Control in finite and infinite dimension

Emmanuel Trélat, notes de cours de M2

May 8, 2019

Contents

Part I

Control in finite dimension

Chapter 1

Controllability

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

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

where $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ 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 Ω 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. Note that this generalized Cauchy-Lipschitz theorem is not exactly the usual one since the dynamics here can be discontinuous (because of the control). For a general version of this existence and uniqueness theorem, we refer to [60, 62]. We stress that the differential equation (1.1) holds for almost every t of 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 whenever the corresponding trajectory $x(\cdot)$, such that $x(0) = x_0$, is well defined over the whole interval $[0, T]$.

We say that the control system is linear whenever $f(t, x, u) = A(t)x + B(t)u + r(t)$, with $A(t)$ a $n \times n$ matrix, $B(t)$ a $n \times m$ matrix, $r(t) \in \mathbb{R}^n$, 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^∞ on every compact interval (actually, L^1 would be enough). The linear control system is said to be instationary whenever the matrices A and B depend on t, and is said to be autonomous if $A(t) \equiv A$ and $B(t) \equiv B$. Note that, for linear control systems, there is no blow-up in finite time (i.e., admissibility holds true on any interval).

Definition 1. *Let $x_0 \in \mathbb{R}^n$ and $T > 0$. The accessible set $\mathrm{Acc}_\Omega(x_0, T)$ from x_0 in time T is defined as the set of all points $x(T)$, where $x(\cdot)$ is a solution of the control system (1.1), with $x(0) = x_0$, associated with some control u, and this set is obtained by considering all possible (admissible) controls $u \in L^\infty(0, T; \Omega)$.*

The control system (1.1) *is said to be controllable from* x_0 *in time* T *whenever* $\mathrm{Acc}_\Omega(x_0, T) = \mathbb{R}^n$.

Of course, other variant notions 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

?⟨sec_cont_linear⟩?

1.1.1 Controllability of autonomous linear systems

?⟨sec_cont_autonomous⟩? In this section, we assume that $A(t) \equiv A$ and $B(t) \equiv B$.

Case without control constraints: Kalman condition

?⟨sec_cont_kalman⟩? The famous Kalman theorem provides a necessary and sufficient condition for autonomous linear control systems without control constraint.

⟨controlabilite⟩ **Theorem 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*) if and only if the Kalman matrix*

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

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

Remark 1. The above Kalman condition does neither depend on T, nor on x_0. This implies that if an autonomous linear control system is controllable in time T, starting at x_0, then it is as well controllable in any time, starting at any point. Note that the Kalman condition is purely algebraic and is easily checkable.

Proof. Let $x_0 \in \mathbb{R}^n$ and $T > 0$ arbitrary. For every $u \in L^\infty(0, T; \mathbb{R}^m)$, the unique solution $x(\cdot)$ of the system, associated with the control u and starting at x_0, satisfies

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

where $L_T : L^\infty(0, T; \mathbb{R}^m) \to \mathbb{R}^n$ is the linear continuous operator defined by

$$L_T u = \int_0^T e^{(T-t)A}Bu(t)\,dt.$$

In terms of this operator, it is clear that 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. *The Kalman matrix $K(A,B)$ is of rank n if and only if L_T is surjective.*

Proof of the lemma. 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,

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

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. □

The theorem is proved. □

It can be noticed that, if a system is controllable in the conditions above, then it can be controlled in arbitrarily small time. This is due to the fact that there are no control constraints. In case of control constraints we cannot hope that L_T be surjective in general.

Remark 2 (Hautus test)**.** The following assertions are equivalent:

(1) The couple (A,B) satisfies Kalman's condition $\operatorname{rank}(K(A,B)) = n$.

(2) $\forall \lambda \in \mathbb{C} \quad \operatorname{rg}(\lambda I - A, B) = n$.

(3) $\forall \lambda \in \operatorname{Spec}(A) \quad \operatorname{rg}(\lambda I - A, B) = n$.

(4) $\forall z$ eigenvector of $A^\top, \quad B^\top z \neq 0$.

(5) $\exists c > 0 \ \mid \ \forall \lambda \in \mathbb{C} \quad \forall z \in \mathbb{R}^n \quad \|(\lambda I - A^\top) z\|^2 + \|B^\top z\|^2 \geqslant c\|z\|^2$.

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.

Case with control constraints

?⟨sec_cont_constraints⟩? An easy adaptation of the proof of Theorem 1 yields the following result.

Proposition 1. *We assume that $r = 0$, that $0 \in \overset{\circ}{\Omega}$, and that the Kalman condition holds true. Then, for every $t > 0$, the accessible set $Acc_\Omega(x_0, t)$ contains a neighborhood of the point $e^{tA}x_0$.*

Global controllability properties can be obtained under strong stability assumptions. For instance we have the following easy result.

Proposition 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. Then, for every $x_0 \in \mathbb{R}^n$, 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 here just 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.

Similar systems

?⟨sec_similar⟩? Let us investigate the effect of a change of basis in linear control systems.

Definition 2. *The linear control systems $\dot{x}_1 = A_1x_1 + B_1u_1$ and $\dot{x}_2 = A_2x_2 + B_2u_2$ are said to be similar whenever there exists $P \in GL_n(\mathbb{R})$ such that $A_2 = PA_1P^{-1}$ and $B_2 = PB_1$. We have then $x_2 = Px_1$ and $u_2 = u_1$.*

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

Remark 3. The Kalman property is instrinsic, that is

$$(B_2, A_2B_2, \dots, A_2^{n-1}B_2) = P(B_1, A_1B_1, \dots, A_1^{n-1}B_1),$$

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

Proposition 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 is the rank of $K(A, B)$ and is as well the rank of $K(A'_1, B'_1)$.

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

$$\begin{aligned} \dot{y}_1 &= A'_1y_1 + B'_1u + A'_3y_2 \\ \dot{y}_2 &= A'_2y_2 \end{aligned}$$

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 uncontrolled and is called the *uncontrollable part* of the system.

Proof. We assume that the rank of $K(A,B)$ is less than n (otherwise there is nothing to prove). The subspace $F = \text{Range } K(A,B) = \text{Range } B + \text{Range } AB + \cdots + \text{Range } A^{n-1}B$ 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 $(f_1, \ldots, f_r)$ be a basis of F and $(f_{r+1}, \ldots, f_n)$ be a basis of G. Let P be the change-of-basis matrix from the basis $(f_1, \ldots, f_n)$ to the canonical basis of $\mathbb{R}^n$. Since F is invariant under A, we get

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

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

⟨thmbrunovski⟩ **Theorem 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). Then the pair (A,B) is similar to the pair $(\tilde{A}, \tilde{B})$, with*

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

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 a companion matrix. This result means that, in the new coordinates, the system is equivalent to the order n scalar differential equation with scalar control $x^{(n)}(t) + a_1 x^{(n-1)}(t) + \cdots + a_n x(t) = u(t)$.

Proof. First of all, let us notice that, if there exists a basis $(f_1, \ldots, f_n)$ in which the pair (A,B) takes the form $(\tilde{A}, \tilde{B})$, then there must hold $f_n = B$ (up to multiplying scalar) and

$$Af_n = f_{n-1} - a_1 f_n, \ \ldots, \ Af_2 = f_1 - a_{n-1}f_n, \ Af_1 = -a_n f_n.$$

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

$$f_n = B, \ f_{n-1} = Af_n + a_1 f_n, \ \ldots, \ f_1 = Af_2 + a_{n-1}f_n.$$

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

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

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

$$\begin{aligned} Af_1 &= A^2 f_2 + a_{n-1} A f_n \\ &= A^2(Af_3 + a_{n-2} f_n) + a_{n-1} A f_n \\ &= A^3 f_3 + a_{n-2} A^2 f_n + a_{n-1} A f_n \\ &\vdots \\ &= A^n f_n + a_1 A^{n-1} f_n + \cdots + a_{n-1} A f_n \\ &= -a_n f_n \end{aligned}$$

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

Remark 4. This theorem can be generalized to the case $m > 1$ but the normal form is not that simple. More precisely, if the pair (A, B) satisfies the Kalman condition, then it is similar to some pair $(\tilde{A}, \tilde{B})$ such that

$$\tilde{A} = \begin{pmatrix} \tilde{A}_1 & * & \cdots & * \\ 0 & \tilde{A}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \tilde{A}_s \end{pmatrix} \quad \text{and} \quad \tilde{B}G = \begin{pmatrix} \tilde{B}_1 \\ \vdots \\ \tilde{B}_s \end{pmatrix}$$

where the matrices $\tilde{A}_i$ are companion matrices, G is a matrix of size $m \times s$, and for every $i \in \{1, \ldots, s\}$, all coefficients of $\tilde{B}_i$ are equal to zero except the one in the last row, column i, which is equal to 1.

1.1.2 Controllability of time-varying linear systems

?⟨sec_controllability_time-varying⟩? In what follows, we denote by $M(\cdot)$ the *resolvent* of the linear system $\dot{x}(t) = A(t)x(t)$, that is, the unique solution of the Cauchy problem $\dot{M}(t) = A(t)M(t)$, $M(0) = I_n$. Note that, in the autonomous case $A(t) \equiv A$, we have $M(t) = e^{tA}$. But in general the resolvent cannot be computed explicitly.

⟨controlabiliteinstationnaire⟩ **Theorem 3.** *We assume that $\Omega = \mathbb{R}^m$ (no constraint 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*

$$G_T = \int_0^T M(t)^{-1} B(t) B(t)^\top M(t)^{-1^\top} dt$$

is invertible.

Remark 5. Note that this condition depends on T but not on the initial point. In other words, if a linear instationary control system is controllable in time T, from the initial point x_0, then it is controllable as well from any other initial point (but with the same time T).

⟨remGT⟩ **Remark 6.** Note that $G_T = G_T^\top$, and that

$$\psi^\top G_T \psi = \langle G_T\psi, \psi\rangle = \int_0^T \|B(t)^\top M(t)^{-1^\top}\psi\|^2\,dt \geqslant 0 \qquad \forall \psi \in \mathbb{R}^n$$

i.e., G_T is a symmetric nonnegative matrix. The above controllability condition says that the system is controllable if and only if G_T is positive definite.

By a diagonalization argument, this is equivalent to say that there exists $C_T > 0$ (which is the smallest eigenvalue of G_T) such that

$$\int_0^T \|B(t)^\top M(t)^{-1^\top}\psi\|^2\,dt \geqslant C_T\|\psi\|^2 \qquad \forall \psi \in \mathbb{R}^n. \tag{1.2}$$ inegobsdimfinie

This is an *observability inequality*.

Although we will not develop here a general theory for observability, it can be noticed that this condition is appropriate to be generalized in the infinite-dimensional setting, and will be of instrumental importance in the derivation of the so-called HUM method (see Part II).

Although we have not defined the concept of observability, we underline that the system is controllable in time T if and only if the above observability inequality holds: this is the well known *controllability – observability duality*.

Proof. 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

$$x(T) = x^* + M(T)\int_0^T M(t)^{-1}B(t)u(t)\,dt$$

with $x^* = M(T)x_0 + M(T)\int_0^T M(t)^{-1}r(t)\,dt$.

Let us first assume that the Gramian G_T is invertible. let $X_1 \in \mathbb{R}^n$ be any target point. Let us search an appropriate control u in the form $u(t) = B(t)^\top M(t)^{-1^\top}\psi$, with $\psi \in \mathbb{R}^n$ to be determined adequately so that $x(T) = x_1$. With such a control, we have $x(T) = x^* + M(T)G_T\psi$, and since G_T is invertible it suffices to take $\psi = G_T^{-1}M(T)^{-1}(x_1 - x^*)$.

Conversely, if G_T is not invertible, then according to Remark 6 there exists $\psi \in \mathbb{R}^n \setminus \{0\}$ such that $\psi^\top G_T\psi = 0$, and therefore, $\int_0^T \|B(t)^\top M(t)^{-1^\top}\psi\|^2 dt = 0$, from which we infer that $\psi^\top M(t)^{-1}B(t) = 0$ for almost every $t \in [0,T]$. As a consequence, we get that $\psi^\top \int_0^T M(t)^{-1}B(t)u(t)\,dt = 0$, for every control $u \in L^\infty(0,T;\mathbb{R}^m)$, and hence $\psi^\top M(T)^{-1}(x_u(T) - x^*) = 0$, which means that $x_u(T)$ belongs to some proper affine subspace of $\mathbb{R}^n$ as u varies. Hence the system is not controllable. □

Remark 7. This theorem can be proved in an easier and more natural way with the Pontryagin maximum principle (with an optimal control viewpoint). 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 8. If the system is autonomous ($A(t) \equiv A$, $B(t) \equiv B$) then $M(t) = e^{tA}$ and thus

$$G_T = \int_0^T e^{-sA} B B^\top e^{-sA^\top} ds.$$

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 is not true anymore in the instationary case.

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

⟨controlabiliteinstationnaire2⟩ **Theorem 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^∞. We define the sequence of matrices

$$B_0(t) = B(t), \quad B_{i+1}(t) = A(t)B_i(t) - \frac{dB_i}{dt}(t), \ i \in \mathbb{N}.$$

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

$$\text{Span } \{B_i(t)v \mid v \in \mathbb{R}^m, \ i \in \mathbb{N}\} = \mathbb{R}^n \tag{1.3}$$ kalman_time

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 (1.3) holds true for every $t \in [0, T]$.*

We do not prove this theorem now. The proof readily follows from the Hamiltonian characterization of singular trajectories (see [7, 62], see also the proof of the weak Pontryagin Maximum Principle in Section 2.2).

There are no simple statements ensuring the controllability of instationary control systems under control constraint. Actually under control constraints one can draw a clear picture of controllability properties by investigating the topology of the accessible set, as studied next.

1.1.3 Topology of accessible sets

?⟨sec_topology⟩? **Theorem 5.** *Consider the control system in $\mathbb{R}^n$*

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

with controls u taking their values in a compact subset Ω of $\mathbb{R}^m$. Let $x_0 \in \mathbb{R}^n$. For every $t \geqslant 0$, the accessible set $\mathrm{Acc}_\Omega(x_0, t)$ is compact, convex and varies continuously with respect to t.

Here, the continuity in time is established for the Hausdorff topology.

Remark 9. Note that the convexity of the accessible set holds true even though Ω is not assumed to be convex. This point is however not obvious and follows from a Lyapunov lemma (itself based on the Krein-Milman theorem in infinite dimension; see [30]). Actually this argument leads to $\mathrm{Acc}_\Omega(x_0,t) = \mathrm{Acc}_{\mathrm{Conv}(\Omega)}(x_0,t)$, where $\mathrm{Conv}(\Omega)$ is the convex closure of Ω. In particular we have $\mathrm{Acc}_{\partial\Omega}(x_0,t) = \mathrm{Acc}_\Omega(x_0,t)$, where $\partial\Omega$ is the boundary of Ω. This result illustrates the so-called *bang-bang principle* (see Section 2.4.1).

Proof. We first assume that Ω is convex. In that case, we notice that

$$\mathrm{Acc}_\Omega(x_0,t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}r(s)\,ds + L_t(L^\infty(0,t;\Omega))$$

where the linear (and continuous) operator $L_t : L^\infty(0,t;\mathbb{R}^m) \to \mathbb{R}^n$ is defined by

$$L_t u = \int_0^t M(t)M(s)^{-1}B(s)u(s)\,ds.$$

The convexity of $\mathrm{Acc}_\Omega(x_0,t)$ follows by linearity from the convexity of $L^\infty(0,t;\Omega)$.

Let us now prove the compactness of $\mathrm{Acc}_\Omega(x_0,t)$. Let $(x_n^1)_{n\in\mathbb{N}}$ be a sequence of points of $\mathrm{Acc}_\Omega(x_0,t)$. 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

$$x_n^1 = x_n(t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}(B(s)u_n(s) + r(s))\,ds. \qquad (1.4)$$ eqdem1

Since Ω 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 [11]), 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 Ω 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 Ω is compact). Besides, using (1.4) 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 $[0,t]$. Passing to the limit in (1.4), we get

$$x(t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}(B(s)u(s) + r(s))\,ds$$

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

Let us prove the continuity in time of $\mathrm{Acc}_\Omega(x_0,t)$. Let $\varepsilon > 0$ arbitrary. We are going to establish that there exists $\delta > 0$ such that, for all nonnegative real numbers t_1 and t_2, if $|t_1 - t_2| \leqslant \delta$ then $d_H(\mathrm{Acc}_\Omega(x_0,t_1), \mathrm{Acc}_\Omega(x_0,t_2)) \leqslant \varepsilon$, where d_H is the Hausdorff distance, defined by

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

for all compact subsets K_1 and K_2 of $\mathbb{R}^n$, and d is the Euclidean distance of $\mathbb{R}^n$. We assume that $0 \leqslant t_1 < t_2$. It suffices to prove that

1. $\forall y \in \mathrm{Acc}_\Omega(x_0, t_2) \quad d(y, \mathrm{Acc}_\Omega(x_0, t_1)) \leqslant \varepsilon$,
2. $\forall y \in \mathrm{Acc}_\Omega(x_0, t_1) \quad d(y, \mathrm{Acc}_\Omega(x_0, t_2)) \leqslant \varepsilon$.

Let us just prove the first point, the second being similar. Let $y \in \mathrm{Acc}_\Omega(x_0, t_2)$. It suffices to prove that there exists $z \in \mathrm{Acc}_\Omega(x_0, t_1)$ such that $d(y, z) \leqslant \varepsilon$. By definition of $\mathrm{Acc}_\Omega(x_0, t_2)$, there exists $u \in L^\infty(0, T; \Omega)$ such that the corresponding trajectory, starting at x_0, satisfies $x(t_2) = y$. Let us prove that $z = x(t_1)$ is suitable. Indeed, we have

$$\begin{aligned} x(t_2) - x(t_1) = M(t_2)x_0 &+ \int_0^{t_2} M(t_2)M(s)^{-1}(B(s)u(s) + r(s))\,ds \\ &- M(t_1)x_0 - \int_0^{t_1} M(t_1)M(s)^{-1}(B(s)u(s) + r(s))\,ds \\ = \; & M(t_2)\int_{t_1}^{t_2} M(s)^{-1}(B(s)u(s) + r(s))\,ds \\ &+ (M(t_2) - M(t_1))\left(x_0 + \int_0^{t_1} M(s)^{-1}(B(s)u(s) + r(s))\,ds\right) \end{aligned}$$

If $|t_1 - t_2|$ is small then the first term of the above sum is small as well by continuity, and the second term is small by continuity of $t \mapsto M(t)$. The result follows.

In the general case where Ω is only compact (but not necessarily convex), the proof is more difficult and uses the Lyapunov lemma in measure theory (see, e.g., [45, Lemma 4A p. 163]) and more generally the Aumann theorem (see, e.g., [30]), from which it follows that

$$\begin{aligned} &\left\{\int_0^T M(t)^{-1}B(t)u(t)\,dt \;\middle|\; u \in L^\infty(0, T; \Omega)\right\} \\ =&\left\{\int_0^T M(t)^{-1}B(t)u(t)\,dt \;\middle|\; u \in L^\infty(0, T; \partial\Omega)\right\} \\ =&\left\{\int_0^T M(t)^{-1}B(t)u(t)\,dt \;\middle|\; u \in L^\infty(0, T; \mathrm{Conv}(\Omega))\right\} \end{aligned}$$

and moreover that these sets are compact and convex. The result follows. □

Remark 10. If the set of control constraints Ω is compact, then the accessible is compact (and convex), and evolves continuously in time. In such conditions obviously $\mathrm{Acc}_\Omega(x_0, t)$ can never be equal to $\mathbb{R}^n$. In other words, the system is never controllable in time t. This is natural in view of the control constraints. Actually the result of the theorem allows one to define the concept of *minimal*

time: given x_0 and x_1, two distinct points of $\mathbb{R}^n$, one cannot steer the control system from x_0 to x_1 in arbitrary small time. A minimal time is due. We will come back on this issue later.

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 $\mathrm{Acc}_\Omega(x_0, t)$, over $t \geqslant 0$, equal to the whole $\mathbb{R}^n$? This question is difficult, even for (instationary) linear control systems under control constraints.

1.2 Controllability of nonlinear systems

⟨sec_nonlinear⟩

1.2.1 Local controllability results

Preliminaries: end-point mapping.

Definition 3. *Let $x_0 \in I\!R^n$ and let $T > 0$ arbitrary. A control $u \in L^\infty(0, T; I\!R^m)$ is said to be admissible on $[0, T]$ if the trajectory $x(\cdot)$, solution of (1.1), corresponding to the control u, and such that $x(0) = x_0$, is well defined on $[0, T]$. The end-point mapping $E_{x_0,T}$ is then defined by $E_{x_0,T}(u) = x(T)$.*

The set of admissible controls on $[0, T]$ is denoted by $\mathcal{U}_{x_0,T,\mathbb{R}^m}$, and the set of admissible controls on $[0, T]$ taking their values in Ω is denoted by $\mathcal{U}_{x_0,T,\Omega}$. The set $\mathcal{U}_{x_0,T,\mathbb{R}^m}$ is nothing else but the set on which $E_{x_0,T}$ is well defined (indeed one has to be careful with blow-up phenomena). It is easy[1] to prove 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 $E_{x_0,T}$ 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 $t \geqslant 0$, the accessible set is $A_{x_0,\Omega}(t) = E_{x_0,t}(\mathcal{U}_{x_0,t,\Omega})$.

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

⟨diffFrechetE⟩

Theorem 6. *Let $x_0 \in I\!R^n$ and let $u \in \mathcal{U}_{x_0,T,I\!R^m}$. The (Fréchet) differential $dE_{x_0,T}(u) : L^\infty(0, T; I\!R^m) \to I\!R^n$ is given by*

$$dE_{x_0,T}(u).\delta u = \delta x(T) = M(T) \int_0^T M^{-1}(t) B(t) \delta u(t)\, dt$$

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 $M(\cdot)$ is the resolvent of the linearized system, defined as the unique $n \times n$ matrix solution of $\dot{M}(t) = A(t)M(t)$, $M(0) = I_n$.

[1] This follows from usual finite-time blow-up arguments on ordinary differential equations, and from the usual Cauchy-Lipschitz theorem with parameters, the parameter being here a control in a Banach set (see for instance [60]).

Proof. By definition, we have $E_{x_0,T}(u+\delta u) = E_{x_0,T}(u) + dE_{x_0,T}(u).\delta u + \mathrm{o}(\delta u)$. In this first-order Taylor expansion, we have $E_{x_0,T}(u) = x_u(T)$ 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

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

so that, at the first order, we identify the linearized system. By integration (note that the remainder terms can be rigorously ruled out by standard Gronwall arguments, not detailed here), we get $\delta x(T) = M(T)\int_0^T M^{-1}(t)B(t)\delta u(t)\,dt$, as expected. Note that this term provides a linear continuous operator and hence is indeed the Fréchet differential of the end-point mapping. □

⟨remark_loccont⟩ **Remark 11.** 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 $u \in \mathcal{U}_{x_0,T,\mathbb{R}^m}$ be arbitrary. According to Remark 11, if the linearized system along $(x_u(\cdot), u(\cdot))$ is controllable in time T, then this exactly means that the end-point mapping of the linearized system is surjective. In other words, the linear continuous mapping $dE_{x_0,T}(u) : L^\infty(0,T;\mathbb{R}^m) \to \mathbb{R}^n$ is surjective. By a standard implicit function argument (surjective mapping theorem), this implies that the end-point mapping itself, $E_{x_0,T}(u) : \mathcal{U}_{x_0,T,\mathbb{R}^m} \to \mathbb{R}^n$, is a *local submersion*, and thus, in particular, is locally surjective and locally open at u.

Note that, here, this 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 Ω of control constraints, provided that we have room to consider local variations of u: this is true as soon as 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 u takes its values in the interior of Ω).

The local surjectivity of $E_{x_0,T}$ exactly means that the general control system (1.1) is locally controllable at $x_u(T)$. We have proved the following result.

⟨thm_localcont_traj⟩ **Theorem 7.** *Let $x_0 \in \mathbb{R}^n$ and let $u \in \mathcal{U}_{x_0,T,\Omega}$. We assume that u 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 $(x_u(\cdot), u(\cdot))$ is controllable in time T, then the nonlinear control system (1.1) is locally controllable in time T from x_0. In other words, there exists a neighborhood V of $x_u(T)$ in $\mathbb{R}^n$ such that, for every $y \in V$, there exists a control $v \in \mathcal{U}_{x_0,T,\Omega}$ such that $y = x_v(T)$.*

Note that, in the above statement, v is close to u for the L^∞ topology, and the trajectories $x_v(\cdot)$ and $x_u(\cdot)$ are close for the C^0 topology. The theorem would actually say more, because $E_{x_0,T}$ is a submersion, meaning that, for some appropriate coordinates (in a chart), $E_{x_0,T}$ is a linear projection.

