

SpringerBriefs on PDEs and Data Science

Emmanuel Trélat

# Control in Finite and Infinite Dimension

# **SpringerBriefs on PDEs and Data Science**

## **Editor-in-Chief**

Enrique Zuazua, Department of Mathematics, University of Erlangen-Nuremberg, Erlangen, Bayern, Germany

## **Series Editors**

Irene Fonseca, Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

Franca Hoffmann, Hausdorff Center for Mathematics, University of Bonn, Bonn, Germany

Shi Jin, Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, Shanghai, China

Juan J. Manfredi, Department of Mathematics, University Pittsburgh, Pittsburgh, PA, USA

Emmanuel Trélat, CNRS, Laboratoire Jacques-Louis Lions, Sorbonne University, PARIS CEDEX 05, Paris, France

Xu Zhang, School of Mathematics, Sichuan University, Chengdu, Sichuan, China

SpringerBriefs on PDEs and Data Science targets contributions that will impact the understanding of partial differential equations (PDEs), and the emerging research of the mathematical treatment of data science.

The series will accept high-quality original research and survey manuscripts covering a broad range of topics including analytical methods for PDEs, numerical and algorithmic developments, control, optimization, calculus of variations, optimal design, data driven modelling, and machine learning. Submissions addressing relevant contemporary applications such as industrial processes, signal and image processing, mathematical biology, materials science, and computer vision will also be considered.

The series is the continuation of a former editorial cooperation with BCAM, which resulted in the publication of 28 titles as listed here: <https://www.springer.com/gp/mathematics/bcam-springerbriefs>

Emmanuel Trélat

# Control in Finite and Infinite Dimension

Emmanuel Trélat  
Laboratoire Jacques-Louis Lions  
Sorbonne Université  
Paris, France

ISSN 2731-7595                      ISSN 2731-7609 (electronic)  
SpringerBriefs on PDEs and Data Science  
ISBN 978-981-97-5947-7              ISBN 978-981-97-5948-4 (eBook)  
<https://doi.org/10.1007/978-981-97-5948-4>

Mathematics Subject Classification: 93C15, 93C20, 93C25

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

If disposing of this product, please recycle the paper.

# Foreword

This short book is the result of various master and summer school courses I have taught. The objective is to introduce the readers to mathematical control theory, both in finite and infinite dimension. In the finite-dimensional context, we consider controlled ordinary differential equations (ODEs); in this context, existence and uniqueness issues are easily resolved thanks to the Picard-Lindelöf (Cauchy-Lipschitz) theorem. In infinite dimension, in view of dealing with controlled partial differential equations (PDEs), the concept of well-posed system is much more difficult and requires to develop a bunch of functional analysis tools, in particular semigroup theory—and this, just for the setting in which the control system is written and makes sense. This is why I have split the book into two parts, the first being devoted to finite-dimensional control systems, and the second to infinite-dimensional ones.

In spite of this splitting, it may be nice to learn basics of control theory for finite-dimensional linear autonomous control systems (e.g., the Kalman condition) and then to see in the second part how some results are extended to infinite dimension, where matrices are replaced by operators, and exponentials of matrices are replaced by semigroups. For instance, the reader will see how the Gramian controllability condition is expressed in infinite dimension, and leads to the celebrated Hilbert Uniqueness Method (HUM).

Except the very last section, in the second part I have only considered linear autonomous control systems (the theory is already quite complicated), providing anyway several references to other textbooks for the several techniques existing to treat some particular classes of nonlinear PDEs. In contrast, in the first part on finite-dimensional control theory, there are much less difficulties to treat general nonlinear control systems, and I give here some general results on controllability, optimal control, and stabilization.

Of course, whether in finite or infinite dimension, there exist much deeper results and methods in the literature, established however for specific classes of control

systems. Here, my objective is to provide the reader with an introduction to control theory and to the main tools allowing to treat general control systems. I hope this will serve as motivation to go deeper into the theory or numerical aspects that are not covered here.

Paris, France  
March 2024

Emmanuel Trélat

# Contents

## Part I Control in Finite Dimension

<b>1</b>	<b>Controllability</b>	3
1.1	Controllability of Linear Systems	4
1.1.1	Controllability of Autonomous Linear Systems	4
1.1.1.1	Without Control Constraints: Kalman Condition	5
1.1.1.2	With Control Constraints	8
1.1.1.3	Similar Systems	9
1.1.2	Controllability of Time-Varying Linear Systems	11
1.1.2.1	Case Without Control Constraints	11
1.1.2.2	Case with Control Constraints	14
1.1.3	Geometry of Accessible Sets	15
1.2	Controllability of Nonlinear Systems	17
1.2.1	Local Controllability Results	17
1.2.2	Geometry of the Accessible Set	23
1.2.3	Global Controllability Results	25
	References	27
<b>2</b>	<b>Optimal Control</b>	29
2.1	Existence of an Optimal Control	30
2.2	Pontryagin Maximum Principle (PMP)	33
2.2.1	General Statement	33
2.2.2	Proof in a Simplified Context	36
2.2.3	Generalizations and Additional Comments	39
2.3	Particular Cases and Examples	41
2.3.1	Minimal Time Problem for Linear Control Systems	41
2.3.2	Linear Quadratic Theory	44
2.3.2.1	The Basic LQ Problem	44
2.3.2.2	Tracking Problem	46
2.3.3	Examples of Nonlinear Optimal Control Problems	48
	References	58



<b>3</b>	<b>Stabilization</b>	61
3.1	Stabilization of Autonomous Linear Systems	61
3.1.1	Reminders on Stability Notions	61
3.1.2	Pole-Shifting Theorem	64
3.2	Stabilization of Instationary Linear Systems	66
3.3	Stabilization of Nonlinear Systems	67
3.3.1	Local Stabilization by Linearization	67
3.3.2	Global Stabilization by Lyapunov Theory	68
	References	74
 <b>Part II Control in Infinite Dimension</b>		
<b>4</b>	<b>Semigroup Theory</b>	77
4.1	Homogeneous Cauchy Problems	79
4.1.1	Semigroups of Linear Operators	79
4.1.2	The Cauchy Problem	85
4.1.2.1	Classical Solutions	85
4.1.2.2	Weak Solutions	87
4.1.3	Scale of Banach Spaces	90
4.2	Nonhomogeneous Cauchy Problems	94
	References	96
<b>5</b>	<b>Linear Control Systems in Banach Spaces</b>	97
5.1	Admissible Control Operators	98
5.1.1	Definition	98
5.1.2	Dual Characterization of the Admissibility	100
5.1.3	Degree of Unboundedness of the Control Operator	102
5.1.4	Examples	104
5.1.4.1	Dirichlet Heat Equation with Internal Control	104
5.1.4.2	Heat Equation with Dirichlet Boundary Control	105
5.1.4.3	Heat Equation with Neumann Boundary Control	107
5.1.4.4	Second-Order Equations	107
5.2	Controllability	109
5.2.1	Definitions	109
5.2.2	Duality Controllability / Observability	111
5.2.3	Hilbert Uniqueness Method (HUM)	113
5.2.4	Example: The Wave Equation	114
5.2.4.1	1D Wave Equation with Dirichlet Boundary Control	115
5.2.4.2	1D Dirichlet Wave Equation with Internal Control	117
5.2.4.3	Multi-D Dirichlet Wave Equation with Internal Control	119
5.2.5	Example: The Heat Equation	120
5.2.5.1	Dirichlet Heat Equation with Internal Control	120
5.2.5.2	Heat Equation with Dirichlet Boundary Control	121

5.2.6	Pontryagin Maximum Principle .....	122
5.3	Further Results .....	124
5.3.1	Kalman Condition in Infinite Dimension .....	124
5.3.2	Necessary Conditions for Exact Controllability .....	125
5.3.3	Moment Method .....	126
5.3.4	Equivalence Between Observability and Exponential Stability .....	129
5.3.5	1D Semilinear Heat Equation .....	133
	References .....	137

**Part I**  
**Control in Finite Dimension**

# Chapter 1

## Controllability



Let  $n$  and  $m$  be two positive integers. In this chapter we consider a control system in  $\mathbb{R}^n$

$$\dot{x}(t) = f(t, x(t), u(t)) \quad (1.1)$$

where  $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  is of class  $C^1$  with respect to  $(x, u)$  and locally integrable with respect to  $t$ , and the controls are measurable essentially bounded functions of time taking their values in some measurable subset  $\Omega$  of  $\mathbb{R}^m$  (set of control constraints).

First of all, given an arbitrary initial point  $x_0 \in \mathbb{R}^n$ , and an arbitrary control  $u$ , we claim that there exists a unique solution  $x(\cdot)$  of (1.1) such that  $x(0) = x_0$ , maximally defined on some open interval of  $\mathbb{R}$  containing 0. We use here a generalization of the usual Picard-Lindelöf theorem (sometimes called Carathéodory theorem), where the dynamics here can be discontinuous (because of the control). For a general version of this existence and uniqueness theorem, we refer to [6, Theorem 5.3] and [12, Appendix C3]. We stress that the differential equation (1.1) holds for almost every  $t$  in the maximal interval. Given a time  $T > 0$  and an initial point  $x_0$ , we say that a control  $u \in L^\infty([0, T], \mathbb{R}^m)$  is admissible if the corresponding trajectory  $x(\cdot)$ , such that  $x(0) = x_0$ , is well defined on  $[0, T]$ .

We say that the control system is *linear* if  $f(t, x, u) = A(t)x + B(t)u + r(t)$ , with  $A(t)$  a  $n \times n$  matrix,  $B(t)$  a  $n \times m$  matrix (with real coefficients),  $r(t) \in \mathbb{R}^n$ , and in that case we will assume that  $t \mapsto A(t)$ ,  $t \mapsto B(t)$  and  $t \mapsto r(t)$  are of class  $L^\infty$  on every compact interval (actually,  $L^1$  would be enough). The linear control system is said to be *autonomous* if  $A(t) \equiv A$  and  $B(t) \equiv B$ , otherwise it is said to be *instationary* or *time-varying*. Note that, for linear control systems, there is no blow-up in finite time (i.e., admissibility holds true on any interval).

**Definition 1.1** Let  $x_0 \in \mathbb{R}^n$  and let  $T > 0$  arbitrary. A control  $u \in L^\infty([0, T], \Omega)$  is said to be *admissible* on  $[0, T]$  if the trajectory  $x_u(\cdot)$ , solution of (1.1), corresponding to the control  $u$ , and such that  $x_u(0) = x_0$ , is well defined on  $[0, T]$ . The *end-point mapping*  $E_{x_0, T}$  is then defined by  $E_{x_0, T}(u) = x_u(T)$ .

The set of admissible controls on  $[0, T]$  is denoted by  $\mathcal{U}_{x_0, T, \Omega}$ . It is the domain of definition of  $E_{x_0, T}$  (indeed one has to be careful with blow-up phenomena), when considering controls taking their values in  $\Omega$ .

**Definition 1.2** The control system (1.1) is said to be (globally) *controllable from*  $x_0$  *in time*  $T$  if  $E_{x_0, T}(\mathcal{U}_{x_0, T, \Omega}) = \mathbb{R}^n$ , i.e., if  $E_{x_0, T}$  is surjective.

Accordingly, defining the *accessible set* from  $x_0$  in time  $T$  by  $\text{Acc}_\Omega(x_0, T) = E_{x_0, T}(\mathcal{U}_{x_0, T, \Omega})$ , the control system (1.1) is (globally) controllable from  $x_0$  in time  $T$  if  $\text{Acc}_\Omega(x_0, T) = \mathbb{R}^n$ .

Since such a global surjectivity property is certainly a very strong property which may not hold in general, it is relevant to define *local controllability*.

**Definition 1.3** Let  $x_1 = E_{x_0, T}(\bar{u})$  for some  $\bar{u} \in \mathcal{U}_{x_0, T, \Omega}$ . The control system (1.1) is said to be *locally controllable from*  $x_0$  *in time*  $T$  *around*  $x_1$  if  $x_1$  belongs to the interior of  $\text{Acc}_\Omega(x_0, T)$ , i.e., if  $E_{x_0, T}$  is locally surjective around  $x_1$ .

Other variants of controllability can be defined. A clear picture will come from the geometric representation of the accessible set.

In this chapter we will provide several tools in order to analyze the controllability properties of a control system, first for linear (autonomous, and then instationary) systems, and then for nonlinear systems.

## 1.1 Controllability of Linear Systems

Throughout this section, we consider the linear control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$ , with  $u \in L^\infty([0, +\infty), \Omega)$ . Since there is no finite-time blow-up for linear systems, we have  $\mathcal{U}_{x_0, T, \Omega} = L^\infty([0, T], \Omega)$  for every  $T > 0$ .

### 1.1.1 Controllability of Autonomous Linear Systems

In this section, we assume that  $A(t) \equiv A$  and  $B(t) \equiv B$ , where  $A$  is a  $n \times n$  matrix and  $B$  is a  $n \times m$  matrix.

### 1.1.1.1 Without Control Constraints: Kalman Condition

In this section, we assume that  $\Omega = \mathbb{R}^m$  (no control constraint). The celebrated Kalman theorem provides a necessary and sufficient condition for autonomous linear control systems without control constraint.

**Theorem 1.1** *We assume that  $\Omega = \mathbb{R}^m$  (no control constraint). The control system  $\dot{x}(t) = Ax(t) + Bu(t) + r(t)$  is controllable (from any initial point, in arbitrary time  $T > 0$ ) if and only if the Kalman matrix*

$$K(A, B) = (B, AB, \dots, A^{n-1}B)$$

(which is of size  $n \times nm$ ) is of maximal rank  $n$ .

**Proof of Theorem 1.1** Given any  $x_0 \in \mathbb{R}^n$ ,  $T > 0$  and  $u \in L^\infty([0, T], \mathbb{R}^m)$ , the Duhamel formula gives

$$E_{x_0, T}(u) = x_u(T) = e^{TA}x_0 + \int_0^T e^{(T-t)A}r(t) dt + L_T u \quad (1.2)$$

where  $L_T : L^\infty([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n$  is the linear continuous operator defined by  $L_T u = \int_0^T e^{(T-t)A}Bu(t) dt$ . Clearly, the system is controllable in time  $T$  if and only if  $L_T$  is surjective. Then to prove the theorem it suffices to prove the following lemma.

**Lemma 1.1** *The Kalman matrix  $K(A, B)$  is of rank  $n$  if and only if  $L_T$  is surjective.*

**Proof of Lemma 1.1** We argue by contraposition. If  $L_T$  is not surjective, then there exists  $\psi \in \mathbb{R}^n \setminus \{0\}$  which is orthogonal to the range of  $L_T$ , that is,

$$\psi^\top \int_0^T e^{(T-t)A}Bu(t) dt = 0 \quad \forall u \in L^\infty([0, T], \mathbb{R}^m).$$

This implies that  $\psi^\top e^{(T-t)A}B = 0$ , for every  $t \in [0, T]$ . Taking  $t = T$  yields  $\psi^\top B = 0$ . Then, derivating first with respect to  $t$ , and taking  $t = T$  then yields  $\psi^\top AB = 0$ . By immediate iteration we get that  $\psi^\top A^k B = 0$ , for every  $k \in \mathbb{N}$ . In particular  $\psi^\top K(A, B) = 0$  and thus the rank of  $K(A, B)$  is less than  $n$ .

Conversely, if the rank of  $K(A, B)$  is less than  $n$ , then there exists  $\psi \in \mathbb{R}^n \setminus \{0\}$  such that  $\psi^\top K(A, B) = 0$ , and therefore  $\psi^\top A^k B = 0$ , for every  $k \in \{0, 1, \dots, n-1\}$ . From the Hamilton-Cayley theorem, there exist real numbers  $a_0, a_1, \dots, a_{n-1}$  such that  $A^n = \sum_{k=0}^{n-1} a_k A^k$ . Therefore we get easily that  $\psi^\top A^n B = 0$ . Then, using the fact that  $A^{n+1} = \sum_{k=1}^n a_k A^k$ , we get as well that  $\psi^\top A^{n+1} B = 0$ . By immediate recurrence, we infer that  $\psi^\top A^k B = 0$ , for every  $k \in \mathbb{N}$ , and therefore, using the series expansion of the exponential, we get that  $\psi^\top e^{(T-t)A}B = 0$ , for

every  $t \in [0, T]$ . We conclude that  $\psi^\top L_T u = 0$  for every  $u \in L^\infty([0, T], \mathbb{R}^m)$  and thus that  $L_T$  is not surjective.  $\square$

Theorem 1.1 is proved.

**Remark 1.1** Note that the Kalman condition is purely algebraic and is easily checkable.

The Kalman condition does neither depend on  $T$  nor on  $x_0$ . Therefore, if an autonomous linear control system is controllable from  $x_0$  in time  $T > 0$ , starting at  $x_0$ , then it is controllable from any other  $x'_0$  in any time  $T' > 0$ , in particular, in arbitrarily small time. This is due to the fact that there are no control constraints. When there are some control constraints one cannot hope to have such a property.

**Example 1.1** Consider an RLC circuit with a resistor with resistance  $R$ , a coil with inductance  $L$  and a capacitor with capacitance  $C$ , connected in series. We control the input voltage  $u(t)$  of the electrical circuit. Denoting by  $i(t)$  the intensity, by additivity of voltages we have

$$Ri(t) + L \frac{di}{dt}(t) + \frac{1}{C} \int^t i(s) ds = u(t).$$

Setting  $x_1(t) = \int^t i(s) ds$ ,  $x_2(t) = \dot{x}_1(t) = i(t)$  and  $x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$ , we find the control system  $\dot{x}(t) = Ax(t) + Bu(t)$  with

$$A = \begin{pmatrix} 0 & 1 \\ \frac{-1}{LC} & \frac{-R}{L} \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The Kalman condition is then easily checked.

This simple but illuminating example shows the importance of the RLC device in electricity. Note that the RLC circuit is paradigmatic for the three main operations in mathematical analysis: multiplication operator, derivation operator and integration operator.

### Example 1.2 (Kalman Condition Computations)

- Let  $m > 0$  and  $d, k \geq 0$ . Prove that the system consisting of the controlled spring:  $m\ddot{x} + d\dot{x} + kx = u$ , is controllable.
- Let  $k_1, k_2 \geq 0$ . Prove that the system of coupled springs (two-car train) given by

$$\ddot{x}_1 = -k_1 x_1 + k_2 (x_2 - x_1), \quad \ddot{x}_2 = -k_2 (x_2 - x_1) + u,$$

is controllable if and only if  $k_2 > 0$ .

- Let  $(b_1, b_2) \in \mathbb{R}^2 \setminus \{(0, 0)\}$ . Prove that the control system

$$\dot{x}_1 = x_2 + b_1 u, \quad \dot{x}_2 = -x_1 + b_2 u,$$

is controllable.

- Prove that the control system

$$\dot{x}_1 = 2x_1 + (\alpha - 3)x_2 + u_1 + u_2, \quad \dot{x}_2 = 2x_2 + \alpha(\alpha - 1)u_1,$$

is controllable if and only if  $\alpha(\alpha - 1) \neq 0$ .

- Let  $N, m \in \mathbb{N} \setminus \{0\}$ , let  $A = (a_{ij})_{1 \leq i, j \leq N}$  be a  $N \times N$  real-valued matrix and  $B = (b_{ij})_{1 \leq i \leq N, 1 \leq j \leq m}$  be a  $N \times m$  real-valued matrix, such that the pair  $(A, B)$  satisfies the Kalman condition. Let  $d \in \mathbb{N} \setminus \{0\}$ . Prove that the control system in  $(\mathbb{R}^d)^N$  given by

$$\dot{v}_i(t) = \sum_{j=1}^N a_{ij}v_j(t) + \sum_{j=1}^m b_{ij}u_j(t), \quad i = 1 \dots N,$$

where  $v_i(t) \in \mathbb{R}^d$  and  $u_j(t) \in \mathbb{R}^d$ , is controllable.

Many other examples are given in [13].

**Remark 1.2 (Hautus Test)** The following assertions are equivalent:

- (1) The pair  $(A, B)$  satisfies Kalman's condition  $\text{rank}(K(A, B)) = n$ .
- (2)  $\forall \lambda \in \mathbb{C} \quad \text{rank}(\lambda I - A, B) = n$ .
- (3)  $\forall \lambda \in \text{Spec}(A) \quad \text{rank}(\lambda I - A, B) = n$ .
- (4)  $\forall z$  eigenvector of  $A^\top, \quad B^\top z \neq 0$ .
- (5)  $\exists c > 0 \mid \forall \lambda \in \mathbb{C} \quad \forall z \in \mathbb{R}^n \quad \|(\lambda I - A^\top)z\|^2 + \|B^\top z\|^2 \geq c\|z\|^2$ .

Indeed, (2)  $\Leftrightarrow$  (3), (2)  $\Leftrightarrow$  (5), and not (4)  $\Rightarrow$  not (1), are easy. We also easily get (3)  $\Leftrightarrow$  (4) by contradiction. The implication not (1)  $\Rightarrow$  not (4) is proved as follows. We set  $N = \{z \in \mathbb{R}^n \mid z^\top A^k B = 0 \quad \forall k \in \mathbb{N}\}$ . It is easy to establish that  $A^\top N \subset N$ . Then, not (1)  $\Rightarrow N \neq \{0\}$ , and then to conclude it suffices to note that  $A^\top$  must have an eigenvector in  $N$ .

The condition (5) of the Hautus test is particularly interesting because it is amenable to generalizations in the infinite-dimensional setting.

**Remark 1.3** Let us finally comment on how to generalize the Kalman condition in infinite dimension. This will be done properly in Lemma 5.12 in Sect. 5.3.1, but we can already anticipate and provide the reader with a flavor of the new difficulties arising in infinite dimension. Let us replace  $\mathbb{R}^n$  with a Banach space  $X$  and  $\mathbb{R}^m$  with a Banach space  $U$ . The matrix  $A$  becomes an operator  $A : D(A) \rightarrow X$  (for instance, a Laplacian) and  $B \in L(U, X)$  is a linear bounded operator (for instance, modelling an internal control for the heat equation). The Duhamel formula (1.2) remains valid provided  $e^{tA}$  is replaced with the semigroup generated by  $A$  (we stress that this is not an exponential if  $A$  is an unbounded operator). At this stage, the reader is advised to admit this notion and continue reading. The essential elements of semigroup theory are recalled in Chap. 4.

The argument of the proof of Theorem 1.1 remains essentially the same, except of course the application of the Hamilton-Cayley theorem, and the Kalman matrix is



replaced with a matrix of operators with an infinite number of columns ( $k$  does not stop at  $n - 1$  but goes to  $+\infty$ ), thus giving the statement of Lemma 5.12. There is however a serious and deep difference: at the beginning of the proof of Lemma 1.1 we have used the fact that, when the vector space  $\text{Ran}(L_T)$  is a proper subset of  $\mathbb{R}^n$ , there exists a nontrivial vector  $\psi$  orthogonal to it: this is a separation argument. In infinite dimension this argument may dramatically fail, because a proper subset of  $X$  might be dense in  $X$ : in such a case, separation is not possible. Then, to make the argument valid, one has to consider the closure  $\overline{\text{Ran}(L_T)}$  of  $\text{Ran}(L_T)$ , as done in Lemma 5.12, leading to a result of *approximate* controllability.

Actually, as we will see in Part II of that book, in infinite dimension we must distinguish (at least) between *approximate* and *exact* controllability. This distinction does not exist in finite dimension. At this stage, let us just say that, for linear autonomous control systems, *exact controllability* means  $\text{Ran}(E_{T,x_0}) = X$  while *approximate controllability* means  $\overline{\text{Ran}(E_{T,x_0})} = X$ . For example, a heat equation settled on a domain with internal control exerted on a proper subset of this domain is approximately controllable in  $X = L^2$  (see Part II) but can never be exactly controllable in this state space because of the smoothing property of the heat semigroup.

### 1.1.1.2 With Control Constraints

An easy adaptation of the proof of Theorem 1.1 yields the following result.

**Proposition 1.1** *We assume that  $r = 0$ , that  $0 \in \overset{\circ}{\Omega}$ , and that the Kalman condition holds true. Let  $x_0 \in \mathbb{R}^n$  be arbitrary. For every  $T > 0$ , the accessible set  $\text{Acc}_\Omega(x_0, T)$  contains an open neighborhood of the point  $e^{TA}x_0$ . In other words, the control system is locally controllable in time  $T$  from  $x_0$  around  $e^{TA}x_0$ .*

*More precisely, for every  $T > 0$  there exists a neighborhood  $V$  of  $e^{TA}x_0$  such that, for every  $x_1 \in V$ , there exists  $u \in L^\infty([0, T], \Omega)$  (which is close to 0 in  $L^\infty$  topology) such that  $E_{x_0, T}(u) = x_1$ . Conversely, this openness property implies the Kalman condition.*

Many variants of controllability properties can be obtained under various additional assumptions. For instance we have the following easy result.

**Proposition 1.2** *We assume that  $r = 0$ , that  $0 \in \overset{\circ}{\Omega}$ , that the Kalman condition holds true, and that all eigenvalues of  $A$  have negative real part. Let  $x_0 \in \mathbb{R}^n$  be arbitrary. There exists a time  $T > 0$  and a control  $u \in L^\infty([0, T], \Omega)$  such that the solution of  $\dot{x}(t) = Ax(t) + Bu(t)$ ,  $x(0) = x_0$ , satisfies  $x(T) = 0$ .*

The time  $T$  in the above result may be large. The strategy of proof consists of taking  $u = 0$  and letting the trajectory converge asymptotically to 0, and then as soon as it is sufficiently close to 0, we apply the controllability result with controls having a small enough norm.

### 1.1.1.3 Similar Systems

Let us investigate the effect of a change of basis in linear autonomous control systems.

**Definition 1.4** The linear control systems  $\dot{x}_1 = A_1 x_1 + B_1 u_1$  and  $\dot{x}_2 = A_2 x_2 + B_2 u_2$  are said to be *similar* whenever there exists  $P \in GL_n(\mathbb{R})$  such that  $A_2 = P A_1 P^{-1}$  and  $B_2 = P B_1$ . We have then  $x_2 = P x_1$  and  $u_2 = u_1$ .

We also say that the pairs  $(A_1, B_1)$  and  $(A_2, B_2)$  are similar.

**Remark 1.4** The Kalman property is intrinsic, that is

$$(B_2, A_2 B_2, \dots, A_2^{n-1} B_2) = P(B_1, A_1 B_1, \dots, A_1^{n-1} B_1),$$

In particular the rank of the Kalman matrix is invariant under a similar transform.

**Proposition 1.3** Let  $A$  be a matrix of size  $n \times n$ , and let  $B$  be a matrix of size  $n \times m$ . Then the pair  $(A, B)$  is similar to the pair  $(A', B')$ , with

$$A' = \begin{pmatrix} A'_1 & A'_3 \\ 0 & A'_2 \end{pmatrix} \quad \text{and} \quad B' = \begin{pmatrix} B'_1 \\ 0 \end{pmatrix}$$

where  $A_1$  is of size  $r \times r$ ,  $B_1$  is of size  $r \times m$ , and  $r = \text{rank } K(A, B) = \text{rank } K(A'_1, B'_1)$ .

In other words, this result says the following. Denoting by  $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$  the new coordinates, with  $y_1$  of dimension  $r$  and  $y_2$  of dimension  $n - r$ , the control system in the new coordinates is written as

$$\begin{aligned} \dot{y}_1 &= A'_1 y_1 + B'_1 u + A'_3 y_2 \\ \dot{y}_2 &= A'_2 y_2 \end{aligned}$$

Since the pair  $(A'_1, B'_1)$  satisfies the Kalman condition, it follows that the part of the system in  $y_1$  is controllable: it is called the *controllable part* of the system. The part in  $y_2$  is uncontrollable and is called the *uncontrollable part* of the system.

**Proof of Proposition 1.3** We assume that the rank of  $K(A, B)$  is less than  $n$  (otherwise there is nothing to prove). The subspace

$$F = \text{Ran } K(A, B) = \text{Ran } B + \text{Ran } AB + \dots + \text{Ran } A^{n-1} B$$

is of dimension  $r$ , and is invariant under  $A$  (this can be seen by using the Hamilton-Cayley theorem). Let  $G$  be a subspace of  $\mathbb{R}^n$  such that  $\mathbb{R}^n = F \oplus G$  and let  $(f_1, \dots, f_r)$  be a basis of  $F$  and  $(f_{r+1}, \dots, f_n)$  be a basis of  $G$ . Let  $P$  be the change-of-basis matrix from the basis  $(f_1, \dots, f_n)$  to the canonical basis of  $\mathbb{R}^n$ .

Since  $F$  is invariant under  $A$ , we get

$$A' = P A P^{-1} = \begin{pmatrix} A'_1 & A'_3 \\ 0 & A'_2 \end{pmatrix}$$

and since  $\text{Ran } B \subset F$ , we must have  $B' = P B = \begin{pmatrix} B'_1 \\ 0 \end{pmatrix}$ . Finally, it is clear that the rank of  $K(A'_1, B'_1)$  is equal to the rank of  $K(A, B)$ .  $\square$

**Theorem 1.2 (Brunovski Normal Form)** *Let  $A$  be a matrix of size  $n \times n$  and let  $B$  be a matrix of size  $n \times 1$  (note that  $m = 1$  here) be such that  $(A, B)$  satisfies the Kalman condition. Then the pair  $(A, B)$  is similar to the pair  $(\tilde{A}, \tilde{B})$ , with*

$$\tilde{A} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix} \quad \text{and} \quad \tilde{B} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

where the coefficients  $a_i$  are those of the characteristic polynomial of  $A$ , that is  $\chi_A(X) = X^n + a_1 X^{n-1} + \cdots + a_{n-1} X + a_n$ .

Note that the matrix  $\tilde{A}$  is the companion matrix of characteristic polynomial  $\chi_A$ . Theorem 1.2 means that, in the new coordinates, the control system is equivalent to the scalar differential equation of order  $n$  with scalar control  $x^{(n)}(t) + a_1 x^{(n-1)}(t) + \cdots + a_n x(t) = u(t)$ .

**Proof of Theorem 1.2** First, let us note that, if there exists a basis  $(f_1, \dots, f_n)$  in which the pair  $(A, B)$  takes the form  $(\tilde{A}, \tilde{B})$ , then we must have  $f_n = B$  (up to scaling) and

$$A f_n = f_{n-1} - a_1 f_n, \dots, A f_2 = f_1 - a_{n-1} f_n, A f_1 = -a_n f_n.$$

Let us then define the vectors  $f_1, \dots, f_n$  by

$$f_n = B, \quad f_{n-1} = A f_n + a_1 f_n, \quad \dots, \quad f_1 = A f_2 + a_{n-1} f_n.$$

The  $n$ -tuple  $(f_1, \dots, f_n)$  is a basis of  $\mathbb{R}^n$ , since

$$\begin{aligned} \text{Span}\{f_n\} &= \text{Span}\{B\} \\ \text{Span}\{f_n, f_{n-1}\} &= \text{Span}\{B, AB\} \\ &\vdots \\ \text{Span}\{f_n, \dots, f_1\} &= \text{Span}\{B, \dots, A^{n-1}B\} = \mathbb{R}^n. \end{aligned}$$

It remains to check that  $Af_1 = -a_n f_n$ . We have

$$\begin{aligned} Af_1 &= A^2 f_2 + a_{n-1} Af_n = A^2 (Af_3 + a_{n-2} f_n) + a_{n-1} Af_n \\ &= \dots = A^n f_n + a_1 A^{n-1} f_n + \dots + a_{n-1} Af_n = -a_n f_n \end{aligned}$$

since by the Hamilton-Cayley theorem we have  $A^n = -a_1 A^{n-1} - \dots - a_n I$ . In the basis  $(f_1, \dots, f_n)$ , the pair  $(A, B)$  takes the form  $(\tilde{A}, \tilde{B})$ .  $\square$

**Remark 1.5** This theorem can be generalized to the case  $m > 1$  but the normal form is not that simple.

### 1.1.2 Controllability of Time-Varying Linear Systems

In what follows, we denote by  $R(t, s)$  the *state-transition matrix* of the linear system  $\dot{x}(t) = A(t)x(t)$ , that is, the unique solution of

$$\partial_t R(t, s) = A(t)R(t, s), \quad R(s, s) = I_n, \quad (1.3)$$

for  $t, s \in \mathbb{R}$ . Note that, in the autonomous case  $A(t) \equiv A$ , we have  $R(t, s) = e^{(t-s)A}$ . But in general the state-transition matrix cannot be computed explicitly. Recall that  $R(t, s)R(s, \tau) = R(t, \tau)$  for all  $t, s, \tau \in \mathbb{R}$ ; in particular,  $R(t, s) = R(s, t)^{-1}$ .

#### 1.1.2.1 Case Without Control Constraints

**Theorem 1.3** *We assume that  $\Omega = \mathbb{R}^m$  (no constraints on the control). The control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$  is controllable in time  $T$  (from any initial point  $x_0$ ) if and only if the Gramian matrix*

$$G_T = \int_0^T R(T, t)B(t)B(t)^\top R(T, t)^\top dt$$

*is invertible.*

**Remark 1.6** The invertibility condition of  $G_T$  depends on  $T$  but not on the initial point. Therefore, if a linear instationary control system is controllable from  $x_0$  in time  $T$  then it is controllable from any other initial point (but with the same time  $T$ ). It may fail to be controllable in time  $T' < T$  (take for instance  $B(t) = 0$  for  $0 \leq t \leq T'$ ). Anyway, controllability in time  $T$  implies controllability in any time  $T' \geq T$ : indeed, at time  $T$  the range of the end-point mapping is equal to the whole  $\mathbb{R}^n$  and it cannot decrease in larger time.

**Remark 1.7** Note that  $G_T = G_T^\top$  and that

$$\psi^\top G_T \psi = \langle G_T \psi, \psi \rangle = \int_0^T \|B(t)^\top R(T, t)^\top \psi\|^2 dt \geq 0 \quad \forall \psi \in \mathbb{R}^n,$$

i.e.,  $G_T$  is a symmetric nonnegative matrix of size  $n \times n$ . Theorem 1.3 states that the system is controllable in time  $T$  if and only if  $G_T$  is positive definite. By a diagonalization argument, this is equivalent to the existence of  $C_T > 0$  (the lowest eigenvalue of  $G_T$ ) such that

$$\int_0^T \|B(t)^\top R(T, t)^\top \psi\|^2 dt \geq C_T \|\psi\|^2 \quad \forall \psi \in \mathbb{R}^n. \quad (1.4)$$

This is an *observability inequality*. The system is controllable in time  $T$  if and only if the above observability inequality holds.

The important concept of observability is not developed in this book. We recall it briefly for a linear control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$  and we refer to [10, 12, 13] for more details. Assume that, at any time  $t$ , we cannot observe the whole state  $x(t)$  but only a part of it,  $y(t) = C(t)x(t)$ , where  $C(t)$  is a  $m \times n$  matrix. The control system, with its output  $y(\cdot)$ , is said to be observable in time  $T$  if a given output  $y(\cdot)$  observed on  $[0, T]$  can be generated by only one initial condition  $x_0$ . While controllability corresponds to a surjectivity property, observability thus corresponds to an injectivity property. Actually, the system  $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$  with output  $y(t) = C(t)x(t)$  is observable in time  $T$  if and only if the control system  $\dot{x}(t) = A(t)^\top x(t) + C(t)u(t)$  is controllable in time  $T$ : this is the so-called controllability – observability duality (see Lemma 5.9 in Part II, Sect. 5.2.2, for a general mathematical statement establishing this duality in a general context). Hence, observability in time  $T$  is characterized by the observability inequality (1.4) with  $B(t)^\top$  replaced by  $C(t)$ .

**Proof of Theorem 1.3** Let  $x_0 \in \mathbb{R}^n$  arbitrary. Any solution of the control system, associated with some control  $u$  and starting at  $x_0$ , satisfies at time  $T$

$$x_u(T) = x^* + \int_0^T R(T, t)B(t)u(t) dt$$

with  $x^* = R(T, 0)x_0 + \int_0^T R(T, t)r(t) dt$ .

Let us assume that  $G_T$  is invertible and let us prove that the control system is controllable in time  $T$ . Let  $x_1 \in \mathbb{R}^n$  be any target point. We seek an appropriate control  $u$  in the form  $u(t) = B(t)^\top R(t, T)^\top \psi$ , with  $\psi \in \mathbb{R}^n$  to be chosen such that  $x_u(T) = x_1$ . With this control, we have  $x_u(T) = x^* + G_T \psi$ , and since  $G_T$  is invertible it suffices to take  $\psi = G_T^{-1}(x_1 - x^*)$ .

Conversely, if  $G_T$  is not invertible, then by Remark 1.7 there exists  $\psi \in \mathbb{R}^n \setminus \{0\}$  such that  $\psi^\top G_T \psi = \int_0^T \|B(t)^\top R(t, T)^\top \psi\|^2 dt = 0$ , hence  $\psi^\top R(T, t)B(t) = 0$

for almost every  $t \in [0, T]$ . It follows that  $\psi^\top \int_0^T R(T, t)B(t)u(t) dt = 0$  for every  $u \in L^\infty([0, T], \mathbb{R}^m)$ , and thus  $\psi^\top (x_u(T) - x^*) = 0$ , which means that  $x_u(T)$  belongs to a proper affine subspace of  $\mathbb{R}^n$  (namely,  $x^* + \psi^\perp$ ) as  $u$  varies. Hence the system is not controllable in time  $T$ .  $\square$

**Remark 1.8** This theorem can be proved in an easier and more natural way with the Pontryagin maximum principle (PMP), within an optimal control viewpoint: anticipating a bit,  $p(t) = R(t, T)^\top \psi$  is the adjoint vector, solution of  $\dot{p}(t) = -A(t)^\top p(t)$  such that  $p(T) = \psi$ , obtained by applying the PMP with the cost being the square of the  $L^2$  norm of the control, and actually the control used in the above proof is optimal for the  $L^2$  norm. The above proof also leads in the infinite-dimensional setting to the *HUM method* (see Part II).

**Remark 1.9** If the system is autonomous ( $A(t) \equiv A$ ,  $B(t) \equiv B$ ) then  $R(t, s) = e^{(t-s)A}$  and thus

$$G_T = \int_0^T e^{(T-t)A} B B^\top e^{(T-t)A^\top} dt = \int_0^T e^{tA} B B^\top e^{tA^\top} dt.$$

In that case, since the controllability (Kalman) condition does not depend on the time, it follows that  $G_{T_1}$  is invertible if and only if  $G_{T_2}$  is invertible, which is not evident from the above integral form (this fact is not true in general in the instationary case).

In the autonomous case, the observability inequality (1.4) can be written as

$$\int_0^T \|B^\top \underbrace{e^{(T-t)A^\top} \psi}_{p(t)}\|^2 dt \geq C_T \underbrace{\|\psi\|^2}_{\|p(T)\|^2} \quad \forall \psi \in \mathbb{R}^n. \quad (1.5)$$

Note that, setting  $\lambda(t) = p(T - t)$ , we have  $\dot{\lambda}(t) = A^\top \lambda(t)$ ,  $\lambda(0) = \psi$ , and (1.5) can also be written as  $\int_0^T \|B^\top \lambda(t)\|^2 dt \geq C_T \|\lambda(0)\|^2$ .

This observability inequality is appropriate to be generalized in the infinite-dimensional setting, replacing  $e^{tA}$  with a semigroup, and will be of instrumental importance in the derivation of the so-called HUM method (see Part II).

Let us now provide an ultimate theorem which generalizes the Kalman condition in the instationary case.

**Theorem 1.4** *We assume that  $\Omega = \mathbb{R}^m$  (no constraint on the control). Consider the control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$  where  $t \mapsto A(t)$  and  $t \mapsto B(t)$  are of class  $C^\infty$ . We define the sequence of matrices*

$$B_0(t) = B(t), \quad B_{k+1}(t) = A(t)B_k(t) - \frac{dB_k}{dt}(t), \quad k \in \mathbb{N}.$$

1. If there exists  $t \in [0, T]$  such that

$$\text{Span} \{B_k(t)v \mid v \in \mathbb{R}^m, k \in \mathbb{N}\} = \mathbb{R}^n \quad (1.6)$$

then the system is controllable in time  $T$ .

2. If  $t \mapsto A(t)$  and  $t \mapsto B(t)$  are moreover analytic (i.e., expandable in a convergent power series at any  $t$ ), then the system is controllable in time  $T$  if and only if (1.6) is satisfied for every  $t \in [0, T]$ .

The proof of Theorem 1.4 readily follows from the Hamiltonian characterization of singular trajectories (see [1, 13], see also the proof of the weak Pontryagin Maximum Principle in Sect. 2.2.2).

**Example 1.3** Thanks to Theorem 1.4, it is easy to prove that the control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$ , with

$$A(t) = \begin{pmatrix} t & 1 & 0 \\ 0 & t^3 & 0 \\ 0 & 0 & t^2 \end{pmatrix}, \text{ and } B(t) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

is controllable in any time  $T > 0$ , while the control system

$$\begin{cases} \dot{x}(t) = -y(t) + u(t) \cos t, \\ \dot{y}(t) = x(t) + u(t) \sin t, \end{cases}$$

is never controllable (Theorem 1.3 can also be applied).

### 1.1.2.2 Case with Control Constraints

When there are some control constraints, we can easily adapt Theorems 1.3 and 1.4, like in Proposition 1.1, to obtain local controllability results.

**Proposition 1.4** We assume that  $r = 0$  and that  $0 \in \overset{\circ}{\Omega}$ . Let  $x_0 \in \mathbb{R}^n$  and  $T > 0$  be arbitrary.

- The control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$  is locally controllable in time  $T$  around the point  $R(T, 0)x_0$  if and only if the Gramian matrix  $G_T$  is invertible.
- Assume that  $t \mapsto A(t)$  and  $t \mapsto B(t)$  are  $C^\infty$ . If (1.6) is satisfied then the control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$  is locally controllable in time  $T$  around  $R(T, 0)x_0$ ; the converse is true if  $t \mapsto A(t)$  and  $t \mapsto B(t)$  are analytic.

### 1.1.3 Geometry of Accessible Sets

**Theorem 1.5** Consider the control system  $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$  in  $\mathbb{R}^n$  with controls  $u$  taking their values in a compact subset  $\Omega$  of  $\mathbb{R}^m$ . For every  $x_0 \in \mathbb{R}^n$ , for every  $t \geq 0$ , the accessible set  $\text{Acc}_\Omega(x_0, t)$  is compact, convex and depends continuously on  $t$  for the Hausdorff topology.<sup>1</sup>

**Remark 1.10** Note that the convexity of the accessible set holds true even though  $\Omega$  is not assumed to be convex. This property is not obvious and follows from a Lyapunov lemma (itself based on the Krein-Milman theorem in infinite dimension; see [7]). Actually this argument leads to  $\text{Acc}_\Omega(x_0, t) = \text{Acc}_{\text{Conv}(\Omega)}(x_0, t) = \text{Acc}_{\partial\Omega}(x_0, t)$ , where  $\text{Conv}(\Omega)$  is the convex closure of  $\Omega$  and  $\partial\Omega$  is the boundary of  $\Omega$ . This illustrates the so-called *bang-bang principle* (see Sect. 2.3.1).

In infinite dimension those questions are much more difficult (see [14]).

**Proof of Theorem 1.5** We first assume that  $\Omega$  is convex. In this case, we have

$$\text{Acc}_\Omega(x_0, t) = R(t, 0)x_0 + \int_0^t R(t, s)r(s) ds + L_t(L^\infty([0, T], \Omega))$$

where the linear continuous operator  $L_t : L^\infty([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n$  is defined by  $L_t u = \int_0^t R(t, s)B(s)u(s) ds$ . The convexity of  $\text{Acc}_\Omega(x_0, t)$  follows by linearity from the convexity of the set  $L^\infty([0, T], \Omega)$ .

Let us now prove the compactness of  $\text{Acc}_\Omega(x_0, t)$ . Let  $(x_n^1)_{n \in \mathbb{N}}$  be a sequence of points of  $\text{Acc}_\Omega(x_0, t)$ . For every  $n \in \mathbb{N}$ , let  $u_n \in L^\infty([0, T], \Omega)$  be a control steering the system from  $x_0$  to  $x_n^1$  in time  $t$ , and let  $x_n(\cdot)$  be the corresponding trajectory. We have

$$x_n^1 = x_n(t) = R(t, 0)x_0 + \int_0^t R(t, s)(B(s)u_n(s) + r(s)) ds. \quad (1.7)$$

Since  $\Omega$  is compact, the sequence  $(u_n)_{n \in \mathbb{N}}$  is bounded in  $L^2([0, T], \mathbb{R}^m)$ . Since this space is reflexive (see [2]), by weak compactness we infer that a subsequence of  $(u_n)_{n \in \mathbb{N}}$  converges weakly to some  $u \in L^2([0, T], \mathbb{R}^m)$ . Since  $\Omega$  is assumed to be convex, we have moreover that  $u \in L^2([0, T], \Omega)$  (note that one has also  $u \in L^\infty([0, T], \Omega)$  because  $\Omega$  is compact). Besides, using (1.7) and the control system we easily see that the sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  is bounded in  $H^1([0, t], \mathbb{R}^n)$ . Since this Sobolev space is reflexive and is compactly imbedded in  $C^0([0, t], \mathbb{R}^n)$ , we deduce that a subsequence of  $(x_n(\cdot))_{n \in \mathbb{N}}$  converges uniformly to some  $x(\cdot)$  on

<sup>1</sup> Denoting by  $d$  the Euclidean distance of  $\mathbb{R}^n$ , given any two compact subsets  $K_1$  and  $K_2$  of  $\mathbb{R}^n$ , the Hausdorff distance  $d_H$  between  $K_1$  and  $K_2$  is defined by

$$d_H(K_1, K_2) = \sup \left( \sup_{y \in K_2} d(y, K_1), \sup_{y \in K_1} d(y, K_2) \right).$$



$[0, t]$ . Passing to the limit in (1.7), we get

$$x(t) = R(t, 0)x_0 + \int_0^t R(t, s)(B(s)u(s) + r(s)) ds$$

and in particular (a subsequence of)  $x_n^1 = x_n(t)$  converges to  $x(t) \in \text{Acc}_\Omega(x_0, t)$ . The compactness property is proved.

Let us prove the continuity in time of  $\text{Acc}_\Omega(x_0, t)$  in Hausdorff topology, i.e., that for any  $\varepsilon > 0$  there exists  $\delta > 0$  such that, for all  $t_1, t_2 \in \mathbb{R}$  satisfying  $0 \leq t_1 < t_2$  and  $t_2 - t_1 \leq \delta$ , we have:

- (1)  $d(y, \text{Acc}_\Omega(x_0, t_1))$  for every  $y \in \text{Acc}_\Omega(x_0, t_2)$ ;
- (2)  $d(y, \text{Acc}_\Omega(x_0, t_2))$  for every  $y \in \text{Acc}_\Omega(x_0, t_1)$ .

Let us prove (1) ((2) being similar). Let  $y \in \text{Acc}_\Omega(x_0, t_2)$ . It suffices to prove that there exists  $z \in \text{Acc}_\Omega(x_0, t_1)$  such that  $d(y, z) \leq \varepsilon$ . By definition of  $\text{Acc}_\Omega(x_0, t_2)$ , there exists  $u \in L^\infty([0, T], \Omega)$  such that the corresponding trajectory, starting at  $x_0$ , satisfies  $x(t_2) = y$ . Then  $z = x(t_1)$  is suitable. Indeed,

$$\begin{aligned} x(t_2) - x(t_1) &= R(t_2, 0)x_0 + \int_0^{t_2} R(t_2, s)(B(s)u(s) + r(s)) ds \\ &\quad - R(t_1, 0)x_0 - \int_0^{t_1} R(t_1, s)(B(s)u(s) + r(s)) ds \\ &= \int_{t_1}^{t_2} R(t_2, s)(B(s)u(s) + r(s)) ds \\ &\quad + (R(t_2, 0) - R(t_1, 0)) \left( x_0 + \int_0^{t_1} R(0, s)(B(s)u(s) + r(s)) ds \right) \end{aligned}$$

where we have used that  $R(t_i, s) = R(t_i, 0)R(0, s)$ . If  $t_2 - t_1$  is small then the first term of the above sum is small by continuity, and the second term is small by continuity of  $t \mapsto R(t, 0)$ . The result follows.

In the general case where  $\Omega$  is only compact (but not necessarily convex), the proof is more difficult and uses the Lyapunov lemma in measure theory (see, e.g., [10, Lemma 4A p. 163]) and more generally the Aumann theorem (see, e.g., [7]), from which, recalling that  $L_T u = \int_0^T R(T, t)B(t)u(t) dt$ , it follows that

$$\begin{aligned} \{L_T u \mid u \in L^\infty([0, T], \Omega)\} &= \{L_T u \mid u \in L^\infty([0, T], \partial\Omega)\} \\ &= \{L_T u \mid u \in L^\infty([0, T], \text{Conv}(\Omega))\} \end{aligned}$$

and moreover that these sets are compact and convex. The result follows.  $\square$

**Remark 1.11** Theorem 1.5 allows one to define the concept of *minimal time*: when there are some compact control constraints, due to the compactness of the accessible

set, one cannot steer the control system from  $x_0$  to another point  $x_1$  in arbitrary small time; a minimal positive time is due.

Another question of interest is to know whether the control system is controllable, in time not fixed, that is: when is the union of all sets  $\text{Acc}_\Omega(x_0, t)$ , over  $t \geq 0$ , equal to the whole  $\mathbb{R}^n$ ? This question is difficult.

## 1.2 Controllability of Nonlinear Systems

### 1.2.1 Local Controllability Results

**Preliminaries: End-Point Mapping** It is easy<sup>2</sup> to establish that the set  $\mathcal{U}_{x_0, T, \mathbb{R}^m}$ , endowed with the standard topology of  $L^\infty([0, T], \mathbb{R}^m)$ , is open, and that the end-point mapping  $E_{x_0, T}$  (see Definition 1.1) is of class  $C^1$  on  $\mathcal{U}_{x_0, T, \mathbb{R}^m}$  (it is  $C^p$  whenever  $f$  is  $C^p$ ).

Note that, for every  $t \geq 0$ , the accessible set is  $\text{Acc}_\Omega(x_0, t) = E_{x_0, t}(\mathcal{U}_{x_0, t, \Omega})$ .

In what follows we often denote by  $x_u(\cdot)$  a trajectory solution of (1.1) corresponding to the control  $u$ .

**Proposition 1.5** *Let  $x_0 \in \mathbb{R}^n$  and let  $u \in \mathcal{U}_{x_0, T, \mathbb{R}^m}$ . The (Fréchet) differential<sup>3</sup>  $dE_{x_0, T}(u) : L^\infty([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n$  is given by*

$$dE_{x_0, T}(u) \cdot \delta u = \delta x(T) = \int_0^T R(T, t) B(t) \delta u(t) dt$$

where  $\delta x(\cdot)$  is the solution of the so-called linearized system along  $(x_u(\cdot), u(\cdot))$ ,

$$\delta \dot{x}(t) = A(t) \delta x(t) + B(t) \delta u(t), \quad \delta x(0) = 0,$$

with

$$A(t) = \frac{\partial f}{\partial x}(t, x_u(t), u(t)), \quad B(t) = \frac{\partial f}{\partial u}(t, x_u(t), u(t)),$$

(which are respectively of size  $n \times n$  and  $n \times m$ ), and  $R(\cdot, \cdot)$  is the state-transition matrix of the linearized system, defined by (1.3).

<sup>2</sup> This follows from usual finite-time blow-up arguments on ordinary differential equations, and from the usual Picard-Lindelöf (Cauchy-Lipschitz) theorem with parameters, the parameter being here a control in a Banach set (see for instance [12]).

<sup>3</sup> We recall that, given two Banach spaces  $X$  and  $Y$ , a mapping  $F : X \rightarrow Y$  is said to be Fréchet differentiable at  $x \in X$  if there exists a linear continuous mapping  $dF(x) : X \rightarrow Y$  such that  $F(x+h) = F(x) + dF(x) \cdot h + o(h)$  for every  $h \in X$ . Here, the notation  $dF(x) \cdot h$  means that the linear operator  $dF(x)$  is applied to  $h$ .

**Proof of Proposition 1.5** We have  $E_{x_0,T}(u + \delta u) = E_{x_0,T}(u) + dE_{x_0,T}(u).\delta u + o(\delta u)$  by definition of the Fréchet differential. In this first-order Taylor expansion,  $E_{x_0,T}(u) = x_u(T)$  and  $E_{x_0,T}(u + \delta u) = x_{u+\delta u}(T)$ . We want to compute  $dE_{x_0,T}(u).\delta u$ , which is equal, at the first order, to  $x_{u+\delta u}(T) - x_u(T)$ . In what follows, we set  $\delta x(t) = x_{u+\delta u}(t) - x_u(t)$ . We have

$$\begin{aligned} \delta \dot{x}(t) &= f(t, x_{u+\delta u}(t), u(t) + \delta u(t)) - f(t, x_u(t), u(t)) \\ &= f(t, x_u(t) + \delta x(t), u(t) + \delta u(t)) - f(t, x_u(t), u(t)) \\ &= \frac{\partial f}{\partial x}(t, x_u(t), u(t)).\delta x(t) + \frac{\partial f}{\partial u}(t, x_u(t), u(t)).\delta u(t) + o(\delta x(t), \delta u(t)) \end{aligned}$$

so that, at the first order, we identify the linearized system. By integration (note that the remainder terms can be rigorously handled by standard Gronwall arguments, not detailed here), we get  $\delta x(T) = \int_0^T R(T, t)B(t)\delta u(t) dt$ , as expected. Note that this term provides a linear continuous operator and thus is indeed the Fréchet differential of the end-point mapping.  $\square$

**Remark 1.12** This theorem says that the differential of the end-point mapping at  $u$  is the end-point mapping of the linearized system along  $(x_u(\cdot), u(\cdot))$ . This is similar to the well known result in dynamical systems theory, stating that the differential of the flow is the flow of the linearized system. This remark has interesting consequences in terms of local controllability properties.

