Ordinary Differential Equations
Partial Differential Equations

Franck Boyer

M1 Enseignement Supérieur et Recherche

Paul Sabatier University - Toulouse 3

September 5, 2023

These notes are in constant improvement. Please, do not hesitate to point out errors or inaccuracies to the address
franck.boyer@math.univ-toulouse.fr
I would like to warmly thank Baalu Ketema that have translated the original French version into English

Contents

Chapter I

Ordinary differential equations

I Introduction

The goal of this chapter is the study of (first order) ordinary differential equations of the form

$$x' = f(t, x),$$

where x is an **unknown function** of the real variable t and the map f is given. For each time t, the value $x(t)$ is called the state of the system at time t and the space in which the function x takes its values is called the *state space*. Since physics is the main origin of the study of such equations, we will often speak of the variable t as the **time variable**. And for a solution $t \mapsto x(t)$ of the equation, we will often refer to it as a **trajectory of the system**.

In this course, the state space will always be a (subset of a) finite dimensional vector space. One can ask several questions about the above equation, for instance:

- Given an initial condition, meaning an initial time t_0 and an initial state x_0 of the system at time t_0, is there one or more solutions to the equation satisfying the condition $x(t_0) = x_0$?
- If such a solution exists, can we compute it explicitly ?
- If such a computation is impossible, can we at least give a qualitative description of the solution ?
- If f is a periodic function in time, are there periodic solutions to the equation ?
- If x and y are two solutions to the equation, respectively associated to the initial states x_0 and y_0 that are close to each other, will the solutions x and y stay close to each other over time ?
- Etc ...

We will only partially answer some of these questions in this chapter. Actually, the answer to many of these questions depends in one way or another on the answer to the first question. This is Cauchy's theory; it essentially says that, given the initial condition (t_0, x_0), to ensure the existence and uniqueness of solutions to the following **Cauchy problem**

$$\begin{cases} x'(t) = f(t, x(t)), \\ x(t_0) = x_0, \end{cases} \tag{I.1}$$

it is sufficient to verify some hypotheses on the map f.

It is possible to come across different types of graphic representations of solutions to ordinary differential equations and it's important to understand their differences right away.

- *Scalar equations.*

  Example of the logistic equation $\quad x' = x(1 - x).$

  One can draw several solutions on the same graph as functions of time (Figure I.1).

- *Non-scalar equations.*

  Example of a Lotka-Voltera model $\quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' = \begin{pmatrix} x_1(0.4 - x_2) \\ -x_2(1 - 3x_1) \end{pmatrix}$

  One can draw a solution as two functions of time on the same graph (Figure I.2) or as a parametrized curve $t \mapsto (x_1(t), x_2(t)) \in \mathbb{R}^2$ in the plane (Figure I.3).

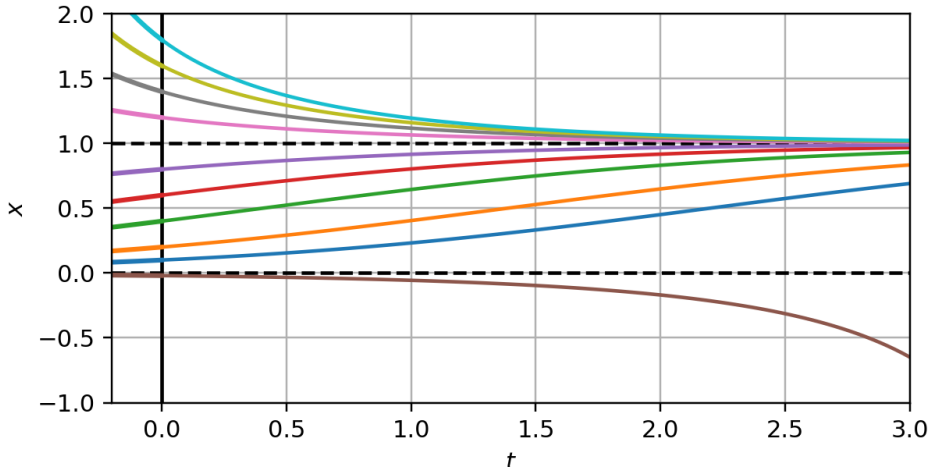


Figure I.1: Several solutions to the logistic equation

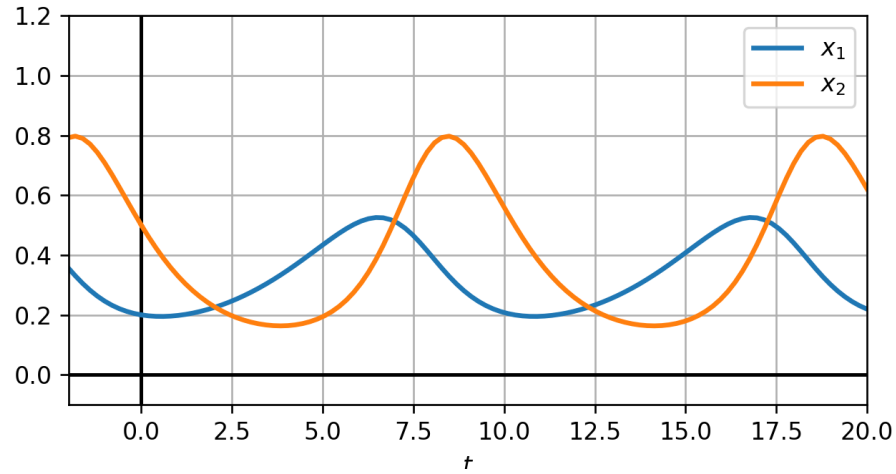


Figure I.2: Only one solution of the Lotka-Voltera system (2 components)

In addition, one also frequently encounters the drawing of a vector field in the form of a set of arrows positioned at certain points on the domain (Figure I.4). There are two versions : the usual version on the left side (each arrow has a length that is proportional to the euclidian norm of the vector $f(t,x)$), the normalised version on the right side (each arrow has a fixed length). This implies that for the second version, at each point on the plane we lost information about the speed of evolution of the solution passing through this point and only preserved the direction.

II Preliminary

II.1 Lipschitz maps

II.1.a Definitions and common examples

All $\mathbb{R}^d$ spaces are endowed with the usual euclidian norm.

Definition I.1

Let $A \subset \mathbb{R}^m$ and $f : A \mapsto \mathbb{R}^n$. The map f is said to be (globally) lipschitz if there exists a constant $L > 0$ such that

$$\forall x, y \in A, \quad \|f(x) - f(y)\| \leq L\|x - y\|.$$

More precisely, we say that f is L-lipschitz. The smallest value for L such that the above property is verified is called the Lipschitz constant of f and is denoted $\mathrm{Lip}(f)$.

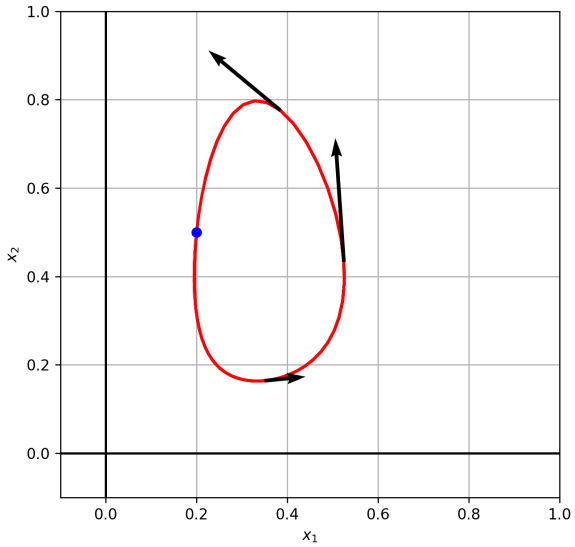


Figure I.3: A trajectory in the state space

Remark I.2

In the above definition, note that the two norms being used are different (one is on $\mathbb{R}^m$ and the other is on $\mathbb{R}^n$). If we take other norms on these spaces, the lipschitzian nature of the map f won't change but the Lipschitz constant might.

Proposition I.3

Lipschitz functions are uniformly continuous (therefore continuous).

A common example of a lipschitz map is given by the following propostion.

Proposition I.4

Let $\Omega \subset \mathbb{R}^m$ be a **convex open set** *of $\mathbb{R}^m$ and $f : \Omega \to \mathbb{R}^n$ a map of class $\mathcal{C}^1$.*
f is lipschitz if and only if it's differential is bounded on Ω. In that case, we have

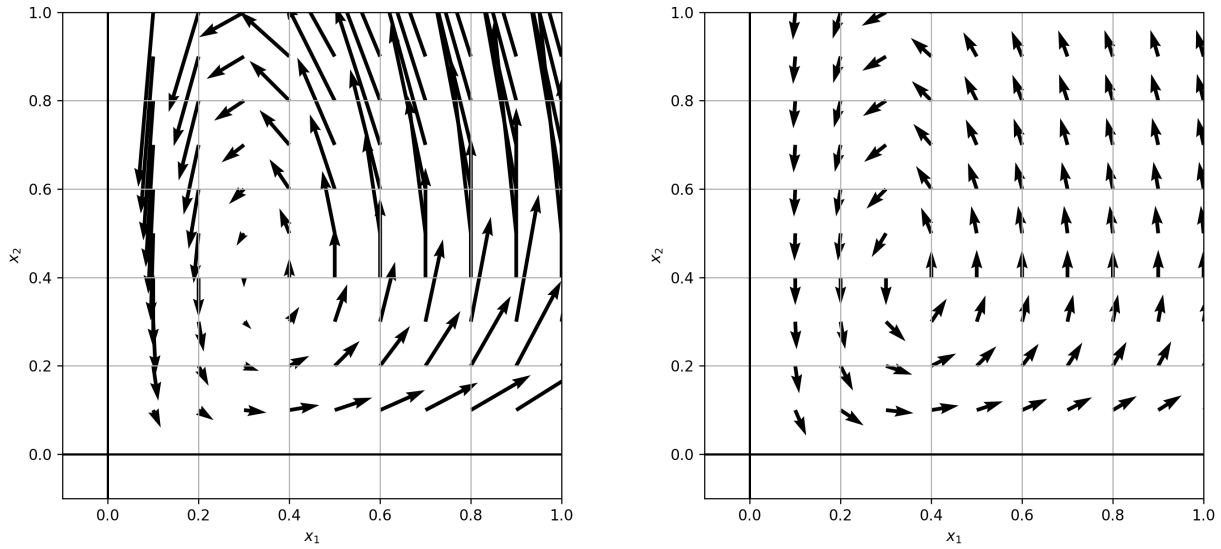
$$\mathrm{Lip}(f) = \sup_{x\in\Omega} \|\mathrm{d}f(x)\|_{L(\mathbb{R}^m,\mathbb{R}^n)}. \tag{I.2}$$

This result can be generalized to a more general class of open sets (not just convex open sets), but it requires precautions and the equality (I.2) is no longer necessarily verified.