Remark 12. The controllability in time T of the linearized system $\delta\dot{x}(t) = A(t)\delta x(t) + B(t)\delta u(t)$ can be characterized thanks to Theorems 3 and 4. We thus get explicit sufficient conditions for local controllability. Note that the conditions are not necessary (for instance, in $1D$, $\dot{x}(t) = u(t)^3$, along $u = 0$).

Let us next provide two important applications of Theorem 7 (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*, that is, 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 $x(t) = \bar{x}$ and $u(t) = \bar{u}$ 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 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.

⟨cor_loc_cont⟩ **Corollary 1.** *With the above notations, assume that* $\operatorname{rank} K(A, B) = n$ *and that* $\bar{u} \in \operatorname{Int}(\Omega)$. *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 exists a neighborhood* V *of* $\bar{x}$ *in* $\mathbb{R}^n$ *such that, for all* $x_0, x_1 \in V$, *there exists a control* $u \in \mathcal{U}_{\bar{x},T,\Omega}$ *such that* $x_u(0) = x_0$ *and* $x_u(T) = x_1$.

Many applications of this result can be found in the literature. It can be noted that the proof, based on the implicit function theorem, is robust and withstands 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, 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 equibrium 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[2] 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 7 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 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 $\bar{u}$ 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.*

Example 1. Consider the Dubins car model

$$\begin{aligned} \dot{x}_1(t) &= \cos\theta(t), & x_1(0) &= 0,\\ \dot{x}_2(t) &= \sin\theta(t), & x_2(0) &= 0,\\ \dot{\theta}(t) &= u(t), & \theta(0) &= 0\ [2\pi]. \end{aligned}$$

In order to prove that this control system is locally controllable at $(0,0,0\ [2\pi])$ (in arbitrary time $T>0$), we 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 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 4) that this system is controllable in any time $T>0$.

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 Topology 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 5.

[2]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 [17].

⟨ensacccompact⟩ **Theorem 8.** *Let $x_0 \in \mathbb{R}^n$ and let $T > 0$. We assume that:*

- *Ω is compact;*
- *there exists $b > 0$ such that, for every admissible control $u \in \mathcal{U}_{x_0,T,\Omega}$, one has $\|x_u(t)\| \leqslant b$ for every $t \in [0,T]$;*
- *there exists $c > 0$ such that $\|f(t,x,u)\| \leqslant c$ for every $t \in [0,T]$, for every $x \in \mathbb{R}^n$ such that $\|x\| \leqslant b$ 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 $\mathrm{Acc}_\Omega(x_0,t)$ is compact and varies continuously in time on $[0,T]$.

Remark 13. 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 Ω is moreover convex.

Proof. First of all, note that $V(t,x)$ is compact, for all (t,x), because Ω is compact. Let us prove that $Acc(x_0,t)$ is compact for every $t \in [0,T]$. It suffices to prove that every sequence (x_n) of points of $Acc(x_0,t)$ 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 [11]). 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 $\lim_{n\to+\infty} x_n(s) = x(s)$ 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 Ω, 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 [11]). 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. For every $\varphi \in L^2(0,t;\mathbb{R})$, we have

$$\int_0^t \varphi(s)g_n(s)ds = \int_0^t \varphi(s)h_n(s)\,ds + \int_0^t \varphi(s)\left(g_n(s) - h_n(s)\right)\,ds. \qquad (1.5)$$ temp1

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 $\|g_n(s) - h_n(s)\| \leqslant C\|x_n(s) - x(s)\|$ 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 $\int_0^t \varphi(s)\left(g_n(s) - h_n(s)\right)\,ds \to 0$ as $n \to +\infty$. Passing to the limit in (1.5), 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 [45, Lemmas 2A, 3A p. 161]).

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

It remains to establish the continuity of the accessible set with respect to time. Let t_1 and t_2 be two real numbers such that $0 < t_1 < t_2 \leqslant T$ and let $x_2 \in \mathrm{Acc}_\Omega(x_0,t_2)$. By definition, there exists a control u taking its values in Ω, 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 $\mathrm{Acc}_\Omega(x_0,t_1)$, and using the assumptions on f, we have $\|x_2 - x_1\| \leqslant C|t_2 - t_1|$. We conclude easily. □

1.2.3 Global controllability results

One can find global controllability results in the existing literature, but they are established for particular classes of control systems. Let us provide, here, controllability results valuable for 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 $f_i : \mathbb{R}^n \to \mathbb{R}^n$, $i = 0,\ldots,m$ 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., $\exp(t_1X)\circ\exp(t_2Y) = \exp(t_2Y)\circ\exp(t_1X)$ 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

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

as $t \to 0$. 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 $Z = X + Y + \frac{1}{2}[X,Y] + \cdots$.

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 [15, 55].

Theorem 9. *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 $f_1,\ldots,f_m$ is equal to $\mathbb{R}^n$ (at any point). Then the system is globally controllable, in any time T.*

Proof. We sketch the proof in the case $n = 3$ and $m = 2$, assuming that $\mathrm{rank}(f_1, f_2, [f_1,f_2]) = 3$ 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, φ_λ is a local diffeomorphism at 0. From the Campbell-Hausdorff formula, we infer that

$$\varphi_\lambda(t_1,t_2,t_3) = \exp(t_1 f_1 + (t_2+t_3)f_2 + \lambda t_3[f_1,f_2] + \cdots),$$

whence

$$\frac{\varphi_\lambda}{\partial t_1}(0) = f_1(x_0),\ \frac{\varphi_\lambda}{\partial t_2}(0) = f_2(x_0),\ \frac{\varphi_\lambda}{\partial t_3}(0) = f_2(x_0) + \lambda[f_1, f_2](x_0) + \mathrm{o}(\lambda).$$

By assumption, it follows that $d\varphi_\lambda$ is an isomorphism, and therefore φ_λ is a local diffeomorphism at 0. We conclude by an easy connectedness argument. □

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 [36] 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. 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 $f_1, \ldots, f_m$ be equal to $\mathbb{R}^n$ (at any point). We also refer the reader to [17] where there is a good survey of controllability results for control-affine systems.

Example 2. 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

$$f_1 = \frac{\partial}{\partial x} + y\frac{\partial}{\partial z}, \quad f_2 = \frac{\partial}{\partial y} - x\frac{\partial}{\partial z}.$$

It is easy to check that $[f_1, f_2] = -2\frac{\partial}{\partial z}$ and thus that the Lie algebra condition is satisfied. Therefore this system is controllable.

Remark 14. The Lie condition above is also called the Hörmander condition, due to the well-known following result due to Hörmander (see [31]). Let $L = \sum_{i=1}^m f_i^2$ be an operator, defined as a sum of squares of vector fields: this is a second-order differential operator, which can be called a *sub-Laplacian*.[3] We say that L is *hypoelliptic* if g is C^∞ whenever Lg is C^∞. Note that the usual Laplacian is hypoelliptic, but the question is not obvious if $m < n$. The famous result by Hörmander states that L is hypoelliptic if and only if the Lie algebra generated by the vector fields $f_1, \ldots, f_m$ is equal to $\mathbb{R}^n$ (at any point).

Here, we recover an idea of controllability because, in order to prove that if Lg is C^∞ then g is C^∞, we have to "control" the derivatives of g in any possible direction of $\mathbb{R}^n$.

[3] If $m = n$ and if the f_i coincide with the canonical basis of $\mathbb{R}^n$ then L is exactly the usual Laplacian. We speak of sub-Laplacian when $m < n$.

Chapter 2

Optimal control

⟨chap_opt⟩ In Chapter 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 $\mathbb{R}^n$

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

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

Let $f^0 : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ be functions of class C^1. For every $x_0 \in \mathbb{R}^n$, for every $t_f \geqslant 0$, and for every admissible control $u \in \mathcal{U}_{x_0,t_f,\Omega}$ (see Section 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

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

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 us now define the optimal control problem that we will consider. 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

$$x(0) \in M_0, \quad x(t_f) \in M_1,$$

and minimizing the cost (2.2) over all possible trajectories steering as well the control system from M_0 to M_1.

Note that this is a general nonlinear optimal control problem, but without any state constraints. We could indeed restrict the set of trajectories by imposing some pointwise constraints on $x(t)$ (typically, such or such region of the state space can be forbidden). Such constraints are however not easily tractable in the Pontryagin maximum principle and make its analysis much more difficult. We will however comment further on this issue, but for simplicity we will mainly ignore state constraints.

2.1 Existence of an optimal control

Although this 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 also no additional difficulty in adding some state constraints.

⟨thmfilippov⟩ **Theorem 10.** *We consider* (**OCP**) *and we assume that:*

- Ω *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 exists* $b > 0$ *such that, for every trajectory* $x(\cdot)$ *defined on* $[0, t_f]$ *and steering the system from* M_0 *to* M_1*, one has* $t_f + \|x(t)\| \leqslant b$ *for every* $t \in [0, t_f]$*;*
- *there exists* $c > 0$ *such that* $\|f(t, x, u)\| + |f^0(t, x, u)| \leqslant c$*, for every* $t \in [0, b]$*, every* $x \in \mathbb{R}^n$ *such that* $\|x\| \leqslant b$*, and every* $u \in \Omega$*;*
- *the epigraph of extended velocities*

$$\tilde{V}(t, x) = \left\{ \begin{pmatrix} f(t, x, u) \\ f^0(t, x, u) + \gamma \end{pmatrix} \;\middle|\; u \in \Omega,\ \gamma \geqslant 0 \right\} \tag{2.3}$$ optconv

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

We assume moreover in (**OCP**) *that the trajectories are subject to state constraints* $c_i(t, x(t)) \leqslant 0$*, 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 has been 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 and M_1 could depend on t, as well as the set Ω (see [45], and more generally see [14] for many variants of existence results).

Such existence results are however often difficult to apply in practice, because in particular the assumption (2.3) is a strong one (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. The proof is similar to the one of Theorem 8.

Let δ 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 $c_i(x(\cdot)) \leqslant 0$, $i = 1, \dots, r$. 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 $C(u_n) \to \delta$ as $n \to +\infty$. 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 $\tilde{F}(\cdot) = \begin{pmatrix} F(\cdot) \\ F^0(\cdot) \end{pmatrix}$ for the weak star topology of $L^\infty(0, b; \mathbb{R}^{n+1})$ (see [11]). Also, up to some subsequence, the sequence $(t_n(u_n))_{n \in \mathbb{N}}$ converges to some $T \geqslant 0$, 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 $\lim_{n \to +\infty} x_n(t) = x(t)$, that is, the sequence $(x_n(\cdot))_{n \in \mathbb{N}}$ converges pointwise to $x(\cdot)$. As in the proof of Theorem 8, the objective is then to prove that the trajectory $x(\cdot)$ is associated with a control u taking its values in Ω, and that this control is moreover optimal.

We set $\tilde{h}_n(t) = \begin{pmatrix} f(t, x(t), u_n(t)) \\ f^0(t, x(t), u_n(t)) \end{pmatrix}$, 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 $\tilde{h}_n(t) = \begin{pmatrix} f(t, x(t), v) \\ f^0(t, x(t), v) \end{pmatrix}$, for some arbitrary $v \in \Omega$. Besides, we define (note that Ω is

compact)

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

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

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

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

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

Proof of Lemma 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 $t \in [0,T]$. □

The set $\tilde{\mathcal{V}}$ is then also convex and weakly closed in $L^2(0,T;\mathbb{R}^{n+1})$ (see [11]). 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 $\tilde{h}$, which belongs to $\tilde{\mathcal{V}}$ since this set is weakly closed.

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

$$\int_0^T \varphi(t)\tilde{F}_n(t)\,dt = \int_0^T \varphi(t)\tilde{h}_n(t)\,dt + \int_0^T \varphi(t)\left(\tilde{F}_n(t) - \tilde{h}_n(t)\right)dt \qquad (2.4)$$ temp1__1

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 $\|\tilde{F}_n(t) - \tilde{h}_n(t)\| \leqslant C\|x_n(t) - x(t)\|$ 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 $\int_0^T \varphi(t)\left(\tilde{F}_n(t) - \tilde{h}_n(t)\right)dt \to 0$ as $n \to +\infty$. 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 $\gamma(t) \geqslant 0$ such that $\tilde{F}(t) = \begin{pmatrix} f(t,x(t),u(t)) \\ f^0(t,x(t),u(t)) + \gamma(t) \end{pmatrix}$. 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 [45, 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 $c_i(x(\cdot)) \leqslant 0$, $i = 1, \dots, r$. Besides, by definition $C(u_n)$ converges to δ, 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 γ takes nonnegative values, this implies that $\int_0^T f^0(t, x(t), u(t))\, dt + g(T, x(T)) \leqslant \int_0^T (f^0(t, x(t), u(t)) + \gamma(t))\, dt + g(T, x(T)) \leqslant C(v)$, for every admissible control v generating a trajectory steering the system from M_0 to M_1 and satisfying all constraints. In other words, the control u is optimal. By the way, note that γ must be equal to 0. □

2.2 Weak Pontryagin maximum principle

⟨sec_weakPMP⟩ In order to understand the main idea underlying the Pontryagin maximum principle (denoted in short PMP), we first study a simplified version of (**OCP**) and state a weaker version of the result that we call "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 is fixed. In that case we denote it rather by T.

These three first simplified assumptions are not serious, and it is actually easy to reduce a given optimal control problem to that case (we refer to [62] for details). In contrast, the following one is by far more serious:

- $\Omega = \mathbb{R}^m$, in other words, there are no control constraints.

Note that, actually, in the following arguments we could as well assume that the control under consideration is in the interior of $L^\infty(0, T; \Omega)$.

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

Let us now analyze the simplified (**OCP**) and come finally out with the weak PMP. Let u be an optimal control (here, we assume its existence, without making any further assumption on the dynamics). We are going to sketch first-order necessary conditions for optimality.

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

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

In this form, this is a nonlinear optimization problem with equality constraints. It is however in infinite dimension, because the unknown v ranges over an infinite dimensional space. Since u is optimal, we can apply the well known Lagrange

multipliers rule to this problem. The argument goes as follows. Let us consider the figure 2.1, in which we draw the image of the mapping F defined by

$$F(u) = (E_{x_0,T}(u), C_{x_0,T}(u))$$

with $E_{x_0,T}(u) \in \mathbb{R}^n$ in abscissa and $C_{x_0,T}(u) \in \mathbb{R}$ in ordinate. The image of F is thus seen as a subset of $\mathbb{R}^n \times \mathbb{R}$, whose topology has no importance. What is important to note is that we are interested in controls steering the system from x_0 to x_1, and hence on the figure they correspond to a point which is in the image of F, vertically above the point x_1. Now the optimal control u corresponds on the figure to the point $F(u)$, which is vertically above x_1 and which is at the boundary of the image of F. In other words, the fundamental property is:

$$u \text{ optimal } \Rightarrow \ F(u) \in \partial F(L^\infty).$$

Indeed, if $F(u)$ were not at the boundary of $F(L^\infty)$ 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.

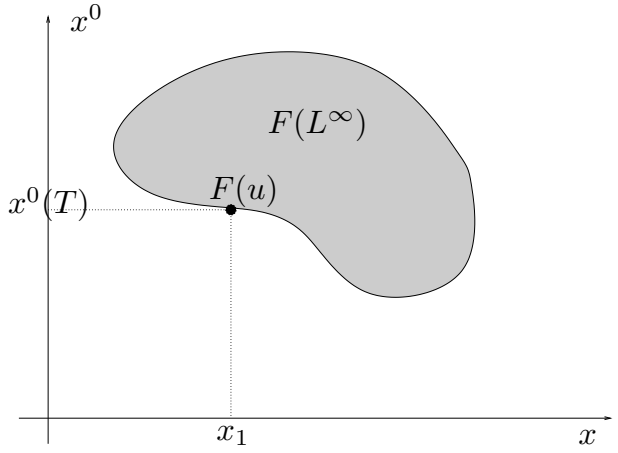

Figure 2.1: Image of the mapping F

⟨ensaccaug⟩

At this step, we use the important simplification $\Omega = \mathbb{R}^m$. Since $F(u) \in \partial F(L^\infty)$, it follows from an implicit function argument (more precisely, the surjective mapping theorem) that the linear continuous mapping

$$dF(u) : L^\infty(0, T; \mathbb{R}^m) \to \mathbb{R}^n \times \mathbb{R}$$

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 $F(L^\infty)$, which would contradict the fact that $F(u) \in \partial F(L^\infty)$). In other words, $\mathrm{Ran}(dF(u))$ is a proper subspace of $\mathbb{R}^n \times \mathbb{R}$.

Note that 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. The usual proof consists of developing what are called needle-like variations (see [54]), 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 [27]).

Now, since $\mathrm{Ran}(dF(u))$ 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 $\tilde{\psi}.dF(u) = 0$ (here, for convenience $\tilde{\psi}$ is written as a row vector and $dF(u)$ as a matrix with $n+1$ rows). In other words, we have obtained the usual Lagrange multipliers relation

$$\psi.dE_{x_0,T}(u) + \psi^0 dC_{x_0,T}(u) = 0. \tag{2.5}$$ lagmu

Now the rest of the game consists of translating mathematically the equation (2.5) (or, more exactly, the equation $\tilde{\psi}.dF(u) = 0$). 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* $\tilde{x} = (x, x^0)$ and the *augmented dynamics* $\tilde{f}(t, \tilde{x}, u) = (f(t, x, u), f^0(t, x, u))$. We consider the *augmented control system*

$$\dot{\tilde{x}}(t) = \tilde{f}(t, \tilde{x}(t), u(t)). \tag{2.6}$$ aug

Note that (**OCP**) is then equivalent to the optimal control problem of steering the system (2.6) from $\tilde{x}_0 = (x_0, 0)$ to $\tilde{x}_1 = (x_1, x^0(T))$ by minimizing $x^0(T)$.

Since $F(u) = (E_{x_0,T}(u), C_{x_0,T}(u))$, it follows that F is exactly the end-point mapping for the augmented control system (2.6).

Note that, then, the image of F (drawn on Figure 2.1) coincides with the accessible set $\tilde{Acc}(\tilde{x}_0, T)$ of the augmented system.

Using Theorem 6, the (Fréchet) differential $dF(u) : L^\infty(0,T;\mathbb{R}^m) \to \mathbb{R}^n$ is given by

$$dF(u).\delta u = \tilde{M}(T) \int_0^T \tilde{M}^{-1}(t)\tilde{B}(t)\delta u(t)\, dt$$

where $\tilde{M}(\cdot)$ is defined as the solution of the Cauchy problem $\dot{\tilde{M}}(t) = \tilde{A}(t)\tilde{M}(t)$, $\tilde{M}(0) = I_{n+1}$, 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 $\tilde{\psi}.dF(u).\delta u = 0$ for every $\delta u \in L^\infty(0,T;\mathbb{R}^m)$, it follows that

$$\tilde{\psi}\tilde{M}(T)\tilde{M}^{-1}(t)\tilde{B}(t) = 0 \tag{2.7}$$ almostpmp

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

$$\dot{\tilde{p}}(t) = -\tilde{p}(t)\tilde{A}(t), \quad \tilde{p}(T) = \tilde{\psi}. \tag{2.8}$$ ptilde

We are almost done. Let us now come back in the initial coordinates in $\mathbb{R}^n \times \mathbb{R}$. We set $\tilde{p}(t) = (p(t), p^0(t))$. Since $\tilde{f}$ does not depend on the (slack) variable x^0, it follows that $\frac{\partial \tilde{f}}{\partial x^0} = 0$, and therefore, using (2.8), that

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

In particular, we have $\dot{p}^0(t) = 0$ and thus $p^0 = \psi^0$ is a constant. Besides, defining the Hamiltonian $H(t, x, p, p^0, u) = p.f(t, x, u) + p^0 f^0(t, x, u)$, we easily get that

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

and from (2.7) we get that

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

almost everywhere on $[0,T]$. We have obtained what we call the weak PMP.

We do not encapsulate it into a summarized statement, because this section was just to motivate the general version of the PMP, that we state hereafter. Besides of the transversality conditions (that can be easily derived with simple considerations, see [62]), as already said the most technical part is to take into account control constraints. Then, the condition $\frac{\partial H}{\partial u} = 0$ becomes a maximization condition of the Hamiltonian.

2.3 Strong Pontryagin maximum principle

Let us now formulate the strong version of the PMP.

The historical proof can be found in [54]. As in [1, 7, 30, 45], 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 [22]), on the Hahn-Banach theorem (see [10]). A concise sketch of proof, still based on an implicit function argument as in the above proof of the weak PMP (and using needle-like variations) can be found in [27]. As discussed in [9], all these different approaches of proof have their specificities. Such or such approach may be preferred to another when trying to derive a PMP in such or such context (for instance, the Ekeland approach is well adapted to derive versions of the PMP with state constraints, or in infinite dimension).

2.3.1 General statement

?⟨PMP⟩? **Theorem 11.** *Let $x(\cdot)$ be a solution of* **(OCP)**, *corresponding to a control u on $[0,t_f]$. Then there exists an absolutely continuous vector-valued function $p(\cdot) : [0,t_f] \longrightarrow \mathbb{R}^n$ called adjoint vector and a real number $p^0 \leqslant 0$, with $(p(\cdot),p^0) \neq (0,0)$, such that*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t,x(t),p(t),p^0,u(t)), \quad \dot{p}(t) = -\frac{\partial H}{\partial x}(t,x(t),p(t),p^0,u(t)), \tag{2.9}$$ systPMP

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

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

and we have the maximization condition

$$H(t,x(t),p(t),p^0,u(t)) = \max_{v\in\Omega} H(t,x(t),p(t),p^0,v) \tag{2.10}$$ contraintePMP

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

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

$$\max_{v\in\Omega} H(t_f,x(t_f),p(t_f),p^0,v) = -p^0\frac{\partial g}{\partial t}(t_f,x(t_f)). \tag{2.11}$$ condannul

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

$$p(0) \perp T_{x(0)}M_0 \tag{2.12}$$?condt1?

$$p(t_f) - p^0\frac{\partial g}{\partial x}(t_f,x(t_f)) \perp T_{x(t_f)}M_1 \tag{2.13}$$ condt2

where the notation T_xM 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).

?⟨rem622⟩? **Remark 15.** In the conditions of the theorem, we have moreover that

$$\frac{d}{dt}H(t,x(t),p(t),p^0,u(t)) = \frac{\partial H}{\partial t}(t,x(t),p(t),p^0,u(t)) \tag{2.14}$$ tez

for almost every $t \in [0,t_f]$. 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 hence from (2.14) it follows that

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

Note that this equality is then valid for every time because the maximized Hamiltonian is a Lipschitz function of t.

Remark 16. If $g = 0$ in (**OCP**) then (2.11) 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 17. If $M_1 = \{x \in \mathbb{R}^n \mid F_1(x) = \cdots = F_p(x) = 0\}$, where the functions F_i are of class C^1 on $\mathbb{R}^n$, then (2.13) implies that

$$\exists \lambda_1, \ldots, \lambda_p \in \mathbb{R} \;\Big|\; p(T) = \sum_{i=1}^{p} \lambda_i \nabla F_i(x(T)) + p^0 \frac{\partial g}{\partial x}(T, x(T)).$$

⟨rem_normalization⟩ **Remark 18.** We have seen in the proof of the weak PMP that $(p(T), p^0) = (\psi, \psi^0)$ is a Lagrange multiplier. It is thus defined up to a multiplying scalar. Therefore, if $(p(\cdot), p^0)$ is a given adjoint vector, then, for every $\lambda > 0$, $(\lambda p(\cdot), \lambda p^0)$ is as well an adjoint vector, for which all equations of the PMP hold true.

At this step, be careful 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 (2.10). Actually, the choice made by Pontryagin is to take $p^0 \leqslant 0$ in the statement: this leads to the *maximum principle* (the choice $p^0 \geqslant 0$ is valid as well but in that case leads to a *minimum principle*).

Definition 4. *A quadruple $(x(\cdot), p(\cdot), p^0, u(\cdot))$ solution of (2.9) and (2.10) is called an extremal.*

- *If $p^0 < 0$, then 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 $p^0 = 0$, then the extremal is said to be abnormal.*

In the proof of the weak PMP, it can be seen that abnormal extremals correspond to $\psi^0 = 0$ and thus to the case where $\psi . dE_{x_0,T}(u) = 0$ (singularity of the end-point mapping).