**Local Controllability Results Along a Trajectory** Let  $x_0 \in \mathbb{R}^n$  and let  $\bar{u} \in \mathcal{U}_{x_0,T,\mathbb{R}^m}$  be arbitrary. According to Remark 1.12, if the linearized system along  $(x_{\bar{u}}(\cdot), \bar{u}(\cdot))$  is controllable in time  $T$ , then the end-point mapping of the linearized system is surjective, meaning that the linear continuous mapping  $dE_{x_0,T}(\bar{u}) : L^\infty([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n$  is surjective. It follows from an implicit function argument (surjective mapping theorem) that the end-point mapping itself,  $E_{x_0,T} : \mathcal{U}_{x_0,T,\mathbb{R}^m} \rightarrow \mathbb{R}^n$ , is locally surjective and locally open at  $u$ .

The above argument works because we have considered controls taking their values in the whole  $\mathbb{R}^m$ . The argument still works whenever one considers a set  $\Omega$  of control constraints, provided that we have room to consider local variations of  $\bar{u}$ : this is true as soon as  $\bar{u}$  is in the interior of  $L^\infty([0, T], \Omega)$  for the topology of  $L^\infty([0, T], \mathbb{R}^m)$  (note that this condition is stronger than requiring that  $\bar{u}$  takes its values in the interior of  $\Omega$ ). We have thus obtained the following result (see Definition 1.3).

**Theorem 1.6** Let  $x_0 \in \mathbb{R}^n$  and let  $\bar{u} \in \mathcal{U}_{x_0,T,\Omega}$ . We denote by  $\bar{x}(\cdot) = x_{\bar{u}}(\cdot)$  the trajectory solution of (1.1), corresponding to the control  $\bar{u}(\cdot)$ , such that  $\bar{x}(0) = x_0$ , and we set  $\bar{x}_1 = \bar{x}(T) = E_{x_0,T}(\bar{u})$ . We assume that the function  $\bar{u}(\cdot)$  is in the interior of  $L^\infty([0, T], \Omega)$  for the topology of  $L^\infty([0, T], \mathbb{R}^m)$ .

If the linearized system along  $(\bar{x}(\cdot), \bar{u}(\cdot))$  is controllable in time  $T$ , then the nonlinear control system (1.1) is locally controllable from  $x_0$  in time  $T$  around

$x_1$ . More precisely, there exists an open neighborhood  $V$  of  $\bar{x}_1$  in  $\mathbb{R}^n$  and an open neighborhood  $U$  of  $\bar{u}(\cdot)$  in  $L^\infty([0, T], \mathbb{R}^m)$ , satisfying  $U \subset \mathcal{U}_{x_0, T, \Omega} \subset L^\infty([0, T], \Omega)$ , such that, for every  $x_1 \in V$ , there exists a control  $u \in U$  such that  $x_u(0) = x_0$  and  $x_1 = x_u(T) = E_{x_0, T}(u)$ .

**Remark 1.13** The controllability in time  $T$  of the linearized system  $\delta\dot{x}(t) = A(t)\delta x(t) + B(t)\delta u(t)$  along  $(\bar{x}(\cdot), \bar{u}(\cdot))$  can be characterized thanks to Theorems 1.3 and 1.4. We thus have explicit sufficient conditions for local controllability. Note that the conditions are not necessary (for instance, in 1D,  $\dot{x}(t) = u(t)^3$ , along  $\bar{u} = 0$ ).

**Example 1.4** Consider the Reeds-Shepp control system

$$\dot{x}(t) = v(t) \cos \theta(t), \quad \dot{y}(t) = v(t) \sin \theta(t), \quad \dot{\theta}(t) = u(t),$$

where the controls  $u$  and  $v$  are subject to the constraints  $|u| \leq 1$  and  $|v| \leq 1$ . We call *segment* any connected piece of trajectory along which  $u = \text{Cst} = 0$  and  $v = \text{Cst} = \bar{v}$  with  $\bar{v} \neq 0$ . We call *arc of a circle* any connected piece of trajectory along which  $u = \text{Cst} = \bar{u}$  and  $v = \text{Cst} = \bar{v}$  with  $\bar{u} \neq 0$  and  $\bar{v} \neq 0$ .

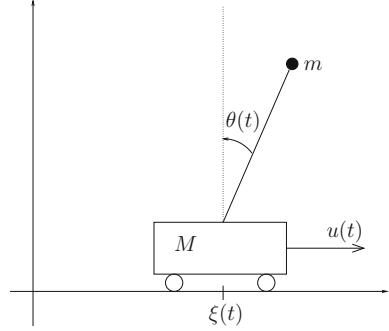
- Prove that the control system is locally controllable along any segment such that  $|\bar{v}| < 1$  (in time equal to that of the segment).
- Prove that the control system is locally controllable along any arc of a circle such that  $|\bar{u}| < 1$  and  $|\bar{v}| < 1$  (in time equal to that of the arc of a circle).
- Deduce that the system is globally controllable: for all  $(x_0, y_0, \theta_0) \in \mathbb{R}^3$  and  $(x_1, y_1, \theta_1) \in \mathbb{R}^3$ , there exist  $T > 0$  and controls  $u$  et  $v$  satisfying the constraints and generating a trajectory steering the system from  $(x_0, y_0, \theta_0)$  to  $(x_1, y_1, \theta_1)$  in time  $T$ . Describe such a strategy in a very simple way.

Let us next provide two important applications of Theorem 1.6 (which are particular cases): local controllability around a point, and the return method.

**Local Controllability Around an Equilibrium Point** Assume that the general control system (1.1) is *autonomous*, i.e., that  $f$  does not depend on  $t$ . Assume that  $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \mathbb{R}^m$  is an equilibrium point of  $f$ , i.e.,  $f(\bar{x}, \bar{u}) = 0$ . In that case, the constant trajectory defined by  $\bar{x}(t) = \bar{x}$  and  $\bar{u}(t) = \bar{u}$  is a solution of (1.1). The linearized system along this (constant) trajectory is given by

$$\delta\dot{x}(t) = A\delta x(t) + B\delta u(t)$$

with  $A = \frac{\partial f}{\partial x}(\bar{x}, \bar{u})$  and  $B = \frac{\partial f}{\partial u}(\bar{x}, \bar{u})$ . It follows from Theorem 1.1 that, if this linear autonomous control system is controllable (in time  $T$ ) then the nonlinear control system is locally controllable in time  $T$  around the point  $\bar{x}$ , i.e.,  $\bar{x}$  can be steered in time  $T$  to any point in some neighborhood. By reversing the time (which is possible because we are here in finite dimension), the converse can be done. We have thus obtained the following result.

**Fig. 1.1** Inverted pendulum

**Corollary 1.1** *With the above notations, assume that  $\text{rank } K(A, B) = n$  and that  $\bar{u} \in \overset{\circ}{\Omega}$  (interior of  $\Omega$ ). Then, for every  $T > 0$ , the control system  $\dot{x}(t) = f(x(t), u(t))$  is locally controllable in time  $T$  around the point  $\bar{x}$  in the following sense: for every  $T > 0$  there exist an open neighborhood  $V$  of  $\bar{x}$  in  $\mathbb{R}^n$  and an open neighborhood  $W$  of  $\bar{u}$  in  $\mathbb{R}^m$ , satisfying  $W \subset \Omega$  and  $L^\infty([0, T], W) \subset \mathcal{U}_{x_0, T, \Omega} \subset L^\infty([0, T], \Omega)$ , such that, for all  $x_0, x_1 \in V$ , there exists a control  $u \in L^\infty([0, T], W)$  such that  $x_u(0) = x_0$  and  $x_u(T) = x_1$ .*

**Example 1.5** Consider the control system

$$\dot{x}(t) = x(t)(1 - x(t))(x(t) - \theta(t)), \quad \dot{\theta}(t) = u(t),$$

where the control  $u$  is subject to the constraint  $|u| \leq 1$ . The state  $x(t)$  represents the density of the population of an ecological system and the state  $\theta(t)$ , of which we control the derivative, is called the *Allee parameter* of the system. Let  $\bar{\theta} \in (0, 1)$  and  $T > 0$ . Prove that the control system is locally controllable in time  $T$  around the equilibrium point  $(x, \theta, u) = (\bar{\theta}, \bar{\theta}, 0)$ .

**Example 1.6** Let us consider the famous example of the inverted pendulum (see Fig. 1.1) of mass  $m$ , attached to a carriage of mass  $M$  whose acceleration  $u(t)$  is controlled, with constraint  $|u(t)| \leq 1$ . The control system is

$$\begin{aligned} \ddot{\xi} &= \frac{ml\dot{\theta}^2 \sin \theta - mg \cos \theta \sin \theta + u}{M + m \sin^2 \theta}, \\ \ddot{\theta} &= \frac{-ml\dot{\theta}^2 \sin \theta \cos \theta + (M + m)g \sin \theta - u \cos \theta}{l(M + m \sin^2 \theta)}, \end{aligned}$$

i.e., this is a system in dimension 4, of state  $x = (\xi, \dot{\xi}, \theta, \dot{\theta})^\top$  (see [5, 13] for the derivation of those equations). The linearized control system at any equilibrium

point  $(\bar{\xi}, 0, 0, 0)^\top$  is given by the pair of matrices

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{lM} & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ \frac{1}{M} \\ 0 \\ -\frac{1}{lM} \end{pmatrix}.$$

The Kalman condition is easily verified. Corollary 1.1 implies that, for any  $T > 0$ , this control system is locally controllable in time  $T$  around the equilibrium.

**Example 1.7** Consider the control system

$$\dot{x}_1(t) = x_2(t) + u_1(t), \quad \dot{x}_2(t) = x_1(t)x_3(t) + u_2(t), \quad \dot{x}_3(t) = -x_1(t)x_2(t),$$

where the state is  $x(t) = (x_1(t), x_2(t), x_3(t))^\top \in \mathbb{R}^3$  and the control is  $u(t) = (u_1(t), u_2(t)) \in \mathbb{R}^2$ . This system stands for the real-valued Maxwell-Bloch equations which model the interaction between light and matter and describe the dynamics of a real-valued two-state quantum system interacting with the electromagnetic mode of an optical resonator.

- (1) First, we see that all equilibrium points  $(\bar{x}, \bar{u})$  of the system are given by the parametrized families  $\mathcal{F}_1 = \{\bar{x} = (0, b, c), \bar{u} = (-b, 0) \mid b, c \in \mathbb{R}\}$  and  $\mathcal{F}_2 = \{\bar{x} = (a, 0, c), \bar{u} = (0, -ac) \mid a, c \in \mathbb{R}\}$ .
- (2) Computing the linearized system at a point of  $\mathcal{F}_1$  (resp., of  $\mathcal{F}_2$ ), we check that the Kalman condition implies that  $b \neq 0$  (resp.,  $a \neq 0$ ) is a necessary and sufficient condition for the controllability of this linearized system.
- (3) It follows from Corollary 1.1 that the Maxwell-Bloch system is locally controllable around any equilibrium point that is not of the form  $\bar{x} = (0, 0, c)$ ,  $\bar{u} = (0, 0)$ .

Many applications of Corollary 1.1 can be found in the literature. It can be noted that the proof, based on the implicit function theorem, is robust and withstands some generalizations in infinite dimension (possibly replacing the implicit function theorem with more appropriate results, such as the Kakutani fixed point theorem). The interested reader will easily find many examples and applications. We do not have room here to list or provide some of them.

**The Return Method** In Corollary 1.1, the sufficient condition is that the linearized system at the equilibrium point be controllable. Assume now that we are in a situation where the linearized system at the equilibrium point is not controllable, and however we would like to prove, using alternative conditions, that the nonlinear

control system is locally controllable. The idea of the so-called *return method*<sup>4</sup> is to assume that there exists a nontrivial loop trajectory of the control system, going from  $x_0$  to  $x_0$  in time  $T$ , along which the linearized control system is controllable. Then, Theorem 1.6 implies that the control system is locally controllable around  $x_0$ .

Note that the method is not restricted to equilibrium points. We have the following corollary.

**Corollary 1.2** *Let  $x_0 \in \mathbb{R}^n$ . Assume that there exists a trajectory  $\bar{x}(\cdot)$  of the control system (1.1), corresponding to a control  $\bar{u}(\cdot)$  on  $[0, T]$ , such that  $\bar{x}(0) = \bar{x}(T) = x_0$ . Assume that the function  $\bar{u}(\cdot)$  is in the interior of  $L^\infty([0, T], \Omega)$  for the topology of  $L^\infty([0, T], \mathbb{R}^m)$ . If the linearized system along  $(\bar{x}(\cdot), \bar{u}(\cdot))$  is controllable in time  $T$ , then the nonlinear control system (1.1) is locally controllable in time  $T$  around the point  $x_0$ : there exists an open neighborhood  $V$  of  $x_0$  in  $\mathbb{R}^n$  and an open neighborhood  $U$  of  $\bar{u}(\cdot)$  in  $L^\infty([0, T], \mathbb{R}^m)$ , satisfying  $U \subset \mathcal{U}_{x_0, T, \Omega} \subset L^\infty([0, T], \Omega)$ , such that, for every  $x_1 \in V$ , there exists a control  $u \in U$  such that  $x_u(0) = x_0$  and  $x_1 = x_u(T) = E_{x_0, T}(u)$ .*

**Example 1.8** Consider the Dubins car model

$$\dot{x}_1(t) = \cos \theta(t), \quad \dot{x}_2(t) = \sin \theta(t), \quad \dot{\theta}(t) = u(t),$$

with initial point  $x_1(0) = x_2(0) = 0$  and  $\theta(0) = 0 \in [2\pi]$ . The control is subject to the constraint  $|u| \leq M$  for some fixed  $M > 0$ . Let us prove that this control system is locally controllable at  $(0, 0, 0 \in [2\pi])$  in any time  $T > \frac{2\pi}{M}$ . Consider the reference trajectory given by

$$\bar{x}_1(t) = \frac{T}{2\pi} \sin \frac{2\pi t}{T}, \quad \bar{x}_2(t) = \frac{T}{2\pi} \left(1 - \cos \frac{2\pi t}{T}\right), \quad \bar{\theta}(t) = \frac{2\pi t}{T}, \quad \bar{u}(t) = \frac{2\pi}{T}$$

(the inequality  $T > \frac{2\pi}{M}$  implies that  $|\bar{u}| < M$ ). The linearized system along this trajectory is represented by the matrices

$$A(t) = \begin{pmatrix} 0 & 0 & -\sin \frac{2\pi t}{T} \\ 0 & 0 & \cos \frac{2\pi t}{T} \\ 0 & 0 & 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and it is easy to check (using Theorem 1.4) that this system is controllable in any time  $T > 0$ . The local controllability property in time  $T > \frac{2\pi}{M}$  follows.

Actually, it can be proved that the Dubins system is *not* locally controllable at  $(0, 0, 0 \in [2\pi])$  in time  $T \in [0, \frac{2\pi}{M}]$ .

---

<sup>4</sup> The return method was invented by J.-M. Coron first with the aim of stabilizing control systems with smooth instationary feedbacks. Then it was applied to many problems of control of PDEs in order to establish controllability properties. We refer the reader to [4].

In the previous paragraphs, we have derived, thanks to simple implicit function arguments, *local* controllability results. The local feature of these results is not a surprise, having in mind that, from a general point of view, it is expected that showing the global surjectivity of the nonlinear mapping  $E_{x_0, T}$  is difficult.

In view of that, we next provide two further issues.

### 1.2.2 Geometry of the Accessible Set

In view of better understanding why it is hopeless, in general, to get global controllability, it is useful to have a clear geometric picture of the accessible set. We have the following result, similar to Theorem 1.5.

**Theorem 1.7** *Let  $x_0 \in \mathbb{R}^n$  and let  $T > 0$ . We assume that:*

- $\Omega$  is compact;
- *there exists  $b > 0$  such that, for every admissible control  $u \in \mathcal{U}_{x_0, T, \Omega}$ , one has  $\|x_u(t)\| \leq b$  for every  $t \in [0, T]$ ;*
- *there exists  $c > 0$  such that  $\|f(t, x, u)\| \leq c$  for every  $t \in [0, T]$ , for every  $x \in \mathbb{R}^n$  such that  $\|x\| \leq b$  and for every  $u \in \Omega$ ;*
- *the set of velocities  $V(t, x) = \{f(t, x, u) \mid u \in \Omega\}$  is convex, for all  $(t, x)$ .*

*Then the set  $\text{Acc}_\Omega(x_0, t)$  is compact and varies continuously in time on  $[0, T]$  (for the Hausdorff topology).*

**Remark 1.14** The second assumption (uniform boundedness of trajectories) is done to avoid blow-up of trajectories. It is satisfied for instance if the dynamics  $f$  is sublinear at infinity. The third assumption is done for technical reasons in the proof, because at the beginning we assumed that  $f$  is locally integrable, only, with respect to  $t$ . The assumption of convexity of  $V(t, x)$  is satisfied for instance for control-affine systems (that is, whenever  $f$  is affine in  $u$ ) and if  $\Omega$  is moreover convex.

**Proof of Theorem 1.7** First of all, note that  $V(t, x)$  is compact, for all  $(t, x)$ , because  $\Omega$  is compact. Let us prove that  $\text{Acc}(x_0, t)$  is compact for every  $t \in [0, T]$ . It suffices to prove that every sequence  $(x_n)$  of points of  $\text{Acc}(x_0, t)$  has a converging subsequence. For every integer  $n$ , let  $u_n \in \mathcal{U}_{x_0, t, \Omega}$  be a control steering the system from  $x_0$  to  $x_n$  in time  $t$  and let  $x_n(\cdot)$  be the corresponding trajectory. We have  $x_n = x_n(t) = x_0 + \int_0^t f(s, x_n(s), u_n(s)) ds$ . Setting  $g_n(s) = f(s, x_n(s), u_n(s))$  for  $s \in [0, t]$ , using the assumptions, the sequence of functions  $(g_n(\cdot))_{n \in \mathbb{N}}$  is bounded in  $L^\infty([0, t], \mathbb{R}^n)$ , and therefore, up to some subsequence, it converges to some function  $g(\cdot)$  for the weak star topology of  $L^\infty([0, t], \mathbb{R}^n)$  (see [2]). For every  $\tau \in [0, t]$ , we set  $x(\tau) = x_0 + \int_0^\tau g(s) ds$ . Clearly,  $x(\cdot)$  is absolutely continuous on  $[0, t]$ , and  $\lim_{n \rightarrow +\infty} x_n(s) = x(s)$  for every  $s \in [0, t]$ , that is, the sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  converges pointwise to  $x(\cdot)$ . The objective is now to prove that  $x(\cdot)$  is



a trajectory associated with a control  $u$  taking its values in  $\Omega$ , that is, to prove that  $g(s) = f(s, x(s), u(s))$  for almost every  $s \in [0, t]$ .

To this aim, for every integer  $n$  and almost every  $s \in [0, t]$ , we set  $h_n(s) = f(s, x(s), u_n(s))$ , and we define the set

$$\mathcal{V} = \{h(\cdot) \in L^2([0, t], \mathbb{R}^n) \mid h(s) \in V(s, x(s)) \text{ for a.e. } s \in [0, t]\}.$$

Note that  $h_n \in \mathcal{V}$  for every integer  $n$ . For all  $(t, x)$ , the set  $V(t, x)$  is compact and convex, and, using the fact that, from any sequence converging strongly in  $L^2$  we can substract a subsequence converging almost everywhere, we infer that  $\mathcal{V}$  is convex and closed in  $L^2([0, t], \mathbb{R}^n)$  for the strong topology. Therefore  $\mathcal{V}$  is closed as well in  $L^2([0, t], \mathbb{R}^n)$  for the weak topology (see [2]). But like  $(g_n)_{n \in \mathbb{N}}$ , the sequence  $(h_n)_{n \in \mathbb{N}}$  is bounded in  $L^2$ , hence up to some subsequence it converges to some function  $h$  for the weak topology, and  $h$  must belong to  $\mathcal{V}$  since  $\mathcal{V}$  is weakly closed.

Finally, let us prove that  $g = h$  almost everywhere. We have

$$\int_0^t \varphi(s) g_n(s) ds = \int_0^t \varphi(s) h_n(s) ds + \int_0^t \varphi(s) (g_n(s) - h_n(s)) ds \quad (1.8)$$

for every  $\varphi \in L^2([0, t], \mathbb{R})$ . By assumption,  $f$  is globally Lipschitz in  $x$  on  $[0, T] \times \bar{B}(0, b) \times \Omega$ , and hence by the mean value inequality, there exists  $C > 0$  such that  $\|g_n(s) - h_n(s)\| \leq C \|x_n(s) - x(s)\|$  for almost every  $s \in [0, t]$ . The sequence  $(x_n)$  converge pointwise to  $x(\cdot)$ , hence, using the dominated convergence theorem, we infer that  $\int_0^t \varphi(s) (g_n(s) - h_n(s)) ds \rightarrow 0$  as  $n \rightarrow +\infty$ . Passing to the limit in (1.8), we obtain  $\int_0^t \varphi(s) g(s) ds = \int_0^t \varphi(s) h(s) ds$  for every  $\varphi \in L^2([0, t], \mathbb{R})$  and therefore  $g = h$  almost everywhere on  $[0, t]$ .

In particular, we have  $g \in \mathcal{V}$ , and hence for almost every  $s \in [0, t]$  there exists  $u(s) \in \Omega$  such that  $g(s) = f(s, x(s), u(s))$ . Applying a measurable selection lemma in measure theory (note that  $g \in L^\infty([0, t], \mathbb{R}^n)$ ),  $u(\cdot)$  can be chosen to be measurable on  $[0, T]$  (see [10, Lemmas 2A, 3A p. 161]).

In conclusion, the trajectory  $x(\cdot)$  is associated on  $[0, t]$  with the control  $u$  taking its values in  $\Omega$ , and  $x(t)$  is the limit of the points  $x_n$ . This shows the compactness of  $\text{Acc}_\Omega(x_0, t)$ .

It remains to establish the continuity of the accessible set with respect to time. Like in Theorem 1.5, let  $t_1$  and  $t_2$  be two real numbers such that  $0 < t_1 < t_2 \leq T$  and let  $x_2 \in \text{Acc}_\Omega(x_0, t_2)$ . By definition, there exists a control  $u$  taking its values in  $\Omega$ , generating a trajectory  $x(\cdot)$ , such that  $x_2 = x(t_2) = x_0 + \int_0^{t_2} f(t, x(t), u(t)) dt$ . The point  $x_1 = x(t_1) = x_0 + \int_0^{t_1} f(t, x(t), u(t)) dt$  belongs to  $\text{Acc}_\Omega(x_0, t_1)$ , and using the assumptions on  $f$ , we have  $\|x_2 - x_1\| \leq \text{Cst}|t_2 - t_1|$ . We conclude easily.  $\square$

### 1.2.3 Global Controllability Results

One can find global controllability results in the existing literature, established for particular classes of control systems. Let us provide here controllability results in the important class of control-affine systems.

We say that a control system is *control-affine* whenever the dynamics  $f$  is affine in  $u$ , in other words the control system is

$$\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t) f_i(x(t))$$

where the mappings  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $i = 0, \dots, m$  are smooth. The term  $f_0(x)$  is called a *drift*. Here, there is a crucial insight coming from differential geometry. We consider the mappings  $f_i$  are vector fields on  $\mathbb{R}^n$ . Such vector fields generate some flows, some integral curves, and at this point geometric considerations come into the picture.

There are many existing results in the literature, providing local or global controllability results under conditions on the Lie brackets of the vector fields.

We recall that the Lie bracket of two vector fields  $X$  and  $Y$  is defined either by  $[X, Y](x) = dY(x).X(x) - dX(x).Y(x)$ , or, recalling that a vector field is a first-order derivation on  $C^\infty(\mathbb{R}^n, \mathbb{R})$  defined by  $(Xf)(x) = df(x).X(x)$  for every  $f \in C^\infty(\mathbb{R}^n, \mathbb{R})$  (Lie derivative), by  $[X, Y] = XY - YX$  (it is obvious to check that it is indeed a first-order derivation). We also mention that, denoting by  $\exp(tX)$  and  $\exp(tY)$  the flows generated by the vector fields  $X$  and  $Y$ , the flows commute, i.e.,  $\exp(t_1 X) \circ \exp(t_2 Y) = \exp(t_2 Y) \circ \exp(t_1 X)$  for all times  $t_1$  and  $t_2$ , if and only if  $[X, Y] = 0$ . If the Lie bracket is nonzero then the flows do not commute, but we have the asymptotic expansion

$$\exp(-tY) \circ \exp(-tX) \circ \exp(tY) \circ \exp(tX)(x) = x + \frac{t^2}{2}[X, Y](x) + o(t^2)$$

as  $t \rightarrow 0$ . The left-hand side of that equality is the point obtained by starting at  $x$ , following the vector field  $X$  during a time  $t$ , then the vector field  $Y$  during a time  $t$ , then  $-X$  during a time  $t$ , and then  $-Y$  during a time  $t$ . What it says is that this loop is not closed! The lack of commutation is measured through the Lie bracket  $[X, Y]$ . For more details on Lie brackets, we refer the reader to any textbook of differential geometry. Without going further, we mention that the Campbell-Hausdorff formula gives a precise series expansion of  $Z$ , defined by  $\exp(X) \circ \exp(Y) = \exp(Z)$ , in terms of iterated Lie brackets of  $X$  and  $Y$ . The first terms are  $Z = X + Y + \frac{1}{2}[X, Y] + \dots$ .

Finally, we recall that the Lie algebra generated by a set of vector fields is the set of all possible iterated Lie brackets of these vector fields.

For control-affine systems without drift, we have the following well-known Chow-Rashevski theorem (also called Hörmander condition, or Lie Algebra Rank Condition), whose early versions can be found in [3, 11].

**Theorem 1.8** *Consider a control-affine system without drift in  $\mathbb{R}^n$ . Assume that  $\Omega = \mathbb{R}^m$  (no constraint on the control) and that the Lie algebra generated by the vector fields  $f_1, \dots, f_m$  is equal to  $\mathbb{R}^n$  (at any point). Then the system is globally controllable, in any time  $T$ .*

**Proof of Theorem 1.8** We sketch the proof in the case  $n = 3$  and  $m = 2$ , assuming that  $\text{rank}(f_1, f_2, [f_1, f_2]) = 3$  at any point. Let  $\lambda \in \mathbb{R}$ . We define the mapping

$$\varphi_\lambda(t_1, t_2, t_3) = \exp(\lambda f_1) \exp(t_3 f_2) \exp(-\lambda f_1) \exp(t_2 f_2) \exp(t_1 f_1)(x_0).$$

We have  $\varphi_\lambda(0) = x_0$ . Let us prove that, for  $\lambda \neq 0$  small enough,  $\varphi_\lambda$  is a local diffeomorphism at 0. From the Campbell-Hausdorff formula, we infer that  $\varphi_\lambda(t_1, t_2, t_3) = \exp(t_1 f_1 + (t_2 + t_3) f_2 + \lambda t_3 [f_1, f_2] + \dots)$ , hence  $\partial_{\partial t_1} \varphi_\lambda(0) = f_1(x_0)$ ,  $\partial_{\partial t_2} \varphi_\lambda(0) = f_2(x_0)$  and  $\partial_{\partial t_3} \varphi_\lambda(0) = f_2(x_0) + \lambda [f_1, f_2](x_0) + o(\lambda)$ . By assumption, it follows that  $d\varphi_\lambda$  is an isomorphism, and therefore  $\varphi_\lambda$  is a local diffeomorphism at 0. We conclude by an easy connectedness argument.  $\square$

We approach here the *geometric control theory*. The theorem above is one of the many existing results that can be obtained with Lie bracket considerations. We refer the reader to the textbook [9] for many results which are of a geometric nature. In particular this reference contains some material in order to treat the case of control-affine systems with drift (see also [4]). Note that, in presence of a drift  $f_0$ , an easy sufficient condition ensuring global controllability is that the Lie algebra generated by the controlled vector fields  $f_1, \dots, f_m$  be equal to  $\mathbb{R}^n$  (at any point); indeed, the rough idea is that, taking large controls, in some sense the drift term can be compensated and then one can apply Theorem 1.8.

**Example 1.9** The Heisenberg system in  $\mathbb{R}^3$

$$\dot{x}(t) = u_1(t), \quad \dot{y}(t) = u_2(t), \quad \dot{z}(t) = u_1(t)y(t) - u_2(t)x(t),$$

is represented by the two vector fields  $f_1 = \partial_x + y\partial_z$  and  $f_2 = \partial_y - x\partial_z$ . We have  $[f_1, f_2] = -2\partial_z$  and thus the Lie algebra condition is satisfied. Therefore this system is controllable.

**Remark 1.15** In practice, Lie brackets are often realized thanks to oscillating functions, like the sine function taken with a sufficiently large frequency (see [4]). One can find in the literature many results concerning the *motion planning problem*, which consists of designing simple enough controls realizing controllability. We refer to [8] for a survey on such techniques.

## References

1. Bonnard, B., Chyba, M.: Singular Trajectories and Their Role in Control Theory. Mathematics And its Applications, vol. 40. Springer-Verlag, Berlin (2003)
2. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext. Springer, New York (2011)
3. Chow, W.-L.: Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung. Math. Ann. **117**, 98–105 (1939)
4. Coron, J.-M.: Control and Nonlinearity. Mathematical Surveys and Monographs, vol. 136, xiv+426 pp. American Mathematical Society, Providence (2007)
5. D'Andréa-Novell, B., De Lara, M.: Control Theory for Engineers. Springer-Verlag, Berlin (2013)
6. Hale, J.K.: Ordinary Differential Equations, 2nd edn., xvi+361 pp. Robert E. Krieger Publishing Co., Inc., Huntington (1980)
7. Hermes, H., LaSalle, J.P.: Functional Analysis and Time Optimal Control. Mathematics in Science and Engineering, vol. 56. Academic Press, Cambridge (1969)
8. Jean, F.: Control of Nonholonomic Systems: From Sub-Riemannian Geometry to Motion Planning. SpringerBriefs in Mathematics, x+104 pp. Springer, Cham (2014)
9. Jurdjevic, V.: Geometric Control Theory. Cambridge Studies in Advanced Mathematics, vol. 52. Cambridge University Press, Cambridge (1997)
10. Lee, E.B., Markus, L.: Foundations of Optimal Control Theory. John Wiley, New York (1967)
11. Rashevski, P.K.: About connecting two points of complete nonholonomic space by admissible curve. Uch. Zapiski Ped. Inst. Libknexa **2**, 83–94 (1938)
12. Sontag, E.D.: Mathematical Control Theory: Deterministic Finite-Dimensional Systems. Texts in Applied Mathematics, vol. 6, 2nd edn., xvi+531. Springer-Verlag, New York (1998)
13. Trélat, E.: Contrôle Optimal (French) [Optimal Control]. Théorie & Applications [Theory and Applications]. Mathématiques Concrètes [Concrete Mathematics], vi+246 pp. Vuibert, Paris (2005)
14. Wang, G., Wang, L., Xu, Y., Zhang, Y.: Time Optimal Control of Evolution Equations. The Progress in Nonlinear Differential Equations and Their Applications, Subseries in Control, vol. 92, xvi+334 pp. Birkhäuser/Springer, Cham (2018)

# Chapter 2

## Optimal Control



In Chap. 1, we have provided controllability properties for general classes of control systems. Considering some control problem of trying to reach some final configuration for the control system (1.1), from some initial configuration, with an admissible control, it happens that, in general, there exists an infinite number of controls making the job (think of all possibilities of realizing a parallel parking, for instance). Among this infinite number of controls, we now would like to select (at least) one control, achieving the desired controllability problem, and moreover minimizing some cost criterion (for instance, one would like to realize the parallel parking by minimizing the time, or by minimizing the fuel consumption). This is then an optimal control problem.

The main objective of this chapter is to formulate the *Pontryagin maximum principle*, which is the milestone of optimal control theory. It provides first-order necessary conditions for optimality, which allow one to compute or at least parametrize the optimal trajectories.

Let us give the general framework that will be used throughout the chapter.

Let  $n$  and  $m$  be two positive integers. We consider a control system in  $\mathbb{R}^n$

$$\dot{x}(t) = f(t, x(t), u(t)) \quad (2.1)$$

where  $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  is of class  $C^1$ , and the controls are measurable essentially bounded functions of time taking their values in some measurable subset  $\Omega$  of  $\mathbb{R}^m$  (set of control constraints).

Let  $f^0 : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$  and  $g : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$  be functions of class  $C^1$ . For every  $x_0 \in \mathbb{R}^n$ , for every  $t_f \geq 0$ , and for every admissible control  $u \in \mathcal{U}_{x_0, t_f, \Omega}$  (see Sect. 1.2), the *cost* of the trajectory  $x(\cdot)$ , solution of (2.1), corresponding to the control  $u$  and such that  $x(0) = x_0$ , is defined by

$$C_{x_0, t_f}(u) = \int_0^{t_f} f^0(t, x(t), u(t)) dt + g(t_f, x(t_f)). \quad (2.2)$$

Many variants of a cost can be given, anyway the one above is already quite general and covers a very large class of problems. If needed, one could easily add some term penalizing the initial point. Note also that the term  $g(t_f, x(t_f))$  could as well be written in integral form and thus be put in the definition of the function  $f^0$ ; however we prefer to keep this formulation that we find convenient in many situations.

Let  $M_0$  and  $M_1$  be two measurable subsets of  $\mathbb{R}^n$ . We consider the optimal control problem (denoted in short **(OCP)** in what follows) of determining a trajectory  $x(\cdot)$ , defined on  $[0, t_f]$  (where the final time  $t_f$  can be fixed or not in **(OCP)**), corresponding to an admissible control  $u \in \mathcal{U}_{x(0), t_f, \Omega}$ , solution of (2.1), such that  $x(0) \in M_0$  and  $x(t_f) \in M_1$  and minimizing the cost (2.2) over all possible trajectories steering the control system from  $M_0$  to  $M_1$  in time  $t_f$ .

This is a general nonlinear optimal control problem, but without any state constraints. We could restrict the set of trajectories by imposing some pointwise constraints on  $x(t)$  (a region of the state space may be forbidden). Such constraints are however not easily tractable in the Pontryagin maximum principle and make the analysis much more difficult (see Sect. 2.2.3).

## 2.1 Existence of an Optimal Control

Although it is not very useful, let us state a general result ensuring the existence of an optimal solution of **(OCP)**. In the theorem below, note that we can assume that  $f^0$  and  $g$  are only continuous. Here, there is no additional difficulty in adding some state constraints.

**Theorem 2.9** *We consider **(OCP)** and we assume that:*

- $\Omega$  is compact,  $M_0$  and  $M_1$  are compact;
- $M_1$  is reachable from  $M_0$ , that is, there exists a trajectory (corresponding to an admissible control) steering the system from  $M_0$  to  $M_1$ ;
- there exist  $b > 0$  and  $c > 0$  such that  $\|f(t, x, u)\| + |f^0(t, x, u)| \leq c$  for every  $t \in [0, b]$ , every  $x \in \mathbb{R}^n$  satisfying  $\|x\| \leq b$ , and every  $u \in \Omega$ ; moreover, for every trajectory  $x(\cdot)$  defined on  $[0, t_f]$  and steering the system from  $M_0$  to  $M_1$ , one has  $t_f \leq b$  and  $\|x(t)\| \leq b$  for every  $t \in [0, t_f]$  (no blow-up);
- the epigraph of extended velocities

$$\tilde{V}(t, x) = \left\{ \left( \begin{array}{c} f(t, x, u) \\ f^0(t, x, u) + \gamma \end{array} \right) \mid u \in \Omega, \gamma \geq 0 \right\} \quad (2.3)$$

is convex, for all  $(t, x)$ .

We assume moreover in **(OCP)** that the trajectories are subject to state constraints  $c_i(t, x(t)) \leq 0$ , where the  $c_i$ ,  $i \in \{1, \dots, r\}$  are continuous functions defined on  $\mathbb{R} \times \mathbb{R}^n$ .

Then **(OCP)** has at least one solution.

If the final time is fixed in **(OCP)** then we assume that  $M_1$  is reachable from  $M_0$  exactly in time  $t_f$ . Note that it is easy to generalize this result to more general situations, for instance the sets  $M_0$ ,  $M_1$  and  $\Omega$  could depend on  $t$  (see [24], and more generally see [11] for many variants of existence results).

Such existence results are however often difficult to apply in practice because of the strong assumption (2.3) (not satisfied in general as soon as  $f$  is “too much” nonlinear). In practice, we often apply the Pontryagin maximum principle (that we will see next), without being sure a priori that there exist an optimal solution. If we can solve the resulting necessary conditions, then this often gives a way for justifying that indeed, a posteriori, there exists an optimal solution.

**Proof of Theorem 2.9** The proof is similar to the one of Theorem 1.7.

Let  $\delta$  be the infimum of costs  $C(u)$  over the set of admissible controls  $u \in L^\infty(0, t(u); \Omega)$  generating trajectories such that  $x(0) \in M_0$ ,  $x(t(u)) \in M_1$  and satisfying the state constraints  $c_i(x(\cdot)) \leq 0$ ,  $i = 1, \dots, r$ . Let us consider a minimizing sequence of trajectories  $x_n(\cdot)$  associated with controls  $u_n$ , that is, a sequence of trajectories satisfying all constraints and such that  $C(u_n) \rightarrow \delta$  as  $n \rightarrow +\infty$ . For every integer  $n$ , we set

$$\tilde{F}_n(t) = \begin{pmatrix} f(t, x_n(t), u_n(t)) \\ f^0(t, x_n(t), u_n(t)) \end{pmatrix} = \begin{pmatrix} F_n(t) \\ F_n^0(t) \end{pmatrix}$$

for almost every  $t \in [0, t(u_n)]$ . From the assumptions, the sequence of functions  $(\tilde{F}_n(\cdot))_{n \in \mathbb{N}}$  (extended by 0 on  $(t_n(u), b]$ ) is bounded in  $L^\infty(0, b; \mathbb{R}^n)$ , and hence, up to some subsequence, it converges to some function  $\tilde{F}(\cdot) = (F(\cdot), F^0(\cdot))^\top$  for the weak star topology of  $L^\infty(0, b; \mathbb{R}^{n+1})$  (see [9]). Also, up to some subsequence, the sequence  $(t_n(u_n))_{n \in \mathbb{N}}$  converges to some  $T \geq 0$ , and we have  $\tilde{F}(t) = 0$  for  $t \in (T, b]$ . Finally, by compactness of  $M_0$ , up to some subsequence, the sequence  $(x_n(0))_{n \in \mathbb{N}}$  converges to some  $x_0 \in M_0$ . For every  $t \in [0, T]$ , we set  $x(t) = x_0 + \int_0^t F(s) ds$ , and then  $x(\cdot)$  is absolutely continuous on  $[0, T]$ . Moreover, for every  $t \in [0, T]$ , we have  $\lim_{n \rightarrow +\infty} x_n(t) = x(t)$ , that is, the sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  converges pointwise to  $x(\cdot)$ . As in the proof of Theorem 1.7, the objective is then to prove that the trajectory  $x(\cdot)$  is associated with a control  $u$  taking its values in  $\Omega$ , and that this control is moreover optimal.

We set  $\tilde{h}_n(t) = (f(t, x(t), u_n(t)), f^0(t, x(t), u_n(t)))^\top$  for every integer  $n$  and for almost every  $t \in [0, t(u_n)]$ . If  $T > t(u_n)$ , then we extend  $\tilde{h}_n$  on  $[0, T]$  by  $\tilde{h}_n(t) = (f(t, x(t), v), f^0(t, x(t), v))^\top$  for some arbitrary  $v \in \Omega$ . Besides, we define (note that  $\Omega$  is compact)

$$\beta = \max\{|f^0(t, x, u)| \mid 0 \leq t \leq b, \|x\| \leq b, u \in \Omega\}.$$

For every  $(t, x) \in \mathbb{R}^{1+n}$ , we then slightly modify the definition of  $\tilde{V}(t, x)$  to make it compact (keeping it convex), by setting

$$\tilde{V}_\beta(t, x) = \left\{ \begin{pmatrix} f(t, x, u) \\ f^0(t, x, u) + \gamma \end{pmatrix} \mid u \in \Omega, \gamma \geq 0, |f^0(t, x, u) + \gamma| \leq \beta \right\}.$$

We define  $\tilde{\mathcal{V}} = \{\tilde{h}(\cdot) \in L^2([0, T], \mathbb{R}^{n+1}) \mid \tilde{h}(t) \in \tilde{V}_\beta(t, x(t)) \text{ for a.e. } t \in [0, T]\}$ . By construction, we have  $\tilde{h}_n \in \tilde{\mathcal{V}}$  for every integer  $n$ . At this step, we need a lemma:

**Lemma 2.2** *The set  $\tilde{\mathcal{V}}$  is convex and strongly closed in  $L^2([0, T], \mathbb{R}^{n+1})$ .*

**Proof of Lemma 2.2** Let us prove that  $\tilde{\mathcal{V}}$  is convex. Let  $\tilde{r}_1, \tilde{r}_2 \in \tilde{\mathcal{V}}$ , and let  $\lambda \in [0, 1]$ . By definition, for almost every  $t \in [0, T]$ , we have  $\tilde{r}_1(t) \in \tilde{V}_\beta(t, x(t))$  and  $\tilde{r}_2(t) \in \tilde{V}_\beta(t, x(t))$ . Since  $\tilde{V}_\beta(t, x(t))$  is convex, it follows that  $\lambda\tilde{r}_1(t) + (1 - \lambda)\tilde{r}_2(t) \in \tilde{V}_\beta(t, x(t))$ . Hence  $\lambda\tilde{r}_1 + (1 - \lambda)\tilde{r}_2 \in \tilde{\mathcal{V}}$ .

Let us prove that  $\tilde{\mathcal{V}}$  is strongly closed in  $L^2([0, T], \mathbb{R}^n)$ . Let  $(\tilde{r}_n)_{n \in \mathbb{N}}$  be a sequence of  $\tilde{\mathcal{V}}$  converging to  $\tilde{r}$  for the strong topology of  $L^2([0, T], \mathbb{R}^n)$ . Let us prove that  $\tilde{r} \in \tilde{\mathcal{V}}$ . Up to some subsequence,  $(\tilde{r}_n)_{n \in \mathbb{N}}$  converges almost everywhere to  $\tilde{r}$ , but by definition, for almost every  $t \in [0, T]$  we have  $\tilde{r}_n(t) \in \tilde{V}_\beta(t, x(t))$ , and  $\tilde{V}_\beta(t, x(t))$  is compact, hence  $\tilde{r}(t) \in \tilde{V}_\beta(t, x(t))$  for almost every  $t \in [0, T]$ . Lemma 2.2 is proved.  $\square$

We now continue the proof of Theorem 2.9. By Lemma 2.2, the set  $\tilde{\mathcal{V}}$  is convex and strongly closed and thus also weakly closed in  $L^2([0, T], \mathbb{R}^{n+1})$  (see [9]). The sequence  $(\tilde{h}_n)_{n \in \mathbb{N}}$  being bounded in  $L^2([0, T], \mathbb{R}^{n+1})$ , up to some subsequence, it converges weakly to some  $\tilde{h} \in \tilde{\mathcal{V}}$  since this set is weakly closed.

Let us prove that  $\tilde{F} = \tilde{h}$  almost everywhere. We have

$$\int_0^T \varphi(t) \tilde{F}_n(t) dt = \int_0^T \varphi(t) \tilde{h}_n(t) dt + \int_0^T \varphi(t) (\tilde{F}_n(t) - \tilde{h}_n(t)) dt \quad (2.4)$$

for every  $\varphi \in L^2(0, T)$ . By assumption, the mappings  $f$  and  $f^0$  are globally Lipschitz in  $x$  on  $[0, T] \times \bar{B}(0, b) \times \Omega$ , hence there exists  $C > 0$  such that  $\|\tilde{F}_n(t) - \tilde{h}_n(t)\| \leq C \|x_n(t) - x(t)\|$  for almost every  $t \in [0, T]$ . Since the sequence  $(x_n(\cdot))_{n \in \mathbb{N}}$  converges pointwise to  $x(\cdot)$ , by the dominated convergence theorem we infer that  $\int_0^T \varphi(t) (\tilde{F}_n(t) - \tilde{h}_n(t)) dt \rightarrow 0$  as  $n \rightarrow +\infty$ . Passing to the limit in (2.4), it follows that  $\int_0^T \varphi(t) \tilde{F}(t) dt = \int_0^T \varphi(t) \tilde{h}(t) dt$  for every  $\varphi \in L^2(0, T)$ , and therefore  $\tilde{F} = \tilde{h}$  almost everywhere on  $[0, T]$ .

In particular,  $\tilde{F} \in \tilde{\mathcal{V}}$ , and hence for almost every  $t \in [0, T]$  there exist  $u(t) \in \Omega$  and  $\gamma(t) \geq 0$  such that  $\tilde{F}(t) = (f(t, x(t), u(t)), f^0(t, x(t), u(t)) + \gamma(t))^T$ . Applying a measurable selection lemma (noting that  $\tilde{F} \in L^\infty([0, T], \mathbb{R}^{n+1})$ ), the functions  $u(\cdot)$  and  $\gamma(\cdot)$  can moreover be chosen to be measurable on  $[0, T]$  (see [24, Lem. 2A, 3A p. 161]).



It remains to prove that the control  $u$  is optimal for **(OCP)**. First of all, since  $x_n(t_n(u_n)) \in M_1$ , by compactness of  $M_1$  and using the convergence properties established above, we get that  $x(T) \in M_1$ . Similarly, we get, clearly, that  $c_i(x(\cdot)) \leq 0$ ,  $i = 1, \dots, r$ . Besides, by definition  $C(u_n)$  converges to  $\delta$ , and, using the convergence properties established above,  $C(u_n)$  converges as well to  $\int_0^T (f^0(t, x(t), u(t)) + \gamma(t)) dt + g(T, x(T))$ . Since  $\gamma$  takes nonnegative values, this implies that  $\int_0^T f^0(t, x(t), u(t)) dt + g(T, x(T)) \leq \int_0^T (f^0(t, x(t), u(t)) + \gamma(t)) dt + g(T, x(T)) \leq C(v)$ , for every admissible control  $v$  generating a trajectory steering the system from  $M_0$  to  $M_1$  and satisfying all constraints. Hence  $u$  is optimal and  $\gamma = 0$ . Theorem 2.9 is proved.  $\square$

## 2.2 Pontryagin Maximum Principle (PMP)

### 2.2.1 General Statement

The Pontryagin maximum principle (in short, PMP) states first-order necessary conditions for optimality.

**Theorem 2.10** *If  $(x(\cdot), u(\cdot))$  is an optimal solution of **(OCP)** on  $[0, t_f]$ , then there exist an absolutely continuous function  $p(\cdot) : [0, t_f] \rightarrow \mathbb{R}^n$  called adjoint vector and a real number  $p^0 \leq 0$ , with  $(p(\cdot), p^0) \neq (0, 0)$ , such that<sup>1</sup>*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t, x(t), p(t), p^0, u(t)), \quad \dot{p}(t) = -\frac{\partial H}{\partial x}(t, x(t), p(t), p^0, u(t)), \quad (2.5)$$

for almost every  $t \in [0, t_f]$ , where the function  $H : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$ , called Hamiltonian of **(OCP)**, is defined by

$$H(t, x, p, p^0, u) = \langle p, f(t, x, u) \rangle + p^0 f^0(t, x, u) \quad (2.6)$$

and we have the maximization condition

$$H(t, x(t), p(t), p^0, u(t)) = \max_{v \in \Omega} H(t, x(t), p(t), p^0, v) \quad (2.7)$$

for almost every  $t \in [0, t_f]$ .

<sup>1</sup> With rigorous notations, we should write (2.5) in the form  $\dot{x} = \frac{\partial H}{\partial p}^\top$  and  $\dot{p} = -\frac{\partial H}{\partial x}^\top$ , or equivalently,  $\dot{x} = \nabla_p H$  and  $\dot{p} = -\nabla_x H$  (as these are vectors in  $\mathbb{R}^n$ ). But we keep the writing (2.5), used in classical mechanics for Hamiltonian systems. In coordinates, this means that  $\dot{x}_i = \frac{\partial H}{\partial p_i}$  and  $\dot{p}_i = -\frac{\partial H}{\partial x_i}$  for every  $i \in \{1, \dots, n\}$ .

If the final time  $t_f$  is not fixed in **(OCP)**, then we have moreover

$$\max_{v \in \Omega} H(t_f, x(t_f), p(t_f), p^0, v) = -p^0 \frac{\partial g}{\partial t}(t_f, x(t_f)). \quad (2.8)$$

Moreover, the adjoint vector can be chosen such that we have the so-called transversality conditions (if they make sense)

$$p(0) \perp T_{x(0)} M_0 \quad (2.9)$$

$$p(t_f) - p^0 \nabla_x g(t_f, x(t_f)) \perp T_{x(t_f)} M_1 \quad (2.10)$$

where the notation  $T_x M$  stands for the usual tangent space to  $M$  at the point  $x$  (these conditions can be written as soon as the tangent space is well defined).

**Remark 2.16** If  $(p(\cdot), p^0)$  is a given adjoint vector satisfying the various conclusions stated in Theorem 2.10, then, for every  $\lambda > 0$ ,  $(\lambda p(\cdot), \lambda p^0)$  is also an adjoint vector satisfying the statements.

Note that we cannot take  $\lambda < 0$  since this would lead to a change of sign in the Hamiltonian and thus would impact the maximization condition (2.7). Actually, the historical choice made by Pontryagin is to take  $p^0 \leq 0$  in the statement: this leads to the *maximum principle* (the choice  $p^0 \geq 0$  is valid as well but in that case leads to a *minimum principle*).

A quadruple  $(x(\cdot), p(\cdot), p^0, u(\cdot))$  satisfying (2.5) and (2.7) is called an *extremal*. The PMP says that every optimal trajectory  $x(\cdot)$ , associated with a control  $u(\cdot)$ , is the projection onto  $\mathbb{R}^n$  of an extremal  $(x(\cdot), p(\cdot), p^0, u(\cdot))$ .

- If  $p^0 < 0$ , the extremal is said to be *normal*. In that case, it is usual (but not mandatory) to *normalize* the adjoint vector so that  $p^0 = -1$ .
- If  $p^0 = 0$ , the extremal is said to be *abnormal*.

The historical proof of the PMP stated in Theorem 2.10 can be found in [27]. As in [1, 5, 18, 24], it is based on the use of needle-like variations combined with a Brouwer fixed point argument. It is interesting to note that there are other proofs, based on the Ekeland variational principle (see [14]), on the Hahn-Banach theorem (see [8]). A concise sketch of proof, based on an implicit function argument (and using needle-like variations) can be found in [16]. As discussed in [7], all these different approaches of proof have their specificities. One approach or another may be preferred when trying to derive a PMP in a given context (for instance, the Ekeland approach is well adapted to derive versions of the PMP with state constraints, or in infinite dimension).

In Sect. 2.2.2, we give a proof of the PMP in the simplified context where  $\Omega = \mathbb{R}^m$  (no control constraint), or at least, under the assumption that the optimal control

$u$  belongs to the interior of  $L^\infty([0, t_f], \Omega)$ . In this case, (2.7) implies that

$$\frac{\partial H}{\partial u}(t, x(t), p(t), p^0, u(t)) = 0 \quad (2.11)$$

almost everywhere on  $[0, t_f]$  and the corresponding statement is sometimes called “weak PMP” (see [5, 29]). By the way, it is interesting to note that, in this context, a control  $u$  that is the projection of an abnormal extremal ( $p^0 = 0$ ) must satisfy  $p(t_f)^\top dE_{x_0, t_f}(u) = 0$  (see Sect. 2.2.2), i.e., is a singularity of the end-point mapping; in other words, the linearized control system along  $(x_u(\cdot), u(\cdot))$  is not controllable in time  $t_f$ .

In the general case where the optimal control  $u$  may saturate the constraints, the proof is more difficult and requires more technical developments such as needle-like variations, or a version of the implicit function theorem under constraints.

**Remark 2.17** In the conditions of Theorem 2.10, we have moreover

$$\frac{d}{dt} \max_{v \in \Omega} H(t, x(t), p(t), p^0, v) = \frac{\partial H}{\partial t}(t, x(t), p(t), p^0, u(t)) \quad (2.12)$$

for almost every  $t \in [0, t_f]$  (this can be proved by the Danskin theorem, using the fact that  $u$  takes its values in a compact subset of  $\Omega$ ).