Example I.5

Some typical examples of lipschitz maps that are not of class $\mathcal{C}^1$:

- *The absolute value function : $x \in \mathbb{R} \mapsto |x|$, or more generally the norm function $x \in \mathbb{R}^m \mapsto \|x\| \in \mathbb{R}$.*
- *The* distance to a set B *:*
$$x \mapsto d(x, B) = \inf_{y\in B} \|x - y\|.$$
- *The projection on a closed convex set $K \subset \mathbb{R}^n$,*
$$x \in \mathbb{R}^n \mapsto P_K x \in \mathbb{R}^n,$$
where $P_K x$ is the unique point of K that realizes the infimum $\inf_{y\in K} \|x - y\|$.



(a) Usual version

(b) Normalised version

Figure I.4: The drawing of the vector field of a Lotka-Voltera model

The proof of the following proposition is left as an exercise to the reader. The reader should also check that the result can be false if one of the maps is not bounded.

Proposition I.6 (The product of bounded lipschitz maps)

Let $A \subset \mathbb{R}^m$. If $f : A \to \mathbb{R}^n$ and $g : A \to \mathbb{R}$ are two **lipschitz** *and* **bounded** *functions, then the product $fg : A \to \mathbb{R}^n$ is also a lipschitz and bounded function.*

II.1.b To go even further

The following properties will be useful later.

Theorem I.7 (of lipschitzian extension)

Let $A \subset \mathbb{R}^m$. If $f : A \to \mathbb{R}^n$ is L-lipschitz, then there exists a $(\sqrt{n}L)$-lipschitz extension of f to the entire space $\mathbb{R}^m$, meaning that there exists a map $F : \mathbb{R}^m \to \mathbb{R}^n$ that verifies:

- *$F_{|A} = f$.*
- *F is $(\sqrt{n}L)$-lipschitz on $\mathbb{R}^m$.*

Moreover, if f is bounded on A, then we can choose F such that it's bounded on $\mathbb{R}^m$.

Remark I.8

- *As a matter of fact, one can show (but it's significantly more difficult) that it's possible to find a lipschitzian extension of f with a Lipschitz constant which is exactly L (Kirszbraun's theorem), even if f is a map between two infinite dimensional Hilbert spaces.*

  But the above result will be very sufficient for our needs.

- *If we assume in addition that A is a convex and bounded set, then the proof of the Kirszbaun's theorem is quite easy :*
  - *First, define $K = \overline{A}$ which is a compact and convex set.*
  - *Then extend f to K, by using the extension theorem of uniformly continuous maps : for all $x \in K$ take a sequence $(x_n)_n$ of points in A that converges to x and define*

    $$\tilde{f}(x) = \lim_{n\to\infty} f(x_n),$$

    by showing that the limit exists and that it's independent of the chosen sequence.
    The constructed function $\tilde{f}$ is indeed L-lipschitz.
  - *Finally, let P_K be the orthogonal projection on K and define*

    $$F(x) := \tilde{f}(P_K(x)), \ \ \forall x \in \mathbb{R}^d.$$

    Clearly, F is an extension of f. And moreover, since P_K is 1-lipschitz, F is L-lipschitz.

Proof :

- Let's start with the case $n = 1$. In this case, we can give an explicit formula for an F that solves the problem

  $$F(x) := \inf_{y\in A} \Big(f(y) + L\|x - y\| \Big).$$

  Indeed:
  - If $x \in A$, the L-lipschitzian nature of f gives

    $$\forall y \in A, \ \ f(x) \leq f(y) + L\|x - y\|,$$

    which proves that

    $$f(x) \leq F(x).$$

    Moreover, $F(x)$ is smaller than $f(x)$ (it suffices to evaluate the quantity in the infimum at $y = x$), therefore we have shown that $f(x) = F(x)$ for $x \in A$.
  - Now let's prove that F is L-lipschitz on $\mathbb{R}^m$. Let $x_1, x_2 \in \mathbb{R}^m$ and $y \in A$, using the triangle inequality we can write

    $$F(x_1) \leq f(y) + L\|x_1 - y\| \leq f(y) + L\|x_2 - y\| + L\|x_1 - x_2\|.$$

    By taking the infimum on y, we obtain

    $$F(x_1) \leq F(x_2) + L\|x_1 - x_2\|.$$

    Finally, since the roles of x_1 and x_2 can be swapped, we can write

    $$|F(x_1) - F(x_2)| \leq L\|x_1 - x_2\|,$$

    which is what we wanted to prove.

- For the case $n > 1$, we can apply the previous formula to each component of the map (with codomain $\mathbb{R}^n$) $f = (f_1, ..., f_n)$. We obtain a lipschitzian extension $F = (F_1, ..., F_n)$ of f such that each component is L-lipschitzian.

  Now for $x_1, x_2 \in \mathbb{R}^m$, we get

  $$\|F(x_1) - F(x_2)\|^2 = \sum_{i=1}^{n} |F_i(x_1) - F_i(x_2)|^2 \leq \sum_{i=1}^{n} L^2\|x_1 - x_2\|^2 = nL^2\|x_1 - x_2\|^2,$$

  and the result follows.

■

We have previously seen that $\mathcal{C}^1$ maps with a bounded differential are necessarily lipschitz, but we have also seen that (example I.5) some lipschitz maps are not necessarily $\mathcal{C}^1$. The following result (that will not be used here and that is only mentioned for the mathematical culture of the reader...) shows that, ultimately, there isn't much difference between these two notions.

Theorem I.9 (Rademacher)

If Ω is an open set of $\mathbb{R}^m$ and $f : \Omega \to \mathbb{R}^n$ is L-lipschitz, then f is differentiable **almost everywhere** *and, outside of a certain negligible set, we have the bound $\|\mathrm{d}f(x)\| \leq L$.*

In the same spirit, we can prove the following easier but still useful result. This result essentially states that, if x is a regular function, then we can (almost) differentiate $|x|$ like if there was no problem.

Proposition I.10

If $x : \mathbb{R} \to \mathbb{R}$ is a function of class $\mathcal{C}^1$, then the function $|x|$ verifies the following integral identity

$$|x(t)| = |x(s)| + \int_s^t x'(\tau)\,\mathrm{sgn}(x(\tau))\,\mathrm{d}\tau, \quad \forall t, s \in \mathbb{R}.$$

where sgn *is the following function*

$$\mathrm{sgn}(u) = \begin{cases} 1 & \text{if } u > 0, \\ 0 & \text{if } u = 0, \\ -1 & \text{if } u < 0. \end{cases}$$

Let us only give a sketch of the proof, the details are left as an exercise to the reader:

Proof :

For all $\varepsilon > 0$, we define the function $\beta_\varepsilon : \mathbb{R} \to \mathbb{R}$ by

$$\beta_\varepsilon(s) = \frac{s^2}{\sqrt{s^2 + \varepsilon}}, \quad \forall s \in \mathbb{R}.$$

We have the following properties on β_ε:

- β_ε is a function of class $\mathcal{C}^1$ on $\mathbb{R}$ and $\|\beta'_\varepsilon\|_\infty \leq 2$ for $\varepsilon > 0$.
- For all $s \in \mathbb{R}$, $\beta_\varepsilon(s) \to |s|$ when $\varepsilon \to 0$.
- For all $s \in \mathbb{R}$, $\beta'_\varepsilon(s) \to \mathrm{sgn}(s)$ when $\varepsilon \to 0$.

We can write the following formula (chain rule)

$$\beta_\varepsilon(x(t)) = \beta_\varepsilon(x(s)) + \int_s^t \beta'_\varepsilon(x(\tau))x'(\tau)\,d\tau,$$

and justify that we can in fact take the limit as $\varepsilon \to 0$ to get the desired result. ■

II.1.c Locally lipschitz maps

In this paragraph, subsets of $\mathbb{R}^m$ will be endowed with the induced topology.

Definition I.11

Let $A \subset \mathbb{R}^m$ and $f : A \to \mathbb{R}^n$ a map. We say that f is **locally lipschitz** *on A if:*
For every point $x_0 \in A$, there exists an open set U of A that contains x_0 such that f is lipschitz on U.

Proposition I.12

1. *Locally lipschitz maps on a set A are continuous on A.*
2. *If A is open, maps of class $\mathcal{C}^1$ on A are locally lipschitz on A.*

Proof :

1. Immediate consequence of the definitions.
2. We apply the mean value inequality. Let $x_0 \in A$ and $r > 0$ such that $K = \bar{B}(x_0, r) \subset A$ (which is possible since A is open). Since f is of class $\mathcal{C}^1$ and K is compact, there exists $L > 0$ such that $\|df(x)\| \leq L$ for all $x \in K$.