Note that 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 $(x(\cdot), p(\cdot), p^0, u(\cdot))$.

Remark 19. Note that, if the optimal control u does not saturate the constraint, then (2.10) implies $\frac{\partial H}{\partial u} = 0$, and we recover the weak PMP.

Remark 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 the implied transversality conditions do coincide.

Remark 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 $\{0\}$.

At the opposite, 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 $p(t_f) = p^0 \frac{\partial g}{\partial x}(t_f, x(t_f))$ (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.

Remark 22 (Generalization of the transversality conditions)**.** Here 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 $\mathbb{R}^n \times \mathbb{R}^n$. Using a simple "copy-paste" of the dynamics, it is then very simple to prove (see [1]) that in that case the transversality conditions become (if they make sense)

$$\left(-p(0), p(t_f) - p^0 \frac{\partial g}{\partial x}(t_f, x(t_f))\right) \perp T_{(x(0),x(t_f))}M.$$

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 $p(0) = p(t_f)$.

Remark 23. A 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 that case, one can still write transversality conditions, by using notions of nonsmooth analysis (see [16, 67]), and there holds

$$\left(p(0), -p(t_f) + p^0 \frac{\partial g}{\partial x}(t_f, x(t_f))\right) \in N_M(x(0), x(t_f))$$

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, in case the subset M has some nonsmooth points.

Remark 24 (Generalizations of the PMP)**.** The PMP withstands many possible generalizations. Anyway its proof (at least, the one of the weak PMP) is quite simple and can be adapted to many situations.

Among the most well known and useful generalizations, one can think of the PMP for (**OCP**) with state constraints (see [12, 16, 54, 67]), for nonsmooth (**OCP**) (see [16, 67]), hybrid (**OCP**) (see [25, 27]), (**OCP**) settled on time scales (see [9]). There exist several possible proofs of the PMP, based either on an implicit function argument (as we did here), or on a (Brouwer) fixed point argument (as in the classical book [54]), or on a Hahn-Banach separation argument (as in [10], or on Ekeland's principle (see [22]). Each of them may or may not be adapted to such or such generalization.

Let us note that, when dealing with state constraints, in general 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, and in that favorable case the adjoint vector is then piecewise absolutely continuous, with possible jumps when touching the state constraint. Unfortunately in the whole generality the structure of the measure can be much more complicated, but such a discussion is outside of the scope of the present manuscript. We refer the reader to [8, 12, 16, 29, 33, 34, 51, 54, 67].

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.

Remark 25. Let us insist of the fact that the PMP is nothing else but a first-order necessary condition for optimality.[1] 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 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 [6] (see references therein) for a survey on that theory and on algorithms in order 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.

Remark 26. In order to compute the optimal trajectories in practice, after having applied the PMP, we have to solve a *shooting problem*, that is, a boundary value problem consisting of computing extremals satisfying certain terminal conditions. This can be done by means of a Newton method, combined with some numerical method for integrating ODEs: this is the shooting method.

We do not have room, here, to describe numerical methods in optimal control. We refer to [5] for a thorough description of so-called direct methods, and to [63] for a survey on indirect methods and the way to implement them in practice (see also [62]).

2.4 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.

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

2.4.1 Minimal time problem for linear control systems

⟨sec241⟩ 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 Chapter 1. Let $x_0 \in \mathbb{R}^n$ be an arbitrary initial point, and let Ω 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, this existence follows immediately from Theorem 5, and we have $t_f = \min\{t \geqslant 0 \ \mid \ x_1 \in \mathrm{Acc}_\Omega(x_0, t)\}$.

With the notations introduced at the beginning of Chapter 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 $g(t,x) = t$). Writing p as a row vector, the Hamiltonian of the optimal control problem is then $H(t,x,p,p^0,u) = pA(t)x + pB(t)u + pr(t) + p^0$.

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

$$p(t)B(t)u(t) = \max_{v\in\Omega} p(t)B(t)v \qquad (2.15)$$ condmaxlineaire

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

Since the function $v \mapsto p(t)B(t)v$ is linear, we expect that the maximum over Ω be reached at the boundary of Ω (unless $p(t)B(t) = 0$). 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 $|u(t)| \leqslant a$. In that case, $B(t)$ is a vector of $\mathbb{R}^n$, and $\varphi(t) = p(t)B(t)$ is a function called *switching function*. The maximization condition (2.15) implies that

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

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 φ 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 φ are isolated). In that case, the zeros of the switching functions are called the switchings of the optimal control. According to the monotonicity of φ, 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 φ vanishes, for instance, along a time subinterval I of $[0, t_f]$, then the maximization condition (2.15) does not provide any immediate information in order to compute the optimal control u. But we can then derivate with respect to time (if this is allowed) the relation $p(t)B(t) = 0$, 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 [45]).

Lemma 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.*

1. *If all eigenvalues of A are real, then any extremal control has at most $n-1$ switchings on $[0, +\infty)$.*

2. *If all eigenvalues of A have a nonzero imaginary part, then any extremal control has an infinite number of switchings on $[0, +\infty)$. As a consequence, for every $N \in \mathbb{N}^*$, there exists $x_0 \in \mathbb{R}^n$ for which the corresponding optimal 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 $|u_1(t)| \leqslant 1$ and $|u_2(t)| \leqslant 1$. As for the case $m = 1$, we set $\varphi_i(t) = p(t)B_i(t)$, for $i = 1, 2$, and the maximization condition (2.15) 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 $u_1(t)^2 + u_2(t)^2 \leqslant 1$. Setting again $\varphi_i(t) = p(t)B_i(t)$, for $i = 1, 2$, the maximization condition (2.15) can be written as

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

  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.

2.4.2 Linear quadratic theory

⟨sec_LQ⟩ 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 [3, 37, 42, 60, 62].

The basic LQ problem

We consider the optimal control problem

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x_0 \\ &\min \int_0^T \Big(x(t)^\top W(t)x(t) + u(t)^\top U(t)u(t) \Big)\, dt + x(T)^\top Q x(T) \end{aligned} \tag{2.16}$$

?systemLQ?

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^∞ 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 let free. The matrices $W(t)$, $U(t)$, and Q, are called weight matrices.

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

$$\int_0^T \|u(t)\|_U^2\, dt \geqslant \alpha \int_0^T u(t)^\top u(t)\, dt \qquad \forall u \in L^2(0, T; \mathbb{R}^m).$$

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

?⟨existenceLQ⟩? **Theorem 12.** *There exists a unique optimal solution.*

This theorem can be proved using classical functional analysis arguments, as in the proof of Theorem 10. The uniqueness comes from the strict convexity of the cost. For a proof, see [62].

Let us apply the PMP to the basic LQ problem. Writing the adjoint vector p as a row vector, the Hamiltonian is

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

and the adjoint equation is $\dot{p}(t) = -p(t)A(t) - 2p^0 x(t)^\top W(t)$. Since the final point is free, the transversality condition on the final adjoint vector yields $p(T) =$

$2p^0 x(T)^\top Q$, 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 18, 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

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

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

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

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

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 $p(t)^\top$ 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 $p(t)^\top = E(t)x(t)$. 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 indeed infer, by Cauchy uniqueness, that $p(t)^\top = E(t)x(t)$. We have obtained the following result.

⟨thmLQ⟩ **Theorem 13.** *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

$$\begin{aligned} \dot{E}(t) &= W(t) - A(t)^\top E(t) - E(t)A(t) - E(t)B(t)U(t)^{-1}B(t)^\top E(t) \\ E(T) &= -Q \end{aligned} \tag{2.17}$$

riccati

Actually, there is a difficulty for finishing the proof of that theorem, by proving that the unique solution $E(\cdot)$ of the Cauchy problem (2.17), 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 we refer the reader, e.g., to [62] for a proof (which uses the optimality status 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.

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 Chapter 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 Figure 2.2).

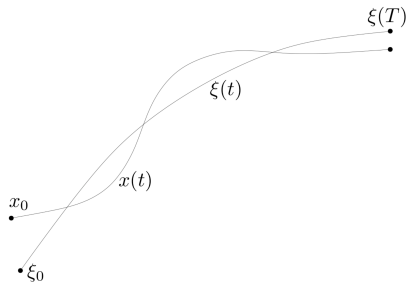

Figure 2.2: Tracking problem

⟨figreg⟩

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

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

with $A(t) = \frac{\partial f}{\partial x}(t, \xi(t), 0)$, $B(t) = \frac{\partial f}{\partial u}(t, \xi(t), 0)$ and $r(t) = f(t, \xi(t), 0) - \dot{\xi}(t) + \mathrm{o}(z(t), u(t))$. It seems reasonable to design the 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 $\mathrm{o}(z(t), u(t))$ 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

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

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 13, there exists a unique optimal control, which is $u(t) = U(t)^{-1}B_1(t)^\top E_1(t)z_1(t)$, 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 4. *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

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

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 $w_0 \geqslant 0$, $u_0 > 0$ and $q_0 \geqslant 0$.

If w_0 is chosen much larger than u_0, then the expected result is that $\|x(t) - \xi(t)\|$ will remain very small (while paying the price of larger values of $u(t)$). Conversely if u_0 is chosen much larger than w_0 then we expect that $u(t)$ take small values, whereas 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 $x(T)$ be very close to $\xi(T)$, at the final time T. A lot of such statements, with numerous possible variants, may be established. We refer to [3, 37, 40, 42, 45, 60, 62] for (many) more precise results.

2.4.3 Examples of nonlinear optimal control problems

Example 3 (Zermelo problem)**.** Let us consider a boat moving with constant speed along a river of constant width ℓ, 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

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

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 Figure 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.

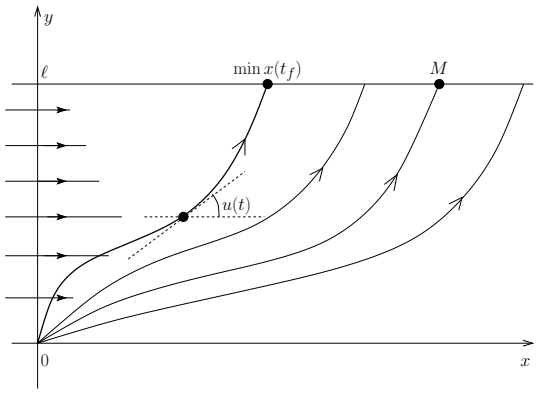

Figure 2.3: Zermelo problem.

⟨fig_zermelo⟩

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

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 $x(t_f)$.
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. Choosing $f^0 = 0$ and $g(t,x,y) = x$, 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

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

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 $c(y) \geqslant v$, this equation is solvable, and we get

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

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 27. The assumption $c(y) \geqslant v$ 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 such and such condition).

?⟨rem_zermelo_reachable⟩? **Remark 28.** The optimal trajectory, minimizing the lateral deport, is represented on Figure 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 $0 \leqslant y \leqslant \ell$ that are at the right of the optimal trajectory.

2. Reaching the opposite side in minimal time.
This time, 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: here, we choose $f^0 = 1$ and $g = 0$, and we obtain $p_x = 0$. It follows that p_y is constant. Besides, we still have $H = 0$ along any extremal. We claim that $p_y \neq 0$ for every time t. 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 29. 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 φ 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 $|p_x v| \leqslant |1 - p_x c(y(t))|$ 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 Figure 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 $c(y) \geqslant v$) 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.

⟨exo7.3.22⟩ **Example 4** (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

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

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

$$0 \leqslant u(t) \leqslant M$$

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 $u(t) \geqslant 0$, we have $\dot{y}(t) \geqslant -cy(t)$, hence $y(t) \geqslant y_0 \mathrm{e}^{-ct} > 0$. 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 $p_y(t) = \frac{b}{c}x(T)(1 - \mathrm{e}^{c(t-T)})$. The maximization condition of the PMP yields

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

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 $x(T) > \frac{c}{b}\frac{1}{1-\mathrm{e}^{-cT}}$. 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 $0 < \alpha \leqslant u(t) \leqslant \beta$.
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 $\alpha \leqslant u_e \leqslant \beta$. 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 $\max_{0\leqslant u\leqslant M}(-p_y y u)$, which gives, since $y(t) > 0$,

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

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 $xp_x = 0$ 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 that $\frac{d}{dt}F_\alpha(x(t), y(t)) = 0$, where

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

(resp., F_β), 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_α at the point $(\alpha, 1)$,

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

we see that F_α has a strict local minimum at the point $(\alpha, 1)$. Moreover, the function F_α 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 α 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 α or to β. It cannot be equal to α, otherwise, since the function F_α 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_α, 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 α or to β. With a similar reasoning as above, we get $u = \alpha$.

Let us now provide a control strategy (actually, optimal; but this requires further analysis) in order to steer the system from any initial point (x_0, y_0) to the point $(\alpha, 1)$.

First, in a neighborhood of the point $(\alpha, 1)$, the level sets of the function F_α 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_β, with respect to the point $(\beta, 1)$ (see Figure 2.4).

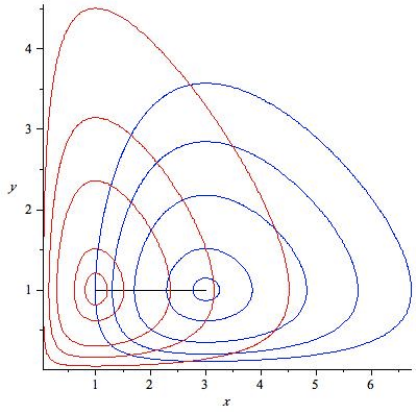

Figure 2.4: Example with $\alpha = 1$ and $\beta = 3$

⟨figlevel⟩

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_α (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_β (thus, "centered" at the point $(\alpha, 1)$) which passes through the target point $(\alpha, 1)$.

Now, if we start from any other point (x_0, y_0), we can determine on Figure 2.4 a sequence of arcs along the level sets, respectively, of F_α and of F_β, that steers the system from the starting point to the target point.

Example 5 (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), the, 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 Figure 2.5.

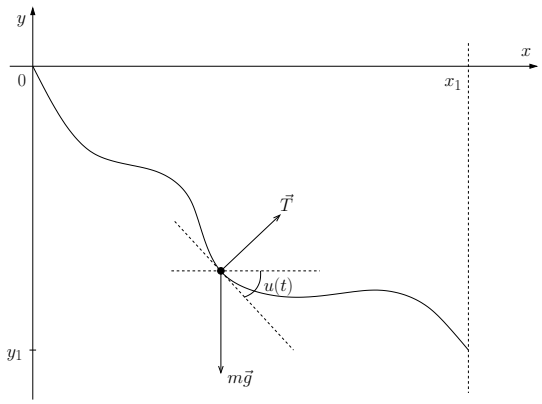

Figure 2.5: The Brachistochrone Problem, seen as an optimal control problem. ⟨fig_tobog⟩

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:

$$\begin{aligned} &\dot{x}(t) = v(t)\cos u(t), && x(0) = 0, && x(t_f) = x_1, \\ &\dot{y}(t) = -v(t)\sin u(t), && y(0) = 0, && y(t_f) = y_1, \\ &\dot{v}(t) = g\sin u(t), && v(0) = 0, && v(t_f) \text{ free.} \end{aligned} \tag{2.18}$$ sys_brach

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 $y_1 \leqslant 0$.

Because of the relation between $y(t)$ and $v(t)$, we can reduce the optimal control problem (2.18) to the minimal time control problem for the following system:

$$\begin{aligned} &\dot{x}(t) = v(t)\cos u(t), && x(0) = 0, && x(t_f) = x_1 > 0 \text{ fixed}, \\ &\dot{v}(t) = g\sin u(t), && v(0) = 0, && v(t_f) \text{ fixed.} \end{aligned} \tag{2.19}$$ sys_brach_reduit

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 (2.19). 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

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

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

We have $\frac{d}{dt}(p_x v(t)) = p_x g \sin u(t)$ et $\frac{d}{dt}(g p_v(t)) = -p_x g \cos u(t)$. As a consequence, if $\varphi(t) = 0$ for every t in some subset I of positive measure, then $p_x = 0$. Therefore $\varphi(t) = (g p_v(t))^2$ 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 φ 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 $H = \sqrt{(p_x v(t))^2 + (g p_v(t))^2} + p^0$. 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 $(p_x v(t))^2 + (g p_v(t))^2 = 1$, and therefore $\cos u(t) = p_x v(t)$ and $\sin u(t) = g p_v(t)$.

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 $v(t) = A \sin(g p_x t)$. 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 $v(t) = \pm\frac{1}{p_x}\sin(g p_x t)$. Now, $\dot{x} = v \cos u = p_x v^2$ and $y = -\frac{1}{2g} v^2$, and by integration we get

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

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 $\sin(g p_x t_f) = 0$ hence $g p_x t_f = \pi + k\pi$ 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 $2g p_x t_f = 2\pi$, we have $x_1 = x(t_f) = \frac{t_f}{2p_x}$, and thus $t_f = \sqrt{\frac{2\pi x_1}{g}}$ and $p_x = \sqrt{\frac{\pi}{2g x_1}}$.

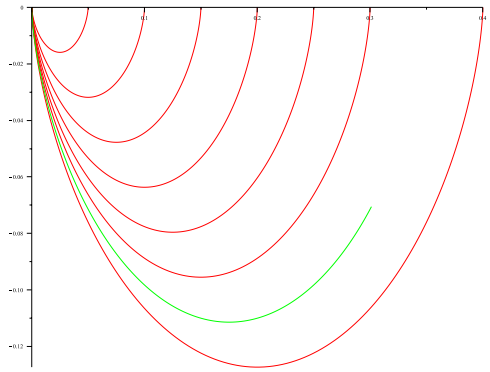

Figure 2.6: Optimal trajectories

⟨fig_brachi⟩

On Figure 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 (2.18) 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 Figure 2.6.

Moreover, if $\dot{y}(t) = 0$ then $2gp_x t = \pi + k\pi$ 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 30. If we investigate the variant of the optimal control problem (2.18), 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 $2gp_x t_f = \pi$, 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 Figure 2.6, the arc of cycloid starting from the origin and reaching $x = x_1$ with an horizontal tangent.

Chapter 3

Stabilization

?⟨chap_stab⟩? 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 $\mathbb{R}^n$

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

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ 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 Ω of $\mathbb{R}^m$ (set of control constraints).

Let $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \mathrm{Int}(\Omega)$[1] be an equilibrium point, that is, $f(\bar{x}, \bar{u}) = 0$. 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 $A \in \mathcal{M}_n(\mathbb{R})$. 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 14. • *If there exists a (complex) eigenvalue λ of A such that $\mathrm{Re}(\lambda) > 0$, then the equilibrium point 0 is unstable, that is, there exists $x_0 \in \mathbb{R}^n$ such that the solution of $\dot{x}(t) = Ax(t)$, $x(0) = x_0$ satisfies $\|x(t)\| \to +\infty$ as $t \to +\infty$.*

[1] Interior of Ω.

- *If all (complex) eigenvalues of A have negative real part, then 0 is asymptotically stable, that is, all solutions of $\dot{x}(t) = Ax(t)$ converge to 0 as $t \to +\infty$.*
- *The equilibrium point 0 is stable if and only if all eigenvalues of A have nonpositive real part and if an eigenvalue λ is such that $\mathrm{Re}(\lambda) = 0$ then λ is a simple root[2] of the minimal polynomial of A.*

Definition 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 χ_A is Hurwitz.

Definition 6. *The Routh table is defined as follows:*

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

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 [57]), which can be proved by means of (nonelementary) complex analysis.

[2] 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.

Theorem 15. *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 16. *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 $a_{n+1} = a_{n+2} = \cdots = a_{2n-1} = 0$, and we define the square matrix of size n

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

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

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

Theorem 17. *[35] If $a_0 > 0$, then P is Hurwitz if and only if $H_i > 0$ for every $i \in \{1, \ldots, n\}$.*

Remark 31. Assume that $a_0 > 0$.

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

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

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

3.1.2 Pole-shifting theorem

Definition 7. *The linear autonomous control system $\dot{x}(t) = Ax(t) + Bu(t)$, with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $A \in \mathcal{M}_n\mathbb{R}$, $B \in \mathcal{M}_{n,m}\mathbb{R}$, is said to be feedback stabilizable if there exists $K \in \mathcal{M}_{m,n}\mathbb{R}$ (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. This is equivalent to requiring that $A+BK$ be Hurwitz.

Remark 32. This concept is invariant through similar transforms $A_1 = PAP^{-1}$, $B_1 = PB$, $K_1 = KP^{-1}$.

⟨thmplacementpoles⟩ **Theorem 18** (pole-shifting theorem)**.** *If (A, B) satisfies the Kalman condition* rank $K(A, B) = n$*, then for every real polynomial of degree n whose leading coefficient is* 1*, there exists $K \in \mathcal{M}_{m,n}\mathbb{R}$ such that $\chi_{A+BK} = P$, that is, the characteristic polynomial of $A + BK$ is equal to P.*

Actually, the converse statement is also true.

Corollary 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 18. We prove the result first in the case $m = 1$. It follows from Theorem 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

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

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 $k_1 = a_n - \alpha_n, \ldots, k_n = a_1 - \alpha_1$.

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

⟨lem_reducedK⟩ **Lemma 4.** *If (A, B) satisfies the Kalman condition, then there exists $y \in \mathbb{R}^m$ and $C \in \mathcal{M}_{m,n}(\mathbb{R})$ such that $(A + BC, By)$ satisfies the Kalman condition.*

It follows from this lemma that, for every polynomial P of degree n whose leading coefficient is 1, there exists $K_1 \in \mathcal{M}_{1,n}(\mathbb{R})$ such that $\chi_{A+BC+ByK_1} = P$, and therefore, setting $K = C + yK_1 \in \mathcal{M}_{m,n}(\mathbb{R})$, we have $\chi_{A+BK} = P$, and the theorem is proved. □

Proof of Lemma 4. Let $y \in \mathbb{R}^m$ be such that $By \neq 0$. Let $x_1 = By$.

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

Indeed otherwise we would have $Ax_1 + \mathrm{Ran}(B) \subset \mathbb{R}x_1$, hence $Ax_1 \in \mathbb{R}x_1$ and $\mathrm{Ran}(B) \subset \mathbb{R}x_1$. Therefore

$$\mathrm{Ran}(AB) = A\mathrm{Ran}(B) \subset \mathbb{R}Ax_1 \subset \mathbb{R}x_1$$

and by immediate iteration, $\mathrm{Ran}(A^k B) \subset \mathbb{R}x_1$, for every integer k. This would imply that

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

which would contradict the Kalman condition.

Claim 2: For every $k \leqslant n$, there exists $x_k \in Ax_{k-1} + \mathrm{Ran}(B)$ (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 $E_k = \mathrm{Span}(x_1, \ldots, x_k)$.

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

Therefore $\mathrm{Ran}(AB) = A\,\mathrm{Ran}(B) \subset AE_{k-1} \subset E_{k-1}$, and similarly we have $\mathrm{Ran}(A^i B) \subset E_{k-1}$ for every integer i. It would follow that

$$\mathrm{Ran}(B, AB, \ldots, A^{n-1}B) \subset E_{k-1}$$

which would contradict the Kalman condition.

We have thus built a basis $(x_1, \ldots, x_n)$ of $\mathbb{R}^n$. We define $C \in \mathcal{M}_{m,n}(\mathbb{R})$ by the relations

$$Cx_1 = y_1,\ Cx_2 = y_2, \ldots,\ Cx_{n-1} = y_{n-1},\ Cx_n \text{ arbitrary.}$$

Then $(A + BC, x_1)$ satisfies the Kalman condition since $(A + BC)x_1 = Ax_1 + By_1 = x_2$, …, $(A + BC)x_{n-1} = Ax_{n-1} + By_{n-1} = x_n$. □

Remark 33. 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 [37]), limited however to $m = 1$ and not very reliable numerically. Better is to use *place.m*, which is a robust pole-shifting routine (see [39]) 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 Section 2.4.2, by taking an infinite horizon of time $T = +\infty$ as said at the end of that section (LQR stabilization).

3.2 Stabilization of instationary linear systems

There is no such theory for instationary linear systems $\dot{x}(t) = A(t)x(t) + B(t)u(t)$.

Let us explain how the difficulties do emerge, by considering the system $\dot{x}(t) = A(t)x(t)$ without any control. A priori one could think that, if the matrix $A(t)$ is Hurwitz for every t, then the system is asymptotically stable. This is however wrong. The statement is wrong as well even under stronger assumptions on $A(t)$. For instance, the system may fail to be asymptotically stable even though the matrix $A(t)$ satisfies the following assumption: there exists $\varepsilon > 0$ such that, for every time t, every (complex) eigenvalue $\lambda(t)$ of $A(t)$ satisfies $\mathrm{Re}\,(\lambda(t)) \leqslant -\varepsilon$.

A simple counterexample is given by

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

with $a \in (1, 2)$ arbitrary. Indeed it can be seen that

$$x(t) = e^{(a-1)t}\begin{pmatrix} \cos t \\ -\sin(t) \end{pmatrix}$$

is a solution of $\dot{x}(t) = A(t)x(t)$, and does not converge to 0 whenever $a \geqslant 1$. 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 varying in time linear systems (see [40]).

3.3 Stabilization of nonlinear systems

3.3.1 Reminders on stability: Lyapunov and Lasalle theorems

Consider the continuous dynamical system $\dot{x}(t) = f(x(t))$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ 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 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 $\|x_0 - \bar{x}\| \leqslant \delta$, one has $\|x(t, x_0) - \bar{x}\| \leqslant \varepsilon$ for every $t \geqslant 0$. It is said to be locally asymptotically stable (in short, LAS) if it is stable and if moreover $x(t, x_0) \to \bar{x}$ as $t \to +\infty$ 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.*

⟨thm_lineariz⟩ **Theorem 19.** *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, not saying anything however on the size of the neighborhoods of stability. Let us now give two important results of Lyapunov theory, providing more knowledge on these neighborhoods, with the concept of Lyapunov function.

Definition 9. *Let Ω be an open subset of $\mathbb{R}^n$ containing the equilibrium point $\bar{x}$. The function $V : \Omega \to \mathbb{R}$ is called a Lyapunov function at $\bar{x}$ on Ω if*

- *V is of class C^1 on Ω;*
- *$V(\bar{x}) = 0$ and $V(x) > 0$ for every $x \in \Omega \setminus \{\bar{x}\}$;*
- *$\langle \nabla V(x), f(x) \rangle \leqslant 0$ for every $x \in \Omega$. If the inequality is strict on $\Omega \setminus \{\bar{x}\}$ then the Lyapunov function 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, ensuring the stability.

Theorem 20 (Lyapunov theorem)**.** *If there exists a Lyapunov function V at $\bar{x}$ on Ω, then $\bar{x}$ is stable. If V is strict then $\bar{x}$ is LAS. If moreover V is proper*[3]*, then $\bar{x}$ is GAS in Ω.*

[3] V is said to be proper whenever $V^{-1}([0, L])$ is a compact subset of Ω, for every $L \in V(\Omega)$; in other words, the inverse image of every compact is compact. When $\Omega = \mathbb{R}^n$, this property is equivalent to say that $V(x) \to +\infty$ as $\|x\| \to +\infty$.

When a Lyapunov function is not strict then one can be more specific and infer that the trajectory converges to some subset.

Theorem 21 (Lasalle principle)**.** *Let V be a Lyapunov function on Ω. Let $\mathcal{I}$ be the largest subset of $\{x \in \Omega \mid \langle \nabla V(x), f(x)\rangle = 0\}$ that is invariant under the flow (in positive time) of the dynamical system. Then all solutions converge to $\mathcal{I}$, in the sense that $d(x(t), \mathcal{I}) \to 0$ as $t \to +\infty$ (with d the Euclidean distance).*

Remark 34. It is interesting to formulate the Lasalle principle in the particular case where the invariant set $\mathcal{I}$ is reduced to the singleton $\{\bar{x}\}$. The statement is then as follows.

Let V be a Lyapunov function such that, if $x(\cdot)$ is a solution of the system such that $\langle \nabla V(x(t)), f(x(t))\rangle = 0$ for every $t \geqslant 0$, then necessarily $x(t) = \bar{x}$. Then the equilibrium point $\bar{x}$ is GAS in Ω.

Example 6. Let $g : \mathbb{R} \to \mathbb{R}$ 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 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 7. Consider the system in $\mathbb{R}^2$

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

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.

⟨Lyapunov_lemma⟩ **Example 8** (Lyapunov lemma and applications)**.** Let A be a real matrix of size n, whose eigenvalues have negative real parts. Then there exists a symmetric positive definite matrix P of size n such that $A^*P + PA = -I_n$. Indeed, it suffices to take $P = \int_0^{+\infty} e^{tA^*} e^{tA}\, dt$.

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 recover Theorem 19.

3.3.2 Application to the stabilization of nonlinear control systems

Consider the general nonlinear control system (3.1), and an equilibrium point $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \mathrm{Int}(\Omega)$, as settled at the beginning of the chapter. We consider the linearized system at that point,

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

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, then Theorem 19 implies a local stabilization result for the nonlinear control system (3.1). In other words, we have the following theorem.

Theorem 22. *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 $(\bar{x}, \bar{u})$. In other words, 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 neighborhood has to be small enough so that the closed-loop control u takes its values in the set Ω.

The Lyapunov and Lasalle theorems can be applied as well to control systems in an evident way, providing knowledge on the stability neighborhoods. For instance, we get the following statement: considering as above the nonlinear control system (3.1), we assume that there exists a function $V : \Omega \to \mathbb{R}^+$ of class C^1, taking positive values in $\Omega \setminus \{\bar{x}\}$, and such that, for every $x \in \Omega$, 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 Ω. Many other similar statements can be easily derived, based on the Lyapunov or Lasalle theorems. Note that there exists an elaborate theory of control Lyapunov functions (see, e.g., [60]). Naturally, what is difficult here is to be able to ensure a good 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.

To illustrate the role of the Lyapunov functions in stabilization, let us next describe a spectacular and widely used (and however very simple) strategy in order to design stabilizing controls thanks to Lyapunov functions.

Jurdjevic-Quinn method.

⟨thm_JQ⟩ **Theorem 23.** *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 $f(\bar{x}) = 0$. In other words, $\bar{x}$ is an equilibrium point of the uncontrolled system (that is, with $u_i = 0$).

We assume that there exists a Lyapunov function at $\bar{x}$ for the uncontrolled system, that is, we assume that there exists a function $V : \mathbb{R}^n \to \mathbb{R}^+$ such that

- *$V(\bar{x}) = 0$ and $V(x) > 0$ for every $x \neq \bar{x}$;*
- *V is proper;*
- *$L_f V(x) = \langle \nabla V(x), f(x) \rangle \leqslant 0$ for every $x \in \mathbb{R}^n$;*[4]
- *the set*

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

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

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

and if $L_F V(x(t)) = 0$ for every $t \geqslant 0$, 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

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

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. □

The idea of the Jurdjevic-Quinn method, that can be seen in the proof above, 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.

The method is much used, for instance, for the stabilization of satellites towards a given orbit. Let us provide hereafter another class of applications in mathematical biology (control of populations in Lotka-Volterra systems).

[4] The notation $L_f V$ is called the Lie derivative of V along f. It is defined by $L_f V(x) = df(x).f(x) = \langle \nabla V(x), f(x) \rangle$. It is the derivative of V along the direction f.

Example 9. 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 23, 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 10 (Generalized Lotka-Volterra system)**.** Consider the generalized Lotka-Volterra system

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

Consider the equilibrium point $\bar{N} = (\bar{N}_1, \ldots, \bar{N}_n)^T$ defined by $b + A\bar{N} = 0$, where $b = (b_1, \ldots, b_n)^\top$ and A is the square matrix of coefficients a_{ij}. Let $c_1, \ldots, c_n$ 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^{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.[5] 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:

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

Then, we compute $\frac{d}{dt}V(N(t)) = \sum_{i=1}^{n-1} \alpha_i(N_i(t) - \bar{N}_i)u(t)$. 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}, $i = 1, \ldots, n - 1$, be nonzero, and that A be invertible.

In the particular case $n = 2$, it is actually possible to ensure moreover that $u(t) \geqslant 0$, by playing with the periodicity of the trajectories (as in Example 4).

[5] 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.

Part II

Control in infinite dimension

⟨part2⟩ In this part we introduce the control theory for infinite-dimensional systems, that is, control systems $\dot{x}(t) = f(x(t), u(t))$ 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 mainly from functional analysis. The reader should then be quite well acquainted with such knowledge, and we refer to the textbook [11] 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 [17] for a complete survey of results on this theory. Throughout this part, 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 Ω, y will denote as well a function of t and x, where t is the time and x the spatial variable, and the system considered throughout is

$$\dot{y} = Ay + Bu \tag{3.2}$$ sys_part2

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 notion of the exponential of A is not straightforward to define and requires 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 (3.2) 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, our main objective is to provide the general framework in which the Hilbert Uniqueness Method (HUM) of J.-L. Lions can be stated.

Chapter 4

Semigroup theory

?⟨chap_semigroup⟩? The objective of this chapter is to prove that, in an appropriate functional setting, the unique solution of the Cauchy problem

$$\dot{y}(t) = Ay(t) + f(t), \qquad y(0) = y_0, \tag{4.1}$$ sys4.1

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

$$y(t) = S(t)y_0 + \int_0^t S(t-s)f(s)\,ds \tag{4.2}$$ f4.2

where $(S(t))_{t\geqslant 0}$ 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 $S(t) = \mathrm{e}^{tA}$, with the usual matrix exponential. In the infinite-dimensional setting this step is far from 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 Ω of $\mathbb{R}^n$.

Most items of the present chapter are borrowed from the textbook [53] on semigroup theory and from [66].

Let us recall several basic notions of functional analysis that are instrumental in what follows (see [11]).

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

The notation X' stands for the (topological) dual of the Banach space X, that is, the vector space of all linear continuous mappings $\ell : X \to \mathbb{R}$ (in other

words, $X' = L(X, \mathbb{R})$). 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 $A : D(A) \to Y$ defined on a vector subspace $D(A) \subset X$ called the *domain* of A. The operator A is said to be *bounded*[1] if $D(A) = X$ and if there exists $C > 0$ such that $\|Ax\|_Y \leqslant C\|x\|_X$ for every $x \in D(A)$.

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

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

Then $D(A^*)$ is a vector subspace of Y'. For every $z \in D(A^*)$, we define the linear form $\ell_z : D(A) \to \mathbb{R}$ by $\ell_z(x) = \langle z, Ax \rangle_{Y',Y}$ for every $x \in D(A)$. By definition of $D(A^*)$ we have $|\ell_z(x)| \leqslant C\|x\|_X$ for every $x \in D(A)$. Since $D(A)$ is dense in X, it follows that the linear form ℓ_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 $A^* : D(A^*) \subset Y' \to X'$, 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' (see [11]).

In the case where X is a Hilbert space, we identify X' with X. A densely defined linear operator $A : D(A) \subset X \to X$ 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.

[1] 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.

Throughout this part, we consider a Banach space X. An operator on X will mean an (unbounded) linear operator $A : D(A) \subset X \to X$. In practice, most of unbounded operators used to model systems of the form (4.1) are operators $A : D(A) \subset X \to X$ 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 [58, 66]).

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 id_X stands for the identity mapping on X.

⟨def_semigroup⟩ **Definition 10.** *A C_0 semigroup of bounded linear operators on X is a one-parameter family $(S(t))_{t\geqslant 0}$ of bounded linear mappings $S(t) : X \to X$ such that*

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

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

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

on the domain $D(A)$ that is the set of $y \in X$ such that the above limit exists, is called the infinitesimal generator of the semigroup $(S(t))_{t\geqslant 0}$.

Note that the above limit is computed in X, hence with the norm $\|\cdot\|_X$.

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

⟨propclosed⟩ **Proposition 5.** *Let $(S(t))_{t\geqslant 0}$ 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;*
- *for every $y \in D(A)$, one has $S(t)y \in D(A)$ and $\dot{S}(t)y = AS(t)y$ for every $t > 0$.*

This proposition (see [53]) 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 10 (see [53]). 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 $(S(t))_{t\geqslant 0}$ satisfies the stronger (uniform convergence) property $\lim_{t\to 0,\ t>0} \|S(t) - \mathrm{id}_X\| = 0$, then it is said to be *uniformly continuous*. The following result is however proved in [53, 58]:

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

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 6. *Let $(S(t))_{t\geqslant 0}$ be a C_0 semigroup. There exist $M \geqslant 1$ and $\omega \in \mathbb{R}$ such that*

$$\|S(t)\| \leqslant M\mathrm{e}^{\omega t} \qquad \forall t \geqslant 0. \tag{4.3}$$ inegsemigroup

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

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

This proposition (see [53]) shows that C_0 semigroups have at most an exponential growth in norm. This property is similar to what happens in finite dimension.

Definition 11. *Let $A : D(A) \to X$ 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 λ such that $\lambda\,\mathrm{id}_X - A : D(A) \to X$ is invertible and $(\lambda\,\mathrm{id}_X - A)^{-1} : X \to X$ is bounded (we say it is boundedly invertible). The resolvent of A is defined by $R(\lambda, A) = (\lambda\,\mathrm{id}_X - A)^{-1}$, for every $\lambda \in \rho(A)$.*

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

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

Remark 35. If $(S(t))_{t\geqslant 0} \in \mathcal{G}(M, \omega)$ then $\{\lambda \in \mathbb{C} \mid \mathrm{Re}\,\lambda > \omega\} \subset \rho(A)$, and

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

for every $x \in X$ and every $\lambda \in \mathbb{C}$ such that $\operatorname{Re}\lambda > \omega$ (Laplace transform). Indeed, integrating by parts, one has

$$\lambda \int_0^{+\infty} \mathrm{e}^{-\lambda t} S(t)\, dt = \mathrm{id}_X + A \int_0^{+\infty} \mathrm{e}^{-\lambda t} S(t)\, dt$$

and thus

$$(\lambda\, \mathrm{id}_X - A) \int_0^{+\infty} \mathrm{e}^{-\lambda t} S(t)\, dt = \mathrm{id}_X.$$

Note that, using the expression (4.4) of $R(\lambda, A)$ with the Laplace transform, it follows that, if $(S(t))_{t\geqslant 0} \in \mathcal{G}(M,\omega)$ then $\|R(\lambda,A)\| \leqslant \frac{M}{\operatorname{Re}\lambda-\omega}$ for every $\lambda \in \mathbb{C}$ such that $\operatorname{Re}\lambda > \omega$. This can be iterated, by derivating $R(\lambda, A)$ with respect to λ, using the resolvent formula and the Laplace transform, and this yields the estimates $\|R(\lambda,A)^n\| \leqslant \frac{M}{(\operatorname{Re}\lambda-\omega)^n}$ for every $n \in \mathbb{N}^*$ and for every $\lambda \in \mathbb{C}$ such that $\operatorname{Re}\lambda > \omega$. Actually, we have the following general result (see [53]).

⟨thm4.1⟩ **Theorem 24.** *A linear operator $A : D(A) \subset X \to X$ is the infinitesimal generator of a C_0 semigroup $(S(t))_{t\geqslant 0} \in \mathcal{G}(M,\omega)$ 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 $\|R(\lambda,A)^n\| \leqslant \frac{M}{(\operatorname{Re}\lambda-\omega)^n}$ for every $n \in \mathbb{N}^*$ and every $\lambda \in \mathbb{C}$ such that $\operatorname{Re}\lambda > \omega$.*

Particular case: contraction semigroups.

Definition 12. *Let $(S(t))_{t\geqslant 0}$ be a C_0 semigroup. Assume that $(S(t))_{t\geqslant 0} \in \mathcal{G}(M,\omega)$ for some $M \geqslant 1$ and $\omega \in \mathbb{R}$. If $\omega \leqslant 0$ and $M = 1$ then $(S(t))_{t\geqslant 0}$ 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 [11, 13]) and in that case Theorem 24 takes the more specific following forms, which are the well-known Hille-Yosida and Lumer-Phillips theorems.

⟨thm_HilleYosida⟩ **Theorem 25** (Hille-Yosida theorem)**.** *A linear operator $A : D(A) \subset X \to X$ 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 $\|R(\lambda,A)\| \leqslant \frac{1}{\lambda}$ for every $\lambda > 0$.*

Remark 36. In the conditions of Theorem 25, one has moreover $\{\lambda \in \mathbb{C} \mid \operatorname{Re}\lambda > 0\} \subset \rho(A)$ and $\|R(\lambda,A)\| \leqslant \frac{1}{\operatorname{Re}\lambda}$ for every $\lambda \in \mathbb{C}$ such that $\operatorname{Re}\lambda > 0$.

Remark 37. Let $A : D(A) \to X$ generating a C_0 semigroup $(S(t))_{t\geqslant 0} \in \mathcal{G}(1,\omega)$. Then the operator $A_\omega = A - \omega \mathrm{id}_X$ (having the same domain) is the infinitesimal generator of $S_\omega(t) = \mathrm{e}^{-\omega t} S(t)$ which is a semigroup of contractions (and conversely). In particular, this implies:

A linear operator $A : D(A) \subset X \to X$ is the infinitesimal generator of a C_0 semigroup $(S(t))_{t\geqslant 0} \in \mathcal{G}(1,\omega)$ 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 $\|R(\lambda, A)\| \leqslant \frac{1}{\lambda-\omega}$ 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 13. *The operator $A : D(A) \subset X \to X$ is said to be:*

- *dissipative if for every $y \in D(A)$ there exists an element $\ell \in F(y)$ such that $\operatorname{Re}\langle \ell, Ay\rangle_{X',X} \leqslant 0$;*
- *m-dissipative if it is dissipative and $\operatorname{Ran}(\mathrm{id}_X - A) = (\mathrm{id}_X - A)D(A) = X$.*

If X is a Hilbert space, then A is dissipative if and only if $\operatorname{Re}(y, Ay)_X \leqslant 0$ for every $y \in D(A)$, where $(\ ,\ )_X$ is the scalar product of X.

Other names are often used in the existing literature (see [11, 13, 58]): 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 38. If $A : D(A) \to X$ is m-dissipative then $\operatorname{Ran}(\lambda\,\mathrm{id}_X - A) = X$ for every $\lambda > 0$.

Remark 39. Let $A : D(A) \to X$ be a m-dissipative operator. If X is reflexive[2] then A is closed and densely defined (i.e., $D(A)$ is dense in X).

?⟨thm_LumerPhillips⟩? **Theorem 26** (Lumer-Phillips theorem)**.** *Let $A : D(A) \subset X \to X$ be a densely defined closed linear operator. Then A is 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 [53]).

⟨prop4.3⟩ **Proposition 7.** *Let $A : D(A) \subset X \to X$ 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.*

[2] There is a canonical injection $\iota : X \to X''$ (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 ι.

Remark 40. 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 [66]).

⟨ex_heat⟩ **Example 11.** Let Ω be an open subset of $\mathbb{R}^n$. The Dirichlet-Laplacian operator $\triangle_{\mathrm{Dir}}$ is defined by $D(\triangle_{\mathrm{Dir}}) = \{f \in H_0^1(\Omega) \mid \triangle f \in L^2(\Omega)\}$ and $\triangle_{\mathrm{Dir}} f = \triangle f$ for every $f \in D(\triangle_{\mathrm{Dir}})$. Note that $H_0^1(\Omega) \cap H^2(\Omega) \subset D(\triangle_{\mathrm{Dir}})$ and that, in general, the inclusion is strict. However, if Ω is an open bounded subset with C^2 boundary, or if Ω is a convex polygon of $\mathbb{R}^2$, then $D(\triangle_{\mathrm{Dir}}) = H_0^1(\Omega) \cap H^2(\Omega)$ (see [26, Chapter 1.5]).

The operator $\triangle_{\mathrm{Dir}}$ is clearly self-adjoint and dissipative in $X = L^2(\Omega)$, and hence by Proposition 7 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 one has $D(\triangle_{\mathrm{Dir}}) = H_0^1(\Omega)$, and actually, if Ω is bounded and of C^1 boundary then $\triangle_{\mathrm{Dir}} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism.

⟨ex_wave⟩ **Example 12.** Anticipating a bit, let us study the operator underlying the wave equation. With the notations of Example 11, we define the operator

$$A = \begin{pmatrix} 0 & \mathrm{id}_{H_0^1(\Omega)} \\ \triangle_{\mathrm{Dir}} & 0 \end{pmatrix}$$

on the domain $D(A) = D(\triangle_{\mathrm{Dir}}) \times H_0^1(\Omega)$, 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 13. 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 $\sup_{n\geqslant 1} \lambda_n < +\infty$ (this is satisfied if $\lambda_n \to -\infty$ as $n \to +\infty$). We define the operator

$$Ay = \sum_{n=1}^{+\infty} \lambda_n (y, e_n)_X e_n \quad \text{on} \quad D(A) = \Big\{ y \in X \;\Big|\; \sum_{n=1}^{+\infty} \lambda_n^2 (y, e_n)_X^2 < +\infty \Big\}.$$

Let us prove that A is self-adjoint and generates the C_0 semigroup defined by

$$S(t)y = \sum_{n=1}^{+\infty} \mathrm{e}^{\lambda_n t} (y, e_n)_X e_n.$$

Firstly, noting that $X_p \subset D(A)$ for every $p \in \mathbb{N}^*$, with $X_p = \{y \in X \mid \exists p \in \mathbb{N}^* \;\; \forall n \geqslant p \;\; (y, e_n) = 0\}$ 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 $y_p \to y \in X$ and $Ay_p \to z \in X$ 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

$\sum_{n=1}^{N} \lambda_n^2 (y_p, e_n)_X^2 \leqslant M$, for every $p \in \mathbb{N}^*$ and every $N \in \mathbb{N}^*$. Letting p tend to $+\infty$ then yields that $\sum_{n=1}^{+\infty} \lambda_n^2 (y, e_n)_X^2 \leqslant M$, 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 $\lambda \mathrm{id}_X - A$ is boundedly invertible if and only if $\inf_{n \geqslant 1} |\lambda - \lambda_n| > 0$. If $\inf_{n \geqslant 1} |\lambda - \lambda_n| > 0$, then let us 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, $A_\lambda : X \to X$ is linear and bounded, and one has $\mathrm{Ran} A_\lambda \subset D(A)$ and $(\lambda \mathrm{id}_X - A) A_\lambda = A_\lambda (\lambda \mathrm{id}_X - A) = \mathrm{id}_X$, hence $\lambda \in \rho(A)$ and $A_\lambda = (\lambda \mathrm{id}_X - A)^{-1}$. Conversely, if $(\lambda \mathrm{id}_X - A)$ is boundedly invertible, then for every $n \in \mathbb{N}^*$ there exists $y_n \in X$ such that $(\lambda \mathrm{id}_X - A) y_n = e_n$, and the sequence $(y_n)_{n \in \mathbb{N}^*}$ is bounded. One has $y_n = \frac{1}{\lambda - \lambda_n} e_n$ and hence $\inf_{n \geqslant 1} |\lambda - \lambda_n| > 0$.

It follows from these arguments that, if $\inf_{n \geqslant 1} |\lambda - \lambda_n| > 0$, then

$$R(\lambda, A)^p y = \sum_{n=1}^{+\infty} \frac{1}{(\lambda - \lambda_n)^p} (y, e_n)_X e_n$$

and hence

$$\|R(\lambda, A)^p\| \leqslant \sup_{n \geqslant 1} \frac{1}{|\lambda - \lambda_n|^p} = \left(\sup_{n \geqslant 1} \frac{1}{|\lambda - \lambda_n|} \right)^p .$$

Let $\omega \geqslant \sup_{n \geqslant 1} \lambda_n$. Then for every $\lambda > \omega$ one has $\inf_{n \geqslant 1} |\lambda - \lambda_n| \geqslant \lambda - \omega$ and hence $\sup_{n \geqslant 1} \frac{\lambda - \omega}{|\lambda - \lambda_n|} \leqslant 1$. 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 14. In order to model the 1D transport equation $\partial_t y + \partial_x y = 0$, with $0 \leqslant x \leqslant 1$, 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 $A - \lambda \mathrm{id}_X$ 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 $A : D(A) \to X$, the adjoint $A^* : D(A^*) \to X'$ 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 $(S(t))_{t \geqslant 0}$ on X, the adjoint family $(S(t)^*)_{t \geqslant 0}$ is a family of bounded operators on X' satisfying the semigroup property but is not necessarily a C_0 semigroup. We have the following result (see [53]).

⟨propS*⟩ **Proposition 8.** *If X is reflexive and if $(S(t))_{t\geqslant 0}$ is a C_0 semigroup on X with generator A then $(S(t)^*)_{t\geqslant 0}$ is a C_0 semigroup on X' with generator A^*.*

4.1.2 The Cauchy problem

?⟨sec_Cauchy⟩? **Classical solutions**

Let $A : D(A) \subset X \to X$ be a densely defined linear operator on the Banach space X. Consider the Cauchy problem

$$\begin{aligned} \dot{y}(t) &= Ay(t) \quad \forall t > 0 \\ y(0) &= y_0 \in D(A). \end{aligned} \tag{4.5}$$ cp

⟨thm4.4⟩ **Theorem 27.** *Assume that A is the infinitesimal generator of a C_0 semigroup $(S(t))_{t\geqslant 0}$ on X. Then the Cauchy problem (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 $t \geqslant 0$. Moreover, the differential equation $\dot{y}(t) = Ay(t)$ makes sense in X.*

This solution is often called *strong solution* in the existing literature.

⟨exheat⟩ **Example 15.** Let Ω be a bounded open subset of $\mathbb{R}^n$ having a C^1 boundary. Having in mind Example 11, let us apply Theorem 27 to the Dirichlet heat equation. The Cauchy problem

$$\partial_t y = \triangle y \ \text{ in } \Omega, \qquad y_{|\partial\Omega} = 0, \qquad y(0) = y_0 \in H_0^1(\Omega),$$

has a unique solution $y(\cdot) \in C^0([0,+\infty); H_0^1(\Omega)) \cap C^1((0,+\infty); H^{-1}(\Omega))$. Moreover, there exist $M \geqslant 1$ and $\omega < 0$ such that $\|y(t,\cdot)\|_{L^2(\Omega)} \leqslant M\mathrm{e}^{\omega t}\|y_0(\cdot)\|_{L^2(\Omega)}$.

⟨exwave⟩ **Example 16.** Let Ω be a bounded open subset of $\mathbb{R}^n$ having a C^1 boundary. Having in mind Example 12, let us apply Theorem 27 to the Dirichlet wave equation. The Cauchy problem

$$\partial_{tt} y = \triangle y \ \text{ in } \Omega, \qquad y_{|\partial\Omega} = 0, \qquad y(0) = y_0 \in H_0^1(\Omega),\ \partial_t y(0) = y_1 \in L^2(\Omega),$$

has a unique solution

$$y(\cdot) \in C^0([0,+\infty); H_0^1(\Omega)) \cap C^1((0,+\infty); L^2(\Omega)) \cap C^2((0,+\infty); H^{-1}(\Omega)).$$

Moreover there holds $\|\partial_t y\|^2_{H^{-1}(\Omega)} + \|y\|^2_{L^2(\Omega)} = \|y_1\|^2_{H^{-1}(\Omega)} + \|y_0\|^2_{L^2(\Omega)}$.

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

$$y(\cdot) \in C^0([0,+\infty); H^2(\Omega) \cap H_0^1(\Omega)) \cap C^1((0,+\infty); H_0^1(\Omega)) \cap C^2((0,+\infty); L^2(\Omega))$$

and $\|\partial_t y\|^2_{L^2(\Omega)} + \|y\|^2_{H_0^1(\Omega)} = \|y_1\|^2_{L^2(\Omega)} + \|y_0\|^2_{H_0^1(\Omega)}$.

⟨rem_regular⟩ **Remark 41.** Concerning the regularity of the solutions of the heat equation (Example 15) and of the wave equation (Example 16), 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 eigenvalues $(\lambda_j)_{j\in\mathbb{N}^*}$.

For the heat equation of Example 15, 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^∞ for $t > 0$ (see [13]), and for every $t > 0$ fixed, the function $x \mapsto y(t,x)$ is (real) analytic on Ω (see [52]). This reflects the so-called *regularizing effect* of the heat equation.

For the wave equation, there is no regularizing effect, but smoothness or analyticity properties can also be established for appropriate initial conditions (see [52]).

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 48 further).

⟨rem4.4⟩ **Remark 42.** 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.

Weak solutions

The objective of this section is to define an extension of the Banach space X, and extensions of C_0 semigroups on X which will provide weaker solutions. A good reference for this part is the textbook [66].

Let $(S(t))_{t\geqslant 0} \in \mathcal{G}(M,\omega)$ be a C_0 semigroup on X, of generator $A : D(A) \to X$. Let $\beta \in \rho(A)$ (if X is real, consider such a real number β).

Definition 14. *Let X_1 denote the Banach space $D(A)$, equipped with the norm $\|y\|_1 = \|(\beta\,\mathrm{id}_X - A)y\|_X$, and let X_{-1} denote the completion of X with respect to the norm $\|y\|_{-1} = \|(\beta\,\mathrm{id}_X - A)^{-1}y\|_X = \|R(\beta,A)y\|_X$.*

Note that, by definition, $\beta\,\mathrm{id}_X - A : X_1 \to X$ and $(\beta\,\mathrm{id}_X - A)^{-1} : X \to X_1$ are surjective isometries (unitary operators). It is then easy to see that the norm $\|\ \|_1$ on X_1 is equivalent to the graph norm $\|y\|_G = \|y\|_X + \|Ay\|_X$. Therefore, from the closed graph theorem, $(X_1, \|\ \|_1)$ 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 $\beta \in \rho(A)$, int he 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 $(\beta\,\mathrm{id}_X - A)(\beta'\,\mathrm{id}_X - A)^{-1} = \mathrm{id}_X + (\beta - \beta')(\beta'\,\mathrm{id}_X - A)^{-1}$, hence

$$(\beta'\,\mathrm{id}_X - A)^{-1} = (\beta\,\mathrm{id}_X - A)^{-1} + (\beta - \beta')(\beta\,\mathrm{id}_X - A)^{-1}(\beta'\,\mathrm{id}_X - A)^{-1}$$

(resolvent identity), and moreover $(\beta\,\mathrm{id}_X - A)^{-1}$ and $(\beta'\,\mathrm{id}_X - A)^{-1}$ commute. The conclusion follows easily.

Remark 43. The injections $X_1 \hookrightarrow X \hookrightarrow X_{-1}$ are, by definition, continuous and dense. They are moreover compact as soon as $\beta\,\mathrm{id}_X - A$ has a compact inverse.

⟨ex4.7⟩ **Example 17.** Let Ω be an open bounded subset of $\mathbb{R}^n$ having a C^2 boundary, and consider the Dirichlet-Laplacian $\triangle_{\rm Dir}$ on $X = L^2(\Omega)$ defined on $D(\triangle_{\rm Dir}) = H^1_0(\Omega) \cap H^2(\Omega)$. Then $X_1 = D(\triangle_{\rm Dir}) = H^1_0(\Omega) \cap H^2(\Omega)$ and, as will follow from Theorem 28 below, $X_{-1} = (H^1_0(\Omega) \cap H^2(\Omega))'$, 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 $\|z\|_1 = \|(\beta\,{\rm id}_{X'} - A^*)z\|_{X'}$ where $\beta \in \rho(A^*) = \rho(A)$, is a Banach space.

⟨thm4.5⟩ **Theorem 28.** *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 $\|y\|_{-1} = \|y\|_{D(A^*)'}$ for every $y \in X$. By definition, we have

$$\begin{aligned}\|y\|_{-1} = \|(\beta\,{\rm id}_X - A)^{-1}y\|_X &= \sup_{f\in X',\ \|f\|_{X'}\leqslant 1} |\langle f, (\beta\,{\rm id}_X - A)^{-1}y\rangle_{X',X}| \\ &= \sup_{f\in X',\ \|f\|_{X'}\leqslant 1} |\langle (\beta\,{\rm id}_{X'} - A^*)^{-1}f, y\rangle_{X',X}|.\end{aligned}$$

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

$$\|y\|_{-1} = \sup_{f\in X',\ \|f\|_{X'}\leqslant 1} |\langle y, (\beta\,{\rm id}_{X'} - A^*)^{-1}f\rangle_{X'',X'}|.$$

Besides, by definition we have

$$\|y\|_{D(A^*)'} = \sup_{z\in D(A^*),\ \|z\|_{D(A^*)}\leqslant 1} |\langle y, z\rangle_{D(A^*)',D(A^*)}|.$$

In this expression we make a change of variable: for every $z \in D(A^*)$ such that $\|z\|_{D(A^*)} \leqslant 1$, there exists $f \in X'$ such that $z = (\beta\,{\rm id}_{X'} - A^*)^{-1}f$, and since $\|z\|_{D(A^*)} = \|(\beta\,{\rm id}_{X'} - A^*)z\|_{X'} = \|f\|_{X'}$ it follows that $\|f\|_{X'} \leqslant 1$. Therefore

$$\|y\|_{D(A^*)'} = \sup_{f\in X',\ \|f\|_{X'}\leqslant 1} |\langle y, (\beta\,{\rm id}_{X'} - A^*)^{-1}f\rangle_{D(A^*)',D(A^*)}|.$$

In the above duality bracket, since $y \in X''$ and $(\beta\,{\rm id}_{X'} - A^*)^{-1}f \in X'$ we can as well use the duality bracket $\langle\ ,\ \rangle_{X'',X'}$. Hence $\|y\|_{-1} = \|y\|_{D(A^*)'}$.

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 5). 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 $D(A^*)'$. □

Theorem 29. *The operator $A : D(A) = X_1 \to X$ can be extended to an operator $A_{-1} : D(A_{-1}) = X \to X_{-1}$, and the C_0 semigroup $(S(t))_{t\geqslant 0}$ on X extends to a semigroup $(S_{-1}(t))_{t\geqslant 0}$ on X_{-1}, generated by A_{-1}.*

Proof. Note that the operator $A : D(A) \to X$ is continuous when one endows $D(A)$ with its norm, because $\|Ay\|_X \leqslant \|y\|_G \leqslant C\|y\|_1$ as already said. By definition of the norm in X_{-1}, we have easily

$$\begin{aligned}\|Ay\|_{-1} = \|(\beta\,\mathrm{id}_X - A)^{-1}Ay\|_X &= \|y - \beta(\beta\,\mathrm{id}_X - A)^{-1}y\|_X \\ &\leqslant \|y\|_X + |\beta|\|(\beta\,\mathrm{id}_X - A)^{-1}y\|_X\end{aligned}$$

for every $y \in D(A)$, and since $(\beta\,\mathrm{id}_X - A)^{-1}$ is bounded it follows that there exists some constant $C_1 > 0$ such that $\|Ay\|_{-1} \leqslant C_1\|y\|_X$ for every $y \in D(A)$. Therefore the operator A has a unique extension $A_{-1} : D(A_{-1}) = X \to X_{-1}$ that is continuous for the respective norms. The fact that $D(A_{-1}) = X$ with equivalent norms follows from the equality $\|y\|_X = \|(\beta\,\mathrm{id}_X - A)^{-1}(\beta\,\mathrm{id}_X - A)y\|_X = \|(\beta\,\mathrm{id}_X - A)y\|_{-1} = \|y\|_{D(A_{-1})}$ for every $y \in D(A)$, and by density this holds true for every $y \in X$. □

Remark 44. Note that, by density, if A is m-dissipative then A_{-1} is m-dissipative as well, and hence $(S_{-1}(t))_{t\geqslant 0}$ is a C_0 semigroup of contractions.

The following result follows from Theorem 27, giving an answer to the question raised in Remark 42.

Corollary 4. *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 $y(t) = S(t)y_0 = S_{-1}(t)y_0$.*

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 27, 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 $\dot{y}(t) = Ay(t)$ 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 sometimes defined by transposition. Indeed, if X is reflexive then $X_{-1} \simeq D(A^*)'$ (see Theorem 28), and hence, considering the differential equation $\dot{y}(t) = A_{-1}y(t)$ in the space X_{-1} exactly means that

$$\langle \dot{y}(t), \varphi\rangle_{(D(A^*))',D(A^*)} = \langle A_{-1}y(t), \varphi\rangle_{(D(A^*))',D(A^*)} \qquad \forall \varphi \in D(A^*).$$

This concept of solution by transposition is often encountered in the existing literature (see, e.g., [17] for control issues).

Example 18. Let $\Omega \subset \mathbb{R}^n$ be a C^2 bounded open set. The Cauchy problem $\partial_t y = \triangle y$ in Ω, $y_{|\partial\Omega} = 0$, $y(0) = y_0 \in L^2(\Omega)$, has a unique solution

$$y \in C^0([0,+\infty); L^2(\Omega)) \cap C^1((0,+\infty); (H_0^1(\Omega) \cap H^2(\Omega))')$$

and there exist $M \geqslant 1$ and $\omega < 0$ such that $\|y(t,\cdot)\|_{L^2(\Omega)} \leqslant M\mathrm{e}^{\omega t}\|y_0(\cdot)\|_{L^2(\Omega)}$.

Example 19. Let $\Omega \subset \mathbb{R}^n$ be a C^2 bounded open set. Consider the Cauchy problem $\partial_{tt} y = \triangle y$ in Ω, $y_{|\partial\Omega} = 0$, $y(0) = y_0$, $\partial_t y(0) = y_1$.

- 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
$$y(\cdot) \in C^0([0,+\infty); H^{-1}(\Omega)) \cap C^1((0,+\infty); (H_0^1(\Omega) \cap H^2(\Omega))').$$
- If $y_0 \in L^2(\Omega)$ and $y_1 \in H^{-1}(\Omega)$, then there is a unique solution
$$y(\cdot) \in C^0([0,+\infty); L^2(\Omega)) \cap C^1((0,+\infty); H^{-1}(\Omega)).$$

4.1.3 Scale of Banach spaces

⟨sec:scale⟩ 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 [23]). Of course, for positive integers n, the corresponding operator $A_n : D(A^n) \to D(A^{n-1})$ is the restriction of A to $D(A^n)$.

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 [61, Sections 3.6 and 3.9] for this general construction (where these spaces are called *rigged spaces*) and for further properties (see also [46, Chapter 2, Section 4.2]). In the sequel we will briefly refer to this continuous scale of Banach spaces.

The Banach space X_α, with α an arbitrary real number, can be defined for instance by symbolic calculus with real powers of the resolvent of A and complex integrals (see [38, 53, 61]), or by Banach spaces interpolation (see [47, 50, 64]), or by Fourier transform when it is possible (see [38]), or with a Hilbert basis when X is a Hilbert space and A is diagonalizable (see Remark 45 below). For instance, the construction with the fractional powers of the resolvent goes as follows, in few words (see [53, 61]). Given $\beta \in \rho(A)$ with $\mathrm{Re}(\beta) > \omega$, for $\alpha > 0$ we define[3]

$$(\beta\,\mathrm{id}_X - A)^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^{+\infty} t^{\alpha-1} e^{-\beta t} S(t)\, dt. \tag{4.6}$$ def_poweralpha

A positive power of the resolvent is defined as the inverse of the corresponding negative power. Then, for $\alpha \geqslant 0$, the space X_α is the image of X under

[3] 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 $\mathrm{Re}(z) > 0$.

$(\beta\,\mathrm{id}_X - A)^{-\alpha}$, and it is endowed with the norm $\|y\|_\alpha = \|(\beta\,\mathrm{id}_X - A)^\alpha y\|_X$. For $\alpha > 0$, the space $X_{-\alpha}$ is the completion of X for the norm $\|y\|_{-\alpha} = \|(\beta\,\mathrm{id}_X - A)^{-\alpha} y\|_X$.

The construction does not depend on the specific choice of β.

In this general framework, the operator $A_\alpha : D(A_\alpha) = X_{\alpha+1} \to X_\alpha$ (with $\alpha \in \mathbb{R}$), which is either the restriction or the extension of $A : D(A) \to X$ (with $X_0 = X$) according to the sign of α, generates the C_0 semigroup $(S_\alpha(t))_{t\geqslant 0}$.

Hereafter, when it is clear from the context, we skip the index α in $S_\alpha(t)$ or in A_α, when referring to the restriction or extension of $S(t)$ or of A to X_α.

Note that if $\alpha_1 > \alpha_2$ then $(\beta\,\mathrm{id}_X - A)^{\alpha_1-\alpha_2} : X_{\alpha_1} \to X_{\alpha_2}$ 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_α 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 $\|y\|_{X_\alpha} \leqslant C\|y\|_{X_n}^{n+1-\alpha}\|y\|_{X_{n+1}}^{\alpha-n}$ for every $y \in X_n$ (see [61]). This is an interpolation inequality, as in [50].

When X is reflexive, the operator $A^* : D(A^*) \to X'$ generates the C_0 semigroup $(S(t)^*)_{t\geqslant 0}$ (see Proposition 8). 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 28, we have

$$X_{-\alpha} = (X^*_\alpha)' \qquad \forall \alpha \in \mathbb{R} \tag{4.7}$$ Xalphastar

where the dual is taken with respect to the pivot space X.

Cauchy problems. The above general construction allows one to generalize in a wide sense the concept of strong or weak solution. Indeed, it follows from Theorem 27 that the Cauchy problem

$$\begin{aligned} \dot{y}(t) &= Ay(t) \\ y(0) &= y_0 \in X_\alpha \end{aligned}$$

has a unique solution

$$y \in C^0([0,+\infty); X_\alpha) \cap C^1((0,+\infty); X_{\alpha-1})$$

given by $y(t) = S(t)y_0$. Here, we have skipped the index α, but it is understood that $A = A_{\alpha-1}$ in the differential equation above. The differential equation is written in $X_{\alpha-1}$ and the derivative is computed with the norm $\|\cdot\|_{X_{\alpha-1}}$.

⟨rem41⟩ **Remark 45.** 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 $A : D(A) \to X$ is a self-adjoint positive operator with A^{-1} compact. Then there exists a normalized Hilbert basis $(e_j)_{j\in\mathbb{N}^*}$ of eigenvectors of A, associated with eigenvalues $(\lambda_j)_{j\in\mathbb{N}^*}$. One has

$$Ay = \sum_{j=1}^{+\infty} \lambda_j (e_j, y)_X e_j \quad \text{on} \quad D(A) = \Big\{ y \in X \;\Big|\; \sum_{j=1}^{+\infty} \lambda_j^2 (e_j, y)_X^2 < +\infty \Big\}$$

and then A^α is defined for every real number α in a spectral way by

$$A^\alpha y = \sum_{j=1}^{+\infty} \lambda_j^\alpha (e_j, y)_X e_j \quad \text{on} \quad D(A^\alpha) = \Big\{ y \in X \ \Big| \ \sum_{j=1}^{+\infty} \lambda_j^{2\alpha} (e_j, y)_X^2 < +\infty \Big\}.$$

⟨example_chain_Dirichlet⟩ **Example 20.** Let us consider the (negative of the) Dirichlet-Laplacian $A = -\triangle$ defined on $D(A) = \{y \in H_0^1(\Omega) \mid \triangle y \in L^2(\Omega)\}$, with $X = L^2(\Omega)$. As already said, we have $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ whenever the boundary of Ω is C^2.

As already seen in the example 17, we have then $X_{-1} = (H_0^1(\Omega) \cap H^2(\Omega))'$, where the dual is taken with respect to the pivot space $X = L^2(\Omega)$. We can define $A^{1/2}$ (often denoted $\sqrt{-\triangle}$) in a spectral way as above.

Assuming that the boundary of Ω is C^∞, the spaces $D(A^{j/2})$, $j \in \mathbb{N}$ (called Dirichlet spaces), are the Sobolev spaces with (the so-called) Navier boundary conditions, defined by

$$\begin{aligned}
D(A^{1/2}) &= H_0^1(\Omega) = \{y \in H^1(\Omega) \ \mid \ y_{|\partial\Omega} = 0\} \\
D(A) &= H_0^1(\Omega) \cap H^2(\Omega) = \{y \in H^2(\Omega) \ \mid \ y_{|\partial\Omega} = 0\} \\
D(A^{3/2}) &= \{y \in H^3(\Omega) \ \mid \ y_{|\partial\Omega} = \triangle y_{|\partial\Omega} = 0\} \\
D(A^2) &= \{y \in H^4(\Omega) \ \mid \ y_{|\partial\Omega} = \triangle y_{|\partial\Omega} = 0\} \\
D(A^{5/2}) &= \{y \in H^5(\Omega) \ \mid \ y_{|\partial\Omega} = \triangle y_{|\partial\Omega} = \triangle^2 y_{|\partial\Omega} = 0\} \\
D(A^3) &= \{y \in H^6(\Omega) \ \mid \ y_{|\partial\Omega} = \triangle y_{|\partial\Omega} = \triangle^2 y_{|\partial\Omega} = 0\}
\end{aligned}$$

in other words we have

$$D(A^{j/2}) = \Big\{ y \in H^j(\Omega) \ \Big| \ y_{|\partial\Omega} = \triangle y_{|\partial\Omega} = \cdots = \triangle^{\left[\frac{j-1}{2}\right]} y_{|\partial\Omega} = 0 \Big\}$$

for every $j \in \mathbb{N}^*$, and moreover the operator $A^{j/2} : D(A^{j/2}) \to L^2(\Omega)$ is an isomorphism (see [66] for other properties). It can be noted that $\|A^{j/2} y\|_{L^2(\Omega)} = \|\triangle^{j/2} y\|_{L^2(\Omega)}$ if j is even and $\|A^{j/2} y\|_{L^2(\Omega)} = \|\triangle^{j/2} y\|_{H_0^1(\Omega)} = \|\nabla \triangle^{j/2} y\|_{L^2(\Omega)}$ if j is odd.

Note that, omitting the indices, we have the scale of Hilbert spaces

$$\cdots \xrightarrow{A^{1/2}} D(A) \xrightarrow{A^{1/2}} D(A^{1/2}) \xrightarrow{A^{1/2}} L^2(\Omega) \xrightarrow{A^{1/2}} D(A^{1/2})' \xrightarrow{A^{1/2}} D(A)' \xrightarrow{A^{1/2}} \cdots$$

with $D(A^{1/2})' = H^{-1}(\Omega)$ and $D(A)' = (H_0^1(\Omega) \cap H^2(\Omega))'$ (with respect to the pivot space $L^2(\Omega)$). All mappings $A^{1/2}$, between the corresponding spaces, are isometric isomorphims.

As in the previous remark, we can even define $X_\alpha = D(A^\alpha)$ (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 [50]), 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 46. Using Proposition 8, if X is reflexive then all these results can be stated as well for the dual operator A^* and the dual C_0 semigroup $S(t)^*$.

4.2 Nonhomogeneous Cauchy problems

⟨chap4_sec4.2⟩ Let $y_0 \in X$. We consider the Cauchy problem

$$\begin{aligned} &\dot{y}(t) = Ay(t) + f(t) \quad \forall t > 0 \\ &y(0) = y_0 \end{aligned} \tag{4.8}$$ cp1

where $A : D(A) \to X$ generates a C_0 semigroup $(S(t))_{t\geqslant 0}$ on X.

⟨thm32⟩ **Proposition 9.** *If $y_0 \in D(A)$ and $f \in L^1_{\mathrm{loc}}(0,+\infty;D(A))$ with $1 \leqslant p \leqslant +\infty$, then (4.8) has a unique solution $y \in C^0([0,+\infty);D(A)) \cap W^{1,1}_{\mathrm{loc}}(0,+\infty;X)$ (often referred to as strong solution of (4.8)) given by*

$$y(t) = S(t)y_0 + \int_0^t S(t-s)f(s)ds. \tag{4.9}$$ eqq1

Morever, the differential equation (4.8) makes sense in X.

Proof. The function y defined by (4.9) is clearly a solution of (4.8). 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 $0 = S(t)z(0) = S(0)z(t) = z(t)$. □

Note that, if $f \in L^1_{\mathrm{loc}}(0,+\infty;X)$, then (4.9) still makes sense. Note also that, using the extension of A (and of $S(t)$) to X_{-1}, Proposition 9 implies that, if $y_0 \in X$ and $f \in L^1_{\mathrm{loc}}(0,+\infty;X)$, then (4.8) has a unique solution $y \in C^0([0,+\infty);X) \cap W^{1,1}_{\mathrm{loc}}(0,+\infty;X_{-1})$ given as well by the Duhamel formula (4.9) (and often referred to as *mild solution* of (4.8)), and the differential equation (4.8) is written in X_{-1} (see, e.g., [53, Chapter 4]). Moreover, for every $T > 0$ there exists $K_T > 0$ (not depending on y_0 and f) such that $\|y(t)\|_X \leqslant C_T(\|y_0\|_X + \|f\|_{L^1(0,T;X)})$.

More generally, using the general scale of Banach spaces $(X_\alpha)_{\alpha\in\mathbb{R}}$ mentioned previously, we have the following result (see [23] or [61, Theorem 3.8.2]).

⟨propXalpha⟩ **Proposition 10.** *If $f \in L^1_{\mathrm{loc}}(0,+\infty;X_\alpha)$ for some $\alpha \in \mathbb{R}$, then for every $y_0 \in X_\alpha$ the Cauchy problem (4.8) has a unique solution*

$$y \in C^0([0,+\infty);X_\alpha) \cap W^{1,1}_{\mathrm{loc}}(0,+\infty;X_{\alpha-1})$$

given as well by (4.9) (called strong solution in X_α in [61, 66]). Here, we have $A = A_{\alpha-1}$ in the equation (4.8) which is written in $X_{\alpha-1}$ almost everywhere, and the integral in (4.9) is done in X_α (with $S(t-s) = S_\alpha(t-s)$ in the integral).

Proposition 9 corresponds to $\alpha = 1$.

Remark 47. The regularity stated in Proposition 10 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 [23, 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 [23, 53, 66]).

⟨lem_moreregular⟩ **Lemma 5.** *If $y_0 \in X_{\alpha+1}$ and $f \in W^{1,1}_{\rm loc}(0,+\infty;X_\alpha)$ then (4.8) has a unique solution $y \in C^0([0,+\infty);X_{\alpha+1}) \cap C^1((0,+\infty);X_\alpha)$ given by (4.9).*

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 $X_{\alpha-1}$.

⟨rem_regular2⟩ **Remark 48.** Let us consider the Cauchy problem

$$\partial_t y = \triangle y + f \ \text{ in } \Omega, \qquad y_{|\partial\Omega} = 0, \qquad y(0) = y_0 \in L^2(\Omega),$$

with $f \in L^2((0,+\infty)\times\Omega)$. The general theory says that the unique solution has the regularity $y \in C^0([0,+\infty);L^2(\Omega)) \cap H^1(0,+\infty;(H^2(\Omega)\cap H^1_0(\Omega))')$.

Actually, by using a spectral expansion as in Remark 41, it is easy to prove that $y \in L^2(0,T;H^1_0(\Omega)) \cap H^1(0,T;H^{-1}(\Omega))$, which is more precise because this set is contained in $C^0([0,T],L^2(\Omega))$. Moreover, if $y_0 \in H^1_0(\Omega)$, then we have the improved regularity $y \in L^2(0,T;H^2(\Omega)\cap H^1_0(\Omega)) \cap H^1(0,T;L^2(\Omega)) \subset C^0([0,T],H^1_0(\Omega))$ (see also [24, Chapter 7.1] where these regularity properties are established by using Galerkin approximations for more general elliptic operators).

As in Remark 41, this remark shows again that the regularity properties obtained by the general semigroup theory may be much improved when using the specific features of the operator under consideration.

Chapter 5

Linear control systems in Banach spaces

?⟨chap_diminfinie⟩? Throughout the chapter, we consider the linear autonomous control system

$$\begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t) \\ y(0) &= y_0 \end{aligned} \tag{5.1}$$ eqE

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, $A : D(A) \to X$ is the generator of a C_0 semigroup $(S(t))_{t\geqslant 0} \in \mathcal{G}(M,\omega)$ 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 (note however that B is a bounded 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 could assume that $B \in L(U, X_{-\alpha})$ for some $\alpha \geqslant 0$. We will comment on that further.

A priori if $u \in L^1_{\mathrm{loc}}(0,+\infty;U)$ then $Bu \in L^1_{\mathrm{loc}}(0,+\infty;X_{-1})$, and since $y_0 \in X$, it follows from the results of Section 4.2 that (5.1) has a unique solution $y \in C^0([0,+\infty);X_{-1}) \cap W^{1,1}_{\mathrm{loc}}(0,+\infty;X_{-2})$, given by

$$y(t;y_0,u) = S(t)y_0 + L_t u \tag{5.2}$$ defyfaible

where

$$L_t u = \int_0^t S(t-s)Bu(s)\,ds. \tag{5.3}$$ def_Lt

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 $y \in C^0([0,+\infty);X) \cap W^{1,1}_{\mathrm{loc}}(0,+\infty;X_{-1})$ 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, under which the situation 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}).

Of course, the notion of admissibility depends on the time-regularity of the inputs u. Since it 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 $1 < p \leqslant +\infty$, see [61] 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 [48, 49] in order to characterize the spaces where exact controllability holds true.

Most of this chapter is borrowed from [66] (see also [49, 61, 71]).

5.1 Admissible control operators

?⟨sec_admissible⟩? As said previously, we have a priori the inclusion $\mathrm{Ran}(L_T) \subset X_{-1}$, for every $T > 0$, and the fact that $L_T \in L(L^2(0,T;U), X_{-1})$.

5.1.1 Definition

Definition 15. *A control operator $B \in L(U, X_{-1})$ is said to be admissible for the C_0 semigroup $(S(t))_{t\geqslant 0}$ if there exists $T > 0$ such that $\mathrm{Ran}(L_T) \subset X$.*

⟨lemequivadmissible⟩ **Lemma 6.** *The following properties are equivalent:*

- *There exists $T > 0$ such that $\mathrm{Ran}(L_T) \subset X$.*
- *For every $T > 0$, one has $\mathrm{Ran}(L_T) \subset X$.*
- *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 $\mathrm{Ran}(L_T) \subset X$. Let us prove that $\mathrm{Ran}(L_T) \subset X$ 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 $\mathrm{Ran}(L_T) \subset X$ then $\mathrm{Ran}(L_T) \subset X$, 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 $\mathrm{Ran}(L_T) \subset X$ then $\mathrm{Ran}(L_{2T}) \subset X$, and by immediate iteration, this implies as well that $\mathrm{Ran}(L_{kT}) \subset X$ for every $k \in \mathbb{N}^*$.

Now, let $t > T$ arbitrary, and let $k \in \mathbb{N}^*$ be such that $kT > t$. Since $\mathrm{Ran}(L_{kT}) \subset X$, it follows from the first part of the proof that $\mathrm{Ran}(L_T) \subset X$.

It remains to prove that, if $\mathrm{Ran}(L_T) \subset X$, then $L_T \in L(L^2(0,T;U),X)$. Note first that the operator L_T is closed. Indeed, we have $L_T u = (\beta\,\mathrm{id}_X - A)\int_0^T S(T-t)(\beta\,\mathrm{id}_X - A)^{-1}Bu(t)\,dt$, for every $u \in L^1(0,T;U)$, with $\beta \in \rho(A)$ arbitrary. By definition of X_{-1}, the operator $(\beta\,\mathrm{id}_X - A)^{-1}B$ is linear and continuous from U to X. Since A is closed (according to Proposition 5), it follows that L_T is closed. A priori, the graph of L_T is contained in X_{-1}. Under the assumption that $\mathrm{Ran}(L_T) \subset X$, 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. □

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 $A : D(A) \to X = L^2(\Omega)$, where Ω is a domain of $\mathbb{R}^n$ and ω is a measurable subset of Ω.

Unbounded control operators are obtained for instance when one considers a control acting along the boundary of Ω (see further for examples).

Remark 49. 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.

?⟨rem_Lp⟩? **Remark 50.** 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 $p \geqslant 1$, 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 [69, 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.

⟨lem_LT*⟩ **Lemma 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 28). 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 [11]), 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 $\beta \in \rho(A)$,

$$\begin{aligned} B_{D(A^*)} &= \{z \in D(A^*) \;\mid\; \|z\|_{D(A^*)} = \|(\beta \mathrm{id}_{X'} - A^*)z\|_{X'} \leqslant 1\} \\ &= \{(\beta \mathrm{id}_{X'} - A^*)^{-1} f \;\mid\; f \in X',\ \|f\|_{X'} \leqslant 1\} \\ &= (\beta \mathrm{id}_{X'} - A^*)^{-1} B_{X'} \end{aligned}$$

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

$$\begin{aligned} \langle L_T u, z\rangle_{D(A^*)',D(A^*)} &= \left\langle \int_0^T S(T-t)Bu(t)\,dt, z \right\rangle_{D(A^*)',D(A^*)} \\ &= \int_0^T \langle S(T-t)Bu(t), z\rangle_{D(A^*)',D(A^*)}\,dt \\ &= \int_0^T \langle B^*S(T-t)^*z, u(t)\rangle_{U',U}\,dt \\ &= \langle t \mapsto B^*S(T-t)^*z, t \mapsto u(t)\rangle_{L^2(0,T;U)',L^2(0,T;U)} \end{aligned}$$

and the conclusion follows. □

The following proposition, providing a dual characterization of admissibility, is then an immediate consequence of Lemmas 6 and 7. Indeed in the admissible case we have $L_T \in L(L^2(0,T;U),X)$ and equivalently $L_T^* \in L(X',L^2(0,T;U)')$.[1]

[1] 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 $L(X',L^2(0,T;U)') = L(X',L^2(0,T;U'))$.

⟨charactdualadm⟩ **Proposition 11.** *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*

$$\int_0^T \|B^* S(T-t)^* z\|_{U'}^2 \, dt \leqslant K_T \|z\|_{X'}^2 \qquad \forall z \in D(A^*). \tag{5.4}$$ ineg_adm

Remark 51. 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 (that is, there is no loss of compactness).

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 52. 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 [66, 68]).

⟨rem_Lebext⟩ **Remark 53.** 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 [68]. 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 Λ-*extension*, defined by

$$B_\Lambda^* z = \lim_{\lambda \to +\infty} B^* \lambda (\lambda \mathrm{id}_X - A^*)^{-1} z$$

also called *strong Yosida extension* in [61, Definition 5.4.1] (note that $(\mathrm{id}_X - A^*)^{-1} z \in D(A^*)$ 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 11 still holds true with B^* replaced with B_Λ^*, with the inequality (5.4) written for every $z \in X'$.

Actually, in the context of Lemma 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 $t \in [0, T]$.

rem_degree_unboundedness⟩ **Remark 54.** As briefly mentioned in the introduction, so far we have focused on control operators $B \in L(U, X_{-1})$. Using the general construction of the scale of Banach spaces $(X_\alpha)_{\alpha \in \mathbb{R}}$ done in Section 4.1.3, we can more generally define (unbounded) control operators such that $B \in L(U, X_{-\alpha})$, for some $\alpha > 0$ (for $\alpha = 0$, the control operator is bounded).

By definition, the *degree of unboundedness* $\alpha(B) \geqslant 0$ of the control operator B (with respect to the spaces X and U) is the infimum of the set of $\alpha \geqslant 0$ such that $B \in L(U, X_{-\alpha})$ (see [56, 61, 66]), that is, such that $(\beta \, \mathrm{id}_X - A)^{-\alpha} B \in$

$L(U,X)$, for some arbitrary $\beta \in \rho(A)$ such that $\mathrm{Re}(\beta) > \omega$, where $(\beta\,\mathrm{id}_X - A)^{-\alpha}$ is defined by (4.6). Equivalently, $\alpha(B)$ is equal to the infimum of the set of $\alpha \geqslant 0$ for which there exists $C_\alpha > 0$ such that

$$\|(\lambda \mathrm{id}_X - A)^{-1}B\| \leqslant \frac{C_\alpha}{\lambda^{1-\alpha}} \qquad \forall \lambda > \omega.$$

When X is reflexive, $\alpha(B)$ is the infimum of the set of $\alpha \geqslant 0$ such that $B^* \in L(X_\alpha^*, U')$ (where $X_\alpha^* = D((\beta\,\mathrm{id}_X - A^*)^\alpha)$ thanks to (4.7)), or equivalently, of the set of $\alpha \geqslant 0$ for which there exists $C_\alpha > 0$ such that

$$\|B^*z\|_{U'} \leqslant C_\alpha \|(\beta\,\mathrm{id}_X - A^*)^\alpha z\|_{X'} \qquad \forall z \in X_\alpha^*. \tag{5.5}$$ inegalpha

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 $\alpha(B) \leqslant 1$. This covers the most usual applications. Internal controls are bounded control operators (thus, $\alpha(B) = 0$). For boundary controls, in general we have $\alpha(B) \leqslant 1$ (see [43, 66] for many examples).

⟨lem_admissible⟩ **Lemma 8.** *1. Assume that X and U are reflexive. If B is admissible then $B \in L(U, X_{-1/2})$ (and thus $\alpha(B) \leqslant 1/2$).*

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.

Actually, the second point of the lemma is true under the more general assumption that the semigroup generated by A is either analytic and normal or invertible (see [70]).

Proof. Let us prove the first point.

First of all, by an obvious change of variable, we have $L_{t+s} = S(s)L_t + L_s$, for all $s, t \geqslant 0$. It follows that $L_n = (S(n-1) + \cdots + S(1) + \mathrm{id}_X)L_1$, for every $n \in \mathbb{N}^*$. Since $(S(t))_{t\geqslant 0} \in \mathcal{G}(M,\omega)$, we have $\|S(k)\| \leqslant Me^{k\omega}$ for every integer k, and therefore, we infer that $\|L_n\| \leqslant K_n\|L_1\|$, 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 $L^2_{\mathrm{loc}}(0,+\infty;U')$ 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 $\|L_{t_1}\| \leqslant \|L_{t_2}\|$.

Now, for an arbitrary $T > 0$, let $n \in \mathbb{N}$ be such that $n \leqslant T < n+1$. Writing $L_T = S(T-n)L_n + L_{T-n}$, we get that $\|L_T\| \leqslant Me^{\omega(T-n)}\|L_n\| + \|L_1\|$.

It finally follows that $\|L_T\| \leqslant Ke^{\omega T}$ if $\omega > 0$, $\|L_T\| \leqslant KT$ if $\omega = 0$, and $\|L_T\| \leqslant K$ 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 $z \in D(A^*)$, we define $(\Psi z)(t) = B^*S(t)^*z$. 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 Ψz. On the one part, we have, by an easy computation as in (4.4), $\mathcal{L}(\Psi z)(s) = \int_0^{+\infty} e^{-st}(\Psi z)(t)\,dt = B^*(s\,\mathrm{id}_X - A^*)z$

for every $s \in \rho(A)$ such that $\mathrm{Re}(s) > \omega$. 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

$$\|\mathcal{L}(\Psi z)(s)\|_{U'} \leqslant \frac{1}{\sqrt{2(\mathrm{Re}(s)-\alpha)}} \|t \mapsto e^{-\alpha t}(\Psi z)(t)\|_{L^2(0,+\infty;U')}.$$

The first point is proved.

Let us prove the second point. By definition, there exists $C > 0$ such that

$$\|B^* z\|_{U'}^2 \leqslant C \|(\beta\,\mathrm{id}_X - A^*)^{1/2} z\|_{X'}^2 = C(z, (\beta\,\mathrm{id}_X - A)z)_X$$

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

$$\int_0^T e^{2\beta t} \|B^* S^*(T-t)\psi\|_{U'}^2\,dt \leqslant C \int_0^T e^{2\beta t}(z(t), (\beta\,\mathrm{id}_X - A)z(t))_X\,dt.$$

Since $\dot{z}(t) = -Az(t)$ and $z(T) = \psi$, we have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left(e^{2\beta t}\|z(t)\|_X^2\right) &= e^{2\beta t}\left(\beta\|z(t)\|_X^2 - (z(t), Az(t))_X\right)\\ &= e^{2\beta t}(z(t), (\beta\,\mathrm{id}_X - A)z(t))_X\end{aligned}$$

and therefore we get

$$\int_0^T e^{2\beta t}\|B^* S^*(T-t)\psi\|_{U'}^2\,dt \leqslant \frac{C}{2}\int_0^T \frac{d}{dt}\left(e^{2\beta t}\|z(t)\|_X^2\right)\,dt \leqslant \frac{C}{2} e^{2\beta T}\|\psi\|_X^2.$$

The lemma follows. □

5.1.3 Examples

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. Consider the heat equation with internal control and Dirichlet boundary conditions

$$\partial_t y = \triangle y + \chi_\omega u \ \text{ in } \Omega, \qquad y_{|\partial\Omega} = 0, \qquad y(0) = y_0 \in L^2(\Omega).$$

We set $X = L^2(\Omega)$, and $A = -\triangle : D(A) \to X$, 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 Ω. It is bounded and therefore admissible, which means that, for every $T > 0$, there exists $K_T > 0$ such that

$$\int_0^T \int_\omega \psi(t,x)^2\,dx\,dt \leqslant K_T \|\psi(0)\|_{L^2(\Omega)}^2$$

for every solution of $\partial_t \psi = \triangle\psi$ in Ω, $\psi_{|\partial\Omega} = 0$, with $\psi(0) \in H^2(\Omega) \cap H_0^1(\Omega)$. By the way, noting that $\psi(t) = S(t)\psi(0)$ with $S(t) \in L(X)$, we recover as well the fact that this inequality holds true.

Heat equation with Dirichlet boundary control. Let $\Omega \subset \mathbb{R}^n$ be a bounded open set having a C^1 boundary. Consider the heat equation with Dirichlet boundary control

$$\partial_t y = \triangle y \ \text{ in } \Omega, \qquad y_{|\partial\Omega} = u, \qquad y(0) = y_0 \in L^2(\Omega). \tag{5.6}$$ boundarycontrolheat

We set $U = L^2(\partial\Omega)$, $X = H^{-1}(\Omega)$, and $A = -\triangle : D(A) \to X$, where $D(A) = X_1 = H^1_0(\Omega)$. The operator A is self-adjoint, hence $D(A^*) = D(A)$. We have $X_{-1} = D(A^*)'$ with respect to the pivot space $H^{-1}(\Omega)$, and thus, if $\partial\Omega$ is smooth then X_{-1} is the dual of $A^{-1}(H^1_0(\Omega)) = \{y \in H^3(\Omega) \ \mid \ y_{|\partial\Omega} = (\triangle y)_{|\partial\Omega} = 0\}$ with respect to the pivot space $L^2(\Omega)$ (see Example 20).

Let us express the control operator $B \in L(U, D(A^*)')$. We proceed by density. Taking a solution y regular enough, associated with a control u (for instance, of class C^1), and using Lemma 5, we have, on the one hand, by integration by parts, $(\partial_t y, A^{-1}\phi)_{L^2(\Omega)} = (y, \phi)_{L^2(\Omega)} - \int_{\partial\Omega} u \frac{\partial\phi}{\partial\nu}_{|\partial\Omega} d\sigma$, and on the other hand, by definition, $\langle \partial_t y, \phi\rangle_{X_{-1},X_1} = \langle Ay, \phi\rangle_{X_{-1},X_1} + \langle Bu, \phi\rangle_{X_{-1},X_1}$ where the duality bracket is considered with respect to the pivot space $X = H^{-1}(\Omega)$, yielding that $\langle Ay, \phi\rangle_{X_{-1},X_1} = (y, A^*\phi)_{H^{-1}(\Omega)} = (y, \phi)_{L^2(\Omega)}$. Hence

$$B^*\phi = -\frac{\partial}{\partial\nu}_{|\partial\Omega}(A^{-1}\phi) \qquad \forall \phi \in D(A^*).$$

This defines B by transposition, by $\langle Bu, \phi\rangle_{X_{-1},X_1} = -\int_{\partial\Omega} u \frac{\partial}{\partial\nu}_{|\partial\Omega}(A^{-1}\phi)\, d\sigma$ for every $u \in L^2(\partial\Omega)$ and every $\phi \in H^1_0(\Omega)$.

If $\partial\Omega$ is C^2 then we can express the operator B by using the Dirichlet map as follows (see [50] and [66, Chapters 10.6 and 10.7]). Assuming that Ω has a C^2 boundary, we define the linear bounded operator $D : L^2(\partial\Omega) \to H^{1/2}(\Omega)$ as the adjoint of $D^* \in L(H^{-1/2}(\Omega), L^2(\partial\Omega))$ given by $D^*g = -\frac{\partial}{\partial\nu}_{|\partial\Omega}(A^{-1}g)$, for every $g \in L^2(\Omega)$. Equivalently, given any $v \in L^2(\partial\Omega)$, Dv is the unique solution in the sense of distributions of the Laplace equation $\triangle(Dv) = 0$ with the boundary condition $(Dv)_{|\partial\Omega} = v$. Note that $Dv \in C^\infty(\Omega)$ by hypoellipticity; in particular, the Dirichlet map D is not surjective. Then a computation shows that $B = AD$ where we consider the extension $A : L^2(\Omega) \to (H^2(\Omega) \cap H^1_0(\Omega))'$ (dual with respect to $L^2(\Omega)$). Note that $X_{1/2} = L^2(\Omega)$ and that $X_{-1/2} = (H^2(\Omega) \cap H^1_0(\Omega))'$. In particular, we have $B \in L(U, X_{-1/2})$ (and $B^* \in L(X_{1/2}, U)$), and thus $\alpha(B) \leqslant 1/2$ (actually, by refining the reasoning, we see that $\alpha(B) = 1/4$). It follows from Lemma 8 that B is admissible, which is, by Proposition 11, equivalent to the fact that, for every $T > 0$, there exists $K_T > 0$ such that

$$\int_0^T \left\| \frac{\partial\psi}{\partial\nu}_{|\partial\Omega}(t) \right\|^2_{L^2(\partial\Omega)} dt \leqslant K_T \|\psi(0)\|^2_{H^1_0(\Omega)}$$

for every solution of $\partial_t\psi = \triangle\psi$ in Ω, $\psi_{|\partial\Omega} = 0$, with $\psi(0) \in H^2(\Omega) \cap H^1_0(\Omega)$. This result says that the heat equation with boundary control (5.6) is well posed in the state space $X = H^{-1}(\Omega)$, with $U = L^2(\partial\Omega)$.

It is interesting to 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, the supremum of $\int_0^T \left\| \frac{\partial \psi}{\partial \nu}_{|\partial\Omega}(t) \right\|^2_{L^2(\partial\Omega)} dt$ over all possible $\psi(0) \in L^2(\Omega)$ such that $\|\psi(0)\|_{L^2(\Omega)} = 1$ is equal to $+\infty$ (see [49, 50]). 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 [43, 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, in accordance with Lemma 8.

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 $A = -\triangle$ 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 the details. Actually, we have $B \in L(U, (D(A^{1/4+\varepsilon}))')$ for every $\varepsilon > 0$ (see [43, Chapter 3.3]), thus $\alpha(B) = 1/4$, and hence B is admissible by Lemma 8.

Second-order equations. Another typical example is provided by second-order equations. The framework is the following (see [66]). Let H be a Hilbert space, and $A_0 : D(A_0) \to H$ be self-adjoint and strictly 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 45). We set $X = D(A_0^{1/2}) \times H$, and we define the skew-adjoint operator $A : D(A) \to X$ on $D(A) = D(A_0) \times D(A_0^{1/2})$ by

$$A = \begin{pmatrix} 0 & I \\ -A_0 & 0 \end{pmatrix}.$$

Let $B_0 \in L(U, D(A_0^{1/2})')$, where U is a Hilbert space, and $D(A_0^{1/2})'$ is the dual of $D(A_0^{1/2})$ with respect to the pivot space H. We consider the second-order control system

$$\partial_{tt} y + A_0 y = B_0 u, \qquad y(0) = y_0, \ \partial_t y(0) = y_1.$$

It can be written in the form

$$\frac{\partial}{\partial t}\begin{pmatrix} y \\ \partial_t y \end{pmatrix} = A\begin{pmatrix} y \\ \partial_t y \end{pmatrix} + Bu \qquad \text{with} \qquad B = \begin{pmatrix} 0 \\ B_0 \end{pmatrix}.$$

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 $B^* \in L(D(A_0) \times D(A_0^{1/2}), U)$ is given by $B^* = \begin{pmatrix} 0 \\ B_0^* \end{pmatrix}$.

Proposition 12. *The following statements are equivalent:*

- *B is admissible.*
- *There exists $K_T > 0$ such that every solution of*

$$\partial_{tt}\psi + A_0\psi = 0, \qquad \psi(0) \in D(A_0), \quad \partial_t\psi(0) \in D(A_0^{1/2})$$

satisfies $\displaystyle\int_0^T \|B_0^*\partial_t\psi(t)\|_{U'}^2\,dt \leqslant K_T\left(\|\psi(0)\|^2_{D(A_0^{1/2})} + \|\partial_t\psi(0)\|_H^2\right).$

- *There exists $K_T > 0$ (the same constant) such that every solution of*

$$\partial_{tt}\psi + A_0\psi = 0, \qquad \psi(0) \in H, \quad \partial_t\psi(0) \in D(A_0^{1/2})'$$

satisfies $\displaystyle\int_0^T \|B_0^*\psi(t)\|_{U'}^2\,dt \leqslant K_T\left(\|\psi(0)\|_H^2 + \|\partial_t\psi(0)\|^2_{D(A_0^{1/2})'}\right).$

?⟨examplewave⟩? **Example 21.** As a particular case, let us consider the wave equation with Dirichlet boundary control

$$\partial_{tt}y = \triangle y \quad \text{in } \Omega, \qquad y_{|\partial\Omega} = u.$$

We assume that Ω has a C^2 boundary. We set $H = H^{-1}(\Omega)$, and we take $A_0 = -\triangle : D(A_0) = H_0^1(\Omega) \to H$ (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 $X_1 = D(A) = H_0^1(\Omega)\times L^2(\Omega)$ and $X_{-1} = H^{-1}(\Omega)\times(H^2(\Omega)\cap H_0^1(\Omega))'$. The spaces X_α can be computed 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 $B_0^*\phi = \frac{\partial}{\partial\nu}(A_0^{-1}\phi)_{|\partial\Omega}$ for every $\phi \in L^2(\Omega)$, or, equivalently, $B_0 \in L(U, D(A_0^{1/2})')$ is given by $B_0u = A_{-1}Du$ for every $u \in U$, where D is the Dirichlet mapping. Then we obtain that $B \in L(U, X_{-1/2})$, and therefore, since A is normal, it follows from Lemma 8 that B is admissible. This is equivalent to the existence of $K_T > 0$ such that

$$\int_0^T \left\|\frac{\partial\psi}{\partial\nu}(t)\right\|^2_{L^2(\partial\Omega)} dt \leqslant K_T\left(\|\psi_0\|^2_{H_0^1(\Omega)} + \|\psi_1\|^2_{L^2(\Omega)}\right)$$

for every $\psi \in C^0([0,T];H^2(\Omega))\cap C^1((0,T);H^1(\Omega))$ solution of

$$\partial_{tt}\psi = \triangle\psi \ \text{ in } \Omega, \qquad \psi_{|\partial\Omega} = 0, \qquad \psi(0) = \psi_0,\ \partial_t\psi(0) = \psi_1.$$

This inequality is usually referred to as a *hidden regularity* property for the Dirichlet wave equation (see [48, 49, 50]).

We refer to [43, 66] (and literature cited therein) for many other examples.

5.2 Controllability

Throughout the section, we consider the linear control system (5.1) with the framework given in the introduction of the chapter. 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 $u \in L^2(0,T;U)$ such that the corresponding solution of (5.1), with $y(0) = y_0$, 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 Ω of $\mathbb{R}^n$, with either an internal or a boundary control. Due to the regularizing effect (see Remark 41, see also [13]), 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.

⟨defnotionscont⟩ **Definition 16.** *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 $\|y(T; y_0, u) - y_1\|_X \leqslant \varepsilon$;*
- *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 $y(T; y_0, u) = 0$.*

⟨remprelimdualitycontobs⟩ **Remark 55.** Using the representation of the solutions $y(T) = S(T)y_0 + L_T u$ (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 $\mathrm{Ran}(L_T) = X$. In particular, if this is true, then B must be admissible and thus $\alpha(B) \leqslant 1/2$.
- The control system (5.1) is approximately controllable in X in time T if and only if $\mathrm{Ran}(L_T) \cap X$ is dense in X.
- The control system (5.1) is exactly null controllable in X in time T if and only if $\mathrm{Ran}(S(T)) \subset \mathrm{Ran}(L_T)$.

Remark 56. Note that, if $\mathrm{Ran}(L_T) = X$ for some $T > 0$, then $\mathrm{Ran}(L_t) = X$ for every $t \geqslant T$. Indeed, taking (as in the proof of Lemma 6) controls such that $u = 0$ on $(0, t-T)$, we have $L_t u = \int_{t-T}^{t} S(t-s)u(s)\,ds = \int_0^T S(T-\tau)u(\tau)\,d\tau = L_T u$. This shows that if the control system (5.1) is exactly controllable in time T then it is exactly controllable as well in any time $t \geqslant T$.

Remark 57. We speak of approximate null controllability in time T when one takes the target $y_1 = 0$ in Definition 16; equivalently, $\mathrm{Ran}(S(T))$ is contained in the closure of $\mathrm{Ran}(L_T)$. Approximate controllability and approximate null controllability (in time T) coincide when $S(T)^*$ is injective, i.e., when $\mathrm{Ran}(S(T))$ is dense in X.

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., [17, 66] 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 55 with the following general lemma of functional analysis (see [11] for the first part and [46, 71] for the last part).

⟨lemgenanafonc⟩ **Lemma 9.** *Let X and Y be Banach spaces, and let $F \in L(X,Y)$. Then:*

- $\mathrm{Ran}(F)$ *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 $z = 0$.*
- $\mathrm{Ran}(F) = Y$ *(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 $\|F^* z\|_{X'} \geqslant C\|z\|_{Y'}$ for every $z \in Y'$.*

Let X, Y and Z be Banach spaces, with Y reflexive, and let $F \in L(X,Z)$ and $G \in L(Y,Z)$. Then $\mathrm{Ran}(F) \subset \mathrm{Ran}(G)$ if and only if there exists $C > 0$ such that $\|F^* z\|_{X'} \leqslant C\|G^* z\|_{Y'}$ 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 9 to the operators L_T and $S(T)$, we get, with Remark 55, the following result.

⟨dualityondes⟩ **Theorem 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 that*

$$\int_0^T \|B^* S^*(T-t)z\|_{U'}^2\,dt \geqslant C_T \|z\|_{X'}^2 \qquad \forall z \in D(A^*). \tag{5.7}$$ inegobs

- *The control system (5.1) is approximately controllable in X in time T if and only if the following implication holds true:*

$$\forall z \in D(A^*) \quad \forall t \in [0,T] \quad B^*S^*(T-t)z = 0 \quad \Rightarrow \quad z = 0. \tag{5.8}$$ prolongunique

- *The control system (5.1) is exactly null controllable in X in time T if and only if there exists $C_T > 0$ such that*

$$\int_0^T \|B^*S^*(T-t)z\|_{U'}^2 \, dt \geqslant C_T \|S(T)^*z\|_{X'}^2, \qquad \forall z \in D(A^*). \tag{5.9}$$ obscont

⟨rem51⟩ **Remark 58.** The inequalities (5.7) and (5.9) are called *observability inequalities*. As in Remark 53, they can be written for every $z \in X'$, provided that B^* be replaced with its Λ-extension B_Λ^*. 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 from below. Proving observability inequalities is a challenging issue in general for PDEs. We will mention some examples further.

Remark 59. The implication (5.8) corresponds exactly, in the context of PDEs, to a *unique continuation* property. A classical tool used to derive such a property is the Holmgren theorem (see [49, 50]).

Remark 60. Setting $\varphi(t) = S^*(T-t)z$, we have

$$\dot{\varphi}(t) = -A^*\varphi(t), \qquad \varphi(T) = z,$$

and the properties above can be interpreted in terms of $\varphi(t)$ which is the infinite-dimensional version of the adjoint vector of the finite-dimensional setting.

Usually, we rather consider $\psi(t) = \varphi(T-t) = S^*(t)z$, and hence we have

$$\dot{\psi}(t) = A^*\psi(t), \qquad \psi(0) = z.$$

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

$$\int_0^T \|B^*\psi(t)\|_{U'}^2 \, dt \geqslant C_T \|\psi(0)\|_{X'}^2$$

for every solution of the adjoint equation, and the exact null controllability is equivalent to the observability inequality

$$\int_0^T \|B^*\psi(t)\|_{U'}^2 \, dt \geqslant C_T \|\psi(T)\|_{X'}^2.$$

Remark 61. As announced in Remark 6 in Chapter 1, the observability inequality (5.7) is the infinite-dimensional version of the observability inequality (1.2) 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 6 and in Theorem 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 7). Identifying $U \simeq U'$ and $L^2(0,T;U) \simeq L^2(0,T;U')$, we define the Gramian operator

$$G_T = L_T L_T^* = \int_0^T S(T-t)BB^*S(T-t)^*\,dt \ \in L(D(A^*),(D(A^*))') \quad (5.10)$$ def_GT

where, in the formula above, BB^* is to be understood as BJB^* where $J : U' \to U$ is the canonical isomorphism.

If 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 we have that $G_T \in L(X',X)$. The expression of G_T is still given by (5.10) when applied to some element $z \in D(A^*)$. Using Remark 53, 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 58, it follows that $\langle G_T z, z\rangle_{D(A^*)',D(A^*)} \geqslant C_T\|z\|^2_{X'}$ for every $z \in D(A^*)$; the converse inequality is satisfied if B is admissible. In other words, we have the following lemma.

⟨lemGT⟩ **Lemma 10.** *The control operator B is admissible and the control system (5.1) is exactly controllable in time T if and only if $G_T : X' \to X$ is an isomorphism satisfying $C_T\|z\|^2_{X'} \leqslant \langle G_T z, z\rangle_{X',X} \leqslant K_T\|z\|^2_{X'}$ for every $z \in X'$.*

5.2.3 Hilbert Uniqueness Method

The Hilbert Uniqueness Method (in short, HUM; see [49]) is based on Lemma 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 $A : D(A) \to Y$ 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 exactly the Banach space for which Lemma 10 is satisfied, i.e., B is admissible and the control system (5.1) is exactly controllable in time T, for this 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 conditions of Lemma 10, the functional J defined by

$$J(\psi) = \frac{1}{2}\langle G_T z, z\rangle_{X',X} + \langle z, S(T)y_0 - y_1\rangle_{X',X} \qquad \forall \psi \in X'$$

is smooth and coercive in X', hence J has a unique minimizer $\bar{z}$, satisfying

$$0 = \nabla J(\bar{\psi}) = G_T\bar{z} + S(T)y_0 - y_1.$$

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 [49]): 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 10 does not hold. In terms of the Gramian operator, the observability inequality (5.9) is written as $\langle G_T z, z\rangle_{D(A^*)',D(A^*)} \geqslant C_T\|S(T)^*z\|^2_{X'}$ 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 X is in general much larger than $D(A^*)$ and may even fail to be a space of distributions (see [49]). Anyway, there is a unique minimizer $\bar{z} \in \mathcal{X}$ of J, satisfying therefore $0 = \nabla J(\bar{\psi}) = G_T\bar{z} + S(T)y_0$, 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.

5.2.4 Further comments

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 11. *We assume that X is reflexive and that $B \in L(U,X)$ is a bounded control operator. We set*

$$U_\infty = \Big\{ u \in U \;\Big|\; Bu \in \bigcap_{n=1}^{+\infty} D(A^n) \Big\}.$$

If the set $\mathcal{K}_T = \mathrm{Span}\{A^n Bu \mid u \in U_\infty,\ n \in \mathbb{N}\}$ *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 $z \in \cap_{n=1}^{+\infty} D((A^*)^n)$ (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 $B^*(A^n)^* z = 0$, hence $\langle z, A^n Bu\rangle_{X',X} = 0$, and therefore $z = 0$ because $\mathcal{K}_T$ is dense in X. □

We refer the reader to [65] for a more precise result (and an almost necessary and sufficient condition).

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 B is admissible (and thus $\alpha(B) \leqslant 1/2$).

1. We assume that the control operator B is bounded, that is, $B \in L(U,X)$, and that B is compact (i.e., the image by B of the unit ball of U is relatively compact).

   It is easy to see that the operator L_T is compact, and hence, since X is infinite dimensional, and by the Riesz compactness lemma, the control system (5.1) cannot be 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. This means that it is impossible to control exactly an infinite-dimensional system with a finite number of controls (see [20, Theorem 4.1.5]).

2. We assume that the control operator B is bounded, that is, $B \in L(U,X)$, and that the semigroup $(S(t))_{t\geqslant 0}$ is compact for every $t > 0$. This is the case for instance for the heat equation with internal control.

   For every $\varepsilon > 0$, we define the operator $L_{T,\varepsilon} : L^2(0,T;U) \to X$ by $L_{T,\varepsilon} u = \int_0^{T-\varepsilon} S(T-t)Bu(t)\,dt$.

   Clearly, $L_{T,\varepsilon}$ converges strongly L_T as ε tends to 0.

   Besides, using the fact that $S(T-t) = S(\varepsilon)S(T-\varepsilon-t)$, 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 as well, and hence, since X is infinite dimensional, the control system (5.1) cannot by exactly controllable in any time $T > 0$ (with controls in $L^2(0,T;U)$).

3. We assume that $X \simeq X'$ and $U \simeq U'$ are Hilbert spaces, and that A is a self-adjoint positive operator with compact inverse. We recall that $X_{1/2} = D(A^{1/2})$ is the completion of $D(A)$ for the norm $\sqrt{\langle Ax, x\rangle}$, and $X_{-1/2} = X'_{1/2}$ with respect to the pivot space X (see also Remark 45). We assume that the control operator is such that $B \in L(U, X_{-1/2})$ (that is, its degree of unboundedness is less than or equal to $1/2$, see Remark 54, and hence in particular B is admissible).

   Let $(\phi_j)_{j\in\mathbb{N}^*}$ an orthonormal basis of (unit) eigenvectors of A associated with eigenvalues $\lambda_j > 0$, with $\lambda_j \to +\infty$ as $j \to +\infty$. Firstly, we have

$$\int_0^T \|B^* S(T-t)^* \phi_j\|_U^2 \, dt = \int_0^T e^{\lambda_j (T-t)} \|B^* \phi_j\|_U^2 \, dt \sim \frac{1}{|\lambda_j|} \|B^* \phi_j\|_U^2 = \|B^* (-A)^{-1/2} \phi_j\|_U^2 \tag{5.11}$$

   ineq21:00

   as $j \to +\infty$.

   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 $(x, \phi_j)_X \to 0$ as $j \to +\infty$, for every $x \in X$. This is exactly the desired weak convergence property.

   Since $B^*(-A)^{-1/2} \in L(X, U)$ is compact and since $\phi_j \rightharpoonup 0$, it follows that $B^*(-A)^{-1/2}\phi_j$ converges strongly to 0. Then, from (5.11), we infer that the observability inequality (5.7) does not hold true.

   We conclude that the control system (5.1) cannot by exactly controllable in any time $T > 0$ (with controls in $L^2(0, T; U)$).

   We refer to [66, Proposition 9.1.1] for such arguments.

5.2.5 Examples

Heat equation with internal control and Dirichlet boundary conditions. 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. Consider the heat equation with internal control and Dirichlet boundary conditions

$$\partial_t y = \triangle y + \chi_\omega u \quad \text{in } \Omega, \qquad y_{|\partial\Omega} = 0, \qquad y(0) = y_0 \in L^2(\Omega).$$

***** Reprise de l'exemple précédent: réf... ***** We set $X = L^2(\Omega)$, and $A = -\triangle : D(A) \to X$, 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 bounded, defined as follows: for every $u \in U = L^2(\omega)$, Bu is the extension of u by 0 to the whole Ω.

***** Note that it follows from Holmgren's theorem that the heat equation is approximately controllable in L^2. Voir Li Yong page 282. *****

Since B is bounded, we already know that it is admissible. Note however that B is an admissible control operator if and only if, for every $T > 0$, there exists $K_T > 0$ such that every solution of

$$\partial_t \psi = \triangle \psi \quad \text{in } \Omega, \qquad \psi_{|\partial\Omega} = 0, \qquad \psi(0) \in H^2(\Omega) \cap H_0^1(\Omega)$$

satisfies

$$\int_0^T \int_\omega \psi(t,x)^2 \, dx \, dt \leqslant K_T \|\psi(0)\|^2_{L^2(\Omega)},$$

which is indeed true. ***** déjà dit !! *****

***** Donner l'inegalité d'obs, et dire qu'elle est toujours vraie par Carleman, cf Coron page 80 ou Tucsnak Weiss. *****

***** heat equation with Dirichlet control: Li Yong page 361. Le pb de regularite est explique page 366, ainsi que dans Lions (vieux bouquin) page 217.**

wave equation: voir Li Yong page 284 ***

***** Pour rappel:**

- Equation chaleur avec controle Dirichlet: bien posée dans H^{-1}, mais pas dans L^2. Ce qui veut dire que, pour une donnée initiale L^2 et un contrôle L^2, la solution vit dans l'espace H^{-1}, mais pas forcément dans l'espace L^2.

- Pourtant, elle est exactement controlable a zero dans L^2

****

One-dimensional 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:

$$\begin{aligned} &\partial_{tt} y = \partial_{xx} y, && t \in (0,T),\ x \in (0,L), \\ &y(t,0) = 0,\ y(t,L) = u(t), && t \in (0,T), \\ &y(0,x) = y_0(x),\ \partial_t y(0,x) = y_1(x), && x \in (0,L), \end{aligned} \tag{5.12}$$

ondes

where the state at time $t \in [0,T]$ is $(y(t,\cdot), y_t(t,\cdot))$ 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 $T \geqslant 2L$.

By Theorem 30, this is equivalent to establishing the following observability inequality: there exists $C_T > 0$ such that, for all $(\psi_0, \psi_1) \in H_0^1(0,L) \times L^2(0,L)$, the solution of

$$\begin{aligned} &\psi_{tt} = \psi_{xx}, \\ &\psi(t,0) = \psi(t,L) = 0, \\ &\psi(0,\cdot) = \psi_0(\cdot),\ \partial_t \psi(0,\cdot) = \psi_1(\cdot) \end{aligned} \tag{5.13}$$

ondesadjoint

satisfies

$$\int_0^T |\partial_x \psi(t,L)|^2\,dt \geqslant C_T \int_0^L \left(|\psi_0'(x)|^2 + |\psi_1(x)|^2\right) dx. \qquad (5.14)$$ inegobsici

Given any $T \geqslant 2L$, let us establish (5.14) by using spectral expansions (Fourier series). Setting

$$\psi_0(x) = \sum_{k=1}^{\infty} a_k \sin\frac{k\pi x}{L} \qquad \text{and} \qquad \psi_1(x) = \sum_{k=1}^{\infty} b_k \sin\frac{k\pi x}{L},$$

the solution of (5.13) is

$$\psi(t,x) = \sum_{k=1}^{\infty} \left(a_k \cos\frac{k\pi t}{L} + \frac{b_k L}{k\pi}\sin\frac{k\pi t}{L}\right)\sin\frac{k\pi x}{L}.$$

Then

$$\begin{aligned}
\int_0^T |\partial_x \psi(t,L)|^2\,dt &\geqslant \int_0^{2L} |\partial_x \psi(t,L)|^2\,dt \\
&= \int_0^{2L} \left|\sum_{k=1}^{\infty} (-1)^k \left(\frac{k\pi}{L} a_k \cos\frac{k\pi t}{L} + b_k \sin\frac{k\pi t}{L}\right)\right|^2 dt \\
&= \sum_{j,k=1}^{\infty} (-1)^{j+k} \int_0^{2L} (a_j\cos(j\pi t) + b_j \sin(j\pi t))(a_k\cos(k\pi t) + b_k\sin(k\pi t))\,dt \\
&\qquad\qquad \times \int_\omega \sin(j\pi x)\sin(k\pi x)\,dx \\
&= \sum_{j=1}^{\infty} (a_j^2 + b_j^2) \int_\omega \sin^2(j\pi x)\,dx
\end{aligned}$$

***** A REPRENDRE..... *****

1. En déduire l'inégalité (5.14) pour $T = 2L$, puis pour $T \geqslant 2L$.
2. En déduire que, pour tout $T \geqslant 2L$, le système (5.12) est exactement contrôlable en temps T dans $L^2(0,L) \times H^{-1}(0,L)$, avec des contrôles $u \in L^2(0,T)$.
3. Démontrer que le système (5.12) n'est pas contrôlable en temps $T < 2L$.

   Indication : pour $T \leqslant 2L - 2\delta$, où $\delta > 0$, considérer la solution de $\psi_{tt} = \psi_{xx},$, $\psi(t,0) = \psi(t,L) = 0$, ayant en $t = T/2$ des données à support contenu dans l'intervalle $]0,\delta[$, et montrer que l'inégalité d'observabilité (5.7) n'a pas lieu, en utilisant le fait que la vitesse de propagation de l'équation des ondes est égale à 1.
4. Décrire rapidement la méthode HUM pour $T = 2\pi$.

One-dimensional semilinear heat equation. Let $L > 0$ be fixed and let $f : \mathbb{R} \to \mathbb{R}$ be a function of class C^2 such that $f(0) = 0$. We consider the 1D semilinear heat equation

$$y_t = y_{xx} + f(y), \qquad y(t,0) = 0,\ y(t,L) = u(t), \tag{5.15}$$ eqcont0

where the state is $y(t,\cdot) : [0,L] \to \mathbb{R}$ and the control is $u(t) \in \mathbb{R}$.

We want to design a feedback control locally stabilizing (5.15) 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 $T > 0$, (5.15) 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 $L^\infty((0,T)\times(0,L))$.[2]

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

$$z_t = z_{xx} + f'(0)z + \frac{x}{L}f'(0)u - \frac{x}{L}v + r(t,x), \qquad z(t,0) = z(t,L) = 0, \tag{5.16}$$ reducedproblem2

with $z(0,x) = y(0,x)$ and

$$r(t,x) = \left(z(t,x) + \frac{x}{L}u(t)\right)^2 \int_0^1 (1-s) f''\left(sz(s,x) + s\frac{x}{L}u(s)\right) ds.$$

Note that, given $B > 0$ arbitrary, there exist positive constants C_1 and C_2 such that, if $|u(t)| \leqslant B$ and $\|z(t,\cdot)\|_{L^\infty(0,L)} \leqslant B$, then

$$\|r(t,\cdot)\|_{L^\infty(0,L)} \leqslant C_1(u(t)^2 + \|z(t,\cdot)\|^2_{L^\infty(0,L)}) \leqslant C_2(u(t)^2 + \|z(t,\cdot)\|^2_{H^1_0(0,L)}).$$

In the sequel, $r(t,x)$ will be considered as a remainder.

We define the operator $A = \triangle + f'(0)\mathrm{id}$ on $D(A) = H^2(0,L) \cap H^1_0(0,L)$, so that (5.16) is written as

$$\dot u = v, \qquad z_t = Az + au + bv + r, \qquad z(t,0) = z(t,L) = 0, \tag{5.17}$$ reducedproblem3

[2] 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

$$\sum_{j,k} |c_{jk}| \leqslant \left(\sum_{j,k} \frac{1}{1+j^2+k^4}\right)^{1/2} \left(\sum_{j,k} (1+j^2+k^4)|c_{jk}|^2\right)^{1/2}$$

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 (5.15) 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 $a(x) = \frac{x}{L} f'(0)$ and $b(x) = -\frac{x}{L}$.

Since A is self-adjoint and has a compact resolvent, there exists a Hilbert basis $(e_j)_{j \geqslant 1}$ 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 $(\lambda_j)_{j \geqslant 1}$ such that $-\infty < \cdots < \lambda_n < \cdots < \lambda_1$ and $\lambda_n \to -\infty$ as $n \to +\infty$.

Any solution $z(t, \cdot) \in H^2(0, L) \cap H_0^1(0, L)$ of (5.16), 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 $j \geqslant 1$,

$$\dot{z}_j(t) = \lambda_j z_j(t) + a_j u(t) + b_j v(t) + r_j(t), \tag{5.18}$$?sysdiag?

with

$$a_j = \frac{f'(0)}{L} \int_0^L x e_j(x)\, dx, \quad b_j = -\frac{1}{L} \int_0^L x e_j(x)\, dx, \quad r_j(t) = \int_0^L r(t, x) e_j(x)\, dx.$$

Setting, for every $n \in \mathbb{N}^*$,

$$X_n(t) = \begin{pmatrix} u(t) \\ z_1(t) \\ \vdots \\ z_n(t) \end{pmatrix}, \; A_n = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_1 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_n & 0 & \cdots & \lambda_n \end{pmatrix}, \; B_n = \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}, \; R_n(t) = \begin{pmatrix} 0 \\ r_1(t) \\ \vdots \\ r_n(t) \end{pmatrix},$$

we have, then,

$$\dot{X}_n(t) = A_n X_n(t) + B_n v(t) + R_n(t).$$

Lemma 12. *The pair (A_n, B_n) satisfies the Kalman condition.*

Proof. We compute

$$\det\left(B_n, A_n B_n, \ldots, A_n^n B_n\right) = \prod_{j=1}^{n} (a_j + \lambda_j b_j)\ \mathrm{VdM}(\lambda_1, \ldots, \lambda_n) \tag{5.19}$$ deter

where $\mathrm{VdM}(\lambda_1, \ldots, \lambda_n)$ is a Van der Monde determinant, and thus is never equal to zero since the λ_i, $i = 1 \ldots n$, 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 (5.19) is never equal to zero. □

By the pole-shifting theorem (Theorem 18), there exists $K_n = (k_0, \ldots, k_n)$ 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 8), there exists a symmetric positive definite matrix P_n of size $n+1$ such that

$$P_n\left(A_n + B_nK_n\right) + \left(A_n + B_nK_n\right)^\top P_n = -I_{n+1}.$$

Therefore, as shown in Example 8, 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 $\dot{X}_n(t) = (A_n + B_nK_n)X_n(t)$.

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

$$V(u,z) = \gamma\, X_n^\top P_n X_n - \frac{1}{2}\langle z, Az\rangle_{L^2(0,L)} = \gamma\, X_n^\top P_n X_n - \frac{1}{2}\sum_{j=1}^{\infty} \lambda_j z_j^2 \qquad (5.20)$$ defLyapV1216

where $X_n \in \mathbb{R}^{n+1}$ is defined by $X_n = (u, z_1, \dots, z_n)^\top$ and $z_j = \langle z(\cdot), e_i(\cdot)\rangle_{L^2(0,L)}$ for every j.

Using that $\lambda_n \to -\infty$ as $n \to +\infty$, 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

$$\begin{aligned} C_3\left(u^2 + \|z\|^2_{H_0^1(0,L)}\right) \;\leqslant\; V(u,z) \;\leqslant\; C_4\left(u^2 + \|z\|^2_{H_0^1(0,L)}\right), \\ V(u,z) \;\leqslant\; C_5\left(\|X_n\|_2^2 + \|Az\|^2_{L^2(0,L)}\right), \qquad \gamma C_6 \|X_n\|_2^2 \;\leqslant\; V(u,z), \end{aligned}$$

for all $(u,z) \in \mathbb{R} \times (H^2(0,L) \cap H_0^1(0,L))$. Here, $\|\ \|_2$ designates the Euclidean norm of $\mathbb{R}^{n+1}$.

Our objective is now to prove that V is a Lyapunov function for the system (5.17) in closed-loop with the control $v = K_nX_n$.

In what follows, we thus take $v = K_nX_n$ and u defined by $\dot{u} = v$ and $u(0) = 0$. We compute

$$\begin{aligned} \frac{d}{dt}V(u(t),z(t)) = -\gamma\,\|X_n(t)\|_2^2 - \|Az(t,\cdot)\|^2_{L^2} - \langle Az(t,\cdot), a(\cdot)\rangle_{L^2} u(t) \\ - \langle Az(t,\cdot), b(\cdot)\rangle_{L^2} K_nX_n(t) - \langle Az(t,\cdot), r(t,\cdot)\rangle_{L^2} \\ + \gamma\left(R_n(t)^\top P_nX_n(t) + X_n(t)^\top P_nR_n(t)\right). \end{aligned} \qquad (5.21)$$ dVuzdt

Let us estimate the terms at the right-hand side of (5.21). Under the a priori estimates $|u(t)| \leqslant B$ and $\|z(t,\cdot)\|_{L^\infty(0,L)} \leqslant B$, there exist positive constants C_7, C_8 and C_9 such that

$$\begin{aligned} &|\langle Az, a\rangle_{L^2} u| + |\langle Az, b\rangle_{L^2} K_nX_n| \leqslant \frac{1}{4}\|Az\|^2_{L^2} + C_7\|X_n\|_2^2, \\ &|\langle Az, r\rangle_{L^2}| \leqslant \frac{1}{4}\|Az\|^2_{L^2} + C_8V^2, \qquad \|R_n\|_\infty \leqslant \frac{C_2}{C_3}V, \\ &|\gamma\left(R_n^\top P_nX_n + X_n^\top P_nR_n\right)| \leqslant \frac{C_2}{C_3\sqrt{C_6}}\sqrt{\gamma}\,V^{3/2}. \end{aligned}$$

We infer that, if $\gamma > 0$ is large enough, then there exist positive constants C_{10} and C_{11} such that $\frac{d}{dt}V \leqslant -C_{10}V + C_{11}V^{3/2}$. We easily conclude the local asymptotic stability of the system (5.17) in closed-loop with the control $v = K_n X_n$.

Remark 62. Of course, the above local asymptotic stability may be achieved with other procedures, for instance, by using the Riccati theory (see [71] for Riccati operators in the parabolic case). However, the procedure developed here is much more efficient because it consists of stabilizing a finite-dimensional part of the system, mainly, the part that is not naturally stable. We refer to [18] for examples and for more details. Actually, we have proved in that reference that, thanks to such a strategy, we 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 is the following fact, already used in the remarkable early paper [59]. 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 (5.20).

This idea has been used as well to treat other parabolic problems. But it is interesting to note that it does not work only for parabolic systems: this idea has been as well used in [19] for the 1D semilinear equation

$$y_{tt} = y_{xx} + f(y), \qquad y(t,0) = 0,\ y_x(t,L) = u(t),$$

with the same assumptions on f as before. We first note that, if $f(y) = cy$ is linear (with $c \in L^\infty(0,L)$), then, setting $u(t) = -\alpha y_t(t,L)$ with $\alpha > 0$ yields an exponentially decrease of the energy $\int_0^L (y_t(t,x)^2 + y_x(t,x)^2)\,dt$, and moreover, the eigenvalues of the corresponding operator have a real part tending to $-\infty$ as α tends to 1. Therefore, in the general case, if α 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 same kind of method as the one developed above can therefore be applied, and yields a feedback based on a finite number of modes, that stabilizes locally the semilinear wave equation, asymptotically to equilibrium.

Equivalence between observability and exponential stability. The following result is a generalization of the main result of [28].

Theorem 31. *Let X be a Hilbert space, let $A : D(A) \to X$ be a densely defined skew-adjoint operator, let B be a bounded self-adjoint nonnegative operator on X. We have equivalence of:*

1. *There exist $T > 0$ and $C > 0$ such that every solution of the conservative equation $\dot\phi(t) + A\phi(t) = 0$ satisfies the observability inequality*

$$\|\phi(0)\|_X^2 \leqslant C \int_0^T \|B^{1/2}\phi(t)\|_X^2 \, dt.$$

2. *There exist $C_1 > 0$ and $\delta > 0$ such that every solution of the damped equation $\dot y(t) + Ay(t) + By(t) = 0$ satisfies*

$$E_y(t) \leqslant C_1 E_y(0) \mathrm{e}^{-\delta t},$$

where $E_y(t) = \frac{1}{2}\|y(t)\|_X^2$.

Proof. Let us first prove that the first property implies the second one: we want to prove that every solution of $\dot y + Ay + By = 0$ satisfies

$$E_y(t) = \frac{1}{2}\|y(t)\|_X^2 \leqslant E_y(0)\mathrm{e}^{-\delta t} = \frac{1}{2}\|y(0)\|_X^2 \mathrm{e}^{-\delta t}.$$

Consider ϕ solution of $\dot\phi + A\phi = 0$, $\phi(0) = y(0)$. Setting $\theta = y - \phi$, we have $\dot\theta + A\theta + By = 0$, $\theta(0) = 0$. Then, taking the scalar product with θ, since A is skew-adjoint, we get $\langle \dot\theta + By, \theta\rangle_X = 0$. But, setting $E_\theta(t) = \frac{1}{2}\|\theta(t)\|_X^2$, we have $\dot E_\theta = -\langle By, \theta\rangle_X$. 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. Hence, thanks to the Young inequality $ab \leqslant \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2$ with $\alpha = 2$, we infer that

$$\begin{aligned}\frac{1}{2}\int_0^T \|\theta(t)\|_X^2 \, dt &\leqslant T\|B^{1/2}\| \int_0^T \|B^{1/2}y(t)\|_X \|\theta(t)\|_X \, dt \\ &\leqslant T^2\|B^{1/2}\|^2 \int_0^T \|B^{1/2}y(t)\|_X^2 \, dt + \frac{1}{4}\int_0^T \|\theta(t)\|_X^2 \, dt,\end{aligned}$$

and therefore,

$$\int_0^T \|\theta(t)\|_X^2 \, dt \leqslant 4T^2\|B^{1/2}\|_X^2 \int_0^T \|B^{1/2}y(t)\|_X^2 \, dt.$$

Now, since $\phi = y - \theta$, it follows that

$$\int_0^T \|B^{1/2}\phi(t)\|_X^2\,dt \leqslant 2\int_0^T \|B^{1/2}y(t)\|_X^2\,dt + 2\int_0^T \|B^{1/2}\theta(t)\|_X^2\,dt$$
$$\leqslant (2 + 8T^2\|B^{1/2}\|^4)\int_0^T \|B^{1/2}y(t)\|_X^2\,dt.$$

Finally, since

$$E_y(0) = E_\phi(0) = \frac{1}{2}\|\phi(0)\|_X^2 \leqslant \frac{C}{2}\int_0^T \|B^{1/2}\phi(t)\|_X^2\,dt$$

it follows that $E_y(0) \leqslant C(1 + 4T^2\|B^{1/2}\|^4)\int_0^T \|B^{1/2}y(t)\|_X^2\,dt$. Besides, one has $E_y'(t) = -\|B^{1/2}y(t)\|_X^2$, and then $\int_0^T \|B^{1/2}y(t)\|_X^2 dt = E_y(0) - E_y(T)$. Therefore

$$E_y(0) \leqslant C(1 + 4T^2\|B^{1/2}\|^4)(E_y(0) - E_y(T)) = C_1(E_y(0) - E_y(T))$$

and hence

$$E_y(T) \leqslant \frac{C_1 - 1}{C_1}E_y(0) = C_2E_y(0),$$

with $C_2 < 1$.

Actually this can be done on every interval $[kT, (k+1)T]$, and it yields $E_y((k+1)T) \leqslant C_2E_y(kT)$ for every $k \in \mathbb{N}$, and hence $E_y(kT) \leqslant E_y(0)C_2^k$.

For every $t \in [kT, (k+1)T)$, noting that $k = \left[\frac{t}{T}\right] > \frac{t}{T} - 1$, and that $\ln\frac{1}{C_2} > 0$, it follows that

$$C_2^k = \exp(k\ln C_2) = \exp(-k\ln\frac{1}{C_2}) \leqslant \frac{1}{C_2}\exp(\frac{-\ln\frac{1}{C_2}}{T}t)$$

and hence $E_y(t) \leqslant E_y(kT) \leqslant \delta E_y(0)\exp(-\delta t)$ for some $\delta > 0$.

Let us now prove the converse: assume the exponential decrease, and let us prove the observability property.

From the exponential decrease inequality, one has

$$\int_0^T \|B^{1/2}y(t)\|_X^2\,dt = E_y(0) - E_y(T) \geqslant (1 - C_1\mathrm{e}^{-\delta T})E_y(0) = C_2E_y(0), \quad (5.22)$$ orl1841

and for $T > 0$ large enough there holds $C_2 = 1 - C_1\mathrm{e}^{-\delta T} > 0$.

Then we make the same proof as before, starting from $\dot\phi + A\phi = 0$, that we write in the form $\dot\phi + A\phi + B\phi = B\phi$, and considering the solution of $\dot y + Ay + By = 0$, $y(0) = \phi(0)$. Setting $\theta = \phi - y$, we have $\dot\theta + A\theta + B\theta = B\phi$, $\theta(0) = 0$. Taking the scalar product with θ, 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 $\langle B\theta, \theta\rangle_X = \|B^{1/2}\theta\|_X \geqslant 0$, it follows that $\dot E_\theta \leqslant \langle B\phi, \theta\rangle_X$. As before we apply $\int_0^T\int_0^t$ and hence, since $E_\theta(0) = 0$,

$$\int_0^T E_\theta(t)\,dt \leqslant \int_0^T\int_0^t \langle B\phi(s), \theta(s)\rangle_X\,ds\,dt = \int_0^T (T-t)\langle B^{1/2}\phi(t), B^{1/2}\theta(t)\rangle_X\,dt.$$

Thanks to the Young inequality, we get, exactly as before,

$$\frac{1}{2}\int_0^T \|\theta(t)\|_X^2\,dt \leqslant T\|B^{1/2}\|\int_0^T \|B^{1/2}\phi(t)\|_X\|\theta(t)\|_X\,dt$$
$$\leqslant T^2\|B^{1/2}\|_X^2\int_0^T \|B^{1/2}\phi(t)\|_X^2\,dt + \frac{1}{4}\int_0^T \|\theta(t)\|_X^2\,dt,$$

and finally,

$$\int_0^T \|\theta(t)\|_X^2\,dt \leqslant 4T^2\|B^{1/2}\|_X^2\int_0^T \|B^{1/2}\phi(t)\|_X^2\,dt.$$

Now, since $y = \phi - \theta$, it follows that

$$\int_0^T \|B^{1/2}y(t)\|_X^2\,dt \leqslant 2\int_0^T \|B^{1/2}\phi(t)\|_X^2\,dt + 2\int_0^T \|B^{1/2}\theta(t)\|_X^2\,dt$$
$$\leqslant (2+8T^2\|B^{1/2}\|^4)\int_0^T \|B^{1/2}\phi(t)\|_X^2\,dt.$$

Now, using (5.22) and noting that $E_y(0) = E_\phi(0)$, we infer that

$$C_2E_\phi(0) \leqslant (2+8T^2\|B^{1/2}\|^4)\int_0^T \|B^{1/2}\phi(t)\|_X^2 dt.$$

This is the desired observability inequality. □

Remark 63. 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 [28] for second-order equations, but the proof works exactly in the same way for more general first-order systems, as shown here. However, this proof uses in a crucial way the fact that the operator B is bounded. We refer to [2] for a generalization for unbounded operators with degree of unboudedness $\leqslant 1/2$, and only for second-order equations, with a proof using Laplace transforms, and under a condition on the unboundedness of B that is not easy to check (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).

***** Moment method? *****

Bibliography

AgrachevSachkov [1] A. Agrachev, Y. Sachkov, *Control theory from the geometric viewpoint*, Encyclopaedia of Mathematical Sciences, 87, Control Theory and Optimization, II, Springer-Verlag, Berlin, 2004.

AmmariTucsnak [2] K. Ammari, M. Tucsnak, *Stabilization of second order evolution equations by a class of unbounded operators*, ESAIM: Cont. Optim. Calc. Var. **6** (2001), 361–386.

AndersonMoore [3] B.D. Anderson, J.B. Moore, *Optimal filtering*, Prentice hall, Englewood Cliffs, 1979.

?BardosLebeauRauch? [4] C. Bardos, G. Lebeau, J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Cont. Optim. **30** (1992), 1024–1065.

Betts [5] J.T. Betts, *Practical methods for optimal control and estimation using nonlinear programming*, Second edition, Advances in Design and Control, 19, SIAM, Philadelphia, PA, 2010.

nardCaillauTrelat_COCV2007 [6] B. Bonnard, J.-B. Caillau, E. Trélat, *Second order optimality conditions in the smooth case and applications in optimal control*, ESAIM Control Optim. Calc. Var. **13** (2007), no. 2, 207–236.

BonnardChyba [7] B. Bonnard, M. Chyba, *Singular trajectories and their role in control theory*, Math. & Appl. (Berlin), 40. Springer-Verlag, Berlin, 2003.

BonnardFaubourgTrelat [8] B. Bonnard, L. Faubourg, E. Trélat, *Mécanique céleste et contrôle de systèmes spatiaux*, Math. & Appl. **51**, Springer Verlag (2006), 276 pages.

BourdinTrelat [9] L. Bourdin, E. Trélat, *Pontryagin Maximum Principle for finite dimensional nonlinear optimal control problems on time scales*, SIAM J. Control Optim. **51** (2013), no. 5, 3781–3813.

BressanPiccoli [10] A. Bressan, B. Piccoli, *Introduction to the mathematical theory of control*, AIMS Series on Applied Mathematics, 2, Springfield, MO, 2007.

Brezis [11] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Universitext, Springer, New York, 2011.

BrysonHo [12] A. Bryson, Y.C. Ho, *Applied optimal control*, Hemisphere Pub. Corporation, 1975.

CazenaveHaraux [13] T. Cazenave, A. Haraux, *An introduction to semilinear evolution equations*, Translated from the 1990 French original by Yvan Martel and revised by the authors. Oxford Lecture Series in Mathematics and its Applications, 13. The Clarendon Press, Oxford University Press, New York, 1998.

Cesari [14] L. Cesari, *Optimization – theory and applications. Problems with ordinary differential equations*, Applications of Mathematics, 17, Springer-Verlag, 1983.

Chow [15] W.-L. Chow, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann. **117** (1939), 98–105.

Clarke [16] F.H. Clarke, *Optimization and nonsmooth analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley & Sons, Inc., New York, 1983.

Coron [17] J.-M. Coron, *Control and nonlinearity*, Mathematical Surveys and Monographs, 136. American Mathematical Society, Providence, RI, 2007, xiv+426 pp.

CoronTrelat [18] J.-M. Coron, E. Trélat, *Global steady-state controllability of 1-D semilinear heat equations*, SIAM J. Control Optim. **43** (2004), no. 2, 549–569.

CoronTrelat_CCM2006 [19] J.-M. Coron, E. Trélat, *Global steady-state stabilization and controllability of 1-D semilinear wave equations*, Commun. Contemp. Math. **8** (2006), no. 4, 535–567.

CurtainZwart [20] R.F. Curtain, H. Zwart, *An introduction to infinite-dimensional linear systems theory*, Texts in Applied Mathematics **21**, Springer-Verlag, New York, 1995.

?Andrea? [21] B. D'Andréa-Novel, M. De Lara, *Control theory for engineers*, Springer-Verlag, 2013.

Ekeland [22] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl. **47** (1974), 324–353.

Nagel [23] K.-J. Engel, R. Nagel, *One-parameter semigroups for linear evolution equations*, Graduate Texts Math. **194**, Springer-Verlag, 2000.

Evans [24] Lawrence C. Evans, *Partial differential equations*, Graduate Studies in Mathematics, 19, American Mathematical Society, Providence, RI, 1998.

GaravelloPiccoli [25] M. Garavello, B. Piccoli, *Hybrid necessary principle*, SIAM Journal on Control and Optimization **43** (2005), no. 5, 1867–1887.

Grisvard [26] P. Grisvard, *Elliptic problems in nonsmooth domains*, Monographs and Studies in Mathematics, 24, Pitman, Boston, MA, 1985.

HaberkornTrelat [27] T. Haberkorn, E. Trélat, *Convergence results for smooth regularizations of hybrid nonlinear optimal control problems*, SIAM J. Control Optim. **49** (2011), no. 4, 1498–1522.

Haraux [28] A. Haraux, *Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps*, Portugal. Math. **46** (1989), no. 3, 245–258.

HartlSethi [29] R.F. Hartl, S.P. Sethi, R.G. Vickson, *A survey of the maximum principles for optimal control problems with state constraints* SIAM Rev. **37** (1995), no. 2, 181–218.

HermesLasalle [30] H. Hermes, J.P. LaSalle, *Functional analysis and time optimal control*, Mathematics in Science and Engineering, Vol. 56, Academic Press, New York-London, 1969.

Ho-67 [31] L. Hörmander, *Hypoelliptic second order differential equations*, Acta Math. **119** (1967), 147–171.

?Imanuvilov? [32] O. Yu. Imanuvilov, *Controllability of parabolic equations*, Sb. Math. **186** (1995), no. 6, 879–900.

IoffeThikhomirov [33] A.D. Ioffe, V.M. Tihomirov, *Theory of extremal problems*, Studies in Mathematics and its Applications, 6, North-Holland Publishing Co., 1979.

Jacobson [34] D.H. Jacobson, M.M. Lele, J.L. Speyer, *New necessary conditions of optimality for control problems with state-variable inequality constraints*, J. Math. Anal. Appl. **35** (1971), 255–284.

Hurwitz [35] A. Hurwitz, *Uber die Bedingungen, unter welchen einer Gleichung nur Wurzeln mit negativen Reelen Teilen Besitzt*, Math. Ann. **146** (1895), 273–284.

Jurdjevic [36] V. Jurdjevic, *Geometric control theory*, Cambridge Studies in Advanced Mathematics, 52, Cambridge University Press, 1997.

Kailath [37] T. Kailath, *Linear Systems*, Prentice-Hall, 1980.

Kato [38] T. Kato, *Perturbation theory for linear operators*, Reprint of the 1980 edition, Classics in Mathematics, Springer-Verlag, Berlin, 1995.

Kautsky [39] J. Kautsky, N.K. Nichols, *Robust pole assignment in linear state feedback*, Int. J. Control **41** (1985), 1129–1155.

Khalil [40] H.K. Khalil, *Nonlinear systems*, Macmillan Publishing Company, New York, 1992.

?Komornik? [41] V. Komornik, *Exact controllability and stabilization, the multiplier method*, Wiley, Masson, Paris, 1994.

KwakernaakSivan [42] H. Kwakernaak, R. Sivan, *Linear optimal control systems*, John Wiley, New-York, 1972.

LasieckaTriggiani [43] I. Lasiecka, R. Triggiani, *Control theory for partial differential equations: continuous and approximation theories. I. Abstract parabolic systems*, Encyclopedia of Mathematics and its Applications, 74, Cambridge University Press, Cambridge, 2000.

?LebeauRobbiano? [44] G. Lebeau, L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations **20** (1995), 335–356.

LeeMarkus [45] E.B. Lee, L. Markus, *Foundations of optimal control theory*, John Wiley, New York, 1967.

LiYong [46] X. Li, J. Yong, *Optimal control theory for infinite-dimensional systems* Systems & Control: Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, 1995.

Lions_interp [47] J.-L. Lions, *Espaces d'interpolation et domaines de puissances fractionnaires d'opérateurs*, J. Math. Soc. Japan **14**, no. 2 (1962), 233–241.

Lions_SIREV [48] J.-L. Lions, *Exact controllability, stabilizability and perturbations for distributed systems*, SIAM Rev. **30** (1988), 1–68.

Lions_HUM [49] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués*, Tome 1, Recherches en Mathématiques Appliquées, 8, Masson, 1988.

LionsMagenes [50] J.-L. Lions, E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1, Travaux et Recherches Mathématiques, No. 17, Dunod, Paris, 1968.

Maurer [51] H. Maurer, *On optimal control problems with bounded state variables and control appearing linearly*, SIAM J. Cont. Optim. **15** (1977), 345–362.

Nelson [52] E. Nelson, *Analytic vectors*, Ann. Math. **70** (1959), 572–615.

Pazy [53] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Applied Mathematical Sciences, 44, Springer-Verlag, New York, 1983, viii+279

Pontryagin [54] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze, E.F. Mishchenko, *The mathematical theory of optimal processes*, Inc. New York-London 1962, viii+360 pp.

Rashevski [55] P.K. Rashevski, *About connecting two points of complete nonholonomic space by admissible curve*, Uch. Zapiski Ped. Inst. Libknexta **2** (1938), 83–94.

RebarberWeiss [56] R. Rebarber, G. Weiss, *Necessary conditions for exact controllability with a finite-dimensional input space*, Syst. Cont. Letters **40** (2000), 217–227.

Routh [57] E.J. Routh, *A treatise on the stability of a given state of motion*, Macmillan & Co.,London, 1877.

Rudin [58] W. Rudin, *Functional analysis*, Second edition, International Series in Pure and Applied Mathematics, McGraw-Hill, Inc., New York, 1991, xviii+424 pp.

Russell [59] D.L. Russell, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev. **20** (1978), no. 4, 639–739.

Sontag [60] E.D. Sontag, *Mathematical control theory. Deterministic finite-dimensional systems*, Second edition, Texts in Applied Mathematics, 6, Springer-Verlag, New York, 1998, xvi+531

Staffans [61] O. Staffans, *Well-posed linear systems*, Encyclopedia of Mathematics and its Applications, 103, Cambridge University Press, Cambridge, 2005, xviii+776 pp.

Trelat [62] E. Trélat, *Contrôle optimal. (French) [Optimal control] Théorie & applications. [Theory and applications]*, Mathématiques Concrètes. [Concrete Mathematics] Vuibert, Paris, 2005, vi+246 pp.

Trelat_JOTA [63] E. Trélat, *Optimal control and applications to aerospace: some results and challenges*, J. Optim. Theory Appl. **154** (2012), no. 3, 713–758.

Triebel [64] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

Triggiani_SICON1976 [65] R. Triggiani, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, SIAM J. Control Optimization **14** (1976), no. 2, 313–338.

TucsnakWeiss [66] M. Tucsnak, G. Weiss, *Observation and control for operator semigroups*, Birkhäuser Advanced Texts, Birkhäuser Verlag, Basel, 2009, xii+483 pp.

Vinter [67] R. Vinter, *Optimal control. Systems & Control: Foundations & Applications*, Birkäuser, Boston, 2000.

Weiss_IJM1989 [68] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math. **65** (1989), no. 1, 17–43.

Weiss_SICON1989 [69] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim. **27** (1989), no. 3, 527–545.

Weiss_1991 [70] G. Weiss, *Two conjectures on the admissibility of control operators*, Estimation and control of distributed parameter systems (Vorau, 1990), 367–378, Internat. Ser. Numer. Math., 100, Birkhäuser, Basel, 1991.

Zabczyk [71] J. Zabczyk, *Mathematical control theory: an introduction*, Systems & Control: Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, 1992.