In particular if **(OCP)** is autonomous, that is, if  $f$  and  $f^0$  do not depend on  $t$ , then  $H$  does not depend on  $t$  as well, and it follows from (2.12) that

$$\max_{v \in \Omega} H(x(t), p(t), p^0, v) = \text{Cst} \quad \forall t \in [0, t_f].$$

Note that this equality is valid for every (not only for almost every) time  $t \in [0, t_f]$  because the function  $t \mapsto \max_{v \in \Omega} H(x(t), p(t), p^0, v)$  is Lipschitz.

Note also that, in (2.7), the maximum over  $\Omega$  exists even when  $\Omega$  is not compact. This is part of the result and this is due to the fact that we have assumed that there exists an optimal solution.

**Remark 2.18** If  $g$  does not depend on  $t$  then (2.8) says that, roughly, if  $t_f$  is free then the (maximized) Hamiltonian vanishes at  $t_f$ . Note that if **(OCP)** is autonomous then this implies that  $H = 0$  along every extremal.

**Remark 2.19** If  $M_1 = \{x \in \mathbb{R}^n \mid F_1(x) = \dots = F_p(x) = 0\}$ , where the functions  $F_i$  are of class  $C^1$  on  $\mathbb{R}^n$ , then (2.10) implies that

$$\exists \lambda_1, \dots, \lambda_p \in \mathbb{R} \mid p(t_f) = \sum_{i=1}^p \lambda_i \nabla F_i(x(t_f)) + p^0 \nabla_x g(t_f, x(t_f)).$$

**Remark 2.20** The minimal time problem corresponds to choose either  $f^0 = 1$  and  $g = 0$ , or  $f^0 = 0$  and  $g(t, x) = t$ . In both cases it can be checked that the implied transversality conditions coincide.

**Remark 2.21 (On the Transversality Conditions)** Note that if  $M_0$  (or  $M_1$ ) is the singleton  $\{x_0\}$ , which means that the initial point is fixed in **(OCP)**, then the corresponding transversality condition is empty (since the tangent space is then reduced to the singleton  $\{0\}$ ).

On the contrary, if for instance  $M_1 = \mathbb{R}^n$ , which means that the final point is free in **(OCP)**, then the corresponding transversality condition yields that  $p(t_f) = p^0 \nabla_x g(t_f, x(t_f))$  (since the tangent space is then equal to  $\mathbb{R}^n$ ). In particular, in that case  $p^0$  cannot be equal to 0, for otherwise we would get  $p^0 = 0$  and  $p(t_f) = 0$ , which contradicts the fact that  $(p(\cdot), p^0)$  is nontrivial.

## 2.2.2 Proof in a Simplified Context

In this section, we consider a simplified version of **(OCP)** and we derive a weaker version of the PMP called “weak PMP”. The simplified framework is the following:

- $M_0 = \{x_0\}$  and  $M_1 = \{x_1\}$ , where  $x_0$  and  $x_1$  are two given points of  $\mathbb{R}^n$ . In other words, we consider a “point to point” control problem.
- $g = 0$  in the definition of the cost (2.2).
- The final time  $t_f = T$  is fixed.<sup>2</sup>

These three first simplifications are minor: it is easy to reduce a given optimal control problem to that case (see [29]). In contrast, the following one is major:

- $\Omega = \mathbb{R}^m$ , i.e., there are no control constraints; or, if there are some control constraints, we assume that the optimal control  $u$  is in the interior of  $L^\infty([0, T], \Omega)$  for the topology of  $L^\infty([0, T], \mathbb{R}^m)$ .

The latter assumption is the most important simplification. We will shortly comment further on the difficulties coming from control constraints.

First of all, we note that **(OCP)** is equivalent to the optimization problem

$$\min_{E_{x_0, T}(v)=x_1} C_{x_0, T}(v)$$

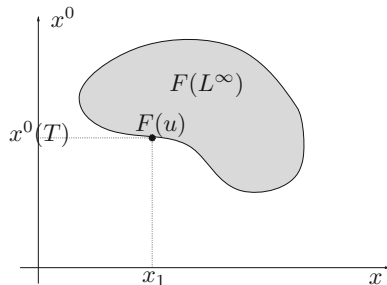
where  $E_{x_0, T}$  is the end-point mapping (see Definition 1.1) and  $C_{x_0, T}$  is the cost defined by 2.2. In this form, this is a nonlinear optimization problem with  $n$  equality constraints, in the infinite dimensional space of controls  $v \in \mathcal{U}_{x_0, T, \mathbb{R}^m} \subset L^\infty([0, T], \mathbb{R}^m)$ .

Let  $u \in \mathcal{U}_{x_0, T, \mathbb{R}^m}$  be an optimal control (here, we assume its existence, without making any further assumption on the dynamics). In order to derive first-order necessary conditions for optimality, we apply the well known Lagrange multipliers

---

<sup>2</sup> We are used to denote the final time by  $T$  when it is fixed in **(OCP)**, and by  $t_f$  when it is let free.

**Fig. 2.1** Range of the mapping  $F$



rule, which we recover as follows. Let us consider Fig. 2.1, in which we draw the range of the mapping  $F$  defined by  $F(v) = (E_{x_0, T}(v), C_{x_0, T}(v))$  with  $E_{x_0, T}(v) \in \mathbb{R}^n$  in abscissa and  $C_{x_0, T}(v) \in \mathbb{R}$  in ordinate. The range of  $F$  is thus seen as a subset of  $\mathbb{R}^n \times \mathbb{R}$ , whose shape is not important. We are interested in controls steering the system from  $x_0$  to  $x_1$ ; on the figure this corresponds to a point that is in the range of  $F$ , projecting onto  $x_1$ . Now the optimal control  $u$  corresponds on Fig. 2.1 to the point  $F(u)$ , which projects onto  $x_1$  and is at the boundary of the range of  $F$ . In other words, the necessary condition for optimality is:

$$u \text{ optimal} \Rightarrow F(u) \in \partial F(L^\infty([0, T], \Omega)).$$

Indeed, if  $F(u)$  were not at the boundary of  $F(L^\infty([0, T], \Omega))$  then this would imply that one can find another control, steering the system from  $x_0$  to  $x_1$  with a lower cost, which would contradict the optimality of  $u$ .

At this step, we use the important simplification  $\Omega = \mathbb{R}^m$ . Since  $F(u) \in \partial F(L^\infty([0, T], \mathbb{R}^m))$ , it follows from an implicit function argument (more precisely, the surjective mapping theorem) that  $dF(u) : L^\infty([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n \times \mathbb{R}$  is not surjective. Indeed, otherwise, the surjective mapping theorem would imply that  $F$  be locally surjective: in other words there would exist a neighborhood of  $F(u)$  in  $\mathbb{R}^n \times \mathbb{R}$  contained in  $F(L^\infty([0, T], \mathbb{R}^m))$ , which would contradict the fact that  $F(u) \in \partial F(L^\infty([0, T], \mathbb{R}^m))$ . Therefore,  $\text{Ran}(dF(u))$  is a proper subspace of  $\mathbb{R}^n \times \mathbb{R}$ .

Note that, when there are some control constraints, the above argument works as well provided  $u$  belongs to the interior of  $L^\infty([0, T], \Omega)$  for the topology of  $L^\infty([0, T], \mathbb{R}^m)$ . The argument is however no more valid whenever the control saturates the constraint, that is, whenever for instance the trajectory contains some sub-arc such that  $u(t) \in \partial\Omega$ . At least, to make it work we would need rather to use an implicit function theorem allowing one to take into account some constraints. Here is actually the main technical difficulty that one has to deal with in order to derive the strong version of the PMP. A usual proof consists of developing needle-like variations (see [27]), but except this (important) technical point, the structure of the proof remains the same, in particular an implicit function argument can still be used (see the sketch of proof in [16]).

Now, since  $\text{Ran}(dF(u))$  is a proper subspace of  $\mathbb{R}^n \times \mathbb{R}$ , there must exist  $\tilde{\psi} = (\psi, \psi^0) \in \mathbb{R}^n \times \mathbb{R} \setminus \{(0, 0)\}$  such that  $\tilde{\psi} \perp \text{Ran } dF(u)$ , i.e.,  $\tilde{\psi}^\top dF(u) = 0$  (here, for convenience  $dF(u)$  is identified to a matrix with  $n+1$  rows). In other words, we have obtained the usual Lagrange multipliers relation

$$\psi^\top dE_{x_0, T}(u) + \psi^0 dC_{x_0, T}(u) = 0. \quad (2.13)$$

Let us now exploit (2.13) (or, more exactly, the equation  $\tilde{\psi}^\top dF(u) = 0$ ). We define a new coordinate  $x^0$  and consider the differential equation  $\dot{x}^0(t) = f^0(t, x(t), u(t))$ , with the initial condition  $x^0(0) = 0$ . Therefore we have  $x^0(T) = C_{x_0, T}(u)$ . We define the *augmented state*  $\tilde{x} = (x, x^0) \in \mathbb{R}^{n+1}$  and the *augmented dynamics*  $\tilde{f}(t, \tilde{x}, v) = \begin{pmatrix} f(t, x, v) \\ f^0(t, x, v) \end{pmatrix}$ . We consider the *augmented control system* in  $\mathbb{R}^{n+1}$

$$\dot{\tilde{x}}(t) = \tilde{f}(t, \tilde{x}(t), v(t)). \quad (2.14)$$

Note that **(OCP)** is then equivalent to the optimal control problem of steering the system (2.14) from  $\tilde{x}_0 = (x_0, 0)$  to  $\tilde{x}_1 = (x_1, x^0(T))$  by minimizing  $x^0(T)$ .

Since  $F(v) = (E_{x_0, T}(v), C_{x_0, T}(v))$ , it follows that  $F$  is the end-point mapping for the augmented control system (2.14). Note that the range of  $F$  (drawn on Fig. 2.1) is the accessible set from  $x_0$  in time  $T$  for the augmented control system.

Using Proposition 1.5, the (Fréchet) differential  $dF(u) : L^\infty([0, T], \mathbb{R}^m) \rightarrow \mathbb{R}^n$  is given by

$$dF(u). \delta u = \int_0^T \tilde{R}(T, t) \tilde{B}(t) \delta u(t) dt \quad \forall \delta u \in L^\infty([0, T], \mathbb{R}^m)$$

where the (augmented) state transition matrix  $\tilde{R}(\cdot, \cdot)$  is defined as the solution of the Cauchy problem  $\partial_t \tilde{R}(t, s) = \tilde{A}(t) \tilde{R}(t, s)$ ,  $\tilde{R}(s, s) = I_{n+1}$ , with

$$\tilde{A}(t) = \frac{\partial \tilde{f}}{\partial \tilde{x}}(t, \tilde{x}(t), u(t)), \quad \tilde{B}(t) = \frac{\partial \tilde{f}}{\partial u}(t, \tilde{x}(t), u(t)).$$

Since  $\tilde{\psi}^\top dF(u). \delta u = 0$  for every  $\delta u \in L^\infty([0, T], \mathbb{R}^m)$ , it follows that

$$\tilde{B}(t)^\top \tilde{R}(T, t)^\top \tilde{\psi} = 0 \quad (2.15)$$

for almost every  $t \in [0, T]$ . We set  $\tilde{p}(t) = \tilde{R}(T, t)^\top \tilde{\psi}$ . By derivating with respect to  $t$  the relation  $\tilde{R}(T, t) \tilde{R}(t, T) = I_{n+1}$ , it is easy to establish that  $\frac{d}{dt} \tilde{R}(T, t) = -\tilde{R}(T, t) \tilde{A}(t)$ . We infer that  $\tilde{p}(\cdot)$  is the unique solution of the Cauchy problem

$$\dot{\tilde{p}}(t) = -\tilde{A}(t)^\top \tilde{p}(t), \quad \tilde{p}(T) = \tilde{\psi}. \quad (2.16)$$

We are almost done. Let us now come back to the initial coordinates in  $\mathbb{R}^n \times \mathbb{R}$ . We set  $\tilde{p}(t) = \begin{pmatrix} p(t) \\ p^0(t) \end{pmatrix}$ . Since  $\tilde{f}$  does not depend on the (slack) variable  $x^0$ , we have  $\frac{\partial \tilde{f}}{\partial x^0} = 0$  and therefore, using (2.16),

$$\begin{pmatrix} \dot{p}(t) \\ \dot{p}^0(t) \end{pmatrix} = - \begin{pmatrix} \frac{\partial f}{\partial x}(t, x(t), u(t))^\top & \frac{\partial f^0}{\partial x}(t, x(t), u(t))^\top \\ 0 & 0 \end{pmatrix} \begin{pmatrix} p(t) \\ p^0(t) \end{pmatrix},$$

with  $p(T) = \psi$  and  $p^0(T) = \psi^0$ . In particular, we have  $\dot{p}^0(t) = 0$  and thus  $p^0 = \psi^0$  is a constant. Defining the Hamiltonian by (2.6), the latter equations give (2.5), and from (2.15) we infer (2.11). This is the “weak PMP”.

**Remark 2.22** In the above proof, we have constructed the adjoint vector so that  $(p(T), p^0) = (\psi, \psi^0)$  is a Lagrange multiplier. It is defined up to scaling (see Remark 2.16 and the subsequent comments).

### 2.2.3 Generalizations and Additional Comments

The PMP withstands many possible generalizations.

**More General Transversality Conditions** In Theorem 2.10, we have given transversality conditions for “decoupled” terminal conditions  $x(0) \in M_0$  and  $x(t_f) \in M_1$ . Assume that, instead, we have the coupled terminal conditions  $(x(0), x(t_f)) \in M$ , where  $M$  is a subset of  $\mathbb{R}^n \times \mathbb{R}^n$ . In this case, using a simple “copy-paste” of the dynamics, it is then easy to prove (see [1]) that the transversality conditions become (if they make sense)

$$(-p(0), p(t_f) - p^0 \nabla_x g(t_f, x(t_f))) \perp T_{(x(0), x(t_f))} M.$$

An important case is the one of periodic terminal conditions  $x(0) = x(T)$ : then  $M = \{(x, x) \mid x \in \mathbb{R}^n\}$ , and, if moreover  $g = 0$  then  $p(0) = p(t_f)$ .

Another useful generalization is when  $M$  is a general closed subset of  $\mathbb{R}^n \times \mathbb{R}^n$ , but is not necessarily a manifold, at least, locally around  $(x(0), x(t_f))$ . In this case, one can still write transversality conditions, by using notions of nonsmooth analysis (see [12, 31]), and we have

$$(p(0), -p(t_f) + p^0 \nabla_x g(t_f, x(t_f))) \in N_M(x(0), x(t_f))$$

where  $N_M(x, y)$  is the *limiting normal cone* to  $M$  at  $(x, y)$ . This generalized condition can be useful to provide sign conditions on the adjoint vector, whenever the subset  $M$  is not smooth.

**Infinite Time Horizon** The statement of the PMP remains the same when  $t_f = +\infty$ , under the assumption that the limit of the optimal trajectory  $x(t)$  exists when  $t \rightarrow +\infty$ ; in particular, the result then asserts that the limit of  $p(t)$  exists (see [24]).

**State Constraints, Hybrid Optimal Control Problems** Among the most well known and useful generalizations, one can think of the PMP for (OCP) with state constraints (see [10, 12, 27, 31]), for nonsmooth (OCP) (see [12, 31]), hybrid (OCP) (see [15, 16]), (OCP) settled on time scales (see [7]). There exist several possible proofs of the PMP (see [13]), based either on an implicit function argument (as we did here), or on a (Brouwer) fixed point argument (as in the classical book [27]), or on a Hahn-Banach separation argument (as in [8], or on Ekeland's principle (see [14]). Each of them may or may not be adapted to such or such generalization.

Let us note that, when dealing with state constraints, in full generality the adjoint vector becomes a measure. The generic situation that one has in mind is the case where this measure has only a finite number of atoms: in this favorable case the adjoint vector is then piecewise absolutely continuous, with possible jumps when touching the state constraint. Unfortunately the structure of the measure might be much more complicated, but such a discussion is outside of the scope of the present manuscript. We refer the reader to [6, 10, 12, 17, 19, 20, 26, 27, 31].

Although it is then not exact, it can be noted that state constraints may be tackled with usual penalization considerations, so as to deal rather with an (OCP) without state constraint. In some cases where getting the true optimal trajectory is not the main objective, this may be useful.

**PMP in Infinite Dimension** We refer to Sect. 5.2.6 in Part II and to [25, Chapter 4] for a generalization of the PMP in infinite dimension. To comment briefly on this extension, we notice that the argument to prove the weak PMP remains valid when replacing  $\mathbb{R}^n$  with a Banach space  $X$ , at the exception of one notable difficulty: in the argument by contraposition, we have to ensure that  $\text{Ran}(dF(u))$  is contained in a closed proper subspace of  $X \times \mathbb{R}$  (thus, with codimension  $\geq 1$ ). Here is the main difference with finite dimension. Indeed, the fact that  $\text{Ran}(dF(u))$  is a proper subspace of  $X \times \mathbb{R}$  is not enough to ensure a separation argument: it could happen that  $\text{Ran}(dF(u)) \subsetneq X \times \mathbb{R}$  be dense in  $X \times \mathbb{R}$ . Such a situation corresponds to *approximate* (but not *exact*) controllability, as we will see in Part II, and in this case the PMP fails to be true (see Example 5.33 in Sect. 5.2.6). Sufficient assumptions under which the PMP is still valid in infinite dimension are given in Sect. 5.2.6.

**Further Comments: Second-Order Conditions** Let us emphasize the fact that the PMP is a *first-order necessary condition* for optimality.<sup>3</sup> As already stressed, the PMP states that every optimal trajectory  $x(\cdot)$ , associated with a control  $u(\cdot)$ , is the projection onto  $\mathbb{R}^n$  of an extremal  $(x(\cdot), p(\cdot), p^0, u(\cdot))$ . However, conversely, an extremal (i.e., a solution of the equations of the PMP) is not necessarily optimal. The study of the optimality status of extremals can be done with the theory of

---

<sup>3</sup> This is an elaborated version of the first-order necessary condition  $\nabla f(x) = 0$  when minimizing a  $C^1$  function over  $\mathbb{R}^n$ !



*conjugate points*. More precisely, as in classical optimization where extremal points are characterized by a first-order necessary condition (vanishing of some appropriate derivative), there exists in optimal control a theory of second-order conditions for optimality, which consists of investigating a quadratic form that is the intrinsic second-order derivative of the end-point mapping: if this quadratic form is positive definite then this means that the extremal under consideration is locally optimal (for some appropriate topology), and if it is indefinite then the extremal is not optimal; conversely if the extremal is optimal then this quadratic form is nonnegative. Times at which the index of this quadratic form changes are called conjugate times. The optimality status of an extremal is then characterized by its first conjugate time. We refer to [4] (see references therein) for a survey on theory and algorithms to compute conjugate times. Much could be written on conjugate time theory (which has nice extensions in the bang-bang case), but this is beyond of the scope of the present book.

**Numerical Computation** The application of the PMP leads to a *shooting problem*, that is a boundary value problem consisting of computing extremals satisfying certain terminal conditions. It can be solved numerically by implementing a Newton method combined with an ODE integrator: this approach is called the *shooting method*.

We do not have room enough, here, to describe numerical methods in optimal control. We refer to [3] for a thorough description of so-called direct methods, and to [30] for a survey on indirect methods and the way to implement them in practice (see also [29] and references cited therein).

## 2.3 Particular Cases and Examples

In this section, we first specify the PMP for two important particular classes of examples: the minimal time problem for linear control systems, which yields the bang-bang principle; linear control systems with a quadratic cost, leading to the well-known “Linear Quadratic” (LQ) theory. Finally, we provide several examples of application of the PMP to nonlinear optimal control problems.

### 2.3.1 Minimal Time Problem for Linear Control Systems

Let us assume that the control system is linear, of the form

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$$

with the notations and regularity assumptions made in the introduction of Chap. 1. Let  $x_0 \in \mathbb{R}^n$  be an arbitrary initial point, and let  $\Omega$  be a compact subset of  $\mathbb{R}^m$ . For

any target point  $x_1 \in \mathbb{R}^n$ , we investigate the problem of steering the system from  $x_0$  to  $x_1$  in minimal time, under the control constraint  $u(t) \in \Omega$ .

It can be noticed that, if  $x_1$  is accessible from  $x_0$ , then there exists a minimal time trajectory steering the system from  $x_0$  to  $x_1$ , in a minimal time denoted by  $t_f$ . Indeed, by Theorem 1.5,  $\text{Acc}_\Omega(x_0, t)$  is a compact convex set depending continuously on  $t$ , thus  $t_f = \min\{t \geq 0 \mid x_1 \in \text{Acc}_\Omega(x_0, t)\}$ .

With the notations introduced at the beginning of Chap. 2, we have  $f(t, x, u) = A(t)x + B(t)u + r(t)$ ,  $f^0(t, x, u) = 1$  and  $g = 0$  (note that we could as well take  $f^0 = 0$  and  $g(t, x) = t$ ). The Hamiltonian of the optimal control problem is then  $H(t, x, p, p^0, u) = p^\top A(t)x + p^\top B(t)u + p^\top r(t) + p^0$ .

Let  $(x(\cdot), u(\cdot))$  be an optimal trajectory on  $[0, t_f]$ . According to the PMP, there exist  $p^0 \leq 0$  and an absolutely continuous mapping  $p(\cdot) : [0, t_f] \rightarrow \mathbb{R}^n$ , with  $(p(\cdot), p^0) \neq (0, 0)$ , such that  $\dot{p}(t) = -A(t)^\top p(t)$  for almost every  $t \in [0, t_f]$ , and the maximization condition yields

$$\langle B(t)^\top p(t), u(t) \rangle = \max_{v \in \Omega} \langle B(t)^\top p(t), v \rangle \quad (2.17)$$

for almost every  $t \in [0, t_f]$ .

Since the function  $v \mapsto \langle B(t)^\top p(t), v \rangle = p(t)^\top B(t)v$  is linear, we expect that the maximum over  $\Omega$  be reached at the boundary of  $\Omega$  (unless  $B(t)^\top p(t) = 0$ ). This is the contents of the *bang-bang principle*.

Let us consider particular but important cases.

**Case  $m = 1$  (Scalar Control)** Let us assume that  $m = 1$ , and that  $\Omega = [-a, a]$  with  $a > 0$ . This means that the control must satisfy the constraints  $|u(t)| \leq a$ . In that case,  $B(t)$  is a vector of  $\mathbb{R}^n$ , and  $\varphi(t) = p(t)^\top B(t)$  is called *switching function*. The maximization condition (2.17) implies that

$$u(t) = a \operatorname{sign}(\varphi(t))$$

as soon as  $\varphi(t) \neq 0$ . Here, we see that the structure of the optimal control  $u$  is governed by the switching function. We say that the control is *bang-bang* if the switching function  $\varphi$  does not vanish identically on any subset of positive measure of  $[0, t_f]$ . For instance, this is the case under the assumption:  $\varphi(t) = 0 \Rightarrow \dot{\varphi}(t) \neq 0$  (because then the zeros of  $\varphi$  are isolated). In that case, the zeros of the switching functions are called the switchings of the optimal control. According to the monotonicity of  $\varphi$ , we see that the optimal control switches between the two values  $\pm a$ . This is the typical situation of a bang-bang control.

In contrast, if the switching function  $\varphi$  vanishes, for instance, along a time subinterval  $I$  of  $[0, t_f]$ , then the maximization condition (2.17) does not provide any immediate information in order to compute the optimal control  $u$ . But we can then differentiate with respect to time (if this is allowed) the relation  $p(t)^\top B(t) = 0$ , and try to recover some useful information. This is a usual method in order to prove by contradiction, when it is possible, that optimal controls are bang-bang. An important

example where this argument is successful is the following result, that we let as an exercise (see [24]).

**Lemma 2.3** *Let us assume that  $A(t) \equiv A$ ,  $B(t) \equiv B$ ,  $r(t) \equiv 0$  (autonomous control system), and that the pair  $(A, B)$  satisfies the Kalman condition. Then any extremal control is bang-bang, and*

- *has at most  $n - 1$  switchings on  $[0, +\infty)$  if all eigenvalues of  $A$  are real;*
- *has an infinite number of switchings on  $[0, +\infty)$  if all eigenvalues of  $A$  have a nonzero imaginary part. In this case, for every  $N \in \mathbb{N}^*$ , there exists  $x_0 \in \mathbb{R}^n$  for which the corresponding minimal time control, steering  $x_0$  to 0, has more than  $N$  switchings.*

**Case  $m = 2$  (Two Scalar Controls  $u_1$  and  $u_2$ )** Let us assume that  $m = 2$ . In that case,  $B(t) = (B_1(t), B_2(t))$ , where  $B_1(t)$  and  $B_2(t)$  are vectors of  $\mathbb{R}^n$ . Let us show how to make explicit the extremal controls from the maximization condition of the PMP, for two important constraints very often considered in practice.

- Assume that  $\Omega = [-1, 1] \times [-1, 1]$ , the unit square of  $\mathbb{R}^2$ . This means that the controls  $u_1$  and  $u_2$  must satisfy the constraints  $|u_1(t)| \leq 1$  and  $|u_2(t)| \leq 1$ . As for the case  $m = 1$ , we set  $\varphi_i(t) = p(t)^\top B_i(t)$ , for  $i = 1, 2$ , and the maximization condition (2.17) implies that  $u_i(t) = \text{sign}(\varphi_i(t))$  as soon as  $\varphi_i(t) \neq 0$ , for  $i = 1, 2$ .
- Assume that  $\Omega = \bar{B}(0, 1)$ , the closed unit ball of  $\mathbb{R}^2$ . This means that the controls  $u_1$  and  $u_2$  must satisfy the constraints  $u_1(t)^2 + u_2(t)^2 \leq 1$ . Setting again  $\varphi_i(t) = p(t)^\top B_i(t)$ , for  $i = 1, 2$ , the maximization condition (2.17) can be written as

$$\varphi_1(t)u_1(t) + \varphi_2(t)u_2(t) = \max_{v_1^2 + v_2^2 \leq 1} \left\langle \begin{pmatrix} \varphi_1(t) \\ \varphi_2(t) \end{pmatrix}, \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \right\rangle$$

and it follows from the Cauchy-Schwarz inequality that

$$u_i(t) = \frac{\varphi_i(t)}{\sqrt{\varphi_1(t)^2 + \varphi_2(t)^2}}$$

for  $i = 1, 2$ , as soon as  $\varphi_1(t)^2 + \varphi_2(t)^2 \neq 0$ .

In these two cases, the comments done previously are still in force in the degenerate case where the switching functions vanish identically on some subset of positive measure. We do not insist on such difficulties at this step. Note that what is done here with  $m = 2$  could be written as well for any value of  $m$ .

### 2.3.2 Linear Quadratic Theory

In this chapter we make an introduction to the well known LQ (linear quadratic) theory, which has many applications in concrete applications, such as Kalman filtering or regulation problems. We first study and solve the basic LQ problem, and then we provide an important application to the tracking problem. For other applications (among which the Kalman filter), see [2, 21, 23, 28, 29].

#### 2.3.2.1 The Basic LQ Problem

We consider the optimal control problem

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \quad x(0) = x_0 \\ \min \int_0^T &\left( x(t)^\top W(t)x(t) + u(t)^\top U(t)u(t) \right) dt + x(T)^\top Qx(T) \end{aligned} \quad (2.18)$$

where  $x_0 \in \mathbb{R}^n$  and  $T > 0$  are fixed (arbitrarily),  $W(t)$  and  $Q$  are symmetric nonnegative matrices of size  $n$ ,  $U(t)$  is a symmetric positive definite matrix of size  $m$ . The dependence in time of the matrices above is assumed to be  $L^\infty$  on  $[0, T]$ . The controls are all possible functions of  $L^2([0, T], \mathbb{R}^m)$ .

We call this problem the basic LQ problem. Note that the final point is left free. The matrices  $W(t)$ ,  $U(t)$ , and  $Q$ , are called weight matrices.

We assume that there exists  $\alpha > 0$  such that

$$\int_0^T u(t)^\top U(t)u(t) dt \geq \alpha \int_0^T u(t)^\top u(t) dt \quad \forall u \in L^2([0, T], \mathbb{R}^m).$$

For instance, this assumption is satisfied if  $t \mapsto U(t)$  is continuous on  $[0, T]$ . In practice, the weight matrices are often constant.

**Theorem 2.11** *There exists a unique optimal solution to (2.18).*

This theorem can be proved using classical functional analysis arguments, as in the proof of Theorem 2.9. The uniqueness comes from the strict convexity of the cost. For a proof, see [29].

Let us apply the PMP to the basic LQ problem. The Hamiltonian is

$$H(t, x, p, p^0, u) = p^\top A(t)x + p^\top B(t)u + p^0(x^\top W(t)x + u^\top U(t)u)$$

and the adjoint equation is  $\dot{p}(t) = -A(t)^\top p(t) - 2p^0 W(t)x(t)$ . Since the final point is free, the transversality condition on the final adjoint vector yields  $p(T) = 2p^0 Qx(T)$ , and hence necessarily  $p^0 \neq 0$  (otherwise we would have  $(p(T), p^0) = (0, 0)$ , which is a contradiction with the PMP). According to Remark 2.16, we choose, here, to normalize the adjoint vector so that  $p^0 = -\frac{1}{2}$  (this will be

convenient when derivating the squares...). Now, since there is no constraint on the control, we have

$$0 = \frac{\partial H}{\partial u}(t, x(t), p(t), p^0, u(t)) = B(t)^\top p(t) - U(t)u(t)$$

and hence  $u(t) = U(t)^{-1}B(t)^\top p(t)$ .

Summing up, we have obtained that, for the optimal solution  $(x(\cdot), u(\cdot))$  of the basic LQ problem, we have

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)U(t)^{-1}B(t)^\top p(t), & x(0) &= x_0, \\ \dot{p}(t) &= -A(t)^\top p(t) + W(t)x(t), & p(T) &= -Qx(T). \end{aligned}$$

At this step, things are already nice, and we could implement a shooting method in order to solve the above problem. But the structure of the problem is so particular that we are actually able, here, to express  $p(t)$  linearly in function of  $x(t)$ . This property is very remarkable but also very specific to that problem. We claim that we can search  $p(t)$  in the form  $p(t) = E(t)x(t)$ . Replacing in the above equations, we easily obtain a relation of the form  $R(t)x(t) = 0$ , with

$$R(t) = \dot{E}(t) - W(t) + A(t)^\top E(t) + E(t)A(t) + E(t)B(t)U(t)^{-1}B(t)^\top E(t)$$

and  $E(T)x(T) = -Qx(T)$ . Therefore, “simplifying by  $x$ ”, we see that, if we assume that  $R(t) = 0$  by definition, then we can go back in the reasoning and infer, by Cauchy uniqueness, that  $p(t) = E(t)x(t)$ . We have obtained the following result.

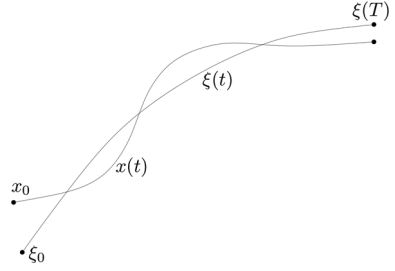
**Theorem 2.12** *The optimal solution  $x(\cdot)$  of the basic LQ problem is associated with the control*

$$u(t) = U(t)^{-1}B(t)^\top E(t)x(t)$$

where  $E(t) \in \mathcal{M}_n(\mathbb{R})$  is the unique solution on  $[0, T]$  of the Riccati matrix differential equation

$$\begin{aligned} \dot{E}(t) &= W(t) - A(t)^\top E(t) - E(t)A(t) - E(t)B(t)U(t)^{-1}B(t)^\top E(t) \\ E(T) &= -Q \end{aligned} \tag{2.19}$$

Actually, there is a difficulty for finishing the proof of that theorem, by proving that the unique solution  $E(\cdot)$  of the Cauchy problem (2.19), which is a priori defined in a neighborhood of  $T$ , is indeed well defined over the whole interval  $[0, T]$ . Indeed, such a Riccati differential equation may produce blow-up, and the well-posedness over the whole  $[0, T]$  is not obvious. We do not prove that fact here, and

**Fig. 2.2** Tracking problem

we refer the reader, e.g., to [29] for a proof (which uses the optimality property of  $u(\cdot)$ ).

It can be noted that  $E(t)$  is symmetric (this is easy to see by Cauchy uniqueness). The result above is interesting because, in that way, the problem is completely solved, without having to compute any adjoint vector for instance by means of a shooting method. Moreover the optimal control is expressed in a feedback form,  $u(t) = K(t)x(t)$ , well adapted to robustness issues.

This is because of that property that the LQ procedures are so much used in practical problems and industrial issues. We are next going to give an application to tracking.

### 2.3.2.2 Tracking Problem

Let us consider the general control system  $\dot{x}(t) = f(t, x(t), u(t))$ , with initial condition  $x(0) = x_0$ , with the regularity assumptions done at the beginning of Chap. 1. Let  $t \mapsto \xi(t)$  be a trajectory in  $\mathbb{R}^n$ , defined on  $[0, T]$ , and which is arbitrary (in particular, this is not necessarily a solution of the control system). We assume however that  $\xi(\cdot)$  is Lipschitz (or, at least, absolutely continuous). The objective is to design a control  $u$  generating a trajectory  $x(\cdot)$  that tracks the trajectory  $\xi(\cdot)$  in the “best possible way” (see Fig. 2.2).

We proceed as follows. We set  $e(t) = x(t) - \xi(t)$ , and we will try to design  $u$  so that  $e(\cdot)$  remains as small as possible. Using a first-order expansion, we have

$$\dot{e}(t) = f(t, \xi(t) + e(t), u(t)) - \dot{\xi}(t) = A(t)e(t) + B(t)u(t) + r(t)$$

with

$$A(t) = \frac{\partial f}{\partial x}(t, \xi(t), 0), \quad B(t) = \frac{\partial f}{\partial u}(t, \xi(t), 0),$$

$$r(t) = f(t, \xi(t), 0) - \dot{\xi}(t) + o(e(t), u(t)).$$

It seems reasonable to seek a control  $u$  minimizing the cost

$$C(u) = \int_0^T \left( z(t)^\top W(t) z(t) + u(t)^\top U(t) u(t) \right) dt + z(T)^\top Q z(T)$$

for the control system

$$\dot{z}(t) = A(t)z(t) + B(t)u(t) + r_1(t), \quad z(0) = x_0 - \xi(0),$$

with  $r_1(t) = f(t, \xi(t), 0) - \dot{\xi}(t)$ , where  $W(t)$ ,  $U(t)$  and  $Q$  are weight matrices that are chosen by the user. We hope that the resulting control will be such that the term  $o(z(t), u(t))$  is small. In any case, the choice above produces a control, which hopefully tracks the trajectory  $\xi(t)$  as closely as possible. Note that, when linearizing the system, we have linearized at  $u = 0$ , considering that  $u$  will be small. We could have linearized along a given  $\bar{u}(t)$ : we then obtain one of the many possible variants of the method.

Let us now solve the above optimal control problem. In order to absorb the perturbation term  $r_1$ , we consider an augmented system, by adding one dimension. We set

$$z_1 = \begin{pmatrix} z \\ 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} A & r_1 \\ 0 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix}, \quad W_1 = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix},$$

and hence we want to minimize the cost

$$C(u) = \int_0^T \left( z_1(t)^\top W_1(t) z_1(t) + u(t)^\top U(t) u(t) \right) dt + z_1(T)^\top Q_1 z_1(T)$$

for the control system  $\dot{z}_1(t) = A_1(t)z_1(t) + B_1(t)u(t)$ , with  $z_1(0)$  fixed. In this form, this is a basic LQ problem, as studied in the previous section. According to Theorem 2.12, there exists a unique optimal control, given by  $u(t) = U(t)^{-1} B_1(t)^\top E_1(t) z_1(t)$ , where  $E_1(t)$  is the solution of the Cauchy problem  $\dot{E}_1 = W_1 - A_1^\top E_1 - E_1 A_1 - E_1 B_1 U^{-1} B_1^\top E_1$ ,  $E_1(T) = -Q_1$ . Setting

$$E_1(t) = \begin{pmatrix} E(t) & h(t) \\ h(t)^\top & \alpha(t) \end{pmatrix}$$

with  $E(t)$  square matrix of size  $n$ ,  $h(t) \in \mathbb{R}^n$  and  $\alpha(t) \in \mathbb{R}$ , we obtain the following result.

**Proposition 2.6** *The optimal (in the sense above) tracking control is*

$$u(t) = U(t)^{-1} B(t)^\top E(t) (x(t) - \xi(t)) + U(t)^{-1} B(t)^\top h(t)$$

where

$$\begin{aligned}\dot{E} &= W - A^\top E - EA - EBU^{-1}B^\top E, & E(T) &= -Q, \\ \dot{h} &= -A^\top h - E(f(t, \xi, 0) - \dot{\xi}) - EBU^{-1}B^\top h, & h(T) &= 0.\end{aligned}$$

It is interesting to note that the control is written in a feedback form  $u(t) = K(t)(x(t) - \xi(t)) + H(t)$ .

As said at the beginning of the section, there are many other applications of the LQ theory, and many possible variants. For instance, one can easily adapt the above tracking procedure to the problem of output tracking: in that case we track an observable. It is also very interesting to let the horizon of time  $T$  go to  $+\infty$ . In that case, we can expect to obtain stabilization results. This is indeed the case for instance when one considers a linear autonomous control system (regulation over an infinite horizon); the procedure is referred to as LQR in practice and is very much used for stabilization issues.

In practice, we often make the choice of constant diagonal weight matrices  $W(t) = w_0 I_n$ ,  $U(t) = u_0 I_m$ , and  $Q = q_0 I_n$ , with  $w_0 \geq 0$ ,  $u_0 > 0$  and  $q_0 \geq 0$ . If  $w_0$  is chosen much larger than  $u_0$ , then it is expected that  $\|x(t) - \xi(t)\|$  will remain small (while paying the price of larger values of  $u(t)$ ). Conversely if  $u_0$  is chosen much larger than  $w_0$  then it is expected that  $u(t)$  will take small values, while the tracking error  $\|x(t) - \xi(t)\|$  may take large values. Similarly, if  $q_0$  is taken very large then it is expected (at least, under appropriate controllability assumptions) that  $x(T)$  will be close to  $\xi(T)$ . A lot of such statements, with numerous possible variants, may be established. We refer to [2, 21–24, 28, 29] for (many) more precise results.

### 2.3.3 Examples of Nonlinear Optimal Control Problems

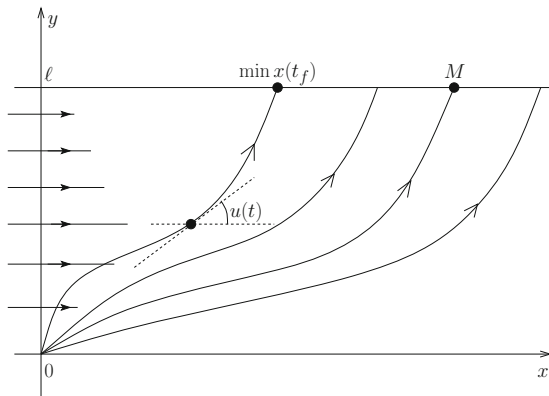
**Example 2.10 (Zermelo Problem)** Let us consider a boat moving with constant speed along a river of constant width  $\ell$ , in which there is a current  $c(y)$  (assumed to be a function of class  $C^1$ ). The movement of the center of mass of the boat is governed by the control system

$$\begin{aligned}\dot{x}(t) &= v \cos u(t) + c(y(t)), & x(0) &= 0, \\ \dot{y}(t) &= v \sin u(t), & y(0) &= 0,\end{aligned}$$

where  $v > 0$  is the constant speed, and the control is the angle  $u(t)$  of the axis of the boat with respect to the axis  $(0x)$  (see Fig. 2.3). We investigate three variants of optimal control problems with the objective of reaching the opposite side: the final condition is  $y(t_f) = \ell$ , where the final time  $t_f$  is free.

1. Assuming that  $c(y) \geq v$  for every  $y \in [0, \ell]$  (strong current), compute the optimal control minimizing the drift  $x(t_f)$ .



**Fig. 2.3** Zermelo problem

2. Compute the minimal time control.
3. Compute the minimal time control for the problem of reaching a precise point  $M = (x_1, \ell)$  of the opposite side.

1. *Reaching the opposite side by minimizing the drift  $x(t_f)$ .*

We choose  $f^0 = 0$  and  $g(t, x, y) = x$ . The Hamiltonian is

$$H = p_x(v \cos u + c(y)) + p_y v \sin u.$$

The adjoint equations are  $\dot{p}_x = 0$ ,  $\dot{p}_y = -p_x c'(y)$ . In particular  $p_x$  is constant. Since the target is  $M_1 = \{y = \ell\}$ , the transversality condition on the adjoint vector yields  $p_x = p^0$ . The maximization condition of the Hamiltonian leads to

$$\cos u(t) = \frac{p_x}{\sqrt{p_x^2 + p_y(t)^2}}, \quad \sin u(t) = \frac{p_y(t)}{\sqrt{p_x^2 + p_y(t)^2}},$$

for almost every  $t$ , provided that the function  $\varphi(t) = p_x^2 + p_y(t)^2$  does not vanish on any subset of positive measure. This condition is proved by contradiction: if  $\varphi(t) \equiv 0$  on  $I$ , then  $p_x = 0$  and  $p_y(t) = 0$  on  $I$ , but then also  $p^0 = p_x = 0$ , and we get a contradiction (because the adjoint  $(p_x, p_y, p^0)$  must be non trivial). Finally, since  $t_f$  is free and the problem is autonomous, we get that  $H = 0$  along any extremal, that is,  $H = v\sqrt{p_x^2 + p_y^2} + p_x c(y) = 0$  along any extremal.

We must have  $p^0 \neq 0$ . Indeed, otherwise,  $p^0 = 0$  implies that  $p_x = 0$ , and from  $H = 0$  we infer that  $p_y = 0$  as well. This is a contradiction. Hence, we can take  $p^0 = -1$ , and therefore  $p_x = -1$ .

From  $H = 0$ , we have  $\sqrt{1 + p_y^2} = \frac{c(y)}{v}$ , and hence  $\cos u = -\frac{v}{c(y)}$ . Since  $c(y) \gg v$ , this equation is solvable, and we get

$$u(t) = \text{Arccos}\left(-\frac{v}{c(y(t))}\right).$$

Note that we have thus determined the optimal control in a *feedback form*, which is the best possible one (in practice, such a control can be made fully automatic, provided one can measure the position  $y$  at any time).

**Remark 2.23** The assumption  $c(y) \gg v$  means that the current is strong enough. Without this assumption, the optimal control problem consisting of minimizing the drift  $x(t_f)$  would be ill-posed: there would not exist any optimal solution (at least, in finite time), because if, for some  $y$ , we have  $c(y) < v$ , then, along this  $y$ , the boat can go against the current towards  $x = -\infty$ .

We also realize that, if we had not made this assumption, then the above equation would not be solvable. This remark provides a way for showing that the optimal control problem has no solution, by contradiction (recall that the PMP says that **if** a trajectory is optimal **then** it must satisfy the various conditions stated in the PMP).

**Remark 2.24** The optimal trajectory, minimizing the lateral deport, is represented on Fig. 2.3. It is interesting to note that any other trajectory is necessarily at the right of that optimal trajectory. In particular, this gives the reachable set (in any time): the reachable set consists of all points such that  $0 \leq y \leq \ell$  that are at the right of the optimal trajectory.

## 2. Reaching the opposite side in minimal time.

We choose  $f^0 = 1$  and  $g = 0$ . Now, the Hamiltonian is

$$H = p_x(v \cos u + c(y)) + p_y v \sin u + p^0,$$

the adjoint equations are the same as previously, as well as the extremal controls provided that  $\varphi(t) \neq 0$ . The transversality condition on the adjoint vector is different: we now obtain  $p_x = 0$ . It follows that  $p_y$  is constant. Besides, we still have  $H = 0$  along any extremal. We claim that  $p_y \neq 0$ . Indeed, otherwise,  $H = 0$  implies that  $p^0 = 0$ , and then  $(p_x, p_y, p^0) = (0, 0, 0)$ , which is a contradiction. Hence, we get that  $\cos u(t) = 0$  and  $\sin u(t) = \text{sign}(p_y)$ , and thus  $u(t) = \frac{\pi}{2}$  for every time  $t$  (the sign of  $u$  is given by the fact that, at least at the beginning, the boat must leave the initial riverbank with  $\dot{y} > 0$ ).

**Remark 2.25** Actually, the fact that the minimal time control is  $u = \frac{\pi}{2}$  is obvious by inspecting the equation in  $y$ . Indeed, since  $t_f = \int_0^{t_f} dt = \int_0^\ell \frac{dy}{\dot{y}}$ , it easily follows that, in minimal time, we must have  $\dot{y} = 1$ .

Note that we do not need to assume, here, that the current is strong enough. The calculations above are valid, whatever the function  $c(y)$  may be.

### 3. Reaching a precise point of the opposite side in minimal time.

This is a variant of the second problem, in which we lose the transversality condition on the adjoint vector, because the target point is fixed. The constant  $p_x$  is thus undetermined at this step. We still have that  $H = 0$  along any extremal. By contradiction, it is still true that the function  $\varphi$  cannot vanish on any subset of positive measure (indeed otherwise  $p_x = 0$  and  $p_y = 0$ , and from  $H = 0$  we get that  $p^0 = 0$ : contradiction).

This third variant is interesting because both the normal case  $p^0 \neq 0$  and the abnormal case  $p^0 = 0$  may occur. Let us analyze them.

- Normal case:  $p^0 = -1$ . In that case, using that  $H = v\sqrt{p_x^2 + p_y^2} + p_x c(y) - 1 = 0$  along any extremal, we get

$$\cos u(t) = \frac{p_x v}{1 - p_x c(y(t))}.$$

Note that, for this equation to be solvable,  $p_x$  must be such that  $|p_x v| \leq |1 - p_x c(y(t))|$  for every time  $t$ . We have thus obtained all possible optimal controls, parametrized by  $p_x$ . The parameter  $p_x$  is the “shooting parameter”, that is, the degree of freedom that is required in order to tune the final condition  $x(t_f) = x_1$ , i.e., in order to reach the target point  $M$ .

All optimal trajectories are represented on Fig. 2.3 in the case of a strong current. To go further, we could specify a current function  $c(y)$ , and either implement a shooting method, or try to make explicit computations if this is possible.

- Abnormal case:  $p^0 = 0$ . Using  $H = 0$ , we get  $\cos u = -\frac{v}{c(y)}$ . In the case where the current is strong enough, we see that we recover exactly the solution of the first variant, that is, the optimal trajectory with minimal drift.

Then, two cases may occur: either the target point  $M$  is different from the end-point of the minimal drift trajectory, and then, the trajectory is not solution of our problem and the abnormal does not occur; or, by chance, the target point  $M$  exactly coincides with the end-point of the minimal drift trajectory, and then (under the assumption  $c(y) \geq v$ ) the abnormal case indeed occurs, and the optimal trajectory coincides with the minimal drift trajectory.

This example is interesting because it gives a very simple situation where we may have an abnormal minimizer and how to interpret it.

**Example 2.11 (Optimal Control of Damaging Insects by Predators)** In order to eradicate as much as possible a population  $x_0 > 0$  of damaging pests, we introduce in the ecosystem a population  $y_0 > 0$  of (nondamaging) predator insects killing the pests.

*First Part* In a first part, we assume that the predator insects that we introduce are infertile, and thus cannot reproduce themselves. The control consists of the

continuous introduction of predator insects. The model is

$$\begin{aligned}\dot{x}(t) &= x(t)(a - by(t)), & x(0) &= x_0, \\ \dot{y}(t) &= -cy(t) + u(t), & y(0) &= y_0,\end{aligned}$$

where  $a > 0$  is the reproduction rate of pests,  $b > 0$  is the predation rate,  $c > 0$  is the natural death rate of predators. The control  $u(t)$  is the rate of introduction of new predators at time  $t$ . It must satisfy the constraint  $0 \leq u(t) \leq M$  where  $M > 0$  is fixed. Let  $T > 0$  be fixed. We want to minimize, at the final time  $T$ , the number of pests, while minimizing as well the number of introduced predators. We choose to minimize the cost

$$x(T) + \int_0^T u(t) dt.$$

Throughout, we denote by  $p = (p_x, p_y)$  and by  $p^0$  the adjoint variables.

First, we claim that  $x(t) > 0$  and  $y(t) > 0$  along  $[0, T]$ , for every control  $u$ .

Indeed, since  $u(t) \geq 0$ , we have  $\dot{y}(t) \geq -cy(t)$ , hence  $y(t) \geq y_0 e^{-ct} > 0$ . For  $x(t)$ , we argue by contradiction: if there exists  $t_1 \in [0, T]$  such that  $x(t_1) = 0$ , then  $x(t) = 0$  for every time  $t$  by Cauchy uniqueness; this raises a contradiction with the fact that  $x(0) = x_0 > 0$ .

The Hamiltonian of the optimal control problem is  $H = p_x x(a - by) + p_y(-cy + u) + p^0 u$  and the adjoint equations are  $\dot{p}_x = -p_x(a - by)$ ,  $\dot{p}_y = bp_x x + cp_y$ . The transversality conditions yield  $p_x(T) = p^0$  and  $p_y(T) = 0$ . It follows that  $p^0 \neq 0$  (otherwise we would have  $(p(T), p^0) = (0, 0)$ , which is a contradiction). In what follows, we set  $p^0 = -1$ .

We have  $\frac{d}{dt} x(t)p_x(t) = x(t)p_x(t)(a - by(t)) - x(t)p_x(t)(a - by(t)) = 0$ , hence  $x(t)p_x(t) = \text{Cst} = -x(T)$  because  $p_x(T) = p^0 = -1$ . It follows that  $\dot{p}_y = -bx(T) + cp_y$ , and since  $p_y(T) = 0$ , we infer, by integration, that  $p_y(t) = \frac{b}{c}x(T)(1 - e^{c(t-T)})$ . The maximization condition of the PMP yields

$$u(t) = \begin{cases} 0 & \text{if } p_y(t) - 1 < 0 \\ M & \text{if } p_y(t) - 1 > 0 \end{cases}$$

unless the function  $t \mapsto p_y(t) - 1$  vanishes identically on some subinterval. But this is not possible because we have seen above that the function  $p_y$  is decreasing. We conclude that the optimal control is bang-bang. Moreover, at the final time we have  $p_y(T) - 1 = -1$ , hence, by continuity, there exists  $\varepsilon > 0$  such that  $p_y(t) - 1 < 0$  along  $[T - \varepsilon, T]$ , and hence  $u(t) = 0$  along a subinterval containing the final time.

We can be more precise: we claim that, actually, the optimal control has at most one switching along  $[0, T]$  (and is 0 at the end).

Indeed, the function  $p_y$  is decreasing (because  $x(T) > 0$ , as we have seen at the beginning), hence the function  $t \mapsto p_y(t) - 1$ , which is equal to  $-1$  at  $t = T$ , vanishes at most one time. If there is such a switching, necessarily it occurs at some

time  $t_1 \in [0, T]$  such that  $p_y(t_1) = 1$ , which yields

$$t_1 = T + \frac{1}{c} \ln \left( 1 - \frac{c}{bx(T)} \right).$$

Note that this switching can occur only if  $t_1 > 0$  (besides, we have  $t_1 < T$ ), hence, only if  $x(T) > \frac{c}{b} \frac{1}{1-e^{-cT}}$ . By integrating backwards the equations, we could even express an implicit condition on the initial conditions, ensuring that this inequality be true, hence, ensuring that there is a switching.

*Second Part* We now assume that the predators that we introduce are fertile, and reproduce themselves with a rate that is proportional to the number of pests. The control is now the death rate of predators. In order to simplify, we assume that the variables are normalized so that all other rates are equal to 1. The model is

$$\begin{aligned} \dot{x}(t) &= x(t)(1 - y(t)), & x(0) &= x_0, \\ \dot{y}(t) &= -y(t)(u(t) - x(t)), & y(0) &= y_0, \end{aligned}$$

where the control  $u(t)$  satisfies the constraint  $0 < \alpha \leq u(t) \leq \beta$ .

First, as before we have  $x(t) > 0$  and  $y(t) > 0$  along  $[0, T]$ , for every control  $u$ .

All equilibrium points of the system are given by  $x_e = u_e$ ,  $y_e = 1$ , for every  $\alpha \leq u_e \leq \beta$ . In the quadrant, we have a whole segment of equilibrium points.

Let us investigate the problem of steering the system in minimal time  $t_f$  to the equilibrium point  $x(t_f) = a$ ,  $y(t_f) = 1$ .

The Hamiltonian is  $H = p_x x(1 - y) - p_y y(u - x) + p^0$  and the adjoint equations are  $\dot{p}_x = -p_x(1 - y) - p_y y$ ,  $\dot{p}_y = p_x x + p_y(u - x)$ . The transversality condition on the final time gives  $H(t_f) = 0$ , and since the system is autonomous, it follows that the Hamiltonian is constant along any extremal, equal to 0.

The maximization condition of the PMP is  $\max_{0 \leq u \leq M} (-p_y y u)$ , which gives, since  $y(t) > 0$ ,

$$u(t) = \begin{cases} \alpha & \text{if } p_y(t) > 0 \\ \beta & \text{if } p_y(t) < 0 \end{cases}$$

unless the function  $t \mapsto p_y(t)$  vanishes identically along a subinterval. If this is the case then  $p_y(t) = 0$  for every  $t \in I$ . Derivating with respect to time, we get that  $x p_x = 0$  and thus  $p_x = 0$  along  $I$ . Therefore, along  $I$ , we have  $H = p^0$ , and since  $H = 0$  we infer that  $p^0 = 0$ , which raises a contradiction. Therefore, we conclude that the optimal control is bang-bang.

Along an arc where  $u = \alpha$  (resp.,  $u = \beta$ ), we compute  $\frac{d}{dt} F_\alpha(x(t), y(t)) = 0$ , where

$$F_\alpha(x, y) = x + y - \alpha \ln x - \ln y$$

(resp.,  $F_\beta$ ), that is,  $F_\alpha(x(t), y(t))$  is constant along such a bang arc.

It can be noted that, formally, this integral of the movement can be obtained by computing  $\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{-y}{1-y} \frac{\alpha-x}{x}$  and by integrating this separated variables one-form.

Considering a second-order expansion of  $F_\alpha$  at the point  $(\alpha, 1)$ ,

$$F_\alpha(\alpha + h, 1 + k) = \alpha - \alpha \ln \alpha + 1 + \frac{1}{2} \left( \frac{h^2}{\alpha} + k^2 \right) + o(h^2 + k^2),$$

we see that  $F_\alpha$  has a strict local minimum at the point  $(\alpha, 1)$ . Moreover, the function  $F_\alpha$  is (strictly) convex, because its Hessian  $\begin{pmatrix} \frac{a}{x^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}$  is symmetric positive definite at any point such that  $x > 0, y > 0$ . It follows that the minimum is global.

For any controlled trajectory (with a control  $u$ ), we have

$$\frac{d}{dt} F_\alpha(x(t), y(t)) = (u(t) - \alpha)(1 - y(t)).$$

Let us prove that there exists  $\varepsilon > 0$  such that  $u(t) = \alpha$ , for almost every  $t \in [t_f - \varepsilon, t_f]$  (in other words, the control  $u$  is equal to  $\alpha$  at the end).

Indeed, at the final time, we have either  $p_y(t_f) = 0$  or  $p_y(t_f) \neq 0$ .

If  $p_y(t_f) = 0$ , then, using the differential equation in  $p_y$ , we have  $\dot{p}_y(t_f) = p_x(t_f)a$ . We must have  $p_x(t_f) \neq 0$  (otherwise we would get a contradiction, noticing as previously that  $H(t_f) = p^0 = 0$ ). Hence  $\dot{p}_y(t_f) \neq 0$ , and therefore the function  $p_y$  has a constant sign along some interval  $[t_f - \varepsilon, t_f[$ . Hence, along this interval, the control is constant, either equal to  $\alpha$  or to  $\beta$ . It cannot be equal to  $\alpha$ , otherwise, since the function  $F_\alpha$  is constant along this arc, and since this arc must reach the point  $(\alpha, 1)$ , this constant would be equal to the minimum of  $F_\alpha$ , which would imply that the arc is constant, equal to the point  $(\alpha, 1)$ : this is a contradiction because we consider a minimal time trajectory reaching the point  $(\alpha, 1)$ .

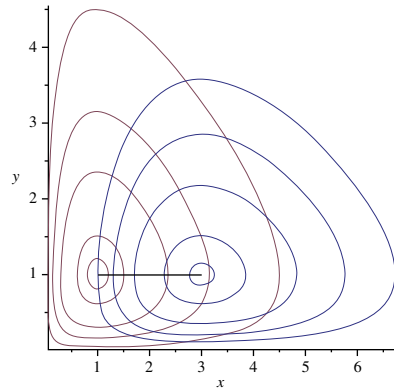
If  $p_y(t_f) \neq 0$ , then the function  $p_y$  has a constant sign along some interval  $[t_f - \varepsilon, t_f[$ , and hence, along this interval, the control is constant, equal either to  $\alpha$  or to  $\beta$ . With a similar reasoning as above, we get  $u = \alpha$ .

Let us now provide a control strategy (which can actually be proved to be optimal) in order to steer the system from any initial point  $(x_0, y_0)$  to  $(\alpha, 1)$ .

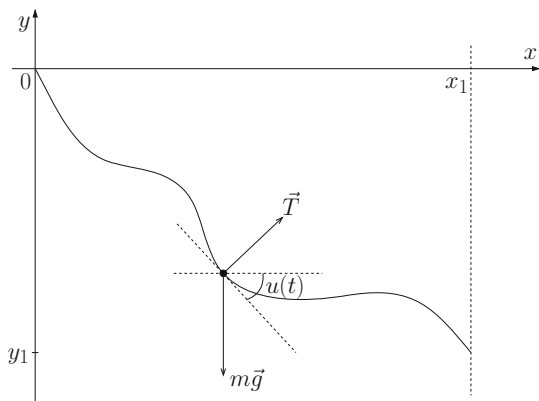
First, in a neighborhood of the point  $(\alpha, 1)$ , the level sets of the function  $F_\alpha$  look like circles. Farther from that point, the level sets look more and more like rectangle triangles, asymptotic to the coordinate axes. Similarly for the level sets of  $F_\beta$ , with respect to the point  $(\beta, 1)$  (see Fig. 2.4).

Let us start from a point  $(x_0, y_0)$ , which is located on the segment joining the two points  $(\alpha, 1)$  and  $(\beta, 1)$ , that is, such that  $\alpha < x_0 < \beta$  and  $y_0 = 1$ . We start with the control  $u = \alpha$ , and we remain along a level set of the function  $F_\alpha$  (thus, “centered” at the point  $(\alpha, 1)$ ). At some time, we switch on the control  $u = \beta$ , and then we remain along a level set of  $F_\beta$  (thus, “centered” at the point  $(\alpha, 1)$ ) which passes through the target point  $(\alpha, 1)$ .

**Fig. 2.4** Example with  $\alpha = 1$  and  $\beta = 3$



**Fig. 2.5** The Brachistochrone Problem, seen as an optimal control problem



Now, if we start from any other point  $(x_0, y_0)$ , we can determine on Fig. 2.4 a sequence of arcs along the level sets, respectively, of  $F_\alpha$  and of  $F_\beta$ , that steers the system from the starting point to the target point.

**Example 2.12 (The Brachistochrone Problem)** The objective is to determine what is the optimal shape of a children's slide (between two given altitudes) so that, if a ball slides along it (with zero initial speed), it arrives at the other extremity in minimal time.

This problem can be modeled as the following optimal control problem. In the Euclidean plane, the slide is modeled as a continuous curve, starting from the origin, and arriving at a given fixed point  $(x_1, y_1)$ , with  $x_1 > 0$ . We consider a ball of mass  $m > 0$  sliding along the curve. We denote by  $(x(t), y(t))$  its position at time  $t$ . The ball is subject to the gravity force  $m\vec{g}$  and to the reaction force of the children's slide. At time  $t$ , we denote by  $u(t)$  the (oriented) angle between the unit horizontal vector and the velocity vector  $(\dot{x}(t), \dot{y}(t))$  of the ball (which is collinear to the tangent to the curve). See Fig. 2.5.

Seeking the curve is equivalent to seeking the angle  $u(t)$ . Therefore, we stipulate that  $u$  is a control. By projecting the equations given by the fundamental principle

of dynamics onto the tangent to the curve, we obtain the following control system:

$$\begin{aligned}\dot{x}(t) &= v(t) \cos u(t), & x(0) &= 0, & x(t_f) &= x_1, \\ \dot{y}(t) &= -v(t) \sin u(t), & y(0) &= 0, & y(t_f) &= y_1, \\ \dot{v}(t) &= g \sin u(t), & v(0) &= 0, & v(t_f) &\text{ free.}\end{aligned}\tag{2.20}$$

The control is  $u(t) \in \mathbb{R}$ ,  $g > 0$  is a constant. We want to minimize the final time  $t_f$ .

First of all, noticing that  $\dot{y} = -\frac{1}{g}v\dot{v}$ , we get, by integration, that  $y(t) = -\frac{1}{2g}v(t)^2$ , for every control  $u$ . This implies that any final point such that  $y_1 > 0$  is not reachable. Therefore, from now on, we will assume that  $y_1 \leq 0$ .

Because of the relation between  $y(t)$  and  $v(t)$ , we can reduce the optimal control problem (2.20) to the minimal time control problem for the following system:

$$\begin{aligned}\dot{x}(t) &= v(t) \cos u(t), & x(0) &= 0, & x(t_f) &= x_1 > 0 \text{ fixed}, \\ \dot{v}(t) &= g \sin u(t), & v(0) &= 0, & v(t_f) &\text{ fixed.}\end{aligned}\tag{2.21}$$

Note that, since  $y(t_f) = y_1$  is fixed, it follows that  $v(t_f) = \pm\sqrt{-2gy_1}$ .

Let us apply the Pontryagin maximum principle to the optimal control problem (2.21). The Hamiltonian is  $H = p_x v \cos u + p_v g \sin u + p^0$ . The adjoint equations are  $\dot{p}_x = 0$  et  $\dot{p}_v = -p_x \cos u$ . In particular,  $p_x$  is constant. Since the final time is free and the problem is autonomous, we have  $H = 0$  along any extremal. The maximization condition of the Hamiltonian yields

$$\cos u(t) = \frac{p_x v(t)}{\sqrt{(p_x v(t))^2 + (gp_v(t))^2}}, \quad \sin u(t) = \frac{gp_v(t)}{\sqrt{(p_x v(t))^2 + (gp_v(t))^2}},$$

provided that  $\varphi(t) = (p_x v(t))^2 + (gp_v(t))^2 \neq 0$ .

We have  $\frac{d}{dt}(p_x v(t)) = p_x g \sin u(t)$  et  $\frac{d}{dt}(gp_v(t)) = -p_x g \cos u(t)$ . As a consequence, if  $\varphi(t) = 0$  for every  $t$  in some subset  $I$  of positive measure, then  $p_x = 0$ . Therefore  $\varphi(t) = (gp_v(t))^2$  and thus  $p_v(t) = 0$  for every  $t \in I$ . Since  $H = 0$ , we infer that  $p^0 = 0$ . We have obtained  $(p_x, p_v(t), p^0) = (0, 0, 0)$ , which is a contradiction. We conclude that the function  $\varphi$  never vanishes on any subset of  $[0, t_f]$  of positive measure. Therefore the above expression of the controls is valid almost everywhere.

The maximized Hamiltonian is  $H = \sqrt{(p_x v(t))^2 + (gp_v(t))^2} + p^0$ . Since  $H = 0$  along any extremal, we infer, by contradiction, that  $p^0 \neq 0$  (indeed otherwise we would infer that  $\varphi \equiv 0$ , which leads to a contradiction). Hence, from now on, we take  $p^0 = -1$ .

Since  $H = 0$  along any extremal, we get that  $(p_x v(t))^2 + (gp_v(t))^2 = 1$ , and therefore  $\cos u(t) = \frac{p_x v(t)}{1}$  and  $\sin u(t) = \frac{gp_v(t)}{1}$ .

If  $p_x$  were equal to 0, then we would have  $\cos u(t) = 0$  and thus  $\dot{x}(t) = 0$ , and then we would never reach the point  $x_1 > 0$ . Therefore  $p_x \neq 0$ .



Let us now integrate the trajectories. We have  $\dot{v} = g^2 p_v$  and  $\dot{p}_v = -p_x^2 v$  and thus  $\ddot{v} + g^2 p_x^2 v = 0$ , and since  $v(0) = 0$  we get that  $v(t) = A \sin(gp_x t)$ . Since  $H = 0$  and  $v(0) = 0$ , we have  $g^2 p_v(0)^2 = 1$  hence  $p_v(0) = \pm \frac{1}{g}$ , and thus  $\dot{v}(0) = \pm g$ . We infer that  $A = \pm \frac{1}{p_x}$ , and hence that  $v(t) = \pm \frac{1}{p_x} \sin(gp_x t)$ . Now,  $\dot{x} = v \cos u = p_x v^2$  and  $y = -\frac{1}{2g} v^2$ , and by integration we get

$$\begin{aligned} x(t) &= \frac{1}{2p_x} t - \frac{1}{4gp_x^2} \sin(2gp_x t), \\ y(t) &= -\frac{1}{2gp_x^2} \sin^2(gp_x t) = -\frac{1}{4gp_x^2} (1 - \cos(2gp_x t)). \end{aligned}$$

Note that, since  $\dot{x} = p_x v^2$ , we must have  $p_x > 0$ .

Representing in the plane the parametrized curves  $(x(t), y(t))$  (with parameters  $p_x$  and  $t$ ), we get *cycloid curves*.

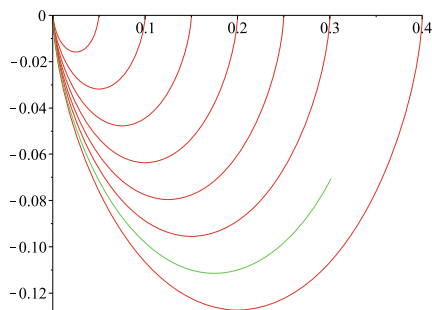
Let us now prove that there is a unique optimal trajectory joining  $(x_1, y_1)$ , and it has at most one cycloid arch.

Let us first compute  $p_x$  and  $t_f$  in the case where  $y_1 = 0$ . If  $y_1 = 0$  then  $\sin(gp_x t_f) = 0$  hence  $gp_x t_f = \pi + k\pi$  with  $k \in \mathbb{Z}$ , but since  $t_f$  must be minimal, we must have  $k = 0$  (and this is what will imply that optimal trajectories have at most one arch). Hence  $p_x = \frac{t_f}{2x_1}$ . Since  $2gp_x t_f = 2\pi$ , we have  $x_1 = x(t_f) = \frac{t_f}{2p_x}$ , and thus  $t_f = \sqrt{\frac{2\pi x_1}{g}}$  and  $p_x = \sqrt{\frac{\pi}{2gx_1}}$ .

On Fig. 2.6, we have represented all optimal trajectories joining points  $(x_1, 0)$ , with  $x_1 > 0$ .

Now, we note that, if a trajectory is optimal on the interval  $[0, t_f]$ , then, for any  $t_1 \in ]0, t_f[$ , it is optimal as well on the interval  $[0, t_1]$  for the problem of reaching the point  $(x(t_1), y(t_1))$  in minimal time (this is the dynamic programming principle). From that remark, we deduce that any optimal trajectory of the problem (2.20) is the truncation of an optimal trajectory reaching a point of the abscissa axis. In particular, we get the desired uniqueness property, and the fact that such a trajectory has at most one point at which  $\dot{y} = 0$  (hence, at most one arch). See Fig. 2.6.

**Fig. 2.6** Optimal trajectories



Moreover, if  $\dot{y}(t) = 0$  then  $2gp_xt = \pi + k\pi$  with  $k \in \mathbb{Z}$ , and necessarily (by optimality)  $k = 0$ , hence  $t = \frac{\pi}{2gp_x}$ . Therefore the set of points where  $\dot{y}(t) = 0$  is the parametrized curve  $x(p_x) = \frac{\pi}{4gp_x^2}$ ,  $y(p_x) = -\frac{1}{2gp_x^2}$ , that is the graph  $y = -\frac{2}{\pi}x$ . Therefore, we have proved that the optimal trajectory  $(x(t), y(t))$  reaching  $(x_1, y_1)$  is such that

- $y(t)$  passes through a minimum if  $y_1 > -\frac{2}{\pi}x_1$ ,
- $y(t)$  is decreasing if  $y_1 < -\frac{2}{\pi}x_1$ .

**Remark 2.26** If we investigate the variant of the optimal control problem (2.20), for which we minimize the final time  $t_f$  with  $y(t_f)$  free, then in the reduced problem we have moreover that  $v(t_f)$  is free, and hence we gain the transversality condition  $p_v(t_f) = 0$ , hence  $\dot{v}(t_f) = 0$ . This gives  $2gp_xt_f = \pi$ , in other words, we find exactly the final points of the previously computed optimal trajectories, stopping when  $\dot{y} = 0$ .

This means that, if  $y(t_f)$  is free (with  $x_1$  fixed), we minimize the time  $t_f$  by choosing, on Fig. 2.6, the arc of cycloid starting from the origin and reaching  $x = x_1$  with an horizontal tangent.

## References

1. Agrachev, A., Sachkov, Y.: Control theory from the geometric viewpoint. Encyclopaedia of Mathematical Sciences. Control Theory and Optimization, II, vol. 87. Springer-Verlag, Berlin (2004)
2. Anderson, B.D., Moore, J.B.: Optimal Filtering. Prentice Hall, Hoboken (1979)
3. Betts, J.T.: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming. Advances in Design and Control, vol. 19, 2nd edn. SIAM, Philadelphia (2010)
4. Bonnard, B., Caillaud, J.-B., Trélat, E.: Second order optimality conditions in the smooth case and applications in optimal control. ESAIM Control Optim. Calc. Var. **13**(2), 207–236 (2007)
5. Bonnard, B., Chyba, M.: Singular Trajectories and Their Role in Control Theory. Mathematics And Its Applications, vol. 40. Springer-Verlag, Berlin (2003)
6. Bonnard, B., Faubourg, L., Trélat, E.: Mécanique céleste et contrôle de Systèmes Spatiaux. Mathematics And Its Applications, vol. 51, 276 pages. Springer Verlag, Berlin (2006)
7. Bourdin, L., Trélat, E.: Pontryagin Maximum Principle for finite dimensional nonlinear optimal control problems on time scales. SIAM J. Control Optim. **51**(5), 3781–3813 (2013)
8. Bressan, A., Piccoli, B.: Introduction to the Mathematical Theory of Control. AIMS Series on Applied Mathematics, vol. 2, Springfield, Madrid (2007)
9. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext. Springer, New York (2011)
10. Bryson, A., Ho, Y.C.: Applied Optimal Control. Hemisphere Publishing Corporation, London (1975)
11. Cesari, L.: Optimization – Theory and Applications. Problems with Ordinary Differential Equations. Applications of Mathematics, vol. 17. Springer-Verlag, Berlin (1983)
12. Clarke, F.H.: Optimization and Nonsmooth Analysis. Canadian Mathematical Society Series of Monographs and Advanced Texts. John Wiley & Sons, Inc., New York (1983)
13. Dmitruk, A.V.: On the development of Pontryagin's maximum principle in the works of A. Ya. Dubovitskii and A. A. Milyutin. Control Cybernet. **38**(4A), 923–957 (2009)
14. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**, 324–353 (1974)

15. Garavello, M., Piccoli, B.: Hybrid necessary principle. *SIAM J. Control Optim.* **43**(5), 1867–1887 (2005)
16. Haberkorn, T., Trélat, E.: Convergence results for smooth regularizations of hybrid nonlinear optimal control problems. *SIAM J. Control Optim.* **49**(4), 1498–1522 (2011)
17. Hartl, R.F., Sethi, S.P., Vickson, R.G.: A survey of the maximum principles for optimal control problems with state constraints. *SIAM Rev.* **37**(2), 181–218 (1995)
18. Hermes, H., LaSalle, J.P.: *Functional Analysis and Time Optimal Control*. Mathematics in Science and Engineering, vol. 56. Academic Press, Cambridge (1969)
19. Ioffe, A.D., Tihomirov, V.M.: *Theory of Extremal Problems*. Studies in Mathematics and its Applications, vol. 6. North-Holland Publishing Co., Amsterdam (1979)
20. Jacobson, D.H., Lele, M.M., Speyer, J.L.: New necessary conditions of optimality for control problems with state-variable inequality constraints. *J. Math. Anal. Appl.* **35**, 255–284 (1971)
21. Kailath, T.: *Linear Systems*. Prentice Hall, Hoboken (1980)
22. Khalil, H.K.: *Nonlinear Systems*. Macmillan Publishing Company, New York (1992)
23. Kwakernaak, H., Sivan, R.: *Linear Optimal Control Systems*. John Wiley, Hoboken (1972)
24. Lee, E.B., Markus, L.: *Foundations of Optimal Control Theory*. John Wiley, New York (1967)
25. Li, X., Yong, J.: *Optimal Control Theory for Infinite-Dimensional Systems*. Systems & Control: Foundations & Applications. Birkhäuser, Boston (1995)
26. Maurer, H.: On optimal control problems with bounded state variables and control appearing linearly. *SIAM J. Cont. Optim.* **15**, 345–362 (1977)
27. Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.: *The Mathematical Theory of Optimal Processes*, viii+360 pp. John Wiley & Sons, New York-London (1962)
28. Sontag, E.D.: *Mathematical Control Theory: Deterministic Finite-Dimensional Systems*. Texts in Applied Mathematics, vol. 6, , 2nd edn., xvi+531. Springer-Verlag, New York (1998)
29. Trélat, E.: *Contrôle Optimal (French) [Optimal control] Théorie & Applications [Theory and applications]*. Mathématiques Concrètes [Concrete Mathematics], vi+246 pp. Vuibert, Paris (2005)
30. Trélat, E.: Optimal control and applications to aerospace: some results and challenges. *J. Optim. Theory Appl.* **154**(3), 713–758 (2012)
31. Vinter, R.: *Optimal Control. Systems & Control: Foundations & Applications*. Birkhäuser, Boston (2000)

## Chapter 3

# Stabilization



In this chapter, our objective will be to stabilize a possibly unstable equilibrium point by means of a feedback control.

Let  $n$  and  $m$  be two positive integers. In this chapter we consider an autonomous control system in  $\mathbb{R}^n$

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

where  $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  is of class  $C^1$  with respect to  $(x, u)$ , and the controls are measurable essentially bounded functions of time taking their values in some measurable subset  $\Omega$  of  $\mathbb{R}^m$  (set of control constraints).

Let  $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \mathbb{R}^m$  be an equilibrium point, that is,  $f(\bar{x}, \bar{u}) = 0$ , such that  $\bar{u} \in \overset{\circ}{\Omega}$  (interior of  $\Omega$ ). Our objective is to be able to design a feedback control  $u(x)$  stabilizing locally the equilibrium  $(\bar{x}, \bar{u})$ , that is, such that the closed-loop system  $\dot{x}(t) = f(x(t), u(x(t)))$  be locally asymptotically stable at  $\bar{x}$ .

### 3.1 Stabilization of Autonomous Linear Systems

#### 3.1.1 Reminders on Stability Notions

Consider the linear system  $\dot{x}(t) = Ax(t)$ , with  $A$  a  $n \times n$  matrix. The point 0 is of course an equilibrium point of the system (it is the only one if  $A$  is invertible). We have the following well-known result (easy to prove with simple linear algebra considerations).

### Theorem 3.13

- If there exists a (complex) eigenvalue  $\lambda$  of  $A$  such that  $\operatorname{Re}(\lambda) > 0$ , then the equilibrium point  $0$  is unstable, meaning that there exists  $x_0 \in \mathbb{R}^n$  such that the solution of  $\dot{x}(t) = Ax(t)$ ,  $x(0) = x_0$  satisfies  $\|x(t)\| \rightarrow +\infty$  as  $t \rightarrow +\infty$ .
- If all (complex) eigenvalues of  $A$  have negative real part, then  $0$  is asymptotically stable, meaning that all solutions of  $\dot{x}(t) = Ax(t)$  converge to  $0$  as  $t \rightarrow +\infty$ .
- The equilibrium point  $0$  is stable (i.e., not unstable)<sup>1</sup> if and only if all eigenvalues of  $A$  have nonpositive real part and if an eigenvalue  $\lambda$  is such that  $\operatorname{Re}(\lambda) = 0$  then  $\lambda$  is a simple root<sup>2</sup> of the minimal polynomial of  $A$ .

**Definition 3.5** The matrix  $A$  is said to be *Hurwitz* if all its eigenvalues have negative real part.

We are next going to see two classical criteria ensuring that a given matrix (with real coefficients) is Hurwitz. These criteria are particularly remarkable because they are purely algebraic (polynomial conditions on the coefficients of the matrix), and they do not require the calculation of the roots of the characteristic polynomial of the matrix (which is impossible to achieve algebraically in general, for degrees larger than 5, as is well known from Galois theory).

**Routh Criterion** We consider the complex polynomial

$$P(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n$$

with real coefficients  $a_i$ , and we are going to formulate some conditions under which all roots of this polynomial have negative real part (in that case we also say that  $P$  is Hurwitz). Note that  $A$  is Hurwitz if and only if its characteristic polynomial  $\chi_A$  is Hurwitz.

**Definition 3.6** The *Routh table* is defined as follows:

$$\begin{array}{llllll}
 a_0 & a_2 & a_4 & a_6 & \cdots & \text{completed by } 0 \\
 a_1 & a_3 & a_5 & a_7 & \cdots & \text{completed by } 0 \\
 b_1 & b_2 & b_3 & b_4 & \cdots & \text{where } b_1 = \frac{a_1 a_2 - a_0 a_3}{a_1}, b_2 = \frac{a_1 a_4 - a_0 a_5}{a_1}, \dots \\
 c_1 & c_2 & c_3 & c_4 & \cdots & \text{where } c_1 = \frac{b_1 a_3 - a_1 b_2}{b_1}, c_2 = \frac{b_1 a_5 - a_1 b_3}{b_1}, \dots \\
 \vdots & \vdots & \vdots & \vdots & & 
 \end{array}$$

<sup>1</sup> See also the general Definition 3.8 further.

<sup>2</sup> Equivalently,  $\ker(A - \lambda I_n) = \ker(A - \lambda I_n)^2$ , or, equivalently, the Jordan decomposition of  $A$  does not have any strict Jordan block.

The process goes on as long as the first element of the row is not equal to 0. The process stops when we have built  $n + 1$  rows.

The Routh table is said to be *complete* if it has  $n + 1$  rows whose first coefficient is not equal to 0.

We have the two following theorems (stated in [9]), which can be proved by means of (nonelementary) complex analysis.

**Theorem 3.14** *All roots of  $P$  have negative real part if and only if the Routh table is complete and the elements in the first column have the same sign.*

**Theorem 3.15** *If the Routh table is complete then  $P$  has no purely imaginary root, and the number of roots with positive real part is equal to the number of sign changes in the first column.*

**Hurwitz Criterion** We set  $a_{n+1} = a_{n+2} = \dots = a_{2n-1} = 0$ , and we define

$$H = \begin{pmatrix} a_1 & a_3 & a_5 & \cdots & a_{2n-1} \\ a_0 & a_2 & a_4 & \cdots & a_{2n-2} \\ 0 & a_1 & a_3 & \cdots & a_{2n-3} \\ 0 & a_0 & a_2 & \cdots & a_{2n-4} \\ 0 & 0 & a_1 & \cdots & a_{2n-5} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & * & \cdots & a_n \end{pmatrix}$$

where  $*$  =  $a_0$  or  $a_1$  according to the parity of  $n$ . Let  $(H_i)_{i \in \{1, \dots, n\}}$  be the principal minors of  $H$ , defined by

$$H_1 = a_1, \quad H_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix}, \quad H_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ a_0 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix}, \quad \dots, \quad H_n = \det H.$$

**Theorem 3.16 ([4])** *If  $a_0 > 0$ , then  $P$  is Hurwitz if and only if  $H_i > 0$  for every  $i \in \{1, \dots, n\}$ .*

**Remark 3.27** Assume that  $a_0 > 0$ .

If all roots of  $P$  have nonpositive real part, then  $a_k \geq 0$  and  $H_k \geq 0$ , for every  $k \in \{1, \dots, n\}$ .

If  $n \leq 3$ ,  $a_k \geq 0$  and  $H_k \geq 0$  for every  $k \in \{1, 2, 3\}$ , then all roots of  $P$  have nonpositive real part.

A necessary condition for stability is that  $a_k \geq 0$  for every  $k \in \{1, \dots, n\}$ . This condition is however not sufficient (take  $P(z) = z^4 + z^2 + 1$ ).

### 3.1.2 Pole-Shifting Theorem

**Definition 3.7** The linear autonomous control system  $\dot{x}(t) = Ax(t) + Bu(t)$ , with  $x(t) \in \mathbb{R}^n$ ,  $u(t) \in \mathbb{R}^m$ ,  $A$  a  $n \times n$  matrix and  $B$  a  $n \times m$  matrix, is said to be *feedback stabilizable* if there exists a  $m \times n$  matrix  $K$  (called gain matrix) such that the closed-loop system with the (linear) feedback  $u(t) = Kx(t)$ ,

$$\dot{x}(t) = (A + BK)x(t)$$

is asymptotically stable, i.e., equivalently,  $A + BK$  is Hurwitz.

**Remark 3.28** This concept is invariant under similar transforms  $A_1 = PAP^{-1}$ ,  $B_1 = PB$ ,  $K_1 = KP^{-1}$ .

**Theorem 3.17 (Pole-Shifting Theorem)** *If  $(A, B)$  satisfies the Kalman condition  $\text{rank } K(A, B) = n$ , then for every real polynomial  $P$  of degree  $n$  whose leading coefficient is 1, there exists a  $m \times n$  matrix  $K$  such that the characteristic polynomial  $\chi_{A+BK}$  of  $A + BK$  is equal to  $P$ .<sup>3</sup>*

**Corollary 3.3** *If the linear control system  $\dot{x}(t) = Ax(t) + Bu(t)$  is controllable then it is stabilizable.*

To prove the corollary, it suffices to take for instance  $P(X) = (X + 1)^n$  and to apply the pole-shifting theorem.

**Proof of Theorem 3.17** We prove the result first in the case  $m = 1$ . It follows from Theorem 1.2 (Brunovski normal form) that the system is similar to

$$A = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Setting  $K = (k_1 \cdots k_n)$  and  $u = Kx$ , we have

$$A + BK = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ k_1 - a_n & k_2 - a_{n-1} & \cdots & k_n - a_1 \end{pmatrix}$$

---

<sup>3</sup> Actually, the converse statement is also true.

and thus  $\chi_{A+BK}(X) = X^n + (a_1 - k_n)X^{n-1} + \cdots + (a_n - k_1)$ . Therefore, for every polynomial  $P(X) = X^n + \alpha_1 X^{n-1} + \cdots + \alpha_n$ , it suffices to choose  $k_1 = a_n - \alpha_n, \dots, k_n = a_1 - \alpha_1$ .

Let us now prove that the general case  $m \geq 1$  can be reduced to the case  $m = 1$ . We have the following lemma.

**Lemma 3.4** *If  $(A, B)$  satisfies the Kalman condition, then there exists  $y \in \mathbb{R}^m$  and a  $m \times n$  matrix  $C$  such that  $(A + BC, By)$  satisfies the Kalman condition.*

The proof of this lemma is done hereafter. It follows from Lemma 3.4 that, for every polynomial  $P$  of degree  $n$  whose leading coefficient is 1, there exists a  $1 \times n$  matrix  $K_1$  such that  $\chi_{A+BC+ByK_1} = P$ , and therefore, defining the  $m \times n$  matrix  $K = C + yK_1$ , we have  $\chi_{A+BK} = P$ , and Theorem 3.17 is proved.  $\square$

**Proof of Lemma 3.4** Let  $y \in \mathbb{R}^m$  be such that  $By \neq 0$ . Let  $x_1 = By$ .

*Claim 1* There exists  $x_2 \in Ax_1 + \text{Ran}(B)$  (and thus there exists  $y_1 \in \mathbb{R}^m$  such that  $x_2 = Ax_1 + By_1$ ) such that  $\dim(\text{Span}(x_1, x_2)) = 2$ .

Indeed, otherwise,  $Ax_1 + \text{Ran}(B) \subset \mathbb{R}x_1$ , hence  $Ax_1 \in \mathbb{R}x_1$  and  $\text{Ran}(B) \subset \mathbb{R}x_1$ . Therefore  $\text{Ran}(AB) = A\text{Ran}(B) \subset \mathbb{R}Ax_1 \subset \mathbb{R}x_1$  and by immediate iteration,  $\text{Ran}(A^k B) \subset \mathbb{R}x_1$ , for every integer  $k$ . This implies that

$$\text{Ran}(B, AB, \dots, A^{n-1}B) = \text{Ran}(B) + \text{Ran}(AB) + \cdots + \text{Ran}(A^{n-1}B) \subset \mathbb{R}x_1$$

which contradicts the Kalman condition.

*Claim 2* For every  $k \leq n$ , there exists  $x_k \in Ax_{k-1} + \text{Ran}(B)$  (and thus there exists  $y_{k-1} \in \mathbb{R}^m$  such that  $x_k = Ax_{k-1} + By_{k-1}$ ) such that  $\dim(E_k) = k$ , where  $E_k = \text{Span}(x_1, \dots, x_k)$ .

Indeed, otherwise,  $Ax_{k-1} + \text{Ran}(B) \subset E_{k-1}$ , and hence  $Ax_{k-1} \subset E_{k-1}$  and  $\text{Ran}(B) \subset E_{k-1}$ . Let us then prove that  $AE_{k-1} \subset E_{k-1}$ . Indeed, note that  $Ax_1 = x_2 - By_1 \in E_{k-1} + \text{Ran}(B) \subset E_{k-1}$ , and similarly for  $Ax_2$ , etc,  $Ax_{k-2} = x_{k-1} - By_{k-1} \in E_{k-1} + \text{Ran}(B) \subset E_{k-1}$ , and finally,  $Ax_{k-1} \in E_{k-1}$ .

Therefore  $\text{Ran}(AB) = A\text{Ran}(B) \subset AE_{k-1} \subset E_{k-1}$ , and similarly we have  $\text{Ran}(A^i B) \subset E_{k-1}$  for every integer  $i$ . Hence  $\text{Ran}(B, AB, \dots, A^{n-1}B) \subset E_{k-1}$ , which contradicts the Kalman condition.

We have thus built a basis  $(x_1, \dots, x_n)$  of  $\mathbb{R}^n$ . We define the  $m \times n$  matrix  $C$  by the relations  $Cx_1 = y_1, Cx_2 = y_2, \dots, Cx_{n-1} = y_{n-1}$ , and  $Cx_n$  arbitrary. Then  $(A + BC, x_1)$  satisfies the Kalman condition since  $(A + BC)x_1 = Ax_1 + By_1 = x_2, \dots, (A + BC)x_{n-1} = Ax_{n-1} + By_{n-1} = x_n$ . Lemma 3.4 is proved.  $\square$

**Remark 3.29** To stabilize a linear control system in practice, one has the following solutions:

- If  $n$  is not too large, one can apply the Routh or Hurwitz criteria and thus determine an algebraic necessary and sufficient condition on the coefficients of  $K$  ensuring the desired stabilization property. Note that the characteristic



polynomial of  $A + BK$  can be computed with a formal computations software like *Maple*.

- There exist many numerical routines in order to compute numerical gain matrices. In the *Matlab Control Toolbox*, we quote *acker.m*, based on Ackermann's formula (see [6]), limited however to  $m = 1$  and not very reliable numerically. Better is to use *place.m*, which is a robust pole-shifting routine (see [7]) based on spectral considerations (but in which the desired poles have to be distinct two by two).
- Another way consists of applying the LQ theory, elements of which have been given in Sect. 2.3.2, by taking an infinite horizon of time  $T = +\infty$  as quickly mentioned at the end of that section (LQR stabilization).

## 3.2 Stabilization of Instationary Linear Systems

For instationary linear systems  $\dot{x}(t) = A(t)x(t) + B(t)u(t)$ , the situation is much more complicated and there is no simple and definitive theory as in the autonomous case.

Let us explain which difficulties appear, by considering the system  $\dot{x}(t) = A(t)x(t)$  without any control. A priori one could expect that, if the matrix  $A(t)$  is Hurwitz for every  $t$ , then the system is asymptotically stable. This is however wrong. The statement remains wrong even under stronger assumptions on  $A(t)$ , such as assuming that there exists  $\varepsilon > 0$  such that, for every time  $t$ , every (complex) eigenvalue  $\lambda(t)$  of  $A(t)$  satisfies  $\text{Re}(\lambda(t)) \leq -\varepsilon$ . Indeed, for example, take

$$A(t) = \begin{pmatrix} -1 + a \cos^2 t & 1 - a \sin t \cos t \\ -1 - a \sin t \cos t & -1 + a \sin^2 t \end{pmatrix}$$

with  $a \in [1, 2)$  arbitrary. Then  $x(t) = e^{(a-1)t} (\cos t, -\sin t)^\top$  is a solution of  $\dot{x}(t) = A(t)x(t)$ , and does not converge to 0 whenever  $a \geq 1$ . Besides, it can be shown that if  $a < 1$  then the system is asymptotically stable.

Let us explain the reason of this failure. A simple way to understand is the following (not so much restrictive) case. Let us assume that, for every  $t$ ,  $A(t)$  is diagonalizable, and that there exists  $P(t)$  invertible such that  $P(t)^{-1}A(t)P(t) = D(t)$ , with  $D(t)$  diagonal, and  $P(\cdot)$  and  $D(\cdot)$  of class  $C^1$ . Setting  $x(t) = P(t)y(t)$ , we get immediately that

$$\dot{y}(t) = (D(t) - P(t)^{-1}\dot{P}(t))y(t).$$

If the term  $P(t)^{-1}\dot{P}(t)$  were equal to 0 (as it is the case in the autonomous case), then, obviously, the asymptotic stability would hold true as soon as the eigenvalues (diagonal of  $D(t)$ ) would have negative real parts. But, even if  $D(t)$  is Hurwitz, the

term  $P(t)^{-1}\dot{P}(t)$  can destabilize the matrix and imply the failure of the asymptotic stability.

In other words, what may imply the divergence is the fact that the eigenvectors (making up the columns of  $P(t)$ ) may evolve quickly in time, thus implying that the norm of  $\dot{P}(t)$  be large.

To end up however with a positive result, it can be noted that, if the matrix  $A(t)$  is slowly varying in time, then the norm of the term  $P(t)^{-1}\dot{P}(t)$  is small, and then if one is able to ensure that this norm is small enough with respect to the diagonal  $D(t)$ , then one can ensure an asymptotic stability result. This is the theory of slowly time-varying linear systems (see [8]).

### 3.3 Stabilization of Nonlinear Systems

#### 3.3.1 Local Stabilization by Linearization

**Reminders** Consider the system  $\dot{x}(t) = f(x(t))$ , where  $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is of class  $C^1$ . We denote by  $x(\cdot, x_0)$  the unique solution of this system such that  $x(0, x_0) = x_0$ . We assume that  $\bar{x}$  is an *equilibrium point*, that is,  $f(\bar{x}) = 0$ .

**Definition 3.8** The equilibrium point  $\bar{x}$  is said to be *stable* if, for every  $\varepsilon > 0$ , there exists  $\delta > 0$  such that, for every initial point  $x_0$  such that  $\|x_0 - \bar{x}\| \leq \delta$ , one has  $\|x(t, x_0) - \bar{x}\| \leq \varepsilon$  for every  $t \geq 0$ . It is said to be *locally asymptotically stable* (in short, LAS) if it is stable and if moreover  $x(t, x_0) \rightarrow \bar{x}$  as  $t \rightarrow +\infty$  for every  $x_0$  in some neighborhood of  $\bar{x}$ . If the neighborhood is the whole  $\mathbb{R}^n$  then we speak of *global asymptotic stability* (in short, GAS). If an asymptotic stability result is established in some neighborhood  $V$  of  $\bar{x}$ , then we say that  $\bar{x}$  is GAS in  $V$ .

**Theorem 3.18** Let  $A$  be the Jacobian matrix of  $f$  at the equilibrium point  $\bar{x}$ . If all eigenvalues of  $A$  have negative real parts (that is, if  $A$  is Hurwitz), then  $\bar{x}$  is LAS. If  $A$  has an eigenvalue with a positive real part then  $\bar{x}$  is not LAS.

The above linearization theorem is an easy first result (see Example 3.17, below, for a proof), not saying anything however, for the moment, on the size of the neighborhoods of stability.

**Application: Local Stabilization of Nonlinear Control Systems** Consider the general nonlinear control system (3.1),  $\dot{x} = f(x, u)$ , and an equilibrium point  $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \hat{\Omega}$ , as settled at the beginning of Chap. 3. Setting  $x(t) = \bar{x} + \delta x(t)$  and  $u(t) = \bar{u} + \delta u(t)$  and keeping the terms of order 1, we obtain (as already discussed for controllability issues) the linearized system at  $(\bar{x}, \bar{u})$ ,

$$\delta \dot{x}(t) = A \delta x(t) + B \delta u(t)$$

where

$$A = \frac{\partial f}{\partial x}(\bar{x}, \bar{u}) \quad \text{and} \quad B = \frac{\partial f}{\partial u}(\bar{x}, \bar{u}).$$

If one can stabilize the linearized system, that is, find a matrix  $K$  of size  $m \times n$  such that  $A + BK$  is Hurwitz (and take  $\delta u = K\delta x$ ), then Theorem 3.18 implies a local stabilization result for the nonlinear control system (3.1). We thus have the following theorem.

**Theorem 3.19** *If the pair  $(A, B)$  satisfies the Kalman condition, then there exists a matrix  $K$  of size  $m \times n$  such that the feedback  $u = K(x - \bar{x}) + \bar{u}$  stabilizes asymptotically the control system (3.1) locally around  $(\bar{x}, \bar{u})$ : the closed-loop system  $\dot{x}(t) = f(x(t), K(x(t) - \bar{x}) + \bar{u})$  is LAS at  $\bar{x}$ .*

Note that the stability neighborhood must be small enough so that the closed-loop control  $u$  takes its values in the set  $\Omega$ .

**Example 3.13** Consider the inverted pendulum system given in Example 1.6. Applying the Routh criterion (Theorem 3.14) and then Theorem 3.19, we establish that a sufficient condition on  $K = (k_1, k_2, k_3, k_4)$  to stabilize the inverted pendulum locally at the unstable equilibrium  $(\xi, 0, 0, 0)^\top$  is

$$\begin{aligned} k_1 &> 0, & k_4 - k_2L &> 0, & k_3 - k_1L - (m + M)g &> 0, \\ k_2((k_4 - k_2L)(k_3 - k_1L - (m + M)g) - MLgk_2) &> k_1(k_4 - k_2L)^2. \end{aligned}$$

**Example 3.14** Consider the Maxwell-Bloch system given in Example 1.7. Let us stabilize locally the system at the equilibrium point  $\bar{x} = (a, 0, 0)$ ,  $\bar{u} = (0, 0)$ , where  $a \neq 0$  is fixed.

By Example 1.7, (2), the linearized system (with  $c = 0$  and  $a \neq 0$ ) at this point is controllable and thus stabilizable with a linear feedback of matrix  $K$ . We seek a particular matrix  $K$  stabilizing the system, of the form  $K = \begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \end{pmatrix}$ . By the Routh criterion, it is easy to see that  $A + BK$  is Hurwitz if and only if  $k_1 < 0$  and  $k_2 < 0$ . We infer from Theorem 3.19 that the Maxwell-Bloch system is locally stabilizable around  $\bar{x}$ , with feedbacks  $u_1 = k_1(x_1 - a)$ ,  $u_2 = k_2x_2$  where  $k_1, k_2 < 0$ .

### 3.3.2 Global Stabilization by Lyapunov Theory

**Reminders: Lyapunov and LaSalle Theorems** Let  $\mathcal{D}$  be an open subset of  $\mathbb{R}^n$  and let  $f : \mathcal{D} \rightarrow \mathbb{R}^n$  continuous. Following [2, Chapter X] (see also [3]) we assume that, for any  $x_0 \in \mathcal{D}$ , the system  $\dot{x}(t) = f(x(t))$  has a unique solution  $x(\cdot)$  such that  $x(0) = x_0$  (this is of course true if  $f$  is  $C^1$ ) and satisfying  $x(t) \in \mathcal{D}$  for every  $t \geq 0$

(i.e.,  $\mathcal{D}$  is an invariant domain under the dynamics). We assume that there exists an *equilibrium point*  $\bar{x} \in \mathcal{D}$ , i.e., satisfying  $f(\bar{x}) = 0$ .

Let us recall two important results of Lyapunov theory, providing knowledge on the stability neighborhoods, with the concept of Lyapunov function.

**Definition 3.9** A function  $V : \mathcal{D} \rightarrow \mathbb{R}$  is a *Lyapunov function* at  $\bar{x}$  on  $\mathcal{D}$  if

- $V$  is of class  $C^1$  on  $\mathcal{D}$ ;
- $V(\bar{x}) = 0$  and  $V(x) > 0$  for every  $x \in \mathcal{D} \setminus \{\bar{x}\}$ ;
- $\langle \nabla V(x), f(x) \rangle \leq 0$  for every  $x \in \mathcal{D}$ . If the inequality is strict on  $\mathcal{D} \setminus \{\bar{x}\}$  then the Lyapunov function  $V$  is said to be strict.

Note that, along a given trajectory of the dynamical system, one has

$$\frac{d}{dt} V(x(t)) = \langle \nabla V(x(t)), f(x(t)) \rangle.$$

Therefore if  $V$  is a Lyapunov function then the value of  $V$  is nonincreasing along any trajectory. A Lyapunov function can be seen as a potential well, ensuring stability.

**Theorem 3.20 (Lyapunov Theorem, see [2])** *If there exists a Lyapunov function  $V$  at  $\bar{x}$  on  $\mathcal{D}$  then  $\bar{x}$  is stable, and if  $V$  is strict then  $\bar{x}$  is LAS. If  $V$  is strict and proper<sup>4</sup> then  $\bar{x}$  is GAS in  $\mathcal{D}$ .*

When a Lyapunov function is not strict then one can be more specific and infer that trajectories converge to some invariant subset. The following result is even more general and does not assume the existence of an equilibrium point.

**Theorem 3.21 (LaSalle Principle, see [2])** *Assume that  $V : \mathcal{D} \rightarrow [0, +\infty)$  is a proper  $C^1$  function such that  $\langle \nabla V(x), f(x) \rangle \leq 0$  for every  $x \in \mathcal{D}$ . Let  $\mathcal{I}$  be the largest subset of  $\{x \in \mathcal{D} \mid \langle \nabla V(x), f(x) \rangle = 0\}$  that is invariant under the flow (in positive time) of the dynamical system. Then all trajectories starting in  $\mathcal{D}$  converge to  $\mathcal{I}$ , in the sense that  $d(x(t), \mathcal{I}) \rightarrow 0$  (Euclidean distance) as  $t \rightarrow +\infty$ .*

**Remark 3.30** It is interesting to formulate the LaSalle principle in the particular case where the invariant set  $\mathcal{I}$  is reduced to a singleton (which must then be an equilibrium point). The statement is then as follows.

Assume that  $V$  is a proper Lyapunov function at  $\bar{x}$  on  $\mathcal{D}$  and that, if  $x(\cdot)$  is a solution of the system such that  $\langle \nabla V(x(t)), f(x(t)) \rangle = 0$  for every  $t \geq 0$ , then  $x(t) = \bar{x}$ . Then the equilibrium point  $\bar{x}$  is GAS in  $\mathcal{D}$ .

**Example 3.15** Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a function of class  $C^1$  such that  $g(0) = 0$  and  $xg(x) > 0$  if  $x \neq 0$ , and satisfying  $\int_0^{+\infty} g = +\infty$  and  $\int_{-\infty}^0 g = -\infty$ . By

<sup>4</sup>  $V$  is said to be proper whenever  $V^{-1}([0, L])$  is a compact subset of  $\mathcal{D}$ , for every  $L \geq 0$ ; in other words, the inverse image of every compact is compact. When  $\mathcal{D} = \mathbb{R}^n$ ,  $V$  is proper if and only if  $V(x) \rightarrow +\infty$  as  $\|x\| \rightarrow +\infty$ .

considering the Lyapunov function  $V(x, y) = \frac{1}{2}y^2 + \int_0^x g(s) ds$ , it is easy to prove that the point  $x = \dot{x} = 0$  is GAS for the system  $\ddot{x} + \dot{x} + g(x) = 0$  (which has to be written as a first-order system).

**Example 3.16** Consider the system in  $\mathbb{R}^2$

$$\dot{x} = \alpha x - y - \alpha x(x^2 + y^2), \quad \dot{y} = x + \alpha y - \alpha y(x^2 + y^2),$$

with  $\alpha > 0$  arbitrary. The equilibrium point  $(0, 0)$  is not stable. With the LaSalle principle, it is easy to prove that the unit circle  $x^2 + y^2 = 1$  is globally attractive in  $\mathbb{R}^2 \setminus \{(0, 0)\}$ , in the sense that a trajectory, starting from any point different from  $(0, 0)$ , converges to the circle in infinite time. Indeed, note that, setting  $V(x, y) = \frac{1}{2}(x^2 + y^2)$ , we have

$$\frac{d}{dt} V(x(t), y(t)) = \alpha(x(t)^2 + y(t)^2)(1 - x(t)^2 - y(t)^2)$$

and one can see that  $\frac{d}{dt} V(x(t), y(t))$  is positive inside the unit disk (except at the origin), and negative outside of the unit disk. It is then easy to design (by translation) Lyapunov functions inside the punctured unit disk, and outside of the unit disk, and to conclude by the LaSalle principle.

**Example 3.17 (Lyapunov Lemma and Applications)** Let  $A$  be a  $n \times n$  real matrix, whose eigenvalues have negative real parts. Then there exists a symmetric positive definite  $n \times n$  matrix  $P$  such that  $A^\top P + PA = -I_n$ . Indeed, it suffices to take  $P = \int_0^{+\infty} e^{tA^\top} e^{tA} dt$ .

Clearly, the function  $V(x) = \langle x, Px \rangle$  is then a strict Lyapunov function for the system  $\dot{x}(t) = Ax(t)$ . We recover the fact that 0 is GAS.

Using a first-order expansion and norm estimates, it is then easy to prove Theorem 3.18 and even to obtain stability neighborhoods thanks to  $V$ .

**Application to the Stabilization of Nonlinear Control Systems** As Theorem 3.18 implied Theorem 3.19, the Lyapunov and LaSalle theorems can be applied to control systems, providing knowledge on the stability neighborhoods. For instance, we get the following statement: consider the nonlinear control system (3.1) and assume that there exists a function  $V : \mathcal{D} \rightarrow \mathbb{R}^+$  of class  $C^1$ , taking positive values in  $\mathcal{D} \setminus \{\bar{x}\}$ , such that for every  $x \in \mathcal{D}$  there exists  $u(x) \in \Omega$  such that  $\langle \nabla V(x), f(x, u(x)) \rangle < 0$ ; then the feedback control  $u$  stabilizes the control system, globally in  $\mathcal{D}$ .

Many other similar statements can be easily derived, based on the Lyapunov or LaSalle theorems. There exists an elaborate theory of control Lyapunov functions (see, e.g., [10]). A difficulty in this theory is to ensure a nice regularity of the feedback control. We do not discuss further this difficult question but we mention that it has raised a whole field of intensive researches (see [1, 10]).

To illustrate the role of the Lyapunov functions in stabilization, and as an application of the idea described above, we next provide a spectacular and widely used (and however very simple) strategy in order to design stabilizing controls thanks to Lyapunov functions.

**Theorem 3.22 (Jurdjevic-Quinn Method [5])** *Consider the control-affine system in  $\mathbb{R}^n$*

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^m u_i(t) g_i(x(t))$$

where  $f$  and the  $g_i$ 's are smooth vector fields in  $\mathbb{R}^n$ . Let  $\bar{x}$  be such that  $f(\bar{x}) = 0$ , i.e.,  $\bar{x}$  is an equilibrium point of the uncontrolled system (that is, with  $u_i = 0$ ).

We assume that there exists a proper Lyapunov function  $V$  at  $\bar{x}$  on  $\mathbb{R}^n$  for the uncontrolled system, i.e., satisfying

- $V(\bar{x}) = 0$  and  $V(x) > 0$  for every  $x \in \mathbb{R}^n \setminus \{\bar{x}\}$ ;
- $V$  is proper;
- $L_f V(x) = \langle \nabla V(x), f(x) \rangle \leq 0$  for every  $x \in \mathbb{R}^n$ ;<sup>5</sup>
- the set

$$\{x \in \mathbb{R}^n \mid L_f V(x) = 0 \text{ and } L_{g_i}^k L_{g_i} V(x) = 0 \quad \forall i \in \{1, \dots, m\} \quad \forall k \in \mathbb{N}\}$$

is reduced to the singleton  $\{\bar{x}\}$ .

Then the equilibrium point  $\bar{x}$  is GAS in  $\mathbb{R}^n$  for the control system in closed-loop with the feedback control defined by  $u_i(x) = -L_{g_i} V(x)$ ,  $i = 1, \dots, m$ .

**Proof** Let  $F(x) = f(x) - \sum_{i=1}^m L_{g_i} V(x) g_i(x)$  be the dynamics of the closed-loop system. First of all, we note that  $F(\bar{x}) = 0$ , that is,  $\bar{x}$  is an equilibrium point for the closed-loop system. Indeed,  $V$  is smooth and reaches its minimum at  $\bar{x}$ , hence  $\nabla V(\bar{x}) = 0$ , and therefore  $L_{g_i} V(\bar{x}) = 0$  for  $i = 1, \dots, m$ . Moreover, we have  $f(\bar{x}) = 0$ . Besides, we have

$$L_F V(x) = \langle \nabla V(x), F(x) \rangle = L_f V(x) - \sum_{i=1}^m (L_{g_i} V(x))^2 \leq 0$$

and if  $L_F V(x(t)) = 0$  for every  $t \geq 0$ , then  $L_f V(x(t)) = 0$  and  $L_{g_i} V(x(t)) = 0$ ,  $i = 1, \dots, m$ . Derivating with respect to  $t$ , we infer that

$$0 = \frac{d}{dt} L_{g_i} V(x(t)) = L_f L_{g_i} V(x(t))$$

<sup>5</sup> The notation  $L_f V$  is called the Lie derivative of  $V$  along  $f$ . It is defined by  $L_f V(x) = dV(x) \cdot f(x) = \langle \nabla V(x), f(x) \rangle$ , which is the derivative of  $V$  along the direction  $f$  at  $x$ .

since  $L_{g_i} V(x(t)) = 0$ . Therefore, clearly, we get that  $L_f^k L_{g_i} V(x(t)) = 0$ , for every  $i \in \{1, \dots, m\}$  and for every  $k \in \mathbb{N}$ . By assumption, it follows that  $x(t) = \bar{x}$ , and the conclusion follows from the LaSalle principle.  $\square$

The idea of the Jurdjevic-Quinn method, that can be seen in the above proof, is very simple. The uncontrolled system has a Lyapunov function, which may be not strict. Then, to get an asymptotic stabilization result, we compute the derivative of  $V$  along solutions of the control system, we see that the control enters linearly into the resulting expression, and we design the controls so as to get the desired decrease.

**Remark 3.31** It is remarkable that the Jurdjevic-Quinn method also allows one to design globally stabilizing feedback controls, which satisfy moreover some constraints. For instance, in the framework of Theorem 3.22, let us add the requirement that  $|u_i| \leq 1, i = 1, \dots, m$ . Then, with the feedback

$$u_i = \text{sat}(-1, -L_{g_i} V(x), 1) = \begin{cases} -1 & \text{if } -L_{g_i} V(x) \leq -1 \\ -L_{g_i} V(x) & \text{if } -1 \leq -L_{g_i} V(x) \leq 1 \\ 1 & \text{if } -L_{g_i} V(x) \geq 1 \end{cases}$$

the equilibrium point  $\bar{x}$  is GAS. Indeed, the above proof is easily adapted, and the dynamics  $F(x)$  of the closed-loop system is locally Lipschitz.

The Jurdjevic-Quinn method is much used, for instance, for the stabilization of satellites along a given orbit. We next give some applications in mathematical biology (control of populations in Lotka-Volterra systems).

**Example 3.18** Consider the controlled predator-prey system

$$\dot{x} = x(1 - y + u), \quad \dot{y} = -y(1 - x).$$

and the equilibrium point  $(x = 1, y = 1)$ . Prove that the function  $V(x, y) = x - 1 - \ln(x) + y - 1 - \ln(y)$  satisfies all assumptions of Theorem 3.22, and deduce a feedback control such that the equilibrium point is GAS in  $x > 0, y > 0$ . Note that the function  $x \mapsto x - 1 - \ln(x)$  is nonnegative on  $(0, +\infty)$  and vanishes only at  $x = 1$ .

**Example 3.19 (Generalized Lotka-Volterra System)** Consider the generalized Lotka-Volterra system

$$\dot{N}_i = N_i \left( b_i + \sum_{j=1}^n a_{ij} N_j \right), \quad i = 1, \dots, n.$$

Consider the equilibrium point  $\bar{N} = (\bar{N}_1, \dots, \bar{N}_n)^T$  defined by  $b + A\bar{N} = 0$ , where  $b = (b_1, \dots, b_n)^T$  and  $A$  is the square matrix of coefficients  $a_{ij}$ . Let  $c_1, \dots, c_n$  be

some real numbers. Let  $C$  be the diagonal matrix whose coefficients are the  $c_i$ 's. We set

$$V(N) = \sum_{i=1}^n c_i \left( N_i - \bar{N}_i - \bar{N}_i \ln \frac{N_i}{\bar{N}_i} \right).$$

An easy computation shows that

$$\frac{d}{dt} V(N(t)) = \sum_{i=1}^n c_i (N_i - \bar{N}_i) (b_i + (AN)_i)$$

where  $(AN)_i$  is the  $i^{\text{th}}$  component of the vector  $AN$ . By noticing that  $b_i + (A\bar{N})_i = 0$ , we easily deduce that

$$\frac{d}{dt} V(N(t)) = \frac{1}{2} \langle N - \bar{N}, (A^\top C + CA)(N - \bar{N}) \rangle.$$

If there exists a diagonal matrix  $C$  such that  $A^\top C + CA$  be negative definite, then we infer that  $\bar{N}$  is GAS.<sup>6</sup> Assume for instance that  $A$  be skew-symmetric, and take  $C = I_n$ . Then  $V(N(t))$  is constant. We introduce some controls, for instance, in the  $n - 1$  first equations:

$$\dot{N}_i = N_i \left( b_i + \sum_{j=1}^n a_{ij} N_j + \alpha_i u_i \right), \quad i = 1, \dots, n - 1.$$

Then, we compute  $\frac{d}{dt} V(N(t)) = \sum_{i=1}^{n-1} \alpha_i (N_i(t) - \bar{N}_i) u_i(t)$ . It is then easy to design a feedback control stabilizing globally (by the LaSalle principle) the system to the equilibrium  $\bar{N}$ , under the assumption that at least one of the coefficients  $a_{in}$ ,  $i = 1, \dots, n - 1$ , be nonzero, and that  $A$  be invertible.

In the particular case  $n = 2$ , it is actually possible to ensure moreover that  $u(t) \geq 0$ , by playing with the periodicity of the trajectories (as in Example 2.11).

**Example 3.20** We consider the bilinear control system in  $\mathbb{R}^2$

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = -x_1(t) + u(t)x_1(t)$$

where the control is subject to the constraint  $|u| \leq 1$ . Let us stabilize globally this system at  $(0, 0)$ . Setting  $V(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ , we have  $\frac{d}{dt} V(x_1(t), x_2(t)) = u(t)x_1(t)x_2(t)$ . We choose the feedback  $u(x_1, x_2) = \text{sat}(-1, -x_1x_2, 1)$ . To prove

<sup>6</sup> Note that a necessary condition for  $A^\top C + CA$  to be negative definite is that the diagonal coefficients  $a_{ii}$  of  $A$  be negative. If at least one of them is zero then  $A^\top C + CA$  is not definite.



the asymptotic stability we apply the LaSalle invariance principle. If  $\dot{V} \equiv 0$  then either  $x_1 \equiv 0$  (and then by derivation we also have  $x_2 \equiv 0$ ), or  $x_2 \equiv 0$  and then by derivation we also find  $x_1 \equiv 0$ . In all cases the maximal invariant set is  $\{0, 0\}$ , which yields the conclusion.

**Example 3.21** Consider the control system

$$\dot{x}(t) = -y(t) + v(t) \cos \theta(t), \quad \dot{y}(t) = x(t) + v(t) \sin \theta(t), \quad \dot{\theta}(t) = u(t),$$

where the controls  $u$  and  $v$  are subject to the constraints  $|u| \leq 1$  and  $|v| \leq 1$ . Let us stabilize globally this system at  $(0, 0, 0)$ . Setting  $V = \frac{1}{2}(x^2 + y^2 + \theta^2)$ , we have  $\dot{V} = v(x \cos \theta + y \sin \theta) + \theta u$ . We choose the feedback controls

$$v = -\text{sat}(-1, x \cos \theta + y \sin \theta, 1) \quad \text{et} \quad u = -\text{sat}(-1, \theta, 1)$$

If  $\dot{V} = 0$  along a trajectory then  $\theta = 0$ ,  $u = 0$ ,  $0 = x \cos \theta + y \sin \theta = x$ ,  $v = 0$ , and thus also  $0 = \dot{x} = -y$ . Therefore the invariant set in the LaSalle principle is reduced to the equilibrium point. This yields the conclusion.

## References

1. Coron, J.-M.: Control and Nonlinearity. Mathematical Surveys and Monographs, vol. 136, xiv+426 pp. American Mathematical Society, Providence (2007)
2. Hale, J.K.: Ordinary Differential Equations, 2nd edn., xvi+361 pp. Robert E. Krieger Publishing Co., Inc., Huntington (1980)
3. Hartman, P.: Ordinary Differential Equations, xv+612 pp. Birkhäuser, Boston (1982)
4. Hurwitz, A.: Über die Bedingungen, unter welchen einer Gleichung nur Wurzeln mit negativen Reellen Teilen Besitzt. Math. Ann. **146**, 273–284 (1895)
5. Jurdjevic, V., Quinn, J.P.: Controllability and stability. J. Differ. Equ. **28**(3), 381–389 (1978)
6. Kailath, T.: Linear Systems. Prentice Hall, Hoboken (1980)
7. Kautsky, J., Nichols, N.K.: Robust pole assignment in linear state feedback. Int. J. Control **41**, 1129–1155 (1985)
8. Khalil, H.K.: Nonlinear Systems. Macmillan Publishing Company, New York (1992)
9. Routh, E.J.: A Treatise on the Stability of a Given State of Motion. Macmillan & Co., London (1877)
10. Sontag, E.D.: Mathematical Control Theory: Deterministic Finite-Dimensional Systems. Texts in Applied Mathematics, vol. 6, 2nd edn., xvi+531. Springer-Verlag, New York (1998)

## Part II

# Control in Infinite Dimension

In this part we introduce the control theory for infinite-dimensional systems, that is, control systems where the state  $x(t)$  evolves in an infinite-dimensional Banach space. Controlled partial differential equations enter into this category.

As we will see, the tools used to analyze such systems significantly differ from the ones used in finite dimension. The techniques are strongly use tools of functional analysis. The reader should then be quite well acquainted with such knowledge, and we refer to the textbook [1] on functional analysis.

The study of the control of nonlinear partial differential equations is beyond the scope of the present monograph, and we refer the reader to [2] for a complete survey. Throughout this part, except at the end, we will focus on linear autonomous infinite-dimensional control systems of the form  $\dot{x}(t) = Ax(t) + Bu(t)$  where  $A$  and  $B$  are operators (which can be viewed, at a first step, as infinite-dimensional matrices).

Since such systems involve partial differential equations, throughout this part the state  $x(t)$  will be rather denoted as  $y(t)$ . For PDEs settled on some domain  $\Omega$ ,  $y$  is a function of  $t$  and  $x$ , where  $t$  is the time and  $x$  the spatial variable. The control system considered throughout is

$$\dot{y} = Ay + Bu \tag{1}$$

where  $\dot{y}$  means  $\partial_t y(t, x)$  when  $y$  is a function of  $(t, x)$ .

The first step is to define the concept of a solution, which in itself is far from being obvious, in contrast to the finite-dimensional setting. In infinite dimension the exponential of  $tA$  is replaced with the concept of *semigroup*. Hence, in this part, a whole chapter is devoted to semigroup theory, with the objective of giving a rigorous sense to the solution of (1) with  $y(0) = y_0$ .

$$y(t) = S(t)y_0 + \int_0^t S(t-s)Bu(s) ds$$

where  $S(t)$  is a semigroup, generalizing  $e^{tA}$ .

There are plenty of ways to introduce the theory of controlled PDEs. Here, one of our main objectives is to provide the general framework in which the Hilbert Uniqueness Method (HUM) of J.-L. Lions can be stated.

## Reference

1. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext. Springer, New York (2011)
2. Coron, J.-M.: Control and Nonlinearity. Mathematical Surveys and Monographs, vol. 136, xiv+426 pp. American Mathematical Society, Providence (2007)

# Chapter 4

## Semigroup Theory



The objective of this chapter is to establish that, in an appropriate functional setting, there is a unique solution of the Cauchy problem

$$\dot{y}(t) = Ay(t) + f(t), \quad y(0) = y_0, \quad (4.1)$$

where  $A$  is a linear operator on a Banach space  $X$ , and where  $y(t)$  and  $f(t)$  evolve in  $X$ , which is given by

$$y(t) = S(t)y_0 + \int_0^t S(t-s)f(s)ds \quad (4.2)$$

where  $(S(t))_{t \geq 0}$  is the semigroup generated by the operator  $A$ .

In finite dimension (that is, if  $X = \mathbb{R}^n$ ), this step is easy and one has  $S(t) = e^{tA}$ , with the usual matrix exponential. In infinite dimension this step is far from being obvious and requires to define rigorously the concept of (unbounded) operator and of semigroup. The reader can keep in mind the example where the operator  $A$  is the Dirichlet-Laplacian, defined on a domain  $\Omega$  of  $\mathbb{R}^n$ .

Most results of the present chapter are borrowed from the textbooks [4] and [11] on semigroup theory and from [15], and are given without proof.

Let us recall several basic notions of functional analysis that are instrumental in what follows (see [1]).

Let  $X$  be a Banach space, endowed with a norm denoted by  $\| \cdot \|_X$ , or simply by  $\| \cdot \|$  when there is no ambiguity. Let  $Y$  be another Banach space. The norm of a bounded (i.e., continuous) linear mapping  $g : X \rightarrow Y$  is denoted as well by  $\|g\|$  and is defined as usual by

$$\|g\| = \sup_{x \in X \setminus \{0\}} \frac{\|g(x)\|_Y}{\|x\|_X}.$$

The set of bounded linear mappings from  $X$  to  $Y$  is denoted by  $L(X, Y)$ .

Let  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{C}$ . The notation  $X'$  stands for the (topological) dual of the Banach space  $X$ , that is, the vector space of all linear continuous mappings  $\ell : X \rightarrow \mathbb{K}$  (in other words,  $X' = L(X, \mathbb{K})$ ). Endowed with the norm of linear continuous forms defined above, it is a Banach space. The duality bracket is defined as usual by  $\langle \ell, x \rangle_{X', X} = \ell(x)$ , for every  $\ell \in X'$  and every  $x \in X$ .

In what follows, the word *operator* is a synonym for *mapping*. By definition, an unbounded linear operator  $A$  from  $X$  to  $Y$  is a linear mapping  $A : D(A) \rightarrow Y$  defined on a vector subspace  $D(A) \subset X$  called the *domain* of  $A$ . The operator  $A$  is said to be *bounded*<sup>1</sup> if  $D(A) = X$  and if there exists  $C > 0$  such that  $\|Ax\|_Y \leq C\|x\|_X$  for every  $x \in D(A)$ .

The operator  $A : D(A) \subset X \rightarrow Y$  is said to be *closed* whenever its *graph*

$$G(A) = \{(x, Ax) \mid x \in D(A)\}$$

is a closed subset of  $X \times Y$ . By the closed graph theorem,  $A$  is a continuous linear mapping from  $X$  to  $Y$  if and only if  $D(A) = X$  and  $G(A)$  is closed.

Let  $A : D(A) \subset X \rightarrow Y$  be a densely defined linear operator (that is,  $D(A)$  is dense in  $X$ ). The *adjoint operator*  $A^* : D(A^*) \subset Y' \rightarrow X'$  is defined as follows. We set

$$D(A^*) = \{z \in Y' \mid \exists C \geq 0 \text{ such that } \forall x \in D(A) \quad |\langle z, Ax \rangle_{Y', Y}| \leq C\|x\|_X\}.$$

Then  $D(A^*)$  is a vector subspace of  $Y'$ . For every  $z \in D(A^*)$ , we define the linear form  $\ell_z : D(A) \rightarrow \mathbb{K}$  by  $\ell_z(x) = \langle z, Ax \rangle_{Y', Y}$  for every  $x \in D(A)$ . By definition of  $D(A^*)$  we have  $|\ell_z(x)| \leq C\|x\|_X$  for every  $x \in D(A)$ . Since  $D(A)$  is dense in  $X$ , it follows that the linear form  $\ell_z$  can be extended in a unique way to a continuous linear form on  $X$ , denoted by  $\tilde{\ell}_z \in X'$  (classical continuous extension argument of uniformly continuous mappings on complete spaces). Then we set  $A^*z = \tilde{\ell}_z$ . This defines the unbounded linear operator  $A^* : D(A^*) \subset Y' \rightarrow X'$ , called the *adjoint* of  $A$ . The fundamental property of the adjoint is that

$$\langle z, Ax \rangle_{Y', Y} = \langle A^*z, x \rangle_{X', X}$$

for every  $x \in D(A)$  and every  $z \in D(A^*)$ . Note that:

- $A$  is bounded if and only if  $A^*$  is bounded, and in this case their norms are equal;
- $A^*$  is closed;
- $D(A^*)$  is not necessarily dense in  $Y'$  (even if  $A$  is closed), however if  $A$  is closed and if  $Y$  is reflexive then  $D(A^*)$  is dense in  $Y'$ .

---

<sup>1</sup> Note that the terminology is paradoxal, since an unbounded linear operator can be bounded! Actually, “unbounded operator” usually underlies that  $A$  is defined on a domain  $D(A)$  that is a proper subset of  $X$ .

When  $X$  is a (real or complex) Hilbert space, we identify  $X'$  with  $X$ . A densely defined linear operator  $A : D(A) \subset X \rightarrow X$  is said to be *self-adjoint* (resp. *skew-adjoint*) whenever  $D(A^*) = D(A)$  and  $A^* = A$  (resp.,  $A^* = -A$ ). Note that self-adjoint and skew-adjoint operators are necessarily closed.

More generally, given two Hilbert spaces  $X$  and  $Z$  such that  $Z \hookrightarrow X$ , i.e.,  $Z$  is continuously embedded in  $X$ , we have  $X' \hookrightarrow Z'$ . Now we can decide, by the Riesz theorem, to identify  $X$  with  $X'$ ; in this case, we have the triple  $Z \hookrightarrow X \hookrightarrow Z'$  (but then we cannot identify  $Z$  with  $Z'$ ). Then, for any  $x \in X \subset Z'$  and any  $z \in Z \subset X$ , we have  $\langle x, z \rangle_{Z', Z} = (x, z)_X$ , i.e., the duality bracket  $\langle \cdot, \cdot \rangle_{Z', Z}$ , standing for the application of a linear continuous mapping on  $Z$  (that is an element of  $Z'$ ) to an element of  $Z$  is identified to the scalar product on  $X$ , when both elements are in  $X$ . We say that  $X$  is the *pivot space*.

Throughout this part, we consider a Banach space  $X$ . An operator on  $X$  will mean an (unbounded) linear operator  $A : D(A) \subset X \rightarrow X$ . In practice, most of unbounded operators used to model systems of the form (4.1) are operators  $A : D(A) \subset X \rightarrow X$  that are closed and whose domain  $D(A)$  is dense in  $X$ .

When an integral is considered over the Banach space  $X$  (like in (4.2)), it is understood that it is in the usual sense of the Bochner integral (see [12, 15]).

## 4.1 Homogeneous Cauchy Problems

We first focus on the homogeneous Cauchy problem, that is (4.1) with  $f = 0$ , with the objective of giving a sense to the unique solution  $y(t) = S(t)y_0$ .

### 4.1.1 Semigroups of Linear Operators

In the sequel, the notation  $\text{id}_X$  stands for the identity mapping on  $X$ .

**Definition 4.10** A  $C_0$  semigroup of bounded linear operators on  $X$  is a one-parameter family  $(S(t))_{t \geq 0}$  of bounded linear mappings  $S(t) \in L(X)$  such that

1.  $S(0) = \text{id}_X$ ;
2.  $S(t + s) = S(t)S(s)$  for all  $(t, s) \in [0, +\infty)^2$  (semigroup property);
3.  $\lim_{t \rightarrow 0, t > 0} S(t)y = y$  for every  $y \in X$ .

The linear operator  $A : D(A) \subset X \rightarrow X$ , defined by

$$Ay = \lim_{t \rightarrow 0^+} \frac{S(t)y - y}{t}$$

on the domain  $D(A)$  that is the set of  $y \in X$  such that the above limit (computed in  $X$ , i.e., with the norm  $\|\cdot\|_X$ ) exists, is called the *infinitesimal generator* of the semigroup  $(S(t))_{t \geq 0}$ .

The semigroup is said to be a group if the second property holds true for all  $(t, s) \in \mathbb{R}^2$ .

**Proposition 4.7 ([11, Section 1.2])** *Let  $(S(t))_{t \geq 0}$  be a  $C_0$  semigroup. Then*

- *the mapping  $t \in [0, +\infty) \mapsto S(t)y$  is continuous for every  $y \in X$ ;*
- *$A$  is closed and  $D(A)$  is dense in  $X$ ;*
- *$S(t)y \in D(A)$  and  $\dot{S}(t)y = AS(t)y$  for all  $y \in D(A)$  and  $t > 0$ .*

This proposition shows that the notion of  $C_0$  semigroup is adapted to solve the homogeneous Cauchy problem.

Actually, more generally, semigroups are defined with the two first items of Definition 4.10 (see [11]). The additional third property characterizes so-called  $C_0$  semigroups (also called strongly continuous semigroups). It is a simple convergence property. If the  $C_0$  semigroup  $(S(t))_{t \geq 0}$  satisfies the stronger (uniform convergence) property  $\lim_{t \rightarrow 0, t > 0} \|S(t) - \text{id}_X\| = 0$ , then it is said to be *uniformly continuous*. The following result is however proved in [11, Theorem 1.2]:

*A linear operator  $A : D(A) \rightarrow X$  is the infinitesimal generator of a uniformly continuous semigroup if and only if  $A$  is bounded and  $D(A) = X$ . In that case, moreover,  $S(t) = e^{tA} = \sum_{n=0}^{+\infty} \frac{t^n}{n!} A^n$ .*

This result shows that, as soon as a given control process in infinite dimension involves unbounded operators (i.e.,  $D(A) \subsetneq X$ , like for instance a PDE with a Laplacian), then the underlying semigroup is not uniformly continuous. Actually as soon as an operator involves a derivation then it is unbounded. In what follows we focus on  $C_0$  semigroups.

**Proposition 4.8 ([11, Section 1.2])** *Let  $(S(t))_{t \geq 0}$  be a  $C_0$  semigroup. There exist  $M \geq 1$  and  $\omega \in \mathbb{R}$  such that*

$$\|S(t)\| \leq M e^{\omega t} \quad \forall t \geq 0. \quad (4.3)$$

*We say that  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$ . The infimum  $\omega^*$  of all possible real numbers  $\omega$  such that (4.3) is satisfied for some  $M \geq 1$  is the growth bound of the semigroup, and is given by*

$$\omega^* = \inf_{t > 0} \frac{1}{t} \ln \|S(t)\|.$$

This proposition shows that  $C_0$  semigroups have at most an exponential growth in norm. This property is similar to what happens in finite dimension.

**Definition 4.11** Let  $A : D(A) \rightarrow X$  be a linear operator on  $X$  defined on the domain  $D(A) \subset X$ . The *resolvent set*  $\rho(A)$  of  $A$  is defined as the set of complex numbers  $\lambda$  such that  $\lambda \text{id}_X - A : D(A) \rightarrow X$  is invertible and  $(\lambda \text{id}_X - A)^{-1} : X \rightarrow$

$X$  is bounded (we say it is boundedly invertible). The *resolvent* of  $A$  is defined by  $R(\lambda, A) = (\lambda \text{id}_X - A)^{-1}$ , for every  $\lambda \in \rho(A)$ .

Notice the so-called *resolvent identity*, often instrumental in some proofs:

$$R(\lambda, A) - R(\mu, A) = (\mu - \lambda)R(\lambda, A)R(\mu, A) \quad \forall (\lambda, \mu) \in \rho(A)^2.$$

**Remark 4.32** If  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  then  $\{\lambda \in \mathbb{C} \mid \text{Re}(\lambda) > \omega\} \subset \rho(A)$ , and

$$R(\lambda, A)y = (\lambda \text{id}_X - A)^{-1}y = \int_0^{+\infty} e^{-\lambda t} S(t)y \, dt \quad (4.4)$$

for every  $x \in X$  and every  $\lambda \in \mathbb{C}$  such that  $\text{Re}(\lambda) > \omega$  (Laplace transform). Indeed, integrating by parts, one has  $\lambda \int_0^{+\infty} e^{-\lambda t} S(t) \, dt = \text{id}_X + A \int_0^{+\infty} e^{-\lambda t} S(t) \, dt$  and thus  $(\lambda \text{id}_X - A) \int_0^{+\infty} e^{-\lambda t} S(t) \, dt = \text{id}_X$ .

Note that, using the expression (4.4) of  $R(\lambda, A)$  with the Laplace transform, it follows that, if  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  then  $\|R(\lambda, A)\| \leq \frac{M}{\text{Re}(\lambda) - \omega}$  for every  $\lambda \in \mathbb{C}$  such that  $\text{Re}(\lambda) > \omega$ . This can be iterated, by derivating  $R(\lambda, A)$  with respect to  $\lambda$ , using the resolvent formula and the Laplace transform, and this yields the estimates  $\|R(\lambda, A)^n\| \leq \frac{M}{(\text{Re}(\lambda) - \omega)^n}$  for every  $n \in \mathbb{N}^*$  and for every  $\lambda \in \mathbb{C}$  such that  $\text{Re}(\lambda) > \omega$ . Actually, we have the following general result.

**Theorem 4.23** ([11, Section 1.5, Theorem 5.2]) *A linear operator  $A : D(A) \subset X \rightarrow X$  is the infinitesimal generator of a  $C_0$  semigroup  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  if and only if the following conditions are satisfied:*

- $A$  is closed and  $D(A)$  is dense in  $X$ ;
- $(\omega, +\infty) \subset \rho(A)$  and  $\|R(\lambda, A)^n\| \leq \frac{M}{(\text{Re}(\lambda) - \omega)^n}$  for every  $n \in \mathbb{N}^*$  and every  $\lambda \in \mathbb{C}$  such that  $\text{Re}(\lambda) > \omega$ .

### Particular Case: Contraction Semigroups

**Definition 4.12** Let  $(S(t))_{t \geq 0}$  be a  $C_0$  semigroup. Assume that  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  for some  $M \geq 1$  and  $\omega \in \mathbb{R}$ . If  $\omega \leq 0$  and  $M = 1$  then  $(S(t))_{t \geq 0}$  is said to be a *semigroup of contractions*.

Semigroups of contractions are of great importance and cover many applications. They are mostly considered in many textbooks (such as [1, 2]) and in that case Theorem 4.23 takes the more specific following forms, which are the well-known Hille-Yosida and Lumer-Phillips theorems.

**Theorem 4.24** (Hille-Yosida Theorem, [11, Section 1.3, Theorem 3.1]) *A linear operator  $A : D(A) \subset X \rightarrow X$  is the infinitesimal generator of a  $C_0$  semigroup of contractions if and only if the following conditions are satisfied:*

- $A$  is closed and  $D(A)$  is dense in  $X$ ;
- $(0, +\infty) \subset \rho(A)$  and  $\|R(\lambda, A)\| \leq \frac{1}{\lambda}$  for every  $\lambda > 0$ .



**Remark 4.33** The latter condition can equivalently be replaced by:  $\{\lambda \in \mathbb{C} \mid \operatorname{Re}(\lambda) > 0\} \subset \rho(A)$  and  $\|R(\lambda, A)\| \leq \frac{1}{\operatorname{Re}(\lambda)}$  for every  $\lambda \in \mathbb{C}$  such that  $\operatorname{Re}(\lambda) > 0$ .

**Remark 4.34** Let  $A : D(A) \rightarrow X$  generating a  $C_0$  semigroup  $(S(t))_{t \geq 0} \in \mathcal{G}(1, \omega)$ . Then the operator  $A_\omega = A - \omega \operatorname{id}_X$  (having the same domain) is the infinitesimal generator of  $S_\omega(t) = e^{-\omega t} S(t)$  which is a semigroup of contractions (and conversely). In particular, we obtain the following corollary.

**Corollary 4.4** A linear operator  $A : D(A) \subset X \rightarrow X$  is the infinitesimal generator of a  $C_0$  semigroup  $(S(t))_{t \geq 0} \in \mathcal{G}(1, \omega)$  if and only if the following conditions are satisfied:

- $A$  is closed and  $D(A)$  is dense in  $X$ ;
- $(\omega, +\infty) \subset \rho(A)$  and  $\|R(\lambda, A)\| \leq \frac{1}{\lambda - \omega}$  for every  $\lambda > \omega$ .

Before providing the statement of the Lumer-Phillips theorem, which is another characterization of  $C_0$  semigroups of contractions, let us recall some important definitions.

For every  $y \in X$ , we define  $F(y) = \{\ell \in X' \mid \langle \ell, y \rangle_{X', X} = \|y\|_X^2 = \|\ell\|_{X'}^2\}$ . It follows from the Hahn-Banach theorem that  $F(y)$  is nonempty. In the important case where  $X$  is a Hilbert space, one has  $y \in F(y)$  (identifying  $X'$  with  $X$ ).

**Definition 4.13** The operator  $A : D(A) \subset X \rightarrow X$  is said to be:

- *dissipative* if for every  $y \in D(A)$  there exists an element  $\ell \in F(y)$  such that  $\operatorname{Re} \langle \ell, Ay \rangle_{X', X} \leq 0$ ;
- *m-dissipative* if it is dissipative and  $\operatorname{Ran}(\operatorname{id}_X - A) = (\operatorname{id}_X - A)D(A) = X$ .

If  $X$  is a Hilbert space, then  $A$  is dissipative if and only if  $\operatorname{Re} (y, Ay)_X \leq 0$  for every  $y \in D(A)$ , where  $(\cdot, \cdot)_X$  is the scalar product of  $X$ .

Other names are often used in the existing literature (see [1, 2, 12]):  $A$  is dissipative if and only if  $A$  is *accretive*, if and only if  $-A$  is *monotone*, and  $A$  is m-dissipative if and only if  $-A$  is *maximal monotone* (the letter  $m$  stands for *maximal*).

**Remark 4.35** If  $A : D(A) \rightarrow X$  is m-dissipative then  $\operatorname{Ran}(\lambda \operatorname{id}_X - A) = X$  for every  $\lambda > 0$ .

**Remark 4.36** Let  $A : D(A) \rightarrow X$  be a m-dissipative operator. If  $X$  is reflexive<sup>2</sup> then  $A$  is closed and densely defined (i.e.,  $D(A)$  is dense in  $X$ ).

**Theorem 4.25 (Lumer-Phillips Theorem, [11, Section 1.4, Theorem 4.3])** Let  $A : D(A) \subset X \rightarrow X$  be a densely defined closed linear operator. Then  $A$  is

<sup>2</sup> There is a canonical injection  $\iota : X \rightarrow X''$  (the bidual of the Banach space  $X$ ), defined by  $\langle \iota y, \ell \rangle_{X'', X'} = \langle \ell, y \rangle_{X', X}$  for every  $y \in X$  and every  $\ell \in X'$ , which is a linear isometry, so that  $X$  can be identified with a subspace of  $X''$ . The Banach space  $X$  is said to be *reflexive* whenever  $\iota(X) = X''$ ; in this case,  $X''$  is identified with  $X$  with the isomorphism  $\iota$ .

the infinitesimal generator of a  $C_0$  semigroup of contractions if and only if  $A$  is  $m$ -dissipative.

Note that it is not necessary to assume that  $A$  is closed and densely defined in this theorem if  $X$  is reflexive. A statement which is very often useful is the following one (see [11, Section 1.4, Corollary 4.4]).

**Proposition 4.9** *Let  $A : D(A) \subset X \rightarrow X$  be a densely defined operator. If  $A$  is closed and if both  $A$  and  $A^*$  are dissipative then  $A$  is the infinitesimal generator of a  $C_0$  semigroup of contractions; the converse is true if  $X$  is moreover reflexive.*

**Remark 4.37** If  $A$  is skew-adjoint then it is closed and dissipative. Actually,  $A$  is skew-adjoint if and only if  $A$  and  $-A$  are  $m$ -dissipative (see [15, Proposition 3.7.2]).

**Example 4.22** Let  $\Omega$  be an open subset of  $\mathbb{R}^n$ . The Dirichlet-Laplacian operator  $\Delta_D$  is defined on  $D(\Delta_D) = \{f \in H_0^1(\Omega) \mid \Delta f \in L^2(\Omega)\}$  by  $\Delta_D f = \Delta f$  for every  $f \in D(\Delta_D)$ , where  $\Delta$  is the usual Laplacian differential operator. Note that  $H_0^1(\Omega) \cap H^2(\Omega) \subset D(\Delta_D)$  and that, in general, the inclusion is strict. However, if  $\Omega$  is an open bounded subset with  $C^2$  boundary, or if  $\Omega$  is a convex polygon of  $\mathbb{R}^2$ , then  $D(\Delta_D) = H_0^1(\Omega) \cap H^2(\Omega)$  (see [6, Chapter 1.5], see also [15, Theorem 3.6.2 and Remark 3.6.6]).

The operator  $\Delta_D$  is self-adjoint and dissipative in  $X = L^2(\Omega)$ , hence by Proposition 4.9 it generates a semigroup of contractions, called the *heat semigroup*.

The Dirichlet-Laplacian can be as well defined in the space  $X = H^{-1}(\Omega)$  (which is the dual of  $H_0^1(\Omega)$  with respect to the pivot space  $L^2(\Omega)$ ). In that case, assuming that  $\Omega$  is an open bounded subset of  $\mathbb{R}^d$ , one has  $D(\Delta_D) = H_0^1(\Omega)$ , and  $\Delta_D : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$  is an isomorphism (see [15, Section 3.6]).

**Example 4.23** Anticipating a bit, let us study the operator underlying the wave equation. With the notations of Example 4.22, we define the operator

$$A = \begin{pmatrix} 0 & \text{id}_{H_0^1(\Omega)} \\ \Delta_D & 0 \end{pmatrix}$$

on the domain  $D(A) = D(\Delta_D) \times H_0^1(\Omega)$ , in the Hilbert space  $X = H_0^1(\Omega) \times L^2(\Omega)$ . Then it is easy to see that  $A$  is closed, densely defined, skew-adjoint and thus  $m$ -dissipative, as well as  $-A$ . Hence  $A$  and  $-A$  both generate a semigroup of contractions, therefore  $A$  generates a group of contractions. The fact that it is a group reflects the fact that the wave equation is time-reversible.

**Example 4.24** Let  $X$  be a Hilbert space, let  $(e_n)_{n \in \mathbb{N}^*}$  be a Hilbert basis of  $X$ , and let  $(\lambda_n)_{n \in \mathbb{N}^*}$  be a sequence of real numbers such that  $\sup_{n \geq 1} \lambda_n < +\infty$  (this is satisfied if  $\lambda_n \rightarrow -\infty$  as  $n \rightarrow +\infty$ ). We define the operator

$$Ay = \sum_{n=1}^{+\infty} \lambda_n (y, e_n)_X e_n \quad \text{on} \quad D(A) = \left\{ y \in X \mid \sum_{n=1}^{+\infty} \lambda_n^2 (y, e_n)_X^2 < +\infty \right\}.$$

Let us prove that  $A$  is self-adjoint and generates the  $C_0$  semigroup defined by

$$S(t)y = \sum_{n=1}^{+\infty} e^{\lambda_n t} (y, e_n)_X e_n.$$

Firstly, noting that  $X_p \subset D(A)$  for every  $p \in \mathbb{N}^*$ , with  $X_p = \{y \in X \mid \exists p \in \mathbb{N}^* \forall n \geq p (y, e_n)_X = 0\}$  and that  $\cup_{p \in \mathbb{N}^*} X_p$  is dense in  $X$ , it follows that  $D(A)$  is dense in  $X$ .

Secondly, let us prove that  $A$  is closed. Let  $(y_p)_{p \in \mathbb{N}^*}$  be a sequence of  $D(A)$  such that  $y_p \rightarrow y \in X$  and  $Ay_p \rightarrow z \in X$  as  $p$  tends to  $+\infty$ . In particular  $(y_p)_{p \in \mathbb{N}^*}$  is bounded in  $X$  and thus there exists  $M > 0$  such that  $\sum_{n=1}^N \lambda_n^2 (y_p, e_n)_X^2 \leq M$ , for every  $p \in \mathbb{N}^*$  and every  $N \in \mathbb{N}^*$ . Letting  $p$  tend to  $+\infty$  then yields that  $\sum_{n=1}^{+\infty} \lambda_n^2 (y, e_n)_X^2 \leq M$ , and thus  $y \in D(A)$ . It remains to prove that  $z = Ay$ . Since  $Ay_p = \sum_{n=1}^{+\infty} \lambda_n (y_p, e_n)_X e_n$ , and since, for every  $n \in \mathbb{N}^*$ ,  $(Ay_p, e_n)_X = \lambda_n (y_p, e_n)_X$  converges to  $\lambda_n (y, e_n)_X = (Ay, e_n)_X$ , it follows that  $Ay_p$  converges weakly to  $Ay$  in  $X$ . Then by uniqueness of the limit it follows that  $z = Ay$ .

Now, let us prove that  $\lambda \text{id}_X - A$  is boundedly invertible if and only if  $\inf_{n \geq 1} |\lambda - \lambda_n| > 0$ . If  $\inf_{n \geq 1} |\lambda - \lambda_n| > 0$ , we set

$$A_\lambda y = \sum_{n=1}^{+\infty} \frac{1}{\lambda - \lambda_n} (y, e_n)_X e_n$$

for every  $y \in X$ . Clearly,  $A_\lambda : X \rightarrow X$  is linear and bounded, and one has  $\text{Ran } A_\lambda \subset D(A)$  and  $(\lambda \text{id}_X - A)A_\lambda = A_\lambda(\lambda \text{id}_X - A) = \text{id}_X$ , hence  $\lambda \in \rho(A)$  and  $A_\lambda = (\lambda \text{id}_X - A)^{-1}$ . Conversely, if  $(\lambda \text{id}_X - A)$  is boundedly invertible, then for every  $n \in \mathbb{N}^*$  there exists  $y_n \in X$  such that  $(\lambda \text{id}_X - A)y_n = e_n$ , and the sequence  $(y_n)_{n \in \mathbb{N}^*}$  is bounded. One has  $y_n = \frac{1}{\lambda - \lambda_n} e_n$  and hence  $\inf_{n \geq 1} |\lambda - \lambda_n| > 0$ .

It follows from these arguments that if  $\inf_{n \geq 1} |\lambda - \lambda_n| > 0$  then

$$R(\lambda, A)^p y = \sum_{n=1}^{+\infty} \frac{1}{(\lambda - \lambda_n)^p} (y, e_n)_X e_n$$

and hence

$$\|R(\lambda, A)^p\| \leq \sup_{n \geq 1} \frac{1}{|\lambda - \lambda_n|^p} = \left( \sup_{n \geq 1} \frac{1}{|\lambda - \lambda_n|} \right)^p.$$

Let  $\omega \geq \sup_{n \geq 1} \lambda_n$ . Then for every  $\lambda > \omega$  one has  $\inf_{n \geq 1} |\lambda - \lambda_n| \geq \lambda - \omega$  and hence  $\sup_{n \geq 1} \frac{\lambda - \omega}{|\lambda - \lambda_n|} \leq 1$ . Then, the conclusion follows from the Hille-Yosida theorem.

This example can be applied to a number of situations. Indeed any self-adjoint operator having a compact inverse (like the Dirichlet-Laplacian) is diagonalizable and the general framework of this example can then be applied.

**Example 4.25** In order to model the 1D transport equation  $\partial_t y + \partial_x y = 0$ , with  $0 \leq x \leq 1$ , we define  $X = L^2(\Omega)$  and the operator  $Ay = -\partial_x y$  on the domain  $D(A) = \{y \in X \mid \partial_x y \in X, y(0) = 0\}$ . It is easy to prove that  $A$  is closed and densely defined, and that  $A$  and  $A^*$  are dissipative (it can also be seen that  $A - \lambda \text{id}_X$  is surjective), and hence  $A$  generates a  $C_0$  semigroup of contractions.

Recall that (as said in the introduction of the chapter), given a densely defined operator  $A : D(A) \rightarrow X$ , the adjoint  $A^* : D(A^*) \rightarrow X'$  is a closed operator, and if moreover  $A$  is closed and  $X$  is reflexive then  $D(A^*)$  is dense in  $X'$ . Concerning the semigroup properties, note that, given a  $C_0$  semigroup  $(S(t))_{t \geq 0}$  on  $X$ , the adjoint family  $(S(t)^*)_{t \geq 0}$  is a family of bounded operators on  $X'$  satisfying the semigroup property but is not necessarily a  $C_0$  semigroup.

**Proposition 4.10** ([11, Section 1.10, Corollary 10.6]) *If  $X$  is reflexive and if  $(S(t))_{t \geq 0}$  is a  $C_0$  semigroup on  $X$  with generator  $A$  then  $(S(t)^*)_{t \geq 0}$  is a  $C_0$  semigroup on  $X'$  with generator  $A^*$ .*

## 4.1.2 The Cauchy Problem

Let  $A : D(A) \subset X \rightarrow X$  be a densely defined linear operator on the Banach space  $X$ , generating a  $C_0$  semigroup  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  on  $X$ .

### 4.1.2.1 Classical Solutions

As a consequence of Proposition 4.7, we have the following result.

**Theorem 4.26** *The Cauchy problem*

$$\begin{aligned} \dot{y}(t) &= Ay(t), & t > 0, \\ y(0) &= y_0 \in D(A), \end{aligned} \tag{4.5}$$

has a unique solution  $y \in C^0([0, +\infty); D(A)) \cap C^1((0, +\infty); X)$  given by  $y(t) = S(t)y_0$  for every  $t \geq 0$ . Moreover, the differential equation  $\dot{y}(t) = Ay(t)$  makes sense in  $X$ .

This solution is often called *strong solution* in the existing literature.

**Example 4.26** Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$  having a Lipschitz boundary (to be able to define a trace). Having in mind Example 4.22, let us apply

Theorem 4.26 to the Dirichlet heat equation. The Cauchy problem

$$\partial_t y = \Delta y \text{ in } \Omega, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \in H_0^1(\Omega),$$

has a unique solution  $y \in C^0([0, +\infty); H_0^1(\Omega)) \cap C^1((0, +\infty); H^{-1}(\Omega))$ . Moreover, there exist  $M \geq 1$  and  $\omega < 0$  such that  $\|y(t, \cdot)\|_{L^2(\Omega)} \leq Me^{\omega t} \|y_0(\cdot)\|_{L^2(\Omega)}$ .

**Example 4.27** Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$  having a Lipschitz boundary. Having in mind Example 4.23, let us apply Theorem 4.26 to the Dirichlet wave equation. The Cauchy problem

$$\partial_{tt} y = \Delta y \text{ in } \Omega, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \in H_0^1(\Omega), \quad \partial_t y(0) = y_1 \in L^2(\Omega),$$

has a unique solution

$$y \in C^0([0, +\infty); H_0^1(\Omega)) \cap C^1((0, +\infty); L^2(\Omega)) \cap C^2((0, +\infty); H^{-1}(\Omega)).$$

Moreover, we have conservation of energy:  $\|\partial_t y(t)\|_{H^{-1}(\Omega)}^2 + \|y(t)\|_{L^2(\Omega)}^2 = \|y_1\|_{H^{-1}(\Omega)}^2 + \|y_0\|_{L^2(\Omega)}^2$  (by integration by parts).

If  $\partial\Omega$  is  $C^2$  and if  $y_0 \in H_0^1(\Omega) \cap H^2(\Omega)$  and  $y_1 \in H_0^1(\Omega)$  then

$$y \in C^0([0, +\infty); H^2(\Omega) \cap H_0^1(\Omega)) \cap C^1((0, +\infty); H_0^1(\Omega)) \cap C^2((0, +\infty); L^2(\Omega))$$

$$\text{and } \|\partial_t y(t)\|_{L^2(\Omega)}^2 + \|y(t)\|_{H_0^1(\Omega)}^2 = \|y_1\|_{L^2(\Omega)}^2 + \|y_0\|_{H_0^1(\Omega)}^2.$$

**Remark 4.38** Concerning the regularity of the solutions of the heat equation (Example 4.26) and of the wave equation (Example 4.27), we can actually be much more precise, by expanding the solutions as a series with the eigenfunctions of the Dirichlet-Laplacian. Indeed, let  $(\phi_j)_{j \in \mathbb{N}^*}$  be a Hilbert basis of  $L^2(\Omega)$ , consisting of eigenfunctions of the Dirichlet-Laplacian, corresponding to the (negative) eigenvalues  $(\lambda_j)_{j \in \mathbb{N}^*}$ .

For the heat equation of Example 4.26, if  $y_0 = \sum_{j=1}^{+\infty} a_j \phi_j \in L^2(\Omega)$  then  $y(t, x) = \sum_{j=1}^{+\infty} a_j e^{\lambda_j t} \phi_j(x)$  is a function of  $(t, x)$  of class  $C^\infty$  for  $t > 0$  (see [2]), and for every  $t > 0$  fixed, the function  $x \mapsto y(t, x)$  is (real) analytic on the open set  $\Omega$  (see [10]). This reflects the *smoothing effect* of the heat equation.

For the wave equation, there is no smoothing effect, but smoothness or analyticity properties can also be established for appropriate initial conditions (see [10]).

These remarks show that the regularity properties obtained by the general semigroup theory may be much improved when using the specific features of the operator under consideration (see also Remark 4.45 further).

**Remark 4.39** If  $y_0 \in X \setminus D(A)$  then in general  $y(t) = S(t)y_0 \notin D(A)$ , and hence  $y(t)$  is not a solution of (4.5) in the above sense. Actually,  $y(t)$  is solution of  $\dot{y}(t) = Ay(t)$  in a weaker sense, by replacing  $A$  with an extension on  $A$  to  $X$ , as we are going to see next.

### 4.1.2.2 Weak Solutions

The objective of this section is to define extensions of the Banach space  $X$  and of the semigroup, which will provide weaker solutions. We follow [15]. Let  $\beta \in \rho(A)$  (if  $X$  is real, consider such a real number  $\beta$ ).

**Definition 4.14** Let  $X_1$  denote the Banach space  $D(A)$ , equipped with the norm  $\|y\|_{X_1} = \|(\beta \text{id}_X - A)y\|_X$ , and let  $X_{-1}$  denote the completion of  $X$  with respect to the norm  $\|y\|_{X_{-1}} = \|(\beta \text{id}_X - A)^{-1}y\|_X = \|R(\beta, A)y\|_X$ .

Note that, by definition,  $\beta \text{id}_X - A : X_1 \rightarrow X$  and  $(\beta \text{id}_X - A)^{-1} : X \rightarrow X_1$  are surjective isometries (unitary operators). It is then easy to see that the norm  $\|\cdot\|_1$  on  $X_1$  is equivalent to the graph norm  $\|y\|_G = \|y\|_X + \|Ay\|_X$ . Therefore, from the closed graph theorem,  $(X_1, \|\cdot\|_1)$  is a Banach space, and we clearly get an equivalent norm by considering any other  $\beta' \in \rho(A)$ .

Similarly, the space  $X_{-1}$  does not depend on the specific value of  $\beta \in \rho(A)$ , in the sense that we get an equivalent norm by considering any other  $\beta' \in \rho(A)$ . Indeed, for all  $(\beta, \beta') \in \rho(A)^2$ , we have  $(\beta \text{id}_X - A)(\beta' \text{id}_X - A)^{-1} = \text{id}_X + (\beta - \beta')(\beta' \text{id}_X - A)^{-1}$ , hence

$$(\beta' \text{id}_X - A)^{-1} = (\beta \text{id}_X - A)^{-1} + (\beta - \beta')(\beta \text{id}_X - A)^{-1}(\beta' \text{id}_X - A)^{-1}$$

(resolvent identity), and moreover  $(\beta \text{id}_X - A)^{-1}$  and  $(\beta' \text{id}_X - A)^{-1}$  commute. The conclusion follows easily.

**Remark 4.40** The injections  $X_1 \hookrightarrow X \hookrightarrow X_{-1}$  are, by definition, continuous and dense. They are moreover compact as soon as  $\beta \text{id}_X - A$  has a compact inverse (i.e., as soon as  $A$  has compact resolvents).

**Example 4.28** Let  $\Omega$  be an open bounded subset of  $\mathbb{R}^n$  having a  $C^2$  boundary, and consider the Dirichlet-Laplacian  $\Delta_D$  on  $X = L^2(\Omega)$  defined on  $D(\Delta_D) = H_0^1(\Omega) \cap H^2(\Omega)$ . Then  $X_1 = D(\Delta_D) = H_0^1(\Omega) \cap H^2(\Omega)$  and, as will follow from Theorem 4.27 below,  $X_{-1} = (H_0^1(\Omega) \cap H^2(\Omega))'$ , where the dual is taken with respect to the pivot space  $X = L^2(\Omega)$ .

Let us now provide a general theorem allowing one to identify the space  $X_{-1}$ . Since  $A^*$  is closed,  $D(A^*)$  endowed with the norm  $\|z\|_{X_1} = \|(\beta \text{id}_{X'} - A^*)z\|_{X'}$  where  $\beta \in \rho(A^*) = \rho(A)$ , is a Banach space.

**Theorem 4.27** *If  $X$  is reflexive then  $X_{-1}$  is isomorphic to  $D(A^*)'$ , where the dual is taken with respect to the pivot space  $X$ .*

**Proof** We begin by recalling the following general fact: if  $E$  and  $F$  are two Banach spaces with a continuous injection  $E \hookrightarrow F$ , then we have a continuous injection  $F' \hookrightarrow E'$ . From this general fact, since  $D(A^*) \subset X'$  with a continuous injection, it follows that  $X'' \subset D(A^*)'$  (with a continuous injection). Using the canonical injection from  $X$  to  $X''$ , it follows that every element of  $X$  is (identified with) an element of  $D(A^*)'$ . Let us prove that  $\|y\|_{X_{-1}} = \|y\|_{D(A^*)'}$  for every  $y \in X$ . By

definition, we have

$$\begin{aligned}\|y\|_{X_{-1}} &= \|(\beta \operatorname{id}_X - A)^{-1}y\|_X \\ &= \sup \left\{ \langle f, (\beta \operatorname{id}_X - A)^{-1}y \rangle_{X', X} \mid f \in X', \|f\|_{X'} \leq 1 \right\} \\ &= \sup \left\{ \langle (\beta \operatorname{id}_{X'} - A^*)^{-1}f, y \rangle_{X', X} \mid f \in X', \|f\|_{X'} \leq 1 \right\}.\end{aligned}$$

Using the canonical injection of  $X$  in  $X''$  (and not yet the fact that  $X$  is reflexive),  $y$  can be considered as an element of  $X''$ , and then

$$\|y\|_{X_{-1}} = \sup \left\{ \langle y, (\beta \operatorname{id}_{X'} - A^*)^{-1}f \rangle_{X'', X'} \mid f \in X', \|f\|_{X'} \leq 1 \right\}.$$

Besides, by definition we have

$$\|y\|_{D(A^*)'} = \sup \left\{ \langle y, z \rangle_{D(A^*)', D(A^*)} \mid z \in D(A^*), \|z\|_{D(A^*)} \leq 1 \right\}.$$

In this expression we make a change of variable: for every  $z \in D(A^*)$  such that  $\|z\|_{D(A^*)} \leq 1$ , there exists  $f \in X'$  such that  $z = (\beta \operatorname{id}_{X'} - A^*)^{-1}f$ , and since  $\|z\|_{D(A^*)} = \|(\beta \operatorname{id}_{X'} - A^*)z\|_{X'} = \|f\|_{X'}$  it follows that  $\|f\|_{X'} \leq 1$ . Therefore

$$\|y\|_{D(A^*)'} = \sup \left\{ \langle y, (\beta \operatorname{id}_{X'} - A^*)^{-1}f \rangle_{D(A^*)', D(A^*)} \mid f \in X', \|f\|_{X'} \leq 1 \right\}.$$

In the above duality bracket, since  $y \in X''$  and  $(\beta \operatorname{id}_{X'} - A^*)^{-1}f \in X'$  we can as well use the duality bracket  $\langle \cdot, \cdot \rangle_{X'', X'}$ . Hence  $\|y\|_{X_{-1}} = \|y\|_{D(A^*)'}$ .

To conclude the proof, it remains to note that  $X_1$  is dense in  $X$  and that  $X \simeq X''$  is dense in  $D(A^*)'$ . It is a general fact that  $X_1$  is dense in  $X$  (see Proposition 4.7). The fact that  $X$  is dense in  $D(A^*)'$  is ensured by the reflexivity assumption: indeed, since  $X$  is reflexive it follows that  $D(A^*)$  is dense in  $X'$  (with a continuous injection), and hence  $X \simeq X''$  is dense in  $D(A^*)'$ .  $\square$

**Theorem 4.28** *The operator  $A : D(A) = X_1 \rightarrow X$  can be extended to an operator  $A_{-1} : D(A_{-1}) = X \rightarrow X_{-1}$ , and the  $C_0$  semigroup  $(S(t))_{t \geq 0}$  on  $X$  extends to a semigroup  $(S_{-1}(t))_{t \geq 0}$  on  $X_{-1}$ , generated by  $A_{-1}$ .*

**Proof** Note that the operator  $A : D(A) \rightarrow X$  is continuous when one endows  $D(A)$  with its norm, because  $\|Ay\|_X \leq \|y\|_G \leq C\|y\|_1$  as already said. By definition of the norm in  $X_{-1}$ , we have easily

$$\begin{aligned}\|Ay\|_{X_{-1}} &= \|(\beta \operatorname{id}_X - A)^{-1}Ay\|_X = \|y - \beta(\beta \operatorname{id}_X - A)^{-1}y\|_X \\ &\leq \|y\|_X + |\beta| \|(\beta \operatorname{id}_X - A)^{-1}y\|_X\end{aligned}$$

for every  $y \in D(A)$ , and since  $(\beta \operatorname{id}_X - A)^{-1}$  is bounded it follows that there exists some constant  $C_1 > 0$  such that  $\|Ay\|_{X_{-1}} \leq C_1\|y\|_X$  for every  $y \in D(A)$ .

Therefore the operator  $A$  has a unique extension  $A_{-1} : D(A_{-1}) = X \rightarrow X_{-1}$  that is continuous for the respective norms. The fact that  $D(A_{-1}) = X$  with equivalent norms follows from the equality

$$\|y\|_X = \|(\beta \text{id}_X - A)^{-1}(\beta \text{id}_X - A)y\|_X = \|(\beta \text{id}_X - A)y\|_{X_{-1}} = \|y\|_{D(A_{-1})}$$

for every  $y \in D(A)$ , and by density this holds true for every  $y \in X$ .  $\square$

**Remark 4.41** Note that, by density, if  $A$  is  $m$ -dissipative then  $A_{-1}$  is  $m$ -dissipative as well, and hence  $(S_{-1}(t))_{t \geq 0}$  is a  $C_0$  semigroup of contractions.

The following result follows from Theorem 4.26, giving an answer to the question raised in Remark 4.39.

**Corollary 4.5** *For every  $y_0 \in X$ , the Cauchy problem  $\dot{y}(t) = A_{-1}y(t)$ ,  $y(0) = y_0$  has a unique solution  $y \in C^0([0, +\infty); X) \cap C^1((0, +\infty); X_{-1})$  given by  $y(t) = S(t)y_0 = S_{-1}(t)y_0$  for every  $t \geq 0$ .*

Note that, here, the differential equation  $\dot{y}(t) = A_{-1}y(t)$  is written in  $X_{-1}$ . In particular, the derivative is computed in  $X_{-1}$ , with the norm  $\|\cdot\|_{X_{-1}}$ .

In other words, with respect to Theorem 4.26, for a given  $y_0 \in X$ ,  $y(t) = S(t)y_0$  (often called *mild solution* in the existing literature) is still a solution of  $\dot{y}(t) = Ay(t)$  (now written in  $X_{-1}$ ) provided  $A$  is replaced with its extension  $A_{-1}$ . Note that this weaker solution is a strong solution for the operator  $A_{-1}$  in the Banach space  $X_{-1}$ . For these reasons, we shall not insist on naming solutions “strong”, “mild” or “weak”. What is important is to make precise the Banach spaces in which we are working.

Note that the above concept of weak solution corresponds to solutions defined by transposition. Indeed, if  $X$  is reflexive then  $X_{-1} \simeq D(A^*)'$  (see Theorem 4.27), and hence, considering the differential equation  $\dot{y}(t) = A_{-1}y(t)$  in the space  $X_{-1}$  means that

$$\langle \dot{y}(t), \varphi \rangle_{D(A^*)', D(A^*)} = \langle A_{-1}y(t), \varphi \rangle_{D(A^*)', D(A^*)} \quad \forall \varphi \in D(A^*).$$

This concept of solution by transposition is often encountered in the existing literature (see, e.g., [3] for control issues).

**Example 4.29** Let  $\Omega \subset \mathbb{R}^n$  be a bounded open set with  $C^2$  boundary.

- The Cauchy problem  $\partial_t y = \Delta y$  in  $\Omega$ ,  $y|_{\partial\Omega} = 0$ ,  $y(0) = y_0 \in L^2(\Omega)$ , has a unique solution

$$y \in C^0([0, +\infty); L^2(\Omega)) \cap C^1((0, +\infty); (H_0^1(\Omega) \cap H^2(\Omega))').$$

Moreover, there exist  $M \geq 1$  and  $\omega \in \mathbb{R}$  (actually,  $\omega < 0$ ) such that  $\|y(t)\|_{L^2(\Omega)} \leq Me^{\omega t} \|y_0\|_{L^2(\Omega)}$  for every  $t \geq 0$ .



- Consider the Cauchy problem  $\partial_{tt}y = \Delta y$  in  $\Omega$ ,  $y|_{\partial\Omega} = 0$ ,  $y(0) = y_0$ ,  $\partial_t y(0) = y_1$ .

– If  $y_0 \in H^{-1}(\Omega)$  and  $y_1 \in (H_0^1(\Omega) \cap H^2(\Omega))'$ , then there is a unique solution

$$y \in C^0([0, +\infty); H^{-1}(\Omega)) \cap C^1((0, +\infty); (H_0^1(\Omega) \cap H^2(\Omega))').$$

– If  $y_0 \in L^2(\Omega)$  and  $y_1 \in H^{-1}(\Omega)$ , then there is a unique solution

$$y \in C^0([0, +\infty); L^2(\Omega)) \cap C^1((0, +\infty); H^{-1}(\Omega)).$$

### 4.1.3 Scale of Banach Spaces

We can generalize the previous framework and adopt a more abstract (and probably simpler, at the end) point of view.

The construction of  $X_1$  and of  $X_{-1}$  can indeed be iterated, and leads to a sequence of Banach spaces  $(X_n)_{n \in \mathbb{Z}}$  (called “tower of Sobolev spaces” in [4]). For positive integers  $n$ , the operator  $A_n : D(A_n) = D(A^{n+1}) \rightarrow D(A^n)$  is the restriction of  $A$  to  $D(A^{n+1})$ .

The construction can even be generalized in order to obtain a continuous *scale of Banach spaces*  $(X_\alpha)_{\alpha \in \mathbb{R}}$ , with the property that if  $\alpha_1 > \alpha_2$  then the canonical injection  $X_{\alpha_1} \hookrightarrow X_{\alpha_2}$  is continuous and dense, and is compact as soon as the resolvent of  $A$  is compact. We refer to [13, Sections 3.6 and 3.9] for this general construction (where these spaces are called *rigged spaces*) and for further properties (see also [4, Chapter II, Section 5, c]).

The Banach space  $X_\alpha$ , with  $\alpha$  an arbitrary real number, can be defined for instance by symbolic calculus with real powers of the resolvent of  $A$  and complex integrals (see [4, 7, 11, 13]), or by Banach spaces interpolation (see [8, 9, 14]), or by Fourier transform when it is possible (see [7]), or with a Hilbert basis when  $X$  is a Hilbert space and  $A$  is diagonalizable (see Remark 4.42 below). For instance, the construction with the fractional powers of the resolvent goes as follows, in few words (see [13, Section 3.9]), provided  $A$  generates a  $C_0$  semigroup.

Recall that  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  and that  $(\omega, +\infty) \subset \rho(A)$  (see Theorem 4.23). Given  $\beta \in \rho(A)$  with  $\operatorname{Re}(\beta) > \omega$ , given any  $\alpha > 0$  we define<sup>3</sup>

$$(\beta \operatorname{id}_X - A)^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} t^{\alpha-1} e^{-\beta t} S(t) dt \quad (4.6)$$

<sup>3</sup> This formula extrapolates the Laplace transform formula  $\frac{1}{z^\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} t^{\alpha-1} e^{-tz} dt$  valid for any  $z \in \mathbb{C}$  such that  $\operatorname{Re}(z) > 0$ .

and then we define

$$(\beta \operatorname{id}_X - A)^\alpha = ((\beta \operatorname{id}_X - A)^{-\alpha})^{-1}$$

on the domain  $\operatorname{Ran}((\beta \operatorname{id}_X - A)^{-\alpha})$ . We also define  $(\beta \operatorname{id}_X - A)^0 = \operatorname{id}$ . We denote by

$$X_\alpha = \operatorname{Ran}((\beta \operatorname{id}_X - A)^{-\alpha}) = (\beta \operatorname{id}_X - A)^{-\alpha}(X)$$

the Banach space endowed with the norm  $\|y\|_{X_\alpha} = \|(\beta \operatorname{id}_X - A)^\alpha y\|_X$ . The Banach space  $X_{-\alpha}$  is defined as the completion of  $X$  for the norm  $\|y\|_{X_{-\alpha}} = \|(\beta \operatorname{id}_X - A)^{-\alpha} y\|_X$ . Accordingly, we set  $X_0 = X$ , endowed with the norm of  $X$ . We have thus defined the scale of Banach spaces  $(X_\alpha)_{\alpha \in \mathbb{R}}$ . The construction does not depend on the specific choice of  $\beta \in \rho(A)$ .

In this general framework, the operator

$$A_\alpha : D(A_\alpha) = X_{\alpha+1} \rightarrow X_\alpha$$

(with  $\alpha \in \mathbb{R}$ ), which is either the restriction or the extension of  $A : D(A) \rightarrow X$  (with  $X_0 = X$ ) according to the sign of  $\alpha$ , generates the  $C_0$  semigroup  $(S_\alpha(t))_{t \geq 0}$ .

Hereafter, when it is clear from the context, we skip the index  $\alpha$  in  $S_\alpha(t)$  or in  $A_\alpha$ , when referring to the restriction or extension of  $S(t)$  or of  $A$  to  $X_\alpha$ .

Note that, for any  $\alpha_1, \alpha_2 \in \mathbb{R}$ ,

$$(\beta \operatorname{id}_X - A)^{\alpha_1 - \alpha_2} : X_{\alpha_1} \rightarrow X_{\alpha_2}$$

is a surjective isometry (unitary operator), where  $A$  denotes here (without the index) the appropriate restriction or extension of the operator  $A$ .

The spaces  $X_\alpha$  are interpolation spaces between the spaces  $X_n$  with integer indices. It can be noted that there exists  $C > 0$  such that, for every  $n \in \mathbb{Z}$ , for every  $\alpha \in [n, n+1]$ , we have

$$\|y\|_{X_\alpha} \leq C \|y\|_{X_n}^{n+1-\alpha} \|y\|_{X_{n+1}}^{\alpha-n} \quad \forall y \in X_{n+1}$$

(see [13, Lemma 3.9.8]). This is an interpolation inequality, as in [9]. Replacing the operator  $A$  with any real power of  $\beta \operatorname{id}_X - A$ , we infer from those inequalities the following more general interpolation inequalities (see [4, Chapter II, Section 5, Theorem 5.34]): given any real numbers  $\alpha < \beta < \gamma$ , there exists  $C(\gamma - \alpha) > 0$  (only depending on  $\gamma - \alpha$ ) such that

$$\|y\|_{X_\beta} \leq C(\gamma - \alpha) \|y\|_{X_\alpha}^{\frac{\gamma-\beta}{\gamma-\alpha}} \|y\|_{X_\gamma}^{\frac{\beta-\alpha}{\gamma-\alpha}} \quad \forall y \in X_\gamma. \quad (4.7)$$

When  $X$  is reflexive, the operator  $A^* : D(A^*) \rightarrow X'$  generates the  $C_0$  semigroup  $(S(t)^*)_{t \geq 0}$  (see Proposition 4.10). By the construction above where  $A$  is replaced

with  $A^*$ , there exists a scale of Banach spaces denoted by  $(X_\alpha^*)_{\alpha \in \mathbb{R}}$ , with  $X_0^* = X'$ . Similarly as in Theorem 4.27, we have

$$X_{-\alpha} = (X_\alpha^*)' \quad \forall \alpha \in \mathbb{R} \quad (4.8)$$

where the dual is taken with respect to the pivot space  $X$ .

There exist plenty of other constructions of (different) interpolation spaces, such as Favard or abstract Hölder spaces (see [4, Chapter II, Section 5, b]).

**Cauchy Problems** The above general construction allows one to generalize in a wide sense the concept of strong or weak solution. As a consequence of Theorem 4.26, we have the following result.

**Theorem 4.29** *The Cauchy problem*

$$\begin{aligned} \dot{y}(t) &= Ay(t), & t > 0, \\ y(0) &= y_0 \in X_\alpha, \end{aligned} \quad (4.9)$$

has a unique solution  $y \in C^0([0, +\infty); X_\alpha) \cap C^1((0, +\infty); X_{\alpha-1})$  given by  $y(t) = S(t)y_0$  for every  $t \geq 0$ .

Here, we have skipped the index  $\alpha$ , but it is understood that  $A = A_{\alpha-1}$  in (4.9); the differential equation is written in  $X_{\alpha-1}$  and the derivative is computed with respect to the norm  $\|\cdot\|_{X_{\alpha-1}}$ .

**Remark 4.42** It is interesting to perform the construction of the scale of Banach spaces in the following important case of diagonalizable operators.

Assume that  $X$  is a Hilbert space and that  $\mathcal{A} : D(\mathcal{A}) \rightarrow X$  is a self-adjoint positive operator with  $\mathcal{A}^{-1}$  compact (for instance, the negative of the Dirichlet-Laplacian on a bounded domain). There exists a Hilbert basis  $(e_j)_{j \in \mathbb{N}^*}$  of eigenvectors of  $\mathcal{A}$ , associated with (positive) eigenvalues  $(\lambda_j)_{j \in \mathbb{N}^*}$ . One has

$$\mathcal{A}y = \sum_{j=1}^{+\infty} \lambda_j (e_j, y)_X e_j \quad \text{on} \quad D(\mathcal{A}) = \left\{ y \in X \mid \sum_{j=1}^{+\infty} \lambda_j^2 (e_j, y)_X^2 < +\infty \right\}$$

and then  $\mathcal{A}^\alpha$  is defined for every  $\alpha \in \mathbb{R}$  in a spectral way by

$$\mathcal{A}^\alpha y = \sum_{j=1}^{+\infty} \lambda_j^\alpha (e_j, y)_X e_j \quad \text{on} \quad D(\mathcal{A}^\alpha) = \left\{ y \in X \mid \sum_{j=1}^{+\infty} \lambda_j^{2\alpha} (e_j, y)_X^2 < +\infty \right\}.$$

Note that we have used a calligraphic  $\mathcal{A}$ , to avoid the confusion with the operator  $A = -\mathcal{A}$  that one can consider in the differential equation  $\dot{y}(t) = Ay(t)$ .

**Example 4.30** Let us consider the negative of the Dirichlet-Laplacian  $\mathcal{A} = -\Delta_D$  defined on  $D(\mathcal{A}) = \{y \in H_0^1(\Omega) \mid \Delta y \in L^2(\Omega)\}$ , with  $X = L^2(\Omega)$ , where  $\Omega$  is a bounded open subset of  $\mathbb{R}^n$  with  $C^2$  boundary. We have  $D(\mathcal{A}) = H^2(\Omega) \cap H_0^1(\Omega)$  (see Example 4.22) and  $X_{-1} = (H_0^1(\Omega) \cap H^2(\Omega))'$  (see Example 4.28) where the dual is taken with respect to the pivot space  $X = L^2(\Omega)$ . We can define  $\mathcal{A}^{1/2} = \sqrt{-\Delta}$  in a spectral way as above.

Assuming that the boundary of  $\Omega$  is of class  $C^\infty$ , the spaces  $D(\mathcal{A}^{j/2})$ , for  $j \in \mathbb{N}$ , called *Dirichlet spaces*, are the Sobolev spaces with (the so-called) Navier boundary conditions, defined by

$$\begin{aligned} D(\mathcal{A}^{1/2}) &= \{y \in H^1(\Omega) \mid y|_{\partial\Omega} = 0\} = H_0^1(\Omega), \\ D(\mathcal{A}) &= \{y \in H^2(\Omega) \mid y|_{\partial\Omega} = 0\} = H_0^1(\Omega) \cap H^2(\Omega), \\ D(\mathcal{A}^{3/2}) &= \{y \in H^3(\Omega) \mid y|_{\partial\Omega} = (\Delta y)|_{\partial\Omega} = 0\}, \\ D(\mathcal{A}^2) &= \{y \in H^4(\Omega) \mid y|_{\partial\Omega} = (\Delta y)|_{\partial\Omega} = 0\}, \\ D(\mathcal{A}^{5/2}) &= \{y \in H^5(\Omega) \mid y|_{\partial\Omega} = (\Delta y)|_{\partial\Omega} = (\Delta^2 y)|_{\partial\Omega} = 0\}, \\ D(\mathcal{A}^3) &= \{y \in H^6(\Omega) \mid y|_{\partial\Omega} = (\Delta y)|_{\partial\Omega} = (\Delta^2 y)|_{\partial\Omega} = 0\}, \end{aligned}$$

etc; in other words,

$$D(\mathcal{A}^{j/2}) = \left\{ y \in H^j(\Omega) \mid y|_{\partial\Omega} = (\Delta y)|_{\partial\Omega} = \cdots = (\Delta^{\lfloor \frac{j-1}{2} \rfloor} y)|_{\partial\Omega} = 0 \right\}$$

for every  $j \in \mathbb{N}^*$ , where  $\lfloor \cdot \rfloor$  is the floor function. Moreover, the operator  $\mathcal{A}^{j/2} : D(\mathcal{A}^{j/2}) \rightarrow L^2(\Omega)$  is an isomorphism (see [15] for other properties). It can be noted that

$$\|\mathcal{A}^{j/2} y\|_{L^2(\Omega)} = \begin{cases} \|(-\Delta)^{j/2} y\|_{L^2(\Omega)} & \text{if } j \text{ is even,} \\ \|(-\Delta)^{j/2} y\|_{H_0^1(\Omega)} = \|(-\Delta)^{(j+1)/2} y\|_{L^2(\Omega)} & \text{if } j \text{ is odd.} \end{cases}$$

Omitting the indices, we have the scale of Hilbert spaces

$$\cdots \xrightarrow{\mathcal{A}^{1/2}} D(\mathcal{A}) \xrightarrow{\mathcal{A}^{1/2}} D(\mathcal{A}^{1/2}) \xrightarrow{\mathcal{A}^{1/2}} L^2(\Omega) \xrightarrow{\mathcal{A}^{1/2}} D(\mathcal{A}^{1/2})' \xrightarrow{\mathcal{A}^{1/2}} D(\mathcal{A})' \xrightarrow{\mathcal{A}^{1/2}} \cdots$$

with

$$D(\mathcal{A}^{1/2})' = H^{-1}(\Omega) \quad \text{and} \quad D(\mathcal{A})' = (H_0^1(\Omega) \cap H^2(\Omega))'$$

(with respect to the pivot space  $L^2(\Omega)$ ). All mappings  $\mathcal{A}^{1/2} = \sqrt{-\Delta}$ , between the corresponding spaces, are isometric isomorphisms.

As in the previous remark, we can even define  $X_\alpha = D(\mathcal{A}^\alpha)$  (and their duals) in a spectral way, for any  $\alpha \in \mathbb{R}$ , thus obtaining the scale  $(X_\alpha)_{\alpha \in \mathbb{R}}$  of Dirichlet spaces associated with the Dirichlet-Laplacian. By interpolation theory (see [9]), for every  $\alpha \in [0, 1)$ , we have  $X_\alpha = H_0^{2\alpha}(\Omega)$  if  $\alpha \neq 1/4$  and  $X_{1/4} = H_{00}^{1/2}(\Omega)$  (Lions-Magenes space).

**Remark 4.43** Using Proposition 4.10, if  $X$  is reflexive then all these results can be stated as well for the adjoint operator  $A^*$  and the adjoint  $C_0$  semigroup  $S(t)^*$ .

## 4.2 Nonhomogeneous Cauchy Problems

Let  $y_0 \in X$ . We consider the Cauchy problem

$$\dot{y}(t) = Ay(t) + f(t), \quad y(0) = y_0, \quad (4.10)$$

where  $A : D(A) \rightarrow X$  generates a  $C_0$  semigroup  $(S(t))_{t \geq 0}$  on  $X$ .

**Proposition 4.11** *If  $y_0 \in D(A)$  and  $f \in L_{\text{loc}}^p([0, +\infty), D(A))$  with  $1 \leq p \leq +\infty$ , then (4.10) has a unique solution  $y \in C^0([0, +\infty); D(A)) \cap W_{\text{loc}}^{1,p}([0, +\infty), X)$  (often referred to as strong solution of (4.10)) given by*

$$y(t) = S(t)y_0 + \int_0^t S(t-s)f(s)ds. \quad (4.11)$$

Moreover, the differential Eq. (4.10) makes sense in  $X$ .

**Proof** The function  $y$  defined by (4.11) is clearly a solution of (4.10). To prove uniqueness, let  $y_1$  and  $y_2$  be two solutions. Then  $z = y_1 - y_2$  is solution of  $\dot{z}(t) = Az(t)$ ,  $z(0) = 0$ . Since  $\frac{d}{ds}S(t-s)z(s) = -S(t-s)Az(s) + S(t-s)Az(s) = 0$  for every  $s \in [0, t]$ , it follows that  $0 = S(t)z(0) = S(0)z(t) = z(t)$ .  $\square$

Note that, if  $f \in L_{\text{loc}}^p([0, +\infty), X)$ , then (4.11) still makes sense. Note also that, using the extension of  $A$  (and of  $S(t)$ ) to  $X_{-1}$ , Proposition 4.11 implies that, if  $y_0 \in X$  and  $f \in L_{\text{loc}}^p([0, +\infty), X)$ , then (4.10) has a unique solution  $y \in C^0([0, +\infty), X) \cap W_{\text{loc}}^{1,p}([0, +\infty), X_{-1})$  given as well by the Duhamel formula (4.11) (and often referred to as *mild solution* of (4.10)), and the differential Eq. (4.10) is written in  $X_{-1}$  (see, e.g., [11, Chapter 4]). Moreover, for every  $T > 0$  there exists  $K_T > 0$  (not depending on  $y_0$  and  $f$ ) such that

$$\|y(t)\|_X \leq K_T(\|y_0\|_X + \|f\|_{L^p([0,T],X)}).$$

More generally, using the general scale of Banach spaces  $(X_\alpha)_{\alpha \in \mathbb{R}}$  mentioned previously, we have the following result (see [4] or [13, Theorem 3.8.2]).

**Proposition 4.12** *If  $f \in L^p_{\text{loc}}([0, +\infty), X_\alpha)$  for some  $\alpha \in \mathbb{R}$  and  $1 \leq p \leq +\infty$ , then for every  $y_0 \in X_\alpha$  the Cauchy problem (4.10) has a unique solution*

$$y \in C^0([0, +\infty), X_\alpha) \cap W^{1,p}_{\text{loc}}([0, +\infty), X_{\alpha-1})$$

given as well by (4.11) (called strong solution in  $X_\alpha$  in [13, 15]). Here, we have  $A = A_{\alpha-1}$  in the Eq. (4.10) which is written in  $X_{\alpha-1}$  almost everywhere, and the integral in (4.11) is done in  $X_\alpha$  (with  $S(\cdot) = S_\alpha(\cdot)$  in the integral).

Proposition 4.11 corresponds to  $\alpha = 1$ .

**Remark 4.44** The regularity stated in Proposition 4.12 is sharp in general. Given  $y_0 \in X_{\alpha+1}$ , the condition  $f \in C^0([0, +\infty); X_\alpha)$  does not ensure that  $y \in C^0([0, +\infty); X_{\alpha+1}) \cap C^1((0, +\infty); X_\alpha)$  (unless the semigroup is analytic, see [4, Chapter VI, Section 7.b, Corollary 7.17]). Indeed, for  $y_0 = 0$  and for a given  $y_1 \in X_\alpha$ , the solution of the Cauchy problem  $\dot{y}(t) = Ay(t) + S(t)y_1$ ,  $y(0) = 0$  is  $y(t) = \int_0^t S(t-s)S(s)y_1 ds = tS(t)y_1$ . Hence, if  $y_1 \in X_\alpha \setminus X_{\alpha+1}$  then  $S(t)y_1$  may not belong to  $X_{\alpha+1}$ .

It can be however noted that if  $f$  is more regular in time then the solution gains some regularity with respect to the space variable. More precisely we have the following (sometimes useful) result (see [4, 11, 15]).

**Lemma 4.5** *If  $y_0 \in X_{\alpha+1}$  and  $f \in W^{1,p}_{\text{loc}}([0, +\infty), X_\alpha)$  then (4.10) has a unique solution*

$$y \in C^0([0, +\infty); X_{\alpha+1}) \cap C^1((0, +\infty), X_\alpha) \cap W^{2,p}_{\text{loc}}([0, +\infty), X_{\alpha-1})$$

given by (4.11).

The assumption on  $f$  can even be weakened if  $X_{\alpha-1}$  is reflexive, and then it suffices to assume that  $f$  is Lipschitz continuous with values in  $X_{\alpha-1}$ .

**Remark 4.45** Let  $\Omega$  be a bounded open subset of  $\mathbb{R}^n$  with  $C^2$  boundary. Let us consider the Cauchy problem

$$\partial_t y = \Delta y + f \text{ in } \Omega, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \in L^2(\Omega),$$

with  $f \in L^2((0, +\infty) \times \Omega)$ . The general theory implies that there exists a unique solution

$$y \in C^0([0, +\infty), L^2(\Omega)) \cap H^1([0, +\infty), (H^2(\Omega) \cap H^1_0(\Omega))').$$

Actually, by using a spectral expansion as in Remark 4.38, it is easy to prove that

$$y \in L^2([0, +\infty), H^1_0(\Omega)) \cap H^1([0, +\infty), H^{-1}(\Omega)),$$

which is more precise because this set is contained in  $C^0([0, +\infty), L^2(\Omega))$ . Moreover, if  $y_0 \in H_0^1(\Omega)$ , then we have the improved regularity

$$\begin{aligned} y &\in L^2([0, +\infty), H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1([0, +\infty), L^2(\Omega)) \\ &\subset C^0([0, +\infty), H_0^1(\Omega)) \end{aligned}$$

(see also [5, Chapter 7.1] where these regularity properties are established by using Galerkin approximations for more general elliptic operators).

## References

1. Brezis, H.: Functional Analysis, Sobolev Spaces and Partial Differential Equations. Universitext, Springer, New York (2011)
2. Cazenave, T., Haraux, A.: An Introduction to Semilinear Evolution Equations. Oxford Lecture Series in Mathematics and its Applications, vol. 13. The Clarendon Press/Oxford University Press, New York (1998)
3. Coron, J.-M.: Control and Nonlinearity. Mathematical Surveys and Monographs, vol. 136, xiv+426 pp. American Mathematical Society, Providence (2007)
4. Engel, K.-J., Nagel, R.: One-Parameter Semigroups For Linear Evolution Equations. Graduate Texts in Mathematics, vol. 194. Springer-Verlag, Berlin (2000)
5. Evans, L.C.: Partial Differential Equations. Graduate Studies in Mathematics, vol. 19. American Mathematical Society, Providence (1998)
6. Grisvard, P.: Elliptic Problems in Nonsmooth Domains. Monographs and Studies in Mathematics, vol. 24. Pitman, Boston (1985)
7. Kato, T.: Perturbation Theory for Linear Operators. Classics in Mathematics, Reprint of the 1980 edition. Springer-Verlag, Berlin (1995)
8. Lions, J.-L.: Espaces d'interpolation et domaines de puissances fractionnaires d'opérateurs. J. Math. Soc. Jpn. **14**(2), 233–241 (1962)
9. Lions, J.-L., Magenes, E.: Problèmes aux limites non homogènes et applications. Travaux et Recherches Mathématiques, vol. 1, no. 17. Dunod, Paris (1968)
10. Nelson, E.: Analytic vectors. Ann. Math. **70**, 572–615 (1959)
11. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Applied Mathematical Sciences, vol. 44, viii+279 pp. Springer-Verlag, Berlin (1983)
12. Rudin, W.: Functional Analysis. International Series in Pure and Applied Mathematics, 2nd edn., xviii+424 pp. McGraw-Hill, Inc., New York (1991)
13. Staffans, O.: Well-Posed Linear Systems. Encyclopedia of Mathematics and its Applications, vol. 103, xviii+776 pp. Cambridge University Press, Cambridge (2005)
14. Triebel, H.: Interpolation Theory, Function Spaces, Differential Operators. North-Holland, Amsterdam (1978)
15. Tucsnak, M., Weiss, G.: Observation and Control for Operator Semigroups, xii+483 pp. Birkhäuser Advanced Texts. Birkhäuser Verlag, Basel (2009)

## Chapter 5

# Linear Control Systems in Banach Spaces



Throughout the chapter, we consider the linear autonomous control system

$$\begin{aligned}\dot{y}(t) &= Ay(t) + Bu(t) \\ y(0) &= y_0\end{aligned}\tag{5.1}$$

where the state  $y(t)$  belongs to a Banach space  $X$ ,  $y_0 \in X$ , the control  $u(t)$  belongs to a Banach space  $U$ ,  $A : D(A) \rightarrow X$  is the generator of a  $C_0$  semigroup  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$  on  $X$ , and  $B \in L(U, X_{-1})$ . The space  $X_{-1}$  has been defined in the previous chapter.

The control operator  $B$  is said to be *bounded* if  $B \in L(U, X)$ , and is called *unbounded* otherwise (this is the standard wording, although it is a bit ambiguous since  $B$  is bounded as an operator from  $U$  in  $X_{-1}$ ). Unbounded operators appear naturally when dealing with boundary or pointwise control systems. Other choices could be made for the control operator, and we can more generally assume that  $B \in L(U, X_{-\alpha})$  for some  $\alpha \geq 0$ . We will comment on that further, with the concept of degree of unboundedness.

A priori if  $u \in L^1_{\text{loc}}(0, +\infty; U)$  then  $Bu \in L^1_{\text{loc}}(0, +\infty; X_{-1})$ , and since  $y_0 \in X$ , it follows from the results of Sect. 4.2 that (5.1) has a unique solution  $y \in C^0([0, +\infty); X_{-1}) \cap W^{1,1}_{\text{loc}}(0, +\infty; X_{-2})$ , given by

$$y(t; y_0, u) = S(t)y_0 + L_t u\tag{5.2}$$

where

$$L_t u = \int_0^t S(t-s)Bu(s) ds.\tag{5.3}$$

Moreover, the differential equation in (5.1) is written in  $X_{-2}$ . The integral (5.3) is done in  $X_{-1}$ .



Note that, of course, if  $B \in L(U, X)$  is bounded, then the regularity moves up a rung: (5.1) has a unique solution  $y \in C^0([0, +\infty); X) \cap W_{\text{loc}}^{1,1}(0, +\infty; X_{-1})$  given as well by (5.2), and the differential equation is written in  $X_{-1}$ .

For a general (unbounded) control operator  $B \in L(U, X_{-1})$ , it is desirable to have conditions under which all solutions of (5.1) take their values in  $X$  (that is, as when the control operator is bounded). Such control operators will be said to be *admissible*. The admissibility property says that the control system is well posed in  $X$  (note that it is always well posed in  $X_{-1}$ ).

The notion of admissibility depends on the time-regularity of the inputs  $u$ . Since the admissibility property will be characterized by duality, it is necessary, here, to fix once for all the class of controls.

In what follows, and in view of the Hilbert Uniqueness Method, we will actually deal with controls  $u \in L^2([0, T], U)$  (for some  $T > 0$  arbitrary). Of course, we have  $L^2([0, T], U) \subset L^1([0, T], U)$ . Also, the duality will be easier to tackle in  $L^2$  (although easy modifications can be done in what follows to deal with  $L^p$ , at least for  $1 < p \leq +\infty$ , see [25] for exhaustive results).

Hence, from now on, the space of controls is  $L^2([0, T], U)$ .

In this chapter, after having defined admissible operators, we will introduce different concepts of controllability, and show that they are equivalent, by duality, to some observability properties. Finally, we will explain the Hilbert Uniqueness Method (in short, HUM) introduced by J.-L. Lions in [19, 20] in order to characterize the spaces where exact controllability holds true.

Most of this chapter is borrowed from [28] (see also [20, 25, 31]).

## 5.1 Admissible Control Operators

As said previously, we have a priori the inclusion  $\text{Ran}(L_T) \subset X_{-1}$ , for every  $T > 0$ , and the fact that  $L_T \in L(L^2([0, T], U), X_{-1})$ .

### 5.1.1 Definition

**Definition 5.15** A control operator  $B \in L(U, X_{-1})$  is said to be *admissible* for the  $C_0$  semigroup  $(S(t))_{t \geq 0}$  if there exists  $T > 0$  such that  $\text{Ran}(L_T) \subset X$ .

**Lemma 5.6** *The following properties are equivalent:*

- *There exists  $T > 0$  such that  $\text{Ran}(L_T) \subset X$ .*
- *For every  $T > 0$ , one has  $\text{Ran}(L_T) \subset X$ .*
- *For every  $T > 0$ , one has  $L_T \in L(L^2([0, T], U), X)$ .*
- *All solutions (5.2) of (5.1), with  $y_0 \in X$  and  $u \in L^2([0, T], U)$ , take their values in  $X$ .*

**Proof** Assume that  $\text{Ran}(L_T) \subset X$ . Let us prove that  $\text{Ran}(L_T) \subset X$  for every  $t > 0$ .

Let  $t \in (0, T)$  arbitrary. For every control  $u \in L^2([0, t], U)$ , we define the control  $\tilde{u} \in L^2([0, T], U)$  by  $\tilde{u}(s) = 0$  for  $s \in [0, T - t]$  and  $\tilde{u}(s) = u(s - T + t)$  for  $s \in [T - t, T]$ . Then, we have  $L_T \tilde{u} = \int_{T-t}^T S(T-s)Bu(s-T+t)ds = \int_0^t S(t-\tau)Bu(\tau)d\tau = L_t u$  (with  $\tau = s - T + t$ ). It follows that if  $\text{Ran}(L_T) \subset X$  then  $\text{Ran}(L_t) \subset X$ , for every  $t \in (0, T)$ .

Before proving the statement for  $t > T$ , let us note that, for every  $u \in L^2(0, 2T; U)$ , we have  $L_{2T}u = \int_0^{2T} S(2T-t)Bu(t)dt = \int_0^T S(2T-t)Bu(t)dt + \int_T^{2T} S(2T-t)Bu(t)dt = S(T) \int_0^T S(T-t)Bu(t)dt + \int_0^T S(T-s)Bu(s+T)ds = S(T)L_T u_1 + L_T u_2$ , with the controls  $u_1$  and  $u_2$  defined by  $u_1(t) = u(t)$  and  $u_2(t) = u(t+T)$  for almost every  $t \in [0, T]$ . It follows that if  $\text{Ran}(L_T) \subset X$  then  $\text{Ran}(L_{2T}) \subset X$ , and by immediate iteration, this implies as well that  $\text{Ran}(L_{kT}) \subset X$  for every  $k \in \mathbb{N}^*$ .

Now, let  $t > T$  arbitrary, and let  $k \in \mathbb{N}^*$  be such that  $kT > t$ . Since  $\text{Ran}(L_{kT}) \subset X$ , it follows from the first part of the proof that  $\text{Ran}(L_t) \subset X$ .

It remains to prove that, if  $\text{Ran}(L_T) \subset X$ , then  $L_T \in L(L^2([0, T], U), X)$ . Note first that the operator  $L_T$  is closed. Indeed, we have  $L_T u = (\beta \text{id}_X - A) \int_0^T S(T-t)(\beta \text{id}_X - A)^{-1} Bu(t)dt$ , for every  $u \in L^1([0, T], U)$ , with  $\beta \in \rho(A)$  arbitrary. By definition of  $X_{-1}$ , the operator  $(\beta \text{id}_X - A)^{-1}B$  is linear and continuous from  $U$  to  $X$ . Since  $A$  is closed (according to Proposition 4.7), it follows that  $L_T$  is closed. A priori, the graph of  $L_T$  is contained in  $X_{-1}$ . Under the assumption that  $\text{Ran}(L_T) \subset X$ , this graph is contained in  $X$ . Moreover, this graph is closed because the operator  $L_T$  is closed. Then the fact that  $L_T \in L(L^2([0, T], U), X)$  follows from the closed graph theorem.  $\square$

Note that, obviously, every bounded control operator  $B \in L(U, X)$  is admissible. The question is however nontrivial for an unbounded control operator.

Classical examples of bounded control operators are obtained when one considers a controlled PDE with an internal control, that is, a control system of the form  $\dot{y}(t) = Ay(t) + \chi_\omega u$ , with  $A : D(A) \rightarrow X = L^2(\Omega)$ , where  $\Omega$  is a domain of  $\mathbb{R}^n$  and  $\omega$  is a measurable subset of  $\Omega$ .

Unbounded control operators appear for instance when one considers a control acting along the boundary of  $\Omega$  (see further for examples).

**Remark 5.46** Note that, if  $B$  is admissible, then the solution  $y$  of (5.1) takes its values in  $X$ , and the equation  $\dot{y}(t) = Ay(t) + Bu(t)$  is written in the space  $X_{-1}$ , almost everywhere on  $[0, T]$ . The solution  $y$  has the regularity  $y \in C^0([0, T]; X) \cap H^1([0, T], X_{-1})$  whenever  $u \in L^2([0, T], U)$ . Note also that, in the term  $L_T u$ , the integration is done in  $X_{-1}$ , but the result is in  $X$  whenever  $B$  is admissible.

**Remark 5.47** As said in the introduction, we have assumed that the class of controls is  $L^2([0, T], U)$ . We can define as well the concept of admissibility within the class of controls  $L^p([0, T], U)$ , for some  $p \geq 1$ , but we obtain then a different concept, called  $p$ -admissibility (for instance if  $X$  is reflexive and  $p = 1$  then every admissible operator is necessarily bounded, see [30, Theorem 4.8]). Here, we restrict ourselves to  $p = 2$  (in particular in view of HUM), which is the most usually encountered case.

### 5.1.2 Dual Characterization of the Admissibility

Let us compute the adjoint of the operator  $L_T \in L(L^2([0, T], U), X_{-1})$ , and then derive a dual characterization of the admissibility property.

**Lemma 5.7** *Assume that  $X$  and  $U$  are reflexive. The adjoint  $L_T^*$  satisfies  $L_T^* \in L(D(A^*), L^2([0, T], U'))$ , and is given by  $(L_T^* z)(t) = B^* S(T - t)^* z$  for every  $z \in D(A^*)$  and for almost every  $t \in [0, T]$ .*

**Proof** Since  $X$  is reflexive, we have  $X_{-1} = D(A^*)'$  (see Theorem 4.27). Since  $L_T$  is a linear continuous operator from  $L^2([0, T], U)$  to  $D(A^*)'$ , the adjoint  $L_T^*$  is a linear continuous operator from  $D(A^*)''$  to  $L^2([0, T], U)'$ .

On the one part, note that  $L^2([0, T], U)' = L^2([0, T], U')$  because  $U$  is reflexive. On the other part, let us prove that  $D(A^*)$  is reflexive (and hence, that  $D(A^*)'' = D(A^*)$ ). According to the Kakutani theorem (see [3]), it suffices to prove that the closed unit ball of  $D(A^*)$  is compact for the weak topology  $\sigma(D(A^*), D(A^*)')$ . We have, for some  $\beta \in \rho(A)$ ,

$$\begin{aligned} B_{D(A^*)} &= \{z \in D(A^*) \mid \|z\|_{D(A^*)} = \|(\beta \text{id}_{X'} - A^*)z\|_{X'} \leq 1\} \\ &= \{(\beta \text{id}_{X'} - A^*)^{-1} f \mid f \in X', \|f\|_{X'} \leq 1\} \\ &= (\beta \text{id}_{X'} - A^*)^{-1} B_{X'} \end{aligned}$$

and since  $X'$  is reflexive, the closed unit ball  $B_{X'}$  is compact for the weak topology  $\sigma(X', X'')$ . Hence  $D(A^*)$  is reflexive.

Therefore,  $L_T^* \in L(D(A^*), L^2([0, T], U'))$ .

Let  $u \in L^2([0, T], U)$  and  $z \in D(A^*)$ . We have, by definition, and using the duality brackets with respect to the pivot space  $X$ ,

$$\langle L_T u, z \rangle_{D(A^*)', D(A^*)} = \langle L_T^* z, u \rangle_{L^2([0, T], U'), L^2([0, T], U)}.$$

Note that, here, we have implicitly used the fact that  $U$  is reflexive. Now, noticing that  $B \in L(U, D(A^*)')$  and hence that  $B^* \in L(D(A^*), U')$ , we have

$$\begin{aligned}
 \langle L_T u, z \rangle_{D(A^*)', D(A^*)} &= \left\langle \int_0^T S(T-t) B u(t) dt, z \right\rangle_{D(A^*)', D(A^*)} \\
 &= \int_0^T \langle S(T-t) B u(t), z \rangle_{D(A^*)', D(A^*)} dt \\
 &= \int_0^T \langle B^* S(T-t)^* z, u(t) \rangle_{U', U} dt \\
 &= \left\langle t \mapsto B^* S(T-t)^* z, t \mapsto u(t) \right\rangle_{L^2([0, T], U)', L^2([0, T], U)}
 \end{aligned}$$

and the conclusion follows.  $\square$

The following proposition, providing a dual characterization of admissibility, is an immediate consequence of Lemmas 5.6 and 5.7. Indeed in the admissible case we have  $L_T \in L(L^2([0, T], U), X)$  and equivalently  $L_T^* \in L(X', L^2([0, T], U'))$ .<sup>1</sup>

**Proposition 5.13** *Assume that  $X$  and  $U$  are reflexive. The control operator  $B \in L(U, X_{-1})$  (with  $X_{-1} \simeq D(A^*)'$ ) is admissible if and only if, for some  $T > 0$  (and equivalently, for every  $T > 0$ ) there exists  $K_T > 0$  such that*

$$\int_0^T \|B^* S(T-t)^* z\|_{U'}^2 dt \leq K_T \|z\|_{X'}^2 \quad \forall z \in D(A^*). \quad (5.4)$$

**Remark 5.48** The inequality (5.4) is called an *admissibility inequality*. Establishing such an inequality is a way to prove that a control operator is admissible. Showing such energy-like inequalities is a classical issue in PDEs (Strichartz inequalities for instance).

Once again, we stress that the admissibility property means that the control system  $\dot{y}(t) = Ay(t) + Bu(t)$  is *well posed* in  $X$ , which means here that, for a control  $u \in L^2([0, T], U)$  and an initial condition  $y(0) \in X$ , the corresponding solution  $y(t)$  stays in  $X$  indeed (and does not go in a wider space like  $X_{-1}$ ).

The concept of well-posedness of a PDE is in general a difficult issue. In finite dimension, this kind of difficulty does not exist, but in the infinite-dimensional setting, showing the admissibility of  $B$  may already be a challenge (at least, for an unbounded control operator). Examples are provided further.

**Remark 5.49** The inequality (5.4) says that the operator  $B^* \in L(D(A^*), U')$  is an admissible observation operator for the semigroup  $S^*(t)$  (see [28, 29]).

<sup>1</sup> Note that, in the case where the operator  $B \in L(U, X)$  is bounded, we always have  $L_T \in L(L^2([0, T], U), X)$  and hence  $L_T^* \in L(X', L^2([0, T], U'))$ . Moreover, if  $U$  is reflexive then  $L(X', L^2([0, T], U')) = L(X', L^2([0, T], U'))$ .

**Remark 5.50** Note that the inequality (5.4) is stated for every  $z \in D(A^*)$ . Of course, the norm  $\|z\|_{X'}^2$  has a sense for  $z$  belonging to the larger space  $X'$ , and it is natural to ask whether the inequality (5.4) can be written for every  $z \in X'$ .

The question has been studied in [29]. If  $z \in X'$  then  $S(T-t)^*z \in X'$  and then we cannot apply  $B^*$  to this element. Actually, we can replace  $B^*$  with its  $\Lambda$ -extension, defined by

$$B_{\Lambda}^* z = \lim_{\lambda \rightarrow +\infty} B^* \lambda (\lambda \text{id}_X - A^*)^{-1} z$$

also called *strong Yosida extension* in [25, Definition 5.4.1] (note that  $(\text{id}_X - A^*)^{-1}z \in D(A^*)$  for every  $z \in X'$ ) and defined on the domain  $D(B_{\Lambda}^*)$  which is the set of  $z$  for which the above limit exists. Then Proposition 5.13 still holds true with  $B^*$  replaced with  $B_{\Lambda}^*$ , with the inequality (5.4) written for every  $z \in X'$ .

Actually, in the context of Lemma 5.7,  $L_T^*$  is given by  $(L_T^* z)(t) = B_{\Lambda}^* S(T-t)^* z$ , for every  $z \in X'$  and for almost every  $t \in [0, T]$ .

### 5.1.3 Degree of Unboundedness of the Control Operator

Recall that a control operator  $B$  is said to be a *bounded* if  $B \in L(U, X)$ . When  $X \subsetneq \text{Ran}(B)$ , it is said to be *unbounded* (although  $B$  is bounded as an operator from  $U$  to  $X_{-1}$ , if we have assumed that  $B \in L(U, X_{-1})$ ).

Now more generally, using the scale of Banach spaces  $(X_{\alpha})_{\alpha \in \mathbb{R}}$  constructed in Sect. 4.1.3, we can define “unbounded control” operators such that  $B \in L(U, X_{-\alpha})$ , for some  $\alpha > 0$ . This leads to the notion of degree of unboundedness of a control operator (see [23, 25], see also [28, Chapter 4, page 138]).

**Definition 5.16** The *degree of unboundedness*  $\alpha(B) \geq 0$  of the control operator  $B$  (with respect to the spaces  $X$  and  $U$ ) is the infimum of the (not necessarily closed!) set of  $\alpha \geq 0$  such that  $B \in L(U, X_{-\alpha})$ , i.e.,  $(\beta \text{id}_X - A)^{-\alpha} B \in L(U, X)$  for some arbitrary  $\beta \in \rho(A)$  satisfying  $\text{Re}(\beta) > \omega$ , where  $(\beta \text{id}_X - A)^{-\alpha}$  is defined by (4.6). In other words,  $B \in L(U, X_{-\alpha(B)-\varepsilon})$  for every  $\varepsilon > 0$  (but not necessarily for  $\varepsilon = 0$ ).

Equivalently,<sup>2</sup>  $\alpha(B)$  is equal to the infimum of the set of  $\alpha \geq 0$  for which there exists  $C_{\alpha} > 0$  such that

$$\|(\lambda \text{id}_X - A)^{-1} B\| \leq \frac{C_{\alpha}}{\lambda^{1-\alpha}} \quad \forall \lambda > \omega.$$

If  $B$  is bounded then  $\alpha(B) = 0$ . When  $X$  is reflexive,  $\alpha(B)$  is the infimum of the set of  $\alpha \geq 0$  such that  $B^* \in L(X_{\alpha}^*, U')$  (where  $X_{\alpha}^* = D((\beta \text{id}_X - A^*)^{\alpha})$ ) thanks to

<sup>2</sup> This second definition of  $\alpha(B)$  is given in [23]. The equivalence with the first one follows from [11, Chapter II, Section 5, Propositions 5.12, 5.14 and 5.33].

(4.8)), or equivalently, of the set of  $\alpha \geq 0$  for which there exists  $C_\alpha > 0$  such that

$$\|B^*z\|_{U'} \leq C_\alpha \|(\beta \text{id}_X - A^*)^\alpha z\|_{X'} \quad \forall z \in X_\alpha^*. \quad (5.5)$$

Note that (5.5) may fail for  $\alpha = \alpha(B)$ .

Throughout the chapter, as in most of the existing literature, we consider control operators such that  $\alpha(B) \leq 1$ . This covers the most usual applications. Internal controls are bounded control operators (thus,  $\alpha(B) = 0$ ). For boundary controls, in general we have  $0 < \alpha(B) \leq 1$  (see further, and see [16, 28] for many examples).

### Lemma 5.8

1. Assume that  $X$  and  $U$  are reflexive Banach spaces. If  $B$  is admissible then there exists  $C > 0$  such that  $\|(\lambda \text{id}_X - A)^{-1}B\| \leq \frac{C}{\lambda^{1/2}}$  for every  $\lambda > \omega$  (and thus  $\alpha(B) \leq 1/2$ ).
2. Assume that  $X$  is a Hilbert space and that  $A$  is self-adjoint. If  $B \in L(U, X_{-1/2})$  then  $B$  is admissible.

**Proof** Let us prove the first point (adapted from [28, Proposition 4.4.6]).

First of all, by an obvious change of variable, we have  $L_{t+s} = S(s)L_t + L_s$ , for all  $s, t \geq 0$ . It follows that  $L_n = (S(n-1) + \cdots + S(1) + \text{id}_X)L_1$ , for every  $n \in \mathbb{N}^*$ . Since  $(S(t))_{t \geq 0} \in \mathcal{G}(M, \omega)$ , we have  $\|S(k)\| \leq Me^{k\omega}$  for every integer  $k$ , and therefore, we infer that  $\|L_n\| \leq K_n \|L_1\|$ , with  $K_n = M \frac{e^{\omega n} - 1}{e^\omega - 1}$  if  $\omega > 0$ ,  $K_n = Mn$  if  $\omega = 0$ , and  $K_n = M \frac{1}{1 - e^\omega}$  if  $\omega < 0$ . Here, the norm  $\|\cdot\|$  stands for the norm of bounded operators from  $L_{\text{loc}}^2(0, +\infty; U')$  to  $X$  (note that  $B$  is assumed to be admissible).

Besides, for  $0 < t_1 < t_2$  arbitrary, by taking controls that are equal to 0 on  $(t_1, t_2)$ , we easily prove that  $\|L_{t_1}\| \leq \|L_{t_2}\|$ .

Now, for an arbitrary  $T > 0$ , let  $n \in \mathbb{N}$  be such that  $n \leq T < n+1$ . Writing  $L_T = S(T-n)L_n + L_{T-n}$ , we get that  $\|L_T\| \leq Me^{\omega(T-n)}\|L_n\| + \|L_{T-n}\|$ .

It finally follows that  $\|L_T\| \leq Ke^{\omega T}$  if  $\omega > 0$ ,  $\|L_T\| \leq KT$  if  $\omega = 0$ , and  $\|L_T\| \leq K$  if  $\omega < 0$ , for some constant  $K > 0$  that does not depend on  $T$ . By duality, we have the same estimates on the norm of  $L_T^*$ .

For every  $z \in D(A^*)$  and every  $t \geq 0$ , we define  $(\Psi z)(t) = B^*S(t)^*z$ . Noting that, for every  $T > 0$ , we have  $(L_T^*z)(t) = (\Psi z)(T-t)$  for every  $t \in [0, T]$  (see Lemma 5.7), it follows from the above estimates (by letting  $T$  tend to  $+\infty$ ) that, for every  $\alpha > \omega$ , the function  $t \mapsto e^{-\alpha t}(\Psi z)(t)$  belongs to  $L^2(0, +\infty; U')$ .

Let us consider the Laplace transform of  $\Psi z$ . On the one part, we have, by an easy computation as in (4.4),  $\mathcal{L}(\Psi z)(s) = \int_0^{+\infty} e^{-st}(\Psi z)(t) dt = B^*(s \text{id}_X - A^*)z$  for every  $s \in \rho(A)$  such that  $\text{Re}(s) > \omega$ . On the other part, writing  $\mathcal{L}(\Psi z)(s) = \int_0^{+\infty} e^{(\alpha-s)t} e^{-\alpha t}(\Psi z)(t) dt$  and applying the Cauchy-Schwarz inequality, we get

$$\|\mathcal{L}(\Psi z)(s)\|_{U'} \leq \frac{1}{\sqrt{2(\text{Re}(s) - \alpha)}} \|t \mapsto e^{-\alpha t}(\Psi z)(t)\|_{L^2(0, +\infty; U')}.$$

The first point follows.

Let us prove the second point (adapted from [28, Proposition 5.1.3]). By definition, there exists  $C > 0$  such that

$$\|B^*z\|_{U'}^2 \leq C\|(\beta \text{id}_X - A^*)^{1/2}z\|_{X'}^2 = C(z, (\beta \text{id}_X - A)z)_X$$

for every  $z \in D(A)$  (we have used that  $A = A^*$ ), where  $(\cdot, \cdot)_X$  is the scalar product in  $X$ . Applying this inequality to  $z(t) = S^*(T - t)\psi$ , multiplying by  $e^{2\beta t}$ , and integrating over  $[0, T]$ , we get

$$\int_0^T e^{2\beta t} \|B^*S^*(T - t)\psi\|_{U'}^2 dt \leq C \int_0^T e^{2\beta t} (z(t), (\beta \text{id}_X - A)z(t))_X dt.$$

Since  $\dot{z}(t) = -Az(t)$  and  $z(T) = \psi$ , we have

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \left( e^{2\beta t} \|z(t)\|_X^2 \right) &= e^{2\beta t} \left( \beta \|z(t)\|_X^2 - (z(t), Az(t))_X \right) \\ &= e^{2\beta t} (z(t), (\beta \text{id}_X - A)z(t))_X \end{aligned}$$

and therefore we get

$$\int_0^T e^{2\beta t} \|B^*S^*(T - t)\psi\|_{U'}^2 dt \leq \frac{C}{2} \int_0^T \frac{d}{dt} \left( e^{2\beta t} \|z(t)\|_X^2 \right) dt \leq \frac{C}{2} e^{2\beta T} \|\psi\|_X^2.$$

The result follows.  $\square$

## 5.1.4 Examples

### 5.1.4.1 Dirichlet Heat Equation with Internal Control

Let  $\Omega \subset \mathbb{R}^n$  be a bounded open set with  $C^2$  boundary, and let  $\omega \subset \Omega$  be an open subset. Consider the internally controlled Dirichlet heat equation

$$\partial_t y = \Delta y + \chi_\omega u \quad \text{in } \Omega, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \in L^2(\Omega).$$

We set  $X = L^2(\Omega)$ , and we consider the operator  $A = \Delta_D : D(A) \rightarrow X$ , where  $D(A) = X_1 = H_0^1(\Omega) \cap H^2(\Omega)$ . The operator  $A$  is self-adjoint, and  $X_{-1} = D(A^*)' = (H_0^1(\Omega) \cap H^2(\Omega))'$  with respect to the pivot space  $L^2(\Omega)$ . The control operator  $B$  is defined as follows: for every  $u \in U = L^2(\omega)$ ,  $Bu \in L^2(\Omega)$  is the extension of  $u$  by 0 to the whole  $\Omega$ . It is bounded and therefore admissible (by Lemma 5.8), which means (by Proposition 5.13) that, for every  $T > 0$ , there exists

$K_T > 0$  such that

$$\int_0^T \int_{\omega} \psi(t, x)^2 dx dt \leq K_T \|\psi(0)\|_{L^2(\Omega)}^2$$

for every solution of  $\partial_t \psi = \Delta \psi$  in  $\Omega$ ,  $\psi|_{\partial\Omega} = 0$ , with  $\psi(0) \in H^2(\Omega) \cap H_0^1(\Omega)$ . The above inequality can also be established by using that  $\psi(t) = S(t)\psi(0)$  with  $S(t) \in L(X)$ .

### 5.1.4.2 Heat Equation with Dirichlet Boundary Control

Let  $\Omega \subset \mathbb{R}^n$  be a bounded open set with Lipschitz boundary. Consider the heat equation with Dirichlet boundary control

$$\partial_t y = \Delta y \text{ in } \Omega, \quad y|_{\partial\Omega} = u, \quad y(0) = y_0 \in L^2(\Omega). \quad (5.6)$$

We set  $U = L^2(\partial\Omega)$ ,  $X = H^{-1}(\Omega)$ , and we consider the self-adjoint operator  $A = \Delta_D : D(A) \rightarrow X$  where  $D(A) = X_1 = H_0^1(\Omega)$ . We have  $X_{-1} = D(A^*)' = D(A)'$  with respect to the pivot space  $H^{-1}(\Omega)$ . Note that if  $\partial\Omega$  is regular enough then  $X_{-1}$  is the dual of  $A^{-1}(H_0^1(\Omega)) = \{y \in H^3(\Omega) \mid y|_{\partial\Omega} = (\Delta y)|_{\partial\Omega} = 0\}$  with respect to the pivot space  $L^2(\Omega)$  (see Example 4.30).

Let us express the control operator  $B \in L(U, D(A^*)')$ . We preliminarily recall (see Example 4.30, with  $\mathcal{A} = -\Delta_D = -A$ ) that, for every  $f \in H^{-1}(\Omega)$ ,  $\|f\|_{H^{-1}(\Omega)} = \|(-\Delta_D)^{-1/2} f\|_{L^2(\Omega)}$  and

$$(f, g)_{H^{-1}(\Omega)} = ((-\Delta_D)^{-1/2} f, (-\Delta_D)^{-1/2} g)_{L^2(\Omega)} = -(A^{-1} f, g)_{L^2(\Omega)}$$

for all  $f \in H^{-1}(\Omega)$  and  $g \in L^2(\Omega)$ . Taking a solution  $y$  regular enough, associated with a control  $u$  (for instance, of class  $C^1$ ), since the differential equation  $\dot{y} = Ay + Bu$  is written in  $X_{-1}$ , by definition we have

$$\langle \dot{y}, \phi \rangle_{X_{-1}, X_1} = \langle Ay, \phi \rangle_{X_{-1}, X_1} + \langle Bu, \phi \rangle_{X_{-1}, X_1} \quad \forall \phi \in X_1.$$

The duality bracket is considered with respect to the pivot space  $X = H^{-1}(\Omega)$ , hence  $\langle \dot{y}, \phi \rangle_{X_{-1}, X_1} = -(\dot{y}, A^{-1}\phi)_{L^2(\Omega)}$  and  $\langle Ay, \phi \rangle_{X_{-1}, X_1} = (Ay, \phi)_{H^{-1}(\Omega)} = -(y, \phi)_{L^2(\Omega)}$ . Besides, using (5.6) and integrating by parts (Green formula), we have

$$( \partial_t y, \psi )_{L^2(\Omega)} = (y, \Delta \psi)_{L^2(\Omega)} - \left( u, \frac{\partial \psi}{\partial \nu} \right)_{L^2(\partial\Omega)} \quad \forall \psi \in H_0^1(\Omega) \cap H^2(\Omega).$$



Taking  $\psi = -A^{-1}\phi$ , we have  $\Delta\psi = -\phi$  on  $\Omega$  and, by identification,

$$\langle Bu, \phi \rangle_{X_{-1}, X_1} = \left( u, \frac{\partial}{\partial \nu} (A^{-1}\phi) \right)_{L^2(\partial\Omega)} \quad \forall u \in L^2(\partial\Omega) \quad \forall \phi \in H_0^1(\Omega),$$

which defines  $B$  by transposition: since  $\langle Bu, \phi \rangle_{X_{-1}, X_1} = (u, B^*\phi)_U$ , we infer that  $B^*\phi = \frac{\partial}{\partial \nu}|_{\partial\Omega} (A^{-1}\phi)$  for every  $\phi \in H_0^1(\Omega) = D(A^*)$ .

When  $\partial\Omega$  is  $C^2$ , we can express the operator  $B$  by using the *Dirichlet map*  $D$  (see [21] and [28, Chapters 10.6 and 10.7]), which is the linear operator defined as follows: given any  $v \in L^2(\partial\Omega)$ ,  $Dv$  is the unique solution in the sense of distributions of the Laplace equation  $\Delta(Dv) = 0$  in  $\Omega$  such that  $(Dv)|_{\partial\Omega} = v$ . Note that  $Dv \in C^\infty(\Omega)$  by hypoellipticity. Integrating by parts (Green formula), we have  $(Dv, \Delta\varphi)_{L^2(\Omega)} = (v, \frac{\partial\varphi}{\partial\nu})_{L^2(\partial\Omega)}$  for every  $\varphi \in H_0^1(\Omega) \cap H^2(\Omega)$ , and thus, setting  $\phi = \Delta_D\varphi = A\varphi$ , we have  $(Dv, \phi)_{L^2(\Omega)} = (v, \frac{\partial}{\partial\nu} (A^{-1}\phi))_{L^2(\partial\Omega)}$  for every  $\phi \in L^2(\Omega)$ . This implies that  $D : L^2(\partial\Omega) \rightarrow L^2(\Omega)$  (but  $D$  is not surjective). Then  $\langle Bu, \phi \rangle_{X_{-1}, X_1} = (Du, \phi)_{L^2(\Omega)} = -(Du, A\phi)_{H^{-1}(\Omega)} = -\langle ADu, \phi \rangle_{X_{-1}, X_1}$  and therefore

$$B = -AD$$

where we consider the extension  $A : L^2(\Omega) \rightarrow (H^2(\Omega) \cap H_0^1(\Omega))'$  (dual with respect to  $L^2(\Omega)$ ). Note that  $X_{1/2} = L^2(\Omega)$  and that  $X_{-1/2} = (H^2(\Omega) \cap H_0^1(\Omega))'$ . In particular, we have  $B \in L(U, X_{-1/2})$  (and  $B^* \in L(X_{1/2}, U)$ ), and thus  $\alpha(B) \leq 1/2$ .

Actually, by using the finer fact that  $D : L^2(\partial\Omega) \rightarrow H^{1/2}(\Omega)$  (see [21], see also [16, Chapter 3.1]), it can be proved that  $\alpha(B) = 1/4$ .

It follows from Lemma 5.8 that  $B$  is admissible; equivalently, by Proposition 5.13, for every  $T > 0$ , there exists  $K_T > 0$  such that

$$\int_0^T \left\| \frac{\partial\psi}{\partial\nu}|_{\partial\Omega}(t) \right\|_{L^2(\partial\Omega)}^2 dt \leq K_T \|\psi(0)\|_{H_0^1(\Omega)}^2$$

for every solution of  $\partial_t\psi = \Delta\psi$  in  $\Omega$ ,  $\psi|_{\partial\Omega} = 0$ , with  $\psi(0) \in H_0^1(\Omega)$ . This result says that the heat equation with boundary control (5.6) is well posed in the state space  $X = H^{-1}(\Omega)$ , with  $U = L^2(\partial\Omega)$ .

Note that (5.6) is not well posed in the state space  $L^2(\Omega)$ , meaning that, for  $y^0 \in L^2(\Omega)$  and  $u \in L^2([0, T], \partial\Omega)$ , the solution  $y$  of (5.6) may fail to belong to  $C^0([0, T]; L^2(\Omega))$  (even in dimension one). Actually, for every  $T > 0$ ,

$$\sup \left\{ \int_0^T \left\| \frac{\partial\psi}{\partial\nu}|_{\partial\Omega}(t) \right\|_{L^2(\partial\Omega)}^2 dt \mid \psi(0) \in L^2(\Omega), \|\psi(0)\|_{L^2(\Omega)} = 1 \right\} = +\infty$$

(see [20, 21]). If we take  $X = L^2(\Omega)$  then  $B^*\phi = -\frac{\partial\phi}{\partial\nu}|_{\partial\Omega}$  for every  $\phi \in H^2(\Omega) \cap H_0^1(\Omega)$  and  $B = -AD \in L(U, D(A^{3/4+\varepsilon})')$  for every  $\varepsilon > 0$  (see [16, Chapter 3.1]), but the continuity property fails for  $\varepsilon = 0$  (even in dimension one). We have then  $\alpha(B) = 3/4$  and  $B$  is not admissible by Lemma 5.8.

### 5.1.4.3 Heat Equation with Neumann Boundary Control

We replace in (5.6) the Dirichlet control with the Neumann control  $\frac{\partial y}{\partial\nu}|_{\partial\Omega} = u$ . In this case, we set  $X = L^2(\Omega)$ ,  $U = L^2(\partial\Omega)$ , we consider the operator  $A = \Delta_N$  defined on  $D(A) = \{y \in H^2(\Omega) \mid \frac{\partial y}{\partial\nu}|_{\partial\Omega} = 0\}$ , and we obtain  $B^*\phi = \phi|_{\partial\Omega}$  and  $B = -AN$ , where  $N$  is the Neumann map. We do not provide any details. Actually, we have  $B \in L(U, (D(A^{1/4+\varepsilon}))')$  for every  $\varepsilon > 0$  (see [16, Chapter 3.3]), thus  $\alpha(B) = 1/4$ , and hence  $B$  is admissible by Lemma 5.8.

### 5.1.4.4 Second-Order Equations

Another typical example is provided by second-order equations. The framework is the following (see [28]). Let  $H$  be a Hilbert space, and  $A_0 : D(A_0) \rightarrow H$  be self-adjoint and positive. Recall that  $D(A_0^{1/2})$  is the completion of  $D(A_0)$  with respect to the norm  $\|y\|_{D(A_0^{1/2})} = \sqrt{\langle A_0 y, y \rangle_H}$ , and that  $D(A_0) \subset D(A_0^{1/2}) \subset H$ , with continuous and dense embeddings (see also Remark 4.42). We set  $X = D(A_0^{1/2}) \times H$ , and we define the skew-adjoint operator  $A : D(A) \rightarrow X$  on  $D(A) = D(A_0) \times D(A_0^{1/2})$  by

$$A = \begin{pmatrix} 0 & I \\ -A_0 & 0 \end{pmatrix}.$$

Let  $U$  be a Hilbert space and let  $B_0 \in L(U, D(A_0^{1/2})')$ , where  $D(A_0^{1/2})'$  is the dual of  $D(A_0^{1/2})$  with respect to the pivot space  $H$ . The second-order control system

$$\partial_t y + A_0 y = B_0 u$$

can be written in the form

$$\frac{\partial}{\partial t} \begin{pmatrix} y \\ \partial_t y \end{pmatrix} = A \begin{pmatrix} y \\ \partial_t y \end{pmatrix} + Bu \quad \text{with} \quad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}.$$

We have  $X_{-1} = D(A^*)' = H \times D(A_0^{1/2})'$  with respect to the pivot space  $X$ , where  $D(A_0^{1/2})'$  is the dual of  $D(A_0^{1/2})$  with respect to the pivot space  $H$ . Moreover we

have  $B \in L(U, H \times D(A_0^{1/2})')$  and

$$B^* = \begin{pmatrix} 0 \\ B_0^* \end{pmatrix} \in L(D(A_0) \times D(A_0^{1/2}), U).$$

**Proposition 5.14** *The following statements are equivalent:*

- $B$  is admissible.
- There exists  $K_T > 0$  such that every solution of

$$\partial_{tt}\psi + A_0\psi = 0, \quad \psi(0) \in D(A_0), \quad \partial_t\psi(0) \in D(A_0^{1/2})$$

$$\text{satisfies } \int_0^T \|B_0^*\partial_t\psi(t)\|_{U'}^2 dt \leq K_T \left( \|\psi(0)\|_{D(A_0^{1/2})}^2 + \|\partial_t\psi(0)\|_H^2 \right).$$

- There exists  $K_T > 0$  (the same constant) such that every solution of

$$\partial_{tt}\psi + A_0\psi = 0, \quad \psi(0) \in H, \quad \partial_t\psi(0) \in D(A_0^{1/2})'$$

$$\text{satisfies } \int_0^T \|B_0^*\psi(t)\|_{U'}^2 dt \leq K_T \left( \|\psi(0)\|_H^2 + \|\partial_t\psi(0)\|_{D(A_0^{1/2})'}^2 \right).$$

**Example 5.31** Consider the wave equation with Dirichlet boundary control

$$\partial_{tt}y = \Delta y \quad \text{in } \Omega, \quad y|_{\partial\Omega} = u,$$

where  $\Omega$  is a bounded open subset of  $\mathbb{R}^n$  with  $C^2$  boundary. We set  $H = H^{-1}(\Omega)$ , and we take  $A_0 = -\Delta_D : D(A_0) = H_0^1(\Omega) \rightarrow H$  (isomorphism). We have  $D(A_0^{1/2}) = L^2(\Omega)$ , and the dual space  $D(A_0^{1/2})'$  (with respect to the pivot space  $H = H^{-1}(\Omega)$ ) is equal to the dual space  $(H^2(\Omega) \cap H_0^1(\Omega))'$  (with respect to the pivot space  $L^2(\Omega)$ ). The state space is

$$X = D(A_0^{1/2}) \times H = L^2(\Omega) \times H^{-1}(\Omega)$$

and we have

$$X_1 = D(A) = H_0^1(\Omega) \times L^2(\Omega), \quad X_{-1} = H^{-1}(\Omega) \times (H^2(\Omega) \cap H_0^1(\Omega))'.$$

The spaces  $X_\alpha$  can be characterized easily. Setting  $U = L^2(\partial\Omega)$ , the controlled wave equation is written as  $y_{tt} = -A_0y + B_0u$  in  $D(A_0^{1/2})'$ , where

$$B_0^*\phi = \frac{\partial}{\partial\nu|_{\partial\Omega}}(A_0^{-1}\phi) \quad \forall \phi \in L^2(\Omega),$$

or, equivalently,  $B_0 = A_{-1}D \in L(U, D(A_0^{1/2})')$  where  $D$  is the Dirichlet mapping. Then  $B \in L(U, X_{-1/2})$ , but this is the limit case of Lemma 5.8. It is however true that  $B$  is admissible, i.e., for every  $T > 0$  there exists  $K_T > 0$  such that

$$\int_0^T \left\| \frac{\partial \psi}{\partial \nu}(t) \right\|_{L^2(\partial\Omega)}^2 dt \leq K_T \left( \|\psi(0)\|_{H_0^1(\Omega)}^2 + \|\partial_t \psi(0)\|_{L^2(\Omega)}^2 \right)$$

for every solution of  $\partial_{tt}\psi = \Delta\psi$  in  $\Omega$ ,  $\psi|_{\partial\Omega} = 0$ . This is the *hidden regularity property* for the Dirichlet wave equation, proved in [19–21] (see also [28, Theorem 7.1.3]) by using *multipliers*. The multiplier method consists of multiplying the evolution equations by adequate functions and then using integrations by parts.

**Example 5.32** Consider the Dirichlet wave equation with internal control

$$\partial_{tt}y = \Delta y + \chi_\omega u \quad \text{in } \Omega, \quad y|_{\partial\Omega} = 0$$

on a bounded open subset  $\Omega$  with Lipschitz boundary. In this case, we set  $H = L^2(\Omega)$ , we take  $A_0 = -\Delta_D : D(A_0) = H_0^1(\Omega) \cap H^2(\Omega) \rightarrow H$  (isomorphism). We have  $D(A_0^{1/2}) = H_0^1(\Omega)$ , and the dual space  $D(A_0^{1/2})'$  (with respect to the pivot space  $H = L^2$ ) is equal to the dual space  $H^{-1}(\Omega)$ . The state space is  $X = D(A_0^{1/2}) \times H = H_0^1(\Omega) \times L^2(\Omega)$ , and we have  $X_1 = D(A) = H_0^1(\Omega) \cap H^2(\Omega) \times H_0^1(\Omega)$  and  $X_{-1} = L^2(\Omega) \times H^{-1}(\Omega)$ . Setting  $U = L^2(\omega)$ , the bounded control operator  $B_0 \in L(U, X)$  is such that, for every  $u \in U$ ,  $Bu$  is the extension of  $u$  by 0 to the whole  $\Omega$ . Its admissibility (which is obvious) means that, for every  $T > 0$ , there exists  $K_T > 0$  such that

$$\int_0^T \int_\omega \psi(t, x)^2 dx dt \leq K_T \left( \|\psi(0)\|_{L^2(\Omega)}^2 + \|\partial_t \psi(0)\|_{H^{-1}(\Omega)}^2 \right)$$

for every solution of  $\partial_{tt}\psi = \Delta\psi$  in  $\Omega$ ,  $\psi|_{\partial\Omega} = 0$ .

We refer to [16, 28] (and references cited therein) for many other examples.

## 5.2 Controllability

We consider the linear control system (5.1). We do not assume that  $B$  is admissible.

### 5.2.1 Definitions

Let us define the concept of controllability. A priori, the most natural concept is to require that for a given time  $T$ , for all  $y_0$  and  $y_1$  in  $X$ , there exists a control

$u \in L^2([0, T], U)$  and a solution of (5.1) such that  $y(0) = y_0$  satisfies  $y(T) = y_1$ . In finite dimension, a necessary and sufficient condition is the Kalman condition. In infinite dimension, new difficulties appear. Indeed, let us consider a heat equation settled on a domain  $\Omega$  of  $\mathbb{R}^n$ , with either an internal or a boundary control. Due to the smoothing effect (see Remark 4.38, see also [4]), whatever the regularity of the initial condition and of the control may be, the solution  $y(t, \cdot)$  is a smooth function (of  $x$ ) as soon as  $t > 0$ , outside of the control domain. It is therefore hopeless to try to reach a final target  $y(T) = y_1 \in L^2(\Omega)$  in general (unless  $y_1$  is smooth enough, so as to belong to the range of the heat semigroup). However, for such a parabolic equation, it makes sense to reach either  $y(T) = 0$ , or to “almost reach” any  $y_1 \in L^2(\Omega)$ . This motivates the following definitions.

**Definition 5.17** Let  $T > 0$  arbitrary. The control system (5.1) is said to be:

- *exactly controllable* in (the state space)  $X$  in time  $T$  if, for all  $(y_0, y_1) \in X^2$ , there exists  $u \in L^2([0, T], U)$  such that the solution (5.2) of (5.1) satisfies  $y(T; y_0, u) = y_1$ ;
- *approximately controllable* in  $X$  in time  $T$  if, for all  $(y_0, y_1) \in X^2$ , for every  $\varepsilon > 0$ , there exists  $u \in L^2([0, T], U)$  such that  $\|y(T; y_0, u) - y_1\|_X \leq \varepsilon$ ;
- *exactly null controllable* in  $X$  in time  $T$  if, for every  $y_0 \in X$ , there exists  $u(\cdot) \in L^2([0, T], U)$  such that  $y(T; y_0, u) = 0$ .

**Remark 5.51** Using the fact that  $y(T) = S(T)y_0 + L_T u$  (see (5.2)), with  $L_T$  defined by (5.3), we make the following remarks:

- The control system (5.1) is exactly controllable in  $X$  in time  $T$  if and only if  $\text{Ran}(L_T) = X$ . In particular, if this is true, then  $B$  must be admissible and thus  $\alpha(B) \leq 1/2$ .
- The control system (5.1) is approximately controllable in  $X$  in time  $T$  if and only if  $\text{Ran}(L_T) \cap X$  is dense in  $X$ .
- The control system (5.1) is exactly null controllable in  $X$  in time  $T$  if and only if  $\text{Ran}(S(T)) \subset \text{Ran}(L_T)$ .

**Remark 5.52** Note that, if  $\text{Ran}(L_T) = X$  for some  $T > 0$ , then  $\text{Ran}(L_t) = X$  for every  $t \geq T$ . Indeed, taking (as in the proof of Lemma 5.6) controls such that  $u = 0$  on  $(0, t - T)$ , we have  $L_t u = \int_{t-T}^t S(t-s)u(s) ds = \int_0^T S(T-\tau)u(\tau+t-T) d\tau = L_T u(\cdot + t - T)$ . This shows that if the control system (5.1) is exactly controllable in time  $T$  then it is exactly controllable in any time  $t \geq T$ .

**Remark 5.53** We speak of approximate null controllability in time  $T$  when one takes the target  $y_1 = 0$  in Definition 5.17; equivalently,  $\text{Ran}(S(T))$  is contained in the closure of  $\text{Ran}(L_T)$ . Approximate controllability and approximate null controllability (in time  $T$ ) coincide when  $S(T)^*$  is injective, i.e., when  $\text{Ran}(S(T))$  is dense in  $X$  (see [26] for finer results).

There are other notions of controllability, depending on the context and on the needs, for instance: spectral controllability, controllability to finite-dimensional subspaces, controllability to trajectories (see, e.g., [5, 28] and references therein).

### 5.2.2 Duality Controllability / Observability

As for the admissibility, we are going to provide a dual characterization of the controllability properties. To this aim, it suffices to combine Remark 5.51 with the following general lemma of functional analysis (see [3] for the first part and [18, 31] for the last part).

**Lemma 5.9** *Let  $X, Y$  and  $Z$  be Banach spaces, and let  $F \in L(X, Y)$ . Then:*

- *$\text{Ran}(F)$  is dense in  $Y$  (that is,  $F$  is “approximately surjective”) if and only if  $F^* \in L(Y', X')$  is one-to-one, that is: for every  $z \in Y'$ , if  $F^*z = 0$  then  $z = 0$ .*
- *$\text{Ran}(F) = Y$  (that is,  $F$  is surjective) if and only if  $F^* \in L(Y', X')$  is bounded below, in the sense that there exists  $C > 0$  such that  $\|F^*z\|_{X'} \geq C\|z\|_{Y'}$  for every  $z \in Y'$ .*
- *Assume that  $Y$  reflexive. Let  $F \in L(X, Z)$  and  $G \in L(Y, Z)$ . Then  $\text{Ran}(F) \subset \text{Ran}(G)$  if and only if there exists  $C > 0$  such that  $\|F^*z\|_{X'} \leq C\|G^*z\|_{Y'}$  for every  $z \in Z'$ .*

It is interesting to stress the difference with the finite-dimensional setting, in which a proper subset cannot be dense. The fact that, in infinite dimension, a proper subset may be dense explains the fact that the notion of approximate controllability is distinct from the notion of exact controllability.

Now, applying Lemma 5.9 to the operators  $L_T$  and  $S(T)$ , we get, with Remark 5.51, the following result.

**Theorem 5.30** *Assume that  $X$  and  $U$  are reflexive. Let  $T > 0$  arbitrary. The control system (5.1) is:*

- *exactly controllable in  $X$  in time  $T$  if and only if there exists  $C_T > 0$  such that*

$$\int_0^T \|B^*S^*(T-t)z\|_{U'}^2 dt \geq C_T \|z\|_{X'}^2, \quad \forall z \in D(A^*); \quad (5.7)$$

- *approximately controllable in  $X$  in time  $T$  if and only if*

$$\forall z \in D(A^*) \quad \forall t \in [0, T] \quad B^*S^*(T-t)z = 0 \quad \Rightarrow \quad z = 0; \quad (5.8)$$

- *exactly null controllable in  $X$  in time  $T$  if and only if there exists  $C_T > 0$  such that*

$$\int_0^T \|B^*S^*(T-t)z\|_{U'}^2 dt \geq C_T \|S(T)^*z\|_{X'}^2, \quad \forall z \in D(A^*). \quad (5.9)$$

**Remark 5.54** The inequalities (5.7) and (5.9) are called *observability inequalities*. As in Remark 5.50, they can be written for every  $z \in X'$ , provided that  $B^*$  be

replaced with its  $\Lambda$ -extension  $B_\Lambda^*$ . The largest constant  $C_T > 0$  such that (5.7) (or (5.9)) holds true is called the *observability constant*.

Like the admissibility inequality, an observability inequality is an energy-like inequality, but in the converse sense. Proving observability inequalities is a challenging issue in general for PDEs. We will give some examples further.

**Remark 5.55** In the context of PDEs, (5.8) corresponds to a *unique continuation* property, often established thanks to Holmgren's theorem (see [20, Section 1.8]).

**Remark 5.56** Setting  $\varphi(t) = S^*(T - t)z$ , we have

$$\dot{\varphi}(t) = -A^*\varphi(t), \quad \varphi(T) = z,$$

and the properties above can be interpreted in terms of  $\varphi(t)$  which is the adjoint vector in an infinite-dimensional version of the PMP (see Sect. 5.2.6). Usually, we rather consider  $\psi(t) = \varphi(T - t) = S^*(t)z$ , and hence we have

$$\dot{\psi}(t) = A^*\psi(t), \quad \psi(0) = z.$$

This is the adjoint equation.

In terms of this adjoint equation, the approximate controllability property is equivalent to the unique continuation property:  $B^*\psi(t) = 0$  for every  $t \in [0, T]$  implies  $\psi(\cdot) = 0$ . The exact controllability is equivalent to the observability inequality

$$\int_0^T \|B^*\psi(t)\|_{U'}^2 dt \geq C_T \|\psi(0)\|_X^2,$$

for every solution of the adjoint equation, and the exact null controllability is equivalent to the observability inequality

$$\int_0^T \|B^*\psi(t)\|_{U'}^2 dt \geq C_T \|\psi(T)\|_{X'}^2.$$

**Remark 5.57** As announced in Remarks 1.7 and 1.9 (in Chap. 1, Sect. 1.1.2), the observability inequality (5.7) is the infinite-dimensional version of the observability inequality (1.5) obtained in the finite-dimensional setting.

Note that, in finite dimension, the properties (5.7) and (5.8) are equivalent, whereas, in the infinite-dimensional setting, there is a deep difference, due to the fact that a proper subset of an infinite-dimensional space may be dense.

**Gramian Operator** As in Remark 1.7 and in Theorem 1.3, we can similarly define the Gramian operator in the present context of Banach spaces. Its general definition is the following. Assume that  $X$  and  $U$  are reflexive. Recall that, in general, we have  $L_T \in L(L^2([0, T], U), D(A^*)')$  and  $L_T^* \in L(D(A^*), L^2([0, T], U'))$  (see the proof of Lemma 5.7). Identifying  $U \simeq U'$  and  $L^2([0, T], U) \simeq L^2([0, T], U')$ , we

define the Gramian operator

$$G_T = L_T L_T^* = \int_0^T S(T-t) B B^* S(T-t)^* dt \in L(D(A^*), D(A^*)') \quad (5.10)$$

where, in the formula above,  $BB^*$  is to be understood as  $BJB^*$  where  $J : U' \rightarrow U$  is the canonical isomorphism.

If the control operator  $B$  is admissible then  $L_T \in L(L^2([0, T], U), X)$  and  $L_T^* \in L(X', L^2([0, T], U'))$ , and therefore  $G_T \in L(X', X)$ . The expression of  $G_T$  is still given by (5.10) when applied to some element  $z \in D(A^*)$ . Using Remark 5.50, it can be noted that the expression of  $G_T$  on the whole space  $X'$  is given by

$$G_T = \int_0^T S(T-t) B_\Lambda B_\Lambda^* S(T-t)^* dt.$$

If the control system (5.1) is exactly controllable in time  $T$ , then, using (5.7) and Remark 5.54, it follows that  $\langle G_T z, z \rangle_{D(A^*)', D(A^*)} \geq C_T \|z\|_{X'}^2$  for every  $z \in D(A^*)$ ; the converse inequality is satisfied if  $B$  is admissible. In other words, we have the following lemma.

**Lemma 5.10** *The control operator  $B$  is admissible and the control system (5.1) is exactly controllable in time  $T$  if and only if  $G_T : X' \rightarrow X$  is an isomorphism satisfying  $C_T \|z\|_{X'}^2 \leq \langle G_T z, z \rangle_{X', X} \leq K_T \|z\|_{X'}^2$  for every  $z \in X'$ .*

### 5.2.3 Hilbert Uniqueness Method (HUM)

The Hilbert Uniqueness Method (in short, HUM; see [19, 20]) is based on Lemma 5.10 by noticing that, in the context of this lemma, the norm  $\|\cdot\|_{X'}$  is equivalent to the norm given by  $(\langle G_T z, z \rangle_{X', X})^{1/2}$ . This gives a characterization of the state space  $X$  in which we have exact controllability.

HUM can then be stated as follows. Let  $Y$  be a reflexive Banach space, let  $A : D(A) \rightarrow Y$  be an operator generating a  $C_0$  semigroup and let  $(Y_\alpha)_{\alpha \in \mathbb{R}}$  be the associated scale of Banach spaces. Let  $U$  be a fixed reflexive Banach space and let  $B \in L(U, Y_{-\alpha})$  be a control operator. Let  $Z$  be the completion of  $D(A^*)$  for the norm  $(\langle G_T z, z \rangle_{D(A^*)', D(A^*)})^{1/2}$ , and let  $X$  be a Banach space such that  $X' = Z$ . Then  $X$  is the Banach space for which Lemma 5.10 is satisfied, i.e., for which  $B$  is admissible and the control system (5.1) is exactly controllable in time  $T$  in the state space  $X$ .

HUM may as well be restated in the other way round: the (reflexive Banach) state space  $X$  is fixed and one wants to characterize the control Banach space  $U$  for which admissibility and exact controllability are satisfied.



**HUM Functional** In the framework of Lemma 5.10, the so-called HUM functional  $J$ , defined by

$$J(z) = \frac{1}{2} \langle G_T z, z \rangle_{X', X} + \langle z, S(T)y_0 - y_1 \rangle_{X', X} \quad \forall z \in X'$$

is smooth and coercive in  $X'$ , hence  $J$  has a unique minimizer  $\bar{z}$ , satisfying

$$0 = \nabla J(\bar{z}) = G_T \bar{z} + S(T)y_0 - y_1.$$

Defining the so-called HUM control by  $\bar{u}(t) = B^*S(T-t)^*\bar{z} = (L_T^*\bar{z})(t)$ , the above equality says that  $S(T)y_0 + L_T\bar{u} = y_1$ , i.e.,  $y(T; y_0, \bar{u}) = y_1$ . In other words, the control  $\bar{u}$  steers the control system (5.1) from  $y_0$  to  $y_1$  in time  $T$ . Actually, the control  $\bar{u}$  is even the minimal  $L^2$  norm control realizing this controllability property (see [19, 20]): this can also be seen by observing that, when wanting to solve the overdetermined equation  $L_T u = y_1 - S(T)y_0$ , the control of minimal  $L^2$  norm is given by  $u = L_T^\#(y_1 - S(T)y_0)$  where  $L_T^\#$  is the pseudo-inverse of  $L_T$  (this is indeed a well known property of the pseudo-inverse); since  $L_T^\# = L_T^*(L_T L_T^*)^{-1} = L_T^* G_T^{-1}$ , the claim follows.

**HUM for Exact Null Controllability** When wanting to realize an exact null controllability result for a control system that is not exactly controllable (like the heat equation), of course the conclusion of Lemma 5.10 does not hold. In terms of the Gramian operator, the observability inequality (5.9) is written as  $\langle G_T z, z \rangle_{D(A^*)', D(A^*)} \geq C_T \|S(T)^* z\|_{\mathcal{X}'}^2$  for every  $z \in D(A^*)$ .

We can however still write the HUM functional as above (with an additional care) and determine the minimal  $L^2$  norm control steering the control system (5.1) to 0 in time  $T$ . The HUM functional  $J$  is defined as above, for every  $z \in D(A^*)$ , with the duality bracket  $\langle G_T z, z \rangle_{D(A^*)', D(A^*)}$  for the first term. The functional  $J$  is however not coercive in  $X'$ . To recover such a property, we define the Banach space  $\mathcal{X}$  as the completion of  $D(A^*)$  for the norm  $(\langle G_T z, z \rangle_{D(A^*)', D(A^*)})^{1/2}$ . Note that the space  $\mathcal{X}$  is in general much larger than  $D(A^*)$  and may even fail to be a space of distributions (see [20]). Anyway, there is a unique minimizer  $\bar{z} \in \mathcal{X}$  of  $J$ , satisfying therefore  $0 = \nabla J(\bar{z}) = G_T \bar{z} + S(T)y_0$ , and then the HUM control  $\bar{u}(t) = B^*S(T-t)^*\bar{z} = (L_T^*\bar{z})(t)$  steers the control system (5.1) to 0 in time  $T$ , and is the control of minimal  $L^2$  norm doing so.

This approach provides a generalization of Theorem 1.3 (in Sect. 1.1.2) to infinite-dimensional autonomous linear control systems.

## 5.2.4 Example: The Wave Equation

The typical (and historical) example of application of HUM is the wave equation, either with an internal control or with a (Dirichlet or Neumann) boundary control.

In 1D, the analysis is easy thanks to Fourier series, as elaborated below. The multi-D case is much more complicated and can be treated thanks to microlocal analysis (see comments further).

### 5.2.4.1 1D Wave Equation with Dirichlet Boundary Control

Let  $T > 0$  and  $L > 0$  be fixed. We consider the 1D wave equation with Dirichlet boundary control at the right-boundary:

$$\begin{aligned} \partial_{tt}y &= \partial_{xx}y, & t \in (0, T), x \in (0, L), \\ y(t, 0) &= 0, \quad y(t, L) = u(t), & t \in (0, T), \\ y(0, x) &= y_0(x), \quad \partial_t y(0, x) = y_1(x), & x \in (0, L), \end{aligned} \quad (5.11)$$

where the state at time  $t \in [0, T]$  is  $(y(t, \cdot), \partial_t y(t, \cdot))$  and the control is  $u(t) \in \mathbb{R}$ . Let us establish that this equation is exactly controllable in time  $T$  in the space  $L^2(0, L) \times H^{-1}(0, L)$  with controls  $u \in L^2(0, T)$  if and only if  $T \geq 2L$ .

By Theorem 5.30 (see also Example 5.31), this is equivalent to establishing the following observability inequality: there exists  $C_T > 0$  such that any solution of

$$\partial_{tt}\psi = \partial_{xx}\psi, \quad \psi(t, 0) = \psi(t, L) = 0, \quad (5.12)$$

such that  $(\psi(0), \partial_t \psi(0)) \in H_0^1(0, L) \times L^2(0, L)$  satisfies

$$\int_0^T |\partial_x \psi(t, L)|^2 dt \geq C_T \left( \|\psi(0)\|_{H_0^1(0, L)}^2 + \|\partial_t \psi(0)\|_{L^2(0, L)}^2 \right) \quad (5.13)$$

Given any  $T \geq 2L$ , let us establish (5.13) by using spectral expansions (Fourier series). We expand the solutions of (5.12) as

$$\psi(t, x) = \sum_{k=1}^{\infty} \frac{L}{k\pi} \left( a_k \cos \frac{k\pi t}{L} + b_k \sin \frac{k\pi t}{L} \right) \sin \frac{k\pi x}{L}$$

with  $(a_k)_{k \in \mathbb{N}^*} \in \ell^2(\mathbb{R})$  and  $(b_k)_{k \in \mathbb{N}^*} \in \ell^2(\mathbb{R})$ , so that  $\psi(0, x) = \sum_{k=1}^{\infty} \frac{L}{k\pi} a_k \sin \frac{k\pi x}{L}$  and  $\partial_t \psi(0, x) = \sum_{k=1}^{\infty} b_k \sin \frac{k\pi x}{L}$  and thus

$$\begin{aligned} \|\psi(0)\|_{H_0^1}^2 + \|\partial_t \psi(0)\|_{L^2}^2 &= \int_0^L \left( |\partial_x \psi(0, x)|^2 + |\partial_t \psi(0, x)|^2 \right) dx \\ &= \frac{L}{2} \sum_{k=1}^{\infty} (a_k^2 + b_k^2). \end{aligned}$$

Then

$$\begin{aligned}
 \int_0^T |\partial_x \psi(t, L)|^2 dt &\geq \int_0^{2L} |\partial_x \psi(t, L)|^2 dt \\
 &= \int_0^{2L} \left| \sum_{k=1}^{\infty} (-1)^k \left( a_k \cos \frac{k\pi t}{L} + b_k \sin \frac{k\pi t}{L} \right) \right|^2 dt \\
 &= \sum_{j,k=1}^{\infty} (-1)^{j+k} \int_0^{2L} \left( a_j \cos \left( \frac{j\pi t}{L} \right) + b_j \sin \left( \frac{j\pi t}{L} \right) \right) \\
 &\quad \times \left( a_k \cos \left( \frac{k\pi t}{L} \right) + b_k \sin \left( \frac{k\pi t}{L} \right) \right) dt \\
 &= L \sum_{k=1}^{\infty} (a_k^2 + b_k^2) = 2 \left( \|\psi(0)\|_{H_0^1(0,L)}^2 + \|\partial_t \psi(0)\|_{L^2(0,L)}^2 \right),
 \end{aligned}$$

and (5.13) is proved.

Using similar Fourier series expansions, we see that admissibility property (of the Dirichlet control operator) is satisfied for any  $T > 0$ : by Proposition 5.14 (see also Example 5.31), equivalently, for any  $T > 0$  there exists  $K_T > 0$  such that

$$\int_0^T |\partial_x \psi(t, L)|^2 dt \leq K_T \left( \|\psi(0)\|_{H_0^1(0,L)}^2 + \|\partial_t \psi(0)\|_{L^2(0,L)}^2 \right).$$

for any solution of (5.12). Indeed,  $\int_0^T |\partial_x \psi(t, L)|^2 dt \leq \int_0^{2nL} |\partial_x \psi(t, L)|^2 dt$  for some  $n \in \mathbb{N}^*$  we perform the same expansion as above.

We conclude that, for  $T \geq 2L$ , we have the double inequality

$$C_T \|(\psi(0), \partial_t \psi(0))\|_{H_0^1 \times L^2}^2 \leq \int_0^T |\partial_x \psi(t, L)|^2 dt \leq K_T \|(\psi(0), \partial_t \psi(0))\|_{H_0^1 \times L^2}^2$$

for all solutions of (5.12), saying that  $(\int_0^T |\partial_x \psi(t, L)|^2 dt)^{1/2}$  is a norm, equivalent to the norm of  $H_0^1(0, L) \times L^2(0, L)$ . This illustrates Lemma 5.10. The term  $\int_0^T |\partial_x \psi(t, L)|^2 dt$  stands for the Gramian.

Let us finally prove that controllability is lost if  $T < 2L$ . Let  $\delta > 0$  be such that  $T \leq 2L - 2\delta$ . We consider a solution of (5.12) such that  $\psi(T/2, \cdot)$  and  $\partial_x \psi(T/2, \cdot)$  are supported in  $(0, \delta)$ . Then, using the fact that the support of any solution of the wave equation propagates at speed 1, it follows that the observability inequality (5.7) is not satisfied.

Note that the same argument shows that, for  $T < 2L$ , the wave equation is not approximately controllable either.

### 5.2.4.2 1D Dirichlet Wave Equation with Internal Control

Instead of (5.11), we consider

$$\begin{aligned} \partial_{tt} y &= \partial_{xx} y + \chi_\omega u, & t \in (0, T), x \in (0, L), \\ y(t, 0) &= 0, \quad y(t, L) = 0, & t \in (0, T), \\ y(0, x) &= y_0(x), \quad \partial_t y(0, x) = y_1(x), & x \in (0, L), \end{aligned} \quad (5.14)$$

where the control is  $u(t, x) \in \mathbb{R}$  and  $\omega \subset (0, L)$  is a measurable subset of positive Lebesgue measure. Let us establish that (5.14) is exactly controllable in time  $T \geq 2L$  in the space  $H_0^1(0, L) \times L^2(0, L)$  with controls  $u \in L^2((0, T) \times \omega)$ .

By Theorem 5.30, this is equivalent to establishing the following observability inequality: for every  $T \geq 2L$ , for every  $\omega \subset (0, L)$  measurable of positive measure, there exists  $C_T(\omega) > 0$  such that

$$\int_0^T \int_\omega \phi(t, x)^2 dx dt \geq C_T(\omega) \left( \|\phi(0)\|_{L^2(0, L)}^2 + \|\partial_t \phi(0)\|_{H^{-1}(0, L)}^2 \right) \quad (5.15)$$

for every solution  $\phi$  of the adjoint equation

$$\partial_{tt} \phi - \partial_{xx} \phi = 0, \quad \phi(t, 0) = \phi(t, \pi) = 0. \quad (5.16)$$

We consider solutions  $\phi$  of (5.16) expanded as

$$\phi(t, x) = \sum_{j=1}^{\infty} \left( a_j \cos\left(\frac{j\pi t}{L}\right) + b_j \sin\left(\frac{j\pi t}{L}\right) \right) \sin\left(\frac{j\pi x}{L}\right)$$

with  $(a_j)_{j \in \mathbb{N}^*} \in \ell^2(\mathbb{R})$  and  $(b_j)_{j \in \mathbb{N}^*} \in \ell^2(\mathbb{R})$ , so that  $\phi(0, x) = \sum_{j=1}^{\infty} a_j \sin\left(\frac{j\pi x}{L}\right)$  and  $\partial_t \phi(0, x) = \sum_{j=1}^{\infty} \frac{j\pi}{L} b_j \sin\left(\frac{j\pi x}{L}\right)$  and thus

$$\|\phi(0)\|_{L^2(0, L)}^2 + \|\partial_t \phi(0)\|_{H^{-1}(0, L)}^2 = \frac{L}{2} \sum_{j=1}^{\infty} (a_j^2 + b_j^2).$$

For every  $T \geq 2L$ , we have  $\int_0^T \int_\omega \phi(t, x)^2 dx dt \geq \int_0^{2L} \int_\omega \phi(t, x)^2 dx dt$  and

$$\begin{aligned} & \int_0^{2L} \int_\omega \phi(t, x)^2 dx dt \\ &= \frac{L}{\pi} \sum_{j,k=1}^{\infty} \int_0^{2\pi} (a_j \cos(js) + b_j \sin(js))(a_k \cos(ks) + b_k \sin(ks)) ds \\ & \quad \times \int_\omega \sin\left(\frac{j\pi x}{L}\right) \sin\left(\frac{k\pi x}{L}\right) dx \\ &= L \sum_{j=1}^{\infty} (a_j^2 + b_j^2) \int_\omega \sin^2\left(\frac{j\pi x}{L}\right) dx. \end{aligned}$$

We now give two ways to infer the observability inequality (5.15).

**First Way** We observe that  $\sin^2\left(\frac{j\pi x}{L}\right) \rightharpoonup \frac{1}{2}$  as  $j \rightarrow +\infty$  (in weak  $L^2$  topology). It follows that, for any measurable subset  $\omega \subset (0, L)$  of positive measure, there exists  $C(\omega) > 0$  such that

$$\int_\omega \sin^2\left(\frac{j\pi x}{L}\right) dx \geq C(\omega) \quad \forall j \in \mathbb{N}^*,$$

and then (5.15) follows.

Note that we have used here an information on the highfrequency eigenfunctions  $\phi_j(x) = \sqrt{\frac{2}{L}} \sin\left(\frac{j\pi x}{L}\right)$ .

**Second Way** We have the following lemma.

**Lemma 5.11** *Given any measurable subset  $\omega \subset (0, L)$  of positive Lebesgue measure  $|\omega| > 0$ , we have*

$$\int_\omega \sin^2\left(\frac{j\pi x}{L}\right) dx \geq \frac{1}{2} \left( |\omega| - \frac{L}{\pi} \sin\left(\frac{\pi}{L} |\omega| \right) \right) \quad \forall j \in \mathbb{N}^*.$$

**Proof** For a fixed integer  $j$ , consider the problem of minimizing the functional  $K_j(\omega') = \int_{\omega'} \sin^2\left(\frac{j\pi x}{L}\right) dx$  over all possible measurable subsets  $\omega' \subset (0, \pi)$  s.t.  $|\omega'| = |\omega|$ . Identifying the minima (zeros) of  $\sin^2\left(\frac{j\pi x}{L}\right)$ , using a bathtub principle argument, it is quite obvious to see that there exists a unique (up to zero measure subsets) optimal set, characterized as a level set of the function  $x \mapsto \sin^2\left(\frac{j\pi x}{L}\right)$ , which is

$$\omega_j^{\inf} = \left(0, \frac{|\omega|}{2j}\right) \cup \bigcup_{k=1}^{j-1} \left(\frac{kL}{j} - \frac{|\omega|}{2j}, \frac{kL}{j} + \frac{|\omega|}{2j}\right) \cup \left(L - \frac{|\omega|}{2j}, L\right)$$

and we have

$$\begin{aligned} \int_{\omega_j^{\inf}} \sin^2 \left( \frac{j\pi x}{L} \right) dx &= 2j \int_0^{|\omega|/2j} \sin^2 \left( \frac{j\pi x}{L} \right) dx \\ &= \frac{2L}{\pi} \int_0^{\frac{\pi}{2L}|\omega|} \sin^2 u \, du = \frac{1}{2} \left( |\omega| - \frac{L}{\pi} \sin \left( \frac{\pi}{L} |\omega| \right) \right) \end{aligned}$$

for any  $j$ . Since this value does not depend on  $j$ , the lemma follows.  $\square$

We infer from that lemma that

$$\int_0^T \int_{\omega} \phi(t, x)^2 dx dt \geq \left( |\omega| - \frac{L}{\pi} \sin \left( \frac{\pi}{L} |\omega| \right) \right) \left( \|\phi(0)\|_{L^2}^2 + \|\partial_t \phi(0)\|_{H^{-1}}^2 \right)$$

which gives the observability inequality (5.15). It is interesting to note that, for  $T = 2L$ , the inequality is sharp and thus the observability constant is

$$C_{T=2L}(\omega) = |\omega| - \frac{L}{\pi} \sin \left( \frac{\pi}{L} |\omega| \right).$$

Recalling the admissibility property in Example 5.32, we have obtained that, for  $T \geq 2\pi$ ,  $(\int_0^T \int_{\omega} \phi(t, x)^2 dx dt)^{1/2}$  is a norm, equivalent to the norm of  $L^2(0, L) \times H^{-1}(0, L)$ , which illustrates Lemma 5.10.

As in the boundary control case, controllability is lost if  $T$  is too small: using the finite speed propagation of the wave equation, it suffices to consider a solution supported in  $(0, L) \setminus \omega$  over a small enough time interval.

For instance, if  $\omega = (a, b) \subset (0, L)$  then the minimal controllability time is  $T = 2 \max(a, L - b)$ .

### 5.2.4.3 Multi-D Dirichlet Wave Equation with Internal Control

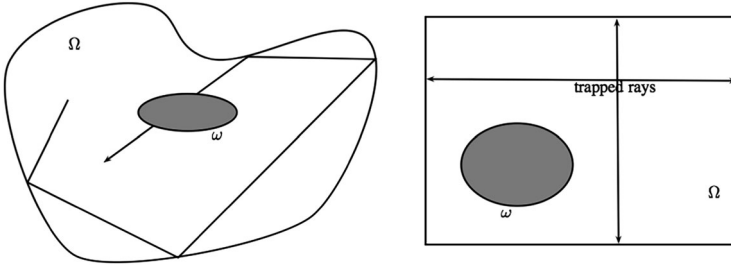
Consider the internally controlled Dirichlet wave equation

$$\begin{aligned} \partial_{tt} y &= \Delta y + \chi_{\omega} u, & t &\in (0, T), \quad x \in \Omega, \\ y(t, x) &= 0, & t &\in (0, T), \quad x \in \partial\Omega \\ y(0, x) &= y_0(x), \quad \partial_t y(0, x) = y_1(x), & x &\in \Omega, \end{aligned} \tag{5.17}$$

on a bounded open domain  $\Omega \subset \mathbb{R}^n$  having a  $C^2$  boundary, and internal control on a measurable subset  $\omega \subset \Omega$  of positive Lebesgue measure.

The admissibility (well-posedness) has been seen in Example 5.32.

It is proved in [2] that, if  $\omega$  is open and if the pair  $(\omega, T)$  satisfies the *Geometric Control Condition* (GCC), then (5.17) is exactly controllable in time  $T$  in the space



**Fig. 5.1** Geodesic rays propagating in  $\Omega$

$H_0^1(\Omega) \times L^2(\Omega)$  with controls  $u \in L^2((0, T) \times \omega)$ ; equivalently, by Theorem 5.30, there exists  $C_T(\omega) > 0$  such that

$$\int_0^T \int_{\omega} \phi(t, x)^2 dx dt \geq C_T(\omega) \left( \|\phi(0)\|_{L^2(\Omega)}^2 + \|\partial_t \phi(0)\|_{H^{-1}(\Omega)}^2 \right) \quad (5.18)$$

for every solution  $\phi$  of the adjoint equation  $\partial_t \phi - \Delta \phi = 0$  with  $\phi = 0$  along the boundary of  $\Omega$ . The GCC stipulates that any geodesic ray, propagating in  $\Omega$  (seen as a billiard) at speed 1 and reflecting at the boundary according to the laws of classical optics (see Fig. 5.1), meets the open set  $\omega$  within time  $T$ . On this figure, on the right,  $\Omega$  is a square in the plane and  $\omega$  is an internal disk: the GCC is never satisfied because of the existence of trapped rays (bouncing balls). The proof of (5.18) requires significantly more elaborate tools than in the 1D case: microlocal analysis, propagation of singularities, defect measures. The “almost-equivalence” between GCC and the observability inequality (5.18) is studied in [10].

We have obtained that, under GCC,  $(\int_0^T \int_{\omega} \phi(t, x)^2 dx dt)^{1/2}$  is a norm, equivalent to the norm of  $L^2(\Omega) \times H^{-1}(\Omega)$ . This illustrates Lemma 5.10.

There are similar results for the boundary control case. Note that, as initially developed in [20], under stronger geometric conditions on  $\omega$ , there are more elementary proofs based on the multiplier method (see [15] and [28, (7.0.3) and Theorem 7.4.1]). Carleman estimates can also be used, with the advantage of allowing to tackle lower-order and/or low regularity terms (see [12]).

## 5.2.5 Example: The Heat Equation

### 5.2.5.1 Dirichlet Heat Equation with Internal Control

Let  $\Omega \subset \mathbb{R}^n$  be a bounded open set having a  $C^2$  boundary, and let  $\omega \subset \Omega$  be an open subset. Like in Sect. 5.1.4.1, we consider the heat equation with internal control and

Dirichlet boundary conditions

$$\partial_t y = \Delta y + \chi_\omega u \quad \text{in } \Omega, \quad y|_{\partial\Omega} = 0, \quad y(0) = y_0 \in L^2(\Omega).$$

The admissibility property of the control operator, as discussed in Sect. 5.1.4.1, is obvious since  $B$  is bounded.

Due to the smoothing effect of the heat semigroup ( $e^{t\Delta}y_0$  is smooth on  $\Omega$  for every  $t > 0$ , whatever the regularity of  $y_0$  may be), the above heat equation cannot be exactly controllable in  $X = L^2(\Omega)$ . But it is approximately controllable in  $X$  and exactly null controllable in  $X$  in any time  $T > 0$ .

Indeed, by Theorem 5.30, the approximate controllability property is equivalent to the following property: given any  $T > 0$ , given any solution  $\psi$  of

$$\partial_t \psi = \Delta \psi \quad \text{in } \Omega, \quad \psi|_{\partial\Omega} = 0 \tag{5.19}$$

such that  $\psi(0) \in H^2(\Omega) \cap H_0^1(\Omega)$ , if  $\psi(t, x) = 0$  for all  $t \in [0, T]$  and  $x \in \omega$  then  $\psi \equiv 0$ . This unique continuation property is a consequence of Holmgren's theorem (see [20, Section 1.8]).

Exact null controllability is equivalent, by Theorem 5.30, to the following observability inequality: given any  $T > 0$ , there exists  $C_T(\omega) > 0$  such that

$$\int_0^T \int_\omega \psi(t, x)^2 dx dt \geq C_T(\omega) \|\psi(T)\|_X^2 \tag{5.20}$$

for any solution of (5.19). The observability inequality (5.20) has been established in [9, 14, 17]. Nowadays, it seems that the most powerful tool in order to establish such inequalities in the parabolic setting is the *Carleman estimates* (see [28, Chapter 9] or [5, Section 2.5]). Even in 1D, the proof of (5.20) by a Carleman estimate is quite technical. It can however be noted that, in 1D, the exact null controllability property can also be proved thanks to harmonic analysis considerations, by applying the *moment method* (see Sect. 5.3.3).

### 5.2.5.2 Heat Equation with Dirichlet Boundary Control

Like in Sect. 5.1.4.2, we consider the heat equation with Dirichlet boundary control

$$\partial_t y = \Delta y \quad \text{in } \Omega, \quad y|_{\partial\Omega} = u, \quad y(0) = y_0 \in L^2(\Omega), \tag{5.21}$$

where  $\Omega \subset \mathbb{R}^n$  is a bounded open subset having a  $C^2$  boundary. We have seen in that section that, setting  $X = H^{-1}(\Omega)$  and  $U = L^2(\partial\Omega)$ , the control operator is admissible, but that admissibility is not true if one takes  $X = L^2(\Omega)$ .

According to [28, Chapter 11.5], the heat Eq. (5.21) is exactly null controllable in  $X = H^{-1}(\Omega)$  in any time  $T > 0$ ; equivalently, for every  $T > 0$  there exists



$C_T > 0$  such that

$$\int_0^T \left\| \frac{\partial \psi}{\partial v} \Big|_{\partial \Omega}(t) \right\|_{L^2(\partial \Omega)}^2 dt \geq C_T \|\psi(T)\|_{H_0^1(\Omega)}^2 \quad (5.22)$$

for any solution of  $\partial_t \psi = \Delta \psi$  in  $\Omega$ ,  $\psi|_{\partial \Omega} = 0$ , with  $\psi(0) \in H_0^1(\Omega)$ .

In most of the existing literature (see, e.g., [14, 17]) one can find the observability inequality (5.22) with the  $L^2$  norm at the right-hand side, thus saying by duality that the heat Eq. (5.21) is exactly null controllable in  $L^2(\Omega)$  in any time  $T > 0$ . Actually, it is exactly controllable in any  $H^s(\mathbb{R})$ , for any  $s \in \mathbb{R}$  (and in particular for any  $s < 0$ ): indeed, taking any  $y_0 \in H^s(\Omega)$ , considering an appropriate extension  $S(t)$  of the Dirichlet heat semigroup, one has  $S(t)y_0 \in H_0^1(\Omega) \cap C^\infty(\Omega)$  for any  $t > 0$  and in particular for  $t = T/2$ ; then apply the controllability property in time  $T/2$  to steer  $S(T/2)y_0$  to 0. Then, by duality, the observability inequality (5.22) remains true when replacing the  $H_0^1$  norm at the right-hand side with the norm of  $X_\alpha$ , for any  $\alpha \in \mathbb{R}$ , where  $(X_\alpha)_{\alpha \in \mathbb{R}}$  is the family of Dirichlet spaces constructed in Example 4.30.

### 5.2.6 Pontryagin Maximum Principle

Formally, HUM is obtained by applying the PMP to the LQ optimal control problem consisting of steering the control system  $\dot{y} = Ay + Bu$  from  $y(0) = y_0$  to  $y(T) = y_1$  in time  $T$ , by minimizing the cost functional  $\int_0^T \|u(t)\|_U^2 dt$ . We have anyway to be careful there. Indeed, as briefly discussed in Sect. 2.2.3, the generalization of the PMP to the infinite-dimensional setting, done for instance in [18, Chapter 4] (proved by using the Ekeland variational principle), requires in general that the final state  $y(T)$  is subject to a finite number only of scalar constraints. There are counterexamples to the statement of the PMP whenever  $y(T) = y_1 \in X$  when  $\dim X = +\infty$ , i.e., when there are an infinite number of final scalar constraints (such counterexamples are easy to design by considering systems enjoying approximate but not exact controllability: see Example 5.33 hereafter). Nevertheless, under exact controllability properties, the PMP is valid for LQ optimal control problems.

The PMP is generalized to infinite-dimensional nonlinear optimal control problems, with a nonlinear cost as in Chap. 2 and a control system of the form  $\dot{y}(t) = Ay(t) + f(t, y(t), u(t))$ , where  $A$  is an operator on a separable or reflexive Banach space  $X$  generating a  $C_0$  semigroup, and we have the same statement as in Theorem 2.10, under the following additional assumption (a more general assumption is given in [18, Chapter 4]): setting  $C$  equal either to  $M_1$  (the target set) or to the Pontryagin cone at the final time (set of variation vectors generated by needle-like variations), we assume that  $\overline{\text{Conv}}(C)$  is of finite codimension in  $X$  and that  $C$  has nonempty interior in  $\overline{\text{Conv}}(C)$ . The first possibility ( $C = M_1$ ) roughly

means that we are allowed to impose only a finite number of scalar constraints on  $y(T)$ ; the second possibility is satisfied for instance if there is no control constraint and the linearized control along the optimal trajectory is exactly controllable, or at least, that the part that is not exactly reachable is of finite dimension.

We do not give more details since, under this additional finite codimensionality assumption, the statement is the same as Theorem 2.10 (with an adjoint  $p(t) \in X'$ ).

**Example 5.33** Following [18, Chapter 4], let us design an example of an optimal control problem in infinite dimension for which the expected PMP statement fails. Let  $X$  be an infinite-dimensional separable Hilbert space. Consider the control system  $\dot{y}(t) = Ay(t) + bu(t)$ , with initial condition  $y(0) = 0$ , where  $A : D(A) \rightarrow X$  is an operator generating a  $C_0$  semigroup  $(S(t))_{t \geq 0}$  and where  $b \in X$  is fixed; here, the control  $u$  is a real-valued function.

The idea is to make assumptions ensuring that we can find a point  $y_1 \in X$  that can be reached from 0 in time 1 with only one control. The constant control  $u = 1$  steers in time 1 the control system to  $y_1 = y(1) = \int_0^1 S(t)b \, dt = (S(1) - \text{id})A^{-1}b$  (the latter formula is obtained by assuming that  $A$  is invertible). Let us assume that  $A$  is self-adjoint negative, so that there exists a Hilbert basis  $(\phi_j)_{j \in \mathbb{N}^*}$  of eigenfunctions, i.e.,  $A\phi_j = \lambda_j \phi_j$  with  $\lambda_j < 0$  for every  $j \in \mathbb{N}^*$ .

For any control  $\bar{u}$  steering the system from  $y(0) = 0$  to  $y(1) = y_1$ , we must have  $\int_0^1 S(1-t)b(\bar{u}(t) - 1) \, dt = 0$ , i.e., expanding  $b = \sum_{j \in \mathbb{N}^*} b_j \phi_j$ ,

$$\sum_{j \in \mathbb{N}^*} \int_0^1 e^{(1-t)\lambda_j} (\bar{u}(t) - 1) \, dt \, b_j \phi_j = 0.$$

Assuming that  $b_j \neq 0$  for every  $j \in \mathbb{N}^*$ , we infer that  $\int_0^1 e^{-t\lambda_j} (\bar{u}(t) - 1) \, dt = 0$  for every  $j \in \mathbb{N}^*$ . Let us now further assume that  $\sum_{j \in \mathbb{N}^*} 1/|\lambda_j| = +\infty$ . Then, by Remark 5.58 in Sect. 5.3.3 further, the family  $(e^{-t\lambda_j})_{j \in \mathbb{N}^*}$  is complete in  $L^2(0, 1)$ . It then follows that  $\bar{u} = 1$ . Hence, under the above assumptions,  $\bar{u} = 1$  is the unique solution steering the system from  $y(0) = 0$  to  $y(1) = y_1$ . Therefore, the control  $\bar{u} = 1$  is optimal for any cost functional.

Let us consider the cost functional

$$C(u) = \int_0^1 (\langle a, y(t) \rangle_X + cu(t)) \, dt$$

for some fixed  $a \in X$  and  $c \in \mathbb{R}$ , and let us assume that the controls are subject, for instance, to the constraint  $|u(t)| \leq 2$  for almost every  $t \in [0, 1]$ . The Hamiltonian of the optimal control problem is

$$H(y, p, p^0, u) = (p, Ay)_X + (p, b)_X u + p^0(a, y)_X + p^0 cu.$$

Let us prove, by contradiction, that the statement of the PMP is not satisfied for the optimal control  $\bar{u} = 1$ . Otherwise, there would exist an adjoint  $p(\cdot)$  satisfying  $\dot{p} =$

$-Ap - p^0 a$ , and thus  $p(t) = S(1-t)p(1) + p^0 \int_t^1 S(1-s)a \, ds = S(1-t)p(1) + p^0(S(1-t) - \text{id})A^{-1}a$ . Besides, the condition  $\frac{\partial H}{\partial u} = 0$  gives  $(p(t), b)_X + p^0 c = 0$  on  $[0, 1]$ , hence, using that  $S(1-t)b = \sum_{j \in \mathbb{N}^*} e^{(1-t)\lambda_j} b_j \phi_j$ ,

$$\sum_{j \in \mathbb{N}^*} e^{(1-t)\lambda_j} b_j (p(1) + p^0 A^{-1}a, \phi_j)_X + p^0 (c - (A^{-1}a, b)_X) = 0$$

for every  $t \in [0, 1]$  and thus (by analyticity) for every  $t \in \mathbb{R}$ . Considering limits at infinity, we easily infer by iteration that all terms above are equal to 0. But then, assuming that  $c \neq \langle A^{-1}a, b \rangle_X$ , it follows that  $p(1) = 0$  and  $p^0 = 0$ , which is a contradiction.

### 5.3 Further Results

#### 5.3.1 Kalman Condition in Infinite Dimension

It is interesting to mention that the unique continuation property implies an infinite-dimensional version of the Kalman condition. A simple sufficient condition is the following.

**Lemma 5.12** *We assume that  $X$  is reflexive and that  $B \in L(U, X)$  is a bounded control operator. We set*

$$U_\infty = \left\{ u \in U \mid Bu \in \bigcap_{k=1}^{+\infty} D(A^k) \right\}.$$

*If the set  $\mathcal{K}_T = \text{Span}\{A^k Bu \mid u \in U_\infty, k \in \mathbb{N}\}$  is dense in  $X$  then the control system (5.1) is approximately controllable in any time  $T > 0$ .*

Note that the set  $\mathcal{K}_T$  is the infinite-dimensional version of the image of the Kalman matrix in finite dimension.

**Proof** We use the equivalence between approximate controllability and (5.8). Note that, in (5.8), it suffices to take  $z$  in any dense subspace of  $D(A^*)$ . Let  $z \in \bigcap_{k=1}^{+\infty} D((A^*)^n)$  (which is dense in  $D(A^*)$ ) be such that  $B^*S(T-t)^*z = 0$  for every  $t \in [0, T]$ . Then, by successive derivations with respect to  $t$ , and taking  $t = T$ , we obtain  $B^*(A^k)^*z = 0$ , hence  $\langle z, A^k Bu \rangle_{X', X} = 0$ , and therefore  $z = 0$  because  $\mathcal{K}_T$  is dense in  $X$ .  $\square$

We refer the reader to [27] for a more precise result (and an almost necessary and sufficient condition).

### 5.3.2 Necessary Conditions for Exact Controllability

Let us assume that  $X$  is of infinite dimension, and let us provide general conditions under which the control system (5.1) is never exactly controllable in finite time  $T$ , with controls in  $L^2([0, T], U)$ . We have already seen that exact controllability implies that the control operator  $B$  is admissible (and thus  $\alpha(B) \leq 1/2$ ).

**Lemma 5.13** *Under any of the following assumptions:*

- $B \in L(U, X)$  is compact;
- $B \in L(U, X)$  is bounded and  $S(t)$  is compact for every  $t > 0$ ;
- $X \simeq X'$  and  $U \simeq U'$  are Hilbert spaces,  $A$  is a self-adjoint negative operator with compact inverse, and  $B \in L(U, X_{-1/2})$  (and thus  $B$  is admissible and  $\alpha(B) \leq 1/2$ , see Sect. 5.1.3);

*the control system (5.1) is not exactly controllable in any time  $T > 0$  (with controls in  $L^2([0, T], U)$ ).*

For instance,  $B$  is compact if  $U$  is finite dimensional. Then, the first point implies that it is impossible to control exactly an infinite-dimensional system with a finite number of controls (see [8, Theorem 4.1.5]). The second point applies for instance to the heat equation with internal control. The third point applies for instance to the heat equation with Neumann boundary control.

**Proof** In the first case where  $B$  is compact, it is easy to see that the operator  $L_T$  is compact. The conclusion follows since  $X$  is infinite dimensional, by the Riesz compactness lemma.

In the second case (compact semigroup), for every  $\varepsilon > 0$ , we define the operator  $L_{T,\varepsilon} : L^2([0, T], U) \rightarrow X$  by  $L_{T,\varepsilon}u = \int_0^{T-\varepsilon} S(T-t)Bu(t)dt$ . Clearly,  $L_{T,\varepsilon}$  converges strongly  $L_T$  as  $\varepsilon$  tends to 0. Besides, using  $S(T-t) = S(\varepsilon)S(T-\varepsilon-t)$ , we get that  $L_{T,\varepsilon} = S(\varepsilon)L_{T-\varepsilon}$ , and hence we infer that  $L_{T,\varepsilon}$  is a compact operator, for every  $\varepsilon > 0$ . Therefore  $L_T$  is compact and the conclusion follows as previously.

In the third case, recall that  $X_{1/2} = D((-A)^{1/2})$  is the completion of  $D(A)$  for the norm  $\sqrt{-(Ax, x)}_X$ , and  $X_{-1/2} = X'_{1/2}$  with respect to the pivot space  $X$  (see also Remark 4.42). Let  $(\phi_j)_{j \in \mathbb{N}^*}$  an orthonormal basis of  $X$ , consisting of eigenvectors of  $A$  associated with the eigenvalues  $\lambda_j < 0$ , with  $\lambda_j \rightarrow -\infty$  as  $j \rightarrow +\infty$ . Firstly, we have

$$\begin{aligned} \int_0^T \|B^*S(T-t)^*\phi_j\|_U^2 dt &= \int_0^T e^{\lambda_j(T-t)} \|B^*\phi_j\|_U^2 dt \\ &\sim \frac{1}{|\lambda_j|} \|B^*\phi_j\|_U^2 = \|B^*(-A)^{-1/2}\phi_j\|_U^2 \end{aligned} \quad (5.23)$$

as  $j \rightarrow +\infty$ . Secondly, since the operator  $(-A)^{-1/2}$  is compact, it follows that the operator  $B^*(-A)^{-1/2} \in L(X, U)$  is compact as well. Thirdly, we claim that the sequence  $(\phi_j)_{j \in \mathbb{N}^*}$  converges to 0 for the weak topology of  $X$ . Indeed, since

$\sum_{j=1}^{+\infty} (x, \phi_j)_X^2 < +\infty$  for every  $x \in X$  (by Parseval for instance), it follows that  $(x, \phi_j)_X \rightarrow 0$  as  $j \rightarrow +\infty$ , for every  $x \in X$ , whence the claim.

Since  $B^*(-A)^{-1/2} \in L(X, U)$  is compact and since  $\phi_j \rightharpoonup 0$  (weak convergence in  $X$ ), it follows that  $B^*(-A)^{-1/2}\phi_j$  converges strongly to 0. Then, from (5.23), we infer that the observability inequality (5.7) does not hold true. We refer to [28, Proposition 9.1.1] for such arguments.  $\square$

### 5.3.3 Moment Method

The moment method, relying on harmonic analysis results, has been used in the 70s to establish the first exact controllability results, essentially in 1D (see [24, Sections 4 and 7]). To explain the moment method, let us consider the 1D heat equation on  $(0, \pi)$  with internal control and with Dirichlet boundary conditions

$$\partial_t y = \partial_{xx} y + \chi_\omega u, \quad y(t, 0) = y(t, \pi) = 0, \quad y(0) = y_0 \in L^2(0, \pi).$$

The eigenfunctions  $\sqrt{\frac{2}{\pi}} \sin(jx)$  of the Dirichlet-Laplacian, associated with the eigenvalues  $\lambda_j = -j^2$ , make an orthonormal basis of  $L^2(0, \pi)$ . Expanding in series, we have  $y_0(x) = \sum_{j=1}^{+\infty} a_j \sin(jx)$  with  $(a_j)_{j \in \mathbb{N}^*} \in \ell^2(\mathbb{R})$ , and writing  $y(t, x) = \sum_{j=1}^{+\infty} y_j(t) \sin(jx)$ , we get  $\dot{y}_j(t) = -j^2 y_j(t) + \int_\omega u(t, x) \sin(jx) dx$  and thus

$$y_j(T) = e^{-j^2 T} a_j + \int_0^T e^{-j^2(T-t)} \int_\omega u(t, x) \sin(jx) dx dt.$$

In order to realize the exact null controllability in time  $T$ , we wish to find some controls  $u$  such that

$$\forall j \in \mathbb{N}^* \quad \int_0^T \int_\omega u(t, x) e^{-j^2(T-t)} \sin(jx) dx dt = -a_j e^{-j^2 T}. \quad (5.24)$$

By Remark 5.58 further, there exists a sequence  $(\theta_T^j)_{j \in \mathbb{N}^*}$  in  $L^2(0, T)$ , spanning a proper subspace of  $L^2(0, T)$ , that is biorthogonal to the sequence of functions  $t \mapsto e^{-j^2 t}$ ,  $j \in \mathbb{N}^*$ , i.e.,  $\int_0^T e^{-j^2 t} \theta_T^k(t) dt = \delta_{jk}$  for all  $j, k \in \mathbb{N}^*$  (see [22] for fine properties). We search controls  $u$  satisfying (5.24), of the form

$$u(t, x) = \sum_{k=1}^{+\infty} b_k \theta_T^k(T-t) \sin(kx),$$

which gives  $b_j \int_{\omega} \sin^2(jx) dx = -a_j e^{-j^2 T}$  for every  $j \in \mathbb{N}^*$ . We conclude that

$$u(t, x) = - \sum_{k=1}^{+\infty} a_k e^{-k^2 T} \theta_T^k(T-t) \frac{\sin(kx)}{\int_{\omega} \sin^2(ky) dy}.$$

Thanks to Lemma 5.11, this function is well defined and is in  $L^2((0, T) \times (0, \pi))$ .

**Abstract Generalization** Let  $X$  and  $U$  be (complex) Hilbert spaces. Let  $A : D(A) \rightarrow X$  be a densely defined operator that generates a  $C_0$  semigroup and such that there exists a Hilbert basis  $(\phi_j)_{j \in \mathbb{N}^*}$  of  $X$  consisting of eigenvectors of  $A$ , corresponding to (complex) eigenvalues  $(\lambda_j)_{j \in \mathbb{N}^*}$  with  $\dots \leq \operatorname{Re}(\lambda_j) \leq \dots \leq \operatorname{Re}(\lambda_1)$ . Note that  $\phi_j \in X_1 = D(A) = D(A^*)$  for every  $j \in \mathbb{N}^*$ . Let  $B \in L(U, D(A^*))$  be an admissible control operator. We consider the control system (5.1) with  $y(0) = y_0 = \sum_{j \in \mathbb{N}^*} a_j \phi_j$ . Expanding  $y(t) = \sum_{j=1}^{+\infty} y_j(t) \phi_j$ , we have  $\dot{y}_j = \lambda_j y_j + \langle Bu, \phi_j \rangle_{X_{-1}, X_1}$  and thus

$$\forall j \in \mathbb{N}^* \quad y_j(T) = e^{\lambda_j T} a_j + \int_0^T e^{\lambda_j(T-t)} (u(t), B^* \phi_j)_U dt, \quad (5.25)$$

and we want to solve  $y_j(T) = 0$  for every  $j \in \mathbb{N}^*$ .

We assume that there exists a sequence  $(\theta_T^j)_{j \in \mathbb{N}^*}$  in  $L^2(0, T)$  that is biorthogonal to the family  $\Lambda$  of functions  $t \mapsto e^{\lambda_j t}$ ,  $j \in \mathbb{N}^*$  (see Remark 5.58 below). We search  $u$  in the particular form

$$u(t) = \sum_{k \in \mathbb{N}^*} b_k \theta_T^k(T-t) B^* \phi_k.$$

Then, solving  $y_j(T) = 0$  for every  $j \in \mathbb{N}^*$  is equivalent to requiring that  $b_j \|B^* \phi_j\|_U^2 = -a_j e^{\lambda_j T}$ , and thus, formally,

$$u(t) = - \sum_{k \in \mathbb{N}^*} a_k e^{\lambda_k T} \theta_T^k(T-t) \frac{B^* \phi_k}{\|B^* \phi_k\|_U^2}.$$

Showing that this formal series converges and gives a well-defined function requires to establish lower estimates of  $\|B^* \phi_k\|_U^2$ .

**Remark 5.58** Such a biorthogonal sequence  $(\theta_T^j)_{j \in \mathbb{N}^*}$  exists if and only if the family  $\Lambda$  is *minimal*, that is, every element  $t \mapsto e^{\lambda_j t}$  lies outside of the closure in  $L^2([0, T], \mathbb{C})$  of the vector space spanned by all other elements  $t \mapsto e^{\lambda_k t}$ , with  $k \neq j$ . If this condition is fulfilled, then the biorthogonal sequence  $(\theta_T^j)_{j \in \mathbb{N}^*}$  is uniquely determined if and only if the family  $\Lambda$  is complete in  $L^2([0, T], \mathbb{C})$ . Note anyway that the biorthogonal sequence is difficult to construct in practice.

It is well known (Müntz-Szász problem, see [24]) that the family  $\Lambda$  is complete in  $L^2([0, T], \mathbb{C})$  (but not independent) if and only if  $\sum_{j \in \mathbb{N}^*} \frac{1}{\lambda - \operatorname{Re}(\lambda_j)} = +\infty$  for some  $\lambda \in \mathbb{R}$  such that  $\lambda - \operatorname{Re}(\lambda_j) > 0$  for every  $j \in \mathbb{N}^*$  (for instance,  $\lambda = 1 + \operatorname{Re}(\lambda_1)$ ). At the opposite, if this series is convergent then the closure of the span of  $\Lambda$  is a proper subspace of  $L^2([0, T], \mathbb{C})$ , moreover  $\Lambda$  is minimal and thus a biorthogonal sequence exists.

Then, here, we are led to assume that the above series is convergent, which is a quite strong restriction on the system under consideration.

**Relationship with HUM** Within the previous abstract general framework, let us solve the moment equations  $y_j(T) = 0$  for every  $j \in \mathbb{N}^*$ , in another way: using (5.25), it suffices to search a control  $u$  such that

$$(u, (t, x) \mapsto e^{\lambda_j(T-t)}(B^*\phi_j)(x))_{L^2([0, T], U)} = -e^{\lambda_j T} a_j \quad \forall j \in \mathbb{N}^*$$

and, generalizing the previous approach, we can solve this moment problem by using, if it exists, a biorthogonal sequence  $(u_j)_{j \in \mathbb{N}^*}$  to the sequence of (time-space) functions  $(t, x) \mapsto e^{\lambda_j(T-t)}(B^*\phi_j)(x)$ , i.e., noting that  $S(T-t)^*\phi_k = e^{\lambda_k(T-t)}\phi_k$ , a sequence satisfying

$$(u_j, B^*S(T-t)^*\phi_k)_{L^2([0, T], U)} = \delta_{jk} \quad \forall j, k \in \mathbb{N}^*. \quad (5.26)$$

Note that, when such a family exists,  $u_j$  is a control steering the control system (5.1) from  $-e^{-\lambda_j T}\phi_j$  to 0 in time  $T$  (this is related to the notion of *spectral controllability*).

There are plenty of ways for designing such controls  $u_j$ , when this is possible. For every  $j \in \mathbb{N}^*$ , let  $u_j \in L^2([0, T], U)$  be the (unique) HUM control steering the initial condition  $-e^{-\lambda_j T}\phi_j$  to 0 (we assume that this is possible): we have  $u_j(t) = B^*S(T-t)^*\psi_j$  where  $G_T\psi_j = \phi_j$  (note that  $u_j$  is the control of minimal  $L^2$  norm, and  $\|u_j\|_{L^2([0, T], U)}^2 = (G_T\psi_j, \psi_j)_U = (\phi_j, \psi_j)_U$ ). Then, obviously, (5.26) holds.

Now, the (unique) HUM control  $u$  such that  $y(T) = 0$  is given by  $u(t) = B^*S(T-t)^*\psi$  with  $S(T)y_0 + G_T\psi = 0$ . Since  $y_0 = \sum_{j \in \mathbb{N}^*} a_j\phi_j$ , it easily follows by linearity that, formally,

$$u = -\sum_{j=1}^{+\infty} a_j e^{\lambda_j T} u_j.$$

Of course, all above computations are formal, and it may be difficult to establish the convergence of the series in practical examples.

### 5.3.4 Equivalence Between Observability and Exponential Stability

The following result is a generalization of the main result of [13].

**Theorem 5.31** *Let  $X$  be a Hilbert space, let  $A : D(A) \rightarrow X$  be a densely defined skew-adjoint operator; let  $B$  be a bounded selfadjoint nonnegative operator on  $X$ . We have equivalence of:*

- (i) *There exist  $T > 0$  and  $C_1 > 0$  such that every solution of the conservative equation*

$$\dot{\phi}(t) + A\phi(t) = 0 \quad (5.27)$$

*(note that  $\|\phi(t)\|_X = \|\phi(0)\|_X$  for every  $t \in \mathbb{R}$ ) satisfies the observability inequality*

$$\|\phi(0)\|_X^2 \leq C_1 \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt.$$

- (ii) *There exist  $T > 0$  and  $C_2 > 0$  such that every solution of the damped equation*

$$\dot{y}(t) + Ay(t) + By(t) = 0 \quad (5.28)$$

*satisfies the observability inequality*

$$\|y(0)\|_X^2 \leq C_2 \int_0^T \|B^{1/2}y(t)\|_X^2 dt.$$

- (iii) *There exist  $C_3 > 0$  and  $\delta > 0$  such that every solution of the damped Eq. (5.28) satisfies*

$$E_y(t) \leq C_3 E_y(0) e^{-\delta t} \quad \forall t \geq 0$$

*where  $E_y(t) = \frac{1}{2} \|y(t)\|_X^2$ .*

**Proof** Let us first prove that (i) implies (ii). Let  $y$  be a solution of (5.28). Let  $\phi$  be the solution of (5.27) with  $\phi(0) = y(0)$ . Setting  $\theta = y - \phi$ , we have

$$\dot{\theta} + A\theta + B\theta = 0, \quad \theta(0) = 0.$$

Then, taking the scalar product with  $\theta$ , since  $A$  is skew-adjoint, we get  $\langle \dot{\theta} + B\theta, \theta \rangle_X = 0$ . Setting  $E_\theta(t) = \frac{1}{2} \|\theta(t)\|_X^2$ , we have  $\dot{E}_\theta = -\langle B\theta, \theta \rangle_X$ . Then, integrating a first time over  $[0, t]$ , and then a second time over  $[0, T]$ , since



$E_\theta(0) = 0$ , we get

$$\begin{aligned} \int_0^T E_\theta(t) dt &= - \int_0^T \int_0^t \langle By(s), \theta(s) \rangle_X ds dt \\ &= - \int_0^T (T-t) \langle B^{1/2}y(t), B^{1/2}\theta(t) \rangle_X dt, \end{aligned}$$

where we have used the Fubini theorem. Using the Cauchy-Schwarz inequality and then the Young inequality  $ab \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2$  with  $\alpha = 2T\|B^{1/2}\|$ , we infer that

$$\begin{aligned} \frac{1}{2} \int_0^T \|\theta(t)\|_X^2 dt &\leq T\|B^{1/2}\| \int_0^T \|B^{1/2}y(t)\|_X \|\theta(t)\|_X dt \\ &\leq T^2\|B^{1/2}\|^2 \int_0^T \|B^{1/2}y(t)\|_X^2 dt + \frac{1}{4} \int_0^T \|\theta(t)\|_X^2 dt \end{aligned}$$

and therefore,

$$\int_0^T \|\theta(t)\|_X^2 dt \leq 4T^2\|B^{1/2}\|^2 \int_0^T \|B^{1/2}y(t)\|_X^2 dt.$$

Now, since  $\phi = y - \theta$ , using the inequality  $(a+b)^2 \leq 2a^2 + 2b^2$ , it follows that

$$\begin{aligned} \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt &\leq 2 \int_0^T \|B^{1/2}y(t)\|_X^2 dt + 2 \int_0^T \|B^{1/2}\theta(t)\|_X^2 dt \\ &\leq (2 + 8T^2\|B^{1/2}\|^4) \int_0^T \|B^{1/2}y(t)\|_X^2 dt. \end{aligned}$$

Finally, since

$$E_y(0) = E_\phi(0) = \frac{1}{2}\|\phi(0)\|_X^2 \leq \frac{C_1}{2} \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt$$

(where we have used the observability inequality of (i)), it follows that

$$E_y(0) \leq C_1(1 + 4T^2\|B^{1/2}\|^4) \int_0^T \|B^{1/2}y(t)\|_X^2 dt,$$

which is the observability inequality stated in (ii), with the constant  $C_2 = 2C_1(1 + 4T^2\|B^{1/2}\|^4)$ .

Let us now prove that (ii) implies (iii). Let  $y$  be a solution of the damped Eq. (5.28). Since  $\dot{E}_y(t) = -\|B^{1/2}y(t)\|_X^2$ , by integration we have then  $\int_0^T \|B^{1/2}y(t)\|_X^2 dt = E_y(0) - E_y(T)$ . Using the observability inequality of (ii), it follows that

$$E_y(0) \leq 2C_2(E_y(0) - E_y(T))$$

and hence

$$E_y(T) \leq \frac{2C_2 - 1}{2C_2} E_y(0) \leq C'_2 E_y(0)$$

with  $0 < C'_2 < 1$ .

Actually this can be done on every interval  $[kT, (k+1)T]$ , and it yields  $E_y((k+1)T) \leq C'_2 E_y(kT)$  for every  $k \in \mathbb{N}$ , and hence  $E_y(kT) \leq E_y(0)(C'_2)^k$ .

For every  $t \in [kT, (k+1)T]$ , noting that  $k = \lfloor \frac{t}{T} \rfloor > \frac{t}{T} - 1$ , and that  $\ln \frac{1}{C'_2} > 0$ , it follows that

$$(C'_2)^k = \exp(k \ln C'_2) = \exp\left(-k \ln \frac{1}{C'_2}\right) \leq \frac{1}{C'_2} \exp\left(\frac{-\ln \frac{1}{C'_2}}{T} t\right)$$

and hence  $E_y(t) \leq E_y(kT) \leq C_3 E_y(0) \exp(-\delta t)$ , with  $C_3 = 1/C'_2$  and  $\delta = \frac{\ln \frac{1}{C'_2}}{T}$ .

Let us finally prove that (iii) implies (i). Let  $\phi$  be a solution of (5.27). Let  $y$  be a solution of (5.28) such that  $y(0) = \phi(0)$ . We infer from the exponential decrease inequality of (iii) that

$$\int_0^T \|B^{1/2}y(t)\|_X^2 dt = E_y(0) - E_y(T) \geq (1 - C_3 e^{-\delta T}) E_y(0) = C_4 E_y(0), \quad (5.29)$$

and for  $T > 0$  large enough there holds  $C_4 = 1 - C_3 e^{-\delta T} > 0$ . Then we make the same proof as before. Writing (5.27) in the form  $\dot{\phi} + A\phi + B\phi = B\phi$  and setting  $\theta = \phi - y$ , we have

$$\dot{\theta} + A\theta + B\theta = B\phi, \quad \theta(0) = 0.$$

Taking the scalar product with  $\theta$ , since  $A$  is skew-adjoint, we get  $\langle \dot{\theta} + B\theta, \theta \rangle_X = \langle B\phi, \theta \rangle_X$ , and therefore

$$\dot{E}_\theta + \langle B\theta, \theta \rangle_X = \langle B\phi, \theta \rangle_X.$$

Since  $\langle B\theta, \theta \rangle_X = \|B^{1/2}\theta\|_X^2 \geq 0$ , it follows that  $\dot{E}_\theta \leq \langle B\phi, \theta \rangle_X$ . As before we apply  $\int_0^T \int_0^t$  and hence, since  $E_\theta(0) = 0$ ,

$$\begin{aligned} \int_0^T E_\theta(t) dt &\leq \int_0^T \int_0^t \langle B\phi(s), \theta(s) \rangle_X ds dt \\ &= \int_0^T (T-t) \langle B^{1/2}\phi(t), B^{1/2}\theta(t) \rangle_X dt. \end{aligned}$$

Using the Cauchy-Schwarz inequality and then the Young inequality, we get, exactly as before,

$$\begin{aligned} \frac{1}{2} \int_0^T \|\theta(t)\|_X^2 dt &\leq T \|B^{1/2}\| \int_0^T \|B^{1/2}\phi(t)\|_X \|\theta(t)\|_X dt \\ &\leq T^2 \|B^{1/2}\|^2 \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt + \frac{1}{4} \int_0^T \|\theta(t)\|_X^2 dt, \end{aligned}$$

and finally,

$$\int_0^T \|\theta(t)\|_X^2 dt \leq 4T^2 \|B^{1/2}\|^2 \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt.$$

Now, since  $y = \phi - \theta$ , it follows that

$$\begin{aligned} \int_0^T \|B^{1/2}y(t)\|_X^2 dt &\leq 2 \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt + 2 \int_0^T \|B^{1/2}\theta(t)\|_X^2 dt \\ &\leq (2 + 8T^2 \|B^{1/2}\|^4) \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt \end{aligned}$$

Now, using (5.29) and noting that  $E_y(0) = E_\phi(0)$ , we infer that

$$C_4 E_\phi(0) \leq (2 + 8T^2 \|B^{1/2}\|^4) \int_0^T \|B^{1/2}\phi(t)\|_X^2 dt$$

which gives the observability inequality of (i) with  $C_1 = \frac{1}{C_4} (2 + 8T^2 \|B^{1/2}\|^4)$ .  $\square$

**Remark 5.59** This result says that the observability property for a linear conservative equation is equivalent to the exponential stability property for the same equation in which a linear damping has been added. This result has been written in [13] for second-order equations, but the proof works exactly in the same way for more general first-order systems, as shown here. The above proof uses in a crucial way the fact that the operator  $B$  is bounded. We refer to [1] for a generalization for unbounded operators with degree of unboundedness  $\leq 1/2$ , and only for second-order equations, with a proof using Laplace transforms, under a condition on the

unboundedness of  $B$  that is related to “hidden regularity” results. For instance this works for waves with a nonlocal operator  $B$  corresponding to a Dirichlet condition, in the state space  $L^2 \times H^{-1}$ , but not for the usual Neumann one, in the state space  $H^1 \times L^2$  (except in 1D).

### 5.3.5 1D Semilinear Heat Equation

Let  $L > 0$  be fixed and let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a function of class  $C^2$  such that  $f(0) = 0$ . We consider the 1D semilinear heat equation

$$\partial_t y = \partial_{xx} y + f(y), \quad y(t, 0) = 0, \quad y(t, L) = u(t), \quad (5.30)$$

where the state is  $y(t, \cdot) : [0, L] \rightarrow \mathbb{R}$  and the control is  $u(t) \in \mathbb{R}$ .

We want to design a feedback control locally stabilizing (5.30) asymptotically to 0. Note that this cannot be global, because we can have other steady-states (a steady-state is a function  $y \in C^2(0, L)$  such that  $y''(x) + f(y(x)) = 0$  on  $(0, L)$  and  $y(0) = 0$ ). By the way, here, without loss of generality we consider the steady-state 0.

Let us first note that, for every  $T > 0$ , (5.30) is well posed in the Banach space  $Y_T = L^2([0, T], H^2(0, L)) \cap H^1([0, T], L^2(0, L))$ , which is continuously embedded in  $L^\infty((0, T) \times (0, L))$ .<sup>3</sup>

First of all, in order to end up with a Dirichlet problem, we set  $z(t, x) = y(t, x) - \frac{x}{L}u(t)$ . Assuming (for the moment) that  $u$  is differentiable, we set  $v(t) = u'(t)$ , and we consider in the sequel  $v$  as a control. We also assume that  $u(0) = 0$ . Then we have

$$\partial_t z = \partial_{xx} z + f'(0)z + \frac{x}{L}f'(0)u - \frac{x}{L}v + r(t, x), \quad z(t, 0) = z(t, L) = 0, \quad (5.31)$$

<sup>3</sup> Indeed, considering  $v \in L^2([0, T], H^2(0, L))$  with  $v_t \in H^1([0, T], L^2(0, L))$ , writing  $v = \sum_{j,k} c_{jk} e^{ijt} e^{ikx}$ , we have

$$\sum_{j,k} |c_{jk}| \leq \left( \sum_{j,k} \frac{1}{1 + j^2 + k^4} \right)^{1/2} \left( \sum_{j,k} (1 + j^2 + k^4) |c_{jk}|^2 \right)^{1/2}$$

and these series converge, whence the embedding, allowing to give a sense to  $f(y)$ .

Now, if  $y_1$  and  $y_2$  are solutions of (5.30) on  $[0, T]$ , then  $y_1 = y_2$ . Indeed,  $v = y_1 - y_2$  is solution of  $v_t = v_{xx} + a v$ ,  $v(t, 0) = v(t, L) = 0$ ,  $v(0, x) = 0$ , with  $a(t, x) = g(y_1(t, x), y_2(t, x))$  where  $g$  is a function of class  $C^1$ . We infer that  $v = 0$ .

with  $z(0, x) = y(0, x)$  and (by performing a second-order Taylor expansion of  $f$  with integral remainder)

$$r(t, x) = \left( z(t, x) + \frac{x}{L} u(t) \right)^2 \int_0^1 (1-s) f'' \left( s z(s, x) + s \frac{x}{L} u(s) \right) ds.$$

Note that, given  $B > 0$  arbitrary, there exist positive constants  $C_1$  and  $C_2$  such that, if  $|u(t)| \leq B$  and  $\|z(t, \cdot)\|_{L^\infty(0,L)} \leq B$ , then

$$\|r(t, \cdot)\|_{L^\infty(0,L)} \leq C_1(u(t)^2 + \|z(t, \cdot)\|_{L^\infty(0,L)}^2) \leq C_2(u(t)^2 + \|z(t, \cdot)\|_{H_0^1(0,L)}^2).$$

In the sequel,  $r(t, x)$  will be considered as a remainder.

We define the operator  $A = \Delta + f'(0)\text{id}$  on  $D(A) = H^2(0, L) \cap H_0^1(0, L)$ , so that (5.31) is written as

$$\dot{u} = v, \quad \partial_t z = Az + au + bv + r, \quad z(t, 0) = z(t, L) = 0, \quad (5.32)$$

with  $a(x) = \frac{x}{L} f'(0)$  and  $b(x) = -\frac{x}{L}$ .

Since  $A$  is self-adjoint and has a compact resolvent, there exists a Hilbert basis  $(e_j)_{j \geq 1}$  of  $L^2(0, L)$ , consisting of eigenfunctions  $e_j \in H_0^1(0, L) \cap C^2([0, L])$  of  $A$ , associated with eigenvalues  $(\lambda_j)_{j \geq 1}$  such that  $-\infty < \dots < \lambda_n < \dots < \lambda_1$  and  $\lambda_n \rightarrow -\infty$  as  $n \rightarrow +\infty$ .

Any solution  $z(t, \cdot) \in H^2(0, L) \cap H_0^1(0, L)$  of (5.31), as long as it is well defined, can be expanded as a series  $z(t, \cdot) = \sum_{j=1}^{\infty} z_j(t) e_j(\cdot)$  (converging in  $H_0^1(0, L)$ ), and then we have, for every  $j \geq 1$ ,

$$\dot{z}_j(t) = \lambda_j z_j(t) + a_j u(t) + b_j v(t) + r_j(t), \quad (5.33)$$

with

$$a_j = \frac{f'(0)}{L} \int_0^L x e_j(x) dx, \quad b_j = -\frac{1}{L} \int_0^L x e_j(x) dx, \\ r_j(t) = \int_0^L r(t, x) e_j(x) dx.$$

Setting, for every  $n \in \mathbb{N}^*$ ,

$$X_n(t) = \begin{pmatrix} u(t) \\ z_1(t) \\ \vdots \\ z_n(t) \end{pmatrix}, \quad A_n = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_1 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_n & 0 & \cdots & \lambda_n \end{pmatrix}, \quad B_n = \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}, \quad R_n(t) = \begin{pmatrix} 0 \\ r_1(t) \\ \vdots \\ r_n(t) \end{pmatrix},$$

we have then

$$\dot{X}_n(t) = A_n X_n(t) + B_n v(t) + R_n(t).$$

**Lemma 5.14** *For every  $n \in \mathbb{N}^*$ , the pair  $(A_n, B_n)$  satisfies the Kalman condition.*

*Proof* We compute

$$\det(B_n, A_n B_n, \dots, A_n^n B_n) = \prod_{j=1}^n (a_j + \lambda_j b_j) \text{VdM}(\lambda_1, \dots, \lambda_n) \quad (5.34)$$

where  $\text{VdM}(\lambda_1, \dots, \lambda_n)$  is a Van der Monde determinant, and thus is never equal to zero since the  $\lambda_i$ ,  $i = 1 \dots n$ , are pairwise distinct. On the other part, using the fact that each  $e_j$  is an eigenfunction of  $A$  and belongs to  $H_0^1(0, L)$ , we compute

$$a_j + \lambda_j b_j = \frac{1}{L} \int_0^L x(f'(0) - \lambda_j)e_j(x) dx = -\frac{1}{L} \int_0^L x e_j''(x) dx = -e_j'(L),$$

and this quantity is never equal to zero since  $e_j(L) = 0$  and  $e_j$  is a nontrivial solution of a linear second-order scalar differential equation. Therefore the determinant (5.34) is never equal to zero.  $\square$

By the pole-shifting theorem (Theorem 3.17), there exists  $K_n = (k_0, \dots, k_n)$  such that the matrix  $A_n + B_n K_n$  has  $-1$  as an eigenvalue of multiplicity  $n + 1$ . Moreover, by the Lyapunov lemma (see Example 3.17), there exists a symmetric positive definite matrix  $P_n$  of size  $n + 1$  such that

$$P_n (A_n + B_n K_n) + (A_n + B_n K_n)^\top P_n = -I_{n+1}.$$

Therefore, as shown in Example 3.17, the function defined by  $V_n(X) = X^\top P_n X$  for any  $X \in \mathbb{R}^{n+1}$  is a Lyapunov function for the closed-loop system  $\dot{X}_n(t) = (A_n + B_n K_n)X_n(t)$ : along this system we have  $\frac{d}{dt} V_n(X_n(t)) = -\|X_n(t)\|_2^2$ . Here,  $\|\cdot\|_2$  stands for the Euclidean norm of  $\mathbb{R}^{n+1}$ .

Let  $\gamma > 0$  and  $n \in \mathbb{N}^*$  to be chosen later. For every  $u \in \mathbb{R}$  and every  $z \in H^2(0, L) \cap H_0^1(0, L)$ , we set

$$V(u, z) = \gamma X_n^\top P_n X_n - \frac{1}{2} \langle z, Az \rangle_{L^2(0, L)} = \gamma X_n^\top P_n X_n - \frac{1}{2} \sum_{j=1}^{\infty} \lambda_j z_j^2 \quad (5.35)$$

where  $X_n = (u, z_1, \dots, z_n)^\top \in \mathbb{R}^{n+1}$  and  $z_j = \langle z, e_j \rangle_{L^2(0, L)}$  for every  $j$ .

Using that  $\lambda_n \rightarrow -\infty$  as  $n \rightarrow +\infty$ , it is clear that, choosing  $\gamma > 0$  and  $n \in \mathbb{N}^*$  large enough, we have  $V(u, z) > 0$  for all  $(u, z) \in \mathbb{R} \times (H^2(0, L) \cap H_0^1(0, L)) \setminus \{(0, 0)\}$ . More precisely, there exist positive constants  $C_3, C_4, C_5$  and  $C_6$  such that

$$\begin{aligned} C_3 \left( u^2 + \|z\|_{H_0^1(0, L)}^2 \right) &\leq V(u, z) \leq C_4 \left( u^2 + \|z\|_{H_0^1(0, L)}^2 \right), \\ V(u, z) &\leq C_5 \left( \|X_n\|_2^2 + \|Az\|_{L^2(0, L)}^2 \right), \quad \gamma C_6 \|X_n\|_2^2 \leq V(u, z), \end{aligned}$$

for all  $(u, z) \in \mathbb{R} \times (H^2(0, L) \cap H_0^1(0, L))$ .

Our objective is now to prove that  $V$  is a Lyapunov function for the system (5.32) in closed-loop with the feedback control  $v = K_n X_n$  and  $u$  defined by  $\dot{u} = v$  and  $u(0) = 0$ . We compute

$$\begin{aligned} \frac{d}{dt} V(u(t), z(t)) &= -\gamma \|X_n(t)\|_2^2 - \|Az(t, \cdot)\|_{L^2}^2 - \langle Az(t, \cdot), a(\cdot) \rangle_{L^2} u(t) \\ &\quad - \langle Az(t, \cdot), b(\cdot) \rangle_{L^2} K_n X_n(t) - \langle Az(t, \cdot), r(t, \cdot) \rangle_{L^2} \\ &\quad + \gamma \left( R_n(t)^\top P_n X_n(t) + X_n(t)^\top P_n R_n(t) \right). \end{aligned} \quad (5.36)$$

Let us estimate the terms at the right-hand side of (5.36). Under the a priori estimates  $|u(t)| \leq B$  and  $\|z(t, \cdot)\|_{L^\infty(0, L)} \leq B$ , there exist positive constants  $C_7, C_8$  and  $C_9$  such that (dropping the dependence in  $t$ )

$$\begin{aligned} |\langle Az, a \rangle_{L^2} u| + |\langle Az, b \rangle_{L^2} K_n X_n| &\leq \frac{1}{4} \|Az\|_{L^2}^2 + C_7 \|X_n\|_2^2, \\ |\langle Az, r \rangle_{L^2}| &\leq \frac{1}{4} \|Az\|_{L^2}^2 + C_8 V^2, \quad \|R_n\|_\infty \leq \frac{C_2}{C_3} V, \\ |\gamma (R_n^\top P_n X_n + X_n^\top P_n R_n)| &\leq \frac{C_2}{C_3 \sqrt{C_6}} \sqrt{\gamma} V^{3/2}. \end{aligned}$$

We infer that, if  $\gamma > 0$  is large enough, then there exist  $C_{10}, C_{11} > 0$  such that  $\frac{d}{dt} V \leq -C_{10} V + C_{11} V^{3/2}$ . We easily conclude the local asymptotic stability of the system (5.32) in closed-loop with the control  $v = K_n X_n$ .

**Remark 5.60** The above local asymptotic stability can be achieved with other procedures, for instance, by using the Riccati theory (see [31] for Riccati operators in the parabolic case). However, the procedure developed above is much more efficient because it consists of stabilizing a finite-dimensional part of the system, namely, the part that is not naturally stable. We refer to [6] for examples and for more details. Actually, it is proved in that reference that, thanks to such a strategy, one can pass from any steady-state to any other one, provided that the two steady-

states belong to a same connected component of the set of steady-states: this is a partially global exact controllability result.

The main idea used above<sup>4</sup> is the following fact, already used in [24]. Considering the linearized system with no control, we have an infinite-dimensional linear system that can be split, through a spectral decomposition, in two parts: the first part is finite-dimensional, and consists of all spectral modes that are unstable (meaning that the corresponding eigenvalues have nonnegative real part); the second part is infinite-dimensional, and consists of all spectral modes that are asymptotically stable (meaning that the corresponding eigenvalues have negative real part). The idea used here then consists of focusing on the finite-dimensional unstable part of the system, and to design a feedback control in order to stabilize that part. Then, we plug this control in the infinite-dimensional system, and we have to check that this feedback indeed stabilizes the whole system (in the sense that it does not destabilize the other infinite-dimensional part). This is the role of the Lyapunov function  $V$  defined by (5.35).

## References

1. Ammari, K., Tucsnak, M.: Stabilization of second order evolution equations by a class of unbounded operators. *ESAIM: Cont. Optim. Calc. Var.* **6**, 361–386. (2001)
2. Bardos, C., Lebeau, G., Rauch, J.: Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary. *SIAM J. Cont. Optim.* **30**, 1024–1065 (1992)
3. Brezis, H.: *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York (2011)
4. Cazenave, T., Haraux, A.: *An Introduction to Semilinear Evolution Equations*. Oxford Lecture Series in Mathematics and Its Applications, vol. 13. The Clarendon Press/Oxford University Press, New York (1998)
5. Coron, J.-M.: *Control and Nonlinearity*. Mathematical Surveys and Monographs, vol. 136, xiv+426 pp. American Mathematical Society, Providence (2007)
6. Coron, J.-M., Trélat, E.: Global steady-state controllability of 1-D semilinear heat equations. *SIAM J. Control Optim.* **43**(2), 549–569 (2004)
7. Coron, J.-M., Trélat, E.: Global steady-state stabilization and controllability of 1-D semilinear wave equations. *Commun. Contemp. Math.* **8**(4), 535–567 (2006)

---

<sup>4</sup> This idea has been used as well to treat other parabolic problems, and even hyperbolic: it has been as well used in [7] for the 1D semilinear equation

$$\partial_{tt} y = \partial_{xx} y + f(y), \quad y(t, 0) = 0, \quad y_x(t, L) = u(t).$$

We first note that, if  $f(y) = cy$  is linear (with  $c \in L^\infty(0, L)$ ), then, taking  $u(t) = -\alpha \partial_t y(t, L)$  with  $\alpha > 0$  yields an exponentially decrease of the energy  $\int_0^L (\partial_t y(t, x))^2 + \partial_x y(t, x)^2 dt$ , and moreover, the eigenvalues of the corresponding operator have a real part tending to  $-\infty$  as  $\alpha \rightarrow 1$ . Therefore, in the general case, if  $\alpha$  is sufficiently close to 1 then at most a finite number of eigenvalues may have a nonnegative real part. Using a Riesz spectral expansion, the above spectral method yields a feedback based on a finite number of modes, which stabilizes locally the semilinear wave equation, asymptotically to equilibrium.



8. Curtain, R.F., Zwart, H.: An Introduction to Infinite-Dimensional Linear Systems Theory. Texts in Applied Mathematics, vol. 21. Springer-Verlag, New York (1995)
9. Fursikov, A.V., Imanuvilov, O.Y.: Controllability of Evolution Equations. Lecture Notes Series, vol. 34, iv+163 pp. Seoul National University, Seoul (1996)
10. Humbert, E., Privat, Y., Trélat, E.: Observability properties of the homogeneous wave equation on a closed manifold. *Commun. Partial Differ. Equ.* **44**(9), 749–772 (2019)
11. Engel, K.-J., Nagel, R.: One-Parameter Semigroups for Linear Evolution Equations. Graduate Texts in Mathematics, vol. 194. Springer-Verlag, Berlin (2000)
12. Fu, X., Lü, Q., Zhang, X.: Carleman Estimates for Second Order Partial Differential Operators and Applications, A Unified Approach. Springer Briefs in Mathematics. BCAM Springer-Briefs, xi+127 pp. Springer, Cham (2019)
13. Haraux, A.: Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps. *Portugal. Math.* **46**(3), 245–258 (1989)
14. Imanuvilov, O.Y.: Controllability of parabolic equations. *Sb. Math.* **186**(6), 879–900 (1995)
15. Komornik, V.: Exact Controllability and Stabilization, the Multiplier Method. Wiley/Masson, Paris (1994)
16. Lasiecka, I., Triggiani, R.: Control Theory for Partial Differential Equations: Continuous and Approximation Theories. I. Abstract Parabolic Systems. Encyclopedia of Mathematics and its Applications, vol. 74. Cambridge University Press, Cambridge (2000)
17. Lebeau, G., Robbiano, L.: Contrôle exact de l'équation de la chaleur. *Commun. Partial Differ. Equ.* **20**, 335–356 (1995)
18. Li, X., Yong, J.: Optimal Control Theory for Infinite-Dimensional Systems. Systems & Control: Foundations & Applications. Birkhäuser, Boston (1995)
19. Lions, J.-L.: Exact controllability, stabilization and perturbations for distributed systems. *SIAM Rev.* **30**, 1–68 (1988)
20. Lions, J.-L.: Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. *Recherches en Mathématiques Appliquées*, 8, Tome 1. Masson, Paris (1988)
21. Lions, J.-L., Magenes, E.: Problèmes aux Limites non Homogènes et Applications. *Travaux et Recherches Mathématiques*, vol. 1, no. 17. Dunod, Paris (1968)
22. Micu, S., Zuazua, E.: Regularity issues for the null-controllability of the linear 1-d heat equation. *Syst. Control Lett.* **60**(6), 406–413 (2011)
23. Rebarber, R., Weiss, G.: Necessary conditions for exact controllability with a finite-dimensional input space. *Syst. Cont. Lett.* **40**, 217–227 (2000)
24. Russell, D.L.: Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions. *SIAM Rev.* **20**(4), 639–739 (1978)
25. Staffans, O.: Well-Posed Linear Systems. Encyclopedia of Mathematics and its Applications, vol. 103, xviii+776 pp. Cambridge University Press, Cambridge (2005)
26. Trélat, E., Wang, G., Xu, Y.: Characterization by observability inequalities of controllability and stabilization properties. *Pure Appl. Anal.* **2**(1), 93–122 (2020)
27. Triggiani, R.: Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators. *SIAM J. Control Optim.* **14**(2), 313–338 (1976)
28. Tucsnak, M., Weiss, G.: Observation and Control for Operator Semigroups. Birkhäuser Advanced Texts, xii+483 pp. Birkhäuser Verlag, Basel (2009)
29. Weiss, G.: Admissible observation operators for linear semigroups. *Isr. J. Math.* **65**(1), 17–43 (1989)
30. Weiss, G.: Admissibility of unbounded control operators. *SIAM J. Control Optim.* **27**(3), 527–545 (1989)
31. Zabczyk, J.: Mathematical Control Theory: An Introduction. Systems & Control: Foundations & Applications. Birkhäuser Boston, Inc., Boston (1992)