   The open set $U = B(x_0, r)$ is convex and therefore the mean value theorem allows us to establish that for all $x, y \in U$, we get
   $$\|f(x) - f(y)\| \leq L\|x - y\|,$$
   which concludes the proof.

■

In practice, the following characterization is often very useful in proofs.

Proposition I.13

We will use the same notations as the previous definition and assume in addition that A is an open set of $\mathbb{R}^m$. The following assertions are equivalent :

1. *f is locally lipschitz on A.*
2. *For all compact sets K of A, f is (globally) lipschitz on K.*

It should be noted that the compact set K can be as big as desired (provided that it is included in A) but of course, the Lipschitz constant of f on K depends on K and can very well tend to infinity as K gets bigger and bigger.

Proof :

1. $\Rightarrow$ 2. Let K be a compact set of A. We will give a proof by contradiction.

   Let us suppose that f is not lipschitz on K. This means that, for all $n \geq 1$, there exist $x_n, y_n \in K$ such that
   $$\|f(x_n) - f(y_n)\| > n\|x_n - y_n\|, \ \forall n \geq 1. \tag{I.3}$$
   Since K is a compact set of a metric space, and therefore $K \times K$ is a compact set of the product space, we can extract common subsequences $(x_{\varphi(n)})_n$, $(y_{\varphi(n)})_n$ that converge to $x^* \in K$ and $y^* \in K$ respectively. Since f is continuous, we know that
   $$f(x_{\varphi(n)}) \xrightarrow[n\to\infty]{} f(x^*),$$
   $$f(y_{\varphi(n)}) \xrightarrow[n\to\infty]{} f(y^*).$$
   Therefore $f(x_{\varphi(n)}) - f(y_{\varphi(n)})$ is bounded, and by using (I.3), we get that
   $$\|x_{\varphi(n)} - y_{\varphi(n)}\| \xrightarrow[n\to\infty]{} 0.$$
   From this, we can deduce that the two limits x^* and y^* are equal. We denote this limit x_0.

   Now we use the assumption on f that says that there exists $\delta > 0$ and $L > 0$ such that
   $$\|f(x) - f(y)\| \leq L\|x - y\|, \ \forall x, y \in A \cap B(x_0, \delta). \tag{I.4}$$
   Since $(x_{\varphi(n)})_n$ and $(y_{\varphi(n)})_n$ converge to x_0, we know that there exists a sufficiently large integer N such that
   $$x_{\varphi(n)} \in A \cap B(x_0, \delta), \text{ and } y_{\varphi(n)} \in A \cap B(x_0, \delta), \text{ for all } n \geq N.$$
   We can now apply (I.4) to obtain
   $$\|f(x_{\varphi(n)}) - f(y_{\varphi(n)})\| \leq L\|x_{\varphi(n)} - y_{\varphi(n)}\|, \ \forall n \geq N.$$
   By comparing this to (I.3), we obtain a contradiction for n large enough.

2. $\Rightarrow$ 1. Let $x_0 \in A$. Since A is open, there exists $\delta > 0$ such that $\bar{B}(x_0, \delta) \subset A$. Since this ball is closed, therefore compact (remember that we are working on a finite dimensional normed vector space !), we can use our assumption to conclude that f is lipschitz on the open ball $B(x_0, \delta)$.

■

By combining the theorem I.7 and the proposition I.13, we obtain the following corollary.

Corollary I.14

Let A be an open set of $\mathbb{R}^m$ and $f : A \to \mathbb{R}^n$ a locally lipschitz map. For any compact set K of A, there exists a globally lipschitz map $f_K : \mathbb{R}^m \to \mathbb{R}^n$ such that

$$f_K = f, \quad \text{on } K.$$

II.1.d Vector field. Time variable. State variable

As we saw in the introduction, an ordinary differential equation is defined with the help of a map f. We will now discuss on some properties of this map.

Definition I.15

Let $I \subset \mathbb{R}$ be an open interval and $\Omega \subset \mathbb{R}^n$ an open set. A map $f : (t, x) \in I \times \Omega \mapsto \mathbb{R}^n$ is called a **vector field** *on $I \times \Omega$.*
We say that this vector field is **autonomous** *if $I = \mathbb{R}$ and if f does not depend expicitly on time (in this case we identify f with a map from Ω to $\mathbb{R}^n$ ...).*

In this lecture, we will only consider continuous vector fields.

Definition I.16

We say that a continuous vector field $f : I \times \Omega \to \mathbb{R}^n$ is **globally lipschitz with respect to the state variable** *if there exists a continuous function $L : t \in I \to L(t) \in \mathbb{R}^+$ such that*

$$\text{For all } t \in I, \text{ the map } x \in \Omega \mapsto f(t, x) \in \mathbb{R}^n \text{ is } L(t)\text{-lipschtiz on } \Omega.$$

Or in other words,

$$\forall t \in I, \forall x_1, x_2 \in \Omega, \quad \|f(t, x_1) - f(t, x_2)\| \leq L(t)\|x_1 - x_2\|.$$

Definition I.17

We say that a continuous vector field $f : I \times \Omega \to \mathbb{R}^n$ is **locally lipschitz with respect to the state variable** *if:*

- *For all $(t_0, x_0) \in I \times \Omega$, there exist $L > 0$, $\delta > 0$ and an open neighborhood U of x_0 such that*

$$\text{for all } t \in I \cap (t_0 - \delta, t_0 + \delta), \text{ the map } x \in U \mapsto f(t, x) \in \mathbb{R}^n \text{ is } L\text{-lipschtiz on } U.$$

Beware of this last definition, it does not simply mean that $f(t, .)$ is locally lipschitz for all t. It is quite stronger than that because we demand the Lipschitz constant to be valid locally uniformly in time. For instance, the function

$$f(t, x) = \begin{cases} t \sin\left(\frac{x}{t^2}\right), & \text{if } t \neq 0, \\ 0, & \text{if } t = 0, \end{cases}$$

is continuous, lipschitz with respect to x for all t, but is not locally lipschitz with respect to x in the sense of the previous definition.

Proposition I.18

A continuous vector field $f : I \times \Omega \to \mathbb{R}^n$ is locally lipschitz with respect to the state variable if and only if: for all compact set $\mathcal{K} \subset I \times \Omega$, there exists $L > 0$ such that

$$\forall t \in I, x_1, x_2 \in \Omega, \quad \text{such that } (t, x_1) \in \mathcal{K} \text{ and } (t, x_2) \in \mathcal{K} \text{ we have } \|f(t, x_1) - f(t, x_2)\| \leq L\|x_1 - x_2\|.$$

Corollary I.19

Let $f : I \times \Omega \to \mathbb{R}^n$ be a continuous vector field that is locally lipschitz with respect to the state variable, K a compact set of Ω and $[a, b] \subset I$ a compact interval contained in I.
There exists a continuous vector field $\tilde{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ that is globally lipschitz *with respect to the state variable such that*

$$\tilde{f} = f, \ \text{on } [a, b] \times K.$$

Proof :

It is sufficient to show that this restriction/extension result is true for all continuous scalar functions $g : I \times \Omega \to \mathbb{R}$ that are locally lipschitz with respect their state variable, and of the general case we can reason in the same way for each components. According to the previous proposition, for the compact set $\mathcal{K} = [a, b] \times K$, there exists a constant $L > 0$ such that

$$\forall t \in [a, b], \forall x_1, x_2 \in K, \ |g(t, x_1) - g(t, x_2)| \leq L\|x_1 - x_2\|.$$

We will start by constructing the sought out map $\tilde{g}$ on the set $[a, b] \times \mathbb{R}^n$, by using, for all fixed $t \in [a, b]$ which will play the role of a parameter, the same formula that we used in the proof of the Theorem I.7:

$$\tilde{g}(t, x) = \inf_{y \in K} \Big(g(t, y) + L\|x - y\| \Big).$$

We have already seen that this function is indeed globally L-lipschitz with respect to the variable x. Now it remains to show that the function $\tilde{g}$ is continuous. To do this, we will use the fact that g is uniformly continuous on the compact set $[a, b] \times K$ (Heine's theorem). So, for a fixed $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\forall t_1, t_2 \in [a, b], \forall x_1, x_2 \in K, \ |t_1 - t_2| + \|x_1 - x_2\| \leq \delta \Longrightarrow |g(t_1, x_1) - g(t_2, x_2)| \leq \varepsilon. \tag{I.5}$$

Now let's take $t_1, t_2 \in [a, b], \forall x_1, x_2 \in \mathbb{R}^n$ such that $|t_1 - t_2| + \|x_1 - x_2\| \leq \delta$ and any $y \in K$. By using (I.5) and the triangle inequality, we have

$$g(t_1, y) + L\|x_1 - y\| \leq g(t_2, y) + \varepsilon + L\|x_1 - x_2\| + L\|x_2 - y\| \leq g(t_2, y) + L\|x_2 - y\| + \varepsilon + L\delta,$$

and by taking the infimum on y, we get

$$\tilde{g}(t_1, x_1) \leq \tilde{g}(t_2, x_2) + \varepsilon + L\delta.$$

Finally, since we can exchange t_1, x_1 with t_2, x_2 in this formula, we obtain that

$$|\tilde{g}(t_1, x_1) - \tilde{g}(t_2, x_2)| \leq \varepsilon + L\delta.$$

This shows that the function $\tilde{g}$ thus constructed is uniformly continuous on $[a, b] \times \mathbb{R}^n$ therefore continuous. Now it remains to extend $\tilde{g}$ to $\mathbb{R} \times \mathbb{R}^n$, which we can easily do by defining

$$\tilde{g}(t, x) = \tilde{g}(a, x), \ \forall t \leq a,$$

$$\tilde{g}(t, x) = \tilde{g}(b, x), \ \forall t \geq b.$$

One can easily verifiy that the extension preserves the expected properties of the function $\tilde{g}$. ■

II.2 Linear scalar ordinary differential equations and inequalities. Gronwall's lemmas.

Before getting into the general study of ordinary differential equations, we can start by noticing that the solutions to the (homogeneous) linear scalar equations are easy to find. This is the case where $d = 1$ and the "vector field" (a scalar vector field ...) is of the form

$$f(t, x) = a(t)x, \ \forall t \in I, \forall x \in \mathbb{R}.$$

Proposition I.20

Let $I \subset \mathbb{R}$ be a non-empty interval and $t \in I \mapsto a(t) \in \mathbb{R}$ a continuous function.
If $x : I \to \mathbb{R}$ is a function of class $\mathcal{C}^1$ that verifies the equation

$$x'(t) = a(t)x(t), \ \forall t \in I,$$

then we have that

$$x(t) = x(s) \exp\left(\int_s^t a(\tau)\, d\tau \right), \ \forall t, s \in I.$$

Proof :
First, let's fix any $s \in I$ and define

$$z_s(t) = x(t) \exp\left(- \int_s^t a(\tau)\, d\tau \right), \ \ \forall t \in I.$$

This function is of class $\mathcal{C}^1$ and verifies

$$z_s'(t) = (x'(t) - a(t)x(t)) \exp\left(- \int_s^t a(\tau)\, d\tau \right),$$

which gives, by using the equation that x verifies,

$$z_s'(t) = 0, \ \ \forall t \in I.$$

In other words, z_s is constant therefore we have that

$$z_s(t) = z_s(s) = x(s), \ \ \forall t \in I.$$

■

Corollary I.21

For all $t_0 \in I$ and $x_0 \in \mathbb{R}$, there exists a unique solution x, defined on the whole interval I, to the following problem

$$\begin{cases} x'(t) = a(t)x(t) \\ x(t_0) = x_0. \end{cases}$$

The solution is given by the following formula

$$x(t) = x_0 \exp\left(\int_{t_0}^t a(\tau)\, d\tau \right), \ \ \forall t \in I.$$

It turns out that the previous computation allows us to obtain an important information on the function x in the case where the ordinary differential equation is replaced by a differential inequality. This is the purpose of the following result.

Proposition I.22

Let $I \subset \mathbb{R}$ be a non-empty interval and $t \in I \mapsto a(t) \in \mathbb{R}$ a continuous function. Let $x : I \to \mathbb{R}$ be a function of class $\mathcal{C}^1$.
If x verifies the following differential inequality

$$x'(t) \leq a(t)x(t), \ \ \forall t \in I,$$

then we have

$$x(t) \leq x(s) \exp\left(\int_s^t a(\tau)\, d\tau \right), \ \ \forall t, s \in I, \text{ s.t. } t \geq s,$$

and

$$x(t) \geq x(s) \exp\left(\int_s^t a(\tau)\, d\tau \right), \ \ \forall t, s \in I, \text{ s.t. } t \leq s.$$

Remark I.23

If x verifies the differential inequality with the opposite inequality sign

$$x'(t) \geq a(t)x(t), \ \ \forall t \in I,$$

then the conlusion remains true but with the opposite inequality signs. To convince oneself, one should exchange the role of x by $-x$.

Proof :
We will take the same steps we took for the proof of the previous proposition. This time, we obtain that

$$z_s'(t) = (x'(t) - a(t)x(t)) \exp(\ldots) \leq 0,$$

In other words, we have shown that z_s is non increasing (and no longer constant). This gives

$$z_s(t) \leq z_s(s) = x(s), \ \ \forall t \in I, t \geq s,$$
$$z_s(t) \geq z_s(s) = x(s), \ \ \forall t \in I, t \leq s.$$

■

One of the **fundamental** tools in the theory of ordinary differential equations is the following result, which states that essentially, the result of the previous proposition persists if the differential inequality is satisfied in an integral form provided, however, that the function a is non-negative. Although it's usual name is *Gronwall's Lemma*, it deserves to be presented as a theorem.

Theorem I.24 (Gronwall's lemma)

Let I be a non-empty interval, $t \in I \mapsto a(t) \in \mathbb{R}$ a **non-negative** *function and $C \in \mathbb{R}$ a constant.*
If $x : I \to \mathbb{R}$ is a **continuous** *function that verifies, for a certain $s \in I$, the following property :*

$$x(t) \leq C + \int_s^t a(\tau)x(\tau)d\tau, \ \ \forall t \in I, t \geq s, \tag{I.6}$$

then we have the inequality

$$x(t) \leq C \exp\left(\int_s^t a(\tau)d\tau\right), \ \ \forall t \in I, t \geq s. \tag{I.7}$$

In some books, the authors add the assumption "x is a non-negative function". It should be noted that this hypothesis is useless (even though in reality, Gronwall's lemma is often used for non-negative functions ...).

Remark I.25

The following is a similar result in the case where $t \leq s$: if x verifies

$$x(t) \leq C + \int_t^s a(\tau)x(\tau)d\tau, \ \ \forall t \in I, t \leq s,$$

then we have

$$x(t) \leq C \exp\left(\int_t^s a(\tau)d\tau\right), \ \ \forall t \in I, t \leq s.$$

It should be noted that the bounds on the integrals have been exchanged.
The proof of this property is an excellent exercise and is left to the reader.

Proof :

If it was possible to "differentiate an inequality"[1], and if x was assumed to be differentiable, then of course we can easily obtain the differential inequality $x' \leq ax$ and conclude by using the previous proposition. The whole idea of the proof is to get around these difficulties by *doing as follows*.

First, let's give a name to the right hand side of the inequality (I.6)

$$y(t) := C + \int_s^t a(\tau)x(\tau)d\tau, \ \forall t \in I.$$

Since x and a are continuous, y is a $\mathcal{C}^1$ function and verifies

$$y'(t) = a(t)x(t).$$

But the hypothesis (I.6) gives us that $x \leq y$ and since a is non-negative[2] we get the following inequality

$$y'(t) \leq a(t)y(t), \ \ \forall t \in I, t \geq s.$$

We can now apply Proposition I.22 to y and conclude that

$$y(t) \leq y(s) \exp\left(\int_s^t a(\tau)d\tau\right), \ \forall t \in I, t \geq s.$$

[1] One should be convinced that this is indeed illegal !!
[2] Note carefully where this crucial assumption comes in

Using the definition of y, we have $y(s) = C$ but since $x \leq y$ by hypothesis, we have proven that

$$x(t) \leq C \exp\left(\int_s^t a(\tau)d\tau\right), \quad \forall t \in I, t \geq s.$$

■

This fundamental lemma has many variants and generalizations. The following result is one example, whose elementary proof is left as an exercise to the reader.

Proposition I.26 (Generalized version of Gronwall's lemma)

Let I be a non-empty interval, $t \in I \mapsto a(t) \in \mathbb{R}$ a **non-negative** *function and $b : I \to \mathbb{R}$* **non decreasing**. *If $x : I \to \mathbb{R}$ is a* **continuous** *function that verifies the following property for a certain $s \in I$:*

$$x(t) \leq b(t) + \int_s^t a(\tau)x(\tau)d\tau, \quad \forall t \in I, t \geq s,$$

then we have the inequality

$$x(t) \leq b(t) \exp\left(\int_s^t a(\tau)d\tau\right), \quad \forall t \in I, t \geq s.$$

II.3 Notion of solution to a Cauchy problem

II.3.a Definitions

Let $f : I \times \Omega \to \mathbb{R}^d$ be a continuous vector field, $t_0 \in I$ be an initial time and $x_0 \in \Omega$ an initial state.

The following is a fundamental definition since it formalizes the idea that the interval on which a solution is defined cannot be known in advance and its determination is an integral part of the problem.

Definition I.27

A solution to the Cauchy problem (I.1) is **a couple** *(J, x) where $J \subset I$ is an interval containing t_0, and $x : J \to \Omega$ is a map of class $\mathcal{C}^1$ that verifies $x(t_0) = x_0$ as well as the differential equation*

$$x'(t) = f(t, x(t)), \quad \forall t \in J.$$

We also need the following definitions since, once again, the interval on which the solutions are defined is *a priori* unknown.

Definition I.28

A solution (J, x) of the Cauchy problem (I.1) is said to be:

- **Maximal :** *if there is no solution $(\tilde{J}, \tilde{x})$ that strictly extends (J, x), in other words it means that we can't have $J \subsetneq \tilde{J}$ and $\tilde{x} = x$ on J. See figure I.5.*
- **Global :** *if $J = I$ (in this case the solution is obviously maximal).*

III Global Cauchy-Lipschitz theorem and applications

In this section, we will assume that the set of all the possible states of the system is the whole space: $\Omega = \mathbb{R}^d$.

The theory exposed here is based on the following simple yet fundamental remark : for $J \subset I$ that contains t_0 and $x : J \to \mathbb{R}^d$ a function, the following properties are equivalent

1. $x \in \mathcal{C}^1(J, \mathbb{R}^d)$, verifies $x(t_0) = x_0$ and $x'(t) = f(t, x(t))$ for all $t \in J$.
2. $x \in \mathcal{C}^0(J, \mathbb{R}^d)$ and verifies

$$x(t) = x_0 + \int_{t_0}^t f(s, x(s))\, ds, \ \forall t \in J.$$

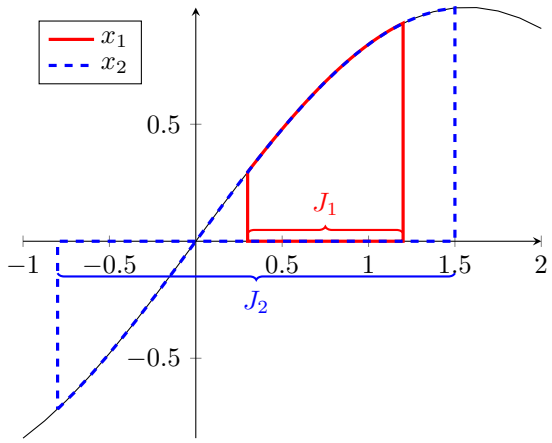


Figure I.5: Notion of extension of a solution. (J_1, x_1) is not a maximal solution.

III.1 Statement and proof of the main theorem

The goal of this subsection is to prove the following theorem.

Theorem I.29 (Cauchy-Lipschitz, global version)

Suppose that the vector field $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ is continuous and **globally lipschitz** *with respect to the state variable.*
Then, for all Cauchy data (t_0, x_0), there **exists** *a* **unique** *solution (I, x), that is* **global***, of the Cauchy problem* (I.1).
Moreover, every other solution $(J, \tilde{x})$ of the Cauchy problem is the restriction of the solution x to J.

Proof :

The proof is structured in several parts.

- *Uniqueness :*

  We will show the following property that will imply the uniqueness of a possible global solution and the last assertion of the theorem : For (J_1, x_1) and (J_2, x_2) two solutions of the Cauchy problem, we will prove that $x_1 = x_2$ on $J = J_1 \cap J_2$.

  The difference $z = x_1 - x_2$ verifies

  $$z'(t) = f(t, x_1(t)) - f(t, x_2(t)), \ \forall t \in J,$$

  as well as $z(t_0) = 0$.

  Now, for all $t \in J$, $t \geq t_0$, we will integrate the above equality between the bounds t_0 and t to obtain

  $$z(t) = \int_{t_0}^{t} \big(f(s, x_1(s)) - f(s, x_2(s))\big)\, ds.$$

  By taking the norm of this identity (here we are using the fact that $t \geq t_0$!)

  $$\|z(t)\| \leq \int_{t_0}^{t} \|f(s, x_1(s)) - f(s, x_2(s))\|\, ds,$$

  and by using the fact that f is globally lipschitz with respect to the state variable (remark that in both terms under the integral sign, f is evaluated at the same time s), we obtain

  $$\|z(t)\| \leq \int_{t_0}^{t} L(s)\|x_1(s) - x_2(s)\|\, ds = \int_{t_0}^{t} L(s)\|z(s)\|\, ds.$$

  But since this is true for all $t \geq t_0$, and since $t \mapsto \|z(t)\|$ is continuous and also since $L \geq 0$ (it's the Lipschitz constant), we can apply Gronwall's lemma (where $C = 0$ here) and deduce that

  $$\|z(t)\| \leq 0 \times \exp\left(\int_{t_0}^{t} L(s)\, dx\right), \ \ \forall t \geq t_0.$$

This shows that $z(t) = 0$ for all $t \geq t_0$, in other words $x_1 \equiv x_2$ on $J \cap [t_0, +\infty)$. A similar reasoning[3] can be done to show that $x_1 \equiv x_2$ on $J \cap (-\infty, t_0]$.

- *An a priori bound:*

  Here, we will assume that the global solution (I, x) exists and we will establish a bound on the solution. To do this, we will once again integrate the differential equation between the bounds t_0 and $t \geq t_0$ to obtain (by using that $x(t_0) = x_0$)

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds, \quad \forall t \geq t_0.$$

  We would like to use Gronwall's lemma to get an estimation on x but for that we should be able to bound the term in f by above and the only assumption that we have on f is the fact that it's globally lipschitz with respect to the state variable. To use that, we will start by rewriting the above identity as follows

$$x(t) - x_0 = \int_{t_0}^{t} f(s, x_0)\, ds + \int_{t_0}^{t} \big(f(s, x(s)) - f(s, x_0)\big)\, ds, \quad \forall t \geq t_0.$$

  The first integral is independent of the solution x and we can use the assumption that f is lipschitz for the second integral. Let's define $z(t) = x(t) - x_0$ and write

$$\|z(t)\| \leq \int_{t_0}^{t} \|f(s, x_0)\|\, ds + \int_{t_0}^{t} L(s)\|z(s)\|\, ds, \quad \forall t \geq t_0.$$

  The generalized version of Gronwall's lemma (Prop. I.26) gives us the following estimation

$$\|z(t)\| \leq \left(\int_{t_0}^{t} \|f(s, x_0)\|\, ds\right) \exp\left(\int_{t_0}^{t} L(s)\, ds\right), \quad \forall t \geq t_0.$$

  By rewriting this in terms of the variable x, we have

$$\|x(t) - x_0\| \leq \left(\int_{t_0}^{t} \|f(s, x_0)\|\, ds\right) \exp\left(\int_{t_0}^{t} L(s)\, ds\right), \quad \forall t \geq t_0.$$

  A similar estimation can be obtained for $t \leq t_0$ (by putting absolute values around the integrals ...) in such a way that, if we define

$$\varphi(t) := 1 + \|x_0\| + \left|\int_{t_0}^{t} \|f(s, x_0)\|\, ds\right|, \forall t \in I,$$

$$\psi(t) := \left|\int_{t_0}^{t} L(s)\, ds\right|, \ \forall t \in I,$$

  which only depend on the given t_0, x_0 and f, then we have obtained the following *a priori* estimate

$$\|x(t)\| \leq \varphi(t) e^{\psi(t)}, \ \forall t \in I. \tag{I.8}$$

  It should be noted that we have chosen to add 1 in the expression of the function φ in such a way that $\varphi(t) \geq 1$. Moreover, observe that φ is non increasing on $I \cap (-\infty, t_0]$ and non decreasing on $I \cap [t_0, +\infty)$.

- *Introducing a good functional space:*

  Thanks to the previous computation, we now know that it is enough to look for x in the set of functions that verifies the estimation (I.8). Actually, we need to weaken this condition a little and introduce the following space

$$E := \left\{ z \in \mathcal{C}^0(I, \mathbb{R}^d), \ \sup_{t \in I} \left(\varphi(t)^{-1} e^{-2\psi(t)} \|z(t)\|\right) < +\infty \right\},$$

  then we endow it with the norm[4]

$$\|z\|_E := \sup_{t \in I} \left(\varphi(t)^{-1} e^{-2\psi(t)} \|z(t)\|\right).$$

  Notice the factor 2 that is added in the exponential term compared to the quantity obtained in the *a priori* bound. This factor will be fundamental for what we will do in the following (actually, any factor greater that 1 would do the job). On the other hand, the division by φ is legal since we made sure that φ was always greater than 1.

[3] The reader is welcome to do it in detail

[4] check, if necessary, that this is indeed a norm !

The proof of the following result is left as an exercise to the reader (the proof is very similar to that of the completeness of the space of continuous and bounded functions equiped with the infinity norm).

Lemma I.30

$(E, \|.\|_E)$ is a Banach space.

- *Introducing a fixed point problem:*

  For all functions $z \in E$, we define a new function $\theta(z) \in \mathcal{C}^0(I, \mathbb{R}^d)$ by the formula

$$\big(\theta(z)\big)(t) := x_0 + \int_{t_0}^t f(s, z(s))\, ds, \quad \forall t \in I.$$

  We will now study some properties of the map θ:

  – We can see that constant functions are in E. Let's calculate (it is a slight abuse of notation, we should have written $\theta(t \mapsto x_0)$):

$$\big(\theta(x_0)\big)(t) = x_0 + \int_{t_0}^t f(s, x_0)\, ds,$$

    which gives us

$$\|\big(\theta(x_0)\big)(t)\| \leq \|x_0\| + \left| \int_{t_0}^t \|f(s, x_0)\|\, ds \right| \leq \varphi(t),$$

    therefore $\theta(x_0) \in E$.

  – Let's show that θ is a map from E into itself: For $z, \tilde{z} \in E$, we can write

$$\big(\theta(z)\big)(t) - \big(\theta(\tilde{z})\big)(t) = \int_{t_0}^t \big(f(s, z(s)) - f(s, \tilde{z}(s))\big)\, ds,$$

    we can bound this from above, by using the fact that f is lipschitz and φ is monotonous, in the following way

$$\begin{aligned}
\|\big(\theta(z)\big)(t) - \big(\theta(\tilde{z})\big)(t)\| &\leq \left| \int_{t_0}^t L(s)\|z(s) - \tilde{z}(s)\|\, ds \right| \\
&= \left| \int_{t_0}^t \varphi(s) L(s) e^{2\psi(s)} \varphi(s)^{-1} e^{-2\psi(s)} \|z(s) - \tilde{z}(s)\|\, ds \right| \\
&\leq \|z - \tilde{z}\|_E \varphi(t) \left| \int_{t_0}^t L(s) e^{2\psi(s)}\, ds \right| \\
&\leq \|z - \tilde{z}\|_E \varphi(t) \left[\frac{e^{2\psi(s)}}{2} \right]_{t_0}^t \\
&\leq \frac{1}{2} \|z - \tilde{z}\|_E \varphi(t) e^{2\psi(t)}.
\end{aligned}$$

    Which gives

$$\sup_{t \in I} \Big(\varphi(t)^{-1} e^{-2\psi(t)} \|\big(\theta(z)\big)(t) - \big(\theta(\tilde{z})\big)(t)\| \Big) \leq \frac{1}{2} \|z - \tilde{z}\|_E. \tag{I.9}$$

    From this we deduce that $\theta(z) - \theta(\tilde{z})$ is an element of E. By taking $\tilde{z} = x_0$, and since we have seen that $\theta(x_0) \in E$, we can deduce that $\theta(z) \in E$ for all $z \in E$, which shows that θ is indeed a map that goes from E into itself.

  – Since we just established that $\theta(E) \subset E$, the inequality (I.9) can be rewritten as

$$\|\theta(z) - \theta(\tilde{z})\|_E \leq \frac{1}{2} \|z - \tilde{z}\|_E, \quad \forall z, \tilde{z} \in E.$$

    This means that θ is a contraction on the space E. One should observe here the importance of the factor 2 in the exponential term of the definition of the norm of the space E. Thanks to this factor we get the contraction rate of $1/2$. To convince oneself, the reader is invited to redo the previous computation by replacing the factor 2 in the definition of the norm of the space E by an arbitrary number $\alpha > 0$ and observe the result.

- *Solving the fixed point problem :*

  We are in the suitable framework to apply the Banach fixed point theorem, which states that the map θ has a unique fixed point in E which we will denote x. This function is continuous and verifies for all $t \in I$

$$x(t) = \big(\theta(x)\big)(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds.$$

  By taking $t = t_0$, we get that $x(t_0) = x_0$. On the other hand, this equality proves that x is the antiderivative of the continuous function $s \mapsto f(s, x(s))$ and therefore x is of class $\mathcal{C}^1$ and verifies

$$x'(t) = f(t, x(t)), \ \ \forall t \in I,$$

  which is indeed the initial ordinary differential equation. The theorem is proved.

■

III.2 Linear ordinary differential equations

A linear ordinary differential equation is one that is associated to a vector field of the form

$$f(t, x) = A(t)x + b(t), \ \ \forall t \in I, \forall x \in \mathbb{R}^d,$$

where $t \in I \mapsto A(t) \in M_d(\mathbb{R})$ is a continuous map that takes its values in the space of square matrices (or endomorphisms if we replace $\mathbb{R}^d$ by a vector space of finite dimension) and $b : t \in I \mapsto b(t) \in \mathbb{R}^d$ a map. If $b \equiv 0$ (resp. $b \not\equiv 0$) we say that the equation is homogeneous (resp. non-homogenous).

It should be noted that, for all $t \in I$, the map $f(t, .)$ is affine and not linear ... which gives rise to a rather small inconsistency of vocabulary that has nonetheless entered into our habits ...

For such a vector field f, we immediately have

$$\forall t \in I, \forall x_1, x_2 \in \mathbb{R}^d, \|f(t, x_1) - f(t, x_2)\| = \|A(t)(x_1 - x_2)\| \leq |||A(t)|||\|x_1 - x_2\|,$$

which shows that f is continuous and globally lipschitz with respect to it's state variable. As a consequence, the global Cauchy-Lipschitz theorem applies and gives

Theorem I.31 (Linear Cauchy-Lipschitz)

Under the above assumptions, for all initial data $(t_0, x_0) \in I \times \mathbb{R}^d$, there exists a unique global solution (I, x) to the Cauchy problem

$$\begin{cases} x'(t) = A(t)x(t) + b(t), \\ x(t_0) = x_0, \end{cases} \tag{I.10}$$

Every other solution to this problem is a restriction of this solution.

Proposition I.32

Let $x_1, ..., x_m : I \to \mathbb{R}^d$ be solutions of the linear homogeneous differential equation $x' = A(t)x$. The following propositions are equivalent:

1. *For all $t \in I$, $(x_1(t), ..., x_m(t))$ is a free family of $\mathbb{R}^d$.*
2. *There exists $t^* \in I$ such that $(x_1(t^*), ..., x_m(t^*))$ is a free family of $\mathbb{R}^d$.*
3. *$(x_1, ..., x_m)$ is a free family of $\mathcal{C}^0(I, \mathbb{R}^d)$.*

We deduce the following immediate yet fundamental corollary.

Corollary I.33 (Structure of solutions of a linear equation)

The set S_0 of all solutions to the homogeneous equation $x' = A(t)x$ is a vector space of dimension d and, for all b, the set S_b of solutions to the more general equation $x' = A(t)x + b(t)$ is an affine space of dimension d directed by S_0.

Proof (of Proposition I.32):

The implications 1. $\Rightarrow$ 2. and 2. $\Rightarrow$ 3. are immediate, and are in fact valid for any family of continuous functions. We will leave them as an exercise to the reader.

The last implication 3. $\Rightarrow$ 1., is only valid because the x_i's are solutions to the differential equation. As a matter of fact, by contradiction, let's assume that there exists $t_0 \in I$ such that $(x_1(t_0), ..., x_m(t_0))$ is a family that is not free. This means that there exist $\alpha_1, ..., \alpha_m \in \mathbb{R}$ that are not zero all at the same time and such that

$$\alpha_1 x_1(t_0) + \cdots + \alpha_m x_m(t_0) = 0. \tag{I.11}$$

Let's define the function $y = \sum_{i=1}^m \alpha_i x_i(t)$. Since each x_i is a solution to the linear homogeneous equation, $x' = A(t)x$, we immediately get that y is also a solution to the equation. But on the other hand, according to (I.11), we have $y(t_0) = 0$. So y is the unique global solution to the equation $x' = A(t)x$ that is associated to the Cauchy data $(t_0, 0)$, but this solution is nothing but the identically zero function.

Therefore, we have shown that $y(t) = 0$ for all $t \in I$, which means that $\sum_{i=1}^m \alpha_i x_i \equiv 0$. But this cannot be true since our assumption was that the family $(x_i)_i$ is free in $\mathcal{C}^0(I, \mathbb{R}^d)$ and the α_i cannot all be zero at the same time. ■

- In the autonomous case, we have a formula at our disposition to compute solutions just like in the scalar case: if $A \in M_d(\mathbb{R})$ and $x_0 \in \mathbb{R}^d$, then the unique solution to the Cauchy problem

$$\begin{cases} x'(t) = Ax(t), \\ x(t_0) = x_0, \end{cases}$$

  is given by the formula

$$x(t) = e^{(t-t_0)A} x_0.$$

- In the non-homogeneous case, the above formula can be generalized by the following formula (called Duhamel's formula, or variation of constants): if $A \in M_d(\mathbb{R})$, $b : I \to \mathbb{R}^d$, and $x_0 \in \mathbb{R}^d$, then the unique solution to the Cauchy problem

$$\begin{cases} x'(t) = Ax(t) + b(t), \\ x(t_0) = x_0, \end{cases}$$

  is given by the formula

$$x(t) = e^{(t-t_0)A} x_0 + \int_{t_0}^t e^{(t-s)A} b(s)\, ds.$$

- In the non-autonomous case (where A is a function of time) where the dimension is $d = 1$, $A(t)$ is a real number that we can write as $a(t)$ to avoid confusion, and we have the following formulae

$$x(t) = \exp\left(\int_{t_0}^t a(\tau) d\tau\right) x_0,$$

  for the solution to the homogeneous Cauchy problem, and

$$x(t) = \exp\left(\int_{t_0}^t a(\tau) d\tau\right) x_0 + \int_{t_0}^t \left[\exp\left(\int_s^t a(\tau) d\tau\right) b(s)\right] ds,$$

  for the solution to the non-homogeneous Cauchy problem.

- At last, in the autonomous case where the dimension d is at least 2, then we can no longer express the solution of the Cauchy problem with the help of matrix exponentials. In particular, the functions $t \mapsto \exp\left(\int_{t_0}^t A(\tau) d\tau\right) x_0$, in most cases, won't solve the equation $x' = A(t)x$.

III.3 Flow of a vector field

Instead of dealing with each solution of the differential equation independently, we propose here to study the action of the differential equation on the different possible initial states to get a more global vision.

III.3.a Definitions

Let's take the hypotheses of the global Cauchy-Lipschitz theorem to be true : f is a continuous vector field and globally lipschitz with respect to the state variable. Theorem I.29 tells us that for all $t_0 \in I$, $x_0 \in \mathbb{R}^d$, there exists a unique global solution to the Cauchy problem (I.1) that we will temporarily denote by $x_{t_0,x_0} \in \mathcal{C}^1(I, \mathbb{R}^d)$ to clearly show the dependency of the solution on the given data.

To deal with these quantities more conveniently, we will introduce the following definition:

Definition I.34 (Flow of a vector field)

Under the previous assumptions, for all t, t_0, x_0 we define the quantity

$$\varphi(t, t_0, x_0) = x_{t_0,x_0}(t).$$

The map φ is called the flow associated to the vector field f.
We also define the following family of maps

$$\Phi(t, t_0) : x_0 \in \mathbb{R}^d \mapsto \varphi(t, t_0, x_0) \in \mathbb{R}^d$$

that we call the flow between the times t_0 and t.

If the initial and final times, t_0 and t respectively, are fixed, then one should see $\Phi(t, t_0)$ as a function that maps each possible initial state to the value of the corresponding solution at the final time t.

III.3.b Main properties

Theorem I.35 (Proprerties of the flow)

With the same previous notations and assumptions, we have the following properties:

- **Group property of the flow:**
  For all $t_0, t_1, t_2 \in I$, we have
  $$\Phi(t_0, t_0) = \mathrm{Id},$$
  $$\Phi(t_2, t_0) = \Phi(t_2, t_1) \circ \Phi(t_1, t_0).$$
  In particular, for all $t_1, t_0 \in I$, the flow $\Phi(t_1, t_0)$ is a bi-lipschitz homeomorphism of $\mathbb{R}^d$ where the inverse map is given by
  $$\Phi(t_1, t_0)^{-1} = \Phi(t_0, t_1).$$
- **Continuity with respect to the data :**
  The map $\varphi : (t, t_0, x_0) \in I \times I \times \mathbb{R}^d \to \mathbb{R}^d$ is locally lipschitz.

Remark I.36 (The autonomous case)

If the vector field f is autonomous (i.e. it does not depend on time), then one can show (exercise) that the map $\Phi(t, t_0)$ only depends on $t - t_0$. In other words, only the time duration counts and not the the actual values of the initial and final times. In this case, we instead use the notation

$$\varphi(t, x_0) := \varphi(t, 0, x_0),$$

$$\Phi(t) := \Phi(t, 0),$$

and the group properties become

$$\Phi(0) = \mathrm{Id},$$
$$\Phi(t + s) = \Phi(t) \circ \Phi(s), \quad \forall t, s \in \mathbb{R},$$
$$(\Phi(t))^{-1} = \Phi(-t), \ \forall t \in \mathbb{R}.$$

Proof :

- First, we remark that, by definition, we have $\Phi(t, t) = \mathrm{Id}$ for all $t \in I$.

Now, let's fix $x_0 \in \mathbb{R}^d$ and t_0, t_1 in I. We define the maps

$$x(t) := \varphi(t, t_0, x_0),\ \forall t \in I,$$

$$y(t) := \varphi(t, t_1, \varphi(t_1, t_0, x_0)),\ \forall t \in I.$$

By definition of φ, these two functions satisfy the same differential equation

$$x'(t) = f(t, x(t)),\ \text{ and }\ y'(t) = f(t, y(t)).$$

On the other hand, we have

$$x(t_1) = \varphi(t_1, t_0, x_0),\ \ y(t_1) = \varphi(t_1, t_1, \varphi(t_1, t_0, x_0)) = \varphi(t_1, t_0, x_0).$$

Hence x and y are solutions of the same Cauchy problem for the vector field f, so the uniqueness property of such a solution implies that x and y are equal. In particular, we have that $x(t_2) = y(t_2)$ which is indeed the desired property.

- Now let's show that $\Phi(t_1, t_0)$ is lipschitz for all $t_0, t_1 \in I$. Fix $t_0 \in I$ and consider $x_0, \tilde{x}_0 \in \mathbb{R}^d$. By writing the Cauchy problem in the integral form

$$\varphi(t, t_0, x_0) = x_0 + \int_{t_0}^{t} f(s, \varphi(s, t_0, x_0))\, ds, \forall t \in I,$$

$$\varphi(t, t_0, \tilde{x}_0) = \tilde{x}_0 + \int_{t_0}^{t} f(s, \varphi(s, t_0, \tilde{x}_0))\, ds, \forall t \in I$$

and by subtracting these two lines and using the triangle inequality (for $t \geq t_0$), we get

$$\|\varphi(t, t_0, x_0) - \varphi(t, t_0, \tilde{x}_0)\| \leq \|x_0 - \tilde{x}_0\| + \int_{t_0}^{t} L(s)\|\varphi(s, t_0, x_0) - \varphi(s, t_0, \tilde{x}_0)\|\, ds,\ \ \forall t \in I, t \geq t_0.$$

Now we use Gronwall's lemma to obtain

$$\|\varphi(t, t_0, x_0) - \varphi(t, t_0, \tilde{x}_0)\| \leq \|x_0 - \tilde{x}_0\| \exp\left(\int_{t_0}^{t} L(s)\, ds\right),\ \ \forall t \in I, t \geq t_0.$$

This shows that, for fixed values of t_0 and $t_1 \geq t_0$, $\Phi(t_1, t_0)$ is lipschitzian. The case $t_1 \leq t_0$ can be proven in a similar way.

- Let's now fix $t, t_0 \in I$ and $x_0 \in \mathbb{R}^d$. Let J be an open interval that contains t and such that $\overline{J} \subset I$. The inequality established in the previous point shows that

$$\|\varphi(t, t_0, x_0) - \varphi(t, t_0, \tilde{x}_0)\| \leq C(J)\|x_0 - \tilde{x}_0\|, \forall t_0, t \in J, \forall x_0 \in \mathbb{R}^n, \tag{I.12}$$

where $C(J) = \exp\left(\int_J L\right)$ only depends on J and the Lipschitz constant L.

The *a priori* estimation that we saw in the proof of the global Cauchy-Lipschitz theorem tells us that, for all $t_1, t_2 \in J$,

$$\|\varphi(t_1, t_2, x_0)\| \leq \left(\|x_0\| + \int_J \|f(s, x_0)\|\, ds\right) \exp\left(\int_J L(s)\, ds\right) := R_1(x_0, J).$$

By combining this with the estimation of the previous point, we get that, for all $\tilde{x}_0 \in B(x_0, 1)$ and for all $t_1, t_2 \in J$,

$$\|\varphi(t_1, t_2, \tilde{x}_0)\| \leq R_1(x_0, J) + \exp\left(\int_J L(s)\, ds\right) := R_2(x_0, J).$$

Now let $M(x_0, J)$ denote a bound of f on the compact set $\overline{J} \times \overline{B}(0, R_2(x_0, J))$. By using the integral form of the equation, we get that, for all $t_1, t_2, s \in J$ and for all $\tilde{x}_0 \in B(x_0, 1)$,

$$\varphi(t_1, s, \tilde{x}_0) = \varphi(t_2, s, \tilde{x}_0) + \int_{t_2}^{t_1} f(\tau, \varphi(\tau, s, \tilde{x}_0))\, d\tau,$$

So

$$\|\varphi(t_1, s, \tilde{x}_0) - \varphi(t_2, s, \tilde{x}_0)\| \leq M(x_0, J)|t_2 - t_1|. \tag{I.13}$$

Now, by using the triangle inequality, the group property of the flow, then the inequality (I.12) and finally (I.13), we can write

$$\begin{aligned}\|\varphi(\tilde{t},\tilde{t}_0,\tilde{x}_0)-\varphi(t,t_0,\tilde{x}_0)\| &\leq \|\varphi(\tilde{t},\tilde{t}_0,\tilde{x}_0)-\varphi(\tilde{t},t_0,\tilde{x}_0)\|+\|\varphi(\tilde{t},t_0,\tilde{x}_0)-\varphi(t,t_0,\tilde{x}_0)\|\\ &= \|\varphi(\tilde{t},\tilde{t}_0,\tilde{x}_0)-\varphi(\tilde{t},\tilde{t}_0,\varphi(\tilde{t}_0,t_0,\tilde{x}_0))\|+\|\varphi(\tilde{t},t_0,\tilde{x}_0)-\varphi(t,t_0,\tilde{x}_0)\|\\ &= \|\varphi(\tilde{t},\tilde{t}_0,\varphi(t_0,t_0,\tilde{x}_0))-\varphi(\tilde{t},\tilde{t}_0,\varphi(\tilde{t}_0,t_0,\tilde{x}_0))\|+\|\varphi(\tilde{t},t_0,\tilde{x}_0)-\varphi(t,t_0,\tilde{x}_0)\|\\ &\leq C(J)\|\varphi(t_0,t_0,\tilde{x}_0)-\varphi(\tilde{t}_0,t_0,\tilde{x}_0)\|+\|\varphi(\tilde{t},t_0,\tilde{x}_0)-\varphi(t,t_0,\tilde{x}_0)\|\\ &\leq C(J)\|\varphi(t_0,t_0,\tilde{x}_0)-\varphi(\tilde{t}_0,t_0,\tilde{x}_0)\|+\|\varphi(\tilde{t},t_0,\tilde{x}_0)-\varphi(t,t_0,\tilde{x}_0)\|\\ &\leq M(x_0,J)(C(J)|t_0-\tilde{t}_0|+|t-\tilde{t}|).\end{aligned}$$

By using (I.12) one last time, we get the following result : for all $\tilde{t},\tilde{t}_0\in J$ and for all $\tilde{x}_0\in B(x_0,1)$, we have the estimation

$$\|\varphi(\tilde{t},\tilde{t}_0,\tilde{x}_0)-\varphi(t,t_0,x)\|\leq C(J)\|\tilde{x}_0-x_0\|+M(x_0,J)(C(J)|t_0-\tilde{t}_0|+|t-\tilde{t}|),$$

which shows that φ is indeed locally lipschitz.

■

III.3.c Dependence with respect to a parameter

In this section, we will suppose that the vector field f depends on a parameter $\alpha\in\mathbb{R}^p$. More precisely, we will suppose that

$$f:(t,x,\alpha)\in I\times\mathbb{R}^d\times\mathbb{R}^p\mapsto f(t,x,\alpha)\in\mathbb{R}^d,$$

is a continuous map that is globally lipschitz with respect to the state variable x and the parameter α.

We consider the following system of differential equations

$$x'=f(t,x,\alpha),$$

and we would like to study the dependence of the solutions on the parameter α.

A useful trick to bring this study back to the previous cases is to consider a new state variable $X=(x,\alpha)\in\mathbb{R}^{d+p}$ and the associated vector field

$$F:(t,X=(x,\alpha))\in I\times(\mathbb{R}^d\times\mathbb{R}^p)\mapsto(f(t,x,\alpha),0)\in\mathbb{R}^d\times\mathbb{R}^p.$$

We can easily verify that F is a continuous vector field that is globally lipschitz on $\mathbb{R}^{d+p}$ and that we can apply the usual Cauchy-Lipschitz theorem and consider the flow $\psi:I\times I\times\mathbb{R}^{d+p}\to\mathbb{R}^{d+p}$.

Let's decompose the flow in two parts

$$\psi(t,s,(x_0,\alpha_0))=(\varphi(t,s,x_0,\alpha_0),\zeta(t,s,x_0,\alpha_0))\in\mathbb{R}^d\times\mathbb{R}^p,$$

and write the equation that ψ verifies

$$\partial_t\psi(t,s,(x_0,\alpha_0))=F(t,\psi(t,s,(x_0,\alpha_0))),$$

which gives us, thanks to the definition of F, the two equations

$$\begin{aligned}\partial_t\varphi(t,s,(x_0,\alpha_0))&=f(t,\varphi(t,s,(x_0,\alpha_0)),\zeta(t,s,(x_0,\alpha_0))),\\ \partial_t\zeta(t,s,(x_0,\alpha_0))&=0,\end{aligned}$$

with the initial conditions

$$\psi(s,s,(x_0,\alpha_0))=x_0,\quad \zeta(s,s,(x_0,\alpha_0))=\alpha_0.$$

The equation on ζ shows that this quantity does not depend on time, so in particular, we have

$$\zeta(t,s,(x_0,\alpha_0))=\alpha_0,\forall t\in I.$$

If we put this in the above equation on φ, we get

$$\partial_t\varphi(t,s,(x_0,\alpha_0))=f(t,\varphi(t,s,(x_0,\alpha_0)),\alpha_0),$$

which shows that $\varphi(.,.,.,\alpha_0)$ is the flow associated to the equation $x'=f(t,x,\alpha)$ for $\alpha=\alpha_0$.

Since Φ is locally lipschitz with respect to all its variables, the same goes for φ. In other words, we have the following result:

Proposition I.37

The flow of the system $x'=f(t,x,\alpha)$ parametrized by α and denoted $\varphi:I\times I\times\mathbb{R}^d\times\mathbb{R}^p\to\mathbb{R}^d$ is a map that is locally lipschitz with respect to all its variables.

In practice, the previous result is often used in the following form:

Corollary I.38

Fix $t_0 \in I$, $x_0 \in \mathbb{R}^d$ and a compact interval $J \subset I$. If we denote $t \mapsto x_\alpha(t)$ the trajectory associated to the system $x' = f(t, x, \alpha)$ for the Cauchy data (t_0, x_0), then the map

$$\alpha \in \mathbb{R}^p \mapsto x_\alpha \in \mathcal{C}^0(J, \mathbb{R}^d),$$

is locally lipschitz, which we can sum up with the sentence "the solution of the Cauchy problem has a regular dependence on the parameter".

III.3.d The linear case

Consider the linear vector field defined by

$$f(t, x) = A(t)x, \ \forall t \in I, \forall x \in \mathbb{R}^d.$$

Definition and Proposition I.39 (Resolvent matrix)

In the linear case, for all $t, s \in I$, the map $\Phi(t, s)$ is linear, and we identify it to a matrix that we denote by $R(t, s)$. We call this the resolvent of the equation.
So we have

$$\varphi(t, s, x) = R(t, s).x, \ \ \forall t, s \in I, \forall x \in \mathbb{R}^d.$$

The properties of the flow can be reinterpreted using the resolvent in the following way:

- *The map $(t, s) \in I \times I \mapsto R(t, s) \in M_d(\mathbb{R})$ is lipschitz.*
- *$R(t, t) = \mathrm{Id}$ for all $t \in I$.*
- *$R(t_1, t_2).R(t_2, t_3) = R(t_1, t_3)$, for all $t_1, t_2, t_3 \in I$.*
- *For all $t, s \in I$, $R(t, s)$ is invertible and we have $R(t, s) = R(s, t)^{-1}$.*

Finally, in the autonomous case (A does not depend on time), we have the explicit expression

$$R(t, s) = e^{(t-s)A}, \ \ \forall t, s \in I.$$

Proposition I.40

- *For all $s \in I$, the map $t \in I \mapsto R(t, s) \in M_d(\mathbb{R})$ is the unique global solution of the following (linear but that takes its values in $M_d(\mathbb{R})$) Cauchy problem*

$$\begin{cases} M'(t) = A(t)M(t), \\ M(s) = \mathrm{Id}. \end{cases}$$

- *For all $t \in I$, the map $s \in I \mapsto R(t, s) \in M_d(\mathbb{R})$ is the unique solution that is global of the following Cauchy problem*

$$\begin{cases} M'(s) = -M(s)A(s), \\ M(t) = \mathrm{Id}. \end{cases}$$

Proposition I.41 (Duhamel's formula)

The unique solution of the following linear and non-homogeneous Cauchy problem

$$\begin{cases} x'(t) = A(t)x(t) + b(t), \\ x(t_0) = x_0, \end{cases}$$

is given by the formula

$$x(t) = R(t, t_0)x_0 + \int_{t_0}^{t} R(t, s)b(s)\, ds.$$

III.3.e Differentiability of the flow

By definition, the flow φ of a vector field is differentiable with respect to the time variable t, we even have the following result

$$t \mapsto \big(\varphi(t, ., .) \in \mathcal{C}^0(I \times \mathbb{R}^d, \mathbb{R}^d)\big),$$

is a $\mathcal{C}^1$ map whose differential at a point $t \in I$ is given by

$$f(t, \varphi(t, ., .)) \in \mathcal{C}^0(I \times \mathbb{R}^d, \mathbb{R}^d).$$

We will now study the differentiability of the flow with respect to x_0, the initial data.

Theorem I.42

Suppose $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ is a vector field that is globally lipschitz with respect to the state variable and $\mathcal{C}^1$ with respect to the same variable.
So, for all $t, t_0 \in I$, the map $\Phi(t, t_0)$ is $\mathcal{C}^1$. We denote $D_{x_0}\big(\Phi(t, t_0)\big)(x_0)$ its Jacobian at each point $x_0 \in \mathbb{R}^d$.
For all t_0 and x_0, the map $t \mapsto D_{x_0}\big(\Phi(t, t_0)\big)(x_0) \in M_d(\mathbb{R})$ is the unique solution of the following linear matrix-valued Cauchy problem

$$\begin{cases} M'(t) = \big(D_x f\big)(t, \varphi(t, t_0, x_0)).M(t), \\ M(t_0) = \mathrm{Id}. \end{cases} \tag{I.14}$$

Remark I.43

Even if the vector field f is autonomous, the Cauchy problem (I.14) satisfied by the Jacobian of the flow is a non-autonomous problem.

Proof :

First of all, note that the equations (I.14) can be obtained by formally differentiating the initial Cauchy problem and by assuming that we can interchange the differentiation operations.

We define M as the unique solution of (I.14) and we will show that M is indeed the Jacobian of the flow with respect to x_0.

Since f is of class $\mathcal{C}^1$, we can define, for all $s \in I, y \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$

$$R(s, y, k) := \frac{1}{\|k\|}\big(f(s, y + k) - f(s, y) - D_x f(s, y).k\big),$$

and, by definition, we have

$$\lim_{k \to 0} \|R(s, y, k)\| = 0, \ \forall s \in I, \ \forall y \in \mathbb{R}^d. \tag{I.15}$$

Moreover, by the mean value theorem and the lipschitz property we get

$$\|R(s, y, k)\| \leq 2L(s), \ \ \forall s \in I, \ \forall y \in \mathbb{R}^d, \forall k \in R^d. \tag{I.16}$$

We have the following integral formulations

$$\begin{cases} \varphi(t,t_0,x_0+h) = x_0+h+\displaystyle\int_{t_0}^t f(s,\varphi(s,t_0,x_0+h))\,ds, \\ \varphi(t,t_0,x_0) = x_0+\displaystyle\int_{t_0}^t f(s,\varphi(s,t_0,x_0))\,ds, \\ M(t) = \mathrm{Id}+\displaystyle\int_{t_0}^t D_xf(s,\varphi(s,t_0,x_0))M(s)\,ds. \end{cases}$$

By combining them, we get

$$\begin{aligned} &\varphi(t,t_0,x_0+h)-\varphi(t,t_0,x_0)-M(t).h \\ &\qquad = \int_{t_0}^t \big(f(s,\varphi(s,t_0,x_0+h))-f(s,\varphi(s,t_0,x_0))-D_xf(s,\varphi(s,t_0,x_0))M(s).h\big)\,ds, \\ &\qquad = \int_{t_0}^t \big(f(s,\varphi(s,t_0,x_0+h))-f(s,\varphi(s,t_0,x_0)+M(s).h)\big)\,ds \\ &\qquad\quad + \int_{t_0}^t \big(f(s,\varphi(s,t_0,x_0)+M(s).h)-f(s,\varphi(s,t_0,x_0))-D_xf(s,\varphi(s,t_0,x_0))M(s).h\big)\,ds. \end{aligned}$$

By using the definition of R and the fact that the vector field is lipschitz, we obtain

$$\begin{aligned} &\|\varphi(t,t_0,x_0+h)-\varphi(t,t_0,x_0)-M(t).h\| \\ &\quad \le \int_{t_0}^t L(s)\|\varphi(s,t_0,x_0+h))-\varphi(s,t_0,x_0)-M(s).h\|\,ds + \|h\|\int_{t_0}^t \|R(s,\varphi(s,t_0,x_0),M(s).h)\|\|M(s)\|\,ds. \end{aligned}$$

Gronwall's lemma gives us

$$\|\varphi(t,t_0,x_0+h)-\varphi(t,t_0,x_0)-M(t).h\| \le \|h\|\left(\int_{t_0}^t \|R(s,\varphi(s,t_0,x_0),M(s).h)\|\|M(s)\|\,ds.\right)\exp\left(\int_0^t L(s)\,ds\right),$$

therefore

$$\frac{1}{\|h\|}\|\varphi(t,t_0,x_0+h)-\varphi(t,t_0,x_0)-M(t).h\| \le C_t\left(\int_{t_0}^t \|R(s,\varphi(s,t_0,x_0),M(s).h)\|\,ds.\right).$$

By using (I.15) and (I.16), we can apply the dominated convergence theorem to pass to the limit in the integral term on the right hand side of the inequality and conclude that

$$\frac{1}{\|h\|}\|\varphi(t,t_0,x_0+h)-\varphi(t,t_0,x_0)-M(t).h\| \xrightarrow[h\to 0]{} 0.$$

■

By using the same format proposed in Section III.3.c, we can transcribe the previous result in terms of $\mathcal{C}^1$ dependence of the flow on a parameter $\alpha \in \mathbb{R}^p$.

IV Local Cauchy-Lipschitz theorem

Let us now treat the most general case where the vector field is locally lipschitz and defined on a set Ω of admissible states.

IV.1 Statement and proof of the main theorem

Theorem I.44 (Local Cauchy-Lipschitