 are verified, then V is continuously differentiable on $[0,T] \times \mathbf{R}^n$.

Proof — Assumptions (85) and (49) yield that for every $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$ there exists a solution z to control system

$$x' = f(t, x(t), u(t)), \quad u(t) \in U(t), \quad x(t_0) = x_0$$

satisfying $V(t_0, x_0) = g(z(T))$.

Fix $t_0 \in [0,T]$, $x_0, x_1 \in \mathbf{R}^n$, $\lambda \in [0,1]$ and consider solutions $x : [t_0, T] \mapsto \mathbf{R}^n$ and $y : [t_0, T] \mapsto \mathbf{R}^n$ to (45) such that

$$V(t_0, x_0) \;=\; g(x(T)), \quad V(t_0, x_1) \;=\; g(y(T))$$

Define the map $z : [t_0, T] \mapsto \mathbf{R}^n$ by $z(t) = \lambda x(t) + (1-\lambda) y(t)$. From (85), we deduce that z is a solution to control system (45) satisfying $z(t_0) = \lambda x_0 + (1-\lambda) x_1$. Thus, from convexity of g,

$$V(t_0, \lambda x_0 + (1-\lambda) x_1) \leq g(z(T)) \;\leq\; \lambda V(t_0, x_0) + (1-\lambda) V(t_0, x_1)$$

and therefore $V(t_0, \cdot)$ is convex.

Next, as $V(t, \cdot)$ is both convex and semiconcave for all $t \in [0,T]$, Proposition 1.13 yields that $V(t, \cdot)$ is continuously differentiable on $\mathbf{R}^n$. The last statement follows from Proposition 1.12. $\diamond$

3.4.2 Regularity of Optimal Feedback

One of the major issues of optimal control theory is to find an "equation" for optimal solutions. Theorem 3.3 provides an inclusion formulation. However, in general, the set-valued map G is not regular enough to make us able to approximate solutions to (59) using, say, Euler's scheme. This is one of the reasons why we have to investigate regularity of the set-valued map G.

Theorem 3.22 *Under all assumptions of Theorem 3.18, suppose that the sets $f(t,x,U)$ are closed. Then G has compact nonempty images and is upper semicontinuous on $[0,T[\times\mathbf{R}^n$.*

Proof — From Theorems 3.18 and 1.11 we know that for every $(t,x) \in [0,T[\times\mathbf{R}^n$ and every $v \in \mathbf{R}^n$ the directional derivative $\frac{\partial V}{\partial (1,v)}(t,x)$ exists. Define the set-valued map

$$\widehat{Q} : [0,T[\times\mathbf{R}^n \rightsquigarrow \mathbf{R}^n$$

by: for every $(t,x) \in [0,T[\times\mathbf{R}^n$, $\widehat{Q}(t,x)$ is equal to

$$\{v \in \mathbf{R}^n \mid \liminf_{\substack{x' \to x,\, h \to 0+ \\ t' \to t,\; t' \geq 0}} \frac{V(t'+h,\; x'+hv) - V(t',x')}{h} \;\leq\; 0\}$$

From Proposition 1.14 follows that $\mathrm{Graph}(\widehat{Q})$ is closed in $[0,T[\times\mathbf{R}^n \times\mathbf{R}^n$. By Proposition 3.2, for all $v \in \overline{co}(f(t,x,U))$, $\frac{\partial V}{\partial(1,v)}(t,x) \geq 0$. Thus

$$G(t,x) \;=\; \widehat{Q}(t,x) \cap f(t,x,U)$$

Consequently, $\mathrm{Graph}(G)$ is closed in $[0,T[\times\mathbf{R}^n \times \mathbf{R}^n$. From Proposition 1.18 we deduce that G is upper semicontinuous on $[0,T[\times\mathbf{R}^n$. $\diamond$

Corollary 3.23 *Under all assumptions of Theorem 3.18 suppose that the sets $f(t,x,U)$ are closed. If G is single-valued on a subset $K \subset [0,T[\times\mathbf{R}^n$, then the map $K \ni (t,x) \mapsto G(t,x)$ is continuous.*

Theorem 3.24 *We posit all assumptions of Theorems 3.18, 3.21 and suppose that g is convex. Then G has convex compact images and is upper semicontinuous. Furthermore, if for every (t,x) the set $f(t,x,U)$ is strictly convex, then G is single valued and continuous on the set*

$$\left\{(t,x) \;\in\; [0,T[\times\mathbf{R}^n \;|\; \frac{\partial V}{\partial x}(t,x) \;\neq\; 0\right\}$$

Proof — By Theorem 3.21, we know that V is continuously differentiable. This and convexity of $f(t,x,U)$ yield that for all $(t,x) \in [0,T[\times\mathbf{R}^n$

$$G(t,x) \;=\; f(t,x,U) \cap \{v \in \mathbf{R}^n \;|\; \langle \nabla V(t,x), (1,v)\rangle \;=\; 0\}$$

is convex. Theorem 3.22 ends the proof of the first statement. From Proposition 3.2 it follows that for all $(t,x) \in [0,T[\times\mathbf{R}^n$

$$\left\{\begin{array}{l} v \;\in\; G(t,x) \iff \\ v \in f(t,x,U) \;\&\; \sup_{u\in U}\left\langle -\frac{\partial V}{\partial x}(t,x),\; f(t,x,u)\right\rangle = \left\langle -\frac{\partial V}{\partial x}(t,x),\; v\right\rangle \end{array}\right.$$

This and strict convexity of $f(t,x,U)$ imply that G is single valued on

$$\left\{(t,x) \in [0,T[\times\mathbf{R}^n \;|\; \frac{\partial V}{\partial x}(t,x) \neq 0\right\}$$

Corollary 3.23 completes the proof. $\diamond$

4 Hamilton-Jacobi-Bellman Equation

Consider the Hamilton-Jacobi-Bellman equation:

$$-\frac{\partial V}{\partial t}(t,x) \;+\; H\left(t,x,-\frac{\partial V}{\partial x}(t,x)\right) \;=\; 0, \quad V(T,\cdot) \;=\; g_K(\cdot) \tag{86}$$

associated to the Mayer problem with end point constraints:

$$\text{minimize } \{\; g(x(T)) \mid \; x(T) \in K \;\}$$

over all solutions to control system

$$x' \;=\; f(t,x,u(t)), \quad u(t) \in U \tag{87}$$

satisfying the initial condition $x(0) = \xi_0$, where K is a given subset (called target) and g_K denotes the restriction of g to K.

In the above equation (86), the Hamiltonian H is given by:

$$H(t,x,p) \;=\; \sup_{u\in U} \langle p, f(t,x,u)\rangle$$

The value function for the constrained Mayer problem is defined by

$$V(t_0,x_0) = \inf\{g(x(T)) \mid \; x \text{ solves (87)}, \; x(t_0) = x_0, \; x(T) \in K\}$$

In general V is merely lower semicontinuous and is equal to $+\infty$ at all points from which it is impossible to reach the target K. In fact one can even avoid using the target K in the definition of minimization problem by setting $g(x) = +\infty$ whenever $x \notin K$.

The value function is non decreasing along solutions to (87) and is constant along optimal solutions. These two properties and the final value $V(T,\cdot) = g(\cdot)$ characterize the value function.

There have been several concepts of "generalized" solutions to Hamilton-Jacobi equation (86): *viscosity solutions* [18], *contingent solutions* [50] and [27, 28], *lower semicontinuous solutions* [7] and [31, 32] . Under quite general assumptions we shall prove that all these concepts of solutions are equivalent and that the value function is the unique solution.

The outline is as follows: In Subsection 1 we state several characterizations of the value function, which are proved in the subsequent subsections. Subsection 2 is devoted to equivalence between contingent solutions and semicontinuous solutions and to the monotone behavior of contingent solutions. Then we show that the value function is the only solution to (86). A comparison to continuous viscosity solutions is provided in Subsection 3.

4.1 Solutions to Hamilton-Jacobi Equation

Consider $T > 0$, a complete separable metric space U and a map $f : [0,T] \times \mathbf{R}^n \times U \mapsto \mathbf{R}^n$. We associate with it the control system

$$x'(t) \;=\; f(t, x(t), u(t)), \quad u(t) \;\in\; U \tag{88}$$

Let an extended function $g : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and $\xi_0 \in \mathbf{R}^n$ be given. Consider the minimization problem (*Mayer's problem*):

$$\min \{g(x(T)) \mid \; x \text{ is a solution to } (88),\; x(0) = \xi_0\} \tag{89}$$

The value function $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ associated with it is defined by: for all $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$

$$V(t_0, x_0) = \inf\{g(x(T)) \mid x \text{ is a solution to } (88),\; x(t_0) = x_0\}$$

We impose the following assumptions

$$\begin{cases} i) & \forall\, R > 0,\; \exists\, c_R \in L^1(0,T) \text{ such that for almost all } t \\ & \forall\, u \in U,\; f(t,\cdot,u) \text{ is } c_R(t) - \text{Lipschitz on } B_R(0) \\ ii) & \exists\, k \in L^1(0,T) \text{ such that for almost all } t \in [0,T], \\ & \forall\, x \in \mathbf{R}^n,\; \sup_{u\in U} \|f(t,x,u)\| \;\le\; k(t)(1 + \|x\|) \\ iii) & \forall\, (t,x) \in [0,T] \times \mathbf{R}^n,\; f(t,x,U) \text{ is convex, compact} \\ iv) & f \text{ is continuous and for all } (t,x) \in [0,T] \times \mathbf{R}^n, \\ & \lim_{(t',x')\to(t,x)} \sup_{u\in U} \|f(t,x,u) - f(t',x',u)\| = 0 \\ v) & \text{g is lower semicontinuous} \end{cases} \tag{90}$$

By these assumptions the control system (88) may be replaced by the differential inclusion

$$x'(t) \;\in\; F(t, x(t)) \;\text{ almost everywhere} \tag{91}$$

where $F(t,x) = f(t,x,U)$. Furthermore, F satisfies the following conditions:

$$\begin{cases} a) & F \text{ is continuous and has nonempty convex compact images} \\ b) & \exists\, k \in L^1(0,T) \text{ such that for almost all } t \in [0,T], \\ & \forall\, x \in \mathbf{R}^n,\ \sup_{v \in F(t,x)} \|v\| \le k(t)(1+\|x\|) \\ c) & \forall\, R > 0,\ \exists\, c_R \in L^1(0,T) \text{ such that for a.e. } t \in [0,T] \\ & F(t,\cdot) \text{ is } c_R(t)\text{-Lipschitz on } B_R(0) \end{cases} \tag{92}$$

Recall that $\mathcal{S}_{[t_0,T]}(x_0)$ denotes the set of absolutely continuous solutions to (91) defined on $[t_0,T]$ and satisfying the initial condition $x(t_0) = x_0$.

We show next that

$$V(t_0,x_0) \;=\; \min\Big\{\, g(x(T)) \mid\ x \in \mathcal{S}_{[t_0,T]}(x_0) \,\Big\} \tag{93}$$

Theorem 4.1 *Consider an extended lower semicontinuous function $g : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ and assume (90).*

Then V takes its values in $\mathbf{R} \cup \{+\infty\}$ and is lower semicontinuous.

This Theorem follows from

Theorem 4.2 *Consider a set-valued map $F : [0,T] \times \mathbf{R}^n \hookrightarrow \mathbf{R}^n$ and an extended lower semicontinuous function $g : \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$.*

Assume (92) and define $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{\pm\infty\}$ by

$$V(t_0,x_0) \;=\; \inf\{g(x(T)) \mid x \text{ solves (91)},\ x(t_0) = x_0\}$$

Then V is lower semicontinuous taking its values in $\mathbf{R} \cup \{+\infty\}$ and (93) holds true.

Proof — Fix $(t_0,x_0) \in [0,T] \times \mathbf{R}^n$. Since the set of solutions $\mathcal{S}_{[t_0,T]}(x_0)$ is compact and g is lower semicontinuous we deduce that $V(t_0,x_0) > -\infty$ and (93). To prove that V is lower semicontinuous consider a sequence $(t_n,x_0^n) \to (t_0,x_0)$ such that $t_n \in [0,T]$

$$\liminf_{(t,x)\to(t_0,x_0),\ t\in[0,T]} V(t,x) \;=\; \lim_{n\to\infty} V(t_n,x_0^n)$$

and let $x_n \in \mathcal{S}_{[t_n,T]}(x_0^n)$ be such that

$$g(x_n(T)) \;\le\; V(t_n,x_0^n) \;+\; \frac{1}{n}$$

For every $n \geq 1$ such that $t_n > t_0$ consider a solution y_n to inclusion

$$x'(s) \ \in \ -F(t_n - s, x(s)), \ \ x(0) \ = \ x_0^n$$

defined on $[0, t_n - t_0]$. Set

$$z_n(t) \ = \ \begin{cases} x_n(t) & \text{if } t \ \geq \ t_n \\ y_n(t_n - t) & \text{otherwise} \end{cases}$$

Then $z_n \in \mathcal{S}_{[t_0,T]}(y_n(t_n - t_0))$. We also observe that (92) b) yields that $y_n(t_n - t_0)$ converge to x_0. By Fillipov's theorem there exist $\bar{z}_n \in \mathcal{S}_{[t_0,T]}(x_0)$ such that $z_n - \bar{z}_n$ converge uniformly to zero. The set $\mathcal{S}_{[t_0,T]}(x_0)$ being compact in the space of continuous functions, there exists a subsequence $\bar{z}_{n_k}$ converging to some $z \in \mathcal{S}_{[t_0,T]}(x_0)$. But then also z_{n_k} converge to z. Since g is lower semicontinuous, $g(z(T)) \leq \lim_{n\to\infty} V(t_n, x_0^n)$. On the other hand, $V(t_0, x_0) \leq g(z(T))$ and the proof follows. $\diamond$

The Hamiltonian associated to control system (88) is the function $H : [0,T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ defined by

$$H(t, x, p) \ = \ \max_{u \in U} \langle p, f(t, x, u) \rangle$$

Consider the Hamilton-Jacobi equation

$$-\frac{\partial V}{\partial t}(t,x) + H\left(t, \ x, \ -\frac{\partial V}{\partial x}(t,x)\right) = 0, \quad V(T, \cdot) = g(\cdot) \tag{94}$$

In the result stated below we use notions of super/subdifferentials and epi/hypo-derivatives introduced in Section 1.

Definition 4.3 *An extended lower semicontinuous function $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ is called a* lower semicontinuous solution *to (94) if it satisfies the following conditions:*

$$\begin{cases} V(T, \cdot) \ = \ g(\cdot) \ \text{ and for all } \ (t,x) \in]0,T[\times \mathbf{R}^n, \\ \forall \ (p_t, p_x) \in \partial_- V(t,x), \quad -p_t + H(t, x, -p_x) = 0 \\ \forall \ (p_t, p_x) \in \partial_- V(0,x), \quad -p_t + H(0, x, -p_x) \geq 0 \\ \forall \ (p_t, p_x) \in \partial_- V(T,x), \quad -p_t + H(T, x, -p_x) \leq 0 \end{cases}$$

Definition 4.4 *An extended lower semicontinuous function $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ is called a* viscosity supersolution *to (94) if for all $t \in]0,T[$ and $x \in \mathbf{R}^n$ such that $(t,x) \in \mathrm{Dom}(V)$ we have*

$$\forall \ (p_t, p_x) \ \in \ \partial_- V(t,x), \ \ -p_t \ + \ H(t, x, -p_x) \ \geq \ 0$$

An extended upper semicontinuous function $V : [0,T] \times \mathbf{R}^n \to \mathbf{R} \cup \{-\infty\}$ *is called a* viscosity subsolution *to (94) if for all* $0 < t < T$ *and* $x \in \mathbf{R}^n$ *such that* $(t,x) \in \mathrm{Dom}(V)$ *we have*

$$\forall\, (p_t, p_x) \ \in\ \partial_+ V(t,x), \quad -p_t \ +\ H(t,x,-p_x) \ \leq\ 0$$

Let $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R}$ *be a continuous function. It is called a* viscosity solution *to (94) if for all* $t \in]0,T[$ *and* $x \in \mathbf{R}^n$

$$\begin{array}{lll} \forall\, (p_t, p_x) & \in\ \partial_- V(t,x), & -p_t \ +\ H(t,x,-p_x) \ \geq\ 0 \\ \forall\, (p_t, p_x) & \in\ \partial_+ V(t,x), & -p_t \ +\ H(t,x,-p_x) \ \leq\ 0 \end{array}$$

Theorem 4.5 *Assume (92) and let* $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ *be an extended lower semicontinuous function.*

Then the following four statements are equivalent:

i) V is the value function given by (93)

ii) V is a lower semicontinuous solution to (94)

iii) V is a contingent solution to (94) in the sense that
$V(T,\cdot) = g(\cdot)$ and for all $(t,x) \in \mathrm{Dom}(V)$,
$0 \ \leq\ t \ <\ T \implies \inf_{v \in F(t,x)} D_\uparrow V(t,x)(1,v) \leq 0$
$0 \ <\ t \ \leq\ T \implies \sup_{v \in F(t,x)} D_\uparrow V(t,x)(-1,-v) \leq 0$

iv) $V(T,\cdot) \ =\ g(\cdot)$ and for all $(t,x) \in]0,T[\times \mathbf{R}^n$,
$\forall\, (p_t, p_x) \in \partial_- V(t,x), \ \ -p_t + H(t,x,-p_x) = 0$
$\forall\, \overline{x} \in \mathbf{R}^n, \ \ V(0,\overline{x}) = \liminf_{t \to 0+,\ x \to \overline{x}} V(t,x)$
$\forall\, \overline{x} \in \mathbf{R}^n, \ \ g(\overline{x}) = \liminf_{t \to T-,\ x \to \overline{x}} V(t,x)$

Finally, if V *is continuous on* $[0,T] \times \mathbf{R}^n$ *then the above statements are equivalent to:*

v) V *is a viscosity solution to (94).*

4.2 Lower Semicontinuous Solutions

4.2.1 Lower Semicontinuous & Contingent Solutions

The equivalence between statements *ii)* and *iii)* of Theorem 4.5 follows from Theorems 4.6 and 4.7 proved below.

Let $T > 0$. Consider a set-valued map $F : [0,T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ with nonempty bounded images and define the Hamiltonian $H : [0,T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ by

$$H(t,x,p) \;=\; \sup_{v \in F(t,x)} < p, v > \tag{95}$$

Then $H(t,x,\cdot)$ is convex and positively homogeneous. Furthermore, if F is continuous, then so is H.

Theorem 4.6 *Consider an extended lower semicontinuous function* $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$. *Assume that* F *is upper semicontinuous and has nonempty convex compact images on* $\mathrm{Dom}(V)$.

Then the following four statements are equivalent :

i) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t < T$ *and for every* $(p_t, p_x, q) \in N^0_{\mathcal{E}p(V)}(t,x,V(t,x))$

$$-p_t + H(t,x,-p_x) \;\geq\; 0 \tag{96}$$

ii) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t < T$ *and all* $y \geq V(t,x)$

$$(\{1\} \times F(t,x) \times \{0\}) \cap T_{\mathcal{E}p(V)}(t,x,y) \;\neq\; \emptyset$$

iii) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t < T$

$$\inf_{v \in F(t,x)} D_{\uparrow} V(t,x)(1,v) \leq 0$$

iv) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t < T$

$$\forall\, (p_t, p_x) \;\in\; \partial_- V(t,x), \quad -p_t + H(t,x,-p_x) \geq 0$$

Proof — Fix $(t,x) \in \mathrm{Dom}(V)$ such that $t < T$ and observe that

$$\forall\, y \;\geq\; V(t,x), \quad T_{\mathcal{E}p(V)}(t,x,V(t,x)) \;\subset\; T_{\mathcal{E}p(V)}(t,x,y) \tag{97}$$

Since the contingent cone to the epigraph is the epigraph of the epiderivative, *ii*) is equivalent to *iii*).

Assume that *iii*) holds true. Fix $(t,x) \in \mathrm{Dom}(V)$ such that $t < T$ and $(p_t, p_x, q) \in [T_{\mathcal{E}p(V)}(t,x,V(t,x))]^-$. Consider $v \in F(t,x)$ satisfying

$$D_{\uparrow} V(t,x)(1,v) \leq 0$$

or, equivalently, $(1,v,0) \in T_{\mathcal{E}p(V)}(t,x,V(t,x))$. Hence $p_t + \langle p_x, v \rangle \leq 0$ and *i*) follows.

Assume next that $i)$ holds true. We claim that

$$(\{1\} \times F(t,x) \times \{0\}) \cap \overline{co}\left(T_{\mathcal{E}p(V)}(t,x,y)\right) \neq \emptyset \tag{98}$$

for all $(t,x) \in \mathrm{Dom}(V)$ such that $t < T$ and $y \geq V(t,x)$.

By (97) it is enough to prove (98) with $y = V(t,x)$. Otherwise, by the separation theorem, there exists

$$(p_t, p_x, q) \in N^0_{\mathcal{E}p(V)}(t,x,V(t,x))$$

such that

$$p_t + \inf_{v \in F(t,x)} \langle p_x, v \rangle > 0$$

Consequently $-p_t + H(t,x,-p_x) < 0$, which contradicts $i)$ and proves (98).

Finally, since F is upper semicontinuous and has convex compact images, Proposition 1.43 and relation (98) imply $ii)$.

By Proposition 1.15, $i)$ yields $iv)$.

Assume next that $iv)$ is verified. Fix $(t,x) \in \mathrm{Dom}(V)$ such that $t < T$ and $(p_t, p_x, q) \in N^0_{\mathcal{E}p(V)}(t,x,V(t,x))$. Since

$$\{0\} \times \{0\} \times \mathbf{R}_+ \subset T_{\mathcal{E}p(V)}(t,x,V(t,x))$$

we have $q \leq 0$. If $q < 0$, then

$$\left(\frac{p_t}{|q|}, \frac{p_x}{|q|}, -1\right) \in N^0_{\mathcal{E}p(V)}(t,x,V(t,x))$$

From Proposition 1.15 and $iv)$ we deduce that

$$-\frac{p_t}{|q|} + H\left(t, x, -\frac{p_x}{|q|}\right) \geq 0$$

Multiplying by $|q|$ the above relations we derive (96).

It remains to consider the case $q = 0$ and $(p_t, p_x) \neq 0$. For this aim it is enough to apply Lemma 1.16 and to use the same arguments as above. $\diamond$

Theorem 4.7 *Consider an extended lower semicontinuous function* $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ *and assume that* F *is lower semicontinuous and has nonempty compact images on* $\mathrm{Dom}(V)$.

Then the following four statements are equivalent :

i) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t > 0$ *and for every*

$(p_t, p_x, q) \in N^0_{\mathcal{E}p(V)}(t, x, V(t,x))$

$$-p_t + H(t, x, -p_x) \ \leq \ 0$$

ii) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t > 0$ *and all* $y \geq V(t,x)$

$$\{-1\} \times (-F(t,x)) \times \{0\} \ \subset \ T_{\mathcal{E}p(V)}(t,x,y)$$

iii) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t > 0$

$$\sup_{v \in F(t,x)} D_{\uparrow} V(t,x)(-1,-v) \leq 0$$

iv) For all $(t,x) \in \mathrm{Dom}(V)$ *such that* $t > 0$

$$\forall\, (p_t, p_x) \ \in \ \partial_- V(t,x), \quad -p_t + H(t,x,-p_x) \leq 0$$

Proof — We deduce using (97) that *ii)* is equivalent to *iii)*. Clearly, *ii)* yields *i)*. We next claim that *i)* implies that

$$\{-1\} \times (-F(t,x)) \times \{0\} \ \subset \ \overline{co}\left(T_{\mathcal{E}p(V)}(t,x,y)\right) \tag{99}$$

for all $(t,x) \in \mathrm{Dom}(V)$ such that $t > 0$ and $y \geq V(t,x)$.

Indeed, by (97) it is enough to consider the case $y = V(t,x)$. If (99) is not satisfied then, by the separation theorem, there exist $v \in F(t,x)$ and $(p_t, p_x, q) \in N^0_{\mathcal{E}p(V)}(t,x,V(t,x))$ such that $-p_t + \langle -p_x, v\rangle > 0$. Consequently $-p_t + H(t,x,-p_x) > 0$, which contradicts *i)* and so inclusion (99) follows.

Since F is lower semicontinuous, (99) and Theorem 1.3 yield that for all $(t,x) \in \mathrm{Dom}(V)$ such that $t > 0$ and all $y \geq V(t,x)$,

$$\begin{aligned} &\{-1\} \times (-F(t,x)) \times \{0\} \\ &\subset \mathrm{Liminf}_{(t',x',y') \to \mathcal{E}p(V)(t,x,y)} \ \overline{co}\left(T_{\mathcal{E}p(V)}(t',x',y')\right) \subset T_{\mathcal{E}p(V)}(t,x,y) \end{aligned}$$

Arguments similar to those of the proof of Theorem 4.6 yield that *i)* is equivalent to *iv)*.

4.2.2 Monotone Behavior of Contingent Solutions

Consider a set-valued map $F : [0,T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and the differential inclusion

$$x'(t) \ \in \ F(t, x(t)) \ \text{ almost everywhere} \tag{100}$$

In this subsection we investigate a relationship between monotone behavior of a function V along solutions to (100) and contingent inequalities *iii)* of Theorem 4.5.

Theorem 4.8 *Let $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous function. Assume that F is upper semicontinuous, that $F(t,x)$ is nonempty convex and compact for all $(t,x) \in \mathrm{Dom}(V)$ and that for some $k \in L^1(0,T)$*

$$\forall\ (t,x)\ \in\ \mathrm{Dom}(V),\quad \sup_{v \in F(t,x)} \|v\| \leq k(t)(1+\|x\|)$$

Then the following two statements are equivalent

$$i)\ \forall\ (t,x) \in \mathrm{Dom}(V)\ \text{ with }\ t < T,\quad \inf_{v \in F(t,x)} D_{\uparrow} V(t,x)(1,v) \leq 0$$

ii) For every $(t_0,x_0) \in [0,T] \times \mathbf{R}^n$, there exists $\overline{x} \in \mathcal{S}_{[t_0,T]}(x_0)$ such that $V(t,\overline{x}(t)) \leq V(t_0,x_0)$ for all $t \in [t_0,T]$.

Proof — Assume that $i)$ holds true and fix $(t_0,x_0) \in \mathrm{Dom}(V)$. Define the upper semicontinuous set-valued map

$$\widehat{F} : \mathbf{R}_+ \times \mathbf{R}^n \times \mathbf{R}\ \hookrightarrow\ \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}$$

by

$$\widehat{F}(t,x,z)\ =\ \begin{cases} \{1\} \times F(t,x) \times \{0\} & \text{when } t\ <\ T \\ [0,1] \times \overline{co}\,(F(T,x) \cup \{0\}) \times \{0\} & \text{when } t\ \geq\ T \end{cases}$$

and consider the viability problem

$$\begin{cases} (t,x,z)' & \in & \widehat{F}(t,x,z) \\ (t,x,z)(t_0) & = & (t_0,x_0,V(t_0,x_0)) \\ (t,x,z) & \in & \mathcal{E}p(V) \end{cases} \tag{101}$$

By Theorem 4.6, for all $(t,x,z) \in \mathcal{E}p(V)$ we have

$$\widehat{F}(t,x,z) \cap T_{\mathcal{E}p(V)}(t,x,z)\ \neq\ \emptyset$$

Since $\widehat{F}$ is upper semicontinuous and has convex compact nonempty images and linear growth on the closed set $\mathcal{E}p(V)$, the Viability Theorem 1.42 yields that problem (101) has a solution

$$[t_0,T]\ \ni\ t\ \mapsto\ (t,\overline{x}(t),z(t))\ \in\ \mathcal{E}p(V)$$

Thus $V(t,\overline{x}(t)) \leq z(t) = V(t_0,x_0)$ for all $t \in [t_0,T]$ and $ii)$ follows.

Conversely, assume that *ii)* is satisfied. Fix $(t_0, x_0) \in \mathrm{Dom}(V)$ with $t_0 < T$ and let $\overline{x}$ be as in *ii)*. Since F is bounded on a neighborhood of (t_0, x_0), we deduce that $\overline{x}(\cdot)$ is Lipschitz at t_0. Let $h_n \to 0+$ be such that $[x(t_0 + h_n) - x(t_0)]/h_n$ converge to some v. Theorem 1.39 yields that $v \in F(t_0, x_0)$. On the other hand

$$D_{\uparrow}V(t_0, x_0)(1, v) \leq \liminf_{n\to\infty} \frac{V(t_0 + h_n, x(t_0 + h_n)) - V(t_0, x_0)}{h_n} \leq 0 \quad \diamond$$

Theorem 4.9 *Let $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semi-continuous function. If F verifies (92), then the following two statements are equivalent:*

$$i) \ \forall\, (t, x) \in \mathrm{Dom}(V) \ \text{ with } \ t > 0, \quad \sup_{v\in F(t,x)} D_{\uparrow}V(t, x)(-1, -v) \leq 0$$

ii) For every $x \in \mathcal{S}_{[t_0,T]}(x_0)$ and all $t \in [t_0, T]$, $V(t_0, x_0) \leq V(t, x(t))$.

Proof — Assume that *i)* holds true. Since *i)* does not involve T, it is enough to prove the inequality in *ii)* for $t = T$. By Theorem 4.7, for all $0 \leq t < T$ and $x \in \mathbf{R}^n$ such that $(T - t, x) \in \mathrm{Dom}(V)$ and for all $z \geq V(T - t, x)$,

$$\{-1\} \times (-F(T - t, x)) \times \{0\} \ \subset \ T_{\mathcal{E}p(V)}(T - t, x, z) \tag{102}$$

Let U denote the closed unit ball in $\mathbf{R}^n$. From Theorem 1.47 there exists a continuous function $f : [0, T] \times \mathbf{R}^n \times U \mapsto \mathbf{R}^n$ and $\alpha > 0$ such that

$$\begin{cases} \forall\, (t, x) \in [0, T] \times \mathbf{R}^n, \ \ F(t, x) \ = \ f(t, x, U) \\ \forall\, u \in U, \ \ f(t, \cdot, u) \ \text{ is } \ \alpha c_R(t) - \text{Lipschitz on } \ B_R(0) \text{ a.e. in } [0, T] \\ \forall\, (t, x) \in [0, T] \times \mathbf{R}^n \ \text{ and for all } \ u, v \ \in \ U, \ \text{ we have} \\ \|f(t, x, u) - f(t, x, v)\| \leq \alpha(\sup_{y\in F(t,x)} \|y\|)\, \|u - v\| \end{cases}$$

Fix $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$ and $x \in \mathcal{S}_{[t_0,T]}(x_0)$. It is enough to consider the case $V(T, x(T)) < \infty$.

Consider a measurable map $u : [t_0, T] \mapsto U$ such that

$$x'(t) \ = \ f(t, x(t), u(t))$$

almost everywhere. Consider a sequence of continuous maps $u_k : [t_0, T] \mapsto U$ converging to u in $L^1(t_0, T; U)$ and let x_k denote the solution to

$$x_k'(t) \ = \ f(t, x_k(t), u_k(t)), \ \ t \ \in \ [t_0, T], \ \ x_k(T) \ = \ x(T)$$

The Gronwall lemma implies that x_k converge uniformly to x. On the other hand, the map $t \mapsto (T-t, x_k(T-t), V(T, x(T)))$ is the only solution to

$$\left\{\begin{array}{l} \gamma'(t) \;=\; -1 \\ y'(t) \;=\; -f(T-t, y(t), u_k(T-t)) \\ z'(t) \;=\; 0 \\ \gamma(0) \;=\; T, \;\; y(0) \;=\; x(T), \;\; z(0) \;=\; V(T, x(T)) \end{array}\right. \tag{103}$$

By (102) we know that

$$\forall\, (\gamma, x, z) \;\in\; \mathcal{E}p(V), \;\; (-1, -f(\gamma, x, u_k(\gamma)), 0) \in T_{\mathcal{E}p(V)}(\gamma, x, z)$$

On the other hand the map

$$(t, x) \;\hookrightarrow\; \{-f(T-t, x, u_k(T-t))\}$$

being continuous, from Viability Theorem we deduce that problem (103) has at least one solution

$$[0, T-t_0] \;\ni\; t \;\mapsto\; (\gamma(t), y(t), z(t)) \in \mathcal{E}p(V)$$

Consequently,

$$\forall\, 0 \;\leq t \;\leq T-t_0, \quad (T-t, x_k(T-t), V(T, x(T))) \in \mathcal{E}p(V)$$

and therefore

$$\forall\, 0 \;\leq\; t \;\leq\; T-t_0, \;\; V(T, x(T)) \geq V(T-t, x_k(T-t))$$

In particular, $V(t_0, x_k(t_0)) \leq V(T, x(T))$. Taking the limit when $k \to \infty$ and using that V is lower semicontinuous, we deduce $ii)$ for $t = T$.

Conversely, assume that $ii)$ is verified. Let $(t_0, x_0) \in \mathrm{Dom}(V)$ be such that $t_0 > 0$. Fix $v \in F(t_0, x_0)$. Corollary 1.34 implies that for some $\overline{h} > 0$ there exist $y_0 \in \mathbf{R}^n$ and $y \in S_{[t_0-\overline{h}, t_0]}(y_0)$ such that $y(t_0) = x_0$ and

$$\lim_{h \to 0+} \frac{y(t_0-h) - x_0}{h} = -v$$

On the other hand, by $ii)$,

$$\forall\, h \;\in\; [0, \overline{h}], \;\; V(t_0-h, y(t_0-h)) \leq V(t_0, x_0)$$

Consequently $D_{\uparrow} V(t_0, x_0)(-1, -v) \leq 0$. Since $v \in F(t_0, x_0)$ is arbitrary $i)$ follows.

4.2.3 Value Function & Contingent Solutions

We prove here equivalence of *i*) and *iii*) of Theorem 4.5.

The Proposition below yields the implication *i*) $\Longrightarrow$ *iii*).

Proposition 4.10 *Assume (92) and let V be defined by (93). Then for all* $(t_0, x_0) \in \mathrm{Dom}(V)$,

$$\begin{cases} t_0 < T & \Longrightarrow \quad \inf_{v \in F(t_0,x_0)} D_{\uparrow} V(t_0, x_0)(1, v) \leq 0 \\ t_0 > 0 & \Longrightarrow \quad \sup_{v \in F(t_0,x_0)} D_{\uparrow} V(t_0, x_0)(-1, -v) \leq 0 \end{cases}$$

Proof — Fix (t_0, x_0) as above. Then, there exists $x \in \mathcal{S}_{[t_0,T]}(x_0)$ such that $V(t, x(t)) \equiv g(x(T))$. Theorem 4.8 ends the proof of the first statement. The second one follows from Theorem 4.9. $\diamond$

The implication *iii*) $\Longrightarrow$ *i*) follows from

Theorem 4.11 *Assume that (92) holds true. Then the function V defined by (93) is the only lower semicontinuous function from* $[0,T] \times \mathbf{R}^n$ *into* $\mathbf{R} \cup \{+\infty\}$ *satisfying*

$$\begin{cases} V(T, \cdot) = g(\cdot) \text{ and for all } (t, x) \in \mathrm{Dom}(V), \\ 0 \leq t < T \implies \inf_{v \in F(t,x)} D_{\uparrow} V(t, x)(1, v) \leq 0 \\ 0 < t \leq T \implies \sup_{v \in F(t,x)} D_{\uparrow} V(t, x)(-1, -v) \leq 0 \end{cases} \tag{104}$$

Proof — Proposition 4.10 implies that V verifies (104). Conversely, consider an extended lower semicontinuous $W : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ satisfying (104). Fix $(t_0, x_0) \in \mathrm{Dom}(W)$ with $t_0 < T$. By Theorem 4.8, there exists $\overline{x} \in S_{[t_0,T]}(x_0)$ such that

$$V(t_0, x_0) \leq g(\overline{x}(T)) = W(T, \overline{x}(T)) \leq W(t_0, x_0)$$

Therefore $W \geq V$. To prove the opposite inequality, consider $(t_0, x_0) \in \mathrm{Dom}(V)$ and $\overline{x} \in \mathcal{S}_{[t_0,T]}(x_0)$ such that $V(t_0, x_0) = g(\overline{x}(T))$. Thus, by Theorem 4.9,

$$W(t_0, x_0) \leq W(T, \overline{x}(T)) = g(\overline{x}(T)) = V(t_0, x_0)$$

Hence $W \leq V$ and the proof is complete.

4.2.4 Regularity of Value Function at Boundary Points

We observe that *i*) yields *iv*). To prove that *iv*) yields *i*), by using the equivalence proved in the preceding subsection, it is enough to show that *iv*) implies *iii*).

Consider a set-valued map $F : [0,T] \times \mathbf{R}^n \hookrightarrow \mathbf{R}^n$.

Theorem 4.12 *Assume (92). If an extended lower semicontinuous function* $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ *satisfies*

$$\left\{ \begin{array}{l} \forall\, (t,x) \in]0,T[\times \mathbf{R}^n, \forall\, (p_t,p_x) \in \partial_- V(t,x), -p_t + H(t,x,-p_x) = 0 \\ \forall\, \overline{x} \in \mathbf{R}^n, \;\; V(0,\overline{x}) \;=\; \liminf_{t\to 0+,\; x\to \overline{x}} V(t,x) \\ \forall\, \overline{x} \in \mathbf{R}^n, \;\; V(T,\overline{x}) \;=\; \liminf_{t\to T-,\; x\to \overline{x}} V(t,x) \end{array} \right.$$

then for all $(t,x) \in \mathrm{Dom}(V)$,

$$\left\{ \begin{array}{rcl} 0 \;<\; t \;\leq\; T & \Longrightarrow & \sup_{v\in F(t,x)} D_\uparrow V(t,x)(-1,-v) \leq 0 \\ 0 \;\leq\; t \;<\; T & \Longrightarrow & \inf_{v\in F(t,x)} D_\uparrow V(t,x)(1,v) \leq 0 \end{array} \right. \tag{105}$$

Proof — From the proofs of Theorems 4.6 and 4.7 we deduce that for all $(t,x) \in \mathrm{Dom}(V)$ with $0 < t < T$ we have

$$\inf_{v\in F(t,x)} D_\uparrow V(t,x)(1,v) \leq 0, \quad \sup_{v\in F(t,x)} D_\uparrow V(t,x)(-1,-v) \leq 0$$

This and Theorems 4.8 and 4.9 yield that for all $0 < t_1 \leq t_2 < T$

$$\forall\, x \;\in\; \mathcal{S}_{[t_1,t_2]}(x_1), \quad V(t_1,x_1) \;\leq\; V(t_2,x(t_2)) \tag{106}$$

and

$$\forall\, (t_1,x_1),\; \exists\, x \in \mathcal{S}_{[t_1,t_2]}(x_1),\; V(t_1,x_1) = V(t_2,x(t_2)) \tag{107}$$

Fix $\overline{x} \in \mathrm{Dom}(V(T,\cdot))$ and let $t_i \to T-,\; x_i \to \overline{x}$ be such that

$$\lim_{i\to\infty} V(t_i,x_i) = V(T,\overline{x})$$

Consider any $x_0 \in \mathbf{R}^n$ and $x \in \mathcal{S}_{[0,T]}(x_0)$ satisfying $x(T) = \overline{x}$. Then we can find $\overline{y}_i$ and $y_i \in \mathcal{S}_{[0,T]}(\overline{y}_i)$ such that $y_i(t_i) = x_i$ and y_i converge to x uniformly on $[0,T]$. Then for all arbitrary, but fixed $0 < t < T$ and all i large enough,

$$V(t,y_i(t)) \leq V(t_i,x_i)$$

Since V is lower semicontinuous,

$$V(t,x(t)) \le \liminf_{i\to\infty} V(t,y_i(t)) \le \lim_{i\to\infty} V(t_i,x_i) = V(T,\overline{x})$$

This, (106) and Theorem 4.9 yield the first inequality in (105).

To prove the second one fix $(0,\overline{x}) \in \mathrm{Dom}(V)$ and consider $t_i \to 0+$, $\overline{x}_i \to \overline{x}$ such that

$$V(0,\overline{x}) \;=\; \lim_{i\to\infty} V(t_i,\overline{x}_i)$$

Let $\overline{y}_i \in \mathbf{R}^n$ and $x_i \in \mathcal{S}_{[0,T]}(\overline{y}_i)$ be such that $x_i(t_i) = \overline{x}_i$ and

$$\forall\, t_i \le t \le T - \frac{1}{i} \;\text{ we have }\; V(t_i,\overline{x}_i) = V(t,x_i(t))$$

Taking a subsequence and keeping the same notations, we may assume that x_i converge uniformly to some $x \in \mathcal{S}_{[0,T]}(\overline{x})$. Then for all $0 < t < T$,

$$V(0,\overline{x}) = \lim_{i\to\infty} V(t,x_i(t)) \ge V(t,x(t))$$

This, (107) and Theorem 4.8 imply the second inequality in (105).

4.3 Viscosity Solutions

In this subsection we prove that statements $i)$ and $v)$ of Theorem 4.5 are equivalent, whenever V is continuous.

Let $F : [0,T] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a set-valued map with nonempty compact images and H be defined by (95).

Consider the Hamilton-Jacobi-Bellman equation

$$-\frac{\partial V}{\partial t}(t,x) \;+\; H\left(t,x,-\frac{\partial V}{\partial x}(t,x)\right) \;=\; 0 \tag{108}$$

Clearly, any V satisfying $ii)$ of Theorem 4.5 is a viscosity supersolution.

Theorem 4.13 *Let $V : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R} \cup \{+\infty\}$ be an extended lower semicontinuous function. Assume that F is upper semicontinuous and has convex compact nonempty images.*

Then the following two statements are equivalent:

i) V is a viscosity supersolution of (108)

ii) For all $0 < t < T$ and $x \in \mathbf{R}^n$ such that $V(t,x) \ne +\infty$, we have

$$\inf_{v\in F(t,x)} D_{\uparrow}V(t,x)(1,v) \le 0 \tag{109}$$

Proof — This follows by the same arguments as in the proof of Theorem 4.6. ◇

Notice next that

$$T_{\mathcal{H}yp(\varphi)}(x_0, \varphi(x_0)) \;=\; \mathcal{H}yp\ (D_{\downarrow}\varphi(x_0))$$

where $\mathcal{H}yp$ denotes the hypograph.

In particular, $p \in \partial_+\varphi(x_0)$ if and only if

$$\forall\, u \;\in\; \mathbf{R}^n,\;\; D_{\downarrow}\varphi(x_0)(u) \leq \langle p, u\rangle \tag{110}$$

Thus

$$p \;\in\; \partial_+\varphi(x_0) \;\;\Longleftrightarrow\;\; (-p,\; +1) \;\in\; N^0_{\mathcal{H}yp\ (\varphi)}(x_0, \varphi(x_0)) \tag{111}$$

Theorem 4.14 *Let $V : [0,T] \times \mathbf{R}^n \to \mathbf{R}$ be continuous. Assume that F satisfies (92).*

Then the following two statements are equivalent

i) V is a viscosity subsolution of (108)

ii) For all $0 < t < T,\; x\;,\; \sup_{v\in F(t,x)} D_{\uparrow}V(t,x)(-1,-v) \;\leq\; 0$

Proof — Assume that *ii)* holds true. Fix $0 < t_0 < T$. By Theorem 4.9, for every $t_0 \leq t_1 < T$ and every $x \in \mathcal{S}_{[t_0,t_1]}(x_0)$ the following holds true:

$$\forall\, t \in [t_0, t_1], \quad V(t_0, x_0) \;\leq\; V(t, x(t))$$

Fix $v \in F(t_0, x_0)$. By Corollary 1.34 there exist $t_0 < t_1 < T$ and $x \in \mathcal{S}_{[t_0,t_1]}(x_0)$ such that $x'(t_0) = v$. The above inequality yields that $0 \leq D_{\downarrow}V(t_0, x_0)(1, v)$. Consequently,

$$\forall\, (p_t, p_x) \;\in\; \partial_+V(t_0, x_0), \quad 0 \leq p_t + \langle p_x, v\rangle$$

Since $v \in F(t_0, x_0)$ is arbitrary, V is a viscosity subsolution.

Assume *i)*. We claim that for all (t, x) such that $0 < t < T$ and all $z \leq V(t, x)$ we have

$$\forall\, (q_t, q_x, q) \;\in\; N^0_{\mathcal{H}yp\ (V)}(t, x, z), \quad q_t \;+\; H(t, x, q_x) \;\leq\; 0 \tag{112}$$

Indeed it is enough to consider the case $z = V(t, x)$. Fix such (q_t, q_x, q). Clearly $q \geq 0$. If $q > 0$ then

$$\left(\frac{q_t}{q},\; \frac{q_x}{q},\; +1\right) \in N^0_{\mathcal{H}yp\ (V)}(t, x, V(t, x))$$

Hence, by (111) and i),

$$\frac{q_t}{q} \ + \ H\left(t,\ x,\ \frac{q_x}{q}\right) \ \leq \ 0$$

and therefore $q_t + H(t,x,q_x) \leq 0$. If $q = 0$, applying Lemma 1.16 to the extended lower semicontinuous function $(s,y) \mapsto -V(s,y)$, we can find a sequence $(t_i, x_i) \to (t,x)$ and a sequence

$$\left(q_t^i,\ q_x^i,\ q^i\right) \ \in \ N^0_{\mathcal{H}yp\,(V)}(t,x,V(t,x))$$

such that $q^i > 0$ and (q_t^i, q_x^i) converge to (q_t, q_x). This and continuity of H yield (112).

We next deduce from (112) and the separation theorem that for all (t,x) such that $0 < t < T$ and all $z \leq V(t,x)$

$$\{1\} \times F(t,x) \times \{0\} \ \subset \ \overline{co}\left(T_{\mathcal{H}yp\,(V)}(t,x,z)\right)$$

This, Theorem 1.3 and lower semicontinuity of F imply that for all (t,x) satisfying $0 < t < T$

$$\begin{aligned}
&\{1\} \times F(t,x) \times \{0\} \\
&\subset \operatorname{Liminf}_{\substack{(t',x',z') \to (t,x,V(t,x)) \\ (t',x',z') \in \mathcal{H}yp\,(V)}} \overline{co}\left(T_{\mathcal{H}yp\,(V)}(t',x',z')\right) \\
&\subset \ T_{\mathcal{H}yp\,(V)}(t,x,V(t,x)) \ = \ \mathcal{H}yp\,(D_\downarrow V(t,x))
\end{aligned}$$

Thus for all (t,x) satisfying $0 < t < T$,

$$\inf_{v \in F(t,x)} D_\downarrow V(t,x)(1,v) \ \geq \ 0$$

Define $W(t,x) = -V(T-t,x)$. Then for all (t,x) such that $0 < t < T$ and for all $v \in F(T-t,x)$, we have

$$D_\uparrow W(t,x)(-1,v) = -D_\downarrow V(T-t,x)(1,v) \ \leq \ 0$$

Applying Theorem 4.9 to W and the set-valued map

$$\widehat{F}(t,x) = -F(T-t,x)$$

we deduce that for every solution y to the inclusion

$$y'(t) \in \widehat{F}(t,y(t)) \ \text{ a.e. in } \ [t_0,t_1]$$

where $0 < t_0 \leq t_1 < T$ we have

$$\forall\, t \;\in\; [t_0, t_1], \quad W(t_0, x_0) \leq W(t, y(t))$$

Fix any $v \in F(t_0, x_0)$ and consider a solution $y(\cdot)$ to the differential inclusion

$$\begin{cases} y' \;\in\; \widehat{F}(t, y) \\ y(T - t_0) \;=\; x_0, \;\; y'(T - t_0) \;=\; -v \end{cases}$$

Then for all small $s > 0$,

$$W(T - t_0, x_0) \leq W(T - t_0 + s, y(T - t_0 + s))$$

and therefore for a sequence $v_s \to v$ we have

$$V(t_0 - s, x_0 - sv_s) \leq V(t_0, x_0)$$

This yields that $D_{\uparrow} V(t_0, x_0)(-1, -v) \leq 0$. Since $v \in F(t_0, x_0)$ is arbitrary, *ii*) follows. $\diamond$

Let $V : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R}$ be a continuous *viscosity solution* to (108). Then, by Theorems 4.13, 4.14 and Proposition 1.15, V verifies *iv*) of Theorem 4.5.

This completes the proof of Theorem 4.5.

5 Value Function of Bolza Problem and Riccati Equations

This Section is concerned with the characteristics of the Hamilton-Jacobi equation

$$-\frac{\partial V}{\partial t} \;+\; H\left(t, x, -\frac{\partial V}{\partial x}\right) \;=\; 0, \quad V(T, \cdot) \;=\; g(\cdot) \tag{113}$$

i.e. solutions to the *Hamiltonian system*

$$\begin{cases} x'(t) \;=\; \dfrac{\partial H}{\partial p}(t, x(t), p(t)), \quad x(T) = x_T \\ \\ -p'(t) \;=\; \dfrac{\partial H}{\partial x}(t, x(t), p(t)), \quad p(T) = \; -\nabla g(x_T) \end{cases} \tag{114}$$

In Section 3 such system arises as "extremal equations" in optimal control, since the Pontryagin maximum principle states that if $x : [t_0, T] \mapsto \mathbf{R}^n$ is

optimal for the Mayer problem and $\nabla g(x(T)) \neq 0$, then there exists $p : [t_0, T] \mapsto \mathbf{R}^n$ such that (x, p) solves (114) with $x_T = x(T)$. This is not however a sufficient condition for optimality because it may happen that to a given $x_0 \in \mathbf{R}^n$ correspond two distinct solutions (x_i, p_i), $i = 1, 2$ of (114) satisfying

$$x_i(t_0) = x_0 \tag{115}$$

and with one of x_i being not optimal. If the solution of (114) is unique for every $x_T \in \mathbf{R}^n$, then

$$p_1(t_0) \neq p_2(t_0) \tag{116}$$

Whenever (115) and (116) hold true for some solutions (x_i, p_i), $i = 1, 2$ of (114), we say that the system (114) has a *shock* at time t_0.

It was already observed in Section 3 that for the Mayer problem the Hamiltonian $H(t, x, \cdot)$ is not differentiable at zero. For this reason the system (114) is not well defined. In this Section we study the Bolza problem:

$$\text{minimize} \int_{t_0}^{T} L(t, x(t), u(t))dt \ + \ g(x(T))$$

over solution-control pairs (x, u) of the control system

$$\begin{cases} x'(t) &= f(t, x(t), u(t)), \quad u(t) \in U \\ x(t_0) &= x_0 \end{cases}$$

The Hamiltonian for this problem is given by

$$H(t, x, p) = \sup_{u \in U}(\langle p, f(t, x, u)\rangle - L(t, x, u))$$

For a class of nonlinear control problems $H(t, \cdot, \cdot)$ is everywhere differentiable. We provide in Subsection 1 an example of such situation.

If shocks never occur on the time interval $[0, T]$, then the solution of (113) can be constructed by simply setting

$$V(t_0, x(t_0)) = g(x(T)) + \int_{t_0}^{T} L(t, x(t), u(t))dt$$

where x solves (114) for some $p(\cdot)$ and the control $u(t) \in U$ is so that $x'(t) = f(t, x(t), u(t))$ almost everywhere in $[t_0, T]$. Then V is the value

function of the above Bolza optimal control problem. Furthermore in this case V is continuously differentiable and

$$\frac{\partial V}{\partial x}(t, x(t)) = -p(t) \quad \& \quad \frac{\partial V}{\partial t}(t, x(t)) = H(t, x(t), p(t))$$

It is well known that shocks do happen. This is the very reason why the value function is not smooth and why one should not expect smooth solutions to the Hamilton-Jacobi-Bellman equation (113). Also it was shown in Section 4 that for the Mayer problem the value function is not smooth at some point (t_0, x_0), where the co-state is non degenerate, if and only if the optimal trajectory is not unique. We shall show under what circumstances a similar statement holds true everywhere for the Bolza problem.

If we could guarantee that on some time interval $[t_0, T]$ there is no shocks, then the value function would be continuously differentiable on $[t_0, T] \times \mathbf{R}^n$ solution of (113). In the same time we have the uniqueness of optimal trajectories and obtain the optimal feedback low $G : [t_0, T] \times \mathbf{R}^n \hookrightarrow U$ by setting

$$G(t, x) = \left\{ u \mid H(t, x, -\frac{\partial V}{\partial x}(t, x)) = \langle -\frac{\partial V}{\partial x}(t, x), f(t, x, u) \rangle - L(t, x, u) \right\}$$

Then the closed loop control system

$$x' = f(t, x, u(t, x)), \quad u(t, x) \in G(t, x), \quad x(t_0) = x_0$$

has exactly one solution and it is optimal for the Bolza problem.

Actually, when the data is smooth, the shocks would not occur till time t_0 if for every (x, p) solving (114) on $[t_0, T]$ the matrix Riccati equation

$$\begin{cases} P' + \dfrac{\partial^2 H}{\partial p \partial x}(t, x(t), p(t))P \ + \ P \ \dfrac{\partial^2 H}{\partial x \partial p}(t, x(t), p(t)) \ + \\ \\ + \ P \ \dfrac{\partial^2 H}{\partial p^2}(t, x(t), p(t))P \ + \ \dfrac{\partial^2 H}{\partial x^2}(t, x(t), p(t)) \ = \ 0 \\ \\ P(T) = -g''(x(T)) \end{cases} \tag{117}$$

has a solution on $[t_0, T]$.

In Subsection 1 we show that the existence of global solutions to Riccati equations (117) implies the absence of shocks. Subsection 2 is devoted to comparison theorems for solutions of (117). In Subsection 3 we relate the

nonexistence of shocks to smoothness of the value function and uniqueness of optimal solutions and then apply the above results to problems with concave-convex Hamiltonians.

5.1 Matrix Riccati Equations and Shocks

In this subsection we relate the absence of shocks of the Hamilton-Jacobi-Bellman equation (113) with the existence of solutions to matrix Riccati equations (114).

Consider $H : [0,T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ and $\psi : \mathbf{R}^n \mapsto \mathbf{R}^n$. We assume that $H(t,\cdot,\cdot)$ is differentiable and associate with these data the Hamiltonian system

$$\begin{cases} x'(t) = \dfrac{\partial H}{\partial p}(t,x(t),p(t)), \quad x(T) = x_T \\ -p'(t) = \dfrac{\partial H}{\partial x}(t,x(t),p(t)), \quad p(T) = \psi(x_T) \end{cases} \tag{118}$$

Definition 5.1 *The system (118) has a* shock *at time t_0 if there exist two solutions $(x_i,p_i)(\cdot)$, $i=1,2$ of (118) such that*

$$x_1(t_0) = x_2(t_0) \quad \& \quad p_1(t_0) \neq p_2(t_0)$$

Definition 5.2 *The* Hamiltonian system *(118) is called* complete *if for every x_T, the solution of (118) is defined on $[0,T]$ and depends continuously on the "initial" state in the following sense:*

Let (x_i,p_i) be solutions of (118) satisfying $x_i(t_i) \to x_0$, $p_i(t_i) \to p_0$ for some $t_i \to t_0$, $x_0 \in \mathbf{R}^n$, $p_0 \in \mathbf{R}^n$. Then (x_i,p_i) converge uniformly to the solution (x,p) of (118) such that $x(t_0) = x_0$ and $p(t_0) = p_0$.

Remark —

a) If the Hamiltonian system (118) is complete, then for all $t_0 \in [0,T]$, $x_0 \in \mathbf{R}^n$, $p_0 \in \mathbf{R}^n$ it has at most one solution (x,p) satisfying $x(t_0) = x_0$, $p(t_0) = p_0$.

b) The Hamiltonian system (118) is complete for instance if for all $r > 0$ there exists $k_r \in L^1(0,T)$ such that the mapping $\frac{\partial H}{\partial(x,p)}(t,\cdot,\cdot)$ is $k_r(t)$-Lipschitz on $B_r(0)$ and has a linear growth: for some $\gamma \in L^1(0,T)$

$$\forall\, x,\ p \in \mathbf{R}^n, \quad \left\| \frac{\partial H}{\partial(x,p)}(t,x,p) \right\| \leq \gamma(t)\,(\|x\| + \|p\| + 1) \quad \diamond$$

Example — Consider

$$f : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R}^n,\ g : [0,T] \times \mathbf{R}^n \mapsto L(U, \mathbf{R}^n),\ l : [0,T] \times \mathbf{R}^n \mapsto \mathbf{R}$$

where U is a finite dimensional space and let $R(t) \in L(U,U)$ be self-adjoint and positive for every $t \in [0,T]$. Define

$$H(t,x,p) = \langle p, f(t,x)\rangle + \sup_{u \in U} \left(\langle p, b(t,x)u\rangle - \frac{1}{2} \langle R(t)u, u\rangle \right) - l(t,x)$$

Then it is not difficult to check that

$$H(t,x,p) = \langle p, f(t,x)\rangle + \left\langle R(t)^{-1} b(t,x)^\star p,\ b(t,x)^\star p \right\rangle - l(t,x)$$

Thus an appropriate smoothness of $f(t,\cdot),\ b(t,\cdot)$ and $l(t,\cdot)$ implies differentiability of $H(t,\cdot,\cdot)$ and completeness of the associated Hamiltonian system. ◇

Theorem 5.3 *Assume that ψ is locally Lipschitz, that $H(t,\cdot,\cdot)$ is twice continuously differentiable and that for every $r > 0$, there exists $k_r \in L^1(0,T)$ satisfying*

$$\frac{\partial H}{\partial (x,p)}(t,\cdot,\cdot) \ \text{ is } \ k_r(t) - \text{Lipschitz on } \ B_r(0)$$

Further assume the Hamiltonian system (118) is complete and define for every $t \in [0,T]$ the set

$$M_t = \{(x(t), p(t)) \mid (x,p) \ \text{ solves } \ (118) \ \text{ for some } \ x_T \in \mathbf{R}^n\}$$

Then the following two statements are equivalent:

i) $\forall\, t \in [0,T]$, *M_t is the graph of a locally Lipschitz function from an open set $\mathcal{D}(t)$ into $\mathbf{R}^n$*

ii) $\forall\ (x,p)$ *solving (118) on $[0,T]$ and $P_T \in \partial^\star \psi(x(T))$, the matrix Riccati equation*

$$\begin{cases} P' + \dfrac{\partial^2 H}{\partial p \partial x}(t,x(t),p(t))P \ + \ P\, \dfrac{\partial^2 H}{\partial x \partial p}(t,x(t),p(t)) + \\ \\ + P\, \dfrac{\partial^2 H}{\partial p^2}(t,x(t),p(t))P \ + \ \dfrac{\partial^2 H}{\partial x^2}(t,x(t),p(t)) \ = \ 0 \\ \\ P(T) = P_T \end{cases} \tag{119}$$

has a solution on $[0,T]$.

Furthermore, if i) (or equivalently ii)) holds true, then

$$\psi \text{ is differentiable } \Longrightarrow \ M_t \text{ is the graph of a differentiable function}$$

$$\psi \in C^1 \ \Longrightarrow \ M_t \text{ is the graph of a } C^1 - \text{function}$$

Corollary 5.4 *Under all assumptions of Theorem 5.3, suppose that for every* (x,p) *solving (118) on* $[0,T]$ *and* $P_T \in \partial^\star \psi(x(T))$, *the matrix Riccati equation (119) has a solution on* $[0,T]$. *Then the Hamiltonian system (118) has no shocks on* $[0,T]$.

To prove the above theorem the following lemma is needed.

Lemma 5.5 *Assume that the Hamiltonian system (118) is complete and for every* $r > 0$, *there exists* $k_r \in L^1(0,T)$ *such that*

$$\frac{\partial H}{\partial(x,p)}(t,\cdot,\cdot) \ \text{ is } \ k_r(t) - \text{Lipschitz on } \ B_r(0)$$

Let $K \subset \mathbf{R}^n$ *be a compact set. Consider a locally Lipschitz function* $\psi : \mathbf{R}^n \mapsto \mathbf{R}^n$ *and the subsets* $M_t(K),\ t \in [0,T]$ *defined by*

$$M_t(K) = \{(x(t),p(t)) \mid (x,p) \ \text{ solves } \ (118),\ x_T \in K\}$$

Then there exists $\delta > 0$ *such that for all* $t \in [T-\delta, T]$, $M_t(K)$ *is the graph of a Lipschitz function.*

Proof — From the completeness of (118) we deduce that the subsets $M_t(K)$ are compact and contained in the ball $B_r(0)$ for some $r > 0$. Set $k(t) = k_r(t)$

We proceed by a contradiction argument. Assume for a moment that there exist $t_i \to T-$ such that $M_{t_i}(K)$ is not the graph of a Lipschitz function. Then for every i we can find two distinct solutions $(x_j^i, p_j^i),\ j = 1,2$ of the Hamiltonian system (118) such that

$$\varepsilon_i \ := \ \frac{\|x_1^i(t_i) - x_2^i(t_i)\|}{\|p_1^i(t_i) - p_2^i(t_i)\|} \ \to \ 0 \ \text{ as } \ i \to +\infty$$

Since for every $s \in [t_i, T]$ we have

$$\left\|x_1^i(s) - x_2^i(s)\right\| \leq$$

$$\varepsilon_i \|p_1^i(t_i) - p_2^i(t_i)\| + \int_{t_i}^s k(\tau)(\left\|x_1^i(\tau) - x_2^i(\tau)\right\| + \left\|p_1^i(\tau) - p_2^i(\tau)\right\|)d\tau$$

the Gronwall lemma implies that for some $C > 0$ independent from i and for all $s \in [t_i, T]$

$$\left\|x_1^i(s) - x_2^i(s)\right\| \leq C(\varepsilon_i \|p_1^i(t_i) - p_2^i(t_i)\| + \int_{t_i}^s k(\tau)\|p_1^i(\tau) - p_2^i(\tau)\|d\tau)$$

Hence for some $C_1 > 0$ and all i large enough and $s \in [t_i, T]$,

$$\left\|p_1^i(s) - p_2^i(s)\right\| \leq$$

$$\left\|p_1^i(t_i) - p_2^i(t_i)\right\| + \int_{t_i}^s k(\tau)(\left\|x_1^i(\tau) - x_2^i(\tau)\right\| + \left\|p_1^i(\tau) - p_2^i(\tau)\right\|)d\tau$$

$$\leq C_1 \left\|p_1^i(t_i) - p_2^i(t_i)\right\| + C_1 \int_{t_i}^s k(\tau) \left\|p_1^i(\tau) - p_2^i(\tau)\right\| d\tau$$

From the Gronwall lemma we deduce that for some $L > 0$ independent from i and all $s \in [t_i, T]$,

$$\left\|p_1^i(s) - p_2^i(s)\right\| \leq L \left\|p_1^i(t_i) - p_2^i(t_i)\right\|$$

This implies that

$$\overline{\varepsilon}_i := \sup_{s \in [t_i, T]} \frac{\left\|x_1^i(s) - x_2^i(s)\right\|}{\left\|p_1^i(t_i) - p_2^i(t_i)\right\|} \text{ converge to zero} \tag{120}$$

We next observe that for all $s \in [t_i, T]$,

$$\left\|p_1^i(s) - p_2^i(s)\right\| \leq$$

$$\leq \left\|p_1^i(T) - p_2^i(T)\right\| + \int_s^T k(\tau)(\left\|x_1^i(\tau) - x_2^i(\tau)\right\| + \left\|p_1^i(\tau) - p_2^i(\tau)\right\|)d\tau$$

$$\leq \left\|p_1^i(T) - p_2^i(T)\right\| + \int_s^T k(\tau)(\left\|p_1^i(\tau) - p_2^i(\tau)\right\| + \overline{\varepsilon}_i \left\|p_1^i(t_i) - p_2^i(t_i)\right\|)d\tau$$

Applying again the Gronwall lemma and taking i large enough we get

$$\left\|p_1^i(t_i) - p_2^i(t_i)\right\| \leq L_1 \left\|p_1^i(T) - p_2^i(T)\right\| + \frac{1}{2} \left\|p_1^i(t_i) - p_2^i(t_i)\right\|$$

for some L_1 independent from i. Hence for all large i

$$\left\|p_1^i(t_i) - p_2^i(t_i)\right\| \leq 2L_1 \left\|p_1^i(T) - p_2^i(T)\right\|$$

and therefore, by (120),

$$\frac{\left\|x_1^i(T) - x_2^i(T)\right\|}{\left\|p_1^i(T) - p_2^i(T)\right\|} = \frac{\left\|x_1^i(T) - x_2^i(T)\right\|}{\left\|p_1^i(t_i) - p_2^i(t_i)\right\|} \times \frac{\left\|p_1^i(t_i) - p_2^i(t_i)\right\|}{\left\|p_1^i(T) - p_2^i(T)\right\|} \to 0$$

Thus

$$\frac{\|\psi(x_1^i(T)) - \psi(x_2^i(T))\|}{\left\|x_1^i(T) - x_2^i(T)\right\|} = \frac{\left\|p_1^i(T) - p_2^i(T)\right\|}{\left\|x_1^i(T) - x_2^i(T)\right\|} \to +\infty$$

which contradicts the Lipschitz continuity of ψ on K. $\diamond$

Proof of Theorem 5.3 — Assume first that for all $t \in [0,T]$, M_t is the graph of a locally Lipschitz function. Consider a solution (x,p) of (118) and the linear system

$$\begin{cases} U' = \dfrac{\partial^2 H}{\partial x \partial p}(t,x(t),p(t))U + \dfrac{\partial^2 H}{\partial p^2}(t,x(t),p(t))V \\ -V' = \dfrac{\partial^2 H}{\partial x^2}(t,x(t),p(t))U + \dfrac{\partial^2 H}{\partial p \partial x}(t,x(t),p(t))V \\ U(T) = Id, \quad V(T) = P_T \end{cases}$$

where $U,\ V : [0,T] \mapsto L(\mathbf{R}^n, \mathbf{R}^n)$ are matrix functions and $P_T \in \partial^\star \psi(x(T))$. Let (x_n, p_n) be solutions of (118) such that

$$\lim_{n\to\infty} x_n(T) = x(T) \quad \& \quad \lim_{n\to\infty} \psi'(x_n(T)) = P_T$$

By completeness of (118), (x_n, p_n) converge uniformly to (x,p).

The variational equation implies that for any $(w(\cdot), q(\cdot))$ solving

$$\begin{cases} w' = \dfrac{\partial^2 H}{\partial x \partial p}(t,x_n(t),p_n(t))w + \dfrac{\partial^2 H}{\partial p^2}(t,x_n(t),p_n(t))q \\ -q' = \dfrac{\partial^2 H}{\partial x^2}(t,x_n(t),p_n(t))w + \dfrac{\partial^2 H}{\partial p \partial x}(t,x_n(t),p_n(t))q \\ w(T) = w_T, \quad q(T) = \psi'(x_n(T))w_T \end{cases} \tag{121}$$

we have $(w(t), q(t)) \in T_{M_t}(x_n(t), p_n(t))$ (contingent cone to M_t at $(x_n(t), p_n(t))$). Because M_t is the graph of a locally Lipschitz function, for some l_t independent from n, $\|q(t)\| \leq l_t \|w(t)\|$.

Taking the limit in (121) we deduce that every solution (w, q) of

$$\left\{\begin{array}{rcl} w' & = & \dfrac{\partial^2 H}{\partial x \partial p}(t, x(t), p(t))w \ + \ \dfrac{\partial^2 H}{\partial p^2}(t, x(t), p(t))q \\ \\ -q' & = & \dfrac{\partial^2 H}{\partial x^2}(t, x(t), p(t))w \ + \ \dfrac{\partial^2 H}{\partial p \partial x}(t, x(t), p(t))q \\ \\ w(T) & = & w_T, \ \ q(T) = P_T w_T \end{array}\right. \tag{122}$$

satisfies $\|q(t)\| \leq l_t \|w(t)\|$.

Thus, by uniqueness of solution to (122), if $w_T \neq 0$, then $w(\cdot)$ never vanishes. Since

$$w(t) = U(t)w_T \quad \& \quad q(t) = V(t)w_T$$

this implies that $U(t)$ is not singular for all $t \in [0, T]$. Setting

$$P(t) = V(t)U(t)^{-1}$$

we check that P satisfies (119).

Conversely let (119) have a solution on $[0, T]$ for all (x, p) solving (118). For every $r > 0$, $t \in [0, T]$ consider the compact sets

$$\Pi_{rt} = \{(x(t), p(t)) \mid (x, p) \ \text{ solves } \ (118), \ x(T) \in B_r(0)\}$$

We first claim that for every $r > 0$ and $t_0 \in [0, T]$, Π_{rt_0} is the graph of a Lipschitz function. Indeed fix r, t_0 as above and assume for a moment that Π_{rt_0} is not the graph of a Lipschitz function.

By Lemma 5.5 for all s near T, Π_{rs} is still the graph of a Lipschitz function. Define

$$\overline{t} \ = \ \inf_{t \in [t_0, T]} \{ \ \forall \ s \in [t, T], \ \ \Pi_{rs} \ \text{ is the graph of a Lipschitz function}\}$$

Then $t_0 \leq \overline{t} < T$ and $\Pi_{r\overline{t}}$ is not the graph of a Lipschitz function, because otherwise, by Lemma 5.5, we could make $\overline{t}$ smaller which would contradict its choice. Define the sets

$$D_r(s) = \{x(s) \mid (x, p) \ \text{ solves } \ (118), \ \|x(T)\| < r\}$$

Observe that for all $r > 0$ and $s \in]\overline{t}, T]$, $D_r(s)$ is open. Its closure is equal to the set

$$\overline{D_r(s)} = \{x(s) \mid (x,p) \text{ solves } (118),\ x(T) \in B_r(0)\}$$

by completeness of (118).

Define next the Lipschitz function $\Phi_{rs} : \overline{D_r(s)} \mapsto \mathbf{R}^n$ by

$$\text{Graph}(\Phi_{rs}) = \Pi_{rs}$$

The Rademacher theorem yields Φ_{rs} is differentiable almost everywhere on $D_r(s)$.

Fix a sequence $t_n \to \overline{t}+$ and observe that the family $\{\Phi_{rt_n}\}_{n\geq 1}$ can not be equilipschitz, because otherwise, using that

$$\Pi_{r\overline{t}} = \text{Lim}_{n\to\infty}\Pi_{rt_n}$$

we would deduce that $\Pi_{r\overline{t}}$ is the graph of a Lipschitz function. Thus there exists a sequence $\overline{x}_n \in D_r(t_n)$ such that $\Phi'_{rt_n}(\overline{x}_n) \to \infty$. Hence

$$\exists\ (u_n, v_n) \in \mathbf{R}^n \times \mathbf{R}^n \text{ satisfying } \Phi'_{t_n}(\overline{x}_n)u_n = v_n,\ \|v_n\| = 1,\ \|u_n\| \to 0$$

Let (x_n, p_n) be a solution of (118) such that $x_n(t_n) = \overline{x}_n$ and $p_n(t_n) = \Phi_{rt_n}(\overline{x}_n)$. Since Φ_{rt_n} is differentiable at $\overline{x}_n$, using variational equation, we deduce that ψ is differentiable at $x_n(T)$. Taking a subsequence and keeping the same notations, by completeness of (118), we may assume that (x_n, p_n) converge uniformly to a solution (x,p) of (118) and for some $P_T \in \partial^\star\psi(x(T))$

$$v_n \to v, \quad \psi'(x_n(T)) \to P_T$$

Consider next the solutions (w_n, q_n) of

$$\begin{cases} w' &= \dfrac{\partial^2 H}{\partial x \partial p}(t, x_n(t), p_n(t))w \ + \ \dfrac{\partial^2 H}{\partial p^2}(t, x_n(t), p_n(t))q \\ -q' &= \dfrac{\partial^2 H}{\partial x^2}(t, x_n(t), p_n(t))w \ + \ \dfrac{\partial^2 H}{\partial p \partial x}(t, x_n(t), p_n(t))q \\ w(t_n) &= u_n, \quad q(t_n) = v_n \end{cases}$$

The variational equation yields $q_n(T) = \psi'(x_n(T))w_n(T)$.

Since $\lim_{n\to\infty}(u_n, v_n) = (0, v)$, passing to the limit in the above system, we deduce that (122) has a solution (w, q) satisfying

$$w(\overline{t}) = 0, \quad q(\overline{t}) \neq 0, \quad q(T) = P_T w(T)$$

In particular $w(T) \neq 0$ and $U(\overline{t})w(T) = 0$. On the other hand, by the previous arguments, $P(t) = V(t)U(t)^{-1}$ solves (119) on $]\overline{t}, T]$. If P is well defined on $[\overline{t}, T]$, then $V(\overline{t}) = P(\overline{t})U(\overline{t})$ and $q(\overline{t}) = V(\overline{t})w(T) = 0$, which leads to a contradiction and proves our claim.

Observe next that for every $s \in [0, T]$ the sequence of open subsets $\{D_r(s)\}_{r>0}$ is non decreasing. Define the open set

$$\mathcal{D}(s) = \bigcup_{k>0} D_k(s)$$

Then

$$\mathcal{D}(s) = \{x \mid \exists\, p \text{ such that } (x, p) \in M_s\}$$

Since $\{\Pi_{rs}\}_{r>0}$ is a non decreasing sequence of graphs of Lipschitz functions, $M_s = \bigcup_{r>0} \Pi_{rs}$ is the graph of a function from $\mathcal{D}(s)$ into $\mathbf{R}^n$.

We next show that M_s is the graph of a locally Lipschitz function. Indeed fix $\overline{x} \in \mathcal{D}(s)$, $r > 0$ such that $B_r(\overline{x}) \subset \mathcal{D}(s)$. Since $B_r(\overline{x})$ is compact and the family of open sets $D_r(s)$ is non decreasing, for some $k > 0$, $B_r(\overline{x}) \subset D_k(s)$. But we already know that $M_s \cap D_k(s) \times \mathbf{R}^n$ is the graph of a Lipschitz function.

The last two statements follow from the variational equation.

5.2 Matrix Riccati Equations

We investigate here matrix differential equations of the following type

$$P' + A(t)^\star P + PA(t) + PE(t)P + D(t) = 0, \quad P(T) = P_T$$

5.2.1 Comparison Theorems

The aim of this subsection is to provide two comparison properties for solutions of Riccati equations.

Theorem 5.6 *Let A, E_i, $D_i : [0, T] \mapsto L(\mathbf{R}^n, \mathbf{R}^n)$, $i = 1, 2$ be integrable. We assume that $E_1(t)$ and $D_1(t)$ are self-adjoint for almost every $t \in [0, T]$ and*

$$D_1(t) \leq D_2(t), \quad E_1(t) \leq E_2(t) \text{ a.e. in } [0, T] \tag{123}$$

Consider self-adjoint operators $P_{iT} \in L(\mathbf{R}^n, \mathbf{R}^n)$ *such that*

$$P_{1T} \leq P_{2T}$$

and solutions $P_i(\cdot) : [t_0, T] \mapsto L(\mathbf{R}^n, \mathbf{R}^n)$ *to the matrix equations*

$$P' + A(t)^\star P + PA(t) + PE_i(t)P + D_i(t) = 0, \quad P_i(T) = P_{iT} \tag{124}$$

for $i = 1, 2$. *If* P_2 *is self-adjoint, then* $P_1 \leq P_2$ *on* $[t_0, T]$.

Proof — From uniqueness of solution to (124), using that $E_1(t)$ and $D_1(t)$ are self-adjoint, it is not difficult to deduce that P_1 is self-adjoint. For all $t \in [t_0, T]$, set

$$Z \ = \ P_2 - P_1, \qquad \mathcal{A}(t) = A(t) + \frac{1}{2}E_1(t)(P_1(t) + P_2(t))$$

Then

$$\mathcal{A}(t)^\star Z(t) + Z(t)\mathcal{A}(t) =$$

$$= A(t)^\star Z(t) + Z(t)A(t) - P_1(t)E_1(t)P_1(t) + P_2(t)E_1(t)P_2(t)$$

Therefore Z solves the Riccati equation

$$Z' + \mathcal{A}(t)^\star Z + Z\mathcal{A}(t) + P_2(t)(E_2(t) - E_1(t))P_2(t) + D_2(t) - D_1(t) = 0$$

Denote by $X(\cdot, t)$ the solution to

$$X' = -\mathcal{A}(s)^\star X, \quad X(t, t) = Id$$

A direct verification yields

$$Z(t) = X(t, T)(P_{2T} - P_{1T})X(t, T)^\star +$$

$$+ \int_t^T X(t, s)(D_2(s) - D_1(s) + P_2(s)(E_2(s) - E_1(s))P_2(s))X(t, s)^\star ds$$

This and assumptions (123) imply $Z \geq 0$ on $[t_0, T]$. $\diamond$

Theorem 5.7 *Let* $A, E_i, D_i : [0, T] \mapsto L(\mathbf{R}^n, \mathbf{R}^n)$, $i = 1, 2$ *be integrable. We assume that* $E_1(t)$, $D_1(t)$ *are self-adjoint for almost all* $t \in [0, T]$ *and*

$$D_1(t) \leq D_2(t), \ \ 0 \leq E_1(t) \leq E_2(t) \ \text{ a.e. in } \ [0, T]$$

Consider self-adjoint operators $P_{iT} \in L(\mathbf{R}^n, \mathbf{R}^n)$ *such that* $P_{1T} \leq P_{2T}$ *and solutions* $P_i(\cdot) : [t_i, T] \mapsto L(\mathbf{R}^n, \mathbf{R}^n)$ *to the matrix equations*

$$P' + A(t)^\star P + PA(t) + PE_i(t)P + D_i(t) = 0, \quad P_i(T) = P_{iT}$$

where $i = 1, 2$. *If* P_2 *is self-adjoint, then the solution* P_1 *is defined at least on* $[t_2, T]$ *and* $P_1 \leq P_2$.

Proof — Consider the square root $B(t)$ of $E_1(t)$, i.e. for almost every $t \in [0, T]$, $E_1(t) = B(t)B(t)^\star$ and set

$$t_0 = \inf_{t \in [0,T]} \{P_1 \text{ is defined on } [t, T]\}$$

Thus either the solution P_1 exists on $[0, T]$ or $\|P_1(t)\| \to \infty$ when $t \to t_0+$. It is enough to show that if $t_2 \leq t_0$, then P_1 is bounded on $]t_0, T]$. So let us assume that $t_2 \leq t_0$. By Theorem 5.6 for every $t_0 < t \leq T$ we have $P_1(t) \leq P_2(t)$. Pick any $x \in \mathbf{R}^n$ of norm one. Since $P_1 = P_1^\star$ we get

$$\langle B(t)^\star P_1(t)x, B(t)^\star P_1(t)x\rangle =$$

$$-\langle P_1'(t)x, x\rangle - \langle A(t)^\star P_1(t)x, x\rangle - \langle P_1(t)A(t)x, x\rangle - \langle D_1(t)x, x\rangle$$

Therefore for every $x \in \mathbf{R}^n$ of norm one and all $t_0 < t \leq T$

$$\int_t^T \|B(s)^\star P_1(s)x\|^2 \, ds \leq$$

$$\leq -\int_t^T < P_1'(s)x, x > +2\int_t^T \|A(s)\| \, \|P_1(s)\| \, ds + \|D_1\|_{L^1(t,T)}$$

$$\leq \langle P_1(t)x, x\rangle + \|P_{1T}\| + 2\int_t^T \|A(s)\| \, \|P_1(s)\| \, ds + \|D_1\|_{L^1(t,T)}$$

$$\leq \|P_2(t)\| + 2\int_t^T \|A(s)\| \, \|P_1(s)\| \, ds + \|P_{1T}\| + \|D_1\|_{L^1(t,T)}$$

$$\leq c + 2\int_t^T \|A(s)\| \, \|P_1(s)\| \, ds$$

for some c independent from t, because P_2 is bounded on $[t_2, T]$.

On the other hand for any $y \in \mathbf{R}^n$ of norm one

$$-\langle P_1'(t)x, y\rangle = \langle P_1(t)B(t)B(t)^\star P_1(t)x, y\rangle + \langle A(t)^\star P_1(t)x, y\rangle +$$

$$+ \langle P_1(t)A(t)x, y\rangle + \langle D_1(t)x, y\rangle$$

Integrating on $[t,T]$ and using the latter inequality and the Hölder inequality, we obtain

$$\langle P_1(t)x,y\rangle \leq \|P_{1T}\| + \|B^\star(\cdot)P_1(\cdot)x\|_{L^2(t,T)}\,\|B^\star(\cdot)P_1(\cdot)y\|_{L^2(t,T)} +$$

$$+2\int_t^T \|A(s)\|\,\|P_1(s)\|\,ds + \|D_1\|_{L^1(t,T)}$$

$$\leq\ c_1 + 2\int_t^T \|A(s)\|\,\|P_1(s)\|\,ds + \left[\left(c + 2\int_t^T \|A(s)\|\,\|P_1(s)\|\,ds\right)^{1/2}\right]^2$$

for some c_1 independent from t. Since this holds true for all x and $y \in \mathbf{R}^n$ of norm one,

$$\forall\ t_0 < t \leq T, \quad \|P_1(t)\|\ \leq\ c + c_1 + 4\int_t^T \|A(s)\|\,\|P_1(s)\|\,ds$$

Applying the Gronwall lemma we deduce that $\|P_1(t)\|$ is bounded on $]t_0,T]$ by a constant independent from t.

5.2.2 Existence of Solutions

We deduce from the previous subsection sufficient conditions for existence of solutions to the matrix Riccati equations.

Theorem 5.8 *Let $A,E,D:[0,T]\mapsto L(\mathbf{R}^n,\mathbf{R}^n)$ be integrable. We assume that $E(t)$, $D(t)$ are self-adjoint and $E(t)\geq 0$ for almost every $t\in[0,T]$. Consider a self-adjoint operator $P_T\in L(\mathbf{R}^n,\mathbf{R}^n)$ and assume that there exists an absolutely continuous $P:[t_0,T]\mapsto L(\mathbf{R}^n,\mathbf{R}^n)$ such that for every $t\in[t_0,T]$, $P(t)$ is self-adjoint and*

$$P'(t) + A(t)^\star P(t) + P(t)A(t) + P(t)E(t)P(t) + D(t) \leq 0 \text{ a.e. in } [t_0,T]$$

and $P_T\leq P(T)$. Then the solution $\overline{P}$ to the equation

$$P' + A(t)^\star P + PA(t) + PE(t)P + D(t) = 0, \quad P(T) = P_T \tag{125}$$

is defined at least on $[t_0,T]$ and $\overline{P}\leq P$ on $[t_0,T]$.

Proof — Set

$$\Gamma(t) = P'(t) + A(t)^\star P(t) + P(t)A(t) + P(t)E(t)P(t) + D(t)$$

Then $\Gamma(t) \leq 0$ and is self-adjoint and P solves the Riccati equation

$$P' + A(t)^\star P + PA(t) + PE(t)P + D(t) - \Gamma(t) = 0$$

where $D(t) - \Gamma(t) \geq D(t)$. By Theorem 5.7, $\overline{P}$ is defined at least on $[t_0, T]$ and $\overline{P} \leq P$. $\diamond$

Corollary 5.9 *Under all assumptions on A, E, D of Theorem 5.8, consider a self-adjoint nonpositive $P_T \in L(\mathbf{R}^n, \mathbf{R}^n)$. If for almost all $t \in [0,T]$, $D(t) \leq 0$, then the solution $\overline{P}$ to the matrix Riccati equation (125) is well defined on $[0,T]$ and $\overline{P} \leq 0$.*

Proof — We apply Theorem 5.8 with $P(\cdot) \equiv 0$. $\diamond$

5.3 Value Function of Bolza Problem

Consider the minimization problem

$$(P) \qquad \text{minimize} \int_{t_0}^{T} L(t, x(t), u(t))dt \; + \; g(x(T))$$

over solution-control pairs (x, u) of the control system

$$\begin{cases} x'(t) &= f(t, x(t), u(t)), \quad u(t) \in U \\ x(t_0) &= x_0 \end{cases} \tag{126}$$

where $t_0 \in [0,T]$, $x_0 \in \mathbf{R}^n$, U is a complete separable metric space,

$$g : \mathbf{R}^n \mapsto \mathbf{R}, \;\; L : [0,T] \times \mathbf{R}^n \times U \mapsto \mathbf{R}, \;\; f : [0,T] \times \mathbf{R}^n \times U \mapsto \mathbf{R}^n$$

The *Hamiltonian* $H : [0,T] \times \mathbf{R}^n \times \mathbf{R}^n \mapsto \mathbf{R}$ is defined by

$$H(t,x,p) = \sup_{u \in U} \left(\langle p, f(t,x,u) \rangle - L(t,x,u) \right)$$

We denote by $\mathcal{U}$ the set of all measurable controls $u : [0,T] \mapsto U$ and by $x(\cdot; t_0, x_0, u)$ the solution of (126) starting at time t_0 from the initial condition x_0 and corresponding to the control $u(\cdot) \in \mathcal{U}$. Of course not to every $u \in \mathcal{U}$ corresponds a solution $x(\cdot; t_0, x_0, u)$ of (126).

For all $(t_0, x_0, u) \in [0,T] \times \mathbf{R}^n \times \mathcal{U}$ set

$$\Phi(t_0, x_0, u) \; = \; \int_{t_0}^{T} L(t, x(t; t_0, x_0, u), u(t))dt + g(x(T; t_0, x_0, u))$$

if this expression is well defined and $\Phi(t_0, x_0, u) = +\infty$ otherwise.

The value function associated to the Bolza problem (P) is defined by

$$V(t_0, x_0) \ = \ \inf_{u \in \mathcal{U}} \Phi(t_0, x_0, u)$$

when (t_0, x_0) range over $[0, T] \times \mathbf{R}^n$.

Proposition 5.10 *Assume that $H(t, \cdot, \cdot)$ is differentiable. Then*

$$\frac{\partial H}{\partial p}(t, x, p) = \{f(t, x, u) \mid \ \langle p, f(t, x, u)\rangle - L(t, x, u) = H(t, x, p)\}$$

and

$$\frac{\partial H}{\partial x}(t, x, p) = \ \left\{ \frac{\partial f}{\partial x}(t, x, u)^\star p - \frac{\partial L}{\partial x}(t, x, u) \mid \right.$$

$$\left. \langle p, f(t, x, u)\rangle - L(t, x, u) = H(t, x, p)\right\}$$

Proof — By Proposition 3.11 applied to the Hamiltonian

$$\mathcal{H}(t, x, (p, q)) = \sup_{u \in U} \langle (f(t, x, u),\ L(t, x, u)),\ (p, q)\rangle$$

at $(p, q) = (p, -1)$ we get

$$\frac{\partial H}{\partial p}(t, x, p) \ = \frac{\partial \mathcal{H}}{\partial p}(t, x, (p, -1)) =$$

$$= \{f(t, x, u) \mid \ \langle p, f(t, x, u)\rangle - L(t, x, u) = H(t, x, p)\}$$

and

$$\frac{\partial H}{\partial x}(t, x, p) = \frac{\partial \mathcal{H}}{\partial x}(t, x, (p, -1)) =$$

$$\left\{ \tfrac{\partial f}{\partial x}(t, x, u)^\star p - \tfrac{\partial L}{\partial x}(t, x, u) \mid \langle p, f(t, x, u)\rangle - L(t, x, u) = H(t, x, p) \right\}$$

$\diamond$

Consider the Hamiltonian system

$$\left\{ \begin{array}{rcll} x'(t) & = & \dfrac{\partial H}{\partial p}(t, x(t), p(t)), & x(T) = x_T \\ \\ -p'(t) & = & \dfrac{\partial H}{\partial x}(t, x(t), p(t)), & p(T) = -\nabla g(x_T) \end{array} \right. \tag{127}$$

Throughout the whole subsection we impose the following hypothesis:

$\mathbf{H}_1$) $f,\ L$ are continuous and $\forall\, r > 0,\ \exists\, k_r \in L^1(0,T)$ such that

$$\forall\, u \in U,\ \ (f(t,\cdot,u), L(t,\cdot,u)) \ \text{ is } \ k_r(t) - \text{Lipschitz on } \ B_r(0)$$

$\mathbf{H}_2$) $f(t,\cdot,u),\ L(t,\cdot,u)$ are differentiable and $g \in \mathcal{C}^1$
$\mathbf{H}_3$) H and $\frac{\partial H}{\partial p}$ are continuous on $[0,T] \times \mathbf{R}^n \times \mathbf{R}^n$
$\mathbf{H}_4$) The Hamiltonian system (127) is complete
$\mathbf{H}_5$) For all $(t,x) \in [0,T] \times \mathbf{R}^n$, the set

$$\{(f(t,x,u), L(t,x,u) + r) \,|\, u \in U,\ r \geq 0\} \ \text{ is closed and convex}$$

5.3.1 Maximum Principle

As in Section 3 to study differentiability of the value function we shall use the maximum principle:

Theorem 5.11 *Assume H_1), H_2) and let $(\overline{x}, \overline{u})$ be an optimal solution-control pair of (P) for some $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$. If $H(t,\cdot,\cdot)$ is differentiable, then there exists $p : [t_0, T] \mapsto \mathbf{R}^n$ such that $(\overline{x}, p)$ solves the Hamiltonian system*

$$\left\{ \begin{array}{rcll} x'(t) & = & \dfrac{\partial H}{\partial p}(t, x(t), p(t)), & x(t_0) = x_0 \\ -p'(t) & = & \dfrac{\partial H}{\partial x}(t, x(t), p(t)), & p(T) = -\nabla g(\overline{x}(T)) \\ p(t_0) & \in & -\partial_+ V_x(t_0, x_0) & \end{array} \right. \tag{128}$$

where $\partial_+ V_x(t_0, x_0)$ denotes the superdifferential of $V(t_0,\cdot)$ at x_0.

Consequently for almost all $t \in [t_0, T]$,

$$H(t, \overline{x}(t), p(t)) = \langle p(t), \overline{x}'(t) \rangle - L(t, \overline{x}(t), \overline{u}(t))$$

Proof — Fix $v \in \mathbf{R}^n$ and let $h_k \to 0+, v_k \to v$ be such that

$$\begin{aligned} D_\downarrow V_x(t_0, x_0)(v) &:= \limsup_{h \to 0+,\ v' \to v} \frac{V(t_0, x_0 + hv') - V(t_0, x_0)}{h} \\ &= \lim_{k \to \infty} \frac{V(t_0, x_0 + h_k v_k) - V(t_0, x_0)}{h_k} \end{aligned}$$

For all k large enough consider the solution $x_k(\cdot)$ of the system

$$\begin{cases} x'(t) &= f(t, x(t), \overline{u}(t)) \\ x(t_0) &= x_0 + h_k v_k \end{cases}$$

The variational equation implies that the sequence $(x_k - \overline{x})/h_k$ converges to the solution $w(\cdot)$ of the linear system

$$w'(t) = \frac{\partial f}{\partial x}(t, \overline{x}(t), \overline{u}(t))w(t), \quad w(t_0) = v$$

Let $X(\cdot)$ denote the fundamental solution of

$$X'(t) = \frac{\partial f}{\partial x}(t, \overline{x}(t), \overline{u}(t))X(t), \quad X(t_0) = Id$$

Then $w(t) = X(t)v$ for all $t \in [t_0, T]$. Thus

$$D_{\downarrow}V_x(t_0, x_0)(v) \leq$$

$$\limsup_{k\to\infty} \frac{\int_{t_0}^{T}(L(t, x_k(t), \overline{u}(t)) - L(t, \overline{x}(t), \overline{u}(t)))dt + g(x_k(T)) - g((\overline{x}(T))}{h_k}$$

$$= \left\langle \int_{t_0}^{T} X(t)^{\star} \frac{\partial L}{\partial x}(t, \overline{x}(t), \overline{u}(t))dt + X(T)^{\star}\nabla g(\overline{x}(T)), v \right\rangle$$

Consider the solution $p(\cdot)$ to the adjoint system

$$\begin{cases} -p' &= \dfrac{\partial f}{\partial x}(t, \overline{x}(t), \overline{u}(t))^{\star} p - \dfrac{\partial L}{\partial x}(t, \overline{x}(t), \overline{u}(t)) \\ p(T) &= -\nabla g(\overline{x}(T)) \end{cases}$$

Then

$$p(t) = -X(t)^{\star^{-1}} \left(X(T)^{\star}\nabla g((\overline{x}(T)) + \int_t^T X(s)^{\star} \frac{\partial L}{\partial x}(s, \overline{x}(s), \overline{u}(s))ds \right)$$

Consequently, for all $v \in \mathbf{R}^n$,

$$D_{\downarrow}V_x(t_0, x_0)(v) \leq \langle -p(t_0), v \rangle$$

and so $p(t_0) \in -\partial_+ V_x(t_0, x_0)$. By the maximum principle for a.e. $t \in [t_0, T]$,

$$\langle p(t), f(t, \overline{x}(t), \overline{u}(t)) \rangle - L(t, \overline{x}(t), \overline{u}(t)) \;=\; H(t, \overline{x}(t), p(t))$$

Since $H(t, \cdot, \cdot)$ is differentiable, we deduce from Proposition 5.10 that $(\overline{x}, p)$ solves the Hamiltonian system (128).

5.3.2 Differentiability of Value Function and Uniqueness of Optimal Solutions

We shall need the following consequence of the maximum principle.

Theorem 5.12 *Assume $H_1)-H_5)$, that V is locally Lipschitz and for every $(t_0,x_0)\in[0,T]\times\mathbf{R}^n$ the problem (P) has an optimal solution. Then for every*

$$\overline{p}\in\partial_x^\star V(t_0,x_0)\ :=\ \mathrm{Limsup}_{x_i\to x_0,\ t_i\to t_0}\left\{\frac{\partial V}{\partial x}(t_i,x_i)\right\}$$

there exists a solution (x,p) of (127) satisfying

$$x(t_0)=x_0\ \ \&\ \ p(t_0)=\overline{p}$$

and x is optimal for problem (P).

In particular if (P) has a unique optimal trajectory $z(\cdot)$, then the set $\partial_x^\star V(t_0,x_0)$ is a singleton. Consequently, $V(t_0,\cdot)$ is differentiable at x_0.

Remark — Various sufficient conditions for local Lipschitz continuity of the value function and for the existence of optimal controls for (P) may be found in many books. They can also be deduced from results of Section 1. We shall not dwell on this question in this section. $\diamond$

Proof — Let $\overline{p}\in\partial_x^\star V(t_0,x_0)$ and $(t_k,x_k)\to(t_0,x_0)$ be such that

$$\lim_{k\to\infty}\frac{\partial V}{\partial x}\,(t_k,x_k)\ =\ \overline{p}$$

Consider optimal solution-control pairs (z_k,u_k) of (P) with (t_0,x_0) replaced by (t_k,x_k). From Theorem 5.11 it follows that there exist absolutely continuous functions p_k such that for all k, (z_k,p_k) solves the following problem

$$\begin{cases} x'(t) &= \dfrac{\partial H}{\partial p}(t,x(t),p(t)),\quad x(t_k)=x_k,\quad p(t_k)=-\dfrac{\partial V}{\partial x}(t_k,x_k)\\[2ex] -p'(t) &= \dfrac{\partial H}{\partial x}(t,x(t),p(t)),\quad p(T)=-\nabla g((z_k(T)) \end{cases}$$

We extend (z_k,p_k) on the time interval $[0,t_k]$ as the solution to the Hamiltonian system

$$\begin{cases} x'(t) &= \dfrac{\partial H}{\partial p}(t,x(t),p(t)),\quad x(t_k)=x_k\\[2ex] -p'(t) &= \dfrac{\partial H}{\partial x}(t,x(t),p(t)),\quad p(t_k)=p_k(t_k) \end{cases}$$

By completeness of (127), (z_k, p_k) converge uniformly to the unique solution (z, p) of the Hamiltonian system

$$\begin{cases} x'(t) &= \dfrac{\partial H}{\partial p}(t, x(t), p(t)), \quad x(t_0) = x_0 \\ -p'(t) &= \dfrac{\partial H}{\partial x}(t, x(t), p(t)), \quad p(t_0) = \overline{p} \end{cases}$$

By Proposition 5.10 for all $k \geq 1$ and almost all $t \in [t_k, T]$,

$$H(t, z_k(t), p_k(t)) = \langle p_k(t), z_k'(t) \rangle - L(t, z_k(t), u_k(t))$$

and from H_3) it follows that $\{z_k'(\cdot)\}$ is bounded in $L^\infty(0, T)$.

We extend $L(\cdot, z_k(\cdot), u_k(\cdot))$ on $[0, t_k[$ by zero function and deduce from the above equality and H_3) that $\{|L(\cdot, z_k(\cdot), u_k(\cdot))|\}_{k \geq 1}$ is bounded in $L^\infty(0, T)$.

Taking a subsequence and keeping the same notations we may assume that

$$(z_k'(\cdot),\ L(\cdot, z_k(\cdot), u_k(\cdot))) \text{ converges weakly in } L^1(0, T) \text{ to } (y(\cdot), \alpha(\cdot))$$

Since for every $t \in [t_k, T]$, $z_k(t) = x_k + \int_{t_k}^t z_k'(s)ds$, taking the limit, we obtain $z(t) = x_0 + \int_{t_0}^T y(s)ds$. Consequently $z'(\cdot) = y(\cdot)$. On the other hand,

$$V(t_k, x_k) = g((z_k(T)) + \int_{t_k}^T L(s, z_k(s), u_k(s))ds$$

Hence, by continuity of V, passing to the limit, we obtain

$$V(t_0, x_0) = g(z(T)) + \int_{t_0}^T \alpha(s)ds$$

By Mazur's theorem and H_1), H_5) for almost all $t \in [t_0, T]$,

$$(y(t), \alpha(t)) \in \{(f(t, z(t), u), L(t, z(t), u) + r) \mid u \in U,\ r \geq 0\}$$

Hence, applying the measurable selection theorem, we can find $\overline{u} \in \mathcal{U}$ and a measurable $r(\cdot) : [t_0, T] \mapsto \mathbf{R}_+$ such that for almost all t,

$$y(t) = f(t, z(t), \overline{u}(t)) \quad \& \quad \alpha(t) = L(t, z(t), \overline{u}(t)) + r(t)$$

This implies that z corresponds to the control $\overline{u} \in \mathcal{U}$. Finally, since $r(t) \geq 0$,

$$V(t_0, x_0) \geq g(z(T)) + \int_{t_0}^T L(s, z(s), \overline{u}(s))ds$$

and therefore $(z, \overline{u})$ is optimal for (P).

To prove the last statement fix $\overline{p}_i \in \partial_x^\star V(t_0, x_0)$, $i = 1, 2$ and let (z_i, p_i), $i = 1, 2$ be solutions of (127) such that $p_i(t_0) = \overline{p}_i$. From the uniqueness of optimal trajectory z and the first claim of our theorem, we deduce that $z^1 = z^2 = z$. Consequently, (z, p_i) solve the Hamiltonian system (127) with $x_T = z(T)$ for $i = 1, 2$. So, by uniqueness, $p_1(t_0) = p_2(t_0)$.

5.3.3 Smoothness of the Value Function

Differentiability of the value function is related to solutions of (127) in the following way.

Theorem 5.13 *Assume $H_1) - H_5)$, that V is locally Lipschitz and for every $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$ the problem (P) has an optimal solution.*

Then the following four statements are equivalent:

i) The value function V is continuously differentiable
ii) For every $t_0 \in [0, T]$, $V(t_0, \cdot)$ is continuously differentiable
iii) $\forall\ (t_0, x_0) \in [0, T] \times \mathbf{R}^n$ the optimal trajectory of (P) is unique
iv) For the Hamiltonian system (127) the set

$$M_t := \{(x(t), p(t)) \mid (x, p) \text{ solves (127) on } [t, T]\}$$

is the graph of a continuous function $\pi_t : \mathbf{R}^n \mapsto \mathbf{R}^n$.

Furthermore, iv) yields that $\pi_t(\cdot) = -\frac{\partial V}{\partial x}(t, \cdot)$ and every solution (x, p) of (127) restricted to $[t_0, T]$ satisfies: x is optimal for (P) with $x_0 = x(t_0)$ and $p(t) = -\frac{\partial V}{\partial x}(t, x(t))$ for all $t \in [0, T]$.

Before proving the above theorem, we shall state few corollaries.

Corollary 5.14 *Under all assumptions of Theorem 5.13, suppose U is a finite dimensional space, that for some $\overline{f} : [0, T] \times \mathbf{R}^n \mapsto \mathbf{R}^n$, $b : [0, T] \times \mathbf{R}^n \mapsto L(U, \mathbf{R}^n)$ we have*

$$\forall\ (t, x) \in [0, T] \times \mathbf{R}^n, \quad f(t, x, u) = \overline{f}(t, x) + b(t, x)u$$

and for every (t, x), $\frac{\partial L}{\partial u}(t, x, \cdot)$ is bijective. Then the (equivalent) statements $i) - iv)$ of Theorem 5.13 are equivalent to

v) For every $(t_0, x_0) \in [0,T] \times \mathbf{R}^n$ *there exists a unique optimal control* $\overline{u}(\cdot)$ *solving the problem (P). Furthermore, if* z *denotes the corresponding optimal trajectory, then for all* $t \in [t_0, T]$,

$$\overline{u}(t) = \left(\frac{\partial L}{\partial u}(t, z(t), \cdot)\right)^{-1} \left(-b(t, z(t))^\star \frac{\partial V}{\partial x}(t, z(t))\right)$$

The above corollary follows from *iii)* of Theorem 5.13 and the fact that $\overline{u}$ verifies

$$H(t, z(t), p(t)) = \langle p(t), z'(t)\rangle - L(t, z(t), \overline{u}(t)) \text{ a.e. in } [t_0, T]$$

where $p(\cdot)$ is the co-state of the maximum principle (see Theorem 5.11).

Our next corollary links results of Subsection 1 and Theorem 5.13.

Corollary 5.15 *Under all assumptions of Theorem 5.13, suppose that* $\nabla g(\cdot)$ *is locally Lipschitz,* $H(t, \cdot, \cdot)$ *is twice continuously differentiable and*

$$\forall\, r > 0,\ \exists\, k_r \in L^1(0,T),\ \frac{\partial H}{\partial(x,p)}(t, \cdot, \cdot) \text{ is } k_r(t) - \text{Lipschitz on } B_r(0)$$

Then the following two statements are equivalent:

i) $\forall\, t \in [0,T]$, $\frac{\partial V}{\partial x}(t, \cdot)$ *is locally Lipschitz*

ii) $\forall\, (x,p)$ *solving (127) on* $[0,T]$ *and every* $P_T \in \partial^\star(\nabla g)(x(T))$, *the matrix Riccati equation*

$$\begin{cases} P' + \dfrac{\partial^2 H}{\partial p \partial x}(t, x(t), p(t))P \ + \ P\, \dfrac{\partial^2 H}{\partial x \partial p}(t, x(t), p(t)) \ + \\ \\ + P\, \dfrac{\partial^2 H}{\partial p^2}(t, x(t), p(t))P \ + \ \dfrac{\partial^2 H}{\partial x^2}(t, x(t), p(t)) \ = \ 0 \\ \\ P(T) = P_T \end{cases}$$

has a solution on $[0,T]$.

Furthermore, if i) (or equivalently ii)) holds true, then

$$\nabla g \text{ is differentiable } \Longrightarrow \ \frac{\partial V}{\partial x}(t, \cdot) \text{ is differentiable}$$

and for every (x,p) *solving (127),* $P(t) = -\frac{\partial^2 V}{\partial x^2}(t, x(t))$. *If moreover* $g \in C^2$, *then* $V(t, \cdot) \in C^2$.

Proof — Let M_t be defined as in Theorem 5.13. If $i)$ holds true, then, by Theorem 5.13, M_t is the graph of a locally Lipschitz function π_t. By the maximum principle (Theorem 5.11), $\pi_t(\cdot) = -\frac{\partial V}{\partial x}(t,\cdot)$. Applying Theorem 5.3, we deduce $ii)$. Conversely, assume that $ii)$ is verified. Thus, by Theorem 5.3, M_t is the graph of a locally Lipschitz function from an open set $\mathcal{D}(t) \subset \mathbf{R}^n$ into $\mathbf{R}^n$. By the maximum principle, $M_t = \text{Graph}(-\frac{\partial V}{\partial x}(t,\cdot))$. Hence $i)$. The last statement follows from Theorem 5.3, because $P(t)$ describes the evolution of tangent space to M_t at $(x(t), p(t))$. $\diamond$

To prove Theorem 5.13 we need the following lemma.

Lemma 5.16 *Under all assumptions of Theorem 5.13 consider* $(t_0, x_0) \in]0, T[\times \mathbf{R}^n$ *such that* V *is differentiable at* (t_0, x_0). *Then*

$$-\frac{\partial V}{\partial t}(t_0, x_0) + H\left(t_0, x_0, -\frac{\partial V}{\partial x}(t_0, x_0)\right) = 0$$

i.e., V *satisfies the Hamilton-Jacobi-Bellman equation almost everywhere in* $[0, T] \times \mathbf{R}^n$.

Proof — Fix any $\overline{u} \in U$ and consider a solution x to

$$\begin{cases} x'(t) &= f(t, x, \overline{u}) \\ x(t_0) &= x_0 \end{cases}$$

Observe that for all small $h > 0$ it is defined on $[t_0, t_0 + h]$ and, by the very definition of the value function,

$$V(t_0 + h, x(t_0 + h)) + \int_{t_0}^{t_0+h} L(s, x(s), \overline{u})ds - V(t_0, x_0) \geq 0$$

Dividing by $h > 0$ and taking the limit we prove

$$\forall\, \overline{u} \in U, \quad \frac{\partial V}{\partial t}(t_0, x_0) + \left\langle \frac{\partial V}{\partial x}(t_0, x_0), f(t_0, x_0, \overline{u}) \right\rangle + L(t_0, x_0, \overline{u}) \geq 0$$

Consider next an optimal solution-control pair $(z, \overline{u})$ of the Bolza problem (P). Then

$$V(t_0 + h, z(t_0 + h)) + \int_{t_0}^{t_0+h} L(s, z(s), u)ds - V(t_0, x_0) = 0 \qquad (129)$$

By Theorem 5.11, $z(\cdot)$ solves the Hamiltonian system (127) with $x_T = z(T)$. Hence, by H_3), $z(\cdot) \in C^1$ $(z(t_0+h) - z(t_0))/h$ converge to some v when $h \to 0+$. By (129), for some $\sigma \in \mathbf{R}$,

$$\lim_{h\to 0+} \frac{1}{h} \int_{t_0}^{t_0+h} L(s, z(s), \overline{u}(s))ds = \sigma$$

By H_5)

$$(v, \sigma) \in \{(f(t_0, x_0, u), L(t_0, x_0, u) + r) \,|\, u \in U,\ r \geq 0\}$$

Thus for some $u_0 \in U$ and $r_0 > 0$

$$(v, \sigma) = (f(t_0, x_0, u_0), L(t_0, x_0, u_0) + r_0)$$

Dividing (129) by h and taking the limit yields

$$\frac{\partial V}{\partial t}(t_0, x_0) + \left\langle \frac{\partial V}{\partial x}(t_0, x_0), f(t_0, x_0, u_0) \right\rangle + L(t_0, x_0, u_0) + r_0 = 0$$

So we proved the existence of $u_0 \in U$ such that

$$\frac{\partial V}{\partial t}(t_0, x_0) + \left\langle \frac{\partial V}{\partial x}(t_0, x_0), f(t_0, x_0, u_0) \right\rangle + L(t_0, x_0, u_0) \leq 0$$

The two inequalities derived above imply the result. $\diamond$

Proof of Theorem 5.13 — Clearly *i*) $\Longrightarrow$ *ii*). Assume next that *ii*) holds true. Fix $0 \leq t_0 < T$, $x_0 \in \mathbf{R}^n$ and let $\overline{x}$ be an optimal solution to problem (P). Then, by Theorem 5.11, there exists $p(\cdot)$ such that $(\overline{x}, p)$ solves the system

$$\begin{cases} x'(t) &= \dfrac{\partial H}{\partial p}(t, x(t), p(t)), \quad x(t_0) = x_0 \\ -p'(t) &= \dfrac{\partial H}{\partial x}(t, x(t), p(t)), \quad p(t_0) = -\dfrac{\partial V}{\partial x}(t_0, x_0) \end{cases}$$

Since the solution to such system is unique, we deduce *iii*).

Conversely assume that *iii*) holds true. Fix $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$. Then, by Lemma 5.12, $\partial_x^\star V(t_0, x_0)$ is a singleton. We claim that the set

$$\partial^\star V(t_0, x_0) := \mathrm{Limsup}_{(t,x)\to(t_0,x_0)} \{\nabla V(t, x)\}$$

is a singleton.

Indeed let $(p_t, p_x) \in \partial^\star V(t_0, x_0)$ and $(t_i, x_i) \to (t_0, x_0)$ be such that $\nabla V(t_i, x_i) \to (p_t, p_x)$. Then $\{p_x\} = \partial_x^\star V(t_0, x_0)$ and, by Lemma 5.16, V satisfies at (t_i, x_i) the Hamilton-Jacobi-Bellman equation

$$-\frac{\partial V}{\partial t}(t_i, x_i) + H\left(t_i, x_i, -\frac{\partial V}{\partial x}(t_i, x_i)\right) = 0$$

Taking the limit we get

$$p_t \;=\; H(t, x, -p_x)$$

So p_t is uniquely defined and, thus, $\partial^\star V(t_0, x_0)$ is a singleton implying that V is differentiable at (t_0, x_0). Since $(t_0, x_0) \in [0, T] \times \mathbf{R}^n$ is arbitrary, we deduce that V is continuously differentiable on $[0, T] \times \mathbf{R}^n$. Hence we proved $iii) \Longrightarrow i)$.

Assume next that $iv)$ holds true. Fix $t_0 \in [0, T]$ and $x_0 \in \mathbf{R}^n$. By Lemma 5.12,

$$(x_0,\ \partial_x^\star V(t_0, x_0)) \;\subset\; \mathrm{Graph}(\pi_{t_0})$$

Thus $\partial_x^\star V(t_0, x_0)$ is a singleton. In particular

$$\mathrm{Limsup}_{x \to x_0} \left\{\frac{\partial V}{\partial x}(t_0, x)\right\} \quad \text{is a singleton}$$

and therefore, $ii)$ is verified.

It remains to show that $ii)$ yields $iv)$. For this aim fix $t_0 \in [0, T]$ and define the continuous mapping $\Psi : \mathbf{R}^n \mapsto \mathbf{R}^n$ in the following way:

For all $x_0 \in \mathbf{R}^n$ consider the solution (x, p) to the system

$$\begin{cases} x'(t) \;=\; \dfrac{\partial H}{\partial p}(t, x(t), p(t)), \quad x(t_0) = x_0 \\ \\ -p'(t) \;=\; \dfrac{\partial H}{\partial x}(t, x(t), p(t)), \quad p(t_0) = -\dfrac{\partial V}{\partial x}(t_0, x_0) \end{cases}$$

and set $\Psi(x_0) = x(T)$. By the maximum principle (Theorem 5.11) we know that $p(T) = -\nabla g(x(T))$. Thus $(x(T), p(T)) \in \mathrm{Graph}(-\nabla g)$. In particular this yields Ψ is one-one. By the Invariance of Domain Theorem, $\Psi(\mathbf{R}^n)$ is open. Thus the set

$$\{(\Psi(x_0), -\nabla g(\Psi(x_0))) \,|\, x_0 \in \mathbf{R}^n\} \text{ is open and closed in } \mathrm{Graph}(-\nabla g)$$

So it coincides with $\mathrm{Graph}(-\nabla g)$. Hence, by uniqueness of solution to the Hamiltonian system (127), $M_t = \mathrm{Graph}(-\frac{\partial V}{\partial x}(t_0, \cdot))$. The proof is complete.

5.3.4 Problems with Concave-Convex Hamiltonians

Observe that in general one has

$$\frac{\partial^2 H}{\partial p^2}\,(t,x(t),p(t)) \;\geq\; 0$$

for every solution (x,p) of the Hamiltonian system

$$\begin{cases} x'(t) &= \dfrac{\partial H}{\partial p}(t,x(t),p(t)), \quad x(T)=x_T \\ \\ -p'(t) &= \dfrac{\partial H}{\partial x}(t,x(t),p(t)), \quad p(T)=-\nabla g(x_T) \end{cases} \tag{130}$$

and that whenever in addition $H(t,\cdot,p(t))$ is concave for all $t\in[0,T]$, then

$$\frac{\partial^2 H}{\partial x^2}\,(t,x(t),p(t)) \;\leq\; 0$$

If g is convex, then every matrix from the generalized Jacobian $\partial^\star g(x(T))$ is nonnegative. From Corollary 5.9 we deduce that for every $P_T\in\partial^\star(\nabla g)(x(T))$, the solution $P(\cdot)$ of the matrix Riccati equation

$$\begin{cases} P' + \dfrac{\partial^2 H}{\partial p\partial x}(t,x(t),p(t))P \;+\; P\,\dfrac{\partial^2 H}{\partial x\partial p}(t,x(t),p(t)) + \\ \\ +\,P\,\dfrac{\partial^2 H}{\partial p^2}(t,x(t),p(t))P \;+\; \dfrac{\partial^2 H}{\partial x^2}(t,x(t),p(t)) \;=\; 0 \\ \\ P(T)=-P_T \end{cases} \tag{131}$$

exists on $[0,T]$. By Theorem 5.3, no shocks of (130) can occur backward in time. Hence we deduce from Theorem 5.13 and Corollary 5.15 the following results.

Theorem 5.17 *Assume $H_1)-H_5)$, that V is locally Lipschitz and for every $(t_0,x_0)\in[0,T]\times\mathbf{R}^n$ the problem (P) has an optimal solution.*

Further assume that $\nabla g(\cdot)$ is locally Lipschitz, $H(t,\cdot,\cdot)$ is twice continuously differentiable and

$$\forall\, r>0,\ \exists\, k_r\in L^1(0,T),\ \frac{\partial H}{\partial(x,p)}(t,\cdot,\cdot) \text{ is } k_r(t)-\text{Lipschitz on } B_r(0)$$

If for every solution (x,p) of (130), $H(t,\cdot,p(t))$ is concave and g is convex, then $V \in C^1$ and $\frac{\partial V}{\partial x}(t,\cdot)$ is locally Lipschitz.

Moreover, every solution (x,p) of (130) is an optimal trajectory-co-state pair. If in addition $g \in C^2$, then $V(t,\cdot) \in C^2$ and, in this case, $P(t) = -\frac{\partial^2 V}{\partial x^2}(t,x(t))$ solves the matrix Riccati equation (131) with $P_T = -g''(x(T))$.

6 Hamilton-Jacobi-Bellman Equation for Problems under State-Constraints

Consider the optimal control problem

$$(\mathrm{P}) \qquad \begin{cases} \text{Minimize } g(x(1)) \\ \text{over } x \in W^{1,1}([0,1];\mathbf{R}^n) \text{ satisfying} \\ x'(t) \in F(t,x(t)) \quad \text{a.e. } t \in [0,1], \\ x(t) \in K \quad \forall\, t \in [0,1], \\ x(0) = x_0, \end{cases}$$

the data for which comprise: a function $g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, a set-valued map $F : [0,1] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$, a closed set $K \subset \mathbf{R}^n$ and a vector $x_0 \in \mathbf{R}^n$. Solutions of the above differential inclusion satisfying the constraints of (P), are called *feasible* arcs (for (P)).

Note that, since g is extended valued, (P) incorporates the endpoint constraint:

$$x(1) \in C$$

where $C := \operatorname{dom} g$.

Denote by $V : [0,1] \times K \to \mathbf{R} \cup \{+\infty\}$ the value function for (P): for each $(t,x) \in [0,1] \times K$, $V(t,x)$ is defined to be the infimum cost for the problem

$$(P_{t,x}) \qquad \begin{cases} \text{Minimize } g(y(1)) \\ \text{over } y \in W^{1,1}([t,1];\mathbf{R}^n) \text{ satisfying} \\ y'(s) \in F(s,y(s)) \quad \text{a.e. } s \in [t,1], \\ y(s) \in K \quad \forall\, s \in [t,1], \\ y(t) = x. \end{cases}$$

Thus

$$V(t,x) = \inf(P_{t,x}).$$

(If $(\mathrm{P}_{t,x})$ has no feasible arcs, we set $V(t,x) = +\infty$.)

In this section we explore the relationship between the value function and the Hamilton-Jacobi Equation:

$$(HJE)\begin{cases} -\frac{\partial V}{\partial t} + H(t,x,-\frac{\partial V}{\partial x}) \;=\; 0 \text{ for } (t,x) \in (0,1)\times \text{int}K \\ V(1,x) \;=\; g(x) \text{ for } x \in K. \end{cases}$$

To get uniqueness of solutions to the above PDE in the constrained case we are lead to impose some kind of constraint qualification on the dynamic constraint at boundary points of the state constraint set.

In [12], Capuzzo-Dolcetta and Lions showed that the value function is continuous and is the unique viscosity solution to (HJE) under hypotheses which include the "inward-pointing" constraint qualification:

$$\min_{v\in F(t,x)} n_x \cdot v < 0 \quad \forall\, x \in \text{bdy}\, K$$

where n_x denotes the unit outward normal vector at the point $x \in \text{bdy}\, K$. Hypotheses of this nature were introduced by Soner [51] to ensure continuity of the value function and to provide a characterization of the value function in terms of viscosity solutions of the relevant Hamilton-Jacobi equation, for an infinite horizon problem.

When the "inward pointing" constraint qualification fails, or when the terminal cost function g is chosen to take account of an endpoint constraint, we can expect that the value function will be discontinuous.

We restrict attention to a special class of state constraints sets, namely a finite intersection of smooth manifolds. (Nonetheless, this is a framework which allows state constraints sets with non smooth boundaries, and covers state constraints encountered in most engineering applications.) A key element is an extension of Filippov's theorem to the constrained case.

6.1 Constrained Hamilton-Jacobi-Bellman Equation

The following theorem provides two characterizations of the value function for optimal control problems with endpoint and state constraints, in terms of lower semicontinuous solutions of the Hamilton-Jacobi equation and one in terms of epiderivative solutions.

It is assumed that the state constraint set K is expressible as

$$K \;=\; \cap_{j=1}^{r}\{x : h_j(x) \le 0\}$$

for a finite family of $C^{1,1}$ functions $\{h_j : \mathbf{R}^n \to \mathbf{R}\}_{j=1}^r$. ($C^{1,1}$ denotes the class of C^1 functions with locally Lipschitz continuous gradients.) The "active set" of index values $I(x)$, at a point $x \in \operatorname{bdy} K$, is

$$I(x) \quad := \quad \{j \in (1, \ldots, r) : h_j(x) = 0\}.$$

Recall the notations $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$ for all real numbers a, b. We write

$$h^+(x) := \left(\max_{j=1,2,\ldots,r} h_j(x) \right) \vee 0.$$

$W^{1,1}([a, b]; \mathbf{R}^n)$ denotes the space of absolutely continuous n-vector valued functions on $[a, b]$, with norm

$$||x||_{W^{1,1}} = \|x(a)\| + \int_a^b \|x'(t)\| dt.$$

Theorem 6.1 *Take a function $V : [0, 1] \times K \to \mathbf{R} \cup \{+\infty\}$. Assume that the following hypotheses are satisfied:*

(H1) F is a continuous set-valued map, which takes values in the space of non-empty, closed, convex sets,

(H2) There exists $c > 0$ such that

$$F(t, x) \subset c(1 + \|x\|)B \quad \forall\, (t, x) \in [0, 1] \times \mathbf{R}^n,$$

(H3) There exists $k \in L^1$ such that

$$F(t, x) \subset F(t, x_1)) + k(t)\|x - x_1\|B \quad \forall\, t \in [0, 1],\ x, x_1 \in \mathbf{R}^n \times \mathbf{R}^n,$$

(H4) g is lower semicontinuous.

Assume furthermore that

(CQ) For all $x \in K$ and $t \in [0, 1]$ there exists $v \in F(t, x)$ such that

$$\forall j \in I(x), \quad \nabla h_j(x) \cdot v > 0.$$

Then assertions (a)-(c) below are equivalent:

(a) V is the value function for (P).

(b) V is lower semicontinuous and

(i) $\forall\ (t,x) \in ([0,1[\times K) \cap \operatorname{dom} V$

$$\inf_{v\in F(t,x)} D_{\uparrow}V(t,x)(1,v) \ \leq\ 0$$

(ii) $\forall\ (t,x) \in]0,1] \times \operatorname{int} K) \cap \operatorname{dom} V$

$$\sup_{v\in F(t,x)} D_{\uparrow}V(t,x)(-1,-v) \ \leq\ 0$$

(iii) $\forall\ x \in K$

$$\liminf_{\{(t',x')\to(1,x):t'<1,x'\in \operatorname{int} K\}} V(t',x') = V(1,x) = g(x)$$

(c) V is lower semicontinuous and

(i) $\forall\ (t,x) \in (]0,1[\times \operatorname{int} K) \cap \operatorname{dom} V,\ (p_t,p_x) \in \partial_- V(t,x)$

$$-p_t + H(t,x,-p_x) = 0.$$

(ii) $\forall\ (t,x) \in (]0,1[\times \operatorname{bdy} K) \cap \operatorname{dom} V,\ (p_t,p_x) \in \partial_- V(t,x)$

$$-p_t + H(t,x,-p_x) \geq 0$$

(iii) $\forall\ x \in K,$

$$\liminf_{\{(t',x')\to(0,x):t'>0\}} V(t',x') = V(0,x)$$

and

$$\liminf_{\{(t',x')\to(1,x):t'<1,x'\in \operatorname{int} K\}} V(t',x') = V(1,x) = g(x).$$

Example Consider the problem

$$\begin{cases} \text{Minimize } g(x(1)) \\ x'(t) \in F(t,x(t)) \\ x(t) \in K \\ x(0) = x_0, \end{cases}$$

in which $n = 1$, $g(x) = x$, $F(t,x) = \{1\}$, $K = \{x : x \leq 0\}$, $x_0 = 0$.

By inspection

$$V(t,x) = \begin{cases} +\infty & \text{if } x > -(1-t) \\ x + (1-t) & \text{if } x \leq -(1-t) \end{cases}$$

The hypotheses for application of Thm. 6.1 are satisfied, including the outward-pointing constraint qualification (CQ). Thm. 6.1 therefore tells us that V is the unique solution of (HJE) (in the sense specified).

Notice that $V(t,x) = +\infty$ at some points in $[0,1] \times K$, despite the fact that g is everywhere finite valued (no endpoint constraints).

6.2 A Neighboring Feasible Trajectories Theorem

A key role in the proof of the Main Theorem is played by an estimate governing the distance of the set of trajectories satisfying a given state constraint from a given trajectory which violates the constraint. This estimate is provided by the following Existence of Feasible Neighboring Trajectories (EFNT) Theorem, which can be regarded as a kind of refined viability theorem, in which the information that a 'viable' solution exists whenever viability condition holds true is supplemented by information about where it is located, in relation to a given solution when a 'strict' viability condition holds true.

As before, we limit attention to state constraint sets K associated with a family of functional inequalities:

$$K = \cap_{j=1}^{r} \{x : h_j(x) \leq 0\},$$

in which the h_j's are given $C^{1,1}$ functions.

Theorem 6.2 *Fix $r_0 > 0$. Assume that for some $c > 0$, $\alpha > 0$ and $k(\cdot) \in L^1$, the following hypotheses are satisfied:*

(i) F takes values in the space of non-empty, closed sets and $F(\cdot, x)$ is measurable for each $x \in \mathbf{R}^n$.

(ii) $F(t,x) \subset c(1 + \|x\|)B \quad \forall\, (t,x) \in [0,1] \times \mathbf{R}^n$.

(iii) $F(t,x) \subset F(t,x')) + k(t)\|x - x'\|B \quad \forall\, t \in [0,1],\ x, x' \in \mathbf{R}^n$.

Assume furthermore that there exists some $\alpha > 0$ such that

$$(CQ)' \qquad \min_{v\in F(t,x)} \max_{j\in I(x)} \nabla h_j(x)\cdot v < -\alpha \quad ,$$
$$x \in B(0, e^c(r_0+c)) \cap \operatorname{bdy} K,\ t\in[0,1].$$

Then there exists a constant ϑ (which depends on r_0, c, α and $k \in L^1$) with the following property: given any $t_0 \in [0,1]$ and any $\hat{x} \in \mathcal{S}_{[t_0,1]}$ such that $\hat{x}(t_0) \in B(0,r_0)\cap K$, an $x \in \mathcal{S}_{[t_0,1]}(\hat{x}(t_0))$ can be found such that

$$x(t) \in K \quad \forall\, t \in [t_0,1]$$

and

$$||x-\hat{x}||_{W^{1,1}([t_0,1];\mathbf{R}^n)} \le \vartheta \max_{t\in[t_0,1]} h^+(\hat{x}(t)).$$

The need to introduce into $(\mathrm{CQ})'$ the positive parameter α arises because it is not hypothesized that F is a continuous multifunction. In the case F is continuous, then $(\mathrm{CQ})'$ is implied by the condition

$$\min_{v\in F(t,x)} \max_{j\in I(x)} \nabla h_j(x)\cdot v < 0 \quad \forall\, x \in B(0,e^c(r_0+c))\cap \operatorname{bdy} K,\ t\in[0,1].$$

Proof. Set $R = e^c(r_0+c)$ and $c_0 = c(1+R)$. Let k_h be a common Lipschitz constant for the h_j's on $B(0,R)$ and let κ be a common Lipschitz constant for the ∇h_j's on $B(0,R)$.

Note that for any $[t',t''] \subset [0,1]$ and any solution $y:[t',t''] \to \mathbf{R}^n$ to our differential inclusion such that $y(t') \in B(0,r_0)$, we have $y(t) \in B(0,R)$. This follows from Gronwall's lemma.

Let $\omega : R^+ \to \mathbf{R}^n$ be a modulus of continuity for $t \to \int_0^t k(s)ds$, i.e. ω is monotone increasing, $\lim_{\sigma\downarrow 0}\omega(\sigma) = 0$, and

$$\omega(t'-t) \ge \int_t^{t'} k(s)ds \quad \forall\, [t,t'] \subset [0,1].$$

Define

$$I_\beta(\xi) \ := \ \{j \in \{1,\dots,r\} : 0 \ge h_j(\xi) \ge -\beta\}.$$

Under the hypotheses, there exists $\beta > 0$ and $\alpha > 0$ such that

$$\forall \xi \in K\cap B(0,R), t\in[0,1] \qquad \min_{v\in F(t,\xi)} \max_{j\in I_\beta(\xi)} \nabla h_j(\xi)\cdot v \ < \ -\alpha.$$

Fix $\alpha' \in (0, \alpha)$. Choose $\eta \in (0,1)$ such that $N := \eta^{-1}$ is an integer and the following conditions are satisfied:

$$\eta < (\alpha - \alpha')(c_0^2 \kappa)^{-1} \tag{132}$$

$$\omega(\eta) < \log(\frac{\alpha - \alpha'}{8 c_0 k_h} + 1) \tag{133}$$

$$\eta(k_h c_0 + \alpha') < \beta, \tag{134}$$

and

$$6\frac{c_0^2}{\alpha'}\kappa e^{\omega(\eta)}\eta + 6\frac{c_0}{\alpha'}k_h(e^{\omega(\eta)} - 1)2c_0 3/\alpha' < 1. \tag{135}$$

Set

$$\vartheta' := \max\left\{\frac{6c_0}{\alpha'}\exp(\int_0^1 k(s)ds), \frac{6}{\alpha'\eta}c_0\right\}. \tag{136}$$

Step 1: We show that, for every $\xi \in K$ and $\tau \in (0, 1-\eta]$ there exists a solution $\tilde{x} : [\tau, \tau+\eta] \to \mathbf{R}^n$ such that $\tilde{x}(\tau) = \xi$ and

$$h_j(\tilde{x}(t)) \leq -\alpha'(t-\tau) \quad \forall\, j,\ \forall t \in [\tau, \tau+\eta].$$

Fix $\xi \in K$ and $\tau \in (0, 1-\eta]$. Consider a measurable function $v : [\tau, \tau+\eta] \to \mathbf{R}^n$ such that $v(t) \in F(t, \xi)$ a.e. $t \in [\tau, \tau+\eta]$ and

$$j \in I_\beta(\xi) \text{ implies } \nabla h_j(\xi) \cdot v(t) < -\alpha \text{ for a.e. } t \in [\tau, \tau+\eta].$$

Set $z(t) = \xi + \int_\tau^t v(s)ds$. We have, for all $j \in I_\beta(\xi)$ and $t \in [\tau, \tau+\eta]$,

$$\begin{aligned}
h_j(z(t)) &= h_j(\xi) + \int_\tau^t \nabla h_j(z(s)) \cdot v(s) ds \\
&\leq 0 + \int_\tau^t \nabla h_j(z(s)) \cdot v(s) ds \\
&\leq \int_\tau^t \nabla h_j(\xi) \cdot v(s) ds + \int_\tau^t \|\nabla h_j(z(t)) - \nabla h_j(\xi)\| \cdot \|v(s)\| ds \\
&\leq -\alpha(t-\tau) + \kappa c_0^2 (t-\tau)^2/2 \\
&\leq (-\alpha + (\alpha - \alpha')/2)(t-\tau) \leq -\left(\frac{\alpha + \alpha'}{2}\right)(t-\tau).
\end{aligned}$$

(We have used (132).)

Fix $j \in I_\beta(\xi)$. By Filippov's Theorem, applied to the reference trajectory z, there exists a solution $\tilde{x} : [\tau, \tau+\eta] \to \mathbf{R}^n$ such that $\tilde{x}(\tau) = \xi$ and, for all $t \in [\tau, \tau+\eta]$,

$$\begin{aligned}
\|\tilde{x}(t) - z(t)\| &\leq \int_\tau^t d_{F(s,z(s))}(z'(s)) \exp\left(\int_s^t k(\sigma)d\sigma\right) ds \\
&\leq c_0(t-\tau)\int_\tau^t k(s)\exp\left(\int_s^t k(\sigma)d\sigma\right) ds \\
&\leq c_0(t-\tau)(\exp\left(\int_\tau^t k(s)ds\right) - 1) \\
&\leq c_0(t-\tau)\left(e^{\omega(t-\tau)} - 1\right) \leq \frac{\alpha-\alpha'}{8k_h}(t-\tau).
\end{aligned}$$

(We have used (133).) But then, for all $t \in [\tau, \tau+\eta]$,

$$\begin{aligned}
h_j(\tilde{x}(t)) &\leq k_h\|\tilde{x}(t) - z(t)\| + h_j(z(t)) \\
&\leq \left(\frac{\alpha-\alpha'}{8} - \frac{\alpha+\alpha'}{2}\right)(t-\tau) \leq -\alpha'(t-\tau).
\end{aligned}$$

On the other hand, if $h_j(\xi) \leq -\beta$, then, for all $t \in [\tau, \tau+\eta]$,

$$h_j(\tilde{x}(t)) \leq -\beta + (t-\tau)k_h c_0 \leq -\alpha'\eta \leq -\alpha'(t-\tau).$$

We see that, in this case too, the inequality is satisfied. Step 1 is complete.

Step 2: Take any $\tau \in [0, 1-\eta]$ and any solution $x_1 : [0,1] \to \mathbf{R}^n$ such that $x_1(t) \in K$ for all $t \in [0, \tau]$. We shall show that there exists an solution $x_2 : [0,1] \to \mathbf{R}^n$ such that $x_2(t) = x_1(t)$ for all $t \in [0, \tau]$,

$$x_2(t) \in K \quad \forall\, t \in [\tau, \tau+\eta]$$

and

$$||x_1 - x_2||_{W^{1,1}([0,1];\mathbf{R}^n)} \leq \vartheta' \max_{t\in[0,1]} h^+(x_1(t)).$$

Set

$$\Delta = \max_{t\in[0,1]} h^+(x_1(t)).$$

Suppose that $\Delta \geq \alpha'\eta/3$. By Step 1, there exists a solution $x_1 : [0, \tau+\eta] \to \mathbf{R}^n$ such that $x_2(t) = x_1(t)$ for all $t \in [0,\tau]$ and $x_2(t) \in K$ for $t \in [0, \tau+\eta]$. We have

$$||x_1 - x_2||_{W^{1,1}([0,1];\mathbf{R}^n)} \leq 2c_0 \leq \vartheta'\left(\frac{\alpha'\eta}{3}\right) \leq \vartheta'\Delta,$$

by (136), as required. We can therefore assume that

$$\Delta \ < \ \alpha'\eta/3.$$

Set

$$\eta' = 3\Delta/\alpha'.$$

Notice that $\eta' \leq \eta$ and $\tau + \eta' \leq 1$. By the results of Step 1, there exists a solution $z : [0, \tau + \eta'] \to \mathbf{R}^n$ such that $z(t) = x_1(t)$ for all $t \in [0, \tau]$,

$$z(t) \in K \quad \forall\, t \in [0, \tau + \eta']$$

and

$$h_j(z(\tau + \eta')) \leq -\alpha'\eta' = -3\Delta \quad \forall j.$$

By Filippov's Theorem, there exists a solution $y : [\tau + \eta', 1] \to \mathbf{R}^n$ such that $y(\tau + \eta') = z(\tau + \eta')$ and, for all $t \in [\tau + \eta', 1]$,

$$\begin{aligned} \|y(t) - x_1(t)\| &\leq \exp\left(\int_{\tau+\eta'}^{t} k(s)ds\right) \|z(\tau + \eta') - x_1(\tau + \eta')\| \\ &\leq \exp\left(\int_{\tau+\eta'}^{t} k(s)ds\right) 2c_0\eta', \qquad (137) \\ \|y'(t) - x_1'(t)\| &\leq k(t)\exp\left(\int_{\tau+\eta'}^{t} k(s)ds\right) \|z(\tau + \eta') - x_1(\tau + \eta')\| \\ &\leq k(t)\exp\left(\int_{\tau+\eta'}^{t} k(s)ds\right) 2c_0\eta'. \qquad (138) \end{aligned}$$

Now concatenate $z : [0, \tau + \eta'] \to \mathbf{R}^n$ and $y : [\tau + \eta', 1] \to \mathbf{R}^n$ to form the solution $x_2 : [0, 1] \to \mathbf{R}^n$.

Since $x_1(0) = x_2(0)$,

$$\begin{aligned} ||x_1 - x_2||_{W^{1,1}} &= ||x_1' - x_2'||_{L^1([0,\tau+\eta'];\mathbf{R}^n)} + ||x_1' - x_2'||_{L^1([\tau+\eta',1];\mathbf{R}^n)} \\ &\leq 2c_0\eta' + 2c_0\eta'(\exp\left(\int_{\tau+\eta'}^{1}(k(s)ds\right) - 1) \\ &= 2c_0\eta'\exp\left(\int_{\tau+\eta'}^{1} k(s)ds\right) \\ &= (6c_0/\alpha')\exp\left(\int_{\tau+\eta'}^{1} k(s)ds\right)\Delta \leq \vartheta'\Delta. \end{aligned}$$

(We have used (138).) It remains to show that

$$x_2(t) \in K \text{ for all } t \in [0, \tau + \eta].$$

The condition is clearly satisfied for any $t \in [0, \tau + \eta']$. On the other hand, for any $t \in [\tau + \eta', \tau + \eta]$ and j, we have from (137) and (138)

$$\begin{aligned} h_j(x_2(t)) &= h_j(x_2(\tau+\eta')) + \int_{\tau+\eta'}^{t} \nabla h_j(x_2(s)) \cdot x_2'(s) ds \\ &= h_j(x_2(\tau+\eta')) + \int_{\tau+\eta'}^{t} \nabla h_j(x_1(s)) \cdot x_1'(s) ds \\ &\quad + \int_{\tau+\eta'}^{t} (\nabla h_j(x_2(s)) - \nabla h_j(x_1(s))) \cdot x_1'(s)) ds \\ &\quad + \int_{\tau+\eta'}^{t} \nabla h_j(x_2(s)) \cdot (x_2'(s) - x_1'(s)) ds \\ &\leq -3\Delta + 2\Delta + (6c_0^2\kappa/\alpha')\eta e^{\omega(\eta)}\Delta + k_h\left(e^{\omega(\eta)} - 1\right) 2c_0(3/\alpha')\Delta \\ &\leq \left(-1 + (6c_0^2\kappa/\alpha')\eta e^{\omega(\eta)} + k_h(e^{\omega(\epsilon')-1})6c_0/\alpha'\right)\Delta \;\leq\; 0. \end{aligned}$$

(We have used (135).) Step 2 is complete.

Take any solution $\hat{x} : [0,1] \to \mathbf{R}^n$ such that $\hat{x}(0) \in B(0, r_0)$. Recall that $N = \eta^{-1}$ is an integer.

Set $x_0 = \hat{x}$. Use the results of Step 2 recursively to generate a finite sequence of solutions $x_0, \ldots, x_N$ on $[0,1]$ with the properties

$$x_i(t) \in K \quad \forall\, t \in [0, i/N]$$

and

$$||x_i - x_{i-1}||_{W^{1,1}([0,1];\mathbf{R}^n)} \;\leq\; \vartheta' \max_{t\in[0,1]} h^+(x_{i-1}(t)).$$

Now set $x = x_N$. Clearly

$$x(t) \in K \quad \forall\, t \in [0,1].$$

It is a routine exercise to show, using the results of Step 2 and the triangle inequality, that

$$||x - \hat{x}||_{W^{1,1}([t_0,1];\mathbf{R}^n)} \;\leq\; \vartheta \max_{t\in[t_0,1]} h^+(\hat{x}(t)), \tag{139}$$

in which $t_0 = 0$ and

$$\vartheta \;:=\; k_h^{-1}[(1 + k_h\vartheta')^N - 1].$$

A special case of the theorem has been proved, in which $t_0 = 0$.

Take any $t_0 \in [0,1]$. Suppose that $\hat{x} : [t_0, 1] \to \mathbf{R}^n$ is a solution such that $\hat{x}(t_0) \in K \cap B(0, r_0)$. Define

$$\bar{F}(t,x) \;=\; \begin{cases} F(t,x) & t \geq t_0 \\ F(t_0,x) \cup \{0\} & t < t_0. \end{cases}$$

and

$$\hat{y}(t) \;=\; \begin{cases} \hat{x}(t) & t \geq t_0 \\ \hat{x}(t_0) & t < t_0. \end{cases}$$

Now apply the earlier construction to $\bar{F}$ and $\hat{y}$ to obtain a solution $x_0 : [0,1] \to \mathbf{R}^n$ to $y' \in \bar{F}(t,y)$. It is a simple matter to check that the solution x obtained by restricting x_0 to $[t_0, 1]$ satisfies $x(t_0) = \hat{x}(t_0)$, $x(t) \in K$ for all $t \in [t_0, 1]$ and also inequality (139) (with the same constant ϑ). The theorem is proved.

6.3 Proof of the Main Theorem

We isolate in the following lemma the steps in the proof of Thm. 6.1 requiring the constraint qualification (CQ).

Lemma 6.3 *(i) Take any point $x_1 \in K$. Then there exists $\delta \in]0,1[$ and a solution $y : [1-\delta, 1] \to \mathbf{R}^n$ such that $y(1) = x_1$ and*

$$y(t) \in \operatorname{int} K \quad \forall\, t \in [1-\delta, 1[.$$

(ii) Take any $t_0 \in [0,1[$ and any solution $x : [t_0, 1] \to K$. Take also a sequence of points $\{(\tau_i, \xi_i)\}$ in $[t_0, 1[\times \operatorname{int} K$ such that $(\tau_i, \xi_i) \to (1, x(1))$. Then there exists a sequence of solutions $\{x_i : [t_0, \tau_i] \to \mathbf{R}^n\}$ such that $x_i(\tau_i) = \xi_i$

$$x_i(t) \in \operatorname{int} K \quad \forall\, t \in [t_0, \tau_i], i = 1, 2, \ldots$$

and

$$||x_i - x||_{L^\infty([t_0,\tau_i];\mathbf{R}^n)} \to 0 \text{ as } i \to \infty.$$

Proof. According to (CQ), there exists $v \in F(1, x_1)$ and $\alpha > 0$ such that

$$\nabla h_j(x_1) \cdot v > \alpha \quad \forall\, j \in I(x_1).$$

For some $\delta \in]0, 1 - t_0]$, whose magnitude will be set presently, define

$$z(t) = x_1 - (1-t)v \quad \text{for } t \in [1-\delta, 1].$$

By Filippov's Theorem, there exists a solution $x : [1-\delta, 1] \to \mathbf{R}^n$ such that $x(1) = x_1$ and

$$\|x(t) - z(t)\| \leq \exp\{\int_0^1 k(t)dt\} \int_t^1 d_{F(s,z(s))}(v)ds$$

for all $t \in [1-\delta, 1]$. We deduce from the continuity of $(t,x) \hookrightarrow F(t,x)$ and the continuous differentiability of the h_j's that there exists a function $\eta : R^+ \to R^+$ such that $\eta(\theta) \downarrow 0$ as $\theta \downarrow 0$,

$$\|x(1-s) - (x_1 - sv)\| \leq \eta(s)s \quad \text{for } s \in [0, \delta]$$

and

$$h_j(x(1-s)) \leq h_j(x_1) + \nabla h_j(x_1) \cdot (x(1-s) - x_1) + \eta(s)s$$

for all $s \in [0, \delta]$ and $j \in I(x_1)$. But then, since $h_j(x_1) = 0$ for all $j \in I(x_1)$, there exists M (M does not depend on s) such that

$$h_j(x(1-s)) \leq -s\nabla h_j(x_1) \cdot v + M\eta(s)s$$

for all $j \in I(x_1)$. Hence

$$s^{-1}h_j(x(1-s)) \leq -\alpha + M\eta(s) \quad \forall\, s \in [0, \delta[,\ j \in I(x_1).$$

It follows that, if we now choose δ such that $M\eta(\delta) < \alpha$, then $h_j(x(t)) < 0$ for all $j \in I(x_1)$. Since $h_j(x_1) < 0$ for all $j \notin I(x_1)$, we can arrange, by a further reduction in the size of δ, that

$$\max_{j \in \{1,\ldots,r\}} h_j(x(t)) < 0 \quad \forall\, t \in [1-\delta, 1[.$$

(ii) Define the sequence of positive numbers

$$\gamma_i := \left(-\max_{j=1,\ldots,r} h_j(\xi_i)\right) \wedge (i^{-1}) \quad \text{for } i = 1, 2, \ldots$$

Since $\{\xi_i\} \subset \operatorname{int} K$ and (CQ) holds true, it follows that $\gamma_i > 0$ for all i. Clearly $\gamma_i \downarrow 0$. For each i define

$$h_j^i(x) := h_j(x) + \gamma_i.$$

Apply Filippov's Theorem to $x' \in F(t,x)$, taking as reference trajectory x restricted to $[t_0, \tau_i]$. This yields a solution $y_i : [t_0, \tau_i] \to \mathbf{R}^n$ satisfying $y_i(\tau_i) = \xi_i$ and

$$||y_i - x||_{L^\infty([t_0,\tau_i];\mathbf{R}^n)} \leq \exp\{\int_0^1 k(t)dt\}\|x(\tau_i) - \xi_i\|.$$

Since $(x(\tau_i) - \xi) \to 0$ as $i \to \infty$, we can conclude that

$$||y_i - x||_{L^\infty([t_0,\tau_i])} \to 0 \quad \text{as } i \to \infty. \tag{140}$$

By the comments following the statement of Theorem 6.2, (CQ) yields (CQ)$'$. So we deduce from Thm.6.2 applied to the set-valued map $-F$ that there exists $\vartheta > 0$ and a sequence of solutions $\{x_i : [t_0, \tau_i] \to \mathbf{R}^n\}$ such that $x_i(\tau_i) = \xi_i$ and

$$||y_i - x_i||_{L^\infty([t_0,\tau_i];\mathbf{R}^n)} \leq \vartheta \left[\max_{t\in[t_0,\tau_i]} \max_j h_j(y_i(t)) + \gamma_i \right]^+$$

$$h_j(x_i(t)) + \gamma_i \leq 0 \quad \forall\, t \in [t_0, \tau_i],\ j \in I(x_i(t)),\ i = 1, 2, \ldots$$

This means that

$$x_i(t) \in \operatorname{int} K \quad \forall\, t \in [t_0, \tau_i],\ i = 1, 2, \ldots$$

Since $h_j(x(t)) \leq 0$ for all $t \in [0, 1]$, we deduce from (140) that

$$||x_i - x||_{L^\infty([t_0,\tau_i])} \to 0 \quad \text{as } i \to \infty.$$

In the next lemma, reference is made to the δ-tube about $\bar{x} : [t_0, t_1] \to \mathbf{R}^n$:

$$T_\delta(\bar{x}) \ := \ \{(t, x) \in [t_0, t_1] \times \mathbf{R}^n : \|x - \bar{x}(t)\| < \delta\}.$$

Lemma 6.4 *Take $[t_0, t_1] \subset [0, 1]$ such that $t_0 < t_1$, a solution $\bar{x} : [t_0, t_1] \to \mathbf{R}^n$, $\delta > 0$ and a lower semicontinuous function $V : [t_0, t_1] \times \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ such that for all $(t, x) \in T_\delta(\bar{x})$ with $t < t_1$,*

$$\forall\, (p_t, p_x) \in \partial_- V(t, x),\ -p_t + H(t, x, -p_x) \leq 0 \tag{141}$$

Then, for any $t_0 \leq t' \leq t'' < t_1$,

$$V(t', \bar{x}(t')) \leq V(t'', \bar{x}(t'')).$$

Proof. We deduce in the same way as in Section 4, proofs of Theorems 4.7 and 4.9 that

$$V(t', \bar{x}(t')) \leq V(t'', \bar{x}(t'')).$$

The fact that $t'' < t_1$ (strict inequality) is important here, since no regularity hypotheses have been imposed on $t \to V(t,\cdot)$ at $t = t_1$.

Proof of Thm. 6.1

(a) $\Rightarrow$ (b). The value function V is lower semicontinuous by the same arguments as those used in Section 4.

Under the hypotheses, $(t,x) \in \operatorname{dom} V$ implies that $(\mathrm{P}_{t,x})$ has a solution. It is a straightforward matter to show that, if y is a minimizer for $(\mathrm{P}_{t,x})$, then $s \to V(s,y(s))$ is constant on $[t,1]$; b(i) can be deduced from this property.

It can also be shown that, if $y : [t,1] \to \mathbf{R}^n$ is a solution satisfying the constraints of $(\mathrm{P}_{t,x})$, then $s \to V(s,y(s))$ is non-decreasing on $[t,1]$; b(ii) can be deduced from this latter property.

Since V is lower semicontinuous, it remains only to verify

$$\liminf_{\{(t',x')\to(1,x):t'<1,x'\in\operatorname{int} K\}} V(t',x') \le V(1,x) \quad \forall\, x \in K.$$

Lemma 6.3 tells us that there exists $\delta \in]0,1[$ and a solution $y : [1-\delta,1] \to \mathbf{R}^n$ such that $y(1) = x$ and

$$y(t) \in \operatorname{int} K \quad \forall\, t \in [1-\delta,1[.$$

But $V(t,y(t)) \le V(1,x)$, a basic monotonicity property of the value function. Since y is continuous,

$$\liminf_{\{(t',x')\to(1,x):t'<1,x'\in\operatorname{int} K\}} V(t',x') \le \limsup_{t\uparrow 1} V(t,y(t)) \le V(1,x).$$

as required.

(b) $\Rightarrow$ (c). This implication is a consequence duality relationships between $\partial_- V$ and $D_\uparrow V$.

(c) $\Rightarrow$ (a). Assume that V satisfies (c). Take any $x_0 \in K$ and $t_0 \in [0,1]$.

Step 1: We show that

$$V(t_0,x_0) \;\ge\; \inf(P_{t_0,x_0}). \tag{142}$$

This inequality is automatically satisfied if $V(t_0,x_0) = +\infty$. So we assume that $V(t_0,x_0) < +\infty$.

Notice that, since $\operatorname{dom} V \subset K$, conditions c(i) and c(ii) imply

$$\forall (t,x) \in]0,1[\times \mathbf{R}^n,\ (p_t, p_x) \in \partial_- V(t,x)$$
$$-p_t + H(t,x,-p_x) \leq 0$$

and

$$\liminf_{\{(t',x')\to(0,x):t'>0\}} V(t',x') = V(0,x) \quad \forall\, x \in \mathbf{R}^n,$$

(We here regard V as a function on $[0,1]\times\mathbf{R}^n$ which takes value $+\infty$ at points $(t,x) \notin [0,1]\times K$.) But then we deduce by applying the same arguments as in Section 4 the existence of a solution $x : [t_0,1] \to \mathbf{R}^n$ such that $x(t_0) = x_0$ and

$$V(t_0,x_0) \geq V(t,x(t)) \quad \forall\, t \in [t_0,1].$$

This inequality implies that $V(t,x(t)) < +\infty$ for all $t \in [t_0,1]$. Since $\operatorname{dom} V \subset K$, we conclude that $x(\cdot)$ satisfies the state constraint. It also implies that

$$V(t_0,x_0) \geq V(1,x(1)) = g(x(1)) \geq \inf(\mathrm{P}_{t,x}).$$

This is the required inequality.

Step 2: We show that

$$V(t_0,x_0) \ \leq\ \inf(P_{t_0,x_0}). \tag{143}$$

This will complete the proof, since (143) combines with (142) to give $V(t_0,x_0) = \inf(P_{t_0,x_0})$.

(143) is automatically satisfied if $\inf(P_{t_0,x_0}) = +\infty$. So we assume that it is finite. In this case, $\inf(P_{t_0,x_0})$ is the infimum of $g(x(1))$ over all feasible arcs of (P_{t_0,x_0}). It therefore suffices to show that

$$V(t_0,x_0) \leq g(\bar{x}(1)),$$

where $\bar{x} \in W^{1,1}([t_0,1];\mathbf{R}^n)$ is an arbitrary feasible arc of (P_{t_0,x_0}).

By hypothesis,

$$g(\bar{x}(1)) = \liminf_{\{(\tau,\xi)\to(1,\bar{x}(1)):\tau<1,\xi\in\operatorname{int} K\}} V(\tau,\xi).$$

There exists, therefore, a sequence $\{(\tau_i,\xi_i)\}$ in $[t_0,1)\times \operatorname{int} K$ such that $\xi_i \to \bar{x}(1)$ and

$$V(\tau_i,\xi_i) \to g(\bar{x}(1)). \tag{144}$$

Lemma 6.3(ii) asserts the existence of a sequence of solutions $x_i : [t_0, \tau_i] \to \mathbf{R}^n$ such that $x_i(\tau_i) = \xi_i$,

$$x_i(t) \in \operatorname{int} K \quad \forall\, t \in [t_0, \tau_i]$$

and

$$||x_i - \bar{x}||_{L^\infty([t_0,\tau_i];\mathbf{R}^n)} \to 0 \quad \text{as } i \to \infty. \tag{145}$$

Filippov's Theorem tells us that x_i can be extended to all of $[t_0, 1]$ (we write the extension also x_i) as a solution to our differential inclusion. Choose $\sigma_i \in]\tau_i, 1[$ and $\epsilon_i > 0$ such that

$$x_i(t) + \epsilon_i B \subset \operatorname{int} K \quad \forall\, t \in [t_0, \sigma_i].$$

Now apply Lemma 6.4 with $\sigma_i = t_1$ and $\bar{x} = x_i$ to conclude that

$$V(t_0, x_i(t_0)) \le V(\tau_i, \xi_i).$$

It follows from (144), (145) and the lower semicontinuity of V that

$$V(t_0, x_0) = V(t_0, \bar{x}(t_0)) \le \liminf_i V(t_0, x_i(t_0)) \le \lim_i V(\tau_i, \xi_i) = g(\bar{x}(1))$$

as required.

References

[1] AUBIN J.-P. & CELLINA A. (1984) DIFFERENTIAL INCLUSIONS, Springer-Velag, Grundlehren der math. Wiss. # 264

[2] AUBIN J.-P. (1981) *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions*, Advances in Mathematics, Supplementary Studies, Ed. Nachbin L., 160-232.

[3] AUBIN J.-P. (1990) *A survey of viability theory*, SIAM J. Control Optim. **28**, 749-788.

[4] AUBIN J.-P. (1991) VIABILITY THEORY, Birkhäuser, Boston.

[5] AUBIN J.-P. & FRANKOWSKA H. (1990) SET-VALUED ANALYSIS, Birkhäuser, Boston.

[6] BARDI, M. & CAPUZZO-DOLCETTA I. (1997) OPTIMAL CONTROL AND VISCOSITY SOLUTIONS OF HAMILTON-JACOBI-BELLMAN EQUATIONS, Birkhäuser, Boston.

[7] BARRON E.N. & JENSEN R. (1990) *Semicontinuous viscosity solutions for Hamilton-Jacobi equations with convex Hamiltonian*, Comm. Partial Differential Equations **15**, 1713-1742.

[8] BUTTAZZO G. (1989) SEMICONTINUITY, RELAXATION AND INTEGRAL REPRESENTATION PROBLEMS IN THE CALCULUS OF VARIATIONS, Pitman Res. Notes Math. Ser., Longman, Harlow.

[9] BYRNES Ch. & FRANKOWSKA H. (1992) *Unicité des solutions optimales et absence de chocs pour les équations d'Hamilton-Jacobi-Bellman et de Riccati,* Comptes-Rendus de l'Académie des Sciences, t. 315, Série 1, Paris, 427-431

[10] BYRNES Ch. & FRANKOWSKA H. (1998) *Uniqueness of optimal trajectories and the nonexistence of shocks for Hamilton-Jacobi-Bellman and Riccati partial differential equations Differential Inclusions and Optimal Control, Lecture Notes in Nonlinear Analysis*, Vol. 2, J. Schauder Center for Nonlinear Studies, Ed. J. Andres, L. Gorniewicz and P. Nistri

[11] CANNARSA P. & FRANKOWSKA H. (1991) *Some characterizations of optimal trajectories in control theory*, SIAM J. on Control, 29, 1322-1347

[12] I. CAPUZZO-DOLCETTA & P.-L. LIONS, *Hamilton Jacobi Equations with State Constraints*, *Trans. Am. Math. Soc.*, 318 (1990), pp.643-685.

[13] CARATHEODORY C. (1935) VARIATIONSRECHNUNG UND PARTIELLE DIFFERENTIALGLEICHUNGEN ERSTER ORDNUNG, Leipzig: Teubner.

[14] CAROFF N. & FRANKOWSKA H. (1996) *Conjugate points and shocks in nonlinear optimal control*, Transactions of AMS, 348, 3133-3153

[15] CESARI L. (1983) OPTIMIZATION THEORY AND APPLICATIONS. PROBLEMS WITH ORDINARY DIFFERENTIAL EQUATIONS, Appl. Math. 17, Springer-Verlag, Berlin.

[16] CLARKE F.H. (1983) OPTIMIZATION AND NONSMOOTH ANALYSIS, Wiley-Interscience

[17] CRANDALL M.G., EVANS L.C. & LIONS P.L. (1984) *Some properties of viscosity solutions of Hamilton-Jacobi equation*, Trans. Amer. Math. Soc., 282(2), 487-502

[18] CRANDALL M.G. & LIONS P.L. (1983) *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277, 1-42

[19] FILIPPOV A.F. (1958) *On some problems of optimal control theory*, Vestnik Moskowskogo Universiteta, Math. no.2, 25-32

[20] FILIPPOV A.F. (1967) *Classical solutions of differential equations with multivalued right hand side*, SIAM J. on Control, 5, 609-621

[21] FLEMING W. H. & RISHEL R. W. (1975) DETERMINISTIC AND STOCHASTIC OPTIMAL CONTROL, Springer-Verlag

[22] FLEMING W. H. & SONER H.M. (1993) CONTROLLED MARKOV PROCESSES AND VISCOSITY SOLUTIONS, Springer-Verlag

[23] FRANKOWSKA H. (1987) *L'équation d'Hamilton-Jacobi contingente*, Comptes-Rendus de l'Académie des Sciences, PARIS, Série 1, 304, 295-298

[24] FRANKOWSKA H. (1987) *The maximum principle for an optimal solution to a differential inclusion with end point constraints*, SIAM J. on Control and Optimization, 25, 145-157

[25] FRANKOWSKA H. (1989) *Contingent cones to reachable sets of control systems*, SIAM J. on Control and Optimization, 27, 170-198

[26] FRANKOWSKA H. (1989) *Hamilton-Jacobi equation: viscosity solutions and generalized gradients*, J. of Math. Analysis and Appl. 141, 21-26)

[27] FRANKOWSKA H. (1989) *Non smooth solutions to an Hamilton-Jacobi equation*, Proceedings of the International Conference Bellman Continuum, Antibes, France, June 13-14, 1988, Lecture Notes in Control and Information Sciences, Springer Verlag

[28] FRANKOWSKA H. (1989) *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Applied Mathematics and Optimization, 19, 291-311

[29] FRANKOWSKA H. (1989) *Set-valued analysis and some control problems*, Proceedings of the International Conference 30 YEARS OF MODERN CONTROL THEORY, Kingston, June 3-6, 1988 E. Roxin Editor, Marcel Dekker

[30] FRANKOWSKA H. (1990) *Some inverse mapping theorems*, Ann. Inst. Henri Poincaré, Analyse Non Linéaire, 3, 183-234

[31] FRANKOWSKA H. (1991) *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, Proceedings of IEEE CDC Conference, Brighton, England, December 1991.

[32] FRANKOWSKA H. (1993) *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim. **31**, 257-272.

[33] FRANKOWSKA H., PLASKACZ S. & RZEŻUCHOWSKI T. (1995) *Measurable viability theorems and Hamilton-Jacobi-Bellman equation*, J. Differential Equations **116**, 265-305.

[34] FRANKOWSKA H. & PLASKACZ S. (1999) *Hamilton-Jacobi equations for infinite horizon control problems with state constraints*, Proceedings of International Conference "Calculus of Variations and related topics", Haifa, March 25-April 1, 1998

[35] FRANKOWSKA H. & PLASKACZ S. (2000) *Semicontinuous solutions of Hamilton-Jacobi-Bellman equations with degenerate state constraints*, JMAA, 251 818-838

[36] FRANKOWSKA H. & RAMPAZZO F., (2000) *Filippov's and Filippov-Ważewski's Theorems on Closed Domains*, J.D.E., 161, 449-478

[37] FRANKOWSKA H. & VINTER R.B. (2000) *Existence of Neighbouring Feasible Trajectories: Applications to Dynamic Programming for State Constrained Optimal Control Problems.*, J. Optimization Theory and Applications, 104, 21-40

[38] GUSEINOV H. G., SUBBOTIN A. I. & USHAKOV V. N. (1985) *Derivatives for multivalued mappings with applications to game theoretical problems of control*, Problems of Control and Information Theory, Vol.14, 155-167

[39] HADDAD G. (1981) *Monotone trajectories of differential inclusions with memory*, Isr. J. Math., 39, 83-100

[40] IOFFE A.D. (1977) *On lower semicontinuity of integral functionals*, SIAM J. Control Optim. **15**, 521-521 and 991-1000.

[41] LIONS P.-L. (1982) GENERALIZED SOLUTIONS OF HAMILTON-JACOBI EQUATIONS, Pitman, Boston

[42] MARCHAUD H. (1934) *Sur les champs de demi-cônes et les équations différentielles du premier ordre*, Bull. Sc. Math., 62, 1-38

[43] MARCHAUD H. (1936) *Sur les champs continus de demi-cônes convexes et leurs intégrales*, Compos. Math. 1, 89-127

[44] NAGUMO M. (1942) *Uber die lage der integralkurven gewhnlicher differentialgleichungen*, Proc. Phys. Math. Soc. Japan, 24, 551-559

[45] OLECH C. (1969) *Existence theorems for optimal control problems involving multiple integrals*, J. Diff. Eq., 6, 512-526

[46] OLECH C. (1969) *Existence theorems for optimal control problems with vector-valued cost functions*, Trans. Am. Math. Soc., 136, 157-180

[47] OLECH C. (1976) *Weak lower semicontinuity of integral functionals*, J. Optim. Theory Appl. **19**, 3-16

[48] ROCKAFELLAR T. (1981) *Proximal subgradients, marginal values and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res. **6**, 424-436.

[49] ROCKAFELLAR T. & WETS R. (1998) VARIATIONAL ANALYSIS, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin.

[50] SUBBOTIN A. I. (1980) *A generalization of the basic equation of the theory of the differential games*, Soviet. Math. Dokl., 22, 358-362.

[51] SONER H.M. (1986) *Optimal Control with State-Space Constraints*, SIAM J. Control and Optimization, 24, pp. 552-561.

[52] WAŻEWSKI T. (1963) *On an optimal control problem*, Proc. Conference DIFFERENTIAL EQUATIONS AND THEIR APPLICATIONS, PRAGUE, 1962, 229-242

[53] ZAREMBA S.C. (1934) *Sur une extension de la notion d'équation diférentielle*, Comptes Rendus Acad. Sc., Paris, 199, 545-548

[54] ZAREMBA S.C. (1936) *Sur les équations au paratingent*, Bull. Sc. Math., 60, 139-160

Return Method: Some Applications to Flow Control

Jean-Michel Coron*

Département de Mathématique, Université Paris-Sud, Orsay, France

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS028010

*Jean-Michel.Coron@math.u-psud.fr

Dedicated to Jean Lévine for his 50^{th} birthday
and to CAS for it's 25^{th} birthday

Abstract

Due to recent progress in advanced technologies in many fields of engineering sciences, applications of flow control are developing very quickly. In this paper we survey only a tiny and theoretical part of the recent results obtained in flow control, namely some results on the controllability and on the stabilizability of the equations of incompressible fluids which have been obtained by means of the return method.

Contents

1 Introduction **659**

2 Return method **659**

3 Controllability of the Euler and Navier-Stokes equations **666**
3.1 Controllability of the Euler equations 668
3.2 Controllability of the Navier-Stokes equations 672
3.2.1 Local results 674
3.2.2 Global results 675

4 Local controllability of a 1-D tank containing a fluid modeled by the Saint-Venant equations **677**

5 Null asymptotic stabilizability of the 2-D Euler control system **692**

References **698**

1 Introduction

For finite dimensional system one knows many powerful sufficient conditions for local controllability of a nonlinear control system. This is not the case in infinite dimension, where, roughly speaking, the only known general result is that if the linearized control system at an equilibrium is controllable, then the nonlinear control system is locally controllable at the equilibrium. The return method, that we have introduced in [11] for a stabilisation problem in finite dimension and first used in infinite dimension for the controllability of the Euler equations in [13], allows in some cases to get the local controllability at the equilibrium of the nonlinear control system even if the linearized control system at the equilibrium is not controllable. The idea of the return consists in the following one. If one can find a trajectory of the nonlinear control system such that

- it starts and ends at the equilibrium,
- the linearized control system around this trajectory is controllable,

then, in general, the implicit function theorem allows to conclude that one can go from any state close to the equilibrium to any other state close to the equilibrium.

In this paper, we sketch some results in flow control which has been obtained by this method, namely

- Global controllability results of the Euler equations of incompressible fluids,
- Global controllability results for the Navier-Stokes equations of incompressible fluids,
- Local controllability of a 1-D tank containing a fluid modeled by the Saint-Venant equations
- Null global asymptotic stabilizability by means of explicit boundary feedback laws for the 2-D inviscid incompressible fluids on simply connected domains

2 Return method

In order to explain this method, let us just consider the problem of local controllability of a control system in finite dimension. So we consider the

control system

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

where $x \in \mathbb{R}^n$ is the state and $u \in \mathbb{R}^m$ is the control ; we assume that f is of class $\mathcal{C}^\infty$ and satisfies

$$f(0, 0) = 0.$$

There are various possible definitions of local controllability. Here we use the following one, called the small time local controllability,

Definition 1 *The control system $\dot{x} = f(x, u)$ is small time locally controllable if for every $T > 0$ there exist $\varepsilon > 0$ in $(0, +\infty)$ such that, for every $x_0 \in \mathbb{R}^n$ and $x_1 \in \mathbb{R}^n$ both of norm less than ε, there exists a bounded measurable function $u : [0, T] \to \mathbb{R}^m$ such that, if x is the (maximal) solution of $\dot{x} = f(x, u(t))$ which satisfies $x(0) = x_0$, then $x(T) = x_1$.*

One does not know any interesting necessary and sufficient condition for small time local controllability but there are many useful necessary conditions and sufficient conditions which have been found during the last thirty years. See for example the papers by A. Agrachev [2], R.M. Bianchini and G. Stefani [5, 6], H. Hermes [38], M. Kawski [43], H.J. Sussmann [71, 72], H.J. Sussmann and V. Jurdjevic [73], and A. Tret'yak [74]. Note that all these conditions rely on Lie bracket and that this geometric tool does not seem to give good results for distributed control systems - in this case x is an infinite dimensional space -. On the other hand for *linear* distributed control systems there are powerful methods to prove controllability - e.g. the H.U.M. method due to J.-L.Lions, see [52]. The return method consists in reducing the local controllability of a nonlinear control system to the existence of - suitable - periodic (or "almost periodic" -see below the cases of the Navier-Stokes control system and of the Saint-Venant equations) trajectories and to the controllability of *linear* systems. The idea is the following one: assume that, for every positive real number T, there exists a measurable bounded function $\bar{u} : [0, T] \to \mathbb{R}^m$ such that, if we denote by $\bar{x}$ the (maximal) solution of $\dot{\bar{x}} = f(\bar{x}, \bar{u}(t))$, $\bar{x}(0) = 0$, then

$$\bar{x}(T) = 0, \tag{2.1}$$

$$\text{the linearized control system around } (\bar{x}, \bar{u}) \text{ is controllable on } [0, T]. \tag{2.2}$$

Then it follows easily from the inverse mapping theorem - see e.g. [66], Theorem 7 p. 126 - that $\dot{x} = f(x, u)$ is small time locally controllable. Let

us recall that the linearized control system around $(\bar{x}, \bar{u})$ is the time-varying control system

$$\dot{y} = A(t)y + B(t)v, \tag{2.3}$$

where the state is $y \in \mathbb{R}^n$, the control is $v \in \mathbb{R}^m$ and

$$A(t) = (\partial f/\partial x)(\bar{x}(t), \bar{u}(t)),\ B(t) = (\partial f/\partial u)(\bar{x}(t), \bar{u}(t)).$$

For the linear control system (2.3), controllability on $[0,T]$ means, by definition, that for every y_0 and y_1 in $\mathbb{R}^n$, there exists a bounded measurable function $v : [0,T] \to \mathbb{R}^m$ such that if $\dot{y} = A(t)y + B(t)v$ and $y(0) = y_0$, then $y(T) = y_1$. There is a well known Kalman-type sufficient condition for the controllability of (1.5) due to Silverman and Meadows [63] -see also [66, Prop. 3.5.16]-. This is the following one.

Proposition 2 *Assume that for some $\bar{t}$ in $[0,T]$*

$$\text{Span}\{B_i(\bar{t})\, v;\ v \in \mathbb{R}^m,\ \ i \geq 0\} = \mathbb{R}^n, \tag{2.4}$$

with

$$B_i = \left(\frac{d}{dt} - A\right)^i B.$$

Then the linear control system (2.3) is controllable on $[0,T]$. Moreover if A and B are analytic on $[0,T]$ and if the linear control system (2.3) is controllable on $[0,T]$, then (2.4) holds for all $\bar{t}$ in $[0,T]$.

Note that if one takes $\bar{u} \equiv 0$, then the above method just gives the well known fact that if the time-invariant linear system

$$\dot{y} = \frac{\partial f}{\partial x}(0,0)y + \frac{\partial f}{\partial u}(0,0)v$$

is controllable, then the nonlinear control system $\dot{x} = f(x,u)$ is small time locally controllable. But it may happen that (2.2) does not hold for $\bar{u} \equiv 0$, but holds for other choices of $\bar{u}$. Let us give simple examples.

Example 3 We take $n = 2$, $m = 1$ and consider the control system

$$\dot{x}_1 = x_2^3,\ \dot{x}_2 = u.$$

Let us take $\bar{u} \equiv 0$; then $\bar{x} \equiv 0$ and the linearized control system around $(\bar{x}, \bar{u})$ is

$$\dot{y}_1 = 0,\ \dot{y}_2 = v.$$

which is clearly not controllable. Let us now take $\bar{u} \in \mathcal{C}^\infty([0,T];\mathbb{R})$ such that

$$\int_0^{T/2} \bar{u}(t)dt = 0,$$
$$\bar{u}(T-t) = \bar{u}(t),\ \forall\, t \in [0,T].$$

Then one easily checks that

$$\bar{x}_2(T/2) = 0,$$
$$\bar{x}_2(T-t) = -\bar{x}_2(t),\ \forall t \in [0,T],$$
$$\bar{x}_1(T-t) = \bar{x}_1(t),\ \forall t \in [0,T].$$

In particular, we have

$$\bar{x}_1(T) = 0\,,\ \bar{x}_2(T) = 0.$$

The linearized control system around $(\bar{x},\bar{u})$ is

$$\dot{y}_1 = 3\bar{x}_2^2(t)y_2\,,\ \dot{y}_2 = v.$$

Hence

$$A(t) = \begin{pmatrix} 0 & 3\bar{x}_2(t)^2 \\ 0 & 0 \end{pmatrix},\ B(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and one easily sees that (2.4) holds if and only if

$$\exists\, i \in \mathbb{N} \text{ such that } \frac{d^i \bar{x}_2}{dt^i}(\bar{t}) \neq 0. \tag{2.5}$$

Note that (2.5) holds for at least a $\bar{t}$ in $[0,T]$ if (and only if) $u \not\equiv 0$. So (2.2) holds if $u \not\equiv 0$.

Example 4 We take $n = 3$, $m = 2$ and the control system is

$$\dot{x}_1 = u_1,\ \dot{x}_2 = u_2,\ \dot{x}_3 = x_1u_2 - x_2u_1. \tag{2.6}$$

Again one can check that the linearized control system around $(\bar{x},\bar{u})$ is controllable on $[0,T]$ if and only if $\bar{u} \not\equiv 0$. Note that, for the control system (2.6), it is easy to achieve the "return condition" (2.1). Indeed, if

$$\bar{u}(T-t) = -\bar{u}(t),\ \forall\, t \in [0,T],$$

then

$$\bar{x}(T-t) = \bar{x}(t),\ \forall\, t \in [0,T]$$

and, in particular,

$$\bar{x}(T) = \bar{x}(0) = 0.$$

Example 5 Let us now give an example, which has some relations with the 1-D tank studied in section 4, where the return method can be used to get large time local controllability. For this example the control system is

$$\dot{x}_1 = x_2,\ \dot{x}_2 = -x_1 + u,\ \dot{x}_3 = x_4,\ \dot{x}_4 = -x_3 + 2x_1x_2, \tag{2.7}$$

where the state is $(x_1, x_2, x_3, x_4) \in \mathbb{R}^4$ and the state is $u \in \mathbb{R}$. Let us first point out that that this control system is not small time locally controllable. Indeed if $(x, u) : [0, T] \to \mathbb{R}^4 \times \mathbb{R}$ is a trajectory of the control system (2.7) such that $x(0) = 0$ then

$$x_3(T) = \int_0^T x_1^2(t)\cos(T - t)dt, \tag{2.8}$$

$$x_4(T) = x_1^2(T) - \int_0^T x_1^2(t)\sin(T - t)dt \tag{2.9}$$

In particular if $x_1(T) = 0$ and $T \leqslant \pi$ then $x_4(T) \leqslant 0$ with equality if and only if $x \equiv 0$. So, if for $T > 0$ we denote by $\mathcal{P}(T)$ the following controllability property

$\mathcal{P}(T)$ There exists $\varepsilon > 0$ in $(0, +\infty)$ such that, for every $x_0 \in \mathbb{R}^n$ and $x_1 \in \mathbb{R}^n$ both of norm less than ε, there exists a bounded measurable function $u : [0, T] \to \mathbb{R}$ such that, if x is the (maximal) solution of (2.7) which satisfies $x(0) = x_0$, then $x(T) = x_1$,

then, for every $T \in (0, \pi]$, $\mathcal{P}(T)$ is false. Let us show how the return method can be used to prove that

$$\mathcal{P}(T) \text{ holds for every } T \in (\pi, +\infty). \tag{2.10}$$

Let $T > \pi$. Let

$$\eta = \frac{1}{2}\text{Min } (T - \pi, \pi).$$

Let $\bar{x}_1 : [0, T] \to \mathbb{R}$ be a function of class $\mathcal{C}^\infty$ such that

$$\bar{x}_1(t) = 0\, \forall t \in [\eta, \pi] \cap [\pi + \eta, T], \tag{2.11}$$

$$\bar{x}_1(t + \pi) = x_1(t)\, \forall t \in [0, \eta]. \tag{2.12}$$

Let $\bar{x}_2 : [0, T] \to \mathbb{R}$ and $\bar{u} : [0, T] \to \mathbb{R}$ be such that

$$\bar{x}_2 = \dot{\bar{x}}_1,\ \bar{u} = \dot{\bar{x}}_2 + \bar{x}_1 + \bar{u},$$

In particular

$$\bar{x}_2(t) = 0 \,\forall t \in [\eta, \pi] \cap [\pi + \eta, T], \tag{2.13}$$

$$\bar{x}_2(t + \pi) = x_2(t) \,\forall t \in [0, \eta]. \tag{2.14}$$

Let $\bar{x}_3 : [0, T] \to \mathbb{R}$ and $\bar{x}_4 : [0, T] \to \mathbb{R}$ be defined by

$$\dot{\bar{x}}_3 = \bar{x}_4,\ \dot{\bar{x}}_4 = -\bar{x}_3 + 2\bar{x}_1\bar{x}_2, \tag{2.15}$$

$$\bar{x}_3(0) = 0,\ \bar{x}_4(0) = 0. \tag{2.16}$$

So $(\bar{x}, \bar{u})$ is a trajectory of the control system (2.7). Then, using (2.8), (2.9), (2.11), (2.13), (2.12), (2.14), one sees that

$$\bar{x}(T) = 0.$$

If $\bar{x}_1 \equiv 0$, $(\bar{x}, \bar{u}) \equiv 0$ and the linearized control system around $(\bar{x}, \bar{u})$ is not controllable. But, as one easily checks using the Kalman-type sufficient condition for the controllability of linear time-varying control sytem due to Silverman and Meadows (Proposition 2), if $\bar{x}_1 \not\equiv 0$ then the linearized control system around $(\bar{x}, \bar{u})$ is controllable. This shows (2.10).

Example 6 This example comes from an exercise proposed by Kawski at this summer school. The control system is now

$$\dot{x}_1 = u,\ \dot{x}_2 = x_1^3,\ \dot{x}_3 = x_1 x_2, \tag{2.17}$$

where the state is $(x_1, x_2, x_3) \in \mathbb{R}^3$ and the control is $u \in \mathbb{R}$. Let us take $\bar{u} \equiv 0$; then $\bar{x} \equiv 0$ and the linearized control system around $(\bar{x}, \bar{u})$ is

$$\dot{y}_1 = v,\ \dot{y}_2 = 0,\ \dot{y}_3 = 0.$$

which is clearly not controllable. Let T be any strictly positive real number and let $\bar{u} \in \mathcal{C}^\infty([0, T]; \mathbb{R})$ be such that

$$\int_0^T \bar{u}(t)dt = 0,$$

$$\bar{u}(T - t) = \bar{u}(t),\ \forall\, t \in [0, T].$$

Then one easily checks that

$$\bar{x}_1(T - t) = -\bar{x}_1(t),\ \forall t \in [0, T],$$

$$\bar{x}_2(T - t) = \bar{x}_2(t),\ \forall t \in [0, T],$$

$$\bar{x}_3(T - t) = \bar{x}_3(t),\ \forall t \in [0, T].$$

In particular, we have

$$\bar{x}_1(T) = 0,\ \bar{x}_2(T) = 0,\ \bar{x}_3(T) = 0.$$

The linearized control system around $(\bar{x}, \bar{u})$ is

$$\dot{y}_1 = v,\ \dot{y}_2 = 3\bar{x}_1 y_1,\ \dot{y}_3 = \bar{x}_2 y_1 + \bar{x}_1 y_2.$$

Hence

$$A(t) = \begin{pmatrix} 0 & 0 & 0 \\ 3\bar{x}_1 & 0 & 0 \\ \bar{x}_2 & \bar{x}_1 & 0 \end{pmatrix},\ B(t) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Hence

$$\text{Det}\ (B_0|B_1|B_2) = -6x_1 d \text{ with } d = x_1^4 + \dot{x}_1 x_2. \tag{2.18}$$

One easily checks that (2.4) holds if (and only if) $u \not\equiv 0$. So (2.2) holds if $u \not\equiv 0$.

One may wonder if the local controllability of $\dot{x} = f(x, u)$ implies the existence of u in $\mathcal{C}^\infty([0,T];\mathbb{R}^m)$ such that (2.1) and (2.2) hold. It has been proved to be true by Sontag in [65]. Let us also remark that the above examples suggest that for many choices of $\bar{u}$ then (2.2) holds. This in fact holds in general. More precisely let us assume that

$$\left\{ h(0); h \in \text{Lie}\left\{ \frac{\partial f}{\partial u^\alpha}(\cdot, 0),\ \alpha \in \mathbb{N}^m \right\} \right\} = \mathbb{R}^n, \tag{2.19}$$

where Lie $\mathcal{F}$ denotes the Lie algebra generated by the vector fields in $\mathcal{F}$; then for generic u in $\mathcal{C}^\infty([0,T];\mathbb{R}^m)$ (2.2) holds; this is proved in [12], and in [68] if f is analytic. Let us recall that by a theorem due to Sussmann and Jurdjevic [73], (2.19) is a necessary condition for local controllability if f is analytic.

The return method does not seem to give any new interesting controllability result if x lies in a finite dimensional space ; in particular

- The small time local controllability in Example 3 follows from the Hermes condition [38, 72],
- The small time local controllability in Example 4 follows from Rashevski-Chow's theorem [59, 9],

- The large time local controllability (more precisely (2.10)) follows from a general result obtained by R. Bianchini about unilateral variations in [7] (one considers the trajectory $(x, u) \equiv (0, 0)$).

But it gives some new results for the controllability of distributed control system as we are now to show in the following sections.

3 Controllability of the Euler and Navier-Stokes equations

Let us introduce some notations. Let $l \in \{2, 3\}$ and let Ω be a bounded nonempty connected open subset of $\mathbb{R}^l$ of class $\mathcal{C}^\infty$. Let Γ_0 be an open subset of $\Gamma := \partial\Omega$ and let Ω_0 be an open subset of Ω. We assume that

$$\Gamma_0 \cup \Omega_0 \neq \emptyset. \tag{3.1}$$

The set Γ_0 is the part of the boundary and Ω_0 is the part of the domain Ω on which the control acts. The fluid that we consider is incompressible so that the velocity field y satisfies

$$\text{div } y = 0.$$

On the part of the boundary $\Gamma\backslash\Gamma_0$ where there is no control the fluid does not cross the boundary: it satisfies

$$y.n = 0 \text{ on } \Gamma\backslash\Gamma_0, \tag{3.2}$$

where n denotes the outward unit normal vector field on Γ. When the fluid is viscous it satifies on $\Gamma\backslash\Gamma_0$, besides (3.2), some extra conditions which will be specified later on. For the moment being, let us just call by **BC** all the boundary conditions (*including* (3.2)) satisfied by the fluid on $\Gamma\backslash\Gamma_0$.

Let us introduce the following definition.

Definition 7 *A trajectory of the Navier-Stokes control system (resp. Euler control system) on the interval of time* $[0, T]$ *is an application* $y : \overline{\Omega}\times[0, T] \to \mathbb{R}^l$ *of class* $\mathcal{C}^\infty$ *such that, for some function* $p : \overline{\Omega} \times [0, T] \to \mathbb{R}$ *of class* $\mathcal{C}^\infty$,

$$\frac{\partial y}{\partial t} - \nu\Delta y + (y \cdot \nabla)y + \nabla p = 0 \text{ in } (\overline{\Omega}\backslash\Omega_0) \times [0, T], \tag{3.3}$$

$$(\textit{resp. } \frac{\partial y}{\partial t} + (y \cdot \nabla)y + \nabla p = 0 \text{ in } (\overline{\Omega}\backslash\Omega_0) \times [0, T]) \tag{3.4}$$

$$div\ y = 0\ in\ \overline{\Omega} \times [0,T], \tag{3.5}$$

$$y(\cdot,t)\ satisfies\ the\ boundary\ conditions\ BC\ on\ \Gamma \backslash \Gamma_0,\ \forall t \in [0,T]. \tag{3.6}$$

The real number $\nu > 0$ appearing in (3.3) is the viscosity. J.–L. Lions' problem of controllability is the following one: let $T > 0$, let y_0 and y_1 in $\mathcal{C}^\infty(\overline{\Omega};\mathbb{R}^l)$ be such that

$$\text{div } y_0 = 0 \text{ in } \overline{\Omega}, \tag{3.7}$$

$$\text{div } y_1 = 0 \text{ in } \overline{\Omega}, \tag{3.8}$$

$$y_0 \text{ satisfies the boundary conditions BC on } \Gamma \backslash \Gamma_0, \tag{3.9}$$

$$y_1 \text{ satisfies the boundary conditions BC on } \Gamma \backslash \Gamma_0, \tag{3.10}$$

does there exist a trajectory y of the Navier-Stokes or the Euler control system such that

$$y(\cdot,0) = y_0 \text{ in } \overline{\Omega}, \tag{3.11}$$

and, for an appropriate topology –see [53, 54]–,

$$y(\cdot,T) \text{ is "close" to } y_1 \text{ in } \overline{\Omega}? \tag{3.12}$$

That is to say, starting with the initial data y_0 for the velocity field, we ask whether there are trajectories of the control system considered (Navier-Stokes if $\nu > 0$, Euler if $\nu = 0$) which, at a fixed time T, are arbitrarly close to the given velocity field y_1. If this problem has always a solution one says that the control system considered is approximately controllable.

Note that (3.3), (3.5), (3.6) and (3.11) have many solutions. In order to have uniqueness one needs to add extra conditions. These extra conditions are the controls.

In the case of the Euler control system one can even require instead of (3.12) the stronger condition

$$y(\cdot,T) = y_1 \text{ in } \overline{\Omega}. \tag{3.13}$$

If y still exists with this stronger condition, one says that the Euler control system is exactly controllable. Of course, due to the smoothing of the Navier-Stokes equations, one cannot expect to have (3.13) instead of (3.12) for general y_1. We will see in subsection 3.2 a way to replace (3.13) in order to recover a natural definition of (exact) controllability of the Navier-Stokes condition.

This section is organized as follows

- In subsection 3.1 we consider the case of the Euler control system,
- In subsection 3.2 we consider the case of the Navier-Stokes control system.

3.1 Controllability of the Euler equations

In this section the boundary conditions BC in (3.6), (3.9), and (3.10) are respectively

$$y(x,t).n(x) = 0, \ \forall (x,t) \in (\Gamma \backslash \Gamma_0) \times [0,T], \tag{3.14}$$

$$y_0(x).n(x) = 0, \quad \forall x \in \Gamma \backslash \Gamma_0, \tag{3.15}$$

$$y_1(x).n(x) = 0, \quad \forall x \in \Gamma \backslash \Gamma_0. \tag{3.16}$$

For simplicity we assume that

$$\Omega_0 = \emptyset,$$

i.e. we study the case of boundary control (see [14] when $\Omega_0 \neq \emptyset$ and $l = 2$). In that case a control is given by $y.n$ on Γ_0 with $\int_{\Gamma_0} y.n = 0$ and by curl y if $l = 2$ and (curl y).n if $l = 3$ at the points of $\Gamma_0 \times [0,T]$ where $y.n < 0$: these boundary conditions, (3.14), and the initial condition (3.11) imply the uniqueness of the solution to the Euler equations (3.4) -up to an arbitrary function of t which may be added to p-; see also [44] for the existence of solution.

Let us first point out that in order to have (exact) controllability one needs that

$$\Gamma_0 \text{ intersects every connected component of } \Gamma. \tag{3.17}$$

Indeed, let $\mathcal{C}$ be a connected component of Γ which does not intersect Γ_0 and assume that, for some smooth Jordan curve γ_0 on $\mathcal{C}$ (if $l = 2$ one takes $\gamma_0 = \mathcal{C}$),

$$\int_{\gamma_0} y_0.ds \neq 0, \tag{3.18}$$

but that

$$y_1(x) = 0, \ \forall x \in \mathcal{C}. \tag{3.19}$$

Then there is no solution to our problem, that is there is no $y \in \mathcal{C}^\infty(\overline{\Omega} \times [0,T]; \mathbb{R}^2)$ and $p \in \mathcal{C}^\infty(\overline{\Omega} \times [0,T]; \mathbb{R})$ such that (3.5), (3.4), (3.11), (3.13), and (3.14) hold. Indeed, if such a solution (y,p) exists, then, by Kelvin's law,

$$\int_{\gamma(t)} y(\cdot,t).ds = \int_{\gamma_0} y_0.ds (\in \mathbb{R}), \tag{3.20}$$

where $\gamma(t)$ is the Jordan curve obtained, at time t, from the points of the fluids which at time 0 where on γ_0; in other words $\gamma(t)$ is the image of γ_0 by the flow map associated to the time-varying vector field y. But (3.13), (3.18), (3.19) and (3.20) are in contradiction.

Conversely, if (3.17) holds, then the Euler control system is exactly controllable:

Theorem 8 *Assume that Γ_0 intersects every connected component of $\partial\Omega$. Then the Euler control system is exactly controllable.*

Theorem 8 has been proved in

- [13] when Ω is simply-connected and $l = 2$,
- [14] when Ω is multi-connected and $l = 2$,
- [34] when Ω is contractible and $l = 3$,
- [35] when Ω is not contractible and $l = 3$.

The strategy of the proof of Theorem 8 relies on the "return method" Applied to the controllability of the Euler control system the return method consists in looking for $(\bar{y}, \bar{p})$ such that (3.5), (3.4), (3.11), (3.13) hold, with $y = \bar{y}, p = \bar{p}$, $y_0 = y_1 = 0$ and such that the linearized control system around around the trajectory $\bar{y}$ is controllable under the assumptions of Theorem 8. With

such a $(\bar{y}, \bar{p})$ one may hope that there exists (y, p) -close to $(\bar{y}, \bar{p})$- satisfying the required conditions, at least if y_0 and y_1 are "small". Finally, by using some scaling argument, one can deduce from the existence of (y, p) when y_0 and y_1 are "small" the existence of (y, p) even if y_0 and y_1 are not "small".

Let us emphasize that one cannot take $(\bar{y}, \bar{p}) = (0, 0)$. Indeed, with such a choice of $(\bar{y}, \bar{p})$, (3.5), (3.4), (3.11), (3.13) hold, with $y = \bar{y}, p = \bar{p}$, $y_0 = y_1 = 0$, but the linearized control system around $\bar{y} = 0$ is not at all controllable. Indeed the linearized control system around $\bar{y} = 0$ is

$$\text{div } z = 0 \text{ in } \overline{\Omega} \times [0, T], \tag{3.21}$$

$$\frac{\partial z}{\partial t} + \nabla \pi = 0 \text{ in } \overline{\Omega} \times [0, T], \tag{3.22}$$

$$z(x, t).n(x) = 0, \ \forall (x, t) \in (\Gamma \backslash \Gamma_0) \times [0, T].$$

Taking the curl of (3.22), one gets

$$\frac{\partial \text{curl } z}{\partial t} = 0,$$

which clearly shows that the linearized control system is not controllable. So one needs to consider other $(\bar{y}, \bar{p})$. Let us briefly explain how one constructs "good" $(\bar{y}, \bar{p})$ when $l = 2$ and Ω is simply connected. In such a case one easily checks the existence of a harmonic function θ in $\mathcal{C}^\infty(\overline{\Omega})$ such that

$$\nabla \theta(x) \neq 0, \ \forall x \in \overline{\Omega},$$

$$\frac{\partial \theta}{\partial n} = 0 \text{ on } \Gamma \backslash \Gamma_0.$$

Let $\alpha \in \mathcal{C}^\infty(0, T)$ vanishing 0 and T. Let

$$(\bar{y}, \bar{p})(x, t) = (\alpha(t) \nabla \theta(x), -\alpha'(t) \theta(x) - \frac{1}{2} \alpha^2(t) \mid \nabla \theta(x) \mid^2).$$

Then (3.5), (3.4), (3.11), (3.13) hold, with $y = \bar{y}, p = \bar{p}$, $y_0 = y_1 = 0$. Moreover, using arguments relying on an extension method analogous to the one introduced by D.L. Russell in [60], one can see that the linearized control system around $\bar{y}$ is controllable.

When Γ_0 does not intersect all the connected components of Γ_0 one can get, if $l = 2$, approximate controllability and even exact controllability outside every arbitrarily small neighborhood of the union $\Gamma^\star$ of the connected components of Γ which does not intersect Γ_0. More precisely, one has

Theorem 9 *[14]. Assume that $l = 2$. There exists a constant c_0 depending only on Ω such that, for every Γ_0 as above, every $T > 0$, every $\varepsilon > 0$, and every y_0, y_1 in $\mathcal{C}^\infty(\overline{\Omega}; \mathbb{R}^2)$ satisfying (3.7), (3.8), (3.15) and (3.16), there exists a trajectory y of the Euler control system on $[0, T]$ satisfying (3.11) such that*

$$y(x, T) = y_1(x), \ \forall x \in \overline{\Omega} \text{ such that } dist(x, \Gamma^\star) \geqslant \varepsilon, \tag{3.23}$$

$$|y(\cdot, T)|_{L^\infty} \leqslant c_0(|y_0|_{L^2} + |y_1|_{L^2} + |curly_0|_{L^\infty} + |curly_1|_{L^\infty}). \tag{3.24}$$

In (3.23), $\text{dist}(x, \Gamma^\star)$ denotes the distance of x to $\Gamma^\star$, i.e.

$$\text{dist } (x, \Gamma^\star) = \text{ Min } \{|x - x^\star|; x^\star \in \Gamma^\star\}. \tag{3.25}$$

We use the convention $\text{dist } (x, \emptyset) = +\infty$ and so Theorem 9 implies Theorem 8. In (3.24) $|\ |_{L^r}$, for $r \in [1, +\infty]$, denotes the L^r-norm on Ω. Let us point out that, y_0, y_1, and T as in Theorem 9 being given, it follows from (3.23) and (3.24) that, for every r in $[1, +\infty)$,

$$\lim_{\varepsilon \to 0^+} |y(0, T) - y_1|_{L^r} = 0; \tag{3.26}$$

that is Theorem 9 implies approximate controllability in the L^r-space for every r in $[1, +\infty)$. Let us notice that, if $\Gamma^\star \neq \emptyset$, then, again by Kelvin's law, approximate controllability for the L^∞-norm does not hold. More precisely let us consider the case $l = 2$ and let us denote by $\Gamma_1^\star, \ldots, \Gamma_k^\star$ the connected components of Γ which does not meet Γ_0. Let y_0, y_1 in $\mathcal{C}^\infty(\overline{\Omega}; \mathbb{R}^2)$ satisfying (3.7), (3.8), (3.15) and (3.16). Assume that for some $i \in \{1, \ldots, k\}$

$$\int_{\Gamma_i^\star} y_0.ds \neq \int_{\Gamma_i^\star} y_1.ds$$

Then for $\varepsilon > 0$ small enough there is no trajectory y of the Euler control system on $[0, T]$ satisfying (3.11) such that

$$|y(\cdot, T) - y_1|_{L^\infty} \leq \varepsilon \tag{3.27}$$

One may wonder if, on the contrary one assumes that

$$\int_{\Gamma_i^\star} y_0.ds = \int_{\Gamma_i^\star} y_1.ds, \ \forall i \in \{1, \ldots, k\}. \tag{3.28}$$

Then O. Glass has proved that one has approximate controllability in L^∞ and even in the Sobolev spaces $W^{1,p}$ for every $p \in [1, +\infty)$. Indeed he has proved in [36]

Theorem 10 *Assume that $l = 2$. For every $T > 0$, and every y_0, y_1 in $\mathcal{C}^\infty(\overline{\Omega}; \mathbb{R}^2)$ satisfying (3.7), (3.8), (3.15), (3.16) and (3.28). there exists a sequence $(y^k)_{k\in\mathbb{N}}$ of trajectories of the Euler control system on $[0,T]$ satisfying (3.11) such that*

$$y^k(x,T) = y_1(x),\ \forall x \in \overline{\Omega} \text{ such that } dist(x,\Gamma^\star) \geqslant 1/k,\ \forall k \in \mathbb{N}, \tag{3.29}$$

$$y^k(\cdot,T) \to y_1 \text{ in } W^{1,p}(\Omega) \text{ as } k \to +\infty,\ \forall p \in [1,+\infty). \tag{3.30}$$

Again the convergence in (3.30) is optimal: since the vorticity curl y is conserved along the trajectories of the vector field y one cannot have the convergence in $W^{1,\infty}$. In order to have convergence in $W^{1,\infty}$ one needs to add a relation between curl y_0 and curl y_1 on the Γ_i for $i \in \{1,\dots,l\}$. In this direction O. Glass has proved in [36]

Theorem 11 *Assume that $l = 2$. Let $T > 0$, and let y_0, y_1 in $\mathcal{C}^\infty(\overline{\Omega}; \mathbb{R}^2)$ be such that (3.7), (3.8), (3.15). (3.16) and (3.28) hold. Assume that, for every $i \in \{1,\dots,l\}$, there exits a diffeomorphism D_i of $\Gamma_i^\star$ preserving the orientation such that*

$$curl\ y_1 = (curl\ y_0) \circ D_i.$$

Then there exists a sequence (y^n) of trajectories of the Euler control system on $[0,T]$ satisfying (3.11), (3.29) and

$$y^k(\cdot,T) \to y_1 \text{ in } W^{2,p}(\Omega) \text{ as } k \to +\infty,\ \forall p \in [1,+\infty). \tag{3.31}$$

Again, one cannot expect a convergence in $W^{2,\infty}$ without an extra assumption on y_0 and y_1 -see [36]-.

3.2 Controllability of the Navier-Stokes equations

In this section $\nu > 0$. We now need to specify the boundary conditions BC. Three types of conditions are considered

- Stokes boundary condition,
- Navier boundary condition,
- curl condition.

The Stokes boundary condition is the well known no-slip boundary condition

$$y = 0 \text{ on } \Gamma \backslash \Gamma_0, \tag{3.32}$$

which of course implies (3.2).

The Navier boundary condition [57] imposes, condition (3.2), which is always assumed, and

$$\overline{\sigma} y.\tau + (1-\overline{\sigma}) n^i \left(\frac{\partial y^i}{\partial x^j} + \frac{\partial y^j}{\partial x^i} \right) \tau^j = 0 \text{ on } \Gamma \backslash \Gamma_0, \tag{3.33}$$

where $\overline{\sigma}$ is a constant in $[0,1)$, $n = (n^1, \dots, n^l)$ and $\tau = (\tau^1, \dots, \tau^l)$ is any tangent vector field on the boundary Γ. In (3.33) we also have used the usual summation convention. Note that the Stokes boundary condition (3.32) corresponds to the case $\overline{\sigma} = 1$, which we will not include in the Navier boundary condition considered here. The boundary condition (3.33) with $\overline{\sigma} = 0$ corresponds to the case where there the fluid slips on the wall without friction. It is the appropriate physical model for some flow problems; see [33] for example. The case $\overline{\sigma} \in (0,1)$ corresponds to a case where there the fluid slips on the wall with friction; it is also used in models of turbulence with rough walls; see, e.g., [49]. Note that in [10] F. Coron has derived rigorously the Navier boundary condition (3.33) from the boundary condition at the kinetic level (Boltzmann equation) for compressible fluids. Let us also recall that C. Bardos, F. Golse, and D. Levermore have derived in [4] the incompressible Navier-Stokes equations from a Boltzmann equation.

Let us point out that, using (3.2), one sees that, if $l = 2$ and if τ is the unit tangent vector field on $\partial\Omega$ such that (τ, n) is a direct basis of $\mathbb{R}^2$, (3.33) is equivalent to

$$\sigma y.\tau + \text{curl } y = 0 \text{ on } \Gamma \backslash \Gamma_0$$

with $\sigma \in \mathcal{C}^\infty(\Gamma; \mathbb{R})$ defined by

$$\sigma(x) = \frac{2(1-\overline{\sigma})\kappa(x) - \overline{\sigma}}{1-\overline{\sigma}}, \ \forall x \in \Gamma, \tag{3.34}$$

where κ is the curvature of Γ defined through the relation $\frac{\partial n}{\partial \tau} = \kappa\tau$. In fact we will not use this particular character of (3.34) in our considerations; Theorem 15 below holds for every $\sigma \in \mathcal{C}^\infty(\Gamma; \mathbb{R})$.

Finally the curl condition is considered in dimension 2 ($l = 2$). This condition is, condition (3.2) which is always assumed, and

$$\text{curl } y = 0 \text{ on } \Gamma \backslash \Gamma_0. \tag{3.35}$$

It corresponds to the case $\sigma = 0$ in (3.34).

As mentioned in the introduction, due to smoothing property of the Navier-Stokes equation, one cannot expect to get (3.13), at least for general y_1. For these equations, the good notion for exact controllability is not passing from a given state (y_0) to another given state (y_1). As proposed by A. Fursikov and O. Yu Imanuvilov in [28, 29], the good definition for exact controllability is passing from a given state (y_0) to a given *trajectory* ($\hat{y}_1$). This leads to the following, still open, problem of exact controllability of the Navier-Stokes equation with the Stokes, or Navier, or curl condition.

Open Problem 12 *Let $T > 0$. Let $\hat{y}_1$ be a trajectory of the Navier-Stokes control system on $[0,T]$. Let $y_0 \in \mathcal{C}^\infty(\overline{\Omega};\mathbb{R}^l)$ satisfying (3.7) and (3.9). Does there exist a trajectory y of the Navier-Stokes control system on $[0,T]$ such that*

$$y(x,0) = y_0(x),\ \forall x \in \overline{\Omega}, \tag{3.36}$$

$$y(x,T) = \hat{y}_1(x,T),\ \forall x \in \overline{\Omega}? \tag{3.37}$$

Let us point out that the (global) approximate controllability of the Navier-Stokes control system is also an open problem. Related to the open problem 12 one knows two types of results

- local results,
- global results,

which we briefly describe in the next subsections

3.2.1 Local results

These results do not rely on the return method, but on the HUM and difficult Carleman's ineaqualities. Let us introduce the following definition.

Definition 13 *The Navier-Stokes control system is locally (for the Sobolev H^1 – norm) exactly controllable along the trajectory $\hat{y}_1$ on $[0,T]$ of the Navier-Stokes control system if there exists $\epsilon > 0$ such that, for every $y_0 \in \mathcal{C}^\infty(\overline{\Omega};\mathbb{R}^l)$ satisfying (3.7), (3.9) and*

$$\mid y_0 - \hat{y}_1(\cdot,0) \mid_{H^1(\Omega)} < \epsilon,$$

there exists a trajectory y of the Navier-Stokes control system on $[0,T]$ satisfying (3.36) and (3.37).

Then one has the following results.

Theorem 14 *The Navier-Stokes control system is locally exactly controllable*

(i) along every trajectory for the curl condition or the Navier boundary condition in dimension 2 (l=2),

(ii) along every trajectory if $\Gamma_0 = \Gamma$*,*

(iii) along every stationary trajectory with compact support for the Stokes condition.

Case (i) has been obtained by A.V. Fursikov and O. Yu Imanuvilov in [29, 30]. Case (ii) has been obtained by A.V. Fursikov in [27]. Case (iii) has been obtained by O. Yu Imanuvilov in [41] and [42].

3.2.2 Global results

Let $d \in \mathcal{C}^0(\overline{\Omega}; \mathbb{R})$ be defined by

$$d(x) = \text{ dist } (x, \Gamma) = \text{ Min } \{|x - x'|; x' \in \Gamma\}.$$

In [15] the following theorem is proved

Theorem 15 *Let* $T > 0$*, let* y_0 *and* y_1 *in* $\mathcal{C}^\infty(\overline{\Omega}, \mathbb{R}^2)$ *be such that (3.7) and (3.8) hold. Let us also assume that* y_0 *and* y_1 *satisfies the Navier boundary condition (3.33). Then there exists a sequence* $(y^k; k \in \mathbb{N})$ *of trajectories of the Navier-Stokes control system on* $[0, T]$ *with the Navier boundary condition (3.33) such that, as* $k \to +\infty$*,*

$$\int_\Omega d^\mu |y^k(\cdot, T) - y_1| \to 0, \ \forall \mu > 0, \tag{3.38}$$

$$|y^k(\cdot, T) - y_1|_{W^{-1,\infty}(\Omega)} \to 0, \tag{3.39}$$

and, for all compact K *included in* $\Omega \cup \Gamma_0$*,*

$$|y^k(\cdot, T) - y_1|_{L^\infty(K)} + |curl\ y^k(\cdot, T) - curl\ y_1|_{L^\infty(K)} \to 0. \tag{3.40}$$

In this theorem $W^{-1,\infty}(\Omega)$ denotes the usual Sobolev space of first derivatives of functions in $L^\infty(\Omega)$ and $|\ |_{W^{-1,\infty}(\Omega)}$ one of it's usual norms, for example the norm given in [1, Section 3.10].

As in the proof of the controllability of the 2-D Euler equations of incompressible inviscid fluids (see section 3.1), one uses the return method. Let us recall that it consists in looking for a trajectory of the Navier-Stokes control system $\bar{y}$ such that

$$\bar{y}(\cdot,0) = \bar{y}(\cdot,T) = 0 \text{ in } \overline{\Omega}, \tag{3.41}$$

and such that the linearized control system around the trajectory $\bar{y}$ has a controllability in a "good" sense. With such a $\bar{y}$ one may hope that there exists y – close to $\bar{y}$ – satisfying the required conditions, at least if y_0 and y_1 are "small". Note that the linearized control system around $\bar{y}$ is

$$\frac{\partial z}{\partial t} - \nu\Delta z + (\bar{y}\cdot\nabla)z + (z\cdot\nabla)\bar{y} + \nabla\pi = 0 \text{ in } (\overline{\Omega}\backslash\Omega_0)\times[0,T], \tag{3.42}$$

$$\mathrm{div} z = 0 \text{ in } \overline{\Omega}\times[0,T], \tag{3.43}$$

$$z.n = 0 \text{ on } (\Gamma\backslash\Gamma_0)\times[0,T], \tag{3.44}$$

$$\sigma z.\tau + \mathrm{curl}\ z = 0 \text{ on } (\Gamma\backslash\Gamma_0)\times[0,T]. \tag{3.45}$$

In [29, 30] A. Fursikov and O. Immanuvilov have proved that this linear control system is controllable (see also [51] for the approximate controllability). Of course it is tempting to consider the case $\bar{y} = 0$. Unfortunately, it is not clear how to deduce from the controllability of the linear system (3.42) with $\bar{y} = 0$, the existence of a trajectory y of the Navier-Stokes control system (with the Navier boundary condition) satisfying (3.11) and (3.12) if y_0 and y_1 are ınot small. For this reason, one does not use $\bar{y} = 0$, but $\bar{y}$ similar to the one constructed in [14] to prove the controllability of the 2-D Euler equations of incompressible inviscid fluids; these $\bar{y}$ are chosen to be "large" so that, in some sense, "Δ" is small compared to "$(\bar{y}\cdot\nabla) + (\ \cdot\nabla)\bar{y}$".

Remark 16 *In fact with the $\bar{y}$ we use, one does not have (3.41): we have only the weaker property*

$$\bar{y}(\cdot,0) = 0,\ \bar{y}(\cdot,T) \text{ is "close" to } 0 \text{ in } \overline{\Omega}. \tag{3.46}$$

But the controllability of the linearized control system around $\bar{y}$ is strong enough to take care of the fact that $\bar{y}(\cdot,T)$ is not equal to 0 but only close to 0.

Note that (3.38), (3.39), and (3.40) are not strong enough to imply

$$|y^k(\cdot, T) - y_1|_{L^2(\Omega)} \to 0, \tag{3.47}$$

i.e. to get the approximate controllability in L^2 of the Navier-Stokes control system. But, in the special case where $\Gamma_0 = \Gamma$, (3.38), (3.39), and (3.40) are strong enough to imply (3.47). Moreover, gluing together the proofs of Theorem 14 and of (ii) of Theorem 15, one gets

Theorem 17 *[19] The open problem 12 has a positive answer when* $\Gamma_0 = \Gamma$ *and* $l = 2$.

This result has been recently generalized by A. Fursikov and O. Immanuvilov in [32] to the case $l = 3$ Let us also mention that, in [24], C. Fabre has obtained, in every dimension, an approximate controllability result of two natural "cut off" Navier-Stokes equations. Her proof relies on a general method introduced by E. Zuazua in [75] to prove approximate controllability of semilinear wave equations. This general method is based on H.U.M. (Hilbert Uniqueness Method, due to J.–L. Lions [52]) and on a fixed point technique; see also [25] where C. Fabre, J.–P. Puel and E. Zuazua use this method to prove approximate controllability of semilinear heat equations.

Remark 18 *It is usually accepted that the viscous Burgers equation provides a realistic simplification of the Navier-Stokes system in fluid Mechanics. But J.I. Diaz has proved in [22] that the viscous Burgers equation is not approximately controllable; see also [28]. For the nonviscous Burgers equation, results have been obtained by F. Ancona and A. Marson in [3] and by Th. Horsin in [40].*

4 Local controllability of a 1-D tank containing a fluid modeled by the Saint-Venant equations

In this section, we consider a 1-D tank containing an inviscid incompressible irrotational fluid. The tank is subject to one-dimensional horizontal moves. We assume that the horizontal acceleration of the tank is small compared to the gravity constant and that the height of the fluid is small compared to the length of the tank. This motivates the use of the Saint-Venant [61] (also called shallow water) equations to describe the motion of the fluid; see e.g.

[21, Sec. 4.2]. Hence the dynamics equations considered are -see [23]-

$$H_t(t,x) + (Hv)_x(t,x) = 0, \tag{4.1}$$

$$v_t(t,x) + \left(gH + \frac{v^2}{2}\right)_x (t,x) = -u(t), \tag{4.2}$$

$$v(t,0) = v(t,L) = 0, \tag{4.3}$$

$$\frac{\mathrm{d}s}{\mathrm{d}t}(t) = u(t), \tag{4.4}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t}(t) = s(t), \tag{4.5}$$

where (see figure 1),

- L is the length of the 1-D tank,
- $H(t,x)$ is the height of the fluid at time t and for $x \in [0,L]$,
- $v(t,x)$ is the horizontal water velocity of the fluid *in a referential attached to the tank* at time t and for $x \in [0,L]$ (in the shallow water model, all the points on the same vertical have the same horizontal velocity),
- $u(t)$ is the horizontal acceleration of the tank in the absolute referential,
- g is the gravity constant,
- s is the horizontal velocity of the tank,
- D is the horizontal displacement of the tank.

This is a control system, denoted Σ, where

- the state is $Y = (H, v, s, D)$,
- the control is $u \in \mathbb{R}$.

Our goal is to study the local controllability of this control system Σ around the equilibrium point

$$(Y_e, u_e) := ((H_e, 0, 0, 0), 0).$$

This problem has been raised by F. Dubois, N. Petit and P. Rouchon in [23].

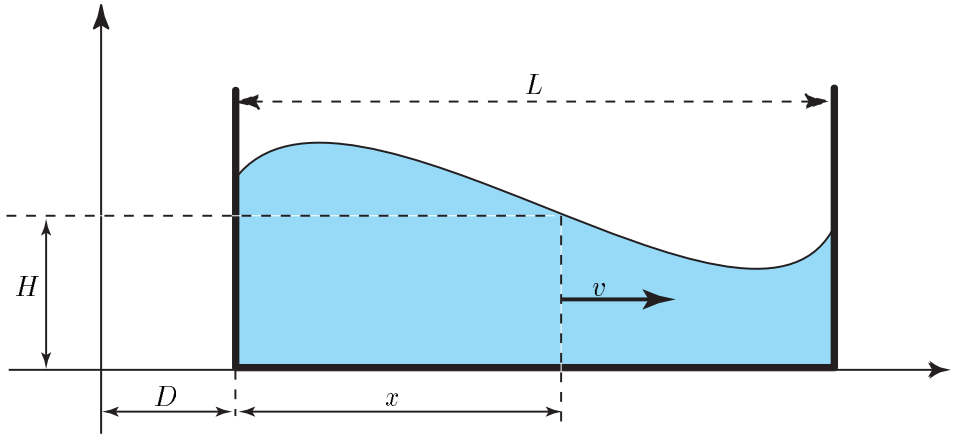


Figure 1: Fluid in the 1-D tank

Of course, the total mass of the fluid is conserved so that, for every solution of (4.1) to (4.3),

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^L H\,(t,x)\,\mathrm{d}x = 0. \tag{4.6}$$

(One gets (4.6) by integrating (4.1) on $[0, L]$ and by using (4.3) together with an integration by parts.) Moreover, if H and v are of class $\mathcal{C}^1$, it follows from (4.2) and (4.3) that

$$H_x(t,0) = H_x(t,L) \quad (= -u\,(t)\,/g). \tag{4.7}$$

Therefore we introduce the vector space E of functions $Y = (H, v, s, D) \in \mathcal{C}^1([0,L]) \times \mathcal{C}^1([0,L]) \times \mathbb{R} \times \mathbb{R}$ such that

$$H_x(0) = H_x(L), \tag{4.8}$$
$$v(0) = v(L) = 0, \tag{4.9}$$

and consider the affine subspace $\mathcal{Y} \subset E$ of $Y = (H, v, s, D) \in E$ satisfying

$$\int_0^L H(x)\mathrm{d}x = LH_e. \tag{4.10}$$

The vector space E is equipped with the usual norm

$$|Y| := |H|_1 + |v|_1 + |s| + |D|,$$

where, for $w \in \mathcal{C}^1([0,L])$,

$$|w|_1 := \mathrm{Max}\{|w(x)| + |w_x(x)|;\ x \in [0,L]\}.$$

With these notations, we can define a trajectory of the control system Σ.

Definition 19 *Let T_1 and T_2 be two real numbers satisfying $T_1 \leqslant T_2$. A function $(Y,u) = ((H,v,s,D),u) : [T_1,T_2] \to \mathcal{Y} \times \mathbb{R}$ is a trajectory of the control system Σ if*

(i) the functions H and v are of class $\mathcal{C}^1$ on $[T_1,T_2] \times [0,L]$,

(ii) the functions s and D are of class $\mathcal{C}^1$ on $[T_1,T_2]$ and the function u is continuous on $[0,T]$,

(iii) the equations (4.1) to (4.5) hold for every $(t,x) \in [T_1,T_2] \times [0,L]$.

Our main result states that the control system Σ is locally controllable around the equilibrium point (Y_e,u_e). More precisely, one has the following theorem.

Theorem 20 *There exists $T > 0$, $C_0 > 0$ and $\eta > 0$ such that, for every $Y_0 = (H_0,v_0,s_0,D_0) \in \mathcal{Y}$, and for every $Y_1 = (H_1,v_1,s_1,D_1) \in \mathcal{Y}$ such that*

$$|H_0 - H_e|_1 + |v_0|_1 < \eta,\ |H_1 - H_e|_1 + |v_1|_1 < \eta,\ |s_1 - s_0| + |D_1 - s_0 T - D_0| < \eta,$$

there exists a trajectory

$$(Y,u) : [0,T] \to \mathcal{Y} \times \mathbb{R},\ t \mapsto ((H(t), v(t), s(t), D(t)), u(t))$$

of the control system Σ such that

$$Y(0) = Y_0 \text{ and } Y(T) = Y_1, \tag{4.11}$$

and, for every $t \in [0,T]$,

$$\begin{aligned} |H(t) - H_e|_1 + |v(t)|_1 + |u(t)| < \\ C_0 \sqrt{|H_0 - H_e|_1 + |v_0|_1 + |H_1 - H_e|_1 + |v_1|_1} \\ + C_0 \left(|s_1 - s_0| + |D_1 - s_0 T - D_0|\right). \end{aligned} \tag{4.12}$$

As a corollary of this theorem, every steady state $Y_1 = (H_e, 0, 0, D_1)$ can be reached from every other steady state $Y_0 = (H_e, 0, 0, D_0)$. More precisely, one has the following corollary.

Corollary 21 *Let T, C_0 and η be as in Theorem 20. Let D_0 and D_1 be two real numbers and let $\eta_1 \in (0, \eta]$. Then, there exists a trajectory $(Y, u) : [0, T(|D_1 - D_0| + \eta_1)/\eta_1] \to \mathcal{Y} \times \mathbb{R}$, $t \mapsto ((H(t), v(t), s(t), D(t)), u(t))$ of the control system Σ such that*

$$Y(0) = (H_e, 0, 0, D_0) \text{ and } Y(T(|D_1 - D_0| + \eta_1)/\eta_1) = (H_e, 0, 0, D_1), \tag{4.13}$$

$$|H(t) - H_e|_1 + |v(t)|_1 + |u(t)| < C_0\eta_1 \ \forall t \in [0, T(|D_1 - D_0| + \eta_1)/\eta_1]. \tag{4.14}$$

Let us give the main steps of the proof of Theorem 20. Let us first point out that by scaling arguments one can assume without loss of generality that

$$L = g = H_e = 1. \tag{4.15}$$

Indeed, if we let

$$H^*(t,x) := \frac{1}{H_e} H\left(\frac{Lt}{\sqrt{H_e g}}, Lx\right), \ v^*(t,x) := \frac{1}{\sqrt{H_e g}} v\left(\frac{Lt}{\sqrt{H_e g}}, Lx\right),$$
$$u^*(t) := \frac{L}{H_e g} u\left(\frac{Lt}{\sqrt{H_e g}}\right), \ s^*(t) := \frac{1}{\sqrt{H_e g}} s\left(\frac{Lt}{\sqrt{H_e g}}\right),$$
$$D^*(t) := \frac{1}{L} D\left(\frac{Lt}{\sqrt{H_e g}}\right),$$

with $x \in [0, 1]$, then equations (4.1) to (4.5) are equivalent to

$$H^*_t(t,x) + (H^* v^*)_x(t,x) = 0,$$
$$v^*_t(t,x) + \left(H^* + \frac{v^{*2}}{2}\right)_x (t,x) = -u^*(t),$$
$$v^*(t,0) = v^*(t,1) = 0,$$
$$\frac{\mathrm{d}s^*}{\mathrm{d}t}(t) = u^*(t),$$
$$\frac{\mathrm{d}D^*}{\mathrm{d}t}(t) = s^*(t).$$

From now on, we always assume that we have (4.15). Since $(Y, u) = ((H, v, s, D), u)$ is a trajectory of the control system Σ if and only if $((H, v, s - a, D -$

$at-b), u)$ is a trajectory of the control system Σ, we may assume without loss of generality that $s_0 = D_0 = 0$.

The proof of Theorem 20 relies again on the return method. So one looks for a trajectory $(\bar{Y}, \bar{u}) : [0, T] \to \mathcal{Y} \times \mathbb{R}$ of the control system Σ satisfying

$$\bar{Y}(0) = \bar{Y}(T) = Y_e, \tag{4.16}$$

$$\text{the linearized control system around } (\bar{Y}, \bar{u}) \text{ is controllable.} \tag{4.17}$$

Let us point out that, as already noticed by F. Dubois, N. Petit and P. Rouchon in [23], property (4.17) does not hold for the natural trajectory $(\bar{Y}, \bar{u}) = (Y_e, u_e)$. Indeed the linearized control system around (Y_e, u_e) is

$$(\Sigma_0) \begin{cases} h_t + v_x = 0, \\ v_t + h_x = -u\,(t)\,, \\ v(t,0) = v(t,1) = 0, \\ \frac{\mathrm{d}s}{\mathrm{d}t}\,(t) = u\,(t)\,, \\ \frac{\mathrm{d}D}{\mathrm{d}t}\,(t) = s\,(t)\,, \end{cases} \tag{4.18}$$

where the state is $(h, v, s, D) \in \mathcal{Y}_0$, with

$$\mathcal{Y}_0 := \left\{ (h, v, s, D) \in E;\ \int_0^L h \mathrm{d}x = 0 \right\},$$

and the control is $u \in \mathbb{R}$. But (4.18) implies that, if

$$h(0, 1-x) = -h(0, x) \text{ and } v(0, 1-x) = v(0, x)\ \forall x \in [0, 1],$$

then

$$h(t, 1-x) = -h(t, x) \text{ and } v(t, 1-x) = v(t, x)\ \forall x \in [0, 1],\ \forall t.$$

Remark 22 *Even if the control system (4.18) is not controllable, one can move, as it is proved in [23], from every steady state* $(h_0, v_0, s_0, D_0) := (0, 0, 0, D_0)$ *to every steady state* $(h_1, v_1, s_1, D_1) := (0, 0, 0, D_1)$ *for this control system (see also [58] when the tank has a non-straight bottom). This does not imply that the related property (move from* $(1, 0, 0, D_0)$ *to* $(1, 0, 0, D_1)$*) also holds for the nonlinear control system* Σ*, but it follows from Corollary 21, that this property indeed also holds for the nonlinear control system* Σ*. Moreover the fact that, for the control system (4.18), it is possible –[23]–, to move from every steady state* $(h_0, v_0, s_0, D_0) := (0, 0, s_0, D_0)$ *to every steady*

state $(h_1, v_1, s_1, D_1) := (0, 0, s_1, D_1)$ *explains why in the right hand side of (4.12) one has*

$$|s_1 - s_0| + |D_1 - s_0 T - D_0|$$

and not

$$(|s_1 - s_0| + |D_1 - s_0 T - D_0|)^{1/2}.$$

As in [11, 13, 14, 19, 32, 34, 35, 67] one has to look for more complicated trajectories $(\bar{Y}, \bar{u})$ in order to have (4.17). In fact, as in [15], one can require instead of (4.16), the weaker property

$$\bar{Y}(0) = Y_e \text{ and } \bar{Y}(T) \text{ is close to } Y_e \tag{4.19}$$

and hope that, as it happens for the Navier-Stokes control system -see above and [15]-, the controllability around $(\bar{Y}, \bar{u})$ will be strong enough to tackle the problem that $\bar{Y}(T)$ is not Y_e but only close to Y_e. Moreover, since as it is proved in [23], one can move for the linear control system Σ_0, from $y_e := (0, 0, 0, 0)$ to $(0, 0, s_1, D_1)$, it is natural to try not to "return" to Y_e, but requires instead (4.19) the property

$$\bar{Y}(0) = Y_e \text{ and } \bar{Y}(T) \text{ is close to } (1, 0, s_1, D_1). \tag{4.20}$$

In order to use this method, one first needs to have trajectories of the control system Σ such that the linearized control system around these trajectories are controllable. Let us give an example of a family of such trajectories. Let us fix a positive real number T^* in $(2, +\infty)$. For $\gamma \in (0, 1]$ and $(a, b) \in \mathbb{R}^2$, let us define $(Y^{\gamma,a,b}, u^\gamma) : [0, T^*] \to \mathcal{Y} \times \mathbb{R}$ by requiring, for every $t \in [0, T^*]$ and for every $x \in [0, 1]$,

$$Y^{\gamma,a,b}(t, x) := \left(1 + \gamma\left(\frac{1}{2} - x\right), 0, \gamma t + a, \gamma\frac{t^2}{2} + at + b\right), \tag{4.21}$$

$$u^\gamma(t) := \gamma. \tag{4.22}$$

Then, $(Y^{\gamma,a,b}, u^\gamma)$ is a trajectory of the control system Σ. The linearized control system around this trajectory is the following control system

$$(\Sigma_\gamma) \begin{cases} h_t + \left(\left(1 + \gamma\left(\frac{1}{2} - x\right)\right) v\right)_x = 0, \\ v_t + h_x = -u(t), \\ v(t, 0) = v(t, 1) = 0, \\ \frac{\mathrm{d}s}{\mathrm{d}t}(t) = u(t), \\ \frac{\mathrm{d}D}{\mathrm{d}t}(t) = s(t) \end{cases} \tag{4.23}$$

where the state is $(h, v, s, D) \in \mathcal{Y}_0$ and the control is $u \in \mathbb{R}$. This linear control system Σ_γ is controllable if $\gamma > 0$ is small enough (see [18] for a proof). Unfortunately the controllability of Σ_γ does not seem to imply directly the local controllability of the control system Σ around the trajectory $(Y^{\gamma,a,b}, u^\gamma)$. Indeed the map from $\mathcal{Y} \times \mathcal{C}^0([0,T])$ into $\mathcal{Y}$ which associates to any initial data $Y_0 = (H_0, v_0, s_0, D_0) \in \mathcal{Y}$ and to any $u \in \mathcal{C}^0([0,T])$ such that

$$H_{0x}(0) = H_{0x}(1) = -u(0)$$

the state $Y(T) \in \mathcal{Y}$, where $Y = (H, v, s, D) : [0,T] \to \mathcal{Y}$ satisfies (4.1) to (4.5) and $Y(0) = Y_0$ is well-defined and continuous on a small open neighborhood of $(Y_e, 0)$ (see e.g. [50]) but is not of class $\mathcal{C}^1$ on this neighborhood. So one cannot use the classical inverse function theorem to get the desired local controllability. To take care of this problem, one adapts the usual iterative scheme used to prove the existence of solutions to hyperbolic systems (see e.g. [20, p. 476-478], [39, p. 54-55], [50, p. 96-107], [55, p. 35-43] or [62, p. 106-116] -see also [13, 14, 19, 32, 34, 35] for the Euler and the Navier control system for incompressible fluids): one uses the following inductive procedure $(h^n, v^n, s^n, D^n, u^n) \mapsto \left(h^{n+1}, v^{n+1}, s^{n+1}, D^{n+1}, u^{n+1}\right)$ so that

$$h_t^{n+1} + v^n h_x^{n+1} + \left(1 + \gamma\left(\frac{1}{2} - x\right) + h^n\right) v_x^{n+1} - \gamma v^{n+1} = 0 \tag{4.24}$$

$$v_t^{n+1} + h_x^{n+1} + v^n v_x^{n+1} = -u^{n+1}(t) \tag{4.25}$$

$$v^{n+1}(t,0) = v^{n+1}(t,L) = 0, \tag{4.26}$$

$$\frac{\mathrm{d}s^{n+1}}{\mathrm{d}t}(t) = u^{n+1}(t), \tag{4.27}$$

$$\frac{\mathrm{d}D^{n+1}}{\mathrm{d}t}(t) = s^{n+1}(t), \tag{4.28}$$

and $\left(h^{n+1}, v^{n+1}, s^{n+1}, D^{n+1}, u^{n+1}\right)$ has the required value for $t = 0$ and for $t = T^*$. Unfortunately we have only been able to prove that the control system (4.24)-(4.28), where the state is $\left(h^{n+1}, v^{n+1}, s^{n+1}, D^{n+1}\right)$ and the control is u^{n+1}, is controllable under a special assumption on (h^n, v^n), see [18]. Hence one has to insure that, at each iterative step, (h^n, v^n) satisfies this condition, which turns out to be possible. So one gets the following proposition, which is proved in [18].

Proposition 23 *There exist $C_1 > 0$,$\mu > 0$ and $\gamma_0 \in (0,1]$ such that, for every $\gamma \in [0, \gamma_0]$, for every $(a,b) \in \mathbb{R}^2$ and for every $(Y_0, Y_1) \in \mathcal{Y}^2$ satisfying*

$$|Y_0 - Y^{\gamma,a,b}(0)| \leqslant \mu\gamma^2 \text{ and } |Y_1 - Y^{\gamma,a,b}(T^*)| \leqslant \mu\gamma^2,$$

there exists a trajectory $(Y,u):\ [0,T^*]\to\mathcal{Y}\times\mathbb{R}$ *of the control system* Σ *such that*

$$Y(0)=Y_0 \text{ and } Y(T^*)=Y_1,$$
$$\left|Y(t)-Y^{\gamma,a,b}(t)\right|+|u(t)|\leqslant C_1\gamma\ \forall t\in[0,T^*]. \tag{4.29}$$

One now needs to construct, for every given $\gamma>0$ small enough, trajectories $(Y,u):\ [0,T^0]\to\mathcal{Y}\times\mathbb{R}$ of the control system Σ satisfying

$$Y(0)=(1,0,0,0) \text{ and } |Y(T^0)-Y^{\gamma,a,b}(0)|\leqslant\mu\gamma^2, \tag{4.30}$$

and trajectories $(Y,u):\ [T^0+T^*,T^0+T^*+T^1]\to\mathcal{Y}\times\mathbb{R}$ of the control system Σ such that

$$Y\left(T^0+T^1+T^*\right)=(1,0,s_1,D_1) \text{ and } \left|Y\left(T^0+T^*\right)-Y^{\gamma,a,b}(T^*)\right|\leqslant\mu\gamma^2, \tag{4.31}$$

for suitable choice of $(a,b)\in\mathbb{R}^2$, $T^0>0$, $T^1>0$. Let us first point out that it follows from [23] that one knows explicit trajectories $(Y^l,u^l):\ [0,T^0]\to\mathcal{Y}\times\mathbb{R}$ of *the linearized control system around* $(0,0)$ (i.e. the control system Σ_0) satisfying $Y^l(0)=0$ and $Y^l(T^0)=Y^{\gamma,a,b}(0)$. (In fact F. Dubois, N. Petit and P. Rouchon have proved in [23] that the linear control system Σ_0 is flat -a notion introduced by M. Fliess, J. Lévine, P. Martin and P. Rouchon in [26]-. They have given a complete explicit parametrization of the trajectories of Σ_0 by means of an arbitrary function and a 1-periodic function.) Then, the idea is that, if one moves "slowly", the same control u^l gives a trajectory $(Y,u):\ [0,T^0]\to\mathcal{Y}\times\mathbb{R}$ of the control system Σ such that (4.30) holds. More precisely, let $f_0\in\mathcal{C}^4([0,4])$ be such that

$$f_0=0 \text{ in } [0,1/2]\cup[3,4], \tag{4.32}$$
$$f_0(t)=s/2\ \forall t\in[1,3/2], \tag{4.33}$$
$$\int_0^4 f_0(t_1)\mathrm{d}t_1=0. \tag{4.34}$$

Similarly, let $f_1 \in \mathcal{C}^4([0,4])$ and $f_2 \in \mathcal{C}^4([0,4])$ be such that

$$f_1 = 0 \text{ in } [0,1/2] \cup [1,3/2] \text{ and } f_1 = 1/2 \text{ in } [3,4], \tag{4.35}$$

$$\int_0^3 f_1(t_1)\mathrm{d}t_1 = 0, \tag{4.36}$$

$$f_2 = 0 \text{ in } [0,1/2] \cup [1,3/2] \cup [3,4], \tag{4.37}$$

$$\int_0^4 f_2(t_1)\mathrm{d}t_1 = 1/2. \tag{4.38}$$

Let

$$\mathbb{D} := \left\{ \left(\bar{s}, \bar{D}\right) \in \mathbb{R}^2;\ |\bar{s}| \leqslant 1,\ \left|\bar{D}\right| \leqslant 1 \right\}.$$

For $\left(\bar{s}, \bar{D}\right) \in \mathbb{D}$, let $f_{\bar{s},\bar{D}} \in \mathcal{C}^4([0,4])$ be defined by

$$f_{\bar{s},\bar{D}} := f_0 + \bar{s} f_1 + \bar{D} f_2. \tag{4.39}$$

For $\epsilon \in (0,1/2]$ and for $\gamma \in \mathbb{R}$, let $u^{\epsilon,\gamma}_{\bar{s},\bar{D}} :\ [0,3/\epsilon] \to \mathbb{R}$ be defined by

$$u^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) := \gamma f'_{\bar{s},\bar{D}}(\epsilon t) + \gamma f'_{\bar{s},\bar{D}}(\epsilon(t+1)). \tag{4.40}$$

Let $\left(h^{\epsilon,\gamma}_{\bar{s},\bar{D}}, v^{\epsilon,\gamma}_{\bar{s},\bar{D}}, s^{\epsilon,\gamma}_{\bar{s},\bar{D}}, D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\right) :\ [0,3/\epsilon] \to \mathcal{C}^1([0,1]) \times \mathcal{C}^1([0,1]) \times \mathbb{R} \times \mathbb{R}$ be such that (4.18) holds for $(h,v,s,D) = \left(h^{\epsilon,\gamma}_{\bar{s},\bar{D}}, v^{\epsilon,\gamma}_{\bar{s},\bar{D}}, s^{\epsilon,\gamma}_{\bar{s},\bar{D}}, D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\right)$, $u = u^{\epsilon,\gamma}_{\bar{s},\bar{D}}$ and

$$\left(h^{\epsilon,\gamma}_{\bar{s},\bar{D}}(0,\cdot), v^{\epsilon,\gamma}_{\bar{s},\bar{D}}(0,\cdot), s^{\epsilon,\gamma}_{\bar{s},\bar{D}}(0), D^{\epsilon,\gamma}_{\bar{s},\bar{D}}(0)\right) = (0,0,0,0).$$

From [23] one gets that

$$h^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t,x) = -\frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon(t+x)) + \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon(t+1-x)), \tag{4.41}$$

$$\begin{aligned} v^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t,x) = \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon(t+x)) + \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon(t+1-x)) \\ - \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon t) - \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon(t+1)), \end{aligned} \tag{4.42}$$

$$s^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) = \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon t) + \frac{\gamma}{\epsilon} f_{\bar{s},\bar{D}}(\epsilon(t+1)), \tag{4.43}$$

$$D^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) = \frac{\gamma}{\epsilon^2} F_{\bar{s},\bar{D}}(\epsilon t) + \frac{\gamma}{\epsilon^2} F_{\bar{s},\bar{D}}(\epsilon(t+1)), \tag{4.44}$$

with

$$F_{\bar{s},\bar{D}}(t) := \int_0^t f_{\bar{s},\bar{D}}(t_1)\mathrm{d}t_1.$$

In particular, using also (4.32) to (4.38), one gets

$$h^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}+t,x\right)=\gamma\left(\frac{1}{2}-x\right)\ \forall t\in\left[0,\frac{1-2\epsilon}{2\epsilon}\right],\ \forall x\in[0,1], \tag{4.45}$$

$$v^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}+t,x\right)=0\ \forall t\in\left[0,\frac{1-2\epsilon}{2\epsilon}\right],\ \forall x\in[0,1], \tag{4.46}$$

$$s^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}+t\right)=\frac{\gamma}{\epsilon}+\frac{\gamma}{2}+\gamma t\ \forall t\in\left[0,\frac{1-2\epsilon}{2\epsilon}\right], \tag{4.47}$$

$$D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}+t\right)=D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right)+\left(\frac{\gamma}{\epsilon}+\frac{\gamma}{2}\right)t+\frac{\gamma}{2}t^2\ \forall t\in\left[0,\frac{1-2\epsilon}{2\epsilon}\right], \tag{4.48}$$

$$h^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{3}{\epsilon},x\right)=0\text{ and }v^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{3}{\epsilon},x\right)=0\ \forall x\in[0,1], \tag{4.49}$$

$$s^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{3}{\epsilon}\right)=\frac{\gamma}{\epsilon}\bar{s}\text{ and }D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{3}{\epsilon}\right)=\frac{\gamma}{2\epsilon}\bar{s}+\frac{\gamma}{\epsilon^2}\bar{D}. \tag{4.50}$$

Let $H^{\epsilon,\gamma}_{\bar{s},\bar{D}}=1+h^{\epsilon,\gamma}_{\bar{s},\bar{D}}$ and $Y^{\epsilon,\gamma}_{\bar{s},\bar{D}}=\left(H^{\epsilon,\gamma}_{\bar{s},\bar{D}},v^{\epsilon,\gamma}_{\bar{s},\bar{D}},s^{\epsilon,\gamma}_{\bar{s},\bar{D}},D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\right)$. Consider

$$a_{\epsilon,\gamma}:=\frac{\gamma}{\epsilon}f_{\bar{s},\bar{D}}(1)+\frac{\gamma}{\epsilon}f_{\bar{s},\bar{D}}(1+\epsilon)=\frac{\gamma}{\epsilon}+\frac{\gamma}{2},$$

$$b^{\bar{s},\bar{D}}_{\epsilon,\gamma}:=\frac{\gamma}{\epsilon^2}F_{\bar{s},\bar{D}}(1)+\frac{\gamma}{\epsilon^2}F_{\bar{s},\bar{D}}(1+\epsilon)=D^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right).$$

Using (4.21), (4.45), (4.46), (4.47), and (4.48), one has

$$Y^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right)=Y^{\gamma,a_{\epsilon,\gamma},b^{\bar{s},\bar{D}}_{\epsilon,\gamma}}(0,\cdot), \tag{4.51}$$

and, if $\epsilon\in(0,1/(2(T^*+1))]$,

$$Y^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}+T^*\right)=Y^{\gamma,a_{\epsilon,\gamma},b^{\bar{s},\bar{D}}_{\epsilon,\gamma}}(T^*). \tag{4.52}$$

The next proposition, which is proved in [18], shows that one can achieve (4.30) with $u=u^{\epsilon,\gamma}_{\bar{s},\bar{D}}$ for suitable choices of T^0, ϵ and γ.

Proposition 24 *There exists a constant $C_2>2$ such that, for every $\epsilon\in(0,1/C_2]$, for every $(\bar{s},\bar{D})\in\mathbb{D}$ and for every $\gamma\in[0,\epsilon/C_2]$, there exists one and only one map $\tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}:\ [0,1/\epsilon]\to\mathcal{Y}$ satisfying the two following conditions*

$$(\tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}},u^{\epsilon,\gamma}_{\bar{s},\bar{D}})\text{ is a trajectory of the control system }\Sigma\ \ (\text{on }[0,1/\epsilon]),$$

$$\tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}(0)=(1,0,0,0)\,,$$

and this unique map $\tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}$ *verifies*

$$\left|\tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) - Y^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t)\right| \leqslant C_2\epsilon\gamma^2 \ \forall t \in [0, 1/\epsilon]. \tag{4.53}$$

In particular, by (4.45) and (4.46),

$$\left|\tilde{v}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right)\right|_1 + \left|\tilde{h}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right) - \gamma\left(\frac{1}{2} - x\right)\right|_1 \leqslant C_2\epsilon\gamma^2. \tag{4.54}$$

Similarly, one has the following proposition, which shows that (4.31) is achieved with $u = u^{\epsilon,\gamma}_{\bar{s},\bar{D}}$ for suitable choices of T^1, ϵ and γ.

Proposition 25 *There exists a constant* $C_3 > 2(T^* + 1)$ *such that, for every* $\epsilon \in (0, 1/C_3]$, *for every* $(\bar{s}, \bar{D}) \in \mathbb{D}$, *and for every* $\gamma \in [0, \epsilon/C_3]$, *there exists one and only one map* $\hat{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}} : [(1/\epsilon) + T^*, 3/\epsilon] \to \mathcal{Y}$ *satisfying the two following conditions*

$$\left(\hat{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}, u^{\epsilon,\gamma}_{\bar{s},\bar{D}}\right) \text{ is a trajectory of the control system } \Sigma \ \ (\text{on } [(1/\epsilon) + T^*, 3/\epsilon]),$$

$$\hat{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{3}{\epsilon}\right) = \left(1, 0, \frac{\gamma}{\epsilon}\bar{s}, \frac{\gamma}{2\epsilon}\bar{s} + \frac{\gamma}{\epsilon^2}\bar{D}\right) = Y^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{3}{\epsilon}\right),$$

and this unique map $\hat{Y}^{\epsilon,\gamma}$ *verifies*

$$\left|\hat{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) - Y^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t)\right| \leqslant C_3\epsilon\gamma^2 \ \forall t \in [(1/\epsilon) + T^*, 3/\epsilon]. \tag{4.55}$$

In particular, by (4.45) and (4.46),

$$\left|\hat{v}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right)\right|_1 + \left|\hat{h}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right) - \gamma\left(\frac{1}{2} - x\right)\right|_1 \leqslant C_2\epsilon\gamma^2. \tag{4.56}$$

Let us choose

$$\epsilon := \ \text{Min}\left(\frac{1}{C_2}, \frac{1}{C_3}, \frac{\mu}{2C_2}, \frac{\mu}{2C_3}\right) \leqslant \frac{1}{2}. \tag{4.57}$$

Let us point out that there exists $C_4 > 0$ such that, for every $(\bar{s}, \bar{D}) \in \mathbb{D}$ and for every $\gamma \in [-\epsilon, \epsilon]$,

$$\left|H^{\epsilon,\gamma}_{\bar{s},\bar{D}}\right|_{\mathcal{C}^2([0,3/\epsilon]\times[0,1])} + \left|v^{\epsilon,\gamma}_{\bar{s},\bar{D}}\right|_{\mathcal{C}^2([0,3/\epsilon]\times[0,1])} \leqslant C_4, \tag{4.58}$$

which, with straightforward estimates, leads to the next proposition, whose proof is omitted.

Proposition 26 *There exists $C_5 > 0$ such that, for every $(\bar{s}, \bar{D}) \in \mathbb{D}$, for every $Y_0 = (H_0, v_0, s_0, D_0) \in \mathcal{Y}$ with*

$$|Y_0 - Y_e| \leqslant \frac{1}{C_5},\ s_0 = 0,\ D_0 = 0$$

and for every $\gamma \in [0, \epsilon/C_2]$, there exists one and only one $Y : [0, 1/\epsilon] \to \mathcal{Y}$ such that

$$(Y, u^{\varepsilon,\gamma}_{\bar{s},\bar{D}} - H_{0x}(0)) \text{ is a trajectory of the control system } \Sigma,$$
$$Y(0) = Y_0,$$

and this unique map Y satisfies

$$\left| Y(t) - \tilde{Y}^{\varepsilon,\gamma}_{\bar{s},\bar{D}}(t) \right| \leqslant C_5 |Y_0 - Y_e|,\ \forall t \in [0, 1/\epsilon].$$

Similarly, (4.58) leads to the following proposition.

Proposition 27 *There exists $C_6 > 0$ such that, for every $(\bar{s}, \bar{D}) \in \mathbb{D}$, for every $\gamma \in [0, \epsilon/C_3]$, and for every $Y_1 = (H_1, v_1, s_1, D_1) \in \mathcal{Y}$ such that*

$$\left| Y_1 - \left(1, 0, \frac{\gamma}{\epsilon}\bar{s}, \frac{\gamma}{2\epsilon}\bar{s} + \frac{\gamma}{\epsilon^2}\bar{D}\right) \right| \leqslant \frac{1}{C_6},\ s_1 = \frac{\gamma}{\epsilon}\bar{s},\ D_1 = \frac{\gamma}{2\epsilon}\bar{s} + \frac{\gamma}{\epsilon^2}\bar{D}$$

there exists one and only one $Y : [(1/\epsilon) + T^*, 3/\epsilon] \to \mathcal{Y}$ such that

$$(Y, u^{\varepsilon,\gamma}_{\bar{s},\bar{D}} - H_{1x}(0)) \text{ is a trajectory of the control system } \Sigma$$
$$Y(3/\epsilon) = Y_1,$$

and this unique map Y satisfies

$$\left| Y(t) - \hat{Y}^{\varepsilon,\gamma}_{\bar{s},\bar{D}}(t) \right| \leqslant C_6 \left| Y_1 - Y^{\varepsilon,\gamma}_{\bar{s},\bar{D}}(3/\epsilon) \right|,\ \forall t \in [(1/\epsilon) + T^*, 3/\epsilon].$$

Finally define

$$T := \frac{3}{\epsilon}, \tag{4.59}$$

$$\eta := \text{ Min} \left(\frac{\epsilon^2\mu}{2C_5(C_3^2 + C_2^2)}, \frac{\epsilon^2\mu}{2C_6(C_3^2 + C_2^2)}, \frac{\epsilon}{C_2}, \frac{\epsilon}{C_3}, \frac{1}{C_5}, \frac{1}{C_6}, \frac{\gamma_0^2\mu}{2C_5}, \frac{\gamma_0^2\mu}{2C_6}, \gamma_0 \right). \tag{4.60}$$

We want to check that Theorem 20 holds with these constants for a large enough C_0. Let $Y_0 = (H_0, v_0, 0, 0) \in \mathcal{Y}$ and $Y_1 = (H_1, v_1, s_1, D_1) \in \mathcal{Y}$ be such that

$$|H_0 - 1|_1 + |v_0|_1 \leqslant \eta,\ |H_1 - 1|_1 + |v_1|_1 \leqslant \eta,\ |s_1| + |D_1| \leqslant \eta. \tag{4.61}$$

Let

$$\gamma := \text{ Max } \left(\sqrt{\frac{2C_5}{\mu}} \sqrt{|H_0 - 1|_1 + |v_0|_1}, \sqrt{\frac{2C_6}{\mu}} \sqrt{|H_1 - 1|_1 + |v_1|_1}, |s_1| + |D_1| \right), \tag{4.62}$$

$$\bar{s} := \frac{\epsilon}{\gamma} s_1,\ \bar{D} := \frac{\epsilon^2}{\gamma} \left(D_1 - \frac{s_1}{2} \right), \tag{4.63}$$

so that, thanks to (4.50),

$$s^{\epsilon,\gamma}_{\bar{s},\bar{D}} = s_1,\ D^{\epsilon,\gamma}_{\bar{s},\bar{D}} = D_1. \tag{4.64}$$

Note that, by (4.57), (4.61), (4.62) and (4.63),

$$\left(\bar{s}, \bar{D}\right) \in \mathbb{D}. \tag{4.65}$$

By (4.60), (4.61) and (4.62), we obtain that

$$\gamma \in \left[0, \text{Min } \left(\frac{\epsilon}{C_2}, \frac{\epsilon}{C_3} \right) \right]. \tag{4.66}$$

Then, by Proposition 26, (4.60), (4.61) and (4.66), there exists a function $Y^0 = (H^0, v^0, s^0, D^0) :\ [0, 1/\epsilon] \to \mathcal{Y}$ such that

$$(Y^0, u^{\epsilon,\gamma}_{\bar{s},\bar{D}} - H_{0x}(0)) \text{ is a trajectory of the control system } \Sigma \text{ on } [0, 1/\epsilon], \tag{4.67}$$

$$Y^0(0) = Y_0, \tag{4.68}$$

$$\left| Y^0(t) - \tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) \right| \leqslant C_5 |Y_0 - Y_e| \ \forall t \in [0, 1/\epsilon]. \tag{4.69}$$

By (4.62) and (4.69),

$$\left| Y^0\left(\frac{1}{\epsilon}\right) - \tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right) \right| \leqslant \frac{\mu\gamma^2}{2}. \tag{4.70}$$

By Proposition 24, (4.57) and (4.66),

$$\left|\tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}\left(\frac{1}{\epsilon}\right) - Y^{\gamma,a_\epsilon,\gamma,b^{\bar{s},\bar{D}}_{\epsilon,\gamma}}(0)\right| \leqslant C_2\epsilon\gamma^2 \leqslant \frac{\mu\gamma^2}{2},$$

which, with (4.70), gives

$$\left|Y^0\left(\frac{1}{\epsilon}\right) - Y^{\gamma,a_\epsilon,\gamma,b^{\bar{s},\bar{D}}_{\epsilon,\gamma}}\right| \leqslant \mu\gamma^2. \tag{4.71}$$

Similarly, by Propositions 25 and 27, (4.57), (4.59), (4.60), (4.61), (4.62), (4.64) and (4.66), there exists $Y^1 = (H^1, v^1, s^1, D^1) : [(1/\epsilon) + T^*, T] \to \mathcal{Y}$ such that

$$(Y^1, u^{\epsilon,\gamma}_{\bar{s},\bar{D}} - H_{1x}(0)) \text{ is a trajectory of the control system } \Sigma \text{ on } [(1/\epsilon) + T^*, T], \tag{4.72}$$

$$Y^1(T) = Y_1, \tag{4.73}$$

$$\left|Y^1(t) - \tilde{Y}^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t)\right| \leqslant C_6\left|Y_1 - (1, 0, s_1, D_1)\right| \ \forall t \in [(1/\epsilon) + T^*, T], \tag{4.74}$$

$$\left|Y^1\left(\frac{1}{\epsilon} + T^*\right) - Y^{\gamma,a_\epsilon,\gamma,b^{\bar{s},\bar{D}}_{\epsilon,\gamma}}(T^*)\right| \leqslant \mu\gamma^2. \tag{4.75}$$

By (4.60), (4.61) and (4.62),

$$\gamma \leqslant \gamma_0. \tag{4.76}$$

From Proposition 23, (4.71), (4.75) and (4.76), there exists a trajectory $(Y^*, u^*) : [0, T^*] \to \mathcal{Y}$ of the control system Σ satisfying

$$Y^*(0) = Y^0\left(\frac{1}{\epsilon}\right), \tag{4.77}$$

$$\left|Y^*(t) - Y^{\gamma,a_\epsilon,\gamma,b^{\bar{s},\bar{D}}_{\epsilon,\gamma}}(t)\right| \leqslant C_1\mu\gamma \ \forall t \in [0, T^*], \tag{4.78}$$

$$Y^*(T^*) = Y^1\left(\frac{1}{\epsilon} + T^*\right). \tag{4.79}$$

The map $(Y, u) : [0, T] \to \mathcal{Y}$ defined by

$$(Y(t), u(t)) = (Y^0(t), u^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) - H_{0x}(0)) \ \forall t \in [0, 1/\epsilon],$$
$$(Y(t), u(t)) = (Y^*(t - (1/\epsilon)), u^*(t - (1/\epsilon))) \ \forall t \in [1/\epsilon, (1/\epsilon) + T^*],$$
$$(Y(t), u(t)) = (Y^1(t), u^{\epsilon,\gamma}_{\bar{s},\bar{D}}(t) - H_{1x}(0)) \ \forall t \in [(1/\epsilon) + T^*, T],$$

is a trajectory of the control system Σ which, by (4.68) and (4.73), satisfies (4.11). Finally the existence of $C_0 > 0$ such that (4.12) holds follows from the construction of (Y, u), (4.7), (4.29), (4.41) to (4.44), (4.53), (4.55), (4.61), (4.62), (4.69), (4.74) and (4.78).

5 Null asymptotic stabilizability of the 2-D Euler control system

In subsection 3.1 we have considered the problem of the controllability of the Euler control system of incompressible inviscid fluid in a bounded domain. In particular we have seen that, if the controls act on an arbitrarily small open subset of the boundary which meets every connected component of this boundary, then the Euler equation are exactly controllable.

For linear control systems, the exact controllability implies the asymptotic stabilizability by means of feedback laws. This is well known for linear control systems of finite dimension and, by M. Slemrod [64], J.–L. Lions [52], I. Lasiecka-R. Triggiani [48] and V. Komornik [46], it also holds in infinite dimension in very general cases. But, as pointed out by H.J. Sussmann in [70], by E.D. Sontag and H.J. Sussmann in [69], and by R.W. Brockett in [8], this is no longer true for *nonlinear* control systems, even of finite dimension. For example (see [8]) the nonlinear control system (2.6) is globally controllable but $0 \in \mathbb{R}^3$ cannot be, even locally, asymptotically stabilized by means of feedback laws. Let us also notice that, as in this counter-example, the linearized control system of the Euler equation around the origin is not controllable.

Therefore it is natural to ask what is the situation for the asymptotic stabilizability of the origin for the 2-D Euler equation of incompressible inviscid fluid in a bounded domain when the controls act on an arbitrarily small open subset of the boundary which meets every connected component of this boundary. In this section we are going to see that the null global asymptotic stabilizability by means of feedback laws holds if the domain is simply connected.

Let Ω be a nonempty bounded connected and simply connected subset of $\mathbb{R}^2$ of class $\mathcal{C}^\infty$ and let Γ_0 be a non empty open subset of the boundary $\partial\Omega$ of Ω. This set Γ_0 is the location of the control. Let y be the velocity field of the inviscid fluid contained in Ω. We assume that the fluid is incompressible,

so that

$$\text{div } y = 0. \tag{5.1}$$

Since Ω is simply connected, y is completely characterized by $\omega := \text{curl } y$ and $y.n$ on $\partial\Omega$ where n denotes the unit outward normal to $\partial\Omega$. For the problem of controllability, one does not really need to specify the control and the state: one considers the "Euler control system" as an under-determined system by requiring $y.n = 0$ on $\partial\Omega \setminus \Gamma_0$ instead of $y.n = 0$ on $\partial\Omega$ as for the uncontrolled usual Euler equation. For the stabilization problem, one needs to specify more precisely the control and the state. In this paper the state is ω. For the control there are at least two natural possibilities

(a) The control is $y.n$ on Γ_0 and the time derivative $\partial\omega/\partial t$ of the vorticity at the points of Γ_0 where $y.n < 0$, i.e. at the points where the fluid enters into the domain Ω,

(b) The control is $y.n$ on Γ_0 and the vorticity ω at the points where $y.n < 0$.

Let us point out that, by (5.1), in both cases $y.n$ has to satisfy $\int_{\partial\Omega} y.n = 0$. In this paper we study only case (a); for case (b), see [16].

Let us give stabilizing feedback laws. Let $g \in \mathcal{C}^\infty(\partial\Omega)$ be such that

$$\text{Support } g \subset \Gamma_0, \tag{5.2}$$

$$\Gamma_0^+ := \{g > 0\} \text{ and } \Gamma_0^- := \{g < 0\} \text{ are connected}, \tag{5.3}$$

$$g \neq 0, \tag{5.4}$$

$$\overline{\Gamma_0^+} \cap \overline{\Gamma_0^-} = \emptyset, \tag{5.5}$$

$$\int_{\partial\Omega} g = 0. \tag{5.6}$$

For every $f \in \mathcal{C}^0(\overline{\Omega})$, we denote

$$|f|_0 = \text{Max } \{|f(x)|\,; x \in \overline{\Omega}\}.$$

Our stabilizing feedback laws are

$$y.n = M\,|\omega|_0\, g \text{ on } \Gamma_0,$$

$$\frac{\partial\omega}{\partial t} = -M\,|\omega|_0\,\omega \text{ on } \Gamma_0^- \text{ if } |\omega|_0 \neq 0,$$

where $M > 0$ is large enough. With these feedback laws, a function $\omega : I \times \overline{\Omega} \to \mathbb{R}$, where I is an interval, is a solution of the closed loop system Σ if

$$\frac{\partial \omega}{\partial t} + \text{ div } (\omega y) = 0 \text{ in } \overset{\circ}{I} \times \Omega, \tag{5.7}$$

$$\text{div } y = 0 \text{ in } \overset{\circ}{I} \times \Omega, \tag{5.8}$$

$$\text{curl } y = \omega \text{ in } \overset{\circ}{I} \times \Omega, \tag{5.9}$$

$$y(t).n = M \, |\omega(t)|_0 \, g \text{ on } \partial\Omega, \, \forall t \in I, \tag{5.10}$$

$$\frac{\partial \omega}{\partial t} = -M \, |\omega(t)|_0 \, \omega \text{ on } \{t; \, \omega(t) \neq 0\} \times \Gamma_0^-. \tag{5.11}$$

where, for $t \in \Omega$, $\omega(t) : \overline{\Omega} \to \mathbb{R}$ and $y(t) : \overline{\Omega} \to \mathbb{R}^2$ are defined by requiring $\omega(t)(x) = \omega(t, x)$ and $y(t)(x) = y(t, x), \forall x \in \overline{\Omega}$. More precisely, the definition of a solution of system Σ is

Definition 28 *Let I be an interval. A function $\omega : I \to \mathcal{C}^0(\overline{\Omega})$ is a solution of system Σ if*

(i) $\omega \in \mathcal{C}^0(I; \mathcal{C}^0(\overline{\Omega}))(\cong \mathcal{C}^0(I \times \overline{\Omega}))$,

(ii) For $y \in \mathcal{C}^0(I \times \overline{\Omega}; \mathbb{R}^2)$ defined by requiring (5.8) and (5.9) in the sense of distributions and (5.10), one has (5.7) in the sense of distributions,

(iii) In the sense of distributions on the open manifold $\{t \in \overset{\circ}{I}; \, \omega(t) \neq 0\} \times \Gamma_0^-$ one has $\partial\omega/\partial t = -M \, |\omega(t)|_0 \, \omega$.

Our first theorem says that, for M large enough, the Cauchy problem for system Σ has at least one solution defined on $[0, +\infty)$ for every initial data in $\mathcal{C}^0(\overline{\Omega})$. More precisely one has

Theorem 29 *There exists $M_0 > 0$ such that, for every $M \geqslant M_0$, the following two properties hold*

(i) For every $\omega_0 \in \mathcal{C}^0(\overline{\Omega})$, there exists a solution of system Σ defined on $[0, +\infty)$ such that $\omega(0) = \omega_0$,

(ii) Any maximal solution of system Σ defined at time 0 is defined on $[0, +\infty)$ (at least).

Remark 30 *a. In this theorem, property (i) is in fact implied by property (ii) and Zorn's lemma. We state (i) in order to emphasize the existence of a solution to the Cauchy problem for system* Σ. *b. We do not know if the solution to the Cauchy problem is unique for positive time. (For negative time, one does not have uniqueness since there are solutions* ω *of system* Σ *defined on* $[0, +\infty)$ *such that* $\omega(0) \neq 0$ *and* $\omega(T) = 0$ *for* $T \in [0, +\infty)$ *large enough.) But let us emphasize that, already for control system in finite dimension, one considers feedback laws which are merely continuous; with these feedback laws, the Cauchy problem for the closed loop system may have many solutions. It turns out that this lack of uniqueness is not a real problem. Indeed, in finite dimension at least, if a point is asymptotically stable for a continuous vector field, then there exists, as in the case of regular vector fields, a (smooth) strict Lyapounov function. This result is due to Kurzweil [47].It is tempting to conjecture that a similar result hold in infinite dimension under reasonable assumptions. The existence of this Lyapounov function insures some robustness to perturbations. This is precisely this robustness which makes the interest of feedback laws compared to open loop controls. We will see that, for our feedback laws, there exists also a strict Lyapounov –see Proposition 34 below– and therefore our feedback laws provide some kind of robustness.*

Our next theorem shows that, at least for M large enough, our feedback laws globally and strongly asymptotically stabilize the origin in $\mathcal{C}^0(\overline{\Omega})$ for system Σ.

Theorem 31 *There exists a positive constant* $M_1 \geqslant M_0$ *such that, for every* $\varepsilon \in (0, 1]$, *every* $M \geqslant M_1/\varepsilon$ *and every maximal solution* ω *of system* Σ *defined at time 0,*

$$|\omega(t)|_0 \leqslant \ Min\ \{|\omega(0)|_0, \frac{\varepsilon}{t}\},\ \forall t > 0. \tag{5.12}$$

Remark 32 *Due to the term* $|\omega(t)|_0$ *appearing in (5.10) and in (5.11) our feedback laws do not depend only on the value of* ω *on* Γ_0. *Let us point out that there is no asymptotically stabilizing feedback law depending only on the value of* ω *on* Γ_0 *such that the origin is asymptotically stable for the closed loop system. In fact, given a nonempty open subset* Ω_0 *of* Ω, *there is no feedback law which does not depend on the values of* ω *on* Ω_0. *This phenomenon is due to the existence of "phantom vortices": there are smooth stationary solutions* $\bar{y} : \overline{\Omega} \to \mathbb{R}^2$ *of the 2-D Euler equations such that Support*

$\bar{y} \subset \Omega_0$ and $\bar{\omega} :=$ curl $\bar{y} \neq 0$; see, e.g., [56]. Then $\omega(t) = \bar{\omega}$ is a solution of the closed loop system if the feedback law does not depend on the values of ω on Ω_0 –and vanishes for $\omega = 0$.

Remark 33 *Let us emphasize that (5.12) implies that*

$$|\omega(t)|_0 \leqslant \varepsilon, \ \forall t \geqslant 1, \tag{5.13}$$

for every maximal solution ω of system Σ defined at time 0 (whatever is $\omega(0)$). It would be intreresting to know if one could have a similar result for the 2-D Navier-Stokes equations of viscous incompressible flows, that is if, given $\varepsilon > 0$, does there exist a feedback law such that (5.13) holds for every solution of the closed loop Navier-Stokes control system? Note that $y = 0$ on Γ_0 is a feedback which leads to asymptotic stabilization of the null solution of the Navier-Stokes control system. But this feedback does not have the required property. One may ask a similar question for the Burgers control system; for the null asymptotic stabilization of this control system, see the paper [45] by M. Krstić and the references therein.

The detailed proofs of Theorem 29 and of Theorem 31 are given in [16]. Let us just mention that Theorem 31 is proved by giving an explicit Lyapounov function. Let us give this Lyapounov function. Let $V : \mathcal{C}^0(\overline{\Omega}) \to [0, +\infty)$ be defined by

$$V(\omega) = |\omega \exp(-\theta)|_0,$$

where $\theta \in \mathcal{C}^\infty(\overline{\Omega})$ satisfies

$$\Delta\theta = 0 \text{ in } \overline{\Omega}, \tag{5.14}$$

$$\frac{\partial\theta}{\partial n} = g \text{ on } \partial\Omega. \tag{5.15}$$

(Let us point out that the existence of θ follows from (5.6).) Theorem 31 is an easy consequence of the following proposition.

Proposition 34 *There exists $M_2 \geqslant M_0$ and $\mu > 0$ such that, for every $M \geqslant M_2$ and every solution $\omega : [0, +\infty) \to \mathcal{C}^0(\overline{\Omega})$ of system Σ, one has, for every $t \in [0, +\infty)$,*

$$[-\infty, 0] \ni \dot{V}(t) := \frac{d}{dt^+} V(\omega(t)) \leqslant -\mu M V^2(\omega(t)), \tag{5.16}$$

where $d/dt^+ V(\omega(t)) := \lim_{\varepsilon \to 0^+} (V(\omega(t+\varepsilon)) - V(\omega(t)))/\varepsilon$.

Let us end this section by some comments for the case where Ω is not simply connected In this case, in order to define the state, one adds to ω the real numbers $\lambda_1, \cdots, \lambda_g$ defined by

$$\lambda_i = \int y.\nabla^\perp \tau_i,$$

where, if one denotes by $\mathcal{C}_0, \mathcal{C}_1, \cdots, \mathcal{C}_g$ the connected components of Γ, the functions $\tau_i \in \mathcal{C}^\infty(\overline{\Omega}), i \in \{1, \cdots, g\}$are defined by

$$\begin{aligned} \Delta\tau_i &= 0 \ , \\ \tau_i &= 0 \text{ on } \partial\Omega \setminus \mathcal{C}_i, \\ \tau_i &= 1 \text{ on } \mathcal{C}_i, \end{aligned}$$

and where $\nabla^\perp \tau_i$ denotes $\nabla\tau_i$ rotated by $\pi/2$. One has the following open problem

Open Problem 35 *Assume that $g \geq 1$ and that Γ_0 meets every connected component of Γ. Does there exist always a feedback law such that $0 \in \mathcal{C}^0(\overline{\Omega}) \times \mathbb{R}^g$ is globally asymptotically stable for the closed loop system?*

Brockett's necessary condition [8] for the existence of asymptotically stabilizing feedback laws cannot be directly applied to our situation since our control system is of infinite dimension. But it leads to the following question.
Question *Assume that Γ_0 meets every connected component of Γ. Let $f \in \mathcal{C}^\infty(\overline{\Omega})$. Does there exit $y \in \mathcal{C}^\infty(\overline{\Omega}; \mathbb{R}^2)$ and $p \in \mathcal{C}^\infty(\overline{\Omega})$ such that*

$$(y.\nabla)y + \nabla p = f \ \textit{in} \ \overline{\Omega} \ , \tag{5.17}$$

$$\text{div } y = 0 \ \textit{in} \ \overline{\Omega} \ , \tag{5.18}$$

$$y.n = 0 \ \textit{on} \ \Gamma \setminus \Gamma_0 ? \tag{5.19}$$

Let us point out that, by scaling arguments, one does not have to assume that f is "small" in this question. It turns out that the answer to this question is indeed positive. This has been proved in [17] if Ω is simply connected and by O. Glass in [37] for the general case.

References

[1] R.A. Adams: *Sobolev spaces*, Academic Press, San Diego, London, 1978.

[2] A. Agrachev, Newton diagrams and tangent cones to attainable sets, in: *Analysis of Controlled Dynamical Systems (Lyon 1990)* (B. Bonnard et al., eds.), Progr. Systems Control Theory **8**, Birkhäuser, Boston, 1991, pp. 1-12.

[3] F. Ancona and A. Marson, On the Attainable Set for Scalar Nonlinear Conservation Laws with Boundary Control, *SIAM J. Control Optim.* 36 (1997) pp. 290-312.

[4] C. Bardos, F. Golse, and D. Levermore, Fluid dynamic limits of kinetic equations I: formal derivations, *J. Statistical Physics*, 63 (1991) pp. 323-344.

[5] R.M. Bianchini and G. Stefani, Sufficient conditions for local controllability, in: *Proc. 25th IEEE Conf. Decision and Control (Athens 1986)*, IEEE, New York, pp. 967-970.

[6] R.M. Bianchini and G. Stefani, Controllability along a trajectory: a variational approach, *SIAM J. Control Optim.* 31 (1993) pp. 900-927.

[7] R.M. Bianchini, Higher order necessary optimality conditions, preprint, 1999.

[8] R.W. Brockett, Asymptotic stability and feedback stabilization, in: *Differential Geometric Control Theory* (R.W. Brockett, R.S. Millman and H.J. Sussmann, eds.), Progr. Math. **27**, Birkhäuser, Basel-Boston, 1983, 181-191.

[9] W.L. Chow, Uber systeme von linearen partiellen differentialgleichung ester ordnung, *Math. Ann.*, 117 (1940-41) pp. 227-232.

[10] F. Coron, Derivation of slip boundary conditions for the Navier-Stokes system from the Boltzmann equation, *J. Statistical Physics*, 54 (1989) pp. 829-857.

[11] J.–M. Coron, Global asymptotic stabilization for controllable systems without drift, *Math. Control Signals Systems*, 5 (1992) pp. 295-312.

[12] J.–M. Coron, Linearized controlled systems and applications to smooth stabilization, *SIAM J. Control Optim.*, 32 (1994) pp. 358-386.

[13] J.–M. Coron, Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles bidimensionnels, *C.R. Acad. Sci. Paris*, 317 (1993) pp. 271-276.

[14] J.–M. Coron, On the controllability of 2-D incompressible perfect fluids, *J. Math. Pures & Appliquées*, 75 (1996) pp. 155-188.

[15] J–M. Coron, On the controllability of the 2-D incompressible Navier-Stokes equations with the Navier slip boundary conditions, *ESAIM: COCV, www.emath.fr/cocv/*, 1 (1996) pp. 35-75.

[16] J–M. Coron, On the null asymptotic stabilization of the 2-D incompressible Euler equations in a simply connected domain, *SIAM J. Control Optim.*, 37 (1999) pp. 1874-1896

[17] J–M. Coron, Sur la stabilisation des fluides parfaits incompressibles bidimensionnels, *Séminaire Équations aux Dérivées Partielles*, 1998-1999, École polytechnique, Centre de Mathématiques, exposé 7.

[18] J–M. Coron, Local controllability of a 1-D tank containing a fluid modeled by the shallow water equations, Preprint 2001-28, Université Paris-Sud, to appear in ESAIM: COCV, 8 (2002).

[19] J–M. Coron and A. Fursikov, Global exact controllabilty of the 2D Navier-Stokes equations on a manifold without boundary, *Russian Journal of Mathematical Physics*, 4 (1996) pp. 429-448.

[20] R. Courant and D. Hilbert, *Methods of mathematical physics, II*, Interscience publishers, John Wiley & Sons, New York London Sydney, 1962.

[21] L. Debnath, *Nonlinear water waves*, Academic Press, San Diego, 1994.

[22] J.I. Diaz, Sobre la controlabilidad approximada de problemas no lineales disipativos, in Proceedings of *Jornadas Hispano-Francesas sobre Control de Sistemas Distribuidos*, A. Valle ed., (1991) pp. 41-48.

[23] F. Dubois, N. Petit and P. Rouchon, Motion planning and nonlinear simulations for a tank containing a fluid, ECC 99.

[24] C. Fabre, Uniqueness results for Stokes equations and their consequences in linear and nonlinear control problems, *ESAIM: COCV, www.emath.fr/cocv/*, 1 (1996) pp. 267-302.

[25] C. Fabre, J.–P. Puel and E. Zuazua, Approximate controllability for semilinear heat equation, Proc. Royal Soc. Edinbourgh, 125A (1995) pp. 31-61.

[26] M. Fliess, J. Lévine, P. Martin and P. Rouchon, Flatness and defect of nonlinear systems: introductory theory and examples, *Internat. J. Control*, 61 (1995) pp. 1327-1361.

[27] A. Fursikov, Exact boundary zero controllability of three-dimensional Navier-Stokes equations, *J. Dynamical Control and Systems*, 1 (1995) pp. 325–350.

[28] A. Fursikov and O. Yu. Imanuvilov, On controllability of certain systems simulating a fluid flow, in: *Flow Control*, IMA vol. in Math. and its Appl., M.D. Gunzburger ed., Springer Verlag, New York, 68 (1994) pp. 149-184.

[29] A. Fursikov and O. Yu. Imanuvilov, Local exact controllability of the Navier-Stokes equations, *C. R. Acad. Sci. Paris*, 323 (1996) pp. 275-280.

[30] A. Fursikov and O. Yu. Imanuvilov, On exact boundary zero controllability of the two-dimensional Navier-Stokes equations *Acta Appl. Math.* 36 (1994) pp. 1-10.

[31] A. Fursikov and O. Yu. Imanuvilov, Local exact controllability for $2-D$ Navier-Stokes equations, *Sbornik Math.*, 187 (1996) (in Russian).

[32] A. V. Fursikov and O. Yu. Imanuvilov, Exact controllability of the Navier-Stokes and Boussinesq equations, *Russian math. surveys*, 54 (1999) pp. 565-618.

[33] G. Geymonat and E. Sanchez-Palencia, On the vanishing viscosity limit for acoustic phenomena in a bounded region, *Arch. Rat. Mechanics and Analysis*, 75 (1981) pp. 257-268.

[34] O. Glass, Contrôlabilité exacte frontière de l'équation d'Euler des fluides parfaits incompressibles en dimension 3, *C.R. Acad. Sci. Paris*, t. 325, Série I, (1997) pp. 987-992.

[35] O. Glass, Exact boundary controllability of 3-D Euler equation, *ESAIM: COCV, www.emath.fr/cocv/*, 5 (2000) pp. 1-44.

[36] O. Glass, An addendum to J.-M. Coron's theorem concerning the controllability of the Euler system for 2D incompressible inviscid fluids, Preprint 1999, Université Paris-Sud, to appear in *J. Math. Pures & Appliquées.*

[37] O. Glass, Existence of steady states for the two-dimensional Euler system for ideal fluids with arbitrary force, Preprint 2000, Université Pierre et Marie Curie.

[38] H. Hermes, Control systems which generate decomposable Lie algebras, *J. Differential Equations*, 44 (1982) pp. 166-187.

[39] L. Hörmander, *Lectures on nonlinear hyperbolic differential equations*, Mathématiques & Applications, 26, Springer-Verlag, Berlin Heidelberg, 1997.

[40] Th. Horsin, On the controllability of the Burgers equation, *ESAIM: COCV, www.emath.fr/cocv/*, 3 (1998) pp. 83-95.

[41] O. Yu. Imanuvilov, On exact controllability for Navier Stokes equations, *ESAIM: COCV, www.emath.fr/cocv/*, 3 (1998), pp. 97-131.

[42] O. Yu. Imanuvilov, Remarks on exact controllability for the Navier-Stokes equations, *ESAIM: COCV, www.emath.fr/cocv/*, 6 (2001), pp. 39-72.

[43] M. Kawski, High-order small time local controllability, in: *Nonlinear Controllability and Optimal Control* (H.J. Sussmann, ed.), Monogr. Textbooks Pure Appl. Math. **113**, Dekker, New York, 1990, pp. 431-467.

[44] A.V. Kazhikov, Note on the formulation of the problem of flow through a bounded region using equations of perfect fluid, *PMM USSR*, 44 (1981) pp. 672-674.

[45] M. Krstić, On global stabilization of Burgers' equation, *Systems and Control Letters.* 37 (1999) pp. 123-141.

[46] V. Komornik, Rapid boundary stabilization of linear distributed systems, *SIAM J. Control Optim.*, 35 (1997) pp. 1591-1613.

[47] J. Kurzweil, On the inversion of Lyapunov's second theorem on stability of motion, *Ann. Math. Soc. Trans.* Ser.2, 24 (1956) pp. 19-77.

[48] I. Lasiecka and R. Triggiani, *Differential and Algebraic Ricatti Equations with Applications to Boundary/Point Control Problems: Continuous and Approximation Theory*, Lecture Notes in Control and Information Sciences, vol. 164, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[49] B.E. Launder and D.B. Spalding, *Mathematical models of turbulence*, Academic Press, 1972.

[50] Li Ta Tsien and Yu Wen-Ci, *Boundary value problems for quasilinear hyperbolic systems*, Mathematics series V, Duke university, Durham, 1985.

[51] J.–L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Gauthier-Villars, Paris, 1968.

[52] J.–L. Lions, Exact controllability, stabilization and perturbations for distributed systems, J. Von Neumann Lecture, Boston 1986, *SIAM review*, 30 (1988) pp.1-68.

[53] J.-L. Lions, Are there connections between turbulence and controllability?, in: *9th INRIA International Conference*, Antibes, June 12-15, 1990.

[54] J.-L. Lions, Exact controllability for distributed systems. Some trends and some problems, in: *Applied and Industrial Mathematics*, R. Spigler ed., Kluwer Academic Publishers, Dodrecht, Boston, London, (1991) pp. 59-84.

[55] A. Majda, *Compressible fluid flow and systems of conservation laws in several space variables*, Applied Math. Sciences, vol. 53, Sringer-Verlag, New York Berlin Heidelberg Tokyo, 1984.

[56] A. Majda, Vorticity and the mathematical theory of incompressible fluid flow, *Comm. Pure Appl. Math.*, 39, special issue, (1986) pp. 187-220.

[57] C.L.M.H. Navier, Sur les lois du mouvement des fluides, *Mem. Acad. R. Sci. Inst. France*, 6 (1823) pp. 389-440.

[58] N. Petit and P. Rouchon, Dynamics and solutions to some control problems for water-tank systems, Preprint, CIT-CDS 00-004.

[59] P.K. Rashevski, About connecting two points of complete nonholonomic space by admissible curve, *Uch Zapiski ped. inst. Libknexta*, 2 (1938), p. 83-94.

[60] D.L. Russell, Exact boundary value controllability theorems for wave and heat processes in star-complemented regions, in *Differential Games and Control Theory*, Roxin, Liu et Sternberg eds., Marcel Dekker, New York, 1974, pp. 291-319.

[61] A.J.C.B. de Saint-Venant, Théorie du mouvement non permanent des eaux, avec applications aux crues des rivières et à l'introduction des marées dans leur lit, *C.R. Acad. Sci. Paris*, 53 (1871) pp. 147-154.

[62] D. Serre, *Systèmes de lois de conservations, I et II*, Diderot Editeur, Arts et Sciences, Paris New York Amsterdam, 1996.

[63] L.M. Silverman, H.E. Meadows, Controllability and observability in time variable linear systems, *SIAM J. on Control and Optimization*, 5 (1967) pp. 64-73.

[64] M. Slemrod, A note on complete controllability and stabilizability for linear control systems in Hilbert space, *SIAM J. Control*, 12 (1974) pp. 500-508.

[65] E.D. Sontag, Finite dimensional open-loop control generator for nonlinear systems, *Int. J. Control*, 47 (1988) pp. 537-556.

[66] E.D. Sontag, *Mathematical Control Theory - Deterministic Finite Dimensional Systems*, 2nd ed., Texts Appl. Math. **6**, Springer-Verlag, New York, 1998.

[67] E. D. Sontag, Control of systems without drift via generic loops, *IEEE Transactions on Automatic Control*, 40 (1995) pp. 1210-1219.

[68] E.D. Sontag, Universal nonsingular controls, *Systems and Control Letters*, 19 (1992), pp. 221-224.

[69] E.D. Sontag and H.J. Sussmann, Remarks on continuous feedback, in: *Proc. IEEE Conf. Decision and Control*, Albuquerque, (1980) pp. 916-921.

[70] H.J. Sussmann, Subanalytic sets and feedback control, *J. Differential Equations*, 31 (1979) pp. 31-52.

[71] H.J. Sussmann, Lie brackets and local controllability : a sufficient condition for scalar-input systems, *SIAM J. Control Optim.*, 21 (1983) pp. 686-713.

[72] H.J. Sussmann, A general theorem on local controllability, *SIAM J. Control Optim.*, 25 (1987) pp. 158-194.

[73] H.J. Sussmann and V. Jurdjevic, Controllability of nonlinear systems, *J. Differential Equations*, 12 (1972) pp. 95-116.

[74] A.I. Tret'yak, On odd-order necessary conditions for optimality in a time-optimal control problem for systems linear in the control, *Math. USSR Sbornik*, 79 (1991) pp. 47-63.

[75] E. Zuazua, Exact controllability for semilinear wave equatios in one space dimension, *Ann. Inst. Henri Poincaré, Nonlinear Analysis*, 10 (1993), pp. 109-129.

Flat Systems

Ph. Martin[1]*, R.M. Murray[2]†and P. Rouchon[1]‡

[1] *Centre Automatique et Systèmes, École des Mines de Paris, Fontainebleau, France*
[2] *Division of Engineering and Applied Science, California Institute of Technology, Pasadena, USA*

Lectures given at the Summer School on Mathematical Control Theory Trieste, 3-28 September 2001

LNS028011

*martin@cas.ensmp.fr
†murray@indra.caltech.edu
‡rouchon@cas.ensmp.fr

Contents

1 Equivalence and flatness **712**

1.1 Control systems as infinite dimensional vector fields 712
1.2 Equivalence of systems 716
1.3 Differential Flatness 719
1.4 Application to motion planning 721
1.5 Motion planning with singularities 723

2 Feedback design with equivalence **724**

2.1 From equivalence to feedback 724
2.2 Endogenous feedback 727
2.3 Tracking: feedback linearization 728
2.4 Tracking: singularities and time scaling 730
2.5 Tracking: flatness and backstepping 731
2.5.1 Some drawbacks of feedback linearization 731
2.5.2 Backstepping 732
2.5.3 Blending equivalence with backstepping 733
2.5.4 Backstepping and time-scaling 737
2.5.5 Conclusion 738

3 Open problems and new perspectives **738**

3.1 Checking flatness: an overview 738
3.1.1 The general problem 738
3.1.2 Known results 740
3.2 Infinite dimension "flat" systems 743
3.2.1 Delay systems 743
3.2.2 Distributed parameters systems 744
3.3 State constraints and optimal control 748
3.3.1 Optimal control 748
3.3.2 State constraints 749
3.4 Symmetries 750
3.4.1 Symmetry preserving flat output 750
3.4.2 Flat outputs as potentials and gauge degree of freedom 752

4 A catalog of flat systems **752**

4.1 Holonomic mechanical systems 752
4.2 Nonholonomic mechanical systems 758
4.3 Electromechanical systems 759

4.4 Chemical systems 760

References **762**

Introduction

Control systems are ubiquitous in modern technology. The use of feedback control can be found in systems ranging from simple thermostats that regulate the temperature of a room, to digital engine controllers that govern the operation of engines in cars, ships, and planes, to flight control systems for high performance aircraft. The rapid advances in sensing, computation, and actuation technologies is continuing to drive this trend and the role of control theory in advanced (and even not so advanced) systems is increasing.

A typical use of control theory in many modern systems is to invert the system dynamics to compute the inputs required to perform a specific task. This inversion may involve finding appropriate inputs to steer a control system from one state to another or may involve finding inputs to follow a desired trajectory for some or all of the state variables of the system. In general, the solution to a given control problem will not be unique, if it exists at all, and so one must trade off the performance of the system for the stability and actuation effort. Often this tradeoff is described as a cost function balancing the desired performance objectives with stability and effort, resulting in an optimal control problem.

This inverse dynamics problem assumes that the dynamics for the system are known and fixed. In practice, uncertainty and noise are always present in systems and must be accounted for in order to achieve acceptable performance of this system. Feedback control formulations allow the system to respond to errors and changing operating conditions in real-time and can substantially affect the operability of the system by stabilizing the system and extending its capabilities. Again, one may formulate the feedback regulation problems as an optimization problem to allow tradeoffs between stability, performance, and actuator effort.

The basic paradigm used in most, if not all, control techniques is to exploit the mathematical structure of the system to obtain solutions to the inverse dynamics and feedback regulation problems. The most common structure to exploit is linear structure, where one approximates the given system by its linearization and then uses properties of linear control systems combined with appropriate cost function to give closed form (or at least numerically computable) solutions. By using different linearizations around different operating points, it is even possible to obtain good results when the system is nonlinear by "scheduling" the gains depending on the operating point.

As the systems that we seek to control become more complex, the use of linear structure alone is often not sufficient to solve the control problems that are arising in applications. This is especially true of the inverse dynamics problems, where the desired task may span multiple operating regions and hence the use of a single linear system is inappropriate.

In order to solve these harder problems, control theorists look for different types of structure to exploit in addition to simple linear structure. In this paper we concentrate on a specific class of systems, called "(differentially) flat systems", for which the structure of the trajectories of the (nonlinear) dynamics can be completely characterized. Flat systems are a generalization of linear systems (in the sense that all linear, controllable systems are flat), but the techniques used for controlling flat systems are much different than many of the existing techniques for linear systems. As we shall see, flatness is particularly well tuned for allowing one to solve the inverse dynamics problems and one builds off of that fundamental solution in using the structure of flatness to solve more general control problems.

Flatness was first defined by Fliess et al. [13, 16] using the formalism of differential algebra, see also [33] for a somewhat different approach. In differential algebra, a system is viewed as a differential field generated by a set of variables (states and inputs). The system is said to be flat if one can find a set of variables, called the flat outputs, such that the system is (non-differentially) algebraic over the differential field generated by the set of flat outputs. Roughly speaking, a system is flat if we can find a set of outputs (equal in number to the number of inputs) such that all states and inputs can be determined from these outputs without integration. More precisely, if the system has states $x \in \mathbb{R}^n$, and inputs $u \in \mathbb{R}^m$ then the system is flat if we can find outputs $y \in \mathbb{R}^m$ of the form

$$y = h(x, u, \dot{u}, \ldots, u^{(r)})$$

such that

$$\begin{aligned} x &= \varphi(y, \dot{y}, \ldots, y^{(q)}) \\ u &= \alpha(y, \dot{y}, \ldots, y^{(q)}). \end{aligned}$$

More recently, flatness has been defined in a more geometric context, where tools for nonlinear control are more commonly available. One approach is to use exterior differential systems and regard a nonlinear control system as a Pfaffian system on an appropriate space [51]. In this context,

flatness can be described in terms of the notion of absolute equivalence defined by E. Cartan [6, 7, 70].

In this paper we adopt a somewhat different geometric point of view, relying on a Lie-Bäcklund framework as the underlying mathematical structure. This point of view was originally described by Fliess et al. in 1993 [14] and is related to the work of Pomet et al. [57, 55] on "infinitesimal Brunovsky forms" (in the context of feedback linearization). It offers a compact framework in which to describe basic results and is also closely related to the basic techniques that are used to compute the functions that are required to characterize the solutions of flat systems (the so-called flat outputs).

Applications of flatness to problems of engineering interest have grown steadily in recent years. It is important to point out that many classes of systems commonly used in nonlinear control theory are flat, see for instance the examples in section 4. As already noted, all controllable linear systems can be shown to be flat. Indeed, any system that can be transformed into a linear system by changes of coordinates, static feedback transformations (change of coordinates plus nonlinear change of inputs), or dynamic feedback transformations is also flat. Nonlinear control systems in "pure feedback form", which have gained popularity due to the applicability of backstepping [29] to such systems, are also flat. Thus, many of the systems for which strong nonlinear control techniques are available are in fact flat systems, leading one to question how the structure of flatness plays a role in control of such systems.

One common misconception is that flatness amounts to dynamic feedback linearization. It is true that any flat system can be feedback linearized using dynamic feedback (up to some regularity conditions that are generically satisfied). However, flatness is a property of a system and does not imply that one intends to then transform the system, via a dynamic feedback and appropriate changes of coordinates, to a single linear system. Indeed, the power of flatness is precisely that it does not convert nonlinear systems into linear ones. When a system is flat it is an indication that the nonlinear structure of the system is well characterized and one can exploit that structure in designing control algorithms for motion planning, trajectory generation, and stabilization. Dynamic feedback linearization is one such technique, although it is often a poor choice if the dynamics of the system are substantially different in different operating regimes.

Another advantage of studying flatness over dynamic feedback linearization is that flatness is a *geometric* property of a system, independent of

coordinate choice. Typically when one speaks of linear systems in a state space context, this does not make sense geometrically since the system is linear only in certain choices of coordinate representations. In particular, it is difficult to discuss the notion of a linear state space system on a manifold since the very definition of linearity requires an underlying linear space. In this way, flatness can be considered the proper geometric notion of linearity, even though the system may be quite nonlinear in almost any natural representation.

Finally, the notion of flatness can be extended to distributed parameters systems with boundary control, see section 3.2.2, and is useful even for controlling linear systems, whereas feedback linearization is yet to be defined in that context.

This paper provides a self-contained description of flat systems. Section 1 introduces the fundamental concepts of equivalence and flatness in a simple geometric framework. This is essentially an open-loop point of view. In section 2 we adopt a closed-loop point of view and relate equivalence and flatness to feedback design. Section 3 is devoted to open problems and new perspectives including developments on symmetries and distributed parameters systems. Finally, section 4 contains a representative catalog of flat systems arising in various fields of engineering.

1 Equivalence and flatness

1.1 Control systems as infinite dimensional vector fields

A system of differential equations

$$\dot{x} = f(x), \quad x \in X \subset \mathbb{R}^n \tag{1}$$

is by definition a pair (X, f), where X is an open set of $\mathbb{R}^n$ and f is a smooth vector field on X. A solution, or *trajectory*, of (1) is a mapping $t \mapsto x(t)$ such that

$$\dot{x}(t) = f(x(t)) \qquad \forall t \geq 0.$$

Notice that if $x \mapsto h(x)$ is a smooth function on X and $t \mapsto x(t)$ is a trajectory of (1), then

$$\frac{d}{dt}h(x(t)) = \frac{\partial h}{\partial x}(x(t)) \cdot \dot{x}(t) = \frac{\partial h}{\partial x}(x(t)) \cdot f(x(t)) \qquad \forall t \geq 0.$$

For that reason the *total derivative*, i.e., the mapping

$$x \mapsto \frac{\partial h}{\partial x}(x) \cdot f(x)$$

is somewhat abusively called the "time-derivative" of h and denoted by $\dot{h}$.

We would like to have a similar description, i.e., a "space" and a vector field on this space, for a control system

$$\dot{x} = f(x, u), \tag{2}$$

where f is smooth on an open subset $X \times U \subset \mathbb{R}^n \times \mathbb{R}^m$. Here f is no longer a vector field on X, but rather an *infinite collection* of vector fields on X parameterized by u: for all $u \in U$, the mapping

$$x \mapsto f_u(x) = f(x, u)$$

is a vector field on X. Such a description is not well-adapted when considering dynamic feedback.

It is nevertheless possible to associate to (2) a vector field with the "same" solutions using the following remarks: given a smooth solution of (2), i.e., a mapping $t \mapsto (x(t), u(t))$ with values in $X \times U$ such that

$$\dot{x}(t) = f(x(t), u(t)) \qquad \forall t \geq 0,$$

we can consider the *infinite* mapping

$$t \mapsto \xi(t) = (x(t), u(t), \dot{u}(t), \dots)$$

taking values in $X \times U \times \mathbb{R}^\infty_m$, where $\mathbb{R}^\infty_m = \mathbb{R}^m \times \mathbb{R}^m \times \dots$ denotes the product of an infinite (countable) number of copies of $\mathbb{R}^m$. A typical point of $\mathbb{R}^\infty_m$ is thus of the form $(u^1, u^2, \dots)$ with $u^i \in \mathbb{R}^m$. This mapping satisfies

$$\dot{\xi}(t) = \big(f(x(t), u(t)), \dot{u}(t), \ddot{u}(t), \dots\big) \qquad \forall t \geq 0,$$

hence it can be thought of as a trajectory of the *infinite* vector field

$$(x, u, u^1, \dots) \mapsto F(x, u, u^1, \dots) = (f(x, u), u^1, u^2, \dots)$$

on $X \times U \times \mathbb{R}^\infty_m$. Conversely, any mapping

$$t \mapsto \xi(t) = (x(t), u(t), u^1(t), \dots)$$

that is a trajectory of this infinite vector field necessarily takes the form $(x(t), u(t), \dot{u}(t), \dots)$ with $\dot{x}(t) = f(x(t), u(t))$, hence corresponds to a solution of (2). Thus F is truly a vector field and no longer a parameterized family of vector fields.

Using this construction, the control system (2) can be seen as the data of the "space" $X \times U \times \mathbb{R}^\infty_m$ together with the "smooth" vector field F on this space. Notice that, as in the uncontrolled case, we can define the "time-derivative" of a smooth function $(x, u, u^1, \dots) \mapsto h(x, u, u^1, \dots, u^k)$ depending on a *finite* number of variables by

$$\begin{aligned}\dot{h}(x, u, u^1, \dots, u^{k+1}) &:= Dh \cdot F \\ &= \frac{\partial h}{\partial x} \cdot f(x, u) + \frac{\partial h}{\partial u} \cdot u^1 + \frac{\partial h}{\partial u^1} \cdot u^2 + \cdots .\end{aligned}$$

The above sum is *finite* because h depends on finitely many variables.

Remark. To be rigorous we must say something of the underlying topology and differentiable structure of $\mathbb{R}^\infty_m$ to be able to speak of smooth objects [76]. This topology is the *Fréchet topology*, which makes things look as if we were working on the product of k copies of $\mathbb{R}^m$ for a "large enough" k. For our purpose it is enough to know that a basis of the open sets of this topology consists of infinite products $U_0 \times U_1 \times \dots$ of open sets of $\mathbb{R}^m$, and that a function is *smooth* if it depends on a *finite* but arbitrary number of variables and is smooth in the usual sense. In the same way a mapping $\Phi : \mathbb{R}^\infty_m \to \mathbb{R}^\infty_n$ is smooth if all of its components are smooth functions.

$\mathbb{R}^\infty_m$ equipped with the Fréchet topology has very weak properties: useful theorems such as the implicit function theorem, the Frobenius theorem, and the straightening out theorem no longer hold true. This is only because $\mathbb{R}^\infty_m$ is a very big space: indeed the Fréchet topology on the product of k copies of $\mathbb{R}^m$ for any finite k coincides with the usual Euclidian topology.

We can also define manifolds modeled on $\mathbb{R}^\infty_m$ using the standard machinery. The reader not interested in these technicalities can safely ignore the details and won't loose much by replacing "manifold modeled on $\mathbb{R}^\infty_m$" by "open set of $\mathbb{R}^\infty_m$".

We are now in position to give a formal definition of a system:

Definition 1. A *system* is a pair $(\mathfrak{M}, F)$ where $\mathfrak{M}$ is a smooth manifold, possibly of infinite dimension, and F is a smooth vector field on $\mathfrak{M}$.

Locally, a control system looks like an open subset of $\mathbb{R}^\alpha$ (α not necessarily finite) with coordinates $(\xi_1, \dots, \xi_\alpha)$ together with the vector field

$$\xi \mapsto F(\xi) = (F_1(\xi), \dots, F_\alpha(\xi))$$

where all the components F_i depend only on a finite number of coordinates. A *trajectory* of the system is a mapping $t \mapsto \xi(t)$ such that $\dot{\xi}(t) = F(\xi(t))$.

We saw in the beginning of this section how a "traditional" control system fits into our definition. There is nevertheless an important difference: we lose the notion of *state dimension*. Indeed

$$\dot{x} = f(x, u), \quad (x, u) \in X \times U \subset \mathbb{R}^n \times \mathbb{R}^m \tag{3}$$

and

$$\dot{x} = f(x, u), \qquad \dot{u} = v \tag{4}$$

now have the same description $(X \times U \times \mathbb{R}^\infty_m, F)$, with

$$F(x, u, u^1, \dots) = (f(x, u), u^1, u^2, \dots),$$

in our formalism: $t \mapsto (x(t), u(t))$ is a trajectory of (3) if and only if $t \mapsto (x(t), u(t), \dot{u}(t))$ is a trajectory of (4). This situation is not surprising since the state dimension is of course not preserved by dynamic feedback. On the other hand we will see there is still a notion of *input dimension*.

Example 1 (The trivial system). The *trivial system* $(\mathbb{R}^\infty_m, F_m)$, with coordinates $(y, y^1, y^2, \dots)$ and vector field

$$F_m(y, y^1, y^2, \dots) = (y^1, y^2, y^3, \dots)$$

describes any "traditional" system made of m chains of integrators of arbitrary lengths, and in particular the direct transfer $y = u$.

In practice we often identify the "system" $F(x, \overline{u}) := (f(x, u), u^1, u^2, \dots)$ with the "dynamics" $\dot{x} = f(x, u)$ which defines it. Our main motivation for introducing a new formalism is that it will turn out to be a natural framework for the notions of equivalence and flatness we want to define.

Remark. It is easy to see that the manifold $\mathfrak{M}$ is finite-dimensional only when there is no input, i.e., to describe a *determined* system of differential equations one needs as many equations as variables. In the presence of inputs, the system becomes *underdetermined*, there are more variables than equations, which accounts for the infinite dimension.

Remark. Our definition of a system is adapted from the notion of *diffiety* introduced in [76] to deal with systems of (partial) differential equations. By definition a diffiety is a pair $(\mathfrak{M}, C\mathfrak{TM})$ where $\mathfrak{M}$ is smooth manifold, possibly of infinite dimension, and $C\mathfrak{TM}$ is an involutive finite-dimensional distribution on $\mathfrak{M}$, i.e., the Lie bracket of any two vector fields of $C\mathfrak{TM}$ is itself in $C\mathfrak{TM}$. The dimension of $C\mathfrak{TM}$ is equal to the number of independent variables.

As we are only working with systems with lumped parameters, hence governed by ordinary differential equations, we consider diffieties with one dimensional distributions. For our purpose we have also chosen to single out a particular vector field rather than work with the distribution it spans.

1.2 Equivalence of systems

In this section we define an equivalence relation formalizing the idea that two systems are "equivalent" if there is an invertible transformation exchanging their trajectories. As we will see later, the relevance of this rather natural equivalence notion lies in the fact that it admits an interpretation in terms of dynamic feedback.

Consider two systems $(\mathfrak{M}, F)$ and $(\mathfrak{N}, G)$ and a smooth mapping $\Psi : \mathfrak{M} \to \mathfrak{N}$ (remember that by definition every component of a smooth mapping depends only on finitely many coordinates). If $t \mapsto \xi(t)$ is a trajectory of $(\mathfrak{M}, F)$, i.e.,

$$\forall \xi, \quad \dot{\xi}(t) = F(\xi(t)),$$

the composed mapping $t \mapsto \zeta(t) = \Psi(\xi(t))$ satisfies the chain rule

$$\dot{\zeta}(t) = \frac{\partial \Psi}{\partial \xi}(\xi(t)) \cdot \dot{\xi}(t) = \frac{\partial \Psi}{\partial \xi}(\xi(t)) \cdot F(\xi(t)).$$

The above expressions involve only finite sums even if the matrices and vectors have infinite sizes: indeed a row of $\frac{\partial \Psi}{\partial \xi}$ contains only a finite number of non zero terms because a component of Ψ depends only on finitely many coordinates. Now, if the vector fields F and G are Ψ-*related*, i.e.,

$$\forall \xi, \quad G(\Psi(\xi)) = \frac{\partial \Psi}{\partial \xi}(\xi) \cdot F(\xi)$$

then

$$\dot{\zeta}(t) = G(\Psi(\xi(t)) = G(\zeta(t)),$$

which means that $t \mapsto \zeta(t) = \Psi(\xi(t))$ is a trajectory of $(\mathfrak{N}, G)$. If moreover Ψ has a smooth inverse Φ then obviously F, G are also Φ-related, and there is a one-to-one correspondence between the trajectories of the two systems. We call such an invertible Ψ relating F and G an *endogenous transformation.*

Definition 2. Two systems $(\mathfrak{M}, F)$ and $(\mathfrak{N}, G)$ are *equivalent at* $(p, q) \in \mathfrak{M} \times \mathfrak{N}$ if there exists an endogenous transformation from a neighborhood of p to a neighborhood of q. $(\mathfrak{M}, F)$ and $(\mathfrak{N}, G)$ are *equivalent* if they are equivalent at every pair of points (p, q) of a dense open subset of $\mathfrak{M} \times \mathfrak{N}$.

Notice that when $\mathfrak{M}$ and $\mathfrak{N}$ have the same *finite* dimension, the systems are necessarily equivalent by the straightening out theorem. This is no longer true in infinite dimensions.

Consider the two systems $(X \times U \times \mathbb{R}^\infty_m, F)$ and $(Y \times V \times \mathbb{R}^\infty_s, G)$ describing the dynamics

$$\dot{x} = f(x, u), \quad (x, u) \in X \times U \subset \mathbb{R}^n \times \mathbb{R}^m \tag{5}$$

$$\dot{y} = g(y, v), \quad (y, v) \in Y \times V \subset \mathbb{R}^r \times \mathbb{R}^s. \tag{6}$$

The vector fields F, G are defined by

$$F(x, u, u^1, \dots) = (f(x, u), u^1, u^2, \dots)$$
$$G(y, v, v^1, \dots) = (g(y, v), v^1, v^2, \dots).$$

If the systems are equivalent, the endogenous transformation Ψ takes the form

$$\Psi(x, u, u^1, \dots) = (\psi(x, \overline{u}), \beta(x, \overline{u}), \dot{\beta}(x, \overline{u}), \dots).$$

Here we have used the short-hand notation $\overline{u} = (u, u^1, \dots, u^k)$, where k is some finite but otherwise arbitrary integer. Hence Ψ is completely specified by the mappings ψ and β, i.e, by the expression of y, v in terms of $x, \overline{u}$. Similarly, the inverse Φ of Ψ takes the form

$$\Phi(y, v, v^1, \dots) = (\varphi(y, \overline{v}), \alpha(y, \overline{v}), \dot{\alpha}(y, \overline{v}), \dots).$$

As Ψ and Φ are inverse mappings we have

$$\begin{aligned} \psi\big(\varphi(y, \overline{v}), \overline{\alpha}(y, \overline{v})\big) &= y \\ \beta\big(\varphi(y, \overline{v}), \overline{\alpha}(y, \overline{v})\big) &= v \end{aligned} \quad \text{and} \quad \begin{aligned} \varphi\big(\psi(x, \overline{u}), \overline{\beta}(x, \overline{u})\big) &= x \\ \alpha\big(\psi(x, \overline{u}), \overline{\beta}(x, \overline{u})\big) &= u. \end{aligned}$$

Moreover F and G Ψ-related implies

$$f\big(\varphi(y,\overline{v}),\alpha(y,\overline{v})\big) = D\varphi(y,\overline{v})\cdot\overline{g}(y,\overline{v})$$

where $\overline{g}$ stands for $(g,v^1,\ldots,v^k)$, i.e., a truncation of G for some large enough k. Conversely,

$$g\big(\psi(x,\overline{u}),\beta(y,\overline{u})\big) = D\psi(x,\overline{u})\cdot\overline{f}(y,\overline{u}).$$

In other words, whenever $t\mapsto(x(t),u(t))$ is a trajectory of (5)

$$t\mapsto(y(t),v(t)) = \big(\varphi(x(t),\overline{u}(t)),\ \alpha(x(t),\overline{u}(t))\big)$$

is a trajectory of (6), and vice versa.

Example 2 (The PVTOL, see example 21). The system generated by

$$\begin{aligned}\ddot{x} &= -u_1\sin\theta+\varepsilon u_2\cos\theta\\ \ddot{z} &= u_1\cos\theta+\varepsilon u_2\sin\theta-1\\ \ddot{\theta} &= u_2.\end{aligned}$$

is globally equivalent to the systems generated by

$$\ddot{y}_1 = -\xi\sin\theta,\qquad \ddot{y}_2 = \xi\cos\theta-1,$$

where ξ and θ are the control inputs. Indeed, setting

$$\begin{aligned}X &:= (x,z,\dot{x},\dot{z},\theta,\dot{\theta})\\ U &:= (u_1,u_2)\end{aligned}\qquad\text{and}\qquad\begin{aligned}Y &:= (y_1,y_2,\dot{y}_1,\dot{y}_2)\\ V &:= (\xi,\theta)\end{aligned}$$

and using the notations in the discussion after definition 2, we define the mappings $Y=\psi(X,\overline{U})$ and $V=\beta(X,\overline{U})$ by

$$\psi(X,\overline{U}) := \begin{pmatrix} x-\varepsilon\sin\theta\\ z+\varepsilon\cos\theta\\ \dot{x}-\varepsilon\dot{\theta}\cos\theta\\ \dot{z}-\varepsilon\dot{\theta}\sin\theta\end{pmatrix}\qquad\text{and}\qquad \beta(X,\overline{U}) := \begin{pmatrix} u_1-\varepsilon\dot{\theta}^2\\ \theta\end{pmatrix}$$

to generate the mapping Ψ. The inverse mapping Φ is generated by the mappings $X=\varphi(Y,\overline{V})$ and $U=\alpha(Y,\overline{V})$ defined by

$$\varphi(Y,\overline{V}) := \begin{pmatrix} y_1+\varepsilon\sin\theta\\ y_2-\varepsilon\cos\theta\\ \dot{y}_1+\varepsilon\dot{\theta}\cos\theta\\ \dot{y}_2-\varepsilon\dot{\theta}\sin\theta\\ \theta\\ \dot{\theta}\end{pmatrix}\qquad\text{and}\qquad \alpha(Y,\overline{V}) := \begin{pmatrix} \xi+\varepsilon\dot{\theta}^2\\ \ddot{\theta}\end{pmatrix}$$

An important property of endogenous transformations is that they preserve the input dimension:

Theorem 1. *If two systems* $(X \times U \times \mathbb{R}^\infty_m, F)$ *and* $(Y \times V \times \mathbb{R}^\infty_s, G)$ *are equivalent, then they have the same number of inputs, i.e.,* $m = s$.

Proof. Consider the truncation Φ_μ of Φ on $X \times U \times (\mathbb{R}^m)^\mu$,

$$\begin{array}{rcl} \Phi_\mu : X \times U \times (\mathbb{R}^{m+k})^\mu & \to & Y \times V \times (\mathbb{R}^s)^\mu \\ (x, u, u^1, \dots, u^{k+\mu}) & \mapsto & (\varphi, \alpha, \dot{\alpha}, \dots, \alpha^{(\mu)}), \end{array}$$

i.e., the first $\mu+2$ blocks of components of Ψ; k is just a fixed "large enough" integer. Because Ψ is invertible, Ψ_μ is a submersion for all μ. Hence the dimension of the domain is greater than or equal to the dimension of the range,

$$n + m(k + \mu + 1) \geq s(\mu + 1) \qquad \forall \mu > 0,$$

which implies $m \geq s$. Using the same idea with Ψ leads to $s \geq m$. □

Remark. Our definition of equivalence is adapted from the notion of equivalence between diffieties. Given two diffieties $(\mathfrak{M}, CT\mathfrak{M})$ and $(\mathfrak{N}, CT\mathfrak{N})$, we say that a smooth mapping Ψ from (an open subset of) $\mathfrak{M}$ to $\mathfrak{N}$ is *Lie-Bäcklund* if its tangent mapping $T\Psi$ satisfies $T\Phi(CT\mathfrak{M}) \subset CT\mathfrak{N}$. If moreover Ψ has a smooth inverse Φ such that $T\Psi(CT\mathfrak{N}) \subset CT\mathfrak{M}$, we say it is a *Lie-Bäcklund isomorphism.* When such an isomorphism exists, the diffieties are said to be *equivalent.* An endogenous transformation is just a special Lie-Bäcklund isomorphism, which preserves the time parameterization of the integral curves. It is possible to define the more general concept of *orbital* equivalence [14, 12] by considering general Lie-Bäcklund isomorphisms, which preserve only the geometric locus of the integral curves (see an example in section 26).

1.3 Differential Flatness

We single out a very important class of systems, namely systems equivalent to a trivial system $(\mathbb{R}^\infty_s, F_s)$ (see example 1):

Definition 3. The system $(\mathfrak{M}, F)$ is *flat at* $p \in \mathfrak{M}$ (resp. *flat*) if it equivalent at p (resp. equivalent) to a trivial system.

We specialize the discussion after definition 2 to a flat system $(X \times U \times \mathbb{R}^\infty_m, F)$ describing the dynamics

$$\dot{x} = f(x,u), \quad (x,u) \in X \times U \subset \mathbb{R}^n \times \mathbb{R}^m.$$

By definition the system is equivalent to the trivial system $(\mathbb{R}^\infty_s, F_s)$ where the endogenous transformation Ψ takes the form

$$\Psi(x,u,u^1,\dots) = (h(x,\overline{u}), \dot{h}(x,\overline{u}), \ddot{h}(x,\overline{u}),\dots). \tag{7}$$

In other words Ψ is the infinite prolongation of the mapping h. The inverse Φ of Ψ takes the form

$$\Psi(\overline{y}) = (\psi(\overline{y}), \beta(\overline{y}), \dot{\beta}(\overline{y}),\dots).$$

As Φ and Ψ are inverse mappings we have in particular

$$\varphi\big(\overline{h}(x,\overline{u})\big) = x \quad \text{and} \quad \alpha\big(\overline{h}(x,\overline{u})\big) = u.$$

Moreover F and G Φ-related implies that whenever $t \mapsto y(t)$ is a trajectory of $y = v$ –i.e., nothing but an *arbitrary* mapping–

$$t \mapsto \big(x(t), u(t)\big) = \big(\psi(\overline{y}(t)),\ \beta(\overline{y}(t))\big)$$

is a trajectory of $\dot{x} = f(x,u)$, and vice versa.

We single out the importance of the mapping h of the previous example:

Definition 4. Let $(\mathfrak{M}, F)$ be a flat system and Ψ the endogenous transformation putting it into a trivial system. The first block of components of Ψ, i.e., the mapping h in (7), is called a *flat (*or *linearizing) output.*

With this definition, an obvious consequence of theorem 1 is:

Corollary 1. *Consider a flat system. The dimension of a flat output is equal to the input dimension, i.e., $s = m$.*

Example 3 (The PVTOL). The system studied in example 2 is flat, with

$$y = h(X, \overline{U}) := (x - \varepsilon \sin\theta, z + \varepsilon\cos\theta)$$

as a flat output. Indeed, the mappings $X = \varphi(\overline{y})$ and $U = \alpha(\overline{y})$ which generate the inverse mapping Φ can be obtained from the implicit equations

$$\begin{aligned} (y_1 - x)^2 + (y_2 - z)^2 &= \varepsilon^2 \\ (y_1 - x)(\ddot{y}_2 + 1) - (y_2 - z)\ddot{y}_1 &= 0 \\ (\ddot{y}_2 + 1)\sin\theta + \ddot{y}_1\cos\theta &= 0. \end{aligned}$$

We first solve for x, z, θ,

$$\begin{aligned} x &= y_1 + \varepsilon \frac{\ddot{y}_1}{\sqrt{\ddot{y}_1^2 + (\ddot{y}_2 + 1)^2}} \\ z &= y_2 + \varepsilon \frac{(\ddot{y}_2 + 1)}{\sqrt{\ddot{y}_1^2 + (\ddot{y}_2 + 1)^2}} \\ \theta &= \arg(\ddot{y}_1, \ddot{y}_2 + 1), \end{aligned}$$

and then differentiate to get $\dot{x}, \dot{z}, \dot{\theta}, u$ in function of the derivatives of y. Notice the only singularity is $\ddot{y}_1^2 + (\ddot{y}_2 + 1)^2 = 0$.

1.4 Application to motion planning

We now illustrate how flatness can be used for solving control problems. Consider a nonlinear control system of the form

$$\dot{x} = f(x, u) \qquad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m$$

with flat output

$$y = h(x, u, \dot{u}, \ldots, u^{(r)}).$$

By virtue of the system being flat, we can write all trajectories $(x(t), u(t))$ satisfying the differential equation in terms of the flat output and its derivatives:

$$\begin{aligned} x &= \varphi(y, \dot{y}, \ldots, y^{(q)}) \\ u &= \alpha(y, \dot{y}, \ldots, y^{(q)}). \end{aligned}$$

We begin by considering the problem of steering from an initial state to a final state. We parameterize the components of the flat output y_i, $i = 1, \ldots, m$ by

$$y_i(t) := \sum_j A_{ij} \lambda_j(t), \tag{8}$$

where the $\lambda_j(t)$, $j = 1, \ldots, N$ are basis functions. This reduces the problem from finding a function in an infinite dimensional space to finding a finite set of parameters.

Suppose we have available to us an initial state x_0 at time τ_0 and a final state x_f at time τ_f. Steering from an initial point in state space to a desired

point in state space is trivial for flat systems. We have to calculate the values of the flat output and its derivatives from the desired points in state space and then solve for the coefficients A_{ij} in the following system of equations:

$$\begin{array}{rclcrcl} y_i(\tau_0) & = & \sum_j A_{ij}\lambda_j(\tau_0) & & y_i(\tau_f) & = & \sum_j A_{ij}\lambda_j(\tau_f) \\ & \vdots & & & & \vdots & \\ y_i^{(q)}(\tau_0) & = & \sum_j A_{ij}\lambda_j^{(q)}(\tau_0) & & y_i^{(q)}(\tau_f) & = & \sum_j A_{ij}\lambda_j^{(q)}(\tau_f). \end{array} \tag{9}$$

To streamline notation we write the following expressions for the case of a *one*-dimensional flat output only. The multi-dimensional case follows by repeatedly applying the one-dimensional case, since the algorithm is decoupled in the component of the flat output. Let $\Lambda(t)$ be the $q+1$ by N matrix $\Lambda_{ij}(t) = \lambda_j^{(i)}(t)$ and let

$$\begin{aligned} \bar{y}_0 &= (y_1(\tau_0), \ldots, y_1^{(q)}(\tau_0)) \\ \bar{y}_f &= (y_1(\tau_f), \ldots, y_1^{(q)}(\tau_f)) \\ \bar{y} &= (\bar{y}_0, \bar{y}_f). \end{aligned} \tag{10}$$

Then the constraint in equation (9) can be written as

$$\bar{y} = \begin{pmatrix} \Lambda(\tau_0) \\ \Lambda(\tau_f) \end{pmatrix} A =: \Lambda A. \tag{11}$$

That is, we require the coefficients A to be in an affine sub-space defined by equation (11). The only condition on the basis functions is that Λ is full rank, in order for equation (11) to have a solution.

The implications of flatness is that the trajectory generation problem can be reduced to simple algebra, in theory, and computationally attractive algorithms in practice. In the case of the towed cable system of example 25, a reasonable state space representation of the system consists of approximately 128 states. Traditional approaches to trajectory generation, such as optimal control, cannot be easily applied in this case. However, it follows from the fact that the system is flat that the feasible trajectories of the system are completely characterized by the motion of the point at the bottom of the cable. By converting the input constraints on the system to constraints on the curvature and higher derivatives of the motion of the bottom of the cable, it is possible to compute efficient techniques for trajectory generation.

1.5 Motion planning with singularities

In the previous section we assumed the endogenous transformation

$$\Psi(x,u,u_1,\dots) := \big(h(x,\overline{u}),\dot{h}(x,\overline{u}),\ddot{h}(x,\overline{u}),\dots\big)$$

generated by the flat output $y = h(x,\overline{u})$ everywhere nonsingular, so that we could invert it and express x and u in function of y and its derivatives,

$$(y,\dot{y},\dots,y^{(q)}) \mapsto (x,u) = \phi(y,\dot{y},\dots,y^{(q)}).$$

But it may well be that a singularity is in fact an interesting point of operation. As ϕ is not defined at such a point, the previous computations do not apply. A way to overcome the problem is to "blow up" the singularity by considering trajectories $t \mapsto y(t)$ such that

$$t \mapsto \phi\big(y(t),\dot{y}(t),\dots,y^{(q)}(t)\big)$$

can be prolonged into a smooth mapping at points where ϕ is not defined. To do so requires a detailed study of the singularity. A general statement is beyond the scope of this paper and we simply illustrate the idea with an example.

Example 4. Consider the flat dynamics

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = u_2u_1, \quad \dot{x}_3 = x_2u_1,$$

with flat output $y := (x_1,x_3)$. When $u_1 = 0$, i.e., $\dot{y}_1 = 0$ the endogenous transformation generated by the flat output is singular and the inverse mapping

$$(y,\dot{y},\ddot{y}) \stackrel{\phi}{\longmapsto} (x_1,x_2,x_3,u_1,u_2) = \left(y_1,\frac{\dot{y}_2}{\dot{y}_1},y_2,\dot{y}_1,\frac{\ddot{y}_2\dot{y}_1-\ddot{y}_1\dot{y}_2}{\dot{y}_1^3}\right),$$

is undefined. But if we consider trajectories $t \mapsto y(t) := \big(\sigma(t),p(\sigma(t))\big)$, with σ and p smooth functions, we find that

$$\frac{\dot{y}_2(t)}{\dot{y}_1(t)} = \frac{\dfrac{dp}{d\sigma}\big(\sigma(t)\big)\cdot\dot{\sigma}(t)}{\dot{\sigma}(t)} \qquad \text{and} \qquad \frac{\ddot{y}_2\dot{y}_1-\ddot{y}_1\dot{y}_2}{\dot{y}_1^3} = \frac{\dfrac{d^2p}{d\sigma^2}\big(\sigma(t)\big)\cdot\dot{\sigma}^3(t)}{\dot{\sigma}^3(t)},$$

hence we can prolong $t \mapsto \phi(y(t),\dot{y}(t),\ddot{y}(t))$ everywhere by

$$t \mapsto \left(\sigma(t),\ \frac{dp}{d\sigma}\big(\sigma(t)\big),\ p(\sigma(t)),\ \dot{\sigma}(t),\ \frac{d^2p}{d\sigma^2}\big(\sigma(t)\big)\right).$$

The motion planning can now be done as in the previous section: indeed, the functions σ and p and their derivatives are constrained at the initial (resp. final) time by the initial (resp. final) point but otherwise arbitrary.

For a more substantial application see [66, 67, 16], where the same idea was applied to nonholonomic mechanical systems by taking advantage of the "natural" geometry of the problem.

2 Feedback design with equivalence

2.1 From equivalence to feedback

The equivalence relation we have defined is very natural since it is essentially a $1-1$ correspondence between trajectories of systems. We had mainly an open-loop point of view. We now turn to a closed-loop point of view by interpreting equivalence in terms of feedback. For that, consider the two dynamics

$$\begin{aligned} \dot{x} &= f(x,u), \quad (x,u) \in X \times U \subset \mathbb{R}^n \times \mathbb{R}^m \\ \dot{y} &= g(y,v), \quad (y,v) \in Y \times V \subset \mathbb{R}^r \times \mathbb{R}^s. \end{aligned}$$

They are described in our formalism by the systems $(X \times U \times \mathbb{R}_m^\infty, F)$ and $(Y \times V \times \mathbb{R}_s^\infty, G)$, with F and G defined by

$$\begin{aligned} F(x,u,u^1,\dots) &:= (f(x,u),u^1,u^2,\dots) \\ G(y,v,v^1,\dots) &:= (g(y,v),v^1,v^2,\dots). \end{aligned}$$

Assume now the two systems are equivalent, i.e., they have the same trajectories. Does it imply that it is possible to go from $\dot{x} = f(x,u)$ to $\dot{y} = g(y,v)$ by a (possibly) dynamic feedback

$$\begin{aligned} \dot{z} &= a(x,z,v), \quad z \in Z \subset \mathbb{R}^q \\ u &= \kappa(x,z,v), \end{aligned}$$

and *vice versa*? The question might look stupid at first glance since such a feedback can only increase the state dimension. Yet, we can give it some sense if we agree to work "up to pure integrators" (remember this does not change the system in our formalism, see the remark after definition 1).

Theorem 2. *Assume* $\dot{x} = f(x,u)$ *and* $\dot{y} = g(y,v)$ *are equivalent. Then* $\dot{x} = f(x,u)$ *can be transformed by (dynamic) feedback and coordinate change into*

$$\dot{y} = g(y,v), \quad \dot{v} = v^1, \quad \dot{v}^1 = v^2, \quad \dots \quad , \quad \dot{v}^\mu = w$$

for some large enough integer μ*. Conversely,* $\dot{y} = g(y,v)$ *can be transformed by (dynamic) feedback and coordinate change into*

$$\dot{x} = f(x,u), \quad \dot{u} = u^1, \quad \dot{u}^1 = u^2, \quad \dots \quad , \quad \dot{u}^\nu = w$$

for some large enough integer ν*.*

Proof [33]. Denote by F and G the infinite vector fields representing the two dynamics. Equivalence means there is an invertible mapping

$$\Phi(y,\overline{v}) = (\varphi(y,\overline{v}), \alpha(y,\overline{v}), \dot{\alpha}(y,\overline{v}), \dots)$$

such that

$$F(\Phi(y,\overline{v})) = D\Phi(y,\overline{v}).G(y,\overline{v}). \tag{12}$$

Let $\tilde{y} := (y, v, v^1, \dots, v^\mu)$ and $w := v^{\mu+1}$. For μ large enough, φ (resp. α) depends only on $\tilde{y}$ (resp. on $\tilde{y}$ and w). With these notations, Φ reads

$$\Phi(\tilde{y},\overline{w}) = (\varphi(\tilde{y}), \alpha(\tilde{y},w), \dot{\alpha}(y,\overline{w}), \dots),$$

and equation (12) implies in particular

$$f(\varphi(\tilde{y}), \alpha(\tilde{y},w)) = D\varphi(\tilde{y}).\tilde{g}(\tilde{y},w), \tag{13}$$

where $\tilde{g} := (g, v^1, \dots, v^k)$. Because Φ is invertible, φ is full rank hence can be completed by some map π to a coordinate change

$$\tilde{y} \mapsto \phi(\tilde{y}) = (\varphi(\tilde{y}), \pi(\tilde{y})).$$

Consider now the dynamic feedback

$$\begin{aligned} u &= \alpha(\phi^{-1}(x,z), w)) \\ \dot{z} &= D\pi(\phi^{-1}(x,z)).\tilde{g}(\phi^{-1}(x,z), w)), \end{aligned}$$

which transforms $\dot{x} = f(x,u)$ into

$$\begin{pmatrix} \dot{x} \\ \dot{z} \end{pmatrix} = \tilde{f}(x,z,w) := \begin{pmatrix} f(x, \alpha(\phi^{-1}(x,z), w)) \\ D\pi(\phi^{-1}(x,z)).\tilde{g}(\phi^{-1}(x,z), w)) \end{pmatrix}.$$

Using (13), we have

$$\tilde{f}\big(\phi(\tilde{y}), w\big) = \begin{pmatrix} f\big(\varphi(\tilde{y}), \alpha(\tilde{y}, w)\big) \\ D\pi(\tilde{y}).\tilde{g}(\tilde{y}, w) \end{pmatrix} = \begin{pmatrix} D\varphi(\tilde{y}) \\ D\pi(\tilde{y}) \end{pmatrix} \cdot \tilde{g}(\tilde{y}, w) = D\phi(\tilde{y}).\tilde{g}(\tilde{y}, w).$$

Therefore $\tilde{f}$ and $\tilde{g}$ are ϕ-related, which ends the proof. Exchanging the roles of f and g proves the converse statement. □

As a flat system is equivalent to a trivial one, we get as an immediate consequence of the theorem:

Corollary 2. *A flat dynamics can be linearized by (dynamic) feedback and coordinate change.*

Remark. As can be seen in the proof of the theorem there are many feedbacks realizing the equivalence, as many as suitable mappings π. Notice all these feedback explode at points where φ is singular (i.e., where its rank collapses).

Further details about the construction of a linearizing feedback from an output and the links with extension algorithms can be found in [35].

Example 5 (The PVTOL). We know from example 3 that the dynamics

$$\begin{aligned} \ddot{x} &= -u_1 \sin\theta + \varepsilon u_2 \cos\theta \\ \ddot{z} &= u_1 \cos\theta + \varepsilon u_2 \sin\theta - 1 \\ \ddot{\theta} &= u_2 \end{aligned}$$

admits the flat output

$$y = (x - \varepsilon \sin\theta, z + \varepsilon \cos\theta).$$

It is transformed into the linear dynamics

$$y_1^{(4)} = v_1, \qquad y_2^{(4)} = v_2$$

by the feedback

$$\begin{aligned} \ddot{\xi} &= -v_1 \sin\theta + v_2 \cos\theta + \xi\dot{\theta}^2 \\ u_1 &= \xi + \varepsilon\dot{\theta}^2 \\ u_2 &= \frac{-1}{\xi}(v_1 \cos\theta + v_2 \sin\theta + 2\dot{\xi}\dot{\theta}) \end{aligned}$$

and the coordinate change

$$(x, z, \theta, \dot{x}, \dot{z}, \dot{\theta}, \xi, \dot{\xi}) \mapsto (y, \dot{y}, \ddot{y}, y^{(3)}).$$

The only singularity of this transformation is $\xi = 0$, i.e., $\ddot{y}_1^2 + (\ddot{y}_2 + 1)^2 = 0$. Notice the PVTOL is not linearizable by static feedback (see section 3.1.2).

2.2 Endogenous feedback

Theorem 2 asserts the existence of a feedback such that

$$\begin{aligned} \dot{x} &= f(x, \kappa(x, z, w)) \\ \dot{z} &= a(x, z, w). \end{aligned} \tag{14}$$

reads, up to a coordinate change,

$$\dot{y} = g(y, v), \quad \dot{v} = v^1, \quad \ldots \quad , \quad \dot{v}^\mu = w. \tag{15}$$

But (15) is trivially equivalent to $\dot{y} = g(y, v)$ (see the remark after definition 1), which is itself equivalent to $\dot{x} = f(x, u)$. Hence, (14) is equivalent to $\dot{x} = f(x, u)$. This leads to

Definition 5. Consider the dynamics $\dot{x} = f(x, u)$. We say the feedback

$$\begin{aligned} u &= \kappa(x, z, w) \\ \dot{z} &= a(x, z, w) \end{aligned}$$

is *endogenous* if the open-loop dynamics $\dot{x} = f(x, u)$ is equivalent to the closed-loop dynamics

$$\begin{aligned} \dot{x} &= f(x, \kappa(x, z, w)) \\ \dot{z} &= a(x, z, w). \end{aligned}$$

The word "endogenous" reflects the fact that the feedback variables z and w are in loose sense "generated" by the original variables $x, \overline{u}$ (see [33, 36] for further details and a characterization of such feedbacks)

Remark. It is also possible to consider at no extra cost "generalized" feedbacks depending not only on w but also on derivatives of w.

We thus have a more precise characterization of equivalence and flatness:

Theorem 3. *Two dynamics $\dot{x} = f(x, u)$ and $\dot{y} = g(y, v)$ are equivalent if and only if $\dot{x} = f(x, u)$ can be transformed by endogenous feedback and coordinate change into*

$$\dot{y} = g(y, v), \quad \dot{v} = v^1, \quad \ldots \quad , \quad \dot{v}^\mu = w. \tag{16}$$

for some large enough integer ν, and vice versa.

Corollary 3. *A dynamics is flat if and only if it is linearizable by endogenous feedback and coordinate change.*

Another trivial but important consequence of theorem 2 is that an endogenous feedback can be "unraveled" by another endogenous feedback:

Corollary 4. *Consider a dynamics*

$$\begin{aligned}\dot{x} &= f(x, \kappa(x, z, w))\\ \dot{z} &= a(x, z, w)\end{aligned}$$

where

$$\begin{aligned}u &= \kappa(x, z, w)\\ \dot{z} &= a(x, z, w)\end{aligned}$$

is an endogenous feedback. Then it can be transformed by endogenous feedback and coordinate change into

$$\dot{x} = f(x, u), \quad \dot{u} = u^1, \quad \dots \quad , \quad \dot{u}^\mu = w. \tag{17}$$

for some large enough integer μ.

This clearly shows which properties are preserved by equivalence: properties that are preserved by adding pure integrators and coordinate changes, in particular controllability.

An endogenous feedback is thus truly "reversible", up to pure integrators. It is worth pointing out that a feedback which is *invertible* in the sense of the standard –but maybe unfortunate– terminology [52] is not necessarily endogenous. For instance the invertible feedback $\dot{z} = v,\ u = v$ acting on the scalar dynamics $\dot{x} = u$ is not endogenous. Indeed, the closed-loop dynamics $\dot{x} = v,\ \dot{z} = v$ is no longer controllable, and there is no way to change that by another feedback!

2.3 Tracking: feedback linearization

One of the central problems of control theory is *trajectory tracking*: given a dynamics $\dot{x} = f(x, u)$, we want to design a controller able to track any reference trajectory $t \mapsto \big(x_r(t), u_r(t)\big)$. If this dynamics admits a flat output $y = h(x, \overline{u})$, we can use corollary 2 to transform it by (endogenous) feedback and coordinate change into the linear dynamics $y^{(\mu+1)} = w$. Assigning then

$$v := y_r^{(\mu+1)}(t) - K\Delta\tilde{y}$$

with a suitable gain matrix K, we get the stable closed-loop error dynamics

$$\Delta y^{(\mu+1)} = -K\Delta\tilde{y},$$

where $y_r(t) := \big(x_r(t), \overline{u}_r(t)\big)$ and $\tilde{y} := (y, \dot{y}, \dots, y^\mu)$ and $\Delta\xi$ stands for $\xi - \xi_{r(t)}$. This control law meets the design objective. Indeed, there is by the definition of flatness an invertible mapping

$$\Phi(\overline{y}) = (\varphi(\overline{y}), \alpha(\overline{y}), \dot{\alpha}(\overline{y}), \dots)$$

relating the infinite dimension vector fields $F(x, \overline{u}) := (f(x,u), u, u^1, \dots)$ and $G(\overline{y}) := (y, y^1, \dots)$. From the proof of theorem 2, this means in particular

$$\begin{aligned} x &= \varphi(\tilde{y}_r(t) + \Delta\tilde{y}) \\ &= \varphi(\tilde{y}_r(t)) + R_\varphi(y_r(t), \Delta\tilde{y}).\Delta\tilde{y} \\ &= x_r(t) + R_\varphi(y_r(t), \Delta\tilde{y}).\Delta\tilde{y} \end{aligned}$$

and

$$\begin{aligned} u &= \alpha(\tilde{y}_r(t) + \Delta\tilde{y}, -K\Delta\tilde{y}) \\ &= \alpha(\tilde{y}_r(t)) + R_\alpha(y_r^{(\mu+1)}(t), \Delta\tilde{y}).\begin{pmatrix} \Delta\tilde{y} \\ -K\Delta\tilde{y} \end{pmatrix} \\ &= u_r(t) + R_\alpha(\tilde{y}_r(t), y_r^{(\mu+1)}(t), \Delta\tilde{y}, \Delta w).\begin{pmatrix} \Delta\tilde{y} \\ -K\Delta\tilde{y} \end{pmatrix}, \end{aligned}$$

where we have used the fundamental theorem of calculus to define

$$\begin{aligned} R_\varphi(Y, \Delta Y) &:= \int_0^1 D\varphi(Y + t\Delta Y)dt \\ R_\alpha(Y, w, \Delta Y, \Delta w) &:= \int_0^1 D\alpha(Y + t\Delta Y, w + t\Delta w)dt. \end{aligned}$$

Since $\Delta y \to 0$ as $t \to \infty$, this means $x \to x_r(t)$ and $u \to u_r(t)$. Of course the tracking gets poorer and poorer as the ball of center $\tilde{y}_r(t)$ and radius Δy approaches a singularity of φ. At the same time the control effort gets larger and larger, since the feedback explodes at such a point (see the remark after theorem 2). Notice the tracking quality and control effort depend only on the mapping Φ, hence on the flat output, and not on the feedback itself.

We end this section with some comments on the use of feedback linearization. A linearizing feedback should always be fed by a *trajectory generator*,

even if the original problem is not stated in terms of tracking. For instance, if it is desired to *stabilize* an equilibrium point, applying directly feedback linearization without first planning a reference trajectory yields very large control effort when starting from a distant initial point. The role of the trajectory generator is to define an *open-loop* "reasonable" trajectory –i.e., satisfying some state and/or control constraints– that the linearizing feedback will then track.

2.4 Tracking: singularities and time scaling

Tracking by feedback linearization is possible only far from singularities of the endogenous transformation generated by the flat output. If the reference trajectory passes through or near a singularity, then feedback linearization cannot be directly applied, as is the case for motion planning, see section 1.5. Nevertheless, it can be used after a *time scaling*, at least in the presence of "simple" singularities. The interest is that it allows exponential tracking, though in a new "singular" time.

Example 6. Take a reference trajectory $t \mapsto y_r(t) = (\sigma(t), p(\sigma(t))$ for example 4. Consider the dynamic time-varying compensator $u_1 = \xi\dot{\sigma}(t)$ and $\dot{\xi} = v_1\dot{\sigma}(t)$. The closed loop system reads

$$x_1' = \xi, \quad x_2' = u_2\xi, \quad x_3' = x_2\xi \quad \xi' = v_1.$$

where $'$ stands for $d/d\sigma$, the extended state is (x_1, x_2, x_3, ξ), the new control is (v_1, v_2). An equivalent second order formulation is

$$x_1'' = v_1, \quad x_3'' = u_2\xi^2 + x_2v_1.$$

When ξ is far from zero, the static feedback $u_2 = (v_2 - x_2v_1)/\xi^2$ linearizes the dynamics,

$$x_1'' = v_1, \quad x_3'' = v_2$$

in σ scale. When the system remains close to the reference, $\xi \approx 1$, even if for some t, $\dot{\sigma}(t) = 0$. Take

$$\begin{array}{l} v_1 = 0 - \mathrm{sign}(\sigma)a_1(\xi - 1) - a_2(x_1 - \sigma) \\ v_2 = \frac{d^2p}{d\sigma^2} - \mathrm{sign}(\sigma)a_1\left(x_2\xi - \frac{dp}{d\sigma}\right)) - a_2(x_3 - p) \end{array} \tag{18}$$

with $a_1 > 0$ and $a_2 > 0$, then the error dynamics becomes exponentially stable in σ-scale (the term $\mathrm{sign}(\sigma)$ is for dealing with $\dot{\sigma} < 0$).

Similar computations for trailer systems can be found in [15, 12].

2.5 Tracking: flatness and backstepping

2.5.1 Some drawbacks of feedback linearization

We illustrate on two simple (and caricatural) examples that feedback linearization may not lead to the best tracking controller in terms of control effort.

Example 7. Assume we want to track any trajectory $t \mapsto \big(x_r(t), u_r(t)\big)$ of

$$\dot{x} = -x - x^3 + u, \quad x \in \mathbb{R}.$$

The linearizing feedback

$$\begin{aligned} u &= x + x^3 - k\Delta x + \dot{x}_r(t) \\ &= u_r(t) + 3x_r(t)\Delta x^2 + \big(1 + 3x_r^2(t) - k\big)\Delta x + \Delta x^3 \end{aligned}$$

meets this objective by imposing the closed-loop dynamics $\Delta\dot{x} = -k\Delta x$.

But a closer inspection shows the open-loop error dynamics

$$\begin{aligned} \Delta\dot{x} &= -\big(1 + 3x_r^2(t)\big)\Delta x - \Delta x^3 + 3x_r(t)\Delta x^2 + \Delta u \\ &= -\Delta x\big(1 + 3x_r^2(t) - 3x_r(t)\Delta x + \Delta x^2\big) + \Delta u \end{aligned}$$

is naturally stable when the open-loop control $u := u_r(t)$ is applied (indeed $1 + 3x_r^2(t) - 3x_r(t)\Delta x + \Delta x^2$ is always strictly positive). In other words, the linearizing feedback does not take advantage of the natural damping effects.

Example 8. Consider the dynamics

$$\dot{x}_1 = u_1, \qquad \dot{x}_2 = u_2(1 - u_1),$$

for which it is required to track an arbitrary trajectory $t \mapsto \big(x_r(t), u_r(t)\big)$ (notice $u_r(t)$ may not be so easy to define because of the singularity $u_1 = 1$). The linearizing feedback

$$\begin{aligned} u_1 &= -k\Delta x_1 + \dot{x}_{1r}(t) \\ u_2 &= \frac{-k\Delta x_2 + \dot{x}_{2r}(t)}{1 + k\Delta x_1 - \dot{x}_{1r}(t)} \end{aligned}$$

meets this objective by imposing the closed-loop dynamics $\Delta\dot{x} = -k\Delta x$. Unfortunately u_2 grows unbounded as u_1 approaches one. This means we must in practice restrict to reference trajectories such that $|1 - u_{1r}(t)|$ is

always "large" –in particular it is impossible to cross the singularity– and to a "small" gain k.

A smarter control law can do away with these limitations. Indeed, considering the error dynamics

$$\begin{aligned}\Delta\dot{x}_1 &= \Delta u_1\\ \Delta\dot{x}_2 &= (1 - u_{1r}(t) - \Delta u_1)\Delta u_2 - u_{2r}(t)\Delta u_1,\end{aligned}$$

and differentiating the positive function $V(\Delta x) := \frac{1}{2}(\Delta x_1^2 + \Delta x_2^2)$ we get

$$\dot{V} = \Delta u_1(\Delta x_1 - u_{2r}(t)\Delta x_2) + (1 - u_{1r}(t) - \Delta u_1)\Delta u_1\Delta u_2.$$

The control law

$$\begin{aligned}\Delta u_1 &= -k(\Delta x_1 - u_{2r}(t)\Delta x_2)\\ \Delta u_2 &= -(1 - u_{1r}(t) - \Delta u_1)\Delta x_2\end{aligned}$$

does the job since

$$\dot{V} = -\big(\Delta x_1 - u_{2r}(t)\Delta x_2\big)^2 - \big((1 - u_{1r}(t) - \Delta u_1)\Delta x_2\big)^2 \leq 0.$$

Moreover, when $u_{1r}(t) \neq 0$, $\dot{V}$ is zero if and only if $\|\Delta x\|$ is zero. It is thus possible to cross the singularity –which has been made an unstable equilibrium of the closed-loop error dynamics– and to choose the gain k as large as desired. Notice the singularity is overcome by a "truly" multi-input design.

It should not be inferred from the previous examples that feedback linearization necessarily leads to inefficient tracking controllers. Indeed, when the trajectory generator is well-designed, the system is always close to the reference trajectory. Singularities are avoided by restricting to reference trajectories which stay away from them. This makes sense in practice when singularities do not correspond to interesting regions of operations. In this case, designing a tracking controller "smarter" than a linearizing feedback often turns out to be rather complicated, if possible at all.

2.5.2 Backstepping

The previous examples are rather trivial because the control input has the same dimension as the state. More complicated systems can be handled by *backstepping*. Backstepping is a versatile design tool which can be helpful in

a variety of situations: stabilization, adaptive or output feedback, etc ([29] for a complete survey). It relies on the simple yet powerful following idea: consider the system

$$\begin{aligned}\dot{x} &= f(x,\xi), \quad f(x_0,\xi_0)=0\\ \dot{\xi} &= u,\end{aligned}$$

where $(x,\xi)\in\mathbb{R}^n\times\mathbb{R}$ is the state and $u\in\mathbb{R}$ the control input, and assume we can asymptotically stabilize the equilibrium x_0 of the subsystem $\dot{x}=f(x,\xi)$, i.e., we know a control law $\xi=\alpha(x)$, $\alpha(x_0)=\xi_0$ and a positive function $V(x)$ such that

$$\dot{V} = DV(x).f(x,\alpha(x)) \leq 0.$$

A key observation is that the "virtual" control input ξ can then "backstepped" to stabilize the equilibrium (x_0,ξ_0) of the complete system. Indeed, introducing the positive function

$$W(x,\xi) := V(x) + \frac{1}{2}(\xi-\alpha(x))^2$$

and the error variable $z:=\xi-\alpha(x)$, we have

$$\begin{aligned}\dot{W} &= DV(x).f(x,\alpha(x)+z) + z\big(u-\dot{\alpha}(x,\xi)\big)\\ &= DV(x).\big(f(x,\alpha(x)) + R(x,z).z\big) + z\big(u - D\alpha(x).f(x,\xi)\big)\\ &= \dot{V} + z\big(u - D\alpha(x).f(x,\xi) + DV(x).R(x,z)\big),\end{aligned}$$

where we have used the fundamental theorem of calculus to define

$$R(x,h) := \int_0^1 \frac{\partial f}{\partial \xi}(x,x+th)dt$$

(notice $R(x,h)$ is trivially computed when f is linear in ξ). As $\dot{V}$ is negative by assumption, we can make $\dot{W}$ negative, hence stabilize the system, by choosing for instance

$$u := -z + D\alpha(x).f(x,\xi) - DV(x).R(x,z).$$

2.5.3 Blending equivalence with backstepping

Consider a dynamics $\dot{y}=g(y,v)$ for which we would like to solve the tracking problem. Assume it is equivalent to another dynamics $\dot{x}=f(x,u)$ for which

we can solve this problem, i.e., we know a tracking control law together with a Lyapunov function. How can we use this property to control $\dot{y} = g(y, v)$? Another formulation of the question is: assume we know a controller for $\dot{x} = f(x, u)$. How can we derive a controller for

$$\begin{aligned} \dot{x} &= f(x, \kappa(x, z, v)) \\ \dot{z} &= a(x, z, v), \end{aligned}$$

where $u = \kappa(x, z, v), \dot{z} = a(x, z, v)$ is an endogenous feedback? Notice backstepping answers the question for the elementary case where the feedback in question is a pure integrator.

By theorem 2, we can transform $\dot{x} = f(x, u)$ by (dynamic) feedback and coordinate change into

$$\dot{y} = g(y, v), \quad \dot{v} = v^1, \quad \ldots \quad , \quad \dot{v}^\mu = w. \tag{19}$$

for some large enough integer μ. We can then trivially backstep the control from v to w and change coordinates. Using the same reasoning as in section 2.3, it is easy to prove this leads to a control law solving the tracking problem for $\dot{x} = f(x, u)$. In fact, this is essentially the method we followed in section 2.3 on the special case of a flat $\dot{x} = f(x, u)$. We illustrated in section 2.5.1 potential drawbacks of this approach.

However, it is often possible to design better –though in general more complicated– tracking controllers by suitably using backstepping. This point of view is extensively developed in [29], though essentially in the single-input case, where general equivalence boils down to equivalence by coordinate change. In the multi-input case new phenomena occur as illustrated by the following examples.

Example 9 (The PVTOL). We know from example 2 that

$$\begin{aligned} \ddot{x} &= -u_1 \sin\theta + \varepsilon u_2 \cos\theta \\ \ddot{z} &= u_1 \cos\theta + \varepsilon u_2 \sin\theta - 1 \\ \ddot{\theta} &= u_2 \end{aligned} \tag{20}$$

is globally equivalent to

$$\ddot{y}_1 = -\xi \sin\theta, \qquad \ddot{y}_2 = \xi \cos\theta - 1,$$

where $\xi = u_1 + \varepsilon\dot{\theta}^2$. This latter form is rather appealing for designing a tracking controller and leads to the error dynamics

$$\begin{aligned}\Delta\ddot{y}_1 &= -\xi\sin\theta + \xi_r(t)\sin\theta_r(t)\\ \Delta\ddot{y}_2 &= \xi\cos\theta - \xi_r(t)\cos\theta_r(t)\end{aligned}$$

Clearly, if θ were a control input, we could track trajectories by assigning

$$\begin{aligned}-\xi\sin\theta &= \alpha_1(\Delta y_1, \Delta\dot{y}_1) + \ddot{y}_{1r}(t)\\ \xi\cos\theta &= \alpha_2(\Delta y_2, \Delta\dot{y}_2) + \ddot{y}_{2r}(t)\end{aligned}$$

for suitable functions α_1, α_2 and find a Lyapunov function $V(\Delta y, \Delta\dot{y})$ for the system. In other words, we would assign

$$\begin{aligned}\xi &= \Xi\big(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)\big) := \sqrt{(\alpha_1 + \ddot{y}_{1r})^2 + (\alpha_2 + \ddot{y}_{2r})^2}\\ \theta &= \Theta\big(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)\big) := \arg(\alpha_1 + \ddot{y}_{1r}, \alpha_2 + \ddot{y}_{2r}).\end{aligned} \tag{21}$$

The angle θ is a priori not defined when $\xi = 0$, i.e., at the singularity of the flat output y. We will not discuss the possibility of overcoming this singularity and simply assume we stay away from it. Aside from that, there remains a big problem: how should the "virtual" control law (21) be understood? Indeed, it seems to be a differential equation: because y depends on θ, hence Ξ and Θ are in fact functions of the variables

$$x, \dot{x}, z, \dot{z}, \theta, \dot{\theta}, y_r(t), \dot{y}_r(t), \ddot{y}_r(t).$$

Notice ξ is related to the actual control u_1 by a relation that also depends on $\dot{\theta}$.

Let us forget this apparent difficulty for the time being and backstep (21) the usual way. Introducing the error variable $\kappa_1 := \theta - \Theta\big(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)\big)$ and using the fundamental theorem of calculus, the error dynamics becomes

$$\begin{aligned}\Delta\ddot{y}_1 &= \alpha_1(\Delta y_1, \Delta\dot{y}_1) - \kappa_1\ R_{\sin}\big(\Theta(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)), \kappa_1\big)\ \Xi\big(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)\big)\\ \Delta\ddot{y}_2 &= \alpha_2(\Delta y_1, \Delta\dot{y}_1) + \kappa_1\ R_{\cos}\big(\Theta(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)), \kappa_1\big)\ \Xi\big(\Delta y, \Delta\dot{y}, \ddot{y}_r(t)\big)\\ \dot{\kappa}_1 &= \dot{\theta} - \dot{\Theta}\big(\kappa_1, \Delta y, \Delta\dot{y}, \ddot{y}_r(t), y_r^{(3)}(t)\big)\end{aligned}$$

Notice the functions

$$\begin{aligned}R_{\sin}(x,h) &= \sin x\frac{\cos h - 1}{h} + \cos x\frac{\sin h}{h}\\ R_{\cos}(x,h) &= \cos x\frac{\cos h - 1}{h} - \sin x\frac{\sin h}{h}\end{aligned}$$

are bounded and analytic. Differentiate now the positive function

$$V_1(\Delta y, \Delta \dot{y}, \kappa_1) := V(\Delta y, \Delta \dot{y}) + \frac{1}{2}\kappa_1^2$$

to get

$$\begin{aligned}\dot{V}_1 &= \frac{\partial V}{\partial \Delta y_1}\Delta \dot{y}_1 + \frac{\partial V}{\partial \Delta \dot{y}_1}(\alpha_1 - \kappa_1 R_{\sin}\Xi) + \\ &\quad \frac{\partial V}{\partial \Delta y_2}\Delta \dot{y}_2 + \frac{\partial V}{\partial \Delta \dot{y}_2}(\alpha_2 + \kappa_1 R_{\cos}\Xi) + \kappa_1\ (\dot{\theta} - \dot{\Theta}) \\ &= \dot{V} + \kappa_1\ \Big(\dot{\theta} - \dot{\Theta} + \kappa_1\ \big(R_{\cos}\frac{\partial V}{\partial \Delta y_1} - R_{\sin}\frac{\partial V}{\partial \Delta y_2}\big)\ \Xi\Big),\end{aligned}$$

where we have omitted arguments of all the functions for the sake of clarity. If $\dot{\theta}$ were a control input, we could for instance assign

$$\begin{aligned}\dot{\theta} &:= -\kappa_1 + \dot{\Theta} - \kappa_1\ \big(R_{\cos}\frac{\partial V}{\partial \Delta y_1} - R_{\sin}\frac{\partial V}{\partial \Delta y_2}\big)\ \Xi \\ &:= \Theta_1\big(\kappa_1, \Delta y, \Delta \dot{y}, \ddot{y}_r(t), y_r^{(3)}(t)\big),\end{aligned}$$

to get $\dot{V}_1 = \dot{V} - \kappa_1^2 \leq 0$. We thus backstep this "virtual" control law: we introduce the error variable

$$\kappa_2 := \dot{\theta} - \Theta_1\big(\kappa_1, \Delta y, \Delta \dot{y}, \ddot{y}_r(t), y_r^{(3)}(t)\big)$$

together with the positive function

$$V_2(\Delta y, \Delta \dot{y}, \kappa_1, \kappa_2) := V_1(\Delta y, \Delta \dot{y}, \kappa_1) + \frac{1}{2}\kappa_2^2.$$

Differentiating

$$\begin{aligned}V_2 &= \dot{V} + \kappa_1(-\kappa_1 + \kappa_2) + \kappa_2(v_2 - \dot{\Theta}_1) \\ &= \dot{V}_1 + \kappa_2(u_2 - \dot{\Theta}_1 + \kappa_2),\end{aligned}$$

and we can easily make $\dot{V}_1$ negative by assigning

$$u_2 := \Theta_2\big(\kappa_1, \kappa_2, \Delta y, \Delta \dot{y}, \ddot{y}_r(t), y_r^{(3)}(t), y_r^{(4)}(t)\big) \tag{22}$$

for some suitable function Θ_2.

A key observation is that Θ_2 and V_2 are in fact functions of the variables

$$x, \dot{x}, z, \dot{z}, \theta, \dot{\theta}, y_r(t), \ldots, y_r^{(4)}(t),$$

which means (22) makes sense. We have thus built a static control law

$$u_1 = \Xi\big(x, \dot{x}, z, \dot{z}, \theta, \dot{\theta}, y_r(t), \dot{y}_r(t), \ddot{y}_r(t)\big) + \varepsilon\dot{\theta}^2$$
$$u_2 = \Theta_2\big(x, \dot{x}, z, \dot{z}, \theta, \dot{\theta}, y_r(t), \ldots, y_r^{(4)}(t)\big)$$

that does the tracking for (20). Notice it depends on $y_r(t)$ up to the fourth derivative.

Example 10. The dynamics

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = x_3(1 - u_1), \quad \dot{x}_3 = u_2,$$

admits (x_1, x_2) as a flat output. The corresponding endogenous transformation is singular, hence any linearizing feedback blows up, when $u_1 = 1$. However, it is easy to backstep the controller of example 8 to build a globally tracking static controller

Remark. Notice that none the of two previous examples can be linearized by *static* feedback (see section 3.1.2). *Dynamic* feedback is necessary for that. Nevertheless we were able to derive *static* tracking control laws for them. An explanation of why this is possible is that a flat system can in theory be linearized by a *quasistatic* feedback [10] –provided the flat output does not depend on derivatives of the input–.

2.5.4 Backstepping and time-scaling

Backstepping can be combined with linearization and time-scaling, as illustrated in the following example.

Example 11. Consider example 4 and its tracking control defined in example 6. Assume, for example, that $\dot{\sigma} \geq 0$. With the dynamic controller

$$\dot{\xi} = v_1\dot{\sigma}, \quad u_1 = \xi\dot{\sigma}, \quad u_2 = (v_2 - x_2 v_1)/\xi^2$$

where v_1 and v_2 are given by equation (18), we have, for the error $e = y - y_r$, a Lyapunov function $V(e, de/d\sigma)$ satisfying

$$dV/d\sigma \leq -aV \tag{23}$$

with some constant $a > 0$. Remember that $de/d\sigma$ corresponds to $(\xi - 1, x_2\xi - dp/d\sigma)$. Assume now that the real control is not (u_1, u_2) but $(\dot{u}_1 := w_1, u_2)$. With the extended Lyapunov function

$$W = V(e, de/d\sigma) + \frac{1}{2}(u_1 - \xi\dot{\sigma})^2$$

we have

$$\dot{W} = \dot{V} + (w_1 - \dot{\xi}\dot{\sigma} - \xi\ddot{\sigma})((u_1 - \xi\dot{\sigma}).$$

Some manipulations show that

$$\dot{V} = (u_1 - \dot{\sigma}\xi)\left(\frac{\partial V}{\partial e_1} + \frac{\partial V}{\partial e_2}x_2 + \frac{\partial V}{\partial e_2'}u_2\xi\right) + \dot{\sigma}\frac{dV}{d\sigma}$$

(remember $\dot{\xi} = v_1\dot{\sigma}$ and (v_1, v_2) are given by (18)). The feedback ($b > 0$)

$$w_1 = -\left(\frac{\partial V}{\partial e_1} + \frac{\partial V}{\partial e_2}x_2 + \frac{\partial V}{\partial e_2'}u_2\xi\right) + \dot{\xi}\dot{\sigma} + \xi\ddot{\sigma} - b(u_1 - \xi\dot{\sigma})$$

achieves asymptotic tracking since $\dot{W} \leq -a\dot{\sigma}V - b(u_1 - \xi\dot{\sigma})^2$.

2.5.5 Conclusion

It is possible to generalize the previous examples to prove that a control law can be backstepped "through" any endogenous feedback. In particular a flat dynamics can be seen as a (generalized) endogenous feedback acting on the flat output; hence we can backstep a control law for the flat output through the whole dynamics. In other words the flat output serves as a first "virtual" control in the backstepping process. It is another illustration of the fact that a flat output "summarizes" the dynamical behavior.

Notice also that in a tracking problem the knowledge of a flat output is extremely useful not only for the tracking itself (i.e., the closed-loop problem) but also for the trajectory generation (i.e., the open-loop problem)

3 Open problems and new perspectives

3.1 Checking flatness: an overview

3.1.1 The general problem

Devising a general computable test for checking whether $\dot{x} = f(x, u), x \in \mathbb{R}^n, u \in \mathbb{R}^m$ is flat remains up to now an open problem. This means there are no systematic methods for constructing flat outputs. This does not make flatness a useless concept: for instance Lyapunov functions and uniform first integrals of dynamical systems are extremely helpful notions both from a theoretical and practical point of view though they cannot be systematically computed.

The main difficulty in checking flatness is that a candidate flat output $y = h(x, u, \dots, u^{(r)})$ may a priori depend on derivatives of u of arbitrary order r. Whether this order r admits an upper bound (in terms of n and m) is at the moment completely unknown. Hence we do not know whether a finite bound exists at all. In the sequel, we say a system is r-flat if it admits a flat output depending on derivatives of u of order at most r.

To illustrate this upper bound might be at least linear in the state dimension, consider the system

$$x_1^{(\alpha_1)} = u_1, \quad x_2^{(\alpha_2)} = u_2, \quad \dot{x}_3 = u_1 u_2$$

with $\alpha_1 > 0$ and $\alpha_2 > 0$. It admits the flat output

$$y_1 = x_3 + \sum_{i=1}^{\alpha_1} (-1)^i x_1^{(\alpha_1 - i)} u_2^{(i-1)}, \quad y_2 = x_2,$$

hence is r-flat with $r := \min(\alpha_1, \alpha_2) - 1$. We suspect (without proof) there is no flat output depending on derivatives of u of order less than $r - 1$.

If such a bound $\kappa(n, m)$ were known, the problem would amount to checking p-flatness for a *given* $p \leq \kappa(n, m)$ and could be solved in theory. Indeed, it consists [33] in finding m functions $h_1, \dots, h_m$ depending on $(x, u, \dots, u^{(p)})$ such that

$$\dim \operatorname{span} \left\{ dx_1, \dots, dx_n, du_1, \dots, du_m, dh_1^{(\mu)}, \dots, dh_m^{(\mu)} \right\}_{0 \leq \mu \leq \nu} = m(\nu + 1),$$

where $\nu := n + pm$. This means checking the integrability of the partial differential system with a transversality condition

$$\begin{aligned} dx_i \wedge dh \wedge \dots \wedge dh^{(\nu)} &= 0, & i &= 1, \dots, n \\ du_j \wedge dh \wedge \dots \wedge dh^{(\nu)} &= 0, & j &= 1, \dots, m \\ dh \wedge \dots \wedge dh^{(\nu)} &\neq 0, \end{aligned}$$

where $dh^{(\mu)}$ stands for $dh_1^{(\mu)} \wedge \dots \wedge dh_m^{(\mu)}$. It is in theory possible to conclude by using a computable criterion [3, 58], though this seems to lead to practically intractable calculations. Nevertheless it can be hoped that, due to the special structure of the above equations, major simplifications might appear.

3.1.2 Known results

Systems linearizable by static feedback. A system which is linearizable by static feedback and coordinate change is clearly flat. Hence the geometric necessary and sufficient conditions in [26, 25] provide sufficient conditions for flatness. Notice a flat system is in general not linearizable by static feedback (see for instance example 3), with the major exception of the single-input case.

Single-input systems. When there is only one control input flatness reduces to static feedback linearizability [8] and is thus completely characterized by the test in [26, 25].

Affine systems of codimension 1. A system of the form

$$\dot{x} = f_0(x) + \sum_{j=1}^{n-1} u_j g_j(x), \qquad x \in \mathbb{R}^n,$$

i.e., with one input less than states and linear w.r.t. the inputs is 0-flat as soon as it is controllable [8] (more precisely strongly accessible for almost every x).

The picture is much more complicated when the system is not linear w.r.t. the control, see [34] for a geometric sufficient condition.

Affine systems with 2 **inputs and** 4 **states.** Necessary and sufficient conditions for 1-flatness of the system can be found in [56]. They give a good idea of the complexity of checking r-flatness even for r small.

Driftless systems. For driftless systems of the form $\dot{x} = \sum_{i=1}^{m} f_i(x)u_i$ additional results are available.

Theorem 4 (Driftless systems with two inputs [38]). *The system*

$$\dot{x} = f_1(x)u_1 + f_2(x)u_2$$

is flat if and only if the generic rank of E_k is equal to $k+2$ for $k = 0, \ldots, n-2n$ where $E_0 := span\{f_1, f_2\}$, $E_{k+1} := span\{E_k, [E_k, E_k]\}$, $k \geq 0$.

A flat two-input driftless system is always 0-flat. As a consequence of a result in [46], a flat two-input driftless system satisfying some additional regularity conditions can be put by *static* feedback and coordinate change into the *chained system* [47]

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = u_2, \quad \dot{x}_3 = x_2 u_1, \quad \ldots, \quad \dot{x}_n = x_{n-1} u_1.$$

Theorem 5 (Driftless systems, n states, and $n-2$ inputs [39, 40]).

$$\dot{x} = \sum_{i=1}^{n-2} u_i f_i(x), \quad x \in \mathbb{R}^n$$

is flat as soon as it is controllable (i.e., strongly accessible for almost every x). More precisely it is 0*-flat when n is odd, and* 1*-flat when n is even.*

All the results mentioned above rely on the use of exterior differential systems. Additional results on driftless systems, with applications to nonholonomic systems, can be found in [74, 73, 70].

Mechanical systems. For mechanical systems with one control input less than configuration variables, [62] provides a geometric characterization, in terms of the metric derived form the kinetic energy and the control codistribution, of flat outputs depending only on the configuration variables.

A necessary condition. Because it is not known whether flatness can be checked with a finite test, see section 3.1.1, it is very difficult to prove that a system is *not* flat. The following result provides a simple necessary condition.

Theorem 6 (The ruled-manifold criterion [65, 16]). *Assume $\dot{x} = f(x,u)$ is flat. The projection on the p-space of the submanifold $p = f(x,u)$, where x is considered as a parameter, is a ruled submanifold for all x.*

The criterion just means that eliminating u from $\dot{x} = f(x,u)$ yields a set of equations $F(x,\dot{x}) = 0$ with the following property: for all (x,p) such that $F(x,p) = 0$, there exists $a \in \mathbb{R}^n$, $a \neq 0$ such that

$$\forall \lambda \in \mathbb{R}, \quad F(x, p + \lambda a) = 0.$$

$F(x,p) = 0$ is thus a ruled manifold containing straight lines of direction a.

The proof directly derives from the method used by Hilbert [23] to prove the second order Monge equation $\frac{d^2z}{dx^2} = \left(\frac{dy}{dx}\right)^2$ is not solvable without integrals.

A restricted version of this result was proposed in [71] for systems linearizable by a special class of dynamic feedbacks.

As crude as it may look, this criterion is up to now the only way –except for two-input driftless systems– to prove a multi-input system is not flat.

Example 12. The system

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = u_2, \quad \dot{x}_3 = (u_1)^2 + (u_2)^3$$

is not flat, since the submanifold $p_3 = p_1^2 + p_2^3$ is not ruled: there is no $a \in \mathbb{R}^3$, $a \neq 0$, such that

$$\forall \lambda \in \mathbb{R}, p_3 + \lambda a_3 = (p_1 + \lambda a_1)^2 + (p_2 + \lambda a_2)^3.$$

Indeed, the cubic term in λ implies $a_2 = 0$, the quadratic term $a_1 = 0$ hence $a_3 = 0$.

Example 13. The system $\dot{x}_3 = \dot{x}_1^2 + \dot{x}_2^2$ does not define a ruled submanifold of $\mathbb{R}^3$: it is not flat in $\mathbb{R}$. But it defines a ruled submanifold in $\mathbb{C}^3$: in fact it is flat in $\mathbb{C}$, with the flat output

$$y = \big(x_3 - (\dot{x}_1 - \dot{x}_2\sqrt{-1})(x_1 + x_2\sqrt{-1}),\ x_1 + x_2\sqrt{-1}\big).$$

Example 14 (The ball and beam [21]). We now prove by the ruled manifold criterion that

$$\ddot{r} = -Bg\sin\theta + Br\dot{\theta}^2$$
$$(mr^2 + J + J_b)\ddot{\theta} = \tau - 2mr\dot{r}\dot{\theta} - mgr\cos\theta,$$

where $(r, \dot{r}, \theta, \dot{\theta})$ is the state and τ the input, is not flat (as it is a single-input system, we could also prove it is not static feedback linearizable, see section 3.1.2). Eliminating the input τ yields

$$\dot{r} = v_r, \quad \dot{v}_r = -Bg\sin\theta + Br\dot{\theta}^2, \quad \dot{\theta} = v_\theta$$

which defines a ruled manifold in the $(\dot{r}, \dot{v}_r, \dot{\theta}, \dot{v}_\theta)$-space for any r, v_r, θ, v_θ, and we cannot conclude directly. Yet, the system is obviously equivalent to

$$\dot{r} = v_r, \quad \dot{v}_r = -Bg\sin\theta + Br\dot{\theta}^2,$$

which clearly does not define a ruled submanifold for any (r, v_r, θ). Hence the system is not flat.

3.2 Infinite dimension "flat" systems

The idea underlying equivalence and flatness –a one-to-one correspondence between trajectories of systems– is not restricted to control systems described by *ordinary* differential equations. It can be adapted to delay differential systems and to partial differential equations with boundary control. Of course, there are many more technicalities and the picture is far from clear. Nevertheless, this new point of view seems promising for the design of control laws. In this section, we sketch some recent developments in this direction.

3.2.1 Delay systems

Consider for instance the simple differential delay system

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = x_1(t) - x_2(t) + u(t-1).$$

Setting $y(t) := x_1(t)$, we can clearly explicitly parameterize its trajectories by

$$x_1(t) = y(t), \quad x_2(t) = \dot{y}(t), \quad u(t) = \ddot{y}(t+1) + \dot{y}(t+1) - y(t+1).$$

In other words, $y(t) := x_1(t)$ plays the role of a"flat" output. This idea is investigated in detail in [42], where the class of δ-*free* systems is defined (δ is the delay operator). More precisely, [42] considers linear differential delay systems

$$M(d/dt, \delta)w = 0$$

where M is a $(n-m) \times n$ matrix with entries polynomials in d/dt and δ and $w = (w_1, \ldots, w_n)$ are the system variables. Such a system is said to be δ-free if it can be related to the "free" system $y = (y_1, \ldots, y_m)$ consisting of arbitrary functions of time by

$$\begin{aligned} w &= P(d/dt, \delta, \delta^{-1})y \\ y &= Q(d/dt, \delta, \delta^{-1})w, \end{aligned}$$

where P (resp. Q) is a $n \times m$ (resp. $m \times n$) matrix the entries of which are polynomial in d/dt, δ and δ^{-1}.

Many linear delay systems are δ-free. For example, $\dot{x}(t) = Ax(t) + Bu(t-1)$, (A, B) controllable, is δ-free, with the Brunovski output of $\dot{x} = Ax + Bv$ as a "δ-free" output.

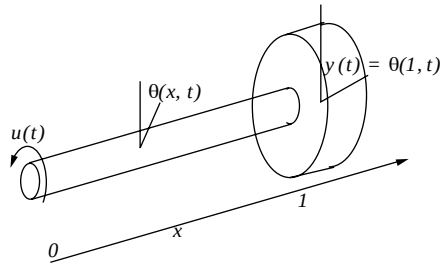


Figure 1: Torsion of a flexible beam

The following systems, commonly used in process control,

$$z_i(s) = \sum_{j=1}^{m} \left\{ \frac{K_i^j \exp(-s\delta_i^j)}{1+\tau_i^j s} \right\} u_j(s), \quad i = 1, \dots p$$

(s Laplace variable, gains K_i^j, delays δ_i^j and time constants τ_i^j between u_j and z_i) are δ-free [54]. Other interesting examples of δ-free systems arise from partial differential equations:

Example 15 (Torsion beam system). The torsion motion of a beam (figure 1) can be modeled in the linear elastic domain by

$$\begin{aligned} \partial_t^2 \theta(x,t) &= \partial_x^2 \theta(x,t), \quad x \in [0,1] \\ \partial_x \theta(0,t) &= u(t) \\ \partial_x \theta(1,t) &= \partial_t^2 \theta(1,t), \end{aligned}$$

where $\theta(x,t)$ is the torsion of the beam and $u(t)$ the control input. From d'Alembert's formula, $\theta(x,t) = \phi(x+t) + \psi(x-t)$, we easily deduce

$$\begin{aligned} 2\theta(t,x) &= \dot{y}(t+x-1) - \dot{y}(t-x+1) + y(t+x-1) + y(t-x+1) \\ 2u(t) &= \ddot{y}(t+1) + \ddot{y}(t-1) - \dot{y}(t+1) + \dot{y}(t-1), \end{aligned}$$

where we have set $y(t) := \theta(1,t)$. This proves the system is δ-free with $\theta(1,t)$ as a "δ-flat" output. See [43, 17] for details and an application to motion planning.

3.2.2 Distributed parameters systems

For partial differential equations with boundary control and mixed systems of partial and ordinary differential equations, it seems possible to describe the

one-to-one correspondence via series expansion, though a sound theoretical framework is yet to be found. We illustrate this original approach to control design on the following two "flat" systems.

Example 16 (Heat equation). Consider the linear heat equation

$$\partial_t \theta(x,t) = \partial_x^2 \theta(x,t), \quad x \in [0,1] \tag{24}$$

$$\partial_x \theta(0,t) = 0 \tag{25}$$

$$\theta(1,t) = u(t), \tag{26}$$

where $\theta(x,t)$ is the temperature and $u(t)$ is the control input. We claim that

$$y(t) := \theta(0,t)$$

is a "flat" output. Indeed, the equation in the Laplace variable s reads

$$s\hat{\theta}(x,s) = \hat{\theta}''(x,s) \quad \text{with} \quad \hat{\theta}'(0,s) = 0, \quad \hat{\theta}(1,s) = \hat{u}(s)$$

($'$ stands for ∂_x and $\hat{}$ for the Laplace transform), and the solution is clearly $\hat{\theta}(x,s) = \cosh(x\sqrt{s})\hat{u}(s)/\cosh(\sqrt{s})$. As $\hat{\theta}(0,s) = \hat{u}(s)/\cosh(\sqrt{s})$, this implies

$$\hat{u}(s) = \cosh(\sqrt{s})\ \hat{y}(s) \quad \text{and} \quad \hat{\theta}(x,s) = \cosh(x\sqrt{s})\ \hat{y}(s).$$

Since $\cosh\sqrt{s} = \sum_{i=0}^{+\infty} s^i/(2i)!$, we eventually get

$$\theta(x,t) = \sum_{i=1}^{+\infty} x^{2i}\ \frac{y^{(i)}(t)}{(2i)!} \tag{27}$$

$$u(t) = \sum_{i=1}^{+\infty} \frac{y^{(i)}(t)}{(2i)!}. \tag{28}$$

In other words, whenever $t \mapsto y(t)$ is an arbitrary function (i.e., a trajectory of the trivial system $y = v$), $t \mapsto \big(\theta(x,t), u(t)\big)$ defined by (27)-(28) is a (formal) trajectory of (24)–(26), and vice versa. This is exactly the idea underlying our definition of flatness in section 1.3. Notice these calculations have been known for a long time, see [75, pp. 588 and 594].

To make the statement precise, we now turn to convergence issues. On the one hand, $t \mapsto y(t)$ must be a smooth function such that

$$\exists\ K, M > 0, \quad \forall i \geq 0, \forall t \in [t_0, t_1], \quad |y^{(i)}(t)| \leq M(Ki)^{2i}$$

to ensure the convergence of the series (27)-(28).

On the other hand $t \mapsto y(t)$ cannot in general be analytic. Indeed, if the system is to be steered from an initial temperature profile $\theta(x, t_0) = \alpha_0(x)$ at time t_0 to a final profile $\theta(x, t_1) = \alpha_1(x)$ at time t_1, equation (24) implies

$$\forall t \in [0, 1], \forall i \geq 0, \quad y^{(i)}(t) = \partial_t^i \theta(0, t) = \partial_x^{2i} \theta(0, t),$$

and in particular

$$\forall i \geq 0, \qquad y^{(i)}(t_0) = \partial_x^{2i} \alpha_0(0) \quad \text{and} \quad y^{(i)}(t_1) = \partial_x^{2i} \alpha_1(1).$$

If for instance $\alpha_0(x) = c$ for all $x \in [0, 1]$ (i.e., uniform temperature profile), then $y(t_0) = c$ and $y^{(i)}(t_0) = 0$ for all $i \geq 1$, which implies $y(t) = c$ for all t when the function is analytic. It is thus impossible to reach any final profile but $\alpha_1(x) = c$ for all $x \in [0, 1]$.

Smooth functions $t \in [t_0, t_1] \mapsto y(t)$ that satisfy

$$\exists\, K, M > 0, \quad \forall i \geq 0, \qquad |y^{(i)}(t)| \leq M (Ki)^{\sigma i}$$

are known as Gevrey-Roumieu functions of order σ [61] (they are also closely related to class S functions [20]). The Taylor expansion of such functions is convergent for $\sigma \leq 1$ and divergent for $\sigma > 1$ (the larger σ is, the "more divergent" the Taylor expansion is). Analytic functions are thus Gevrey-Roumieu of order ≤ 1.

In other words we need a Gevrey-Roumieu function on $[t_0, t_1]$ of order > 1 but ≤ 2, with initial and final Taylor expansions imposed by the initial and final temperature profiles. With such a function, we can then compute open-loop control steering the system from one profile to the other by the formula (27).

For instance, we steered the system from uniform temperature 0 at $t = 0$ to uniform temperature 1 at $t = 1$ by using the function

$$\mathbb{R} \ni t \mapsto y(t) := \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t > 1 \\ \dfrac{\int_0^t \exp\bigl(-1/(\tau(1-\tau))^\gamma\bigr) d\tau}{\int_0^1 \exp\bigl(-1/(\tau(1-\tau))^\gamma\bigr) d\tau} & \text{if } t \in [0, 1], \end{cases}$$

with $\gamma = 1$ (this function is Gevrey-Roumieu functions of order $1 + 1/\gamma$). The evolution of the temperature profile $\theta(x, t)$ is displayed in figure 2 (the Matlab simulation is available upon request at rouchon@cas.ensmp.fr).

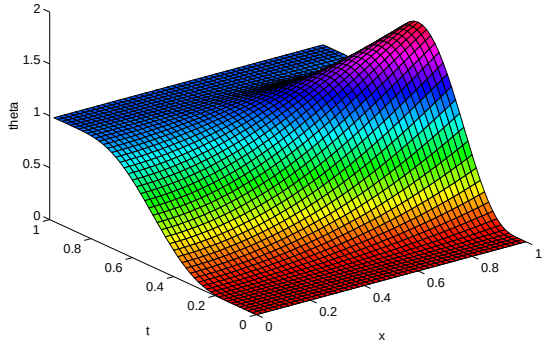


Figure 2: Evolution of the temperature profile for $t \in [0, 1]$

Similar but more involved calculations with convergent series corresponding to Mikunsiński operators are used in [18] to control a flexible rod modeled by an Euler-Bernoulli equation. For nonlinear systems, convergence issues are more involved and are currently under investigation. Yet, it is possible to work –at least formally– along the same line.

Example 17 (Flexion beam system). Consider with [30] the mixed system

$$\begin{aligned} \rho\partial_t^2 u(x,t) &= \rho\omega^2(t)u(x,t) - EI\partial_x^4 u(x,t), \quad x \in [0,1] \\ \dot{\omega}(t) &= \frac{\Gamma_3(t) - 2\omega(t) <u, \partial_t u>(t)}{I_d + <u,u>(t)} \end{aligned}$$

with boundary conditions

$$u(0,t) = \partial_x u(0,t) = 0, \qquad \partial_x^2 u(1,t) = \Gamma_1(t), \qquad \partial_x^3 u(1,t) = \Gamma_2(t),$$

where ρ, EI, I_d are constant parameters, $u(x,t)$ is the deformation of the beam, $\omega(t)$ is the angular velocity of the body and $<f,g>(t) := \int_0^1 \rho f(x,t)g(x,t)dx$. The three control inputs are $\Gamma_1(t)$, $\Gamma_2(t)$, $\Gamma_3(t)$. We claim that

$$y(t) := \big(\partial_x^2 u(0,t),\ \partial_x^3 u(0,t), \omega(t)\big)$$

is a "flat" output. Indeed, $\omega(t)$, $\Gamma_1(t)$, $\Gamma_2(t)$ and $\Gamma_3(t)$ can clearly be expressed in terms of $y(t)$ and $u(x,t)$, which transforms the system into the

equivalent Cauchy-Kovalevskaya form

$$EI\partial_x^4 u(x,t) = \rho y_3^2(t) u(x,t) - \rho \partial_t^2 u(x,t) \qquad \text{and} \qquad \begin{cases} u(0,t) = 0 \\ \partial_x u(0,t) = 0 \\ \partial_x^2 u(0,t) = y_1(t) \\ \partial_x^3 u(0,t) = y_2(t). \end{cases}$$

Set then formally $u(x,t) = \sum_{i=0}^{+\infty} a_i(t) \frac{x^i}{i!}$, plug this series into the above system and identify term by term. This yields

$$a_0 = 0, \qquad a_1 = 0, \qquad a_2 = y_1, \qquad a_3 = y_2,$$

and the iterative relation $\forall i \geq 0,\ EIa_{i+4} = \rho y_3^2 a_i - \rho \ddot{a}_i$. Hence for all $i \geq 1$,

$$a_{4i} = 0 \qquad a_{4i+2} = \frac{\rho}{EI}(y_3^2 a_{4i-2} - \ddot{a}_{4i-2})$$
$$a_{4i+1} = 0 \qquad a_{4i+3} = \frac{\rho}{EI}(y_3^2 a_{4i-1} - \ddot{a}_{4i-1}).$$

There is thus a 1–1 correspondence between (formal) solutions of the system and arbitrary mappings $t \mapsto y(t)$: the system is formally flat.

3.3 State constraints and optimal control

3.3.1 Optimal control

Consider the standard optimal control problem

$$\min_u J(u) = \int_0^T L(x(t), u(t)) dt$$

together with $\dot{x} = f(x,u)$, $x(0) = a$ and $x(T) = b$, for known a, b and T.

Assume that $\dot{x} = f(x,u)$ is flat with $y = h(x, u, \dots, u^{(r)})$ as flat output,

$$x = \varphi(y, \dots, y^{(q)}), \quad u = \alpha(y, \dots, y^{(q)}).$$

A numerical resolution of $\min_u J(u)$ a priori requires a discretization of the state space, i.e., a finite dimensional approximation. A better way is to discretize the flat output space. As in section 1.4, set $y_i(t) = \sum_1^N A_{ij} \lambda_j(t)$. The initial and final conditions on x provide then initial and final constraints

on y and its derivatives up to order q. These constraints define an affine subspace V of the vector space spanned by the the A_{ij}'s. We are thus left with the nonlinear programming problem

$$\min_{A\in V} J(A) = \int_0^T L(\varphi(y,\dots,y^{(q)}),\alpha(y,\dots,y^{(q)}))dt,$$

where the y_i's must be replaced by $\sum_1^N A_{ij}\lambda_j(t)$.

This methodology is used in [50] for trajectory generation and optimal control. It should also be very useful for predictive control. The main expected benefit is a dramatic improvement in computing time and numerical stability. Indeed the exact quadrature of the dynamics –corresponding here to exact discretization via well chosen input signals through the mapping α– avoids the usual numerical sensitivity troubles during integration of $\dot{x} = f(x,u)$ and the problem of satisfying $x(T) = b$.

3.3.2 State constraints

In the previous section, we did not consider state constraints. We now turn to the problem of planning a trajectory steering the state from a to b while satisfying the constraint $k(x,u,\dots,u^{(p)}) \leq 0$. In the flat output "coordinates" this yields the following problem: find $T > 0$ and a smooth function $[0,T] \ni t \mapsto y(t)$ such that $(y,\dots,y^{(q)})$ has prescribed value at $t = 0$ and T and such that $\forall t \in [0,T]$, $K(y,\dots,y^{(\nu)})(t) \leq 0$ for some ν. When $q = \nu = 0$ this problem, known as the *piano mover problem*, is already very difficult.

Assume for simplicity sake that the initial and final states are equilibrium points. Assume also there is a quasistatic motion strictly satisfying the constraints: there exists a *path* (not a trajectory) $[0,1] \ni \sigma \mapsto Y(\sigma)$ such that $Y(0)$ and $Y(1)$ correspond to the initial and final point and for any $\sigma \in [0,1]$, $K(Y(\sigma),0,\dots,0) < 0$. Then, there exists $T > 0$ and $[0,T] \ni t \mapsto y(t)$ solution of the original problem. It suffices to take $Y(\eta(t/T))$ where T is large enough, and where η is a smooth increasing function $[0,1] \ni s \mapsto \eta(s) \in [0,1]$, with $\eta(0) = 0$, $\eta(1) = 1$ and $\frac{d^i\eta}{ds^i}(0,1) = 0$ for $i = 1,\dots,\max(q,\nu)$.

In [64] this method is applied to a two-input chemical reactor. In [60] the minimum-time problem under state constraints is investigated for several mechanical systems. [68] considers, in the context of non holonomic systems, the path planning problem with obstacles. Due to the nonholonomic constraints, the above quasistatic method fails: one cannot set the

y-derivative to zero since they do not correspond to time derivatives but to arc-length derivatives. However, several numerical experiments clearly show that sorting the constraints with respect to the order of y-derivatives plays a crucial role in the computing performance.

3.4 Symmetries

3.4.1 Symmetry preserving flat output

Consider the dynamics $\dot{x} = f(x,u)$, $(x,u) \in X \times U \subset \mathbb{R}^n \times \mathbb{R}^m$. According to section 1 it generates a system $(F, \mathfrak{M})$, where $\mathfrak{M} := X \times U \times \mathbb{R}^\infty_m$ and $F(x,u,u^1,\dots) := (f(x,u),u^1,u^2,\dots)$. At the heart of our notion of equivalence are endogenous transformations, which map solutions of a system to solutions of another system. We single out here the important class of transformations mapping solutions of a system onto solutions of the *same* system:

Definition 6. An endogenous transformation $\Phi_g : \mathfrak{M} \longmapsto \mathfrak{M}$ is a *symmetry* of the system $(F, \mathfrak{M})$ if

$$\forall \xi := (x,u,u^1,\dots) \in \mathfrak{M}, \quad F(\Phi_g(\xi)) = D\Phi_g(\xi) \cdot F(\xi).$$

More generally, we can consider a *symmetry group*, i.e., a collection $\big(\Phi_g\big)_{g\in G}$ of symmetries such that $\forall g_1, g_2 \in G, \Phi_{g_1} \circ \Phi_{g_2} = \Phi_{g_1 * g_2}$, where $(G,*)$ is a group.

Assume now the system is flat. The choice of a flat output is by no means unique, since any endogenous transformation on a flat output gives rise to another flat output.

Example 18 (The kinematic car). The system generated by

$$\dot{x} = u_1 \cos\theta, \quad \dot{y} = u_1 \sin\theta, \quad \dot{\theta} = u_2,$$

admits the 3-parameter symmetry group of planar (orientation-preserving) isometries: for all translation $(a,b)'$ and rotation α , the endogenous mapping generated by

$$\begin{aligned} X &= x\cos\alpha - y\sin\alpha + a \\ Y &= x\sin\alpha + y\cos\alpha + b \\ \Theta &= \theta + \alpha \\ U^1 &= u^1 \\ U^2 &= u^2 \end{aligned}$$

is a symmetry, since the state equations remain unchanged,

$$\dot{X} = U_1 \cos\Theta, \quad \dot{Y} = U_1 \sin\Theta, \quad \dot{\Theta} = U_2.$$

This system is flat $z := (x, y)$ as a flat output. Of course, there are infinitely many other flat outputs, for instance $\tilde{z} := (x, y + \dot{x})$. Yet, z is obviously a more "natural" choice than $\tilde{z}$, because it "respects" the symmetries of the system. Indeed, each symmetry of the system induces a transformation on the flat output z

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \longmapsto \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} z_1 \cos\alpha - z_2 \sin\alpha + a \\ z_1 \sin\alpha + z_2 \cos\alpha + b \end{pmatrix}$$

which does not involve *derivatives* of z, i.e., a *point* transformation. This point transformation generates an endogenous transformation $(z, \dot{z}, \dots) \mapsto (Z, \dot{Z}, \dots)$. Following [19], we say such an endogenous transformation which is the total prolongation of a point transformation is *holonomic*.

On the contrary, the induced transformation on $\tilde{z}$

$$\begin{pmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{pmatrix} \longmapsto \begin{pmatrix} \tilde{Z}_1 \\ \tilde{Z}_2 \end{pmatrix} = \begin{pmatrix} X \\ Y + \dot{X} \end{pmatrix} = \begin{pmatrix} \tilde{z}_1 \cos\alpha + (\dot{\tilde{z}}_1 - \tilde{z}_2)\sin\alpha + a \\ \tilde{z}_1 \sin\alpha + \tilde{z}_2 \cos\alpha + (\ddot{\tilde{z}}_1 - \dot{\tilde{z}}_2)\sin\alpha + b \end{pmatrix}$$

is *not* a point transformation (it involves derivatives of $\tilde{z}$) and does not give to a holonomic transformation.

Consider the system $(F, \mathfrak{M})$ admitting a symmetry Φ_g (or a symmetry group $\left(\Phi_g\right)_{g\in G}$). Assume moreover the system is flat with h as a flat output and denotes by $\Psi := (h, \dot{h}, \ddot{h}, \dots)$ the endogenous transformation generated by h. We then have:

Definition 7 (Symmetry-preserving flat output). The flat output h *preserves* the symmetry Φ_g if the composite transformation $\Psi \circ \Phi_g \circ \Psi^{-1}$ is holonomic.

This leads naturally to a fundamental question: assume a flat system admits the symmetry group $\left(\Phi_g\right)_{g\in G}$. Is there a flat output which preserves $\left(\Phi_g\right)_{g\in G}$?

This question can in turn be seen as a special case of the following problem: view a dynamics $\dot{x} - f(x, u) = 0$ as an *underdetermined differential system* and assume it admits a symmetry group; can it then be reduced to a "smaller" differential system? Whereas this problem has been studied for a long time and received a positive answer in the *determined* case, the underdetermined case seems to have been barely untouched [53].

3.4.2 Flat outputs as potentials and gauge degree of freedom

Symmetries and the quest for potentials are at the heart of physics. To end the paper, we would like to show that flatness fits into this broader scheme.

Maxwell's equations in an empty medium imply that the magnetic field H is divergent free, $\nabla \cdot H = 0$. In Euclidian coordinates (x_1, x_2, x_3), it gives the underdetermined partial differential equation

$$\frac{\partial H_1}{\partial x_1} + \frac{\partial H_2}{\partial x_2} + \frac{\partial H_3}{\partial x_3} = 0$$

A key observation is that the solutions to this equation derive from a vector potential $H = \nabla \times A$: the constraint $\nabla \cdot H = 0$ is automatically satisfied whatever the potential A. This potential parameterizes all the solutions of the underdetermined system $\nabla \cdot H = 0$, see [59] for a general theory. A is a priori not uniquely defined, but up to an arbitrary gradient field, the gauge degree of freedom. The symmetries of the problem indicate how to use this degree of freedom to fix a "natural" potential.

The picture is similar for flat systems. A flat output is a "potential" for the underdetermined differential equation $\dot{x} - f(x, u) = 0$. Endogenous transformations on the flat output correspond to gauge degrees of freedom. The "natural" flat output is determined by symmetries of the system. Hence controllers designed from this flat output can also preserve the physics.

A slightly less esoteric way to convince the reader that flatness is an interesting notion is to take a look at the following small catalog of flat systems.

4 A catalog of flat systems

We give here a (partial) list of flat systems encountered in applications.

4.1 Holonomic mechanical systems

Example 19 (Fully actuated holonomic systems). The dynamics of a holonomic system with as many independent inputs as configuration variables is

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = M(q)u + D(q, \dot{q}),$$

with $M(q)$ invertible. It admits q as a flat output –even when $\frac{\partial^2 L}{\partial \dot{q}^2}$ is singular – : indeed, u can be expressed in function of $q, \dot{q}$ by the *computed torque*

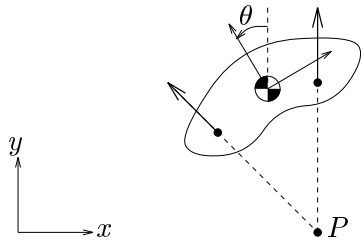


Figure 3: A rigid body controlled by two body fixed forces.

formula

$$u = M(q)^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} - D(q,\dot{q})\right).$$

If q is constrained by $c(q) = 0$ the system remains flat, and the flat output corresponds to the configuration point in $c(q) = 0$.

Example 20 (Planar rigid body with forces). Consider a planar rigid body moving in a vertical plane under the influence of gravity and controlled by two forces having lines of action that are fixed with respect to the body and intersect at a single point (see figure 3). Let (x, y) represent the horizontal and vertical coordinates of center of mass G of the body with respect to a stationary frame, and let θ be the counterclockwise orientation of a body fixed line through the center of mass. Take m as the mass of the body and J as the moment of inertia. Let $g \approx 9.8$ m/sec^2 represent the acceleration due to gravity.

Without loss of generality, we will assume that the lines of action for F_1 and F_2 intersect the y axis of the rigid body and that F_1 and F_2 are perpendicular. The equations of motion for the system can be written as

$$\begin{aligned} m\ddot{x} &= F_1\cos\theta - F_2\sin\theta \\ m\ddot{y} &= F_1\sin\theta + F_2\cos\theta - mg \\ J\ddot{\theta} &= rF_1. \end{aligned}$$

The flat output of this system corresponds to *Huyghens center of oscillation* [16]

$$(x - \frac{J}{mr}\sin\theta, \quad y + \frac{J}{mr}\cos\theta).$$

This example has some practical importance. The PVTOL system, the gantry crane and the robot $2k\pi$ (see below) are of this form, as is the sim-

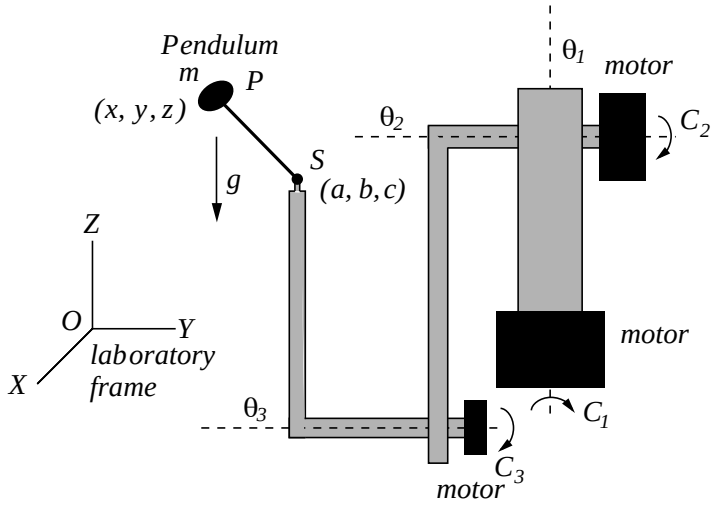


Figure 4: The robot $2k\pi$ carrying its pendulum.

plified *planar ducted fan* [49]. Variations of this example can be formed by changing the number and type of the inputs [45].

Example 21 (PVTOL aircraft). A simplified Vertical Take Off and Landing aircraft moving in a vertical Plane [22] can be described by

$$\begin{aligned}\ddot{x} &= -u_1 \sin\theta + \varepsilon u_2 \cos\theta \\ \ddot{z} &= u_1 \cos\theta + \varepsilon u_2 \sin\theta - 1 \\ \ddot{\theta} &= u_2.\end{aligned}$$

A flat output is $y = (x - \varepsilon \sin\theta, z + \varepsilon \cos\theta)$, see [37] more more details and a discussion in relation with unstable zero dynamics.

Example 22 (The robot $2k\pi$ of Ecole des Mines). It is a robot arm carrying a pendulum, see figure 4. The control objective is to flip the pendulum from its natural downward rest position to the upward position and maintains it there. The first three degrees of freedom (the angles $\theta_1, \theta_2, \theta_3$) are actuated by electric motors, while the two degrees of freedom of the pendulum are not actuated.

The position $P = (x, y, z)$ of the pendulum oscillation center is a flat output. Indeed, it is related to the position $S = (a, b, c)$ of the suspension

point by

$$\begin{aligned}(x-a)(\ddot{z}+g) &= \ddot{x}(z-c)\\ (y-b)(\ddot{z}+g) &= \ddot{y}(z-c)\\ (x-a)^2+(y-b)^2+(z-c)^2 &= l^2,\end{aligned}$$

where l is the distance between S and P. On the other hand the geometry of the robot defines a relation $(a,b,c)=\mathcal{T}(\theta_1,\theta_2,\theta_3)$ between the position of S and the robot configuration. This relation is locally invertible for almost all configurations but is not globally invertible.

Example 23 (Gantry crane [16]). A direct application of Newton's laws provides the implicit equations of motion

$$\begin{aligned}m\ddot{x} &= -T\sin\theta & x &= R\sin\theta + D\\ m\ddot{z} &= -T\cos\theta + mg & z &= R\cos\theta,\end{aligned}$$

where x, z, θ are the configuration variables and T is the tension in the cable. The control inputs are the trolley position D and the cable length R. This system is flat, with the position (x,z) of the load as a flat output.

Example 24 (Conventional aircraft). A conventional aircraft is flat, provided some small aerodynamic effects are neglected, with the coordinates of the center of mass and side-slip angle as a flat output. See [33] for a detailed study.

Example 25 (Towed cable system). Consider the dynamics of a system consisting of an aircraft flying in a circular pattern while towing a cable with a tow body (drogue) attached at the bottom. Under suitable conditions, the cable reaches a relative equilibrium in which the cable maintains its shape as it rotates. By choosing the parameters of the system appropriately, it is possible to make the radius at the bottom of the cable much smaller than the radius at the top of the cable. This is illustrated in Figure 5. The motion of the towed cable system can be approximately represented using a finite element model in which segments of the cable are replaced by rigid links connected by spherical joints. The forces acting on the segment (tension, aerodynamic drag and gravity) are lumped and applied at the end of each rigid link. In addition to the forces on the cable, we must also consider the forces on the drogue and the towplane. The drogue is modeled as a sphere and essentially acts as a mass attached to the last link of the cable, so that the forces acting on it are included in the cable dynamics.

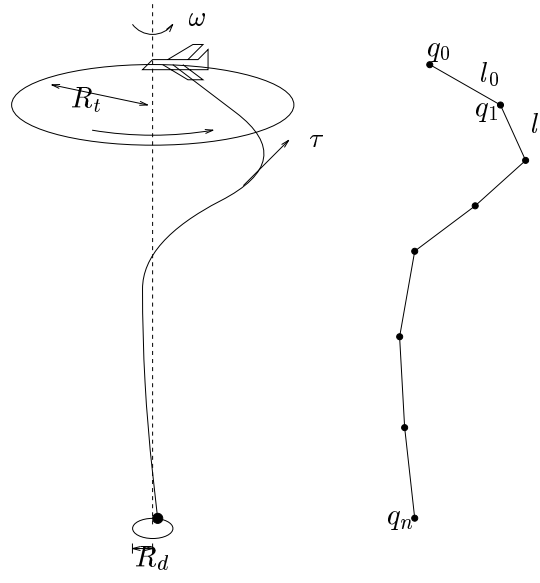


Figure 5: Towed cable system and finite link approximate model.

The external forces on the drogue again consist of gravity and aerodynamic drag. The towplane is attached to the top of the cable and is subject to drag, gravity, and the force of the attached cable. For simplicity, we simply model the towplane as a pure force applied at the top of the cable. Our goal is to generate trajectories for this system that allow operation away from relative equilibria as well as transition between one equilibrium point and another. Due to the high dimension of the model for the system (128 states is typical), traditional approaches to solving this problem, such as optimal control theory, cannot be easily applied. However, it can be shown that this system is differentially flat using the position of the bottom of the cable as the differentially flat output. Thus all feasible trajectories for the system are characterized by the trajectory of the bottom of the cable. See [44] for a more complete description and additional references.

We end this section with a system which is not known to be flat for generic parameter value but still enjoys the weaker property of being *orbitally* flat [14].

Example 26 (Satellite with two controls). Consider with [4] a satellite with

two control inputs u_1, u_2 described by

$$\begin{aligned}
\dot{\omega}_1 &= u_1 \\
\dot{\omega}_2 &= u_2 \\
\dot{\omega}_3 &= a\omega_1\omega_2 \\
\dot{\varphi} &= \omega_1 \cos\theta + \omega_3 \sin\theta \\
\dot{\theta} &= (\omega_1 \sin\theta - \omega_3 \cos\theta)\tan\varphi + \omega_2 \\
\dot{\psi} &= \frac{(\omega_3 \cos\theta - \omega_1 \sin\theta)}{\cos\varphi},
\end{aligned} \tag{29}$$

where $a = (J_1 - J_2)/J_3$ (J_i are the principal moments of inertia); physical sense imposes $|a| \leq 1$. Eliminating u_1, u_2 and ω_1, ω_2 by

$$\omega_1 = \frac{\dot{\varphi} - \omega_3 \sin\theta}{\cos\theta} \quad \text{and} \quad \omega_2 = \dot{\theta} + \dot{\psi}\sin\varphi$$

yields the equivalent system

$$\dot{\omega}_3 = a(\dot{\theta} + \dot{\psi}\sin\varphi)\frac{\dot{\varphi} - \omega_3 \sin\theta}{\cos\theta} \tag{30}$$

$$\dot{\psi} = \frac{\omega_3 - \dot{\varphi}\sin\theta}{\cos\varphi\cos\theta}. \tag{31}$$

But this system is in turn equivalent to

$$\begin{aligned}
\cos\theta\big(\ddot{\psi}\cos\varphi - (1+a)\dot{\psi}\dot{\varphi}\sin\varphi\big) + \sin\theta\big(\ddot{\varphi} + a\dot{\psi}^2\sin\varphi\cos\varphi\big) \\
+ \dot{\theta}(1-a)(\dot{\varphi}\cos\theta - \dot{\psi}\sin\theta\cos\varphi) = 0
\end{aligned}$$

by substituting $\omega_3 = \dot{\psi}\cos\varphi\cos\theta + \dot{\varphi}\sin\theta$ in (30).

When $a = 1$, θ can clearly be expressed in function of φ, ψ and their derivatives. We have proved that (29) is flat with (φ, ψ) as a flat output. A similar calculation can be performed when $a = -1$.

When $|a| < 1$, whether (29) is flat is unknown. Yet, it is *orbitally* flat [63]. To see that, rescale time by $\dot{\sigma} = \omega_3$; by the chain rule $\dot{x} = \dot{\sigma}x'$ whatever the variable x, where $'$ denotes the derivation with respect to σ. Setting then

$$\bar{\omega}_1 := \omega_1/\omega_3, \quad \bar{\omega}_2 := \omega_2/\omega_3, \quad \bar{\omega}_3 := -1/a\omega_3,$$

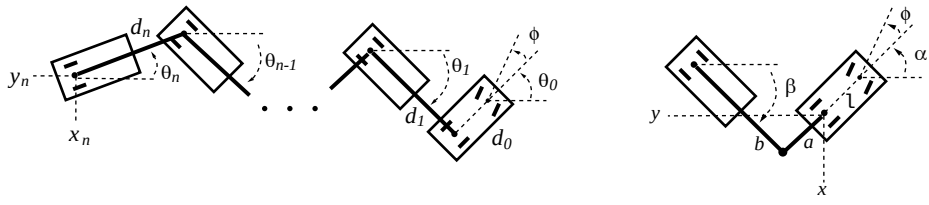


Figure 6: n-trailer system (left) and 1-trailer system with kingpin hitch (right).

and eliminating the controls transforms (29) into

$$\begin{aligned}
\omega_3' &= \bar{\omega}_1 \bar{\omega}_2 \\
\varphi' &= \bar{\omega}_1 \cos\theta + \sin\theta \\
\theta' &= (\bar{\omega}_1 \sin\theta - \cos\theta)\tan\varphi + \bar{\omega}_2 \\
\psi' &= \frac{(\cos\theta - \bar{\omega}_1 \sin\theta)}{\cos\varphi}.
\end{aligned}$$

The equations are now independent of a. This implies the satellite with $a \neq 1$ is orbitally equivalent to the satellite with $a = 1$. Since it is flat when $a = 1$ it is orbitally flat when $a \neq 1$, with (φ, ψ) as an orbitally flat output.

4.2 Nonholonomic mechanical systems

Example 27 (Kinematics generated by two *nonholonomic constraints).* Such systems are flat by theorem 5 since they correspond to driftless systems with n states and $n-2$ inputs. For instance the rolling disc (p. 4), the rolling sphere (p. 96) and the bicycle (p. 330) considered in the classical treatise on nonholonomic mechanics [48] are flat.

Example 28 (Mobile robots). Many mobile robots modeled by rolling without sliding constraints, such as those considered in [5, 47, 74] are flat. In particular, the n-trailer system (figure 6) has for flat output the mid-point P_n of the last trailer axle [67, 16]. The 1-trailer system with kingpin hitch is also flat, with a rather complicated flat output involving elliptic integrals [66, 12], but by theorem 4 the system is *not* flat when there is more than one trailer.

Example 29 (The rolling penny). The dynamics of this Lagrangian system

submitted to a nonholonomic constraint is described by

$$\begin{aligned} \ddot{x} &= \lambda \sin\varphi + u_1 \cos\varphi \\ \ddot{y} &= -\lambda \cos\varphi + u_1 \sin\varphi \\ \ddot{\varphi} &= u_2 \\ \dot{x}\sin\varphi &= \dot{y}\cos\varphi \end{aligned}$$

where x, y, φ are the configuration variables, λ is the Lagrange multiplier of the constraint and u_1, u_2 are the control inputs. A flat output is (x, y): indeed, parameterizing time by the arclength s of the curve $t \mapsto (x(t), y(t))$ we find

$$\cos\varphi = \frac{\mathrm{d}x}{\mathrm{d}s}, \quad \sin\varphi = \frac{\mathrm{d}y}{\mathrm{d}s}, \quad u_1 = \dot{s}, \quad u_2 = \kappa(s)\ \ddot{s} + \frac{\mathrm{d}\kappa}{\mathrm{d}s}\ \dot{s}^2,$$

where κ is the curvature. These formulas remain valid even if $u_1 = u_2 = 0$.

This example can be generalized to any mechanical system subject to m flat nonholonomic constraints, provided there are $n - m$ control forces independent of the constraint forces (n the number of configuration variables), i.e., a "fully-actuated" nonholonomic system as in [5].

All these flat nonholonomic systems have a controllability singularity at rest. Yet, it is possible to "blow up" the singularity by reparameterizing time with the arclength of the curve described by the flat output, hence to plan and track trajectories starting from and stopping at rest as explained in sections 1.5 and 2.4, see [16, 67, 12] for more details.

4.3 Electromechanical systems

Example 30 (DC-to-DC converter). A Pulse Width Modulation DC-to-DC converter can be modeled by

$$\dot{x}_1 = (u-1)\frac{x_2}{L} + \frac{E}{L}, \qquad \dot{x}_2 = (1-u)\frac{x_1}{LC} - \frac{x_2}{RC},$$

where the duty ratio $u \in [0, 1]$ is the control input. The electrical stored energy $y := \dfrac{x_1^2}{2C} + \dfrac{x_2^2}{2L}$ is a flat output [69, 27].

Example 31 (Magnetic bearings). A simple flatness-based solution to motion planning and tracking is proposed in [32]. The control law ensures that only one electromagnet in each actuator works at a time and permits to reduce the number of electromagnets by a better placement of actuators.

Example 32 (Induction motor). The standard two-phase model of the induction motor reads in complex notation (see [31] for a complete derivation)

$$\begin{aligned} R_s i_s + \dot{\psi}_s &= u_s & \psi_s &= L_s i_s + M e^{jn\theta} i_r \\ R_r i_r + \dot{\psi}_r &= 0 & \psi_r &= M e^{-jn\theta} i_s + L_r i_r, \end{aligned}$$

where ψ_s and i_s (resp. ψ_r and i_r) are the complex stator (resp. rotor) flux and current, θ is the rotor position and $j = \sqrt{-1}$. The control input is the voltage u_s applied to the stator. Setting $\psi_r = \rho e^{j\alpha}$, the rotor motion is described by

$$J\frac{d^2\theta}{dt^2} = \frac{n}{R_r}\rho^2\dot{\alpha} - \tau_L(\theta, \dot{\theta}),$$

where τ_L is the load torque.

This system is flat with the two angles (θ, α) as a flat output [41] (see [9] also for a related result).

4.4 Chemical systems

Example 33 (CSTRs). Many simple models of Continuous Stirred Tank Reactors (CSTRs) admit flats outputs with a direct physical interpretation in terms of temperatures or product concentrations [24, 1], as do closely related biochemical processes [2, 11]. In [64] flatness is used to steer a reactor model from a steady state to another one while respecting some physical constraints.

A basic model of a CSTR with *two* chemical species and any number of exothermic or endothermic reactions is

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2) + g_1(x_1, x_2)u \\ \dot{x}_2 &= f_2(x_1, x_2) + g_2(x_1, x_2)u, \end{aligned}$$

where x_1 is a concentration, x_2 a temperature and u the control input (feed-flow or heat exchange). It is obviously linearizable by static feedback, hence flat.

When more chemical species are involved, a single-input CSTR is in general not flat, see [28]. Yet, the addition of another manipulated variable often renders it flat, see [1] for an example on a free-radical polymerization CSTR. For instance basic model of a CSTR with *three* chemical species, any

number of exothermic or and two control inputs is

$$\begin{aligned}\dot{x}_1 &= f_1(x) + g_1^1(x)u_1 + g_1^2(x)u_2\\ \dot{x}_2 &= f_2(x) + g_2^1(x)u_1 + g_2^2(x)u_2\\ \dot{x}_3 &= f_3(x) + g_3^1(x)u_1 + g_3^2(x)u_2,\end{aligned}$$

where x_1, x_2 are concentrations and x_3 is a temperature temperature and u_1, u_2 are the control inputs (feed-flow, heat exchange, feed-composition, ...). Such a system is always flat, see section 3.1.2.

Example 34 (Polymerization reactor). Consider with [72] the reactor

$$\begin{aligned}\dot{C}_m &= \frac{C_{m_{m_s}}}{\tau} - \left(1 + \overline{\varepsilon}\frac{\mu_1}{\mu_1 + M_m C_m}\right)\frac{C_m}{\tau} + R_m(C_m, C_i, C_s, T)\\ \dot{C}_i &= -k_i(T)C_i + u_2\frac{C_{i_{i_s}}}{V} - \left(1 + \overline{\varepsilon}\frac{\mu_1}{\mu_1 + M_m C_m}\right)\frac{C_i}{\tau}\\ \dot{C}_s &= u_2\frac{C_{s_{i_s}}}{V} + \frac{C_{s_{m_s}}}{\tau} - \left(1 + \overline{\varepsilon}\frac{\mu_1}{\mu_1 + M_m C_m}\right)\frac{C_s}{\tau}\\ \dot{\mu}_1 &= -M_m R_m(C_m, C_i, C_s, T) - \left(1 + \overline{\varepsilon}\frac{\mu_1}{\mu_1 + M_m C_m}\right)\frac{\mu_1}{\tau}\\ \dot{T} &= \phi(C_m, C_i, C_s, \mu_1, T) + \alpha_1 T_j\\ \dot{T}_j &= f_6(T, T_j) + \alpha_4 u_1,\end{aligned}$$

where u_1, u_2 are the control inputs and $C_{m_{m_s}}$, M_m, $\overline{\varepsilon}$, τ, $C_{i_{i_s}}$, $C_{s_{m_s}}$, $C_{s_{i_s}}$, V, α_1, α_4 are constant parameters. The functions R_m, k_i, ϕ and f_6 are not well-known and derive from experimental data and semi-empirical considerations, involving kinetic laws, heat transfer coefficients and reaction enthalpies.

The polymerization reactor is flat whatever the functions R_m, k_i, ϕ, f_6 and admits $(C_{s_{i_s}} C_i - C_{i_{i_s}} C_s,\ M_m C_m + \mu_1)$ as a flat output [65].

References

[1] J. Alvarez, R. Suarez, and A. Sanchez. Nonlinear decoupling control of free-radical polymerization continuous stirred tank reactors. *Chem. Engng. Sci*, 45:3341–3357, 1990.

[2] G. Bastin and Dochain. *On-Line Estimation and Adaptive Control of Bioreactors*. Elsevier Science Publishing Co, 1990.

[3] R.L. Bryant, S.S. Chern, R.B. Gardner, H.L. Goldschmidt, and P.A. Griffiths. *Exterior Differential Systems*. Springer, 1991.

[4] C.I. Byrnes and A. Isidori. On the attitude stabilization of rigid spacecraft. *Automatica*, 27:87–95, 1991.

[5] G. Campion, B. d'Andrea Novel, and G. Bastin. Structural properties and classification of kinematic and dynamic models of wheeled mobile robots. *IEEE Trans. Robotics Automation*, 12(1):47–62, 1996.

[6] E. Cartan. Sur l'équivalence absolue de certains systèmes d'équations différentielles et sur certaines familles de courves. *Bull. Soc. Math. France*, 42:12–48, 1914. Also in Œuvres Complètes, part II, vol. 2, pp.1133–1168, CNRS, Paris, 1984.

[7] E. Cartan. Sur l'intégration de certains systèmes indéterminés d'équations différentielles. *J. für reine und angew. Math.*, 145:86–91, 1915. Also in Œuvres Complètes, part II, vol. 2, pp.1164–1174, CNRS, Paris, 1984.

[8] B. Charlet, J. Lévine, and R. Marino. On dynamic feedback linearization. *Systems Control Letters*, 13:143–151, 1989.

[9] J. Chiasson. Dynamic feedback linearization of the induction motor. *IEEE Trans. Automat. Control*, 38:1588–1594, 1993.

[10] E. Delaleau and J. Rudolph. Decoupling and linearization by quasi-static feedback of generalized states. In *Proc. of the 3rd European Control Conf.*, pages 1069–1074, Rome, 1995.

[11] J. El Moubaraki, G. Bastin, and J. Lévine. Nonlinear control of biological processes with growth/production decoupling. *Mathematical Biosciences*, 116:21–44, 1993.

[12] M. Fliess, J. levine, P. Martin, F. Ollivier, and P. Rouchon. Controlling nonlinear systems by flatness. In C.I. Byrnes, B.N. Datta, D.S. Gilliam, and C.F. Martin, editors, *Systems and control in the Twenty-First Century*, Progress in Systems and Control Theory. Birkhauser, 1997.

[13] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon. Sur les systèmes non linéaires différentiellement plats. *C.R. Acad. Sci. Paris*, I–315:619–624, 1992.

[14] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon. Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund. *C.R. Acad. Sci. Paris*, I-317:981–986, 1993.

[15] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon. Design of trajectory stabilizing feedback for driftless flat systems. In *Proc. of the 3rd European Control Conf.*, pages 1882–1887, Rome, 1995.

[16] M. Fliess, J. Lévine, Ph. Martin, and P. Rouchon. Flatness and defect of nonlinear systems: introductory theory and examples. *Int. J. Control*, 61(6):1327–1361, 1995.

[17] M. Fliess, H. Mounier, P. Rouchon, and J. Rudolph. Controllability and motion planning for linear delay systems with an application to a flexible rod. In *Proc. of the 34th IEEE Conf. on Decision and Control*, pages 2046–2051, New Orleans, 1995.

[18] M. Fliess, H. Mounier, P. Rouchon, and J. Rudolph. Systèmes linéaires sur les opérateurs de Mikusiński et commande d'une poutre flexible. In *ESAIM Proc. "Élasticité, viscolélasticité et contrôle optimal", 8ème entretiens du centre Jacques Cartier, Lyon*, pages 157–168, 1996.

[19] M. Gromov. *Partial Differential Relations.* Springer-Verlag, 1986.

[20] I.M. Guelfand and G.E. Chilov. *Les Distributions, tome 3.* Dunod, Paris, 1964.

[21] J. Hauser, S. Sastry, and P. Kokotović. Nonlinear control via approximated input-output linearization: the ball and beam example. *IEEE Trans. Automat. Contr.*, 37:392–398, 1992.

[22] J. Hauser, S. Sastry, and G. Meyer. Nonlinear control design for slightly nonminimum phase systems: Application to V/STOL aircraft. *Automatica*, 28(4):665–679, 1992.

[23] D. Hilbert. Über den Begriff der Klasse von Differentialgleichungen. *Math. Ann.*, 73:95–108, 1912. also in Gesammelte Abhandlungen, vol. III, pp. 81–93, Chelsea, New York, 1965.

[24] K.A. Hoo and J.C. Kantor. An exothermic continuous stirred tank reactor is feedback equivalent to a linear system. *Chem. Eng. Commun.*, 37:1–10, 1985.

[25] L.R. Hunt, R. Su, and G. Meyer. Global transformations of nonlinear systems. *IEEE Trans. Automat. Control*, 28:24–31, 1983.

[26] B. Jakubczyk and W. Respondek. On linearization of control systems. *Bull. Acad. Pol. Sci. Ser. Sci. Math.*, 28:517–522, 1980.

[27] L. Karsenti and P. Rouchon. A tracking controller-observer scheme for DC-to-DC converters. In *ECC'97*, 1997.

[28] C. Kravaris and C.B. Chung. Nonlinear state feedback synthesis by global input/output linearization. *AIChE J.*, 33:592–603, 1987.

[29] M. Krstić, I. Kanellakopoulos, and P. Kokotović. *Nonlinear and Adaptive Control Design*. John Wiley & Sons, Inc., 1995.

[30] H. Laousy, C.Z. Xu, and G. Sallet. Boundary feedback stabilization of rotation body-beam system. *IEEE Autom. Control*, 41:1–5, 1996.

[31] W. Leonhard. *Control of Electrical Drives*. Elsevier, 1985.

[32] J. Lévine, J. Lottin, and J.-C. Ponsart. A nonlinear approach to the control of magnetic bearings. *IEEE Trans. Control Systems Technology*, 4:524–544, 1996.

[33] Ph. Martin. *Contribution à l'étude des systèmes diffèrentiellement plats*. PhD thesis, École des Mines de Paris, 1992.

[34] Ph. Martin. A geometric sufficient conditions for flatness of systems with m inputs and $m+1$ states. In *Proc. of the* 32*nd IEEE Conf. on Decision and Control*, pages 3431–3436, San Antonio, 1993.

[35] Ph. Martin. An intrinsic condition for regular decoupling. *Systems & Control Letters*, 20:383–391, 1993.

[36] Ph. Martin. Endogenous feedbacks and equivalence. In *Systems and Networks: Marthematical Theory and Applications (MTNS'93)*, volume II, pages 343–346. Akademie Verlag, Berlin, 1994.

[37] Ph. Martin, S. Devasia, and B Paden. A different look at output feedback: control of a VTOL aircraft. *Automatica*, 32(1):101–108, 1996.

[38] Ph. Martin and P. Rouchon. Feedback linearization and driftless systems. *Math. Control Signal Syst.*, 7:235–254, 1994.

[39] Ph. Martin and P. Rouchon. Any (controllable) driftless system with 3 inputs and 5 states is flat. *Systems Control Letters*, 25:167–173, 1995.

[40] Ph. Martin and P. Rouchon. Any (controllable) driftless system with m inputs and m+2 states is flat. In *Proc. of the* 34*th IEEE Conf. on Decision and Control*, pages 2886–2891, New Orleans, 1995.

[41] Ph. Martin and P. Rouchon. Flatness and sampling control of induction motors. In *Proc. IFAC World Congress*, pages 389–394, San Francisco, 1996.

[42] H. Mounier. *Propriétés structurelles des systèmes linéaires à retards: aspects théoriques et pratiques.* PhD thesis, Université Paris Sud, Orsay, 1995.

[43] H. Mounier, J. Rudolph, M. Petitot, and M. Fliess. A flexible rod as a linear delay system. In *Proc. of the* 3*rd European Control Conf.*, pages 3676–3681, Rome, 1995.

[44] R. M. Murray. Trajectory generation for a towed cable flight control system. In *Proc. IFAC World Congress*, pages 395–400, San Francisco, 1996.

[45] R. M. Murray, M. Rathinam, and W. Sluis. Differential flatness of mechanical control systems: A catalog of prototype systems. In *asmeIMECE*, San Francisco, November 1995.

[46] R.M. Murray. Nilpotent bases for a class on nonintegrable distributions with applications to trajectory generation for nonholonomic systems. *Math. Control Signal Syst.*, 7:58–75, 1994.

[47] R.M. Murray and S.S. Sastry. Nonholonomic motion planning: Steering using sinusoids. *IEEE Trans. Automat. Control*, 38:700–716, 1993.

[48] Iu I. Neimark and N.A. Fufaev. *Dynamics of Nonholonomic Systems*, volume 33 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, 1972.

[49] M. van Nieuwstadt and R. M. Murray. Approximate trajectory generation for differentially flat systems with zero dynamics. In *Proc. of the 34th IEEE Conf. on Decision and Control*, pages 4224–4230, New Orleans, 1995.

[50] M. van Nieuwstadt and R.M. Murray. Real time trajectory generation for differentially flat systems. *Int. Journal of Robust and Nonlinear Control*, 1997. submitted for publication.

[51] M. van Nieuwstadt, M. Rathinam, and R.M. Murray. Differential flatness and absolute equivalence. In *Proc. of the 33rd IEEE Conf. on Decision and Control*, pages 326–332, Lake Buena Vista, 1994.

[52] H. Nijmeijer and A.J. van der Schaft. *Nonlinear Dynamical Control Systems*. Springer-Verlag, 1990.

[53] P.J. Olver. *Applications of Lie groups to differential equations*, volume 107 of *Graduate Texts in Mathematics*. Springer-Verlag, 2nd edition, 1993.

[54] N. Petit, Y. Creff, and P. Rouchon. δ-freeness of a class of linear delayed systems. In *ECC'97*, Bruxelles, 1997.

[55] J.B. Pomet. A differential geometric setting for dynamic equivalence and dynamic linearization. In *Workshop on Geometry in Nonlinear Control, Banach Center Publications, Warsaw*, 1993.

[56] J.B. Pomet. On dynamic feedback linearization of four-dimensional affine control systems with two inputs. *ESAIM-COCV*, 1997. http://www.emath.fr/Maths/Cocv/Articles/articleEng.html.

[57] J.B. Pomet, C. Moog, and E. Aranda. A non-exact Brunovsky form and dynamic feedback linearization. In *Proc. of the 31st IEEE Conf. on Decision and Control*, pages 2012–2017, 1992.

[58] J.F. Pommaret. *Systems of Partial Differential Equations and Lie Pseudogroups.* Gordon & Breach, N.Y., 1978.

[59] J.F. Pommaret. Dualité différentielle et applications. *C.R. Acad. Sci. Paris, Série I*, 320:1225–1230, 1995.

[60] C. Raczy. *Commandes optimales en temps pour les systèmes différentiellement plats.* PhD thesis, Université des Sciences et Technologies de Lille, 1997.

[61] J.P. Ramis. Dévissage Gevrey. *Astérisque*, 59-60:173–204, 1979.

[62] M. Rathinam and R.M. Murray. Configuration flatness of Lagrangian systems underactuated by one control. *SIAM J. Control and Optim.*, 1997. to appear.

[63] A. Reghai. Satellite à deux commandes. Technical report, Ecole Polytechnique, Palaiseau, France, 1995. Mémoire de fin d'études.

[64] R. Rothfuß, J. Rudolph, and M. Zeitz. Flatness based control of a nonlinear chemical reactor model. *Automatica*, 32:1433–1439, 1996.

[65] P. Rouchon. Necessary condition and genericity of dynamic feedback linearization. *J. Math. Systems Estim. Control*, 5(3):345–358, 1995.

[66] P. Rouchon, M. Fliess, J. Lévine, and Ph. Martin. Flatness and motion planning: the car with n-trailers. In *Proc. ECC'93, Groningen*, pages 1518–1522, 1993.

[67] P. Rouchon, M. Fliess, J. Lévine, and Ph. Martin. Flatness, motion planning and trailer systems. In *Proc. of the 32nd IEEE Conf. on Decision and Control*, pages 2700–2705, San Antonio, 1993.

[68] S. Sekhavat. *Planification de Mouvements sans Collision pour Systèmes non Holonomes.* PhD thesis, LAAS-CNRS, Toulouse, 1996.

[69] H. Sira-Ramirez and M. Ilic-Spong. Exact linearzation in switched-mode DC-to-DC power converters. *Int. J. Control*, 50:511–524, 1989.

[70] W.M. Sluis. *Absolute Equivalence and its Application to Control Theory.* PhD thesis, University of Waterloo, Ontario, 1992.

[71] W.M. Sluis. A necessary condition for dynamic feedback linearization. *Systems Control Letters*, 21:277–283, 1993.

[72] M. Soroush and C. Kravaris. Multivariable nonlinear control of a continous polymerization reactor. In *American Control Conferences*, pages 607–614, 1992.

[73] D. Tilbury, O. Sørdalen, L. Bushnell, and S. Sastry. A multisteering trailer system: conversion into chained form using dynamic feedback. *IEEE Trans. Robotics Automation*, 11(6):807, 1995.

[74] D.M. Tilbury. *Exterior differential systems and nonholonomic motion planning.* PhD thesis, University of California, Berkeley, 1994. Memorandum No. UCB/ERL M94/90.

[75] G. Valiron. *Equations Fonctionnelles.* Masson et Cie, Editeurs, Paris, 2nd edition, 1950.

[76] V.V. Zharinov. *Geometrical Aspects of Partial Differential Equations.* World Scientific, Singapore, 1992.

Mass Balance Modelling of Bioprocesses

Olivier Bernard*

COMORE, INRIA, France

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS0280012

*obernard@sophia.inria.fr

Abstract

These lecture notes detail how to design a model for a biological process. The difficulty is due to the fact that, on the contrary to other fields (mechanics, electronics, etc.) there does not exist any validated law to describe the behaviour of biological systems. Nevertheless, these systems satisfy the mass conservation principle. On this basis, the lecture explains how to derive a model which will represent the main mass transfer within the system. The method consists of three steps. First determine the reaction scheme and define a model in which the microbial kinetics are not specified. Then find an analytical expression for the biological kinetics. Finally validate the model trying to test separately the different hypotheses assumed during model design.

Contents

1 Introduction **773**

2 Principle of a bioreactor **774**
2.1 The use of microorganisms 774
2.2 The main types of bioreactors 775
2.3 Working of a bioreactor 776
2.3.1 Presentation 776
2.3.2 Batch mode 776
2.3.3 Fedbatch mode 776
2.3.4 The continuous mode (chemostat) 778
2.3.5 The Sequencing Batch Reactors (SBR) 778

3 The mass balance modelling **778**
3.1 Introduction 778
3.2 Reaction scheme 780
3.3 Choice of the reactions and of the variables 781
3.4 Example 1 781

4 The mass balance models **782**
4.1 Introduction 782
4.2 Example 2 782
4.3 Matrix representation 784
4.3.1 Example 2 (continued) 784
4.3.2 Example 1 (continued) 785
4.4 The gaseous flows 785
4.5 Electro neutrality and affinity constants 786
4.6 Example 1 (continued) 786
4.6.1 Gaseous flows 786
4.6.2 Affinity constants 787
4.6.3 Electro neutrality of the solution 787
4.6.4 Conclusion 787
4.7 Conclusion 788

5 Modelling of the kinetics **789**
5.1 Introduction 789
5.2 The mathematical constraints 789
5.2.1 Positivity of the variables 789

5.2.2 Variables that are necessary for the reaction 790
5.2.3 Example 1 (continued) 790
5.2.4 Phenomenological knowledge 791
5.3 The growth rate 791
5.3.1 The Monod model 791
5.3.2 Haldane model 792
5.3.3 Multiple limitations 792
5.4 Kinetics representation using neural networks 792

6 Model validation **794**
6.1 Introduction 794
6.2 Validation of the reaction scheme 794
6.2.1 Mathematical principle 794
6.2.2 Example 4 795
6.3 Qualitative model validation 797
6.3.1 Example 798
6.4 Global model validation 798
6.4.1 Example 800

7 Mass balance models properties **801**
7.1 Boundness and positivity of the variables 801
7.2 Equilibrium point and local behaviour 804
7.2.1 Introduction 804
7.2.2 Equilibrium points and local stability 804
7.2.3 Global behaviour 805
7.2.4 Asymptotic behaviour 805
7.2.5 Example 4 (continued) 805

8 Conclusion **806**

References **810**

1 Introduction

System modelling in general is difficult and requires time to properly understand the system and identify a model. This exercise is complicated when the system integrates living organisms. On the contrary to domains like physics where laws that are known since centuries (Ohm law, ideal gas relationship, fundamental principle in mechanics, thermodynamic principle, ...) can apply, most of the biological models rely on empirical mathematical expressions. These laws result from *a priori* ideas on the working of the system (metabolism, trophic relationships, etc.) or, in some rare cases, have been estimated from some experiments. Since it is not possible to use laws that are admitted by everybody and that have been extensively validated and used, it is primordial to characterise the reliability of the mathematical expressions used during the model development. This implies that the reliability of the used relationships must be sorted hierarchically during the model development. In this chapter, we will see how to organise the knowledge in the model in order to distinguish a reliable part issued from the mass balance and a more speculative part which will represent the bacterial kinetics.

The model quality and the model structure must above all be determined with respect to the model objectives. Indeed, a model can be developed for very different purposes that must be clearly identified. Will the model be used in order to:

- Reproduce an observed behaviour
- Explain an observed behaviour
- Predict the system evolution
- Understand some of the system mechanisms
- Estimate non measured variables
- Estimate process parameters
- Act on a system to regulate and impose the values for its variables
- Detect anomalies in the process working
- ...

Depending on the modelling objectives and resources, a formalism must be chosen. If the spatial heterogeneity is important and must be taken into account in the model, a parameter distributed model must be written (using e.g. partial differential equations). If the modelling aims at the improvement of a metabolite production during transient phases, the system dynamics must be represented in the model.

Moreover, besides its objectives, the model must also be in adequation with the available data. Indeed a complex model involving a large number of parameters will also require a large amount of data to identify its parameters and to validate the model.

Finally, if we remember that most of the laws used in biology are speculative, the key step in the modelling of bioprocesses is the model validation. This step is often neglected, despite its determinant role to guaranty the model quality. In particular it is crucial to demonstrate that the model reaches properly the goals for which it was developed.

2 Principle of a bioreactor

2.1 The use of microorganisms

The fermentation principle consists in exploiting metabolic reactions that take place in the cell of a micro-organism (bacteria, yeast, phytoplankton, etc.). In order to activate the micro-organisms interesting metabolic pathways, some specific environmental conditions must be applied (temperature, pH, nutrient concentration). The microorganisms generally need nutrients to growth and precursors or activators in order to produce specific molecules. The simplest required reaction is the growth process itself in order to recover the biomass of microorganisms.

In these metabolic reactions, we can distinguish the following biochemical components:

- the substrates S_i, which are necessary for the goal of the fermentation (growth of the microorganisms and/or precursor for the metabolite to be produced). The substrate associated with growth must contain all the elements necessary to sustain growth (*i.e.* N, C, K, P, Fe, ...). In general, these elements are added in excess so that they are never limiting during the cultivation. Only the main nutrients (carbon, nitrogen or phosphorus source) are monitored along the cultivation.

- microbial biomasses (denoted X_i). The microorganisms can be of various type and species (bacteria, phytoplankton, fungi, yeast, etc.);
- the products of the biochemical reactions,(denoted P_i). These products can be in the agro-industrial field (cheese, beer, wine, ...), chemistry (enzymes, colourings...), pharmaceutical industry (antibiotics, hormones, vitamins...) or for energy production (ethanol, biogas...)...
- catalysts: they can neither be produced nor consumed during the reaction, but they are necessary.

Depending on the objectives of the fermentation, specific microorganisms will be grown in order to enhance:

- production of biomass itself. It is for example the case for the production of backer yeast.
- production of a metabolite. The goal is to enhance the cellular synthesis of a particular compounds (ethanol, penicillin, ...).
- substrate uptake. In this case, the substrate degradation itself is the objective. This is more specially used to remove pollutants from a liquid medium. Most of the biological depollution processes are among this category.
- phenomenological studies. In this particular case the fermentation aims a better knowledge of the microorganism. The application can be to better understand how the microorganisms grow in the natural field.

2.2 The main types of bioreactors

There are a great deal of different bioreactors. Depending on the type of microorganisms that are grown, they will need a support to settle or can be free in the liquid. They can resist to more or less intense shearing constraints which will implicate a specific steering system. These two main requirements will determine the type of bioreactor. Two classes can be identified [1]:

- stirred tank reactors (CSTR) in which the medium is homogeneous and each element of volume will represent the concentrations in the whole fermenter

- the bioreactors with non homogeneous concentration along space. In particular the bioreactor for microorganisms using a support to growth (called a "bed") are in this category.

When the medium is homogeneous it can be described by ordinary differential equations. When a strong spatial distribution must be taken into account a model based on partial differential equations are more appropriate. In this lecture we will present only the CSTR modelled with ODE.

2.3 Working of a bioreactor

2.3.1 Presentation

Figure 1 presents a simplified conceptual scheme explaining the principle of a bioreactor. It is mainly a culture vessel of volume V where the microorganisms grow. A pipe feeds the vessel with an influent medium (with flow rate Q_{in}) and another one withdraws the culture medium with a flow rate Q_{out}.

Depending on the way the fermenter is fed and withdrawn, 3 basic working modes can be identified (figure 2).

2.3.2 Batch mode

The system is in batch during the fermentation, and has a constant volume, since no feeding or withdrawal are performed during the fermentation. An inoculum of micro-organisms is introduced at the initial time with all the nutrients and substrates. The biomass or the final product are recovered at the end of the fermentation. The advantage of this approach is that it avoids the contaminations with other bacteria that can come in an open system. The drawback is the limited means of action to act on the fermentation (pH, temperature, aeration...). Therefore the batch mode is often the less optimal from the automatic control point of view to optimise a cost criterion. Nevertheless, this is the most used mode in the industry.

2.3.3 Fedbatch mode

As for the batch mode the duration of a fedbatch is finite. But here the fermenter is fed and starts from a volume V_0 to reach a volume V_f at the end of the fermentation. This mode allows a better control of the growth and

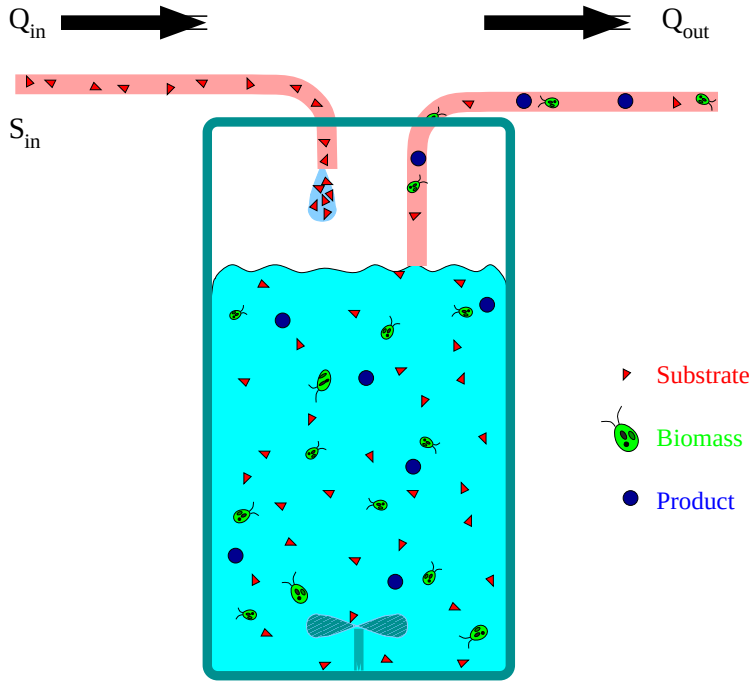


Figure 1: Principle of a bioreactor

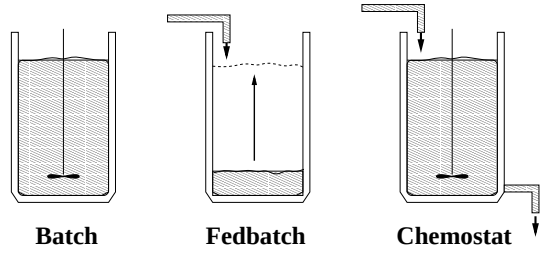


Figure 2: The various working modes of the bioreactors

biotransformation process along the fermentation. The fedbatch processes are often in closed loop. This operating mode is particularly used when the product to be recovered necessitates to empty the bioreactor like e.g. for intracellular components.

2.3.4 The continuous mode (chemostat)

This is the most popular working mode in the field of wastewater treatment. The volume of the bioreactor is constant since the influent flow rate is equal to the effluent flow rate. This mode provides the richest dynamics, and therefore presents the more latitude to optimise the process. It is also often used in laboratories to study the physiology of a microorganism. The advantage is also that it allows important productions in small size reactors.

2.3.5 The Sequencing Batch Reactors (SBR)

It is a combination of the previous working mode. The idea is to recover the biomass before emptying the bioreactor. For this, the agitation is stopped to let the biomass settle. The different steps used for wastewater treatment are presented on Figure 3.

In the same way, the SFBR (sequencing fedbatch reactor) is a SBR with a stage of filling that follows a fedbatch mode.

3 The mass balance modelling

3.1 Introduction

The modelling of biological systems is delicate because it is not based on validated laws, like in other fields (mechanics, electronics, etc.). The evolution of microorganisms is very complex and does not follow any clear law. Nevertheless, this system has to respect some rules, like all the physical systems. For example, the mass conservation, the electro neutrality of the solutions, etc. We will see in this section how to take these aspects into account in the model design. As a result, this mass balance approach will guaranty a certain robustness in the model.

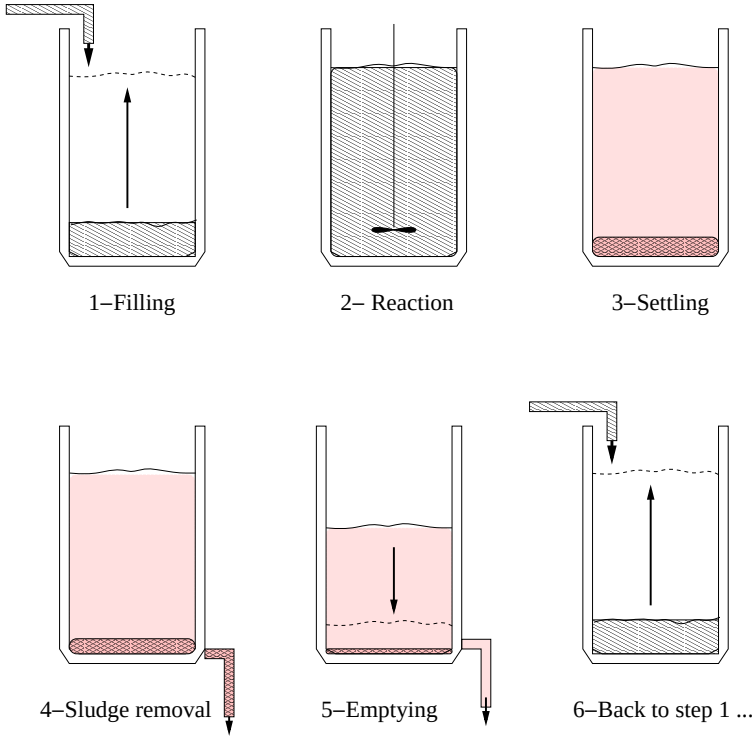


Figure 3: SBR (sequencing batch reactors): representation of the different steps

3.2 Reaction scheme

The reaction scheme of a biochemical process is a macroscopic description of the set of biological and chemical reactions which represents the main mass transfer within the fermenter. A formalism close to this used in chemistry is adopted [2]. A set of substrates S_i are transformed into products P_i following 3 possibilities:

- The reaction is a pure chemical reaction, and no biomass is involved. The reaction is then a classical chemical reaction:

$$S_1 + S_2 + ... + S_p \longrightarrow P_1 + ... + P_q$$

- The reaction is catalysed by a biomass X. The biomass acts only as a catalyser and the reaction is not associated with the growth of the microorganisms:

$$S_1 + S_2 + ... + S_p \xrightarrow{X} P_1 + ... + P_q$$

- The reaction is associated with growth of the microorganisms. Therefore the biomass is also a product of the reaction.

$$S_1 + S_2 + ... + S_p \xrightarrow{X} P_1 + ... + P_q + X$$

The reaction scheme is a concise way to summarise at the macroscopic level a set of reactions that are assumed to determine the process dynamics. The reaction scheme is therefore based on the assumptions related to the available phenomenological knowledge of the process.

In general only the main components of a reaction are represented. Indeed, it would be very difficult to present a real reaction for the growth of a micro-organism since a great deal of components are necessary (Fe, Pb, F, ...).

In the sequel, we will detail the reaction scheme by adding the yield coefficients associated with the consumption (k_i) or the production (k'_i) of each coefficient. Moreover, we will also indicate the rate of the reaction φ:

$$k_1 S_1 + k_2 S_2 + ... + k_p S_p \xrightarrow{\varphi} k'_1 P_1 + ... + k'_q P_q + X$$

The consumption rate of S_i is thus $k_i\varphi$, the production rate P_i is thus $k'_i\varphi$. By convention φ corresponds to the production rate of the biomass.

In the sequel we will assume that the reaction scheme is composed of a set of k biological or chemical reactions. We will considered n variables (chemical concentrations, biomass,...).

3.3 Choice of the reactions and of the variables

The choice of the number of reactions to be taken into account and the choice of the state variables is capital for the modelling purpose. It will be guided by the available knowledge on the reaction scheme on the basis of the available data set. Often the complexity of the model is too high with respect to the amount of data that are available to test and validate the model. It must be chosen with parsimony, keeping in mind the objectives of the model.

The choice of the reactions and of the variables will mainly determine the model structure, it must be considered with care. We will see in section 6 how to validate this reaction scheme.

We briefly present in Appendix A a procedure to determine the number of reactions that must be taken into account with respect to the available data.

In the sequel, we will assume that the reaction scheme:

- represents the main mass and flow repartition between the set of reactions that intervene in the process,
- is a set of reactions whose yield coefficients are constant.

3.4 Example 1

We will consider here the example of anaerobic digestion. This process is used to remove a polluting substrate (S_1) from wastewater thanks to anaerobic bacteria. In fact, this is a very complex process which involves several different bacterial populations [3]. If the modelling objective is to control this intricate ecosystem in order to improve the pollution removal, then we need a rather simple model. This is why, to limit the model complexity, we consider only two main bacterial populations. We assume therefore that the dynamics can be described by two main steps:

- An acidogenesis step (with a rate $r_1(.)$) in which the substrate S_1 is degraded by acidogenic bacteria (X_1) and is transformed into volatile fatty acids (VFA) (S_2) and CO_2:

$$k_1\, S_1 \xrightarrow{r_1(.)} X_1 + k_2\, S_2 + k_4\, CO_2 \tag{1}$$

- A methanogenesis step (with a rate $r_2(.)$), where the volatile fatty acids are degraded into CH_4 and CO_2 by methanogenic bacteria (X_2).

$$k_3\, S_2 \xrightarrow{r_2(.)} X_2\, , + k_5\, CO_2 + k_6\, CH_4 \tag{2}$$

The constants k_1, k_2, k_4, respectively represent the stoichiometric coefficients associated with substrate S_1 consumption, production of VFA and CO_2 during acidogenesis. k_3, k_5 and k_6 respectively represent stoichiometric coefficients associated with VFA consumption and with CO_2 and CH4 production during methanogenesis.

It is worth noting that in some sense this reaction scheme has no biological reality since biomasses X_1 and X_2 represent a set of different species. In the same way for substrates S_1 and S_2 which gathers a set of heterogeneous compounds. A lot of models can be found in the literature for this process [4, 3, 5]. Generally, the description of the processes within the bioreactor are much more detailed [6, 7] but it leads to models difficult to use for control purpose.

4 The mass balance models

4.1 Introduction

We will consider a continuously stirred tank reactor that guarantees a perfect mixing. We will see that independently of the working mode (batch, fedbatch, continuous), the dynamical behaviour of the biological or chemical compounds in the reactors can be directly deduced from the reaction scheme.

We will show on a very simple example how the dynamical model can be established.

4.2 Example 2

We will consider here the very simple example of the growth of a microorganism X on a substrate S with rate $r(.)$:

$$k\, S \xrightarrow{r(.)} X$$

The yield coefficient associated with substrate consumption is denoted k.

We assume that the influent flow rate is Q_{in} and that the effluent flow rate is Q_{out}. We denote by x and s the total amount of biomass and substrate in the volume V of the bioreactor.

Let us consider the evolution of $V(t)$, $x(t)$ and $s(t)$ between two very close time instants t and $t + dt$.

The evolution of the total liquid volume V is rather simple:

$$V(t + dt) = V(t) + Q_{in}dt - Q_{out}dt$$

For the biomass, we have to take into account the new biomass produced between t and $t+dt$. The production term in the whole volume V is $r(.)Vdt$, and thus:

$$x(t + dt) = x(t) + r(.)Vdt - Q_{out}dt\frac{x}{V}$$

Note that, in order to compute the biomass lost in the effluent (in the volume $Q_{out}dt$) we assume that the concentration in the small volume is the same as in the whole bioreactor (*i.e.* $\frac{x}{V}$). At this point the hypothesis of homogeneity in the reactor is crucial.

In the same way, for the substrate, we must also consider the quantity of substrate (with concentration S_{in}) arriving between the two time instant:

$$s(t + dt) = s(t) + Q_{in}S_{in} - kr(.)Vdt - Q_{out}dt\frac{s}{V}$$

For a very small dt, we can then derive the following equations:

$$\begin{cases} \dfrac{dx}{dt} = r(.)V - Q_{out}\dfrac{x}{V} & (3) \\ \dfrac{ds}{dt} = -kr(.)V + Q_{in}S_{in} - Q_{out}\dfrac{s}{V} & (4) \\ \dfrac{dV}{dt} = Q_{in} - Q_{out} & (5) \end{cases}$$

Now, let us rewrite this model in term of concentration *i.e.* using the variables $X = \frac{x}{V}$ and $S = \frac{s}{V}$). It is straightforward to see that we get the following model:

$$\begin{array}{l} \frac{dX}{dt} = r(.) - DX \\ \frac{dS}{dt} = -kr(.) + D(S_{in} - S) \\ \frac{dV}{dt} = Q_{in} - Q_{out} \end{array} \quad (6)$$

where $D = \frac{Q_{in}}{V}$ corresponds to the dilution rate.

Model (6) simplifies for the various working modes:

- **Batch**. In this case we have $Q_{in} = Q_{out} = 0$. The volume is then constant.
- **Fed batch**. Here $Q_{out} = 0$; $\frac{dV}{dt} = Q_{in}$, V is increasing.
- **Continuous mode**. The volume V is constant since $Q_{in} = Q_{out}$.

For sake of simplicity, in the sequel we will not describe the fed batch case and we will concentrate on the batch or continuous mode. This simplifies the equation since we do not need the equation which forecasts the volume evolution.

4.3 Matrix representation

The reaction scheme leads to the following mass balance model which describes **equivalently** the mass flows within the bioreactor [2]:

$$\dot{\xi} = Kr(.) + D(\xi_{in} - \xi) - Q(\xi) \tag{7}$$

Where ξ is the state vector containing all the process compounds and biomasses, ξ_{in} is the vector of the influent concentrations, $r(.)$ is a vector of reaction rates. The matrix K contains the stoichiometric coefficients (yields). $Q(\xi)$, represents the gaseous terms of exchange between the liquid and the gas phase. The dilution rate, D, is the ratio between the influent flow rate Q_{in} and the reactor volume V.

Remark 1 *In the case of the fed batch process, the state vector must also contain the volume V of the reactor. The last equation will describe the volume evolution (cf. equation (5)).*

4.3.1 Example 2 (continued)

Let us consider model (6) working in continuous mode (V is constant, $D = \frac{Q_{in}}{V}$). The model can be rewritten as follows:

$$\begin{pmatrix} X \\ S \end{pmatrix} = \begin{pmatrix} 1 \\ -k \end{pmatrix} r(.) + D\left(\begin{pmatrix} 0 \\ S_{in} \end{pmatrix} - \begin{pmatrix} X \\ S \end{pmatrix}\right)$$

It corresponds exactly to the general model, (7) with:

$$\xi = \begin{pmatrix} X \\ S \end{pmatrix}, \quad K = \begin{pmatrix} 1 \\ -k \end{pmatrix}, \quad \xi_{in} = \begin{pmatrix} 0 \\ S_{in} \end{pmatrix}, \quad Q(\xi) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

4.3.2 Example 1 (continued)

Now let us come back to the anaerobic digestion example (see section 3.4). We will assume that the methane solubility is very low and therefore that it directly goes into the gas phase. The carbon dioxide is stored in the liquid phase where he enters in the inorganic carbon compartment (C).

The mass balance model is then the following:

$$\frac{dX_1}{dt} = r_1(.) - DX_1 \tag{8}$$

$$\frac{dX_2}{dt} = r_2(.) - DX_2 \tag{9}$$

$$\frac{dS_1}{dt} = D(S_{1in} - S_1) - k_1 r_1(.) \tag{10}$$

$$\frac{dS_2}{dt} = D(S_{2in} - S_2) + k_2 r_1(.) - k_3 r_2(.) \tag{11}$$

$$\frac{dC}{dt} = D(C_{in} - C) - q_C(\xi) + k_4 r_1(.) + k_5 r_2(.) \tag{12}$$

where S_{1in}, S_{2in} and C_{in} are respectively the influent concentrations of substrate, VFA and dissolved inorganic carbon. The term $q_C(\xi)$ represents the inorganic carbon flow rate (of CO_2) from the liquid phase to the gaseous phase.

4.4 The gaseous flows

In order to derive the mass balance, we must take into account the compounds which have a gaseous phase. Indeed, the gaseous species can escape the bioreactor after going from the liquid to the gaseous phase (they can also enter into the bioreactor).

We use for this Henry's law which describes the molar flow rate of a compound C from its liquid phase to its gaseous phase:

$$q_c = K_L a(C - C^\star) \tag{13}$$

Remark 2 *If $q_c < 0$, it means that the gaseous flow will take place from the gaseous phase to the liquid phase.*

The transfer coefficient $K_L a$ (1/T) highly depends on the operating conditions and especially from stirring, and the exchange area between the liquid and the gaseous phases (size of the bubbles)[8, 1]. The modelling of this parameter with respect to the operating conditions can be very delicate.

The quantity $C^\star$ is the saturation concentration of dissolved C. This quantity is related to the partial pressure of gaseous C (P_C) thanks to Henry's constant:

$$C^\star = K_H P_C \tag{14}$$

Henry's constant can also vary with respect to the compounds in the culture medium or the temperature.

Moreover, when several gaseous species are simultaneously in the gaseous phase, they must follow the ideal gas law. This will give a relationship of constant ratio between molar flow rates and partial pressures. For m gaseous species $C_1 \dots C_m$:

$$\frac{P_{c1}}{q_{c1}} = \frac{P_{c2}}{q_{c2}} = \dots = \frac{P_{cm}}{q_{cm}} \tag{15}$$

4.5 Electro neutrality and affinity constants

The electro neutrality of the solutions is a second rule that the biological systems must respect: the anions concentrations weighted by the number of electrical charges must equal the concentration of cations with the same weighting.

The chemical reactions are often well known and an affinity constant is generally associated. This constant is generally related to the protons concentration H^+, and therefore to pH.

4.6 Example 1 (continued)

4.6.1 Gaseous flows

The methane flow rate is directly related to methanogenesis:

$$q_M = k_6\, r_2(.) \tag{16}$$

The gaseous CO_2 flow rate follows Henry's law:

$$q_C(\xi) = K_L a(CO_2 - K_H P_C) \tag{17}$$

where P_C is the CO_2 partial pressure.

4.6.2 Affinity constants

In the anaerobic digestion example, we will use the electro neutrality and the chemical affinity constants:

In the usual operating range of pH for these processes ($6 \leq pH \leq 8$) we assume that the VFA are under their ionised form. The dissolved CO_2 is in equilibrium with bicarbonate:

$$CO_2 + H_2O \leftrightarrow HCO_3^- + H^+$$

The affinity constant of this reaction is then

$$K_b = \frac{HCO_3^- \; H^+}{CO_2} \tag{18}$$

4.6.3 Electro neutrality of the solution

The cations (Z), are mainly ions which are not affected by biochemical reactions (Na^+,...). Therefore, their dynamics will simply follow, without modification the cation concentration Z_{in} in the influent, so that:

$$\frac{dZ}{dt} = D(Z_{in} - Z) \tag{19}$$

The anions are mainly represented by the VFA and the bicarbonate. Electro neutrality ensures then that:

$$Z = S_2 + HCO_3^- \tag{20}$$

4.6.4 Conclusion

If we add equation (19), the model can finally be rewritten under the matrix form (7), with :

$$\xi = \begin{bmatrix} X_1 \\ X_2 \\ Z \\ S_1 \\ S_2 \\ C \end{bmatrix}, \; r(.) = \begin{bmatrix} r_1(.) \\ r_2(.) \end{bmatrix}, \; K = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ -k_1 & 0 \\ k_2 & -k_3 \\ k_4 & k_5 \end{bmatrix} \tag{21}$$

$$\xi_{in} = \begin{bmatrix} 0 \\ 0 \\ Z_{in} \\ S_{1in} \\ S_{2in} \\ C_{in} \end{bmatrix}, \; Q = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ q_C(\xi) \end{bmatrix}, \tag{22}$$

An elimination of variables HCO_3^-, CO_2, and P_C using equations (17),(15) (18) and (20), leads to the following expression for $P_C(\xi)$ (cf [9]):

$$P_C(\xi) = \frac{\phi - \sqrt{\phi^2 - 4K_H\, P_T\, (C + S_2 - Z)}}{2K_H} \tag{23}$$

setting: $\phi = C + S_2 - Z + K_H\, P_T + \dfrac{k_6}{k_L a} r_2(.)$, we finally get

$$q_C(\xi) = k_L a(C + S_2 - Z - K_H P_C(\xi)) \tag{24}$$

4.7 Conclusion

At this stage, we end up with a model based on the following physical and chemical principles:

- Mass balance
- Ionic balance
- Affinity constants
- Ideal gas law
- Henry's law

The more important hypothesis (with respect to model reliability) is the mass balance hypothesis deduced from the reaction scheme. This hypothesis will therefore require to be validated in the sequel of the modelling approach.

The mass balance model can be used in this form for monitoring or control purpose. Indeed, using the approaches developed in the framework of systems with unknown inputs [10, 11, 12], the unknown reaction rates can be removed thanks to adequate state transformations [2].

Nevertheless, if the initial objective consists in simulating the system, then the reaction rates $r_i(.)$ must be written with respect to the state variables and to the system inputs (environmental variables). This step is much more delicate and a lot of hypotheses difficult to verify are requested.

5 Modelling of the kinetics

5.1 Introduction

For some specific purposes (optimal control, simulation, predictions, etc.) it is necessary to have an analytical expression relating the reaction rates to the state variables of the system. We have nevertheless to keep in mind that these expressions are most of the time approximate relationships issued from empirical considerations. Therefore we leave the background of physical modelling presented previously.

In this section we will see how to establish a hierarchy between the assumed hypotheses in order to obtain a two reliability level description of the kinetics.

5.2 The mathematical constraints

5.2.1 Positivity of the variables

A priori, some physical constraints that the model must respect are known: The variables must remain positive and they must be bounded if the amount of matter entering in the bioreactor is bounded. These physical constraints will impose constraints on the structure of the $r_i(.)$. Some quantities (percentage, ratios, etc.) must remain between known bounds. To guaranty that the model respects this property, it should verify the following property:

Property 1 (H1) *For each state variable* $\xi_i \in [L_{i\,min}, L_{i\,max}]$, *the field* $\dot{\xi}_i$ *on the boundaries must be directed in the admissible space. In other words, the following conditions must be satisfied:*

$$\xi_i = L_{i\,min} \Rightarrow \dot{\xi}_i \geq 0$$

$$\xi_i = L_{i\,max} \Rightarrow \dot{\xi}_i \leq 0$$

Particular case: We must have $\xi_i = 0 \Rightarrow \dot{\xi}_i \geq 0.$ in order that variable ξ_i remains positive.

5.2.2 Variables that are necessary for the reaction

The second important constraint which must be satisfied by the biochemical kinetics is related to the reaction scheme. A reaction can not take place if one of the reactant necessary for the reaction is missing. This justifies the following property:

Property 2 *If ξ_j is a reactant of reaction i, then ξ_j can be factorised in r_i:*

$$r_i(\xi, u) = \xi_j \, \nu_{ij}(\xi, u)$$

We verify then easily that $\xi_j = 0 \Rightarrow r_i(\xi, u) = 0$

In the same way, for the reactions associated to a biomass X, we have the same property. Therefore a growth reaction can be rewritten

$$r_i(\xi, u) = \mu_i(\xi, u)X$$

The term $\mu_i(\xi, u)$ is called the growth rate.

5.2.3 Example 1 (continued)

Let us consider the anaerobic digestion model given by equations (8) to (11) and let us apply the state positivity principle:

$$X_1 = 0 \Rightarrow r_1(.) \geq 0 \tag{25}$$

$$X_2 = 0 \Rightarrow r_2(.) \geq 0 \tag{26}$$

$$S_1 = 0 \Rightarrow D(S_{1in} - S_1) - k_1 r_1(.) \geq 0 \tag{27}$$

$$S_2 = 0 \Rightarrow D(S_{2in} - S_2) + k_2 r_1(.) - k_3 r_2(.) \geq 0 \tag{28}$$

Equations (25) and (26) are not very informative. In order that (27) and (28) are respected whatever the experimental conditions, it requires:

$$r_1(.) = S_1 \phi_1(.) \text{ and } r_2(.) = S_2 \phi_2(.)$$

Moreover, biomasses X_1 and X_2 are necessary, respectively for reactions 1 and 2, and thus:

$$r_1(.) = \mu_1(.)X_1 \text{ and } r_2(.) = \mu_2(.)X_2$$

Finally, we must have:

$$r_1(.) = S_1 X_1 \nu_1(.) \tag{29}$$

$$r_2(.) = S_2 X_1 \nu_2(.) \tag{30}$$

5.2.4 Phenomenological knowledge

We will exploit the available phenomenological knowledge (even if it is often speculative) in order to propose an expression for the reaction kinetics.

First, the laboratory experiments allows one to determine the variables which act on the reaction rates. We have seen that the reactant and sometimes the biomass must be found among these variables.

Then, we must know whether the reaction is activated or inhibited by these variables. It often happens that a variable is activating and that she becomes inhibiting at high concentrations (toxicity effect).

Now, there remains to propose an analytical expression which will take into account the mathematical constraints so as the phenomenological knowledge on the process. For this, the modelling choices rely on one hand on experimental observations (when they exist!) and on the other hand on the available models in the literature. In all the cases, the parsimony principle will be privileged to guaranty that the models can be identified and validated.

The following paragraph details the list of models that are often found in the literature to describe some typical reactions. These examples are indicative and a very large number of different models can be found in the literature, in particular to describe the growth rate [2, 1].

5.3 The growth rate

5.3.1 The Monod model

The most commonly used model is the Monod [13] model which uses the kinetics identified by Michaëlis-Menten for enzymatic kinetics :

$$\mu(S) = \mu_{max} \frac{S}{K_s + S} \tag{31}$$

μ_{max} is the maximal growth rate and K_s the half saturation constant.

This simple model summarises the two main phases of the growth of a microorganism:

- Unlimited growth, for high values of substrate ($S >> K_S$). The growth rate is then constant, equal to the maximal growth rate μ_{max}
- The limited growth, for small values of substrate. In this case the growth rate is approximately proportional to the substrate.

Note that the similitude between enzymatic reaction and growth of a microorganism are often used to justify the analytical expression of a reaction rate [14, 15].

5.3.2 Haldane model

The Haldane model, initially proposed for an enzymatic reaction can be used to represent a substrate inhibiting the growth at high values [16]:

$$\mu = \mu_{max} \frac{S}{K_s + S + \frac{S^2}{K_i}} \tag{32}$$

where K_i is an inhibition constant. This model predicts that the growth rate is inversely proportional to the growth rate at high concentrations.

5.3.3 Multiple limitations

When two substrates S_1 and S_2 are simultaneously limiting the growth, a usual way of modelling the reaction rates is to take the product of two Michaelis-Menten kinetics:

$$\mu = \mu_{max} \left(\frac{S_1}{K_{S_1} + S_1} \right) \left(\frac{S_2}{K_{S_2} + S_2} \right) \tag{33}$$

where K_{S_1} and K_{S_2} are the half saturation constants associated respectively to substrates S_1 and S_2.

If one of the substrate (say S_1) is at high concentration, the growth rate is then equivalent to a Monod model with respect to the other substrate (i.e. S_2).

5.4 Kinetics representation using neural networks

We expose briefly here an alternative method to represent the kinetics using a neural network. The global model will then be composed of a mass balance model based on O.D.E, and of a neural network for the reaction rates. In this sense it is an hybrid model. No *a priori* hypotheses are performed on the kinetics, except that we take into account some constraints to guaranty that the system trajectory keep an acceptable meaning. The kinetics represented by the neural network are then directly identified along the training step.

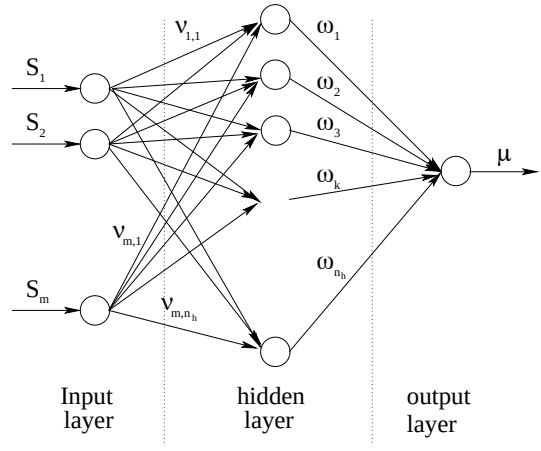


Figure 4: Scheme of a neural network including a single hidden layer

Nevertheless, the variables which influence the kinetics must be determined. These variables will constitute the input of the neural network.

A schematic view of the network is presented on Figure 4 for a single hidden layer. The expression of the output of the network with respect to the inputs is as follows:

$$\mu(S_1, \ldots, S_m) = \sum_{k=1}^{n_h} \omega_k \, \phi(\sum_{i=1}^{m} \upsilon_{ki} S_i) \tag{34}$$

where n_h represents the number of neurons in the hidden layer. The ω_k and υ_{ki} are respectively the weights of the input and outputs layers. Function ϕ is the activating function of the neuron. It is generally chosen among a set of functions (sigmoides, hyperbolic tangent, gaussian, etc.).

The choice of the type of network and of the number of neurons is a rather classical choice and we invite the reader to refer to [17] for more details.

Once the structure of the network has been chosen, the next step is the training phase consisting in identifying the networks weights. This operation is a bit specific for hybrid systems and we refer to [18, 19] for more explanations.

6 Model validation

6.1 Introduction

The last modelling step is certainly the most important, but it is also the most often neglected one. It is all the more important since we have seen that it was necessary to assume a great amount of speculative hypotheses. Before using a model, it is important to validate it properly. This stage follows generally the identification step which is not described here.

The general objective for the validation is to verify that the model fits the objectives that have been fixed. More precisely, we will see how to test **separately** the various hypotheses that have been assumed during the model development:

- the reaction scheme
- the qualitative model predictions
- the model as a whole (reaction scheme+kinetics+parameters)

It is important to note that the validation phase must be performed from a data set which was not used to establish or to identify the model. Moreover the new experiments that must be used to test the model validity must significatively differ from the previously used data set (otherwise it is a test of the experimental reproducibility rather than a test of the model validity). If these conditions are not respected, the model can not pretend to be validated

6.2 Validation of the reaction scheme

6.2.1 Mathematical principle

The proposed procedure relies on an important property, which is a consequence of the mass conservation within the bioreactor. As a result this approach will allow us to check if the obtained mass balance is consistent with the data.

Property 3 *We assume that the $n \times k$ matrix K has more rows than columns ($n > k$). This means that there are more variables than reactions. In this conditions, we have at least $n - k$ independent vectors $v_i \in \mathbf{R}^n$ such that:*

$$v_i^t K = 0_{1 \times k}$$

By convention, we normalise the first component of the vector v_i in order to have $v_{i\,1} = 1$

Consequence : let us consider the real variable $w_i = v_i^t \xi$, this variable satisfies the following equation:

$$\frac{dw_i}{dt} = D(w_{i\,in} - w_i) - v_i^t Q(\xi) \tag{35}$$

with $w_{i\,in} = v_i^t \xi_{in}$. Let us integrate (35) between two time instants t_1 and t_2. We rewrite this equation in order to let the components v_{ij} of vector v_i appear. It leads to:

$$\sum_{j=2}^{n} v_{ij} \phi_{\xi_j}(t_1, t_2) = \phi_{\xi_1}(t_1, t_2) \tag{36}$$

where

$$\phi_{\xi_j}(t_1, t_2) = \xi_j(t_2) - \xi_j(t_1) - \int_{t_1}^{t_2} D(\tau)(\xi_{j\,in}(\tau) - \xi_j(\tau)) - Q_j(\xi(\tau))d\tau$$

The terms $\phi_{\xi_j}(t_1, t_2)$ can be estimated from the experimental measurements of ξ_j along time. An approximation of the integral can be computed *e.g.* using a trapeze formulae. Moreover if the sampling frequency is not sufficient, the data will probably require to be interpolated. We recommend for this task to use spline functions which will at the same time smooth and interpolate the data.

The relationship (36) is a linear relation linking the v_{ij} to the terms $\phi_{\xi_j}(t_1, t_2)$. Since the $\phi_{\xi_j}(t_1, t_2)$ can be computed between various time instants t_1 and t_2, (36) is a linear regression whose validity can be experimentally tested.

Important remark: In fact, relationship (36) is a linear regression which will provide us with an estimate of the v_{ij}. These terms are related with the coefficients of the yield matrix K, and will in general allow to estimate the value of these coefficients.

6.2.2 Example 4

Let us consider here the simple example of the growth of the filamentous fungi *Pycnoporus cinnabarinus* (X) on two substrates, glucose (carbon (C) source) and ammonium (nitrogen (N) source). We assume therefore that the reaction scheme is composed by a single reaction:

$$N + C \longrightarrow X$$

The stoichiometric matrix K associated to this reaction is the following ($\xi = (X\ N\ C)^t$):

$$K = (1\ -k_1\ -k_2)^t, \quad \text{and } \xi_{in} = (0\ N_{in}\ C_{in})^t \tag{37}$$

Let us consider the two following vectors orthogonal to the columns of K:

$$v_1 = (1\ \frac{1}{k_1}\ 0)^t \text{ and } v_2 = (1\ 0\ \frac{1}{k_2})^t$$

We can then define the following quantities:

$$\phi_X(t_1, t_2) = X(t_2) - X(t_1) + \int_{t_1}^{t_2} D(\tau)X(\tau)$$

$$\phi_N(t_1, t_2) = N(t_2) - N(t_1) - \int_{t_1}^{t_2} D(\tau)(N_{in}(\tau) - N(\tau))d\tau$$

$$\phi_C(t_1, t_2) = C(t_2) - C(t_1) - \int_{t_1}^{t_2} D(\tau)(C_{in}(\tau) - C(\tau))d\tau$$

which will allow us to rewrite the following regressions associated with v_1 and v_2:

$$\phi_X(t_1, t_2) = \frac{1}{k_1}\phi_N(t_1, t_2) \tag{38}$$

$$\phi_X(t_1, t_2) = \frac{1}{k_2}\phi_C(t_1, t_2) \tag{39}$$

It is now easy to verify if the relationships (38) and (39) are significative from a statistical point of view.

Figure (5) presents a validation example on the basis of a series of experiment. The obtained regression is highly significative. This means that relations (38) and (39) are valid. As a consequence, the rows of matrix K, which are orthogonal to v_1 and v_2 are necessarily of the type $K = (1\ -\alpha_1\ -\alpha_2)^t$. Therefore the reaction scheme is valid, and subsequently the mass balance model as well.

Note that these techniques lead also to the estimate of the yield coefficients k_1 and k_2.

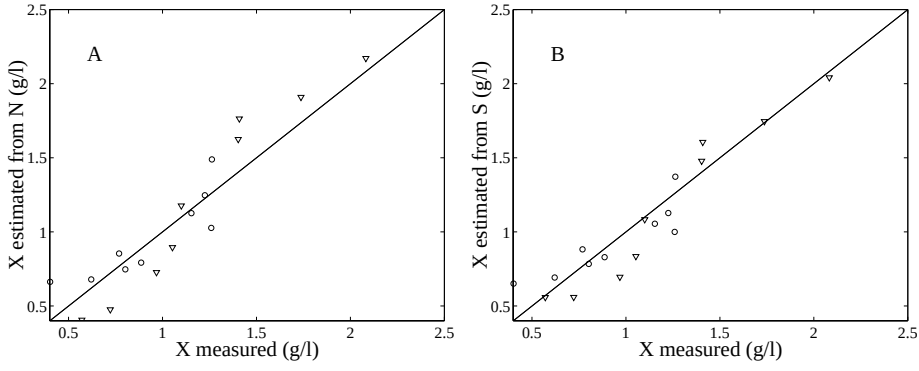


Figure 5: Validation of the linear relationship relating ϕ_X and ϕ_N (A); ϕ_X and ϕ_C (B)

6.3 Qualitative model validation

For the third stage, we assume that the reaction scheme, and therefore the mass balance model has been validated. We will then consider a simulation model consisting of the mass balance model plus the mathematical expression of the kinetics.

The first think to do is to test whether the qualitative properties of the model respect the experimental observations.

The first qualitative behaviour that we expect the model to reproduce is the asymptotic behaviour obtained for constant inputs. Will the model predict an equilibrium, or a more complex behaviour (limit cycle, chaos,...) in agreement with experiments ?

How do these properties evolve when the inputs vary ? For example, the model will predict that an equilibrium in a bioreactor is globally stable for values of the dilution rate lower than a bound, and that for higher values the equilibrium becomes unstable. Does it correspond to the experimental observations ?

More precise qualitative property on the type of transient allowed by the model can also be compared with experimental data. For some specific systems, these transients can be rather precisely determined from a structure analysis [20, 21, 22, 23].

Another qualitative criterion that can be discussed is the response of the system at steady state to a change in an input. Assume for example that an increase of input u_i (which is then kept constant) leads to a decrease in the

steady state value of ξ_j: is it verified from an experimental point of view ?

6.3.1 Example

For example, Hansen and Hubbell (1980) study the competition between two bacterial species in a chemostat. The reaction scheme is composed of two growth reactions:

$$k_1\, S \longrightarrow X_1$$

$$k_2\, S \longrightarrow X_2$$

The growth rate associated to these reactions is assumed to be of Monod type, *i.e.*:

$$\mu_i(S) = \mu_{max\, i} \frac{S}{S + K_{s\, i}}$$

where $\mu_{max\, i}$ and $K_{s\, i}$ are the maximum growth rate and the half saturation constant associated with substrate S for species i.

Hansen and Hubbell showed that the winner of the competition predicted by the model depends on the dilution rate. More precisely, the winner is the species with the smaller ratio $J_i = \frac{K_{s\, i}}{\mu_{max\, i} - D}$. The comparison of the 2 ratios J_1 and J_2 leads to the study of the quantity $r = \frac{\mu_{max\, 1} - \mu_{max\, 2}}{\mu_{max\, 2} - D}$ with respect to the threshold value $\frac{K_{s\, 1}}{K_{s\, 2}} - 1$. If we assume that we are in the case where $D < \mu_{max\, 1} < \mu_{max\, 2}$, then species 2 wins for a dilution rate lower than $D_0 = \frac{\mu_{max\, 2} K_{s\, 1} - \mu_{max\, 1} K_{s\, 2}}{K_{s\, 1} - K_{s\, 2}}$, whereas for higher values, it is species 1 (see figure 6). These qualitative properties are verified experimentally (see Figure 7).

6.4 Global model validation

This is the classical way of validating a model: the simulation results are quantitatively compared to experimental data. The most popular criterion is the least square criterion which is computed as follows for a data set of N measurements:

$$J = \sum_{i}^{N} |\hat{\xi}(t_i) - \xi(t_i)|^2$$

where $\hat{\xi}(ti)$ is the simulated value of the state ξ at the sampling instant t_i. The criterion can be improved by weighting each component of the state ξ_j by a coefficient which takes into account the mean value of ξ_j and the measurement accuracy for this variable.

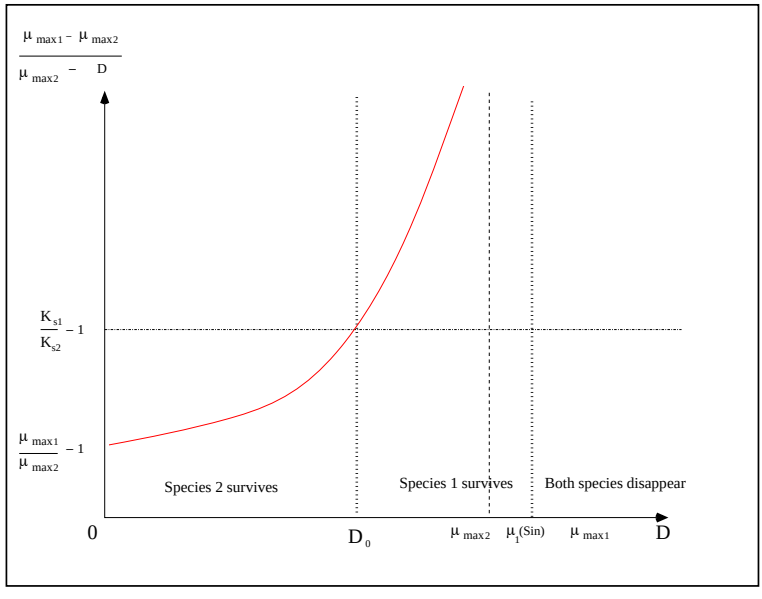


Figure 6: Competition in a chemostat with respect to the dilution rate (discussion of the quantity $\frac{\mu_{max\,1}-\mu_{max\,2}}{\mu_{max\,2}-D}$ with respect to $\frac{k_{s\,1}}{k_{s\,2}}-1$). We consider here the case where $D < \mu_{max\,2} < \mu_{max\,1}$

This criterion should be minimum. In theory, the residuals (*i.e.* $\hat{\xi}-\xi$) must be studied from a statistical point of view. In the ideal case, it should have properties comparable to those of the measurement noise: it should at least be zero on average, and more precisely one can expect a gaussian distribution [25].

In this approach, the model is considered as a whole. If the residual analysis is not good, in the case where the previous validation steps (reaction scheme and qualitative criteria) have not been performed properly it would be impossible to know the cause of the problem. This criterion does not give any clue on the structural validity of the model (underlying reaction scheme, qualitative properties), on the validity of the type of reaction rate modelling used or on the correctness of parameter values.

If the two first validation steps have been successfully fulfilled, the problem is probably due to a an erratic parameter estimation.

In practice, in the framework of biotechnological systems, as it is difficult to validate *stricto sensu* these models, one will be satisfied with a good

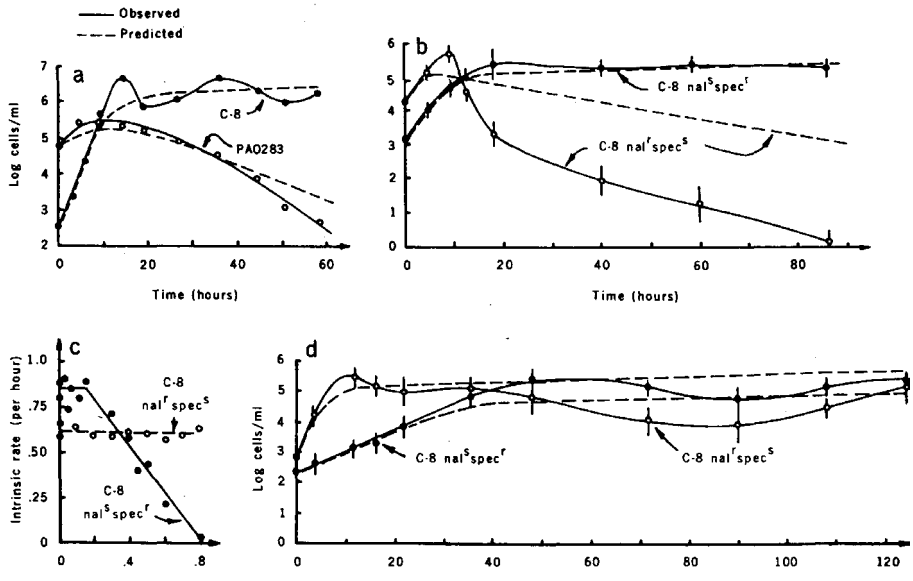


Figure 7: Experimental validation of the qualitative model behaviour. Quantitative model predictions are represented as well. The qualitative model predictions are verified for: a) Two species (*Escherichia coli*, strain C-8 and *Pseudomonas aeruginosa*, strain PA0283 which differ from their half-saturation constants. b) Two strains of *Escherichia coli* which differ from their maximal growth rates. d) Coexistence obtained with 2 strains of *Escherichia coli* which have the same parameter J_i. Figure c) represents the effect of nalidixic acid on the maximal growth rate for the 2 considered strains C-8. (from [24])

visual adequation between simulations and data. This subjective criterion can be reinforced by an analysis of the correlation between predictions and measurements.

6.4.1 Example

The following validation example presents the results obtained with the anaerobic digestion model exposed throughout the paper. Figures 8 and 9 present model simulations compared to direct measurements [9]. The periods of time considered for the calibration step are shown on the figures.

The model correctly reproduces the behaviour of the system for the considered period in spite of the fact that it has been calibrated only using

steady state measurements.

Indeed Figure 8 shows that the continuously measured variables (*i.e.* gaseous flow rate and pH) are well predicted. It is worth noting that these simulations also correctly reproduce the effect of the disturbances induced by pump failures (around day 45). Remark also that the pH predictions match quite well the direct measurements although pH measurements have not been used to calibrate the model parameters. However the model predicts a more severe pH drop during the destabilisation phase (days 21-25). This may be due to an underestimation of the buffer capacity (*i.e.* the alkalinity of the system). It can be noticed that during the destabilisation period the gases are underestimated by the model.

The model simulations are also in good agreement with the off-line data (Figure 9). Even if S_1 is a variable that stands for the various components of the COD that can be rather different along the experiment, the adequacy between model and measurements is good. The reaction of the model to the overloading produced on day 68 seems to be slower than the process, so that the accumulation starts less rapidly in the model.

The main quality of the model is its ability to predict the destabilisation of the plant. This was not obvious since only equilibrium data have been used for the model calibration and the data obtained during the destabilisation phases were not used. The quality of the model justifies its integration in an on-line monitoring procedure in order to early detect a possible destabilisation [26]. The model is also used to derive a robust control algorithm, that is insensitive to the main modelling uncertainties and that avoid the plant destabilisation [27].

7 Mass balance models properties

7.1 Boundness and positivity of the variables

We have seen in paragraph 5.2.1 that the models must be designed in order to meet constraints like the positivity of the state variables.

We will see here that the models based on mass balances are of the type BIBS (bounded input bounded state). To show this property, we use the following hypotheses which are verified for the mass balance based systems:

Hypothesis 1 (H2) *There exists a vector* v^+ *whose components are strictly*

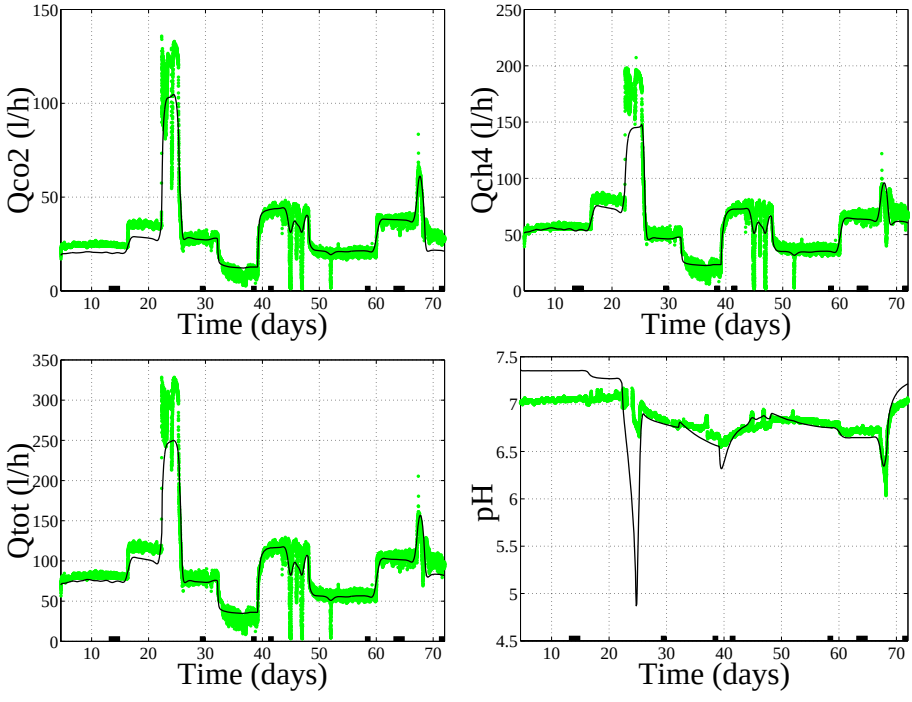


Figure 8: Comparison between simulation results and measurements for the gaseous flow rates and the pH. The periods considered for the calibration step are represented on the time axis

positive, such that:

$$v^+ K = 0_{1\times k}$$

Consequence: Let us consider the scalar quantity $w^+ = v^+\xi$. It verifies the following equation: (35):

$$\frac{dw^+}{dt} = D(w_{in}^+ - w^+) - v^+ Q(\xi) \tag{40}$$

We have to assume an hypothesis for $Q(\xi)$, which is verified in most of the cases:

Hypothesis 2 (H3) *There exists a positive real a and a real b, such that $Q(\xi)$ can be compared to a linear expression as follows:*

$$v^+ Q(\xi) \geq a v^+ \xi + b$$

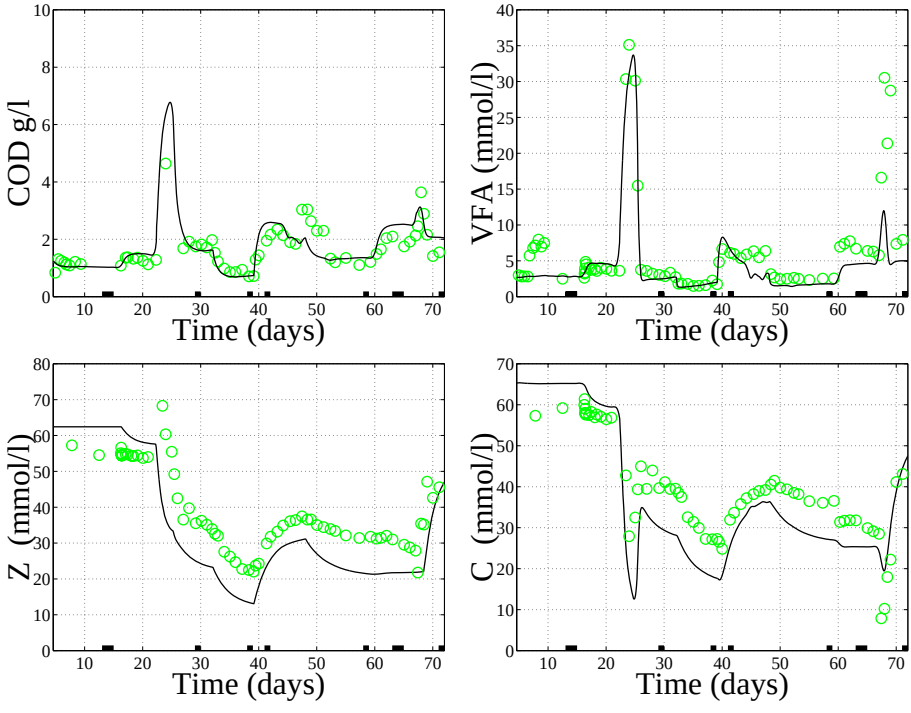


Figure 9: Comparison between simulation results and measurements for COD, VFA, alkalinity and total inorganic carbon. The periods considered for the calibration step are underlined

This hypothesis is verified if $v^+ Q(\xi) = 0$, or if $Q(\xi)$ is described by Henry's law (see section 4.4).

Property 4 *If hypotheses (H1), (H2) and (H3) are verified, then the system is BIBS.*

Proof: The dynamics of w^+ can be bounded as follows:

$$\frac{dw^+}{dt} \leq (D+a)\left(\frac{Dw_{in}^+ - b}{D+a} - w^+\right) \tag{41}$$

if we apply property 1, we can deduce: $w^+ \leq \max(w^+(0), \frac{Dw_{in}^+ - b}{D+a})$.

In other words, $\sum w_i^+ \xi_i$ is bounded. Since $w_i^+ > 0$, the state variables ξ_i are bounded.

7.2 Equilibrium point and local behaviour

7.2.1 Introduction

In this section we briefly recall the principles of the studies of the model properties. We invite the reader to consult [28] for more details.

Generally, the bioreactor models are **non linear** (*e.g.* they often have multiple steady state), and they are of high dimension (large number of state variables). They often have a large number of parameters, which often intervene in nonlinear functions (nonlinearity with respect to the parameters).

Nevertheless, for dimensions greater than 3, it becomes very difficult to characterise the behaviour of a dynamical system. We will however show that the mass balance based model have structural properties which make easier the system understanding.

In this paragraph, we consider a general dynamical system:

$$\frac{d\xi}{dt} = f(\xi, u) \tag{42}$$

We keep in mind that $f(\xi, u) = Kr(\xi) + D(\xi_{in} - \xi) - Q(\xi)$. We will consider here the case where $u = (D, \xi_{in})$ is constant.

7.2.2 Equilibrium points and local stability

The equilibrium points are obtained for $\frac{d\xi}{dt} = 0$ when the inputs are maintained constant.

The non linear systems generically differ from linear systems since they can have multiple equilibrium points.

The first step in the model analysis consists in testing if these equilibrium points are locally stable. We consider the jacobian matrix of the linearised:

$$J(\xi) = \frac{Df}{D\xi}(\xi)$$

The equilibrium ξ_0 is locally stable if and only if all the eigenvalues of $J(\xi_0)$ have a negative real part. If there exists an eigenvalue with positive real part, the equilibrium is unstable. We can not conclude on the system stability if none eigenvalues have a positive real part but one (at least) eigenvalue has a zero real part.

7.2.3 Global behaviour

The dynamics of a nonlinear system can be very complicated, and complex behaviours like limit cycles, chaos, etc. can appear in addition to the equilibria. It is therefore important to test whether a unique locally stable equilibrium is globally stable. In other words if for any initial conditions the trajectories will converge toward this equilibrium.

The standard method to prove that an equilibrium is globally stable relies on the Lyapunov [28] approach. However it is often difficult to find a Lyapunov function for a biological system. One can refer to [29] for constructive methods to find Lyapunov functions in a large class of growth models.

7.2.4 Asymptotic behaviour

We have seen in paragraph 6.2.1 that in the general case where $n > k$, there exists $n - k$ vectors v_i in the kernel of K^T. These vectors allow to compute the quantities $w_i = v_i^t \xi$ whose dynamics satisfies equation (35).

Moreover, there are often q vectors v_i^0 among the v_i which verify:

$$v_i^{0\,t} Q(\xi) = 0 \tag{43}$$

The dynamics of the associated w_i^0 is then very simple:

$$\frac{dw_i^0}{dt} = D(w_{i\,in}^0 - w_i^0) \tag{44}$$

In the conditions that we consider (*i.e.* constant D and ξ_{in}), the solutions of (44) asymptotically converge towards $w_{i\,in}^0$. This means that the solutions of system (42) will converge towards the hyperplane $v_i^{0\,t} \xi = 0$.

The state of the system will then asymptotically converge toward the vectorial subspace of dimension $n - q$, which is orthogonal to the q vectors v_i^0. This allows to simplify the study of the n dimensional system (42) into a $n - q$ dimensional system.

7.2.5 Example 4 (continued)

Let us consider the model of fungal growth (equation 37). We will moreover assume that the kinetics has been represented by a Monod law with respect to the 2 substrates C and N:

$$r(\xi) = \mu_{max} \frac{C}{K_C + C} \frac{N}{K_N + N} X \tag{45}$$

The two vectors v_1 and v_2 identified in paragraph (6.2.2) verify straightforwardly equation (43).

Therefore when $t \to +\infty$, $X + \frac{N}{k_1} \to \frac{N_{in}}{k_1}$ and $X + \frac{C}{k_2} \to \frac{C_{in}}{k_2}$.

The study of the 3 dimensional system is then simplified into the study of the following system in dimension 1:

$$\frac{dX}{dt} = \mu_{max} \frac{C_{in} - k_2 X}{K_C + C_{in} - k_2 X} \frac{N_{in} - k_1 X}{K_N + N_{in} - k_1 X} X - DX \qquad (46)$$

One will verify that this system has three real equilibrium points (one of them being the trivial equilibrium $X = 0$). These equilibria, in increasing order, are respectively locally stable, unstable and locally unstable. With respect to the parameters values, the equilibria will be positive (and therefore admissible) or not. For the parametric domains where there exists a single positive equilibrium, this equilibrium is globally stable.

8 Conclusion

We have presented a constructive and systematic method to develop bioprocess models in 4 steps. Let us recall that the modelling of a bioprocess must be performed in the framework of a clearly identified objective. The modelling must correspond to the quality and the quantity of the available information so that the model can be correctly validated and identified.

The first modelling steps consists in gathering the physical and chemical principles that can apply to the system and to assume a reaction scheme in order to obtain the mass balance model.

In a second step, one must take benefit of the constraints that the model must verify and use the empirical relationships to find an analytical expression for the reaction kinetics.

The third step consists in identify the model parameters by separating those who are related to the mass balances (yield coefficients), those who are related with the used physical principles (affinity constants, transfer constants, etc.) and those who intervene in the reaction rates.

Finally, the ultimate modelling step must not be neglected: namely the model validation. During this last step the model quality must be tested using the more objective as possible criteria. The validity of the model must be assessed along its ability to properly represent the mass balance, to reproduce correctly the qualitative features of the data, and to fit quantitatively

the data. The important point is that the data which must be used for model validation must not have been already used in the model construction phase. During the validation step, not only the quality of the model will be assessed, but also its validity domains: the working domains (in terms of state variable and inputs) where the model is satisfactory.

To conclude, we insist on the fact that the modelling step can be very long and expensive, but the quality of a model is a necessary conditions to ensure that a controller or an observer based on it will properly work.

Appendix A. Theoretical determination of the dimension of K

Let us integrate equation (7) between 2 time instants t and $t+T$:

$$\xi(t+T) - \xi(t) \; - \int_t^{t+T} D(\xi_{in}(\tau) - \xi(\tau))) + Q(\xi(\tau))d\tau \; = \; K \int_t^{t+T} r(\xi(\tau))d\tau \, , \tag{47}$$

Let us denote:

$$v(t) = \xi(t+T) - \xi(t) \; - \int_t^{t+T} D(\xi_{in}(\tau) - \xi(\tau))) + Q(\xi(\tau))d\tau$$

and

$$w(t) = \int_t^{t+T} r(\xi(\tau))d\tau$$

Equation (47) can then be rephrased:

$$v(t) = \; K \, w(t) \tag{48}$$

The vector $v(t)$ can be estimated along time on the basis of the available measurements. The integral value can be estimated *e.g.* with a trapeze approximation.

To avoid conditioning problem and to give the same weighting to all the state variables, we normalise the data vectors $u(t_i)$ as follows:

$$\tilde{v}(t_i) = \frac{v(t_i) - e(v)}{\sqrt{N}\sigma(v)}$$

where $e(v)$ is the average value of $v(t_i)$, and $\sigma(v)$ their standard deviation.

Now the question of the dimension of matrix K can be formulated as follows: what is the dimension of the image of K, in other words, what is

the dimension of the space where $u(t)$ lives. Note that we are looking for a full rank matrix K. Otherwise, it would mean that the same dynamical behaviour could be obtained with a matrix K of lower dimension.

Determining the dimension of the $v(t)$ space is a classical problem in statistical analysis. It corresponds to the principal component analysis that determines the dimension of the vectorial space spanned by the vectors k_i, rows of K. To reach this objective, we consider matrix U obtained from a set of N recording of $v(t)$:

$$V = (\tilde{v}(t_1), \ldots, \tilde{v}(t_N))$$

We will also consider the associated matrix of reaction rates, which is unknown:

$$W = (w(t_1), \ldots, w(t_N))$$

We assume that matrix W is of full rank. This means first that there are more measurements than reactions. It means also that the reactions are independent (none of the reaction rates can be written as a linear combination of the other ones).

Property 5 *For a matrix K of rank k, if W has full rank, then the $n \times n$ matrix $M = VV^T = KWW^TK^T$ has rank k. Since it it is a positive symmetric matrix, it can be written, by:*

$$M = P^t \Sigma P$$

where P is an orthogonal matrix ($P^TP = I$) and

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & & & \ldots & & 0 \\ 0 & \sigma_2 & 0 & & & & 0 \\ \vdots & & \ddots & & & & \\ & & & \sigma_k & & & \\ & & & & 0 & & \\ & & & & & \ddots & \vdots \\ 0 & & \ldots & & & & 0 \end{pmatrix}$$

with $\sigma_{i-1} \geq \sigma_i > 0$ for $i \in \{2, ..., k\}$.

Proof: it is direct application of the singular decomposition theorem [30]. Since rank $(M) =$ rank $(\Sigma) = k$, it provides the result.

Now from a theoretical point of view it is possible to determine the number of reactions in the reaction scheme: it corresponds to the rank of K or, in other words, to the number of non zero singular values of VV^T.

In the reality, the noises due to model approximations, measurement errors or interpolation perturb the analysis. Therefore in practice there are no zero eigenvalues for the matrix $M = V^T V$.

The question is then to determine the number of eigenvectors that must be taken into account in order to represent a reasonable approximation of the data $v(t)$. To solve this problem, let us remark that the eigenvalues σ_i of M correspond to the variance associated with the corresponding eigenvector (inertia axis).

The method will then consist in selecting the p first principal axis which represent a total variance larger than a fixed threshold.

References

[1] J. E. Bailey and D. F. Ollis, *Biochemical engineering fundamentals.* McGraw-Hill, 1986.

[2] G. Bastin and D. Dochain, *On-line estimation and adaptive control of bioreactors.* Amsterdam: Elsevier, 1990.

[3] F. Mosey, "Mathematical modelling of the anaerobic digestion process: regulatory mechanisms for the formation of short-chain volatile acids from glucose," *Water Science and Technology*, vol. 15, pp. 209–232, 1983.

[4] D. Hill and C. Barth, " A dynamic model for simulation of animal waste digestion," *Journal of the Water Pollution Control Association*, vol. 10, pp. 2129–2143, 1977.

[5] R. Moletta, D. Verrier, and G. Albagnac, "Dynamic modelling of anaerobic digestion," *Wat.Res.*, vol. 20, pp. 427–434, 1986.

[6] D. Costello, P. Greenfield, and P. Lee, " Dynamic modelling of a single-stage high-rate anaerobic reactor - I. Model derivation," *Water Research*, vol. 25, pp. 847–858, 1991.

[7] D. Batstone, J. Keller, B. Newell, and M. Newland, "Model development and full scale validation for anaerobic treatment of protein and fat based wastewater," *Water Science and Technology*, vol. 36, pp. 423–431, 1997.

[8] J. Merchuk, "Further considerations on the enhancement factor for oxygen absorption into fermentation broth," *Biotechnol. & Bioeng.*, vol. 19, pp. 1885–1889, 1977.

[9] O. Bernard, Z. Hadj-Sadok, D. Dochain, A. Genovesi, and J. Steyer, " Dynamical model development and parameter identification for an anaerobic wastewater treatment process," *Biotech.Bioeng.*, no. 75, pp. 424–438, 2001.

[10] P. Kudva, N. Viswanadham, and A. Ramakrishna, "Observers for linear systems with unknown inputs," *IEEE Trans. Autom. Contr.*, vol. AC-25, no. 1, pp. 113–115, 1980.

[11] M. Hou and P. Mller, "Design of observers for linear systems with unknown inputs," *IEEE Trans. Autom. Contr.*, vol. AC-37, no. 6, pp. 871–875, 1991.

[12] M. Darouach, "On the novel approach to the design of the unknown input observers," *IEEE Trans. Autom. Cont.*, vol. 39, no. 3, pp. 698–699, 1994.

[13] J. Monod, *Recherches sur la croissance des cultures bactriennes.* Paris, France: Hermes, 1942.

[14] L. A. Segel, *Modeling Dynamic Phenomena in Molecular and Cellular Biology.* Cambridge: Cambridge University Press, 1984.

[15] L. Edelstein, *Mathematical Models in Biology.* New York: Random House, 1988.

[16] J. Andrews, "A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrate," *Biotechnol. & Bioeng.*, vol. 10, pp. 707–723, 1968.

[17] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation.* Addison-Wesley, 1991.

[18] L. Chen, O. Bernard, G. Bastin, and P. Angelov, "Hybrid modelling of biotechnological processes using neural networks," *Contr. Eng. Prcatice*, vol. 8, pp. 821–827, 2000.

[19] A. Karama, O. Bernard, A. Genovesi, D. Dochain, A. Benhammou, and J.-P. Steyer, "Hybrid modelling of anaerobic wastewater treatment processes," *Wat. Sci. Technol.*, vol. 43, no. 1, pp. 43–50, 2001.

[20] C. Jeffries, "Qualitative stability of certain nonlinear systems," *Linear Algebra and its Applications*, vol. 75, pp. 133–144, 1986.

[21] E. Sacks, "A dynamic systems perspective on qualitative simulation," *Artif. Intell.*, vol. 42, pp. 349–362, 1990.

[22] O. Bernard and J.-L. Gouzé, "Transient behavior of biological loop models, with application to the Droop model," *Mathematical Biosciences*, vol. 127, no. 1, pp. 19–43, 1995.

[23] J.-L. Gouzé, "Positive and negative circuits in dynamical systems," *Journal Biol. Syst.*, vol. 6, no. 1, pp. 11–15, 1998.

[24] S. R. Hansen and S. P. Hubbell, "Single-nutrient microbial competition," *Science*, vol. 207, no. 28, pp. 1491–1493, 1980.

[25] E. Walter and L. Pronzato, *Identification de modèles paramétriques.* Masson, 1994.

[26] O. Bernard, Z. Hadj-Sadok, and D. Dochain, "Dynamical modelling and state estimation of anaerobic wastewater treatment plants," in *Proceedings of ECC99 (CDROM)*, Karlsruhe, Germany, 1999.

[27] O. Bernard, M.Polit, Z. Hadj-Sadok, M.Pengov, D. Dochain, M.Estaben, and P.Labat, "Advanced monitoring and control of anaerobic wastewater treatment plants: software sensors and controllers for an anaerobic digestor," *Wat. Sci. Technol*, vol. 43, no. 7, pp. 175–182, 2001.

[28] H. Khalil, *Nonlinear Systems.* Macmillan Publishing Company, 1996.

[29] B. Li, "Global asymptotic behavior of the chemostat: General response functions and different removal rates," *SIAM Journal*, vol. 59, 1998.

[30] R. Horn and C. Johson, *Matrix analysis.* Cambridge Univesity Press, 1992.

State Estimation for Bioprocesses

Olivier Bernard*and Jean-Luc Gouzé

COMORE, INRIA, France

Lectures given at the
Summer School on Mathematical Control Theory
Trieste, 3-28 September 2001

LNS0280013

*obernard@sophia.inria.fr

Abstract

In these lecture notes we explain how to build an observer for a biological system. We review the existing linear and nonlinear observers and we propose criteria to define which is the best observer with respect to the available information. Depending on the model reliability and on the level of noise, we can develop observers which use the full model description (high gain observers) or asymptotic observers which use only a mass balance model where the biological kinetics are considered as unknown inputs. If the bounds on the uncertainties can be characterised, interval observers can be designed. Each observer is illustrated with an example performed on a biological system.

Contents

1 Introduction **817**

2 Notions on system observability **817**
2.1 System observability: definitions 818
2.2 General definition of an observer 819
2.3 How to manage the uncertainties in the model or in the output 821

3 Observers for linear systems **822**
3.1 Luenberger observer 823
3.2 The linear case up to an output injection 824
3.3 Local observation of a nonlinear system around an equilibrium point 824
3.4 PI observer 825
3.5 Kalman filter 825
3.6 The extended Kalman filter 827

4 High gain observers **827**
4.1 Definitions, hypotheses 827
4.2 Change of variable 828
4.3 Fixed gain observer 829
4.4 Variable gain observers (Kalman like observer) 829
4.5 Example: growth of micro-algae 830

5 Observers for mass balance based systems **833**
5.1 Introduction 833
5.2 Definitions, hypotheses 834
5.3 The asymptotic observer 835
5.4 Example 836
5.5 Improvements 838

6 interval observers **840**
6.1 Principle 840
6.2 The linear case up to an output injection 842

7 Interval estimator for an activated sludge process **844**

8 Conclusion **846**

1 Introduction

One of the main limitations to the improvement of monitoring and optimisation of bioreactors is probably due to the difficulty to measure chemical and biological variables. Indeed there are very few sensors which are at the same time cheap and reliable and that can be on-line used. The measurement of some biological variables (biomass, cellular quota, etc.) is sometimes very difficult and can necessitate complicated and sophisticated operations.

The question is to estimate the internal state of a bioreactor when only a few measurements are available. In this lecture we propose methods to build observers which will use the available measurements to estimate non measured state variables (or at least some of them). The principle of this so called "software sensor" is to use the process model to reconstruct asymptotically the state on the basis of the outputs. As it will be detailed in this chapter, the system must be observable, or at least detectable, in order to estimate the internal state.

There are numerous methods to design an observer. They rely on ideas that can be very different. Thus the best observer must be chosen with respect to the type of problem. The choice will then be strongly connected to the quality and the uncertainties of the model and of the data. If the biological kinetics are not precisely known, the mass balance will be the core of the asymptotic observers. If there are bounded uncertainties on the inputs and/or on the parameters, then we will estimate intervals in which the state of the system should lie. If the model has been correctly identified an validated, then we can fully exploit it and -if the output are not corrupted with a high level of noise- we can develop a high gain observer.

The type of observer to be developed must not be based only on the model quality: it must also take into account the objectives to be achieved. Indeed, an observer can have other purposes than monitoring a bioreactor: it can be developed to apply a control action which need an estimate of the internal state. It can also be used to determine if a failure did not happen in the process.

2 Notions on system observability

We will only recall the main useful notions, we will give references for the more technical parts (see [1, 2]).

The observability notion is fundamental in automatic control. Intuitively,

one tries to estimate the state variables from the available measurements. If this is possible from a theoretical point of view, the system is said to be observable. Then the next question is how to derive an observer which is another dynamical system providing a state estimate. Let us mention that the question of observability and of observer design are very different: the observability property does not give any clue on how to build an observer.

The theory is extensively developed in the linear case (see next section) and, in the nonlinear case, has been strongly developed during the last years but for particular classes of models.

2.1 System observability: definitions

We will consider the general continuous time system:

$$(\mathcal{S}) \begin{cases} \frac{dx}{dt}(t) = f(x(t), u(t)) \ ; \ x(t_0) = x_0 \\ y(t) = h(x(t)) \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the input vector, $y \in \mathbb{R}^p$ is the output vector, x_0 is the initial condition for initial time t_0, $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ and $h : \mathbb{R}^n \longrightarrow \mathbb{R}^p$.The functions are assumed to be sufficiently smooth in order to avoid problems of existence and uniqueness of the solutions.

Example: For the bioreactors described by a mass balance model, we have:

$$f(x(t), u(t)) = Kr(x(t)) + D(x_{in}(t) - x(t)) - Q(x(t))$$

Here D and x_{in} stands for the input vector.
We assume therefore that, for system $(\mathcal{S})$,

- the input $u(t)$ is known
- the output $y(t)$ is known
- functions f and h, are known, *i.e.* the model is known (for a bioreactor it means that $r(.)$ is known in the mass balance based modelling).

We want to estimate $x(t)$; the observability is a theoretical notion that states if it is possible.

Definition 1 *Two states x_0 and x_0' are said indiscernible if for any input time function $u(t)$ and for any $t \geq 0$, the outputs $h(x(t, x_0))$ and $h(x(t, x_0'))$ that result are equal.*

Definition 2 *The system is said to be observable if it do not have any distinct couple of initial state* x_0, x_0' *that are indiscernible.*

This means that for any input the initial condition can be uniquely estimated from the output. It can be noticed that generally for nonlinear system the observability depends on the input; a system can be observable for some inputs and not observable for others.

Definition 3 *An input is said to be universal if it can distinguish any couple of initial conditions.*

Definition 4 *A non universal input is said to be singular.*

Even in the case where all the inputs are universal (the system is said to be uniformly observable and can be rewritten under a specific shape, see section 4), this can be insufficient in practice. We impose then that the universal property persists with time, and we obtain (at least for some systems) the notion of regularly persisting input (see Hypothesis 5, paragraph 5.3).

For the linear systems things are much simpler (see next section).

2.2 General definition of an observer

Once the system has been proven to be observable, the next step is the observer building in order to estimate the state variable x from the inputs, the outputs and the model.

The observer principle is presented on Figure 1. It is a second dynamical system that will be coupled to the first one thanks to the measured output.

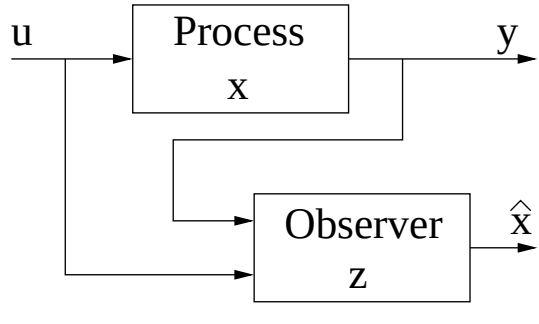


Figure 1: Observer principle

Definition 5 *An observer is an auxiliary system coupled with the original system:*

$$(\mathcal{O}) \begin{cases} \frac{dz}{dt}(t) = \hat{f} : (z(t), u(t), y(t)) \ ; \ z(t_0) = z_0 \\ \hat{x}(t) = \hat{h} : (z(t), u(t), y(t)) \end{cases} \tag{2}$$

with $z \in \mathbb{R}^q$, $\hat{f}\ \mathbb{R}^q \times \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \mathbb{R}^q$ *and* $\hat{h}\ \mathbb{R}^q \times \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \mathbb{R}^n$ *such that*

$$\lim_{t \to \infty} \| \ x(t) - \hat{x}(t) \ \| = 0 \tag{3}$$

It is the classical definition which may be insufficient in some cases. It is stated that the estimation error tends asymptotically toward zero. Indeed one tries to tune the error decreasing rate (convergence rate). Let us explain this with a simple linear example: let us consider the linear system $\frac{dx}{dt} = Ax + Bu$ where $x \in \mathbb{R}^n$ and let us assume that matrix A is stable. A trivial observer can be obtained with a copy of the system: $\frac{d\hat{x}}{dt} = A\hat{x} + Bu$. Indeed, the error $e = x - \hat{x}$ follows the same dynamics $\frac{de}{dt} = Ae$ and therefore converges toward zero. Let us remark that this observer does not necessitate any output. This example shows that the stable internal dynamics is sufficient to estimate the final state. This example highlights a property which will be called detectability for linear systems and which will be the basis of asymptotic observer (section 4) in a different framework. As a consequence, an additional requested property is to be able to tune the convergence rate of the observer in order to be able to reconstruct the state variables more rapidly than the dynamics of the system. Let us remark that the observer variable (z in $\mathcal{O}$) can be of greater dimension than the state variable to be estimated x.

Another property that we wish is that if the observer is properly initiated, *i.e.* with the true value $x(0)$, then its estimation remains equal to $x(t)$ for all t. This suggest a peculiar structure for the observer

Definition 6 *Often, the following observer is taken:*

$$(O) \begin{cases} \frac{d\hat{x}}{dt}(t) = f(\hat{x}(t), u(t)) + k[z(t), h(\hat{x}(t)) - y(t)] \\ \frac{dz}{dt}(t) = \hat{f}(z(t), u(t), y(t)) \quad with \quad k(z(t), 0) = 0 \end{cases}$$

This is a copy of the system with a correcting term depending on the discrepancy between the true measured outputs and the value of the output computed from the observer. The correction amplitude is tuned thanks to the function k that can be seen as a gain (it is an internal tuning of the

observer).

In the ideal case, the gain k can be tuned in order to have a converging rate as large as requested.

Definition 7 *System* (O) *is said to be an exponential observer if, for any positive* λ*, the gain* k *can be tuned such that*

$$\forall (x_0, \hat{x}(0), z_0) \;\; \forall t > 0, \;\; ||\hat{x}(t) - x(t)|| \leq e^{-\lambda t} ||\hat{x}(0) - x(0)||.$$

2.3 How to manage the uncertainties in the model or in the output

In real life- and especially in the biological field- one often considers that there are noises either in the output (measurement noise) or in the state equation (model noise). In general the model noise is assumed to be additive (see section 3.5), which is a strong hypothesis (it could be e.g. multiplicative).

Another important case which often appears in the bioprocesses is when the model integrates some unknown parts. For example the biological kinetics in the mass balance models for bioreactors are generally not precisely known [3].

How to manage these two problems which have some related aspects ?

- Linear filtering, and more specifically Kalman filtering. It is the most popular method. It assumes that the noises are additive and white; it minimises the error variance (see next section).

- The approach L^2, H^2 or H^∞. It consists in assuming that the noises or perturbations $w(t)$ belong to a given class of functions (L^2) and to try to minimise their impact on the output using the transfer function. In the H^2 approach, one tries to minimise the norm of this transfer function. in the approach H^∞, one tries to minimise the input effect in the worst case (see [4]). For example, for a $\gamma > 0$ and R a positive definite matrix, one wants the observer $\hat{x}$ to verify:

$$\sup_{w(.)} \int_0^\infty |\hat{x}(t) - x(t)|_R^2 - \gamma^2 |w(t)|^2 \, dt \;\; \leq 0.$$

- Disturbance rejection. One tries to build observers independent from the unknown perturbation. The disturbance is cancelled for example thanks to linear combinations of variables [5, 6].

  The asymptotic observers are among this class of systems (see section 5).

- Bounds on the perturbations and on the uncertainties. One assumes that uncertainties are bounded, and one tries to design interval observers which provide the best possible bounds for the variables to be estimated. For some cases, one tries to minimise this bounding (section 6).

- One can also use these bounds to design sliding mode observers which have a correcting term of the type $\text{sign}(x - \hat{x})$. Note that the way these observers take the uncertainties into account generates a discontinuous dynamics on the sliding manifolds [7].

Remark: it is possible to construct examples where a system is observable when the model is known and becomes unobservable when a part of the model is unknown. For such cases the requirement for a classical observer may be relaxed. In particular, we will not assume anymore that

$$\lim_{t\to\infty} \parallel x(t) - \hat{x}(t) \parallel = 0 \tag{4}$$

but that the discrepancy tends toward a reasonable value for practical applications.

3 Observers for linear systems

For single output linear stationary systems we have:

$$(\mathcal{S}_L): \left\{ \begin{array}{lcl} \frac{dx}{dt}(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) \end{array} \right. \tag{5}$$

with $A \in \mathcal{M}^{n\times n}(\mathbb{R})$ $(n \geq 2)$, $C \in \mathcal{M}^{1\times n}(\mathbb{R})$.

The well known observability criterion is formulated as follows:

$$(\mathcal{S}_L) \text{ observable } \Leftrightarrow \text{rank} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = n.$$

which relies on the fact that the observability space is generated by the vectors $(C, CA, \ldots, CA^{n-1})$.

The canonical observability forms, that can be obtained after a linear change of coordinates, highlight the observation structure. They will reappear in the nonlinear case for the high gain observer (section 6).

Theorem 1 *If the pair (A, C) is observable, then there exists an invertible matrix P such that:*

$$A_0 = P^{-1}AP, \quad C_0 = CP$$

with

$$A_0 = \begin{pmatrix} -a_n & 1 & 0 & \ldots 0 \\ -a_{n-1} & 0 & 1 & \ldots 0 \\ \vdots & \vdots & \vdots & \vdots \\ -a_2 & 0 & 0 & \ldots 1 \\ -a_1 & 0 & 0 & \ldots 0 \end{pmatrix} \quad C_0 = (1 \quad 0 \cdots 0)$$

What happens if the system is not observable ? One can rewrite it in two parts, as it is shown in the following theorem. Here A_1 and A_3 are two square matrices with dimensions corresponding to x_1 and x_2. The canonical form shows clearly that x_1 can not be estimated from x_2.

Theorem 2 *General canonical form:*

$$\begin{array}{lcl} \frac{dx_1}{dt} & = & A_1x_1 + A_2x_2 + B_1u \\ \frac{dx_2}{dt} & = & A_3x_2 + B_2u \\ y & = & C_2x_2 \end{array}$$

Matrix A_1 imposes the dynamics of the unobservable part; if it is stable, then the dynamics of the total error will be stable, but the unobservable part will tend toward zero with its own dynamics (given by A_1); the system is said to be detectable.

3.1 Luenberger observer

If system (5) is observable, a Luenberger observer [8] can be derived:

$$\frac{d\hat{x}(t)}{dt} = A\hat{x}(t) + Bu(t) + K(C\hat{x}(t) - y(t))$$

where K is a dimension n gain vector, which allows to tune the convergence rate of the observer.

Indeed, the dynamics of the observation error $e = x - \hat{x}$ is:

$$\frac{de}{dt} = (A + KC)e$$

Let us note that this dynamics do not depend on the input. The pole placement theorem sates that the error dynamics can be arbitrarily chosen.

Theorem 3 *If (A, C) is observable, the vector K can be chosen to have an arbitrary linear dynamics of the observation error.*

In particular, the gain vector K can be chosen in order that the error converges rapidly toward zero. But then the observer will be very sensitive to perturbations (measurement noise for example). A good compromise must be chosen between stability and precision. The Kalman filter is a way to manage this compromise.

3.2 The linear case up to an output injection

There is a very simple case for which a linear observer can be designed for a nonlinear system, it is the case where the nonlinearity depends only on the output y.

$$(\mathcal{S}): \quad \left\{ \begin{array}{lcl} \frac{dx}{dt}(t) & = & Ax(t) + \phi(t, y(t)) + Bu(t) \\ y(t) & = & Cx(t) \end{array} \right. \tag{6}$$

ϕ is a nonlinear (known) function which takes its values in $\mathbb{R}^n$. The following " Luenberger like" observer generates a linear observation error equation:

$$\frac{d\hat{x}(t)}{dt} = A\hat{x}(t) + \phi(t, y(t)) + Bu(t) + K(C\hat{x}(t) - y(t))$$

The dynamics can be arbitrarily chosen if the pair (A, C) is observable.

3.3 Local observation of a nonlinear system around an equilibrium point

Let us consider the general system (1), and let us assume that it admits a single equilibrium point (working point) at (x_e, u_e). The system can then be linearised around this point:

Theorem 4 *The linearised system of (1) around (x_e, u_e) is*

$$(\mathcal{S}) \left\{ \begin{array}{l} \frac{dX}{dt}(t) = AX + BU \\ Y(t) = CX \end{array} \right. \tag{7}$$

with

$$A = \frac{\partial f(x,u)}{\partial x}\ B = \frac{\partial f(x,u)}{\partial u}\ C = \frac{\partial h(x)}{\partial x}$$

Matrices A, B, C *are estimated at* x_e, u_e. *Variables* X, U, Y *are deviations toward equilibrium:*

$$X = x - x_e,\ U = u - e_e,\ Y = y - Cx_e$$

If the pair (A, C) *is observable, the nonlinear system is locally observable around the equilibrium.*

3.4 PI observer

The Luenberger observer is based on a correction of the estimations with a term related to the difference between the measured outputs and the predicted outputs.

The idea behind the proportional integral observer is to use the integral of this error term. We consider the auxiliary variable $\hat{w}$:

$$\hat{w} = \int_0^t (C\hat{x}(\tau) - y(\tau))d\tau.$$

The PI observer for system (8) will then be rewritten:

$$\left\{ \begin{array}{lcl} \frac{d\hat{x}}{dt}(t) & = & Ax(t) + Bu(t) + K_I(C\hat{x}(t) - y) + K_P\hat{w} \\ \frac{d\hat{w}}{dt}(t)(t) & = & C\hat{x} - y \end{array} \right. \tag{8}$$

The error equation ($e_x = \hat{x} - x$ and $e_w = \hat{w}$) is then:

$$\left(\begin{array}{c} \frac{de_x}{dt}(t) \\ \frac{de_w}{dt}(t) \end{array} \right) = \left(\begin{array}{cc} A + K_I C & K_P \\ C & 0 \end{array} \right) \left(\begin{array}{c} e_x(t) \\ e_w(t) \end{array} \right) \tag{9}$$

The gains K_I and K_p can be chosen such as to ensure stable error dynamics [9]. The integrator addition provides more robustness to the observer to deal with measurement noise or modelling uncertainties.

3.5 Kalman filter

The Kalman filter (see [10]) is very famous in the framework of linear systems; it can be seen as Luenberger observer with a time varying gain; this allows to minimise the error estimate variance.

A stochastic representation can be given by the observable system:

$$\begin{cases} \frac{dx}{dt}(t) = A\ x(t) + Bu(t) + w(t)\ \ ;\ \ x(t_0) = x_0 \\ y(t) = C\ x(t) + v(t) \end{cases} \tag{10}$$

where $w(t)$ and $v(t)$ are independent centred white noises (Gaussian perturbations), with respective covariances $Q(t)$ and $R(t)$. Let us also assume that the initial distribution is Gaussian, such that:

$$E[x_0] = \hat{x}_0\ \ ;\ \ E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T] = P_0 \tag{11}$$

where E represents the expected value and P_0 is the initial covariance matrix of the error. The filter is written in several steps:

1. Initialisation:

$$E[x_0] = \hat{x}_0\ \ ;\ \ E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T] = P_0 \tag{12}$$

2. Estimation of the state vector:

$$\frac{d\hat{x}}{dt}(t) = A\ \hat{x}(t) + Bu(t) + K(t)\ [y(t) - C\ \hat{x}(t)]\ \ ;\ \ \hat{x}(t_0) = \hat{x}_0 \tag{13}$$

3. Error covariance propagation (Riccati equation):

$$\frac{dP}{dt}(t) = A\ P(t) + P(t)\ A^T - P(t)C^T R(t)^{-1} C P(t) + Q(t) \tag{14}$$

4. Gain computation:

$$K(t) = P(t)\ C^T R(t)^{-1} \tag{15}$$

Some points can be emphasised:

- This filter can still be applied when matrices A and C depend on time (the observability must nevertheless be proven).
- The estimation of the positive definite matrices R, Q, P_0 is often very delicate, especially when the noise properties are not known.
- A deterministic interpretation of this observer can be given: it consists in minimising the integral from 0 to t of the square of the error.
- This observer can be extended by adding a term $-\theta P(t)$ in the Riccati equation. This exponential forgetting factor allows to consider the cases where $Q = 0$.

3.6 The extended Kalman filter

The idea consists in linearising a nonlinear system around its estimated trajectory. Then the problem is equivalent to build a Kalman filter for non stationary system. Let us consider the system

$$\begin{cases} \frac{dx}{dt}(t) = f(x(t)) + w(t) \ ; \ x(t_0) = x_0 \\ y(t) = h(x(t)) + v(t) \end{cases} \tag{16}$$

and the observer is designed as above, with a change in the second step:

2. Estimation of the state vector:

$$\frac{d\hat{x}}{dt}(t) = f(\hat{x}(t)) + K(t)\ [y(t) - h(\hat{x}(t))] \ ; \ \hat{x}(t_0) = \hat{x}_0 \tag{17}$$

and using the matrices of the tangent linearised:

$$A(t) = \left.\frac{\partial f(x(t))}{\partial x(t)}\right|_{x(t)=\hat{x}(t)} \quad C(t) = \left.\frac{\partial h(x(t))}{\partial x(t)}\right|_{x(t)=\hat{x}(t)} \tag{18}$$

This extended filter is often used, even if only few theoretical results guarantee its convergence (see Section 4.4).

4 High gain observers

4.1 Definitions, hypotheses

In this chapter, we will assume that a simulation model of the process is available, (*i.e.* with modelling of the biological kinetics). We also assume that the model has been deeply validated: the high gain observers are dedicated to the nonlinear systems and require a high quality modelling.

We will consider now the systems which are affine with respect to the input, that are described as follows:

$$\frac{d\xi}{dt} = f(\xi) + ug(\xi) \tag{19}$$

We consider here the case where $u \in \mathbb{R}$. For bioreactors, the input corresponds generally to the dilution rate $u = D$. In this case $f(\xi) = Kr(\xi) - Q(\xi)$ and $g(\xi) = \xi_{in} - \xi$.

Moreover, we assume that the output is a function of the state: $y = h(\xi) \in \mathbb{R}$.

Hypothesis 1 *We will state the two following hypotheses:*

- *[i] the system (19) is observable for any input.*
- *[ii] there exists a positively invariant compact $\mathcal{K}$, such that for any time t, $\xi(t) \in \mathcal{K}$.*

We will denote $L_f h(\xi) = \frac{Dh}{D\xi} f(\xi)$, which is the Lie derivative of h along the vector field f. By convention, we will write $L_f^p h(\xi) = L_f L_f^{p-1} h(\xi)$.

4.2 Change of variable

Let us consider the following change of coordinates, defined on the compact set $\mathcal{K}$:

$$\phi : \xi \longrightarrow \zeta = \left[\; h(\xi), \quad L_f h(\xi), \quad ..., \quad L_f^{(n-1)} h(\xi) \; \right]^T \tag{20}$$

This change of variable consists in considering (in the autonomous case) the output y and its $n-1$ first derivatives as new coordinates.

Hypothesis 2 *The mapping ϕ is a global diffeomorphism.*

One can verify [11] that under Hypothesis 2 ϕ transforms (19) into:

$$\frac{d\zeta}{dt} = A\zeta + \tilde{\psi}(\zeta) + \bar{\psi}(\zeta)u \tag{21}$$

$$y = C\zeta \tag{22}$$

with:

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}, \quad C = [1, 0, \dots, 0]$$

$$\tilde{\psi}(\zeta) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ L_f^n h(\phi^{-1}(\zeta)) \end{pmatrix}, \quad \bar{\psi}(\zeta) = \begin{pmatrix} \bar{\psi}_1(\zeta_1) \\ \bar{\psi}_2(\zeta_1, \zeta_2) \\ \vdots \\ \bar{\psi}_2(\zeta_1, \zeta_2, \dots, \zeta_n) \end{pmatrix}$$

where

$$\bar{\psi}_i(z) = \bar{\psi}_i(\zeta_1, .., \zeta_i) = L_g L_f^{(i-1)} h[\phi^{-1}(\zeta)] \tag{23}$$

In this canonical form, all the system nonlinearities have been concentrated in the terms $\tilde{\psi}(\zeta)$ and $\bar{\psi}(\zeta)$. We will present the various observers using this canonical form (let us note that this canonical form is very close to the one in section 3 for the observer pole assignment).

Let us remark that an observer in the new basis will provide an estimate $\hat{\zeta}$ which will estimate ζ, *i.e.* the successive output derivatives. The idea consists in writing the observer in this canonical basis *i.e.* a numerical differentiator of the output. Then, going back to the initial coordinates (applying $\phi^{-1}(\zeta)$), the observer will be expressed in the original basis.

To design a high gain observer, we need an additional technical hypothesis:

Hypothesis 3 *The mappings $\tilde{\psi}$ and $\bar{\psi}$ defined in (21) are global Lipschitz on $\mathcal{K}$.*

Intuitively, this hypothesis will allow us to dominate the non-linear part, imposing that the dynamics of the observer can be faster than the system ones (this explains the idea of the "high gain").

4.3 Fixed gain observer

Property 1 *[12] For a sufficiently high gain θ, and under Hypotheses 1, 2 and 3 the following differential system is an exponential observer of (19):*

$$\frac{d\hat{x}}{dt} = f(\hat{x}) + u\, g(\hat{x}) - \left[\frac{\partial \phi}{\partial x}\right]^{-1}_{x=\hat{x}} S_\theta^{-1}\, C^t (h(\hat{x}) - y) \tag{24}$$

where S_θ, is the solution of the equation $\theta S_\theta + A^t S_\theta + S_\theta A = C^t C$
S_θ can be computed as follows:

$$S_\theta(i,j) = \frac{(-1)^{i+j}}{\theta^{i+j-1}} \frac{(i+j-2)!}{(i-1)!(j-1)!} \tag{25}$$

For the convergence proof and other details we refer to [12].

4.4 Variable gain observers (Kalman like observer)

The extended Kalman filter is often used in a framework where its convergence is not guaranteed (see section 3.5). We show here how to build a high gain observer very close to the Kalman filter (after change of variable), whose convergence is guaranteed.

Property 2 *[13] For a gain θ sufficiently high, and under hypotheses 1, 2 and 3 the following differential system is an exponential observer of (19):*

$$\begin{cases} \frac{d\hat{x}}{dt} = f(\hat{x}) + u\, g(\hat{x}) - \frac{1}{r}\left[\frac{\partial \phi}{\partial x}\right]^{-1}_{x=\hat{x}} S^{-1}\, C^t (h(\hat{x}) - y) \\ \frac{dS}{dt} = -SQ_\theta S - A^{\star\, t}(\hat{x}, u) S - S A^\star(\hat{x}, u) + \frac{1}{r} C^t C \end{cases} \tag{26}$$

with $r > 0$, Q_θ is computed from the two positive definite symmetric matrices Δ_θ and Q:

$$\Delta_\theta = diag(\theta, \theta^2, \ldots, \theta^n) \tag{27}$$

$$Q_\theta = \Delta_\theta Q \Delta_\theta \tag{28}$$

Matrix $A^\star$ can be computed from the diffeomorphism ϕ:

$$A^\star(\hat{\xi}, u) = A + \left[\frac{\partial \phi}{\partial \zeta}\right]_{\zeta=\phi(\hat{\xi})} + u \left[\frac{\partial \psi}{\partial \zeta}\right]_{\zeta=\phi(\hat{\xi})} \tag{29}$$

We refer to [13] for the proof of the convergence of this observer and for more details, especially for the choice of r and of matrix Q.

It is worth noting that, even if the filtering and noise attenuation performances of this extend Kalman filer are *a priori* better, this observer is above all a high gain observer; it will therefore present the same generic high sensitivity with respect to the measurement noises and modelling errors.

The advantages of the Kalman like high gain observer have a price: this observer is heavier to implement. $\frac{n(n+3)}{2}$ differential equations must be integrated instead of n equations for the simple high gain observer.

4.5 Example: growth of micro-algae

We will consider the growth of micro algae in a continuous photobioreactor. The algal development is limited by a nitrogen source (NO_3) denoted S, and uses principally the inorganic dissolved carbon (C), mainly under the form of CO_2. The algal biomass (X) will then correspond to an amount of particulate nitrogen (N).

In order to simultaneously describe the cellular carbon and nitrogen uptake, we will consider the following reaction scheme

$$S \xrightarrow{r_1(.)} N$$

$$C \xrightarrow{r_2(.)} X$$

Setting $\xi = (X,\ N,\ S,\ C)^t$, the mass balance based model (33) can be written with:

$$K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ -1 & 0 \\ 0 & -k_1 \end{pmatrix},\ \xi_{in} = \begin{pmatrix} 0 \\ 0 \\ S_{in} \\ C_{in} \end{pmatrix}\ Q(\xi) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ Q_c(\xi) \end{pmatrix}$$

The units for carbon and nitrogen are the same for biomass and substrate, and moreover the nitrogen uptake yield is assume to be unitary. The nutrient uptake rate is assumed to follow a Michaelis-Menten law [14]:

$$r_1(\xi) = \rho_{max} \frac{S}{S + k_S} X$$

The algal growth from carbon is $r_2(\xi) = \mu(\xi)X$, where the growth rate $\mu(\xi)$ is described by the Droop law [15]:

$$\mu(\xi) = \mu(q) = \bar{\mu}(1 - \frac{k_q}{q}) \tag{30}$$

Variable q represents the internal nitrogen quota defined by the amount of nitrogen per biomass unit: $q = \frac{N}{X}$.

We assume that biomass is measured (it is estimated by its total biovolume), and will be used to design a high gain observer to determine S and q.

In this case, the nitrate concentration in the renewal medium (S_{in}) can be controlled. More precisely, S_{in} can vary as follows:

$$S_{in} = s_{in}(1 + u)$$

where u is the control, and s_{in} the nominal concentration, corresponding to $u = 0$.

In the sequel, we will consider only the 3 first equations of this system, and we will consider the following change of variables:

- $x_1 = \frac{\rho_m X}{s_{in}}$; $x_2 = \frac{N}{X k_q}$; $x_3 = \frac{S}{s_{in}}$
- $a_1 = \frac{k_s}{s_{in}}$; $a_2 = \bar{\mu}$; $a_3 = \frac{\rho_m}{k_q}$

that leads to the following system:

$$\begin{cases} \dfrac{dx}{dt} = f(x) + ug(x) \\ y = h(x_1) \end{cases} \tag{31}$$

with:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \; f(x) = \begin{pmatrix} a_2(1 - \dfrac{1}{x_2})x_1 - Dx_1 \\ a_3 \dfrac{x_3}{a_1 + x_3} - a_2(x_2 - 1) \\ D(1 - x_3) - \dfrac{x_1 x_3}{a_1 + x_3} \end{pmatrix}, \tag{32}$$

$$g(x) = \begin{pmatrix} 0 \\ 0 \\ D \end{pmatrix}, \quad h(x_1) = x_1$$

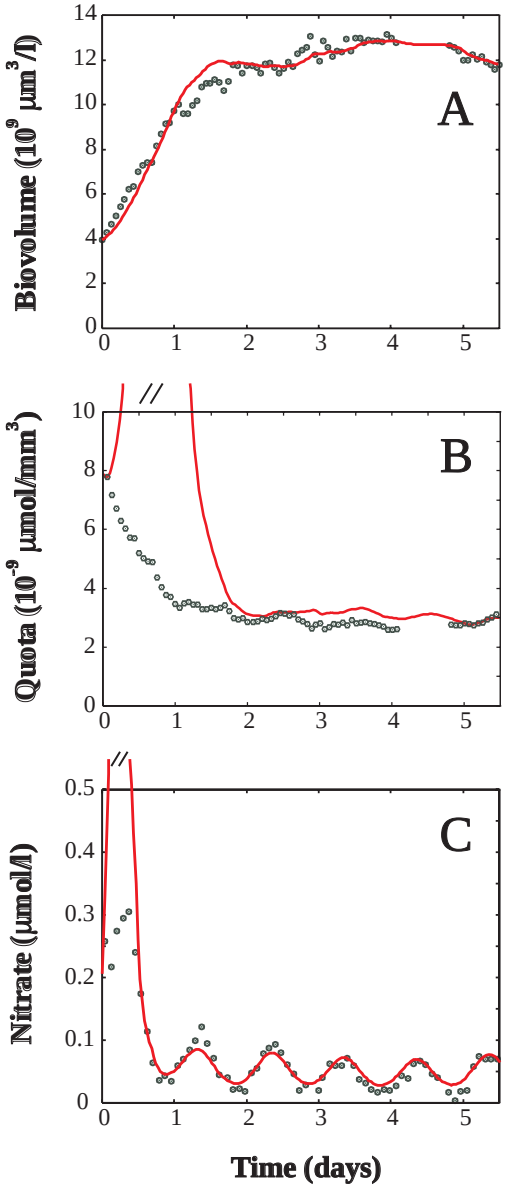


Figure 2: Comparison between direct measurements (•) and observer predictions (—) for model (31): (A) Biomass estimated from total algal biovolume. (B) internal quota. (C) Nitrate concentration

The high gain observer for model (31) is then given by:

$$G(\hat{x}) = \begin{pmatrix} 3\theta \\ \left[3\theta\frac{\hat{x}_2}{\hat{x}_1}\left[1-\left(1-\frac{D}{a_2}\right)\hat{x}_2\right]+3\theta^2\frac{\hat{x}_2^2}{a_2\hat{x}_1}\right] \\ \left[3\theta\hat{B}_{31}+3\theta^2\hat{B}_{32}+\theta^3\frac{\hat{x}_2^2(a_1+\hat{x}_3)^2}{a_1a_2a_3\hat{x}_1}\right] \end{pmatrix}$$

with:

$$\begin{aligned} \hat{B}_{31} &= \frac{1}{a_1a_3\hat{x}_1}\left[\frac{a_3\hat{x}_3}{a_1+\hat{x}_3}+2a_2+\hat{x}_2^2\left(2a_2-3D-\frac{D^2}{a_2}\right)\right. \\ &\quad \left.-\hat{x}_2\left(2\frac{a_3\hat{x}_3}{a_1+\hat{x}_3}\left(1-\frac{D}{a_2}\right)+4a_2-4D\right)\right] \\ \hat{B}_{32} &= \frac{\hat{x}_2(a_1+\hat{x}_3)^2}{a_1a_2a_3\hat{x}_1}\left[\hat{x}_2(2D-3a_2)+4a_2+2\frac{a_3\hat{x}_3}{a_1+\hat{x}_3}\right] \end{aligned}$$

An experiment where u fluctuates sinusoidally was used to validate the observer. Figure 2 proves the observer efficiency when the model is well known. The observer predictions are in agreement with the experimental measurements. For more details on this example, see [16, 17].

5 Observers for mass balance based systems

5.1 Introduction

In the previous sections we have considered the case where the uncertainties were due to noise on the outputs and, in some cases were due to modelling noise. We have seen Chapter 2 that the bioprocess models are often badly known. In particular when the model is written on the basis of a mass balance analysis, a term representing the reaction rates appears. This term which represents the biological kinetics with respect to the model state variable is often speculative. Often the modelling of the reaction rate is not reliable enough to base an observer on it. In this section we will use the results for the observers with unknown inputs [6, 18, 5], whose principle relies on a cancellation of the unknown part after a change of variable in order to build the observer.

We will show how to build an observer for a system represented by a mass balance and for which the kinetics would not have been expressed. We

will see that the main condition to design such an observer is that enough variables are measured. In particular we will not assume any observability property. This is not so surprising since the observability property relies on its full description (including the kinetics) which is not used to build the mass balance observer. In fact it is not really an observer *stricto sensu*, but more precisely a detector, relying on hypothesis that the non observable part is stable.

5.2 Definitions, hypotheses

In this chapter we will consider the biotechnological processes that are modelled with a mass balance model:

$$\frac{d\xi}{dt} = Kr(\xi) - D(t)\xi + D(t)\xi_{in}(t) - Q(\xi) \tag{33}$$

with

$$\xi \in \mathbb{R}^n \quad r \in \mathbb{R}^p \tag{34}$$

We assume that the set of available measurements y can be decomposed into three vectors:

$$y = [y_1 \; y_2 \; y_3]^T \tag{35}$$

where:

- y_1 is a set of q measured state variables. To simplify the notations, we will order the components of the state so that, y_1 corresponds to the q first components of ξ.
- y_2 represents the measured gaseous flow rates: $y_2 = Q(\xi)$
- y_3 represents the other available measurements (pH, conductivity,...) that are related to the state through the following relationship: $y_3 = h(\xi)$

Let us rewrite system (33) after splitting the measured part ($\xi_1 = y_1$) from the other part of the state (ξ_2).

$$\frac{d\xi_1}{dt} = K_1 r(\xi) - D\xi_1 + D\xi_{in1} - Q_1(\xi) \tag{36}$$

$$\frac{d\xi_2}{dt} = K_2 r(\xi) - D\xi_2 + D\xi_{in2} - Q_2(\xi) \tag{37}$$

Matrices K_1 and K_2, vectors ξ_{in1}, ξ_{in2}, Q_1 and Q_2 are such that

$$K = \begin{pmatrix} K_1 \\ K_2 \end{pmatrix},\ \xi_{in} = \begin{pmatrix} \xi_{in1} \\ \xi_{in2} \end{pmatrix},\ Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$$

5.3 The asymptotic observer

In order to build the asymptotic observers we need the two following technical hypotheses:

Hypothesis 4 *(i) There are more measured quantities than reactions: $q \geq p$. (ii) Matrix K_1 is of full rank.*

Hypothesis 4(ii) means that a non zero r cannot cancel $K_1 r$ (a reaction can not compensate the other ones with respect to the measured variables).

Consequences: under Hypothesis 4, the $q \times p$ matrix K_1 admits a left inverse; there exists a $p \times q$ matrix G such that:

$$GK_1 = I_{p\times p} \tag{38}$$

Let us set: $A = -K_2 G$, and let us consider the following linear change of coordinates:

$$\zeta_1 = \xi_1 \tag{39}$$
$$\zeta_2 = A\xi_1 + \xi_2 \tag{40}$$

this change of variable transforms (36) and (37) into:

$$\frac{d\zeta_1}{dt} = K_1 r(T\zeta) - D\zeta_1 + D\zeta_{in1} - Q_1(T\zeta) \tag{41}$$
$$\frac{d\zeta_2}{dt} = -D(\zeta_2 - \zeta_{in2}) - (AQ_1(T\zeta) + Q_2(T\zeta)) \tag{42}$$

with

$$T = \begin{pmatrix} I_p & 0_{p,n-p} \\ -A & I_{n-p} \end{pmatrix},\ M = \begin{pmatrix} A & I_{n-p} \end{pmatrix} \tag{43}$$

and

$$\zeta_{in2} = M\xi_{in} \tag{44}$$

The equation of ζ_2 can be rewritten using the output y_2:

$$\frac{d\zeta_2}{dt} = -D(\zeta_2 - \zeta_{in2}) - My_2 \tag{45}$$

Remark: System (45) is a linear system up to an output injection (cf. Section 3.2).

We can now design an observer for this system, simply after copying equations (45). But we must first state an hypothesis to guarantee the observer convergence:

Hypothesis 5 *The positive scalar variable D is a regularly persisting input i.e. there exists positive constants c_1 and c_2 such that, for all time instant t:*

$$0 < c_1 \leq \int_t^{t+c_2} D(\tau)d\tau$$

In practice, c_2 must be low with respect to the time constant of the system. Moreover $\frac{c_1}{c_2}$ must be high because it determines the minimal converging rate of the observer.

Lemma 1 *(see [3]) Under Hypothesis* **??**, *the solution $\hat{\xi}_2$ of the following asymptotic observer:*

$$\begin{array}{l} \dfrac{d\hat{\zeta}_2}{dt} = -D(\hat{\zeta}_2 - \zeta_{in2}) - My_2 \\ \hat{\xi}_2 = \hat{\zeta}_2 - Ay_1 \end{array} \tag{46}$$

converges asymptotically toward the solution ξ_2 of the reduced system (37).

Proof: it can be easily verified that the estimation error $e_2 = \hat{\xi}_2 - \xi_2 = \hat{\zeta}_2 - \zeta_2$ satisfies:

$$\frac{de_2}{dt} = -De_2. \tag{47}$$

and converges asymptotically toward ξ_2 if Hypothesis **??** is fullfiled. [3].

5.4 Example

We will consider as example the growth of the filamentous fungi *Pycnoporus cinnabarinus* (X) [19]. The fungi uses two substrates to grow: glucose as carbon source (C) and ammonium as nitrogen source (N). The reaction scheme is assumed to be composed by one reaction:

$$N + C \longrightarrow X$$

The model is then of the type (33), with:

$$\xi = [N\ C\ X]^T,\ K = [-k_1\ -k_2\ 1]^T,\ \xi_{in} = [N_{in},\ C_{in},\ 0]^T$$

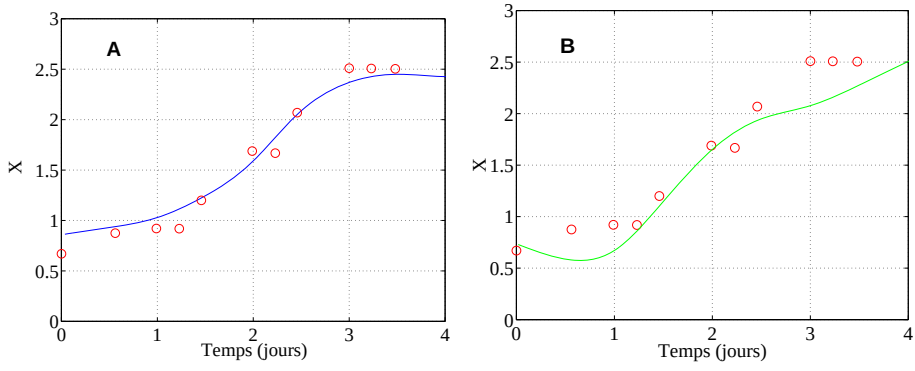


Figure 3: Comparison between direct biomass measurements of *Pycnoporus cinnabarinus* (o) and observer predictions based on the nitrogen measurement (A) or on the carbon measurement (B)

The following measurements are available: $y_1 = [N\ C]^T$.

The state partition will then be the following:

$$\xi_1 = [N\ C]^T,\ \xi_2 = X$$

associated with:

$$K_1 = [-k_1\ \ -k_2]^T,\ K_2 = 1$$

Matrix K_1 has an infinite number of left inverses. We will consider two of them: $G_1 = [-\frac{1}{k_1},\ 0]$ and $G_2 = [0,\ -\frac{1}{k_2}]$. These two matrices will naturally lead to two observers. The first one based on the nitrogen measurements:

$$\begin{array}{l} \dfrac{d\hat{\zeta}_2^1}{dt} = -D(\hat{\zeta}_2^1 - \dfrac{N_{in}}{k_1}) \\ \hat{X}^1 = \hat{\zeta}_2^1 - \frac{N}{k_1} \end{array} \tag{48}$$

and the other one based on carbon:

$$\begin{array}{l} \dfrac{d\hat{\zeta}_2^2}{dt} = -D(\hat{\zeta}_2^2 - \dfrac{C_{in}}{k_2}) \\ \hat{X}^2 = \hat{\zeta}_2^2 - \frac{C}{k_2} \end{array} \tag{49}$$

The results of these observers obtained with experimental data are presented in Figure 5.4. In this case, the observer based on the nitrogen measurements is more reliable.

5.5 Improvements

The asymptotic observers work in open loop. Indeed, their estimate relies on the mass balances and are not corrected by a discrepancy between measured and estimated quantities. It assumes that the mass balance model is ideal. Nevertheless, the yield parameters are difficult to estimate properly, and in some cases (wastewater treatment) the mass inputs in the system are not precisely known. In this case it can be dangerous to base the observer only on the mass balance model without taking into account some measurements on the system that reflect its actual state. It can be possible to on-line estimate these unknown parameters, but we will see here another method aiming at improving the observer robustness with respect to some uncertainties.

In this paragraph, we will see how to exploit the available measurements y_3 to improve the asymptotic observer performances.

We assume here that $y_3 \in \mathbb{R}$. We define the mapping h:

$$h : (\xi_1, \xi_2) \in (\mathbb{R}^p \times \mathbb{R}^{n-p}) \longrightarrow y_3 = h(\xi_1, \xi_2) \in \mathbb{R}$$

We suppose that h satisfies the following hypothesis:

Hypothesis 6 *the mapping h is monotonous with respect to ξ_2, i.e.:$\frac{Dh}{D\xi_2}$ is of fixed sign on the considered domain Ω.*

Example: in the example detailed hereafter, $h(\xi_1, \xi_2) = \alpha S + \beta P$, and thus:

$$\frac{Dh}{D\xi_2} = [\alpha \ \beta] \tag{50}$$

which is of fixed sign. Of course, h can be nonlinear.

Proposition 1 *Let $\lambda \in \mathbb{R}^p$ be a unitary constant vector ($\|\lambda\| = 1$), whose signs are chosen such that* $\text{sign}(\lambda) = \text{sign}(\frac{Dh}{D\xi_2})$*, θ is a positive scalar (which can depend on time) and $z_{in} = M\xi_{in}$. The following system:*

$$\frac{d\hat{z}}{dt} = -D(\hat{z} - z_{in}) - M\, y_2 - \theta\lambda(h(y_1, \hat{z} - Ay_1) - y_3) \tag{51}$$

is an asymptotic observer of the reduced system (45).

For the proof of this property, and for more details, we invite the reader to refer to [20].

Example: we will consider a bacterial biomass (X) growing in a bioreactor. The micro-organisms uptake the substrate S and metabolise a product P:

$$S \longrightarrow X \ + \ P$$

The associated model is then:

$$\left\{ \begin{array}{l} \frac{dX}{dt} = r(\xi) - DX \\ \frac{dS}{dt} = -c_1 r(\xi) + D(S_{in} - S) \\ \frac{dP}{dt} = c_2 r(\xi) - DP \end{array} \right. \tag{52}$$

where S_{in} is the influent substrate, c_1 and c_2 are the yield coefficients.

We assume that the bacterial biomass and the conductivity of the solution can be measured. The conductivity is related to a positive linear combination of the ions in the liquid i.e. S and P. We have therefore:

$$y_1 = X \tag{53}$$

$$y_2 = (0,0,0)^t \tag{54}$$

$$y_3 = \alpha S + \beta P \tag{55}$$

We suppose that the substrate concentration in the influent S_{in} is not precisely known, and we will use an estimate denoted $\hat{S}_{in}$.

Thanks to Proposition 1 we can design the following observer (for sake of clarity we choose $\lambda = [1\,0]$ which satisfies the right hypotheses).

$$\left\{ \begin{array}{l} \frac{d\hat{z}_1}{dt} = D(\hat{z}_{in1} - \hat{z}_1) - \theta(\alpha\hat{S} + \beta\hat{P} - y_3) \\ \frac{d\hat{z}_2}{dt} = -D\hat{z}_2 \\ \hat{S} = \hat{z}_1 - \frac{y_1}{c_1} \\ \hat{P} = \hat{z}_2 + \frac{y_1}{c_2} \end{array} \right. \tag{56}$$

with $\hat{z}_{in1} = \hat{S}_{in}$.

let us show now the robustness properties when the estimate S_{in} is false. If $S^\star$ and $P^\star$ represent equilibrium values of S and P, we denote $\hat{S}^\star$ and $\hat{P}^\star$ the equilibrium values for the closed loop observer. If the observer is in open loop ($\theta = 0$), using $\hat{S}_{in}$, a direct computation provides:

$$\hat{S}^\star = S^\star + \hat{S}_{1in} - S_{1in} \tag{57}$$

The prediction error is thus exactly the error on S_{1in}. With the closed loop observer, the steady state is:

$$\hat{S}^\star = S^\star + \frac{D}{\theta\alpha + D}(\hat{S}_{1in} - S_{1in}) \tag{58}$$

If the gain θ is high, it is easy to see that $\hat{S^\star} \simeq S^\star$: the bias is reduced by the closed loop observer.

6 interval observers

The usual observers implicitly assume that the model is a good approximation of the real system. Nevertheless, we have seen that a model of a bioprocess is often poorly known. In this case, the observation principle must be revisited: generally it will no more be possible to build an exact observer (which would guaranty: $\|e(t)\| = \|\hat{x}(t) - x(t)\| \to 0$ when $t \to \infty$) whose convergence rate could be tuned (as for example an exponential observer). Therefore, in this case the result must be weaken.

We present here a possible way (among others) consisting in bounding the uncertainty on the model. The bound on the variable to be estimated is deduced. To simplify, we will first present the linear (or close to linear) case (see [21, 22]).

6.1 Principle

The idea is to use the known dynamical bounds of the uncertainties:

The dynamical bounds on the model uncertainties allow to derive (in the good cases) the dynamical bounds on the state variable to be estimated.

Figure 4 summarises the philosophy of the interval estimation.

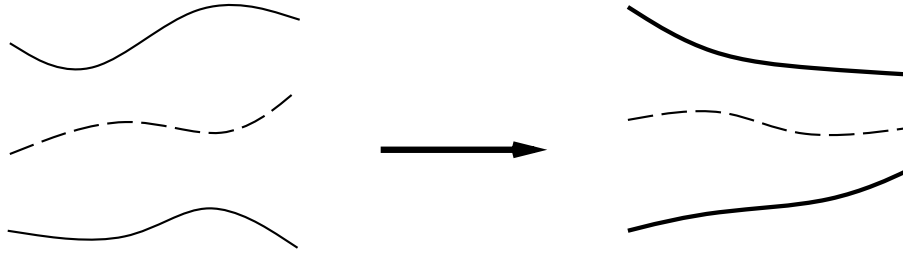


Figure 4: Principle of interval estimation for bounded uncertainties: *a priori* bounds on the uncertainties U provide bounds on the non measured state X

Let us consider the general system:

$$(\mathcal{S}_0) \begin{cases} \frac{dx}{dt}(t) = f(x(t), u(t), w(t)) \quad ; \quad x(t_0) = x_0 \\ y(t) = h(x(t), v(t)) \end{cases} \tag{59}$$

where $x \in \mathbb{R}^n$ is the state vector $y \in \mathbb{R}^p$ is the output vector, $u \in \mathbb{R}^m$ the input vector, x_0 the initial condition at t_0, $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \longrightarrow \mathbb{R}^n$ and $h : \mathbb{R}^n \times \mathbb{R}^s \longrightarrow \mathbb{R}^p$.

The unknown quantities $w \in \mathbb{R}^r$ and $v \in \mathbb{R}^s$ are characterised by their upper and lower bounds:

$$w^-(t) \leq w(t) \leq w^+(t) \quad \forall\ t \geq t_0 \tag{60}$$
$$v^-(t) \leq v(t) \leq v^+(t) \quad \forall\ t \geq t_0 \tag{61}$$

Remark: the operator $\leq$ applies to vectors, it corresponds to inequalities between each component.

Based on the fixed model structure $(\mathcal{S}_0)$ and on the set of known variables, a dynamical auxiliary system can be designed as follows:

$$(\mathcal{O}_0) \begin{cases} \frac{dz}{dt}^- = f^-(z^-, z^+, u, y, w^-, w^+, v^-, v^+) \quad ; \quad z^-(t_0) = g^-(x_0^-, x_0^+) \\ \frac{dz}{dt}^+ = f^+(z^-, z^+, u, y, w^-, w^+, v^-, v^+) \quad ; \quad z^+(t_0) = g^+(x_0^-, x_0^+) \\ x^- = h^-(z^-, z^+, u, y, w^-, w^+, v^-, v^+) \\ x^+ = h^+(z^-, z^+, u, y, w^-, w^+, v^-, v^+) \end{cases} \tag{62}$$

with $z^-, z^+ \in \mathbb{R}^q$, the other functions being defined in the appropriate domains.

Definition 8 (interval estimator) *System $(\mathcal{O}_0)$ is an interval estimator of system $(\mathcal{S}_0)$ if for any pair of initial conditions $x_0^- \leq x_0^+$, there exists bounds $z^-(t_0)$, $z^+(t_0)$ such that the coupled system $(\mathcal{S}_0, \mathcal{O}_0)$ verifies:*

$$x^-(t) \leq x(t) \leq x^+(t) \ ; \ \forall\ t \geq t_0 \tag{63}$$

The interval estimator comes from the coupling between two estimators providing each an under-estimate $x^-(t)$ and an over-estimate $x^+(t)$ of $x(t)$, The estimator provides a dynamical interval $[x^-(t)\ ,\ x^+(t)]$ containing the unknown value $x(t)$ (FIG. 4).

Of course, this interval can be very large and therefore useless. The next step consists in trying to reduce as far as possible this interval and increase the convergence rate toward this interval, for example with an exponential convergence rate. Then, we move back to classical observation problems, with the important difference that we don't require the observation error (the interval amplitude) to tend asymptotically *exactly* toward zero

6.2 The linear case up to an output injection

First, let us take a very simple case. We consider again the following system:

$$(\mathcal{S}): \quad \left\{ \begin{array}{lcl} \frac{dx}{dt}(t) & = & Ax(t) + \phi(t, y(t)) \\ y(t) & = & Cx(t) \end{array} \right. \tag{64}$$

with $A \in \mathcal{M}^{n\times n}(\mathbb{R})$ $(n \geq 2)$, $C \in \mathcal{M}^{1\times n}(\mathbb{R})$. If the mapping $\phi : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}^n$ is known, a Luenberger observer can be designed (Section 3.1).

What happens now if function ϕ is badly known ? We assume that it can be bounded and that the bounds are known. Thus, the functions $\phi^-, \phi^+ : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}^n$, are known, sufficiently smooth, such that:

$$\phi^-(t, y) \leq \phi(t, y) \leq \phi^+(t, y), \quad \forall (t, y) \in \mathbb{R}^+ \times \mathbb{R} \tag{65}$$

Then we will use these bounds to design an upper and a lower estimator:

$$\frac{dx^+}{dt}(t) = Ax^+(t) + \phi^+(t, y(t)) + K(Cx^+(t) - y(t)) \tag{66}$$

$$\frac{dx^-}{dt}(t) = Ax^-(t) + \phi^-(t, y(t)) + K(Cx^-(t) - y(t)). \tag{67}$$

Let us consider now the "upper" error $e^+(t) = x^+(t) - x(t)$, we have:

$$\frac{de^+}{dt} = (A + KC)e^+ + b^+(t)$$

with

$$b^+(t) = \phi^+(t, y(t)) - \phi(t, y(t)).$$

It follows that b^+ is positive, and the following Lemma can be easily proven:

Lemma 2 *If the elements of matrix $(A + KC)$ are positive outside the diagonal (the matrix is said cooperative), then $e^+(0) \geq 0$ implies $e^+(t) \geq 0$ for any positive t.*

Of course, we have the same Lemma for the lower error: $e^-(t) = x(t) - x^-(t)$ and the total error $e(t) = e^-(t) + e^+(t)$. The following theorems can be deduced:

Theorem 5 *If the gains of vector K can be chosen such that matrix $(A + KC)$ is cooperative, and if we have an initial estimate such that*

$$x^-(0) \leq x(0) \leq x^+(0)$$

then equations (66),(67) provide an interval estimator for system (64).

Theorem 6 *If hypotheses of Theorem 5 are verified, if matrix* $(A+KC)$ *is stable, and if moreover the error on* ϕ *can be bounded, i.e. if we have:*

$$b(t) = \phi^+(t,y) - \phi^-(t,y) \leq B$$

where B *is a positive constant, then the error* $e(t)$ *converges asymptotically toward an interval smaller (for each component) than the positive vector:*

$$e_{max} = -(A+KC)^{-1}B$$

In particular, if the components of e_{max} *are zero, then the corresponding components for* $e(t)$ *converge toward zero.*

The proofs are straightforward; the proof of the first theorem follows directly from Lemma 2. The proof of the second theorem is due to the differential inequality

$$(A+KC)e + b(t) \leq (A+KC)e + B$$

which implies (with equal initial conditions)

$$e(t) \leq e_m(t), \quad \forall t \geq 0$$

where $e_m(t)$ is the solution of $\frac{de_m}{dt} = (A+KC)e_m + B$.

Remarks:

- We use in the observer design the fundamental hypothesis that it is possible to derive inequalities between the variables from inequalities on the left hand side of the differential equations. This hypothesis is connected with the comparison of the solutions of differential equations (see appendix). There exists other techniques to estimate the intervals, they are more precise but less explicit [23].

- We need also the assumption that the initial estimate is valid

  $$x^-(0) \leq x(0) \leq x^+(0);$$

  a large estimate can be chosen in practice.

- The problem of the tuning of the convergence rate has not been considered here; is it possible to choose a gain K that will ensure cooperativity, stability, and arbitrary convergence rate ? This is a complicated problem, we invite the reader to consult [22] for more details.

We illustrate this approach with an example of such an estimator for a biochemical process (see [24, 25]).

7 Interval estimator for an activated sludge process

We consider a very simplified model of an activated sludge process, used for the biological wastewater treatment. The objective is to process a wastewater, with an influent flow rate Q_{in} and a pollutant (substrate) concentration s_{in}. We want that the concentration of the effluent is lower than s_{out}. The process is composed by an aerator (bioreactor) followed by a settler separating the liquid and solid (biomass) phases. Then we recycle a part of the biomass toward the aerator. Let us denote x, s, and x_r the three state variables of this simple model, representing respectively the biomass and substrate concentrations in the aerator, and the recycled biomass in the settler. Q_{in}, Q_{out}, Q_r and Q_w are the flow rates, V_a and V_s the volumes (see Figure 5). We suppose that the biological reactions only take place in the aerator.

$$\frac{1}{Y} s \xrightarrow{\mu(.)x} x$$

Y is a yield coefficient, and $\mu(.)$ the bacterial growth rate. If we take into account the biomass recycling, we get:

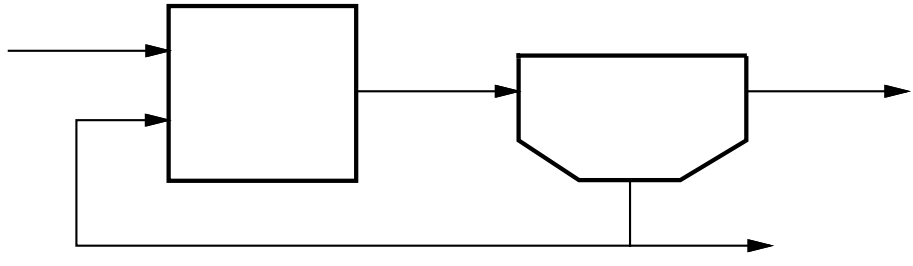


Figure 5: Diagram of an activated sludge process

$$\begin{cases} \dfrac{dx}{dt} = \mu(\cdot)x - (1+r)D(t)x + rD(t)x_r \\ \dfrac{ds}{dt} = -\dfrac{\mu(\cdot)x}{Y} - (1+r)D(t)s + D(t)s_{in}(t) \\ \dfrac{dx_r}{dt} = v(1+r)D(t)x - v(w+r)D(t)x_r \end{cases} \tag{68}$$

with

$$D(t) = \frac{Q_{in}}{V_a} \quad ; \quad r = \frac{Q_r}{Q_{in}} \quad ; \quad w = \frac{Q_w}{Q_{in}} \quad ; \quad v = \frac{V_a}{V_s}$$

We state the following hypotheses:

- We measure only s and we want to estimate x and x_r.

- We assume that $\mu(.)$ is not known, and we want to design an asymptotic observer (see section 5.3)

- We know a bounding (even very loose) of the initial conditions for x and x_r.

- The substrate input for s_{in} fluctuates but is not known. However we know dynamical bounds for $s_{in}(t)$:

$$s_{in}^{-}(t) \leq s_{in}(t) \leq s_{in}^{+}(t) \quad \forall\, t \geq 0.$$

These hypotheses correspond to what happens in a urban wastewater treatment plant. The influent varies but is not measured. But it can be bounded by two functions corresponding to human activities. These bounds will probably evolve with respect to seasons.

We will design an asymptotic interval estimator, which will provide bounds for the variables to be estimated. First we perform a change of variable to eliminate $\mu(.)$ (cf. section 5.3):

$$Z = X + \begin{bmatrix} Y \\ 0 \end{bmatrix} s; \quad Z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \quad ; \quad X = \begin{bmatrix} x \\ x_r \end{bmatrix} \tag{69}$$

and we get the 2-dimensional system:

$$\frac{dZ}{dt} = D(t)\Big(A\ Z + B(s,t)\Big) \quad ; \quad Z_0 = \begin{bmatrix} x_0 + Y s_0 \\ x_{r0} \end{bmatrix}$$

$$A = \begin{bmatrix} -(1+r) & r \\ v(1+r) & -v(w+r) \end{bmatrix} \quad ; \quad B(s,t) = \begin{bmatrix} Y s_{in}(t) \\ -Y v(1+r)s \end{bmatrix}$$

We can now build two estimators (upper and lower) which use unknown bounds on the influent concentration s_{in}:

$$\begin{cases} \frac{d\hat{Z}^+}{dt} = D(t)\Big(A\ \hat{Z}^+ + B^+(s,t)\Big) \quad ; \quad \hat{Z}^+(0) = X_0^+ + \begin{bmatrix} Y \\ 0 \end{bmatrix} s_0 \\ \\ \frac{d\hat{Z}^-}{dt} = D(t)\Big(A\ \hat{Z}^- + B^-(s,t)\Big) \quad ; \quad \hat{Z}^-(0) = X_0^- + \begin{bmatrix} Y \\ 0 \end{bmatrix} s_0 \\ \\ \hat{X}^+ = \hat{Z}^+ - \begin{bmatrix} Y \\ 0 \end{bmatrix} s \\ \\ \hat{X}^- = \hat{Z}^- - \begin{bmatrix} Y \\ 0 \end{bmatrix} s \end{cases} \tag{70}$$

with $B^+(s,t) = \begin{bmatrix} s_{in}^+(t) \\ -v(1+r)s \end{bmatrix} Y \quad ; \quad B^-(s,t) = \begin{bmatrix} s_{in}^-(t) \\ -v(1+r)s \end{bmatrix} Y$

In this simple case, the convergence rate is fixed by the system. On Figure 6, we have represented the influent concentration and its bounds, the measurement s and the two estimates with the bounds. This very simple observer illustrates how to take into account the knowledge on the dynamical bounds.

Under additional hypotheses, it can be shown that this observer can be tuned [24].

8 Conclusion

We have seen a set of methods to design an observer for a bioreactor. Other techniques exist and we do not pretend to be exhaustive. Let us mention for example the methods based on neural networks, where the system and its observer are estimated at the same time. The convergence rate of the obtained neural network can not be tuned.

The presented observers assumed that the model parameters were known. In some cases these parameters can evolve. Algorithms to estimate the parameters must then be used, they lead to adaptive observers. Of course in that case, the convergence of the full system observer-parameter estimator must be demonstrated.

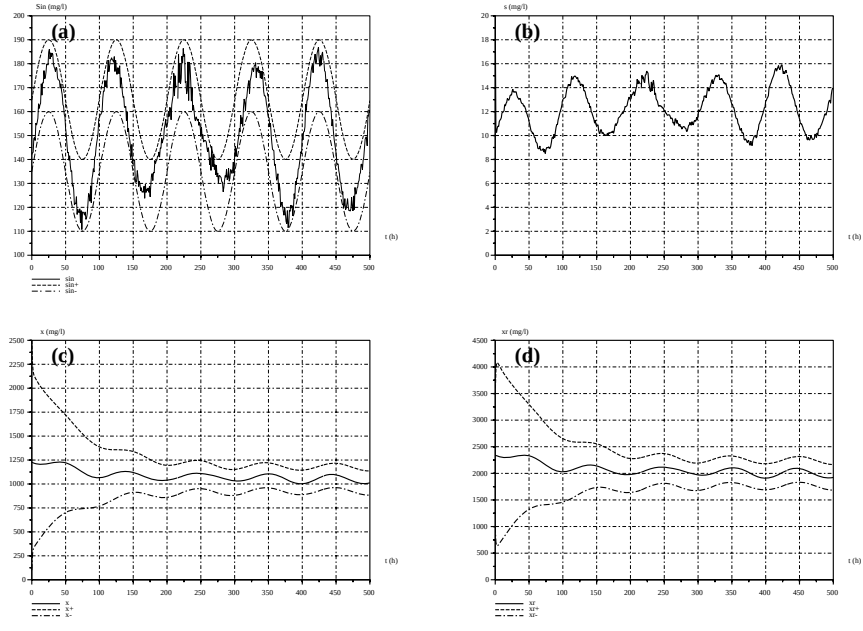


Figure 6: Interval observer. (a) Influent bounds, (b) Measurements of S, (c) interval estimations of X, (d) interval estimations of X_r

The choice of the type of observer to be found must be made above all by considering the reliability of the model and of the available measurements. A triple trade-off must then be managed between robustness with respect to modelling uncertainties, robustness with respect to disturbances and convergence rate.

Finally, to implement an observer in a computer, a discretising phase is required. This step will be based on Euler type algorithms. This step is not difficult, but it requires care. In particular, if the sampling rate is too high for the discretisation rate, continuous/discrete observers must be used [26].

To conclude, we insist that the observers must first be validated before they can be used. For this, there predictions must extensively be compared to direct measurements that were not used during the calibration process.

Appendix: a comparison theorem

We propose here a general theorem in the non linear case. It can be useful to apply the interval estimation techniques. The reader is invited to consult [35] for the details and for a general presentation.

Definition 9 *A non linear system in dimension n is said to be cooperative if its Jacobian matrix is positive outside its diagonal on a convex domain.*

We propose now the comparison theorem between $x(t)$ and $y(t)$ defined by the two systems

$$\begin{aligned} \frac{dx}{dt} &= f(x,t) \quad ; \quad x(0) = x_0 \\ \frac{dy}{dt} &= g(y,t) \quad ; \quad y(0) = y_0 \end{aligned}$$

where $f, g : \; U \times \mathbb{R}^+ \to \mathbb{R}^n$ are sufficiently regular on a convex domain $U \subset \mathbb{R}^n$.

Property 3 *If*
- $\forall z \in U, \; \forall t \geq 0, \; f(z,t) \leq g(z,t)$
- *g is cooperative*
- $x_0 \leq y_0$

then $x(t) \leq y(t)$ for $t > 0$

The inequalities must be considered for each element. It means that, for a cooperative system the order between two solutions is conserved for any time. This property is fundamental for the set up of interval estimators.

References

[1] T. Kailath, *Linear Systems.* London: Prentice-Hall, Inc., Englewood Cliffs, N.J., 1980.

[2] D. G. Luenberger, *Introduction to dynamic systems : theory, models and applications.* Wiley, 1979.

[3] G. Bastin and D. Dochain, *On-line estimation and adaptive control of bioreactors.* Elsevier, 1990.

[4] T. Basar and P. Bernhard, *H^∞-Optimal Control ans Related Minimax Design Problems : a Dynamic Game Approach.* Boston: Birkhaser, 1991.

[5] M. Darouach, M. Zasadzinski, and S. J. Xu, "Full-order observers for linear systems with unknown inputs," *IEEE Trans. Automat. Contr.*, vol. AC-39, pp. 606–609, 1994.

[6] P. Kudva, N. Viswanadham, and A. Ramakrishna, "Observers for linear systems with unknown inputs," *IEEE Trans. Autom. Contr.*, vol. AC-25, no. 1, pp. 113–115, 1980.

[7] Edwards, Christopher and Spurgeon, Sarah K., "On the development of discontinuous observers.," *Int. J. Control 59, No.5, 1211-1229 (1994).*, 1994.

[8] D. Luenberger, "Observers for multivariable systems," *IEEE Trans. Autom. Contr.*, vol. 11, pp. 190–197, 1966.

[9] S. Beale and B. Shafai, "Robust control system design with a proportional integral observer," *Int. J. Contr.*, vol. 50, no. 1, pp. 97–111, 1989.

[10] Anderson, Brian D.O. and Moore, John B., *Optimal control. Linear quadratic methods.* Prentice Hall, Englewood Cliffs, NJ, 1990.

[11] J. Gauthier and G. Bornard, "Observability for any $u(t)$ of a class of nonlinear systems," *IEEE Trans. Autom. Contr.*, vol. 26, no. 4, pp. 922–926, 1981.

[12] J. P. Gauthier, H. Hammouri, and S. Othman, "A simple observer for nonlinear systems applications to bioreactors," *IEEE Trans. Autom. Contr.*, vol. 37, pp. 875–880, 1992.

[13] F. Deza, E. Busvelle, J. Gauthier, and D. Rakotopara, "High gain estimation for nonlinear systems," *System and control letters*, vol. 18, pp. 292–299, 1992.

[14] R. Dugdale, "Nutrient limitation in the sea: dynamics, identification and significance," *Limnol. Oceanogr.*, vol. 12, pp. 685–695, 1967.

[15] M. Droop, "Vitamin B12 and marine ecology. IV. the kinetics of uptake growth and inhibition in Monochrysis lutheri," *J. Mar. Biol. Assoc.*, vol. 48, no. 3, pp. 689–733, 1968.

[16] O. Bernard, G. Sallet, and A. Sciandra, "Nonlinear observers for a class of biological systems. Application to validation of a phytoplanktonic growth model," *IEEE Trans. Autom. Contr.*, vol. 43, pp. 1056–1065, 1998.

[17] O. Bernard, G. Sallet, and A. Sciandra, "Use of nonlinear software sensors to monitor the internal state of a culture of microalgae," in *Proceedings of the IFAC World Congress*, vol. L, pp. 145–150, Beijing, China, 1999.

[18] M. Hou and P. Mller, "Design of observers for linear systems with unknown inputs," *IEEE Trans. Autom. Contr.*, vol. AC-37, no. 6, pp. 871–875, 1991.

[19] O. Bernard, G. Bastin, C. Stentelaire, L. Lesage-Meessen, and M. Asther, "Mass balance modelling of vanillin production from vanillic acid by cultures of the fungus *pycnoporus cinnabarinus* in bioreactors," *Biotech. Bioeng*, pp. 558–571, 1999.

[20] O. Bernard, J.-L. Gouz, and Z. Hadj-Sadok, "Observers for the biotechnological processes with unknown kinetics. application to wastewater treatment," in *Proceedings of CDC 2000*, Sydney, Australia, 2000.

[21] J. L. Gouzé, A. Rapaport, and Z. Hadj-Sadok, "Interval observers for uncertain biological systems," *Ecological modelling*, vol. 133, pp. 45–56, 2000.

[22] A. Rapaport and J. Gouz, "Practical observers for uncertain affine outputs injection systems," in *Proceedings of ECC99 (CDROM)*, Karlsruhe, Germany, 1999.

[23] M. Kieffer, *Estimation ensembliste par analyse par intervalles : application à la localisation d'un véhicule.* PhD thesis, Université de Paris-Sud, 1999.

[24] M. Z. Hadj-Sadok and J. L. Gouzé, "Bounds estimation for uncertain models of wastewater treatment," *IEEE International Conf. on Control and Applications, Trieste, Italy*, pp. 336–340, 1998.

[25] V. Alcaraz-Gonzalez, A. Genovesi, J. Harmand, A. Gonzalez, A. Rapaport, and J.-P. Steyer, "Robust exponential nonlinear observers for a class of lumped models useful in chemical and biochemical engineering - Application to a wastewater treatment," *International Workshop on Application of Interval Analysis to Systems and Control, Girona, Espagne*, pp. 225–235, 1999.

[26] M. Pengov, *Application des observateurs non-linéaires à la commande des bioprocédés.* PhD thesis, Université de Metz, 1998.

[27] K. Busawon, A. ElAssoudi, and H. Hammouri, "Dynamical output feedback stabilization od a clas of nonlinear systems," in *32th CDC San Antonio*, pp. 1966–1971, 1993.

[28] F. Deza, E. Busvelle, J. Gauthier, and D. Rakotopara, "Exponential-lyconverging observers for distillation columns," *Chemical Enginering Science*, vol. 47, no. 4, pp. 3935–3941, 1992.

[29] D. Burmaster, "The unsteady continuous culture of phosphate-limited Monochrysis lutheri Droop: experimental and theoritical analysis," *J. Exp. Mar. Biol. Ecol.*, vol. 39, no. 2, pp. 167–186, 1979.

[30] F. Deza, D. Bossanne, E. Busvelle, J. Gauthier, and D. Rakotopara, "Exponential observers for nonlinear systems," *IEEE trans. autom. contr.*, vol. 38, no. 3, pp. 482–484, 1993.

[31] M. Darouach, "On the novel approach to the design of the unknown input observers," *IEEE Trans. Autom. Cont.*, vol. 39, no. 3, pp. 698–699, 1994.

[32] M. Darouach, M. Zasadzinski, and S.-J. Xu, "Full-order observers for linear systems with unknown inputs," *IEEE Trans. Autom. Cont.*, vol. 39, no. 3, pp. 1068–1072, 1994.

[33] M. Darouach, M. Zasadzinski, and M. Hayar, "Reduced-order observer design for descriptor systems with unknown inputs," *IEEE Trans. Autom. Cont.*, vol. 41, no. 7, pp. 1068–1072, 1996.

[34] J. Gauthier and I. A. K. Kupka, "Observability and observers for nonlinear systems.," *SIAM J. Control and Optimization*, vol. 32, no. 4, pp. 975–994, 1994.

[35] H. L. Smith, *Monotone Dynamical Dystems: an Introduction to the Theory of Competitive and Cooperative Systems.* Providence, Rhode Island: American Mathematical Society, 1995.

[36] D'Andrea-Novel, Brigitte and Cohen de Lara, Michel, *Commande lineaire des systemes dynamiques. (Linear control of dynamic systems).* Paris: Masson, 1994.

[37] Fossard, A.J.(ed.) and Normand-Cyrot, D.(ed.) and Mouyon, Ph.(ed.), *Systemes non lineaires. 1. Modélisation - Estimation.* Paris: Masson, 1993.

[38] J. Andrews, "A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrate," *Biotechnol. & Bioeng.*, vol. 10, pp. 707–723, 1968.

[39] R. Bajpai and M. Reuß, "Evaluation of feeding strategies in carbon-regulated secondary metabolite production through mathematical modelling," *Biotechnol. & Bioeng.*, vol. 23, pp. 717–738, 1981.

[40] J. E. Bailey and D. F. Ollis, *Biochemical engineering fundamentals.* McGraw-Hill, 1986.

[41] O. Bernard, Z. Hadj-Sadok, D. Dochain, A. Genovesi, and J.-P. Steyer, "Dynamical model development and parameter identification for an anaerobic wastewater treatment process," *Biotech.Bioeng.*, to appear.

[42] O. Bernard and J. Gouzé, "Transient behavior of biological loop models with application to the droop model," *Mathematical Biosciences*, vol. 127, pp. 19–43, 1995.

[43] L. Chen, O. Bernard, G. Bastin, and P. Angelov, "Hybrid modelling of biotechnological processes using neural networks," *Contr. Eng. Prcatice*, vol. 8, pp. 821–827, 2000.

[44] D. Dochain, *On line parameter estimation, adaptive state estimation and control of fermentation processes.* PhD thesis, Universit Catholique de Louvain-la-Neuve, Belgique, 1986.

[45] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation.* Addison-Wesley, 1991.

[46] R. Horn and C. Johson, *Matrix analysis.* Cambridge Univesity Press, 1992.

[47] H. Khalil, *Nonlinear Systems.* Macmillan Publishing Company, 1996.

[48] A. Karama, O. Bernard, A. Genovesi, D. Dochain, A. Benhammou, and J.-P. Steyer, "Hybrid modelling of anaerobic wastewater treatment processes," *Wat. Sci. Technol.*, vol. 43, no. 1, pp. 43–50, 2001.

[49] J. Lee and D. Meyrick, "Gas-liquid interfacial areas in salt solutions in an agitated tank," *Trans. Instn. Chem. Engrs.*, vol. 48, pp. T37–T45, 1970.

[50] H. Lim, Y. Tayeb, J. Modak, and P. Bonte, "Computational algorithms for optimal feed rates for a class of fed-batch fermentation: numerical results for penicillin and cell mass production," *Biotechnol. & Bioeng.*, vol. 28, pp. 1408–1420, 1986.

[51] B. Li, "Global asymptotic behavior of the chemostat: General response functions and different removal rates," *SIAM Journal*, vol. 59, 1998.

[52] J. Merchuk, "Further considerations on the enhancement factor for oxygen absorption into fermentation broth," *Biotechnol. & Bioeng.*, vol. 19, pp. 1885–1889, 1977.

[53] J. Monod, *Recherches sur la croissance des cultures bactriennes.* Paris, France: Hermes, 1942.

[54] A. Schumpe, "Gas solubilities in biomedia," *Advances in Biochemical Engineering/Biotechnology*, vol. 2, pp. 159–170, 1985.

[55] H. Smith and P. Waltman, *The theory of the chemostat: dynamics of microbial competition.* Cambridge University Press, 1995.

[56] K.-W. Wang, B. Baltzis, and G. Lewandowski, "Kinetics of phenol biodegradation in the presence of glucose," *Biotechnol. & Bioeng.*, vol. 51, pp. 87–94, 1996.

[57] E. Sacks, "A dynamic systems perspective on qualitative simulation," *Artif. Intell.*, vol. 42, pp. 349–362, 1990.

[58] O. Bernard and J.-L. Gouzé, "Transient behavior of biological loop models, with application to the Droop model," *Mathematical Biosciences*, vol. 127, no. 1, pp. 19–43, 1995.

[59] R. M. Nisbet and W. S. C. Gurney, *Modelling fluctuating populations.* Wiley, 1982.

[60] R. Thomas, "On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations," in *Numerical methods in the study of critical phenomena* (J. Della-Dora, J. Demongeot, and B. Lacolle, eds.), vol. 9 of *Springer Series in Synergetics*, pp. 180–193, Springer-Verlag, 1981.

[61] A. Pavé, *Modélisation en biologie et en écologie.* Lyon: Aléas, 1994.

[62] C. Jeffries, "Qualitative stability of certain nonlinear systems," *Linear Algebra and its Applications*, vol. 75, pp. 133–144, 1986.

[63] L. A. Segel, *Modeling Dynamic Phenomena in Molecular and Cellular Biology.* Cambridge: Cambridge University Press, 1984.

[64] J. Monod, *Recherches sur la Croissance des Cultures Bactériennes.* Paris: Hermann, 1942.

[65] L. Edelstein, *Mathematical Models in Biology.* New York: Random House, 1988.

[66] S. R. Hansen and S. P. Hubbell, "Single-nutrient microbial competition," *Science*, vol. 207, no. 28, pp. 1491–1493, 1980.

[67] J.-L. Gouzé, "Positivity, space scale, and convergence towards the equilibrium," *Journal of Biological Systems*, vol. 3, no. 2, pp. 613–620, 1995.

[68] J.-L. Gouzé, "Positive and negative circuits in dynamical systems," *Journal Biol. Syst.*, vol. 6, no. 1, pp. 11–15, 1998.

[69] E. Walter and L. Pronzato, *Identification de modèles paramétriques.* Masson, 1994.

[70] L. Chen, O. Bernard, G. Bastin, and P. Angelov, "Hybrid modelling of biotechnological processes using neural networks," *Contr. Eng. Prcatice*, vol. 8, pp. 821–827, 2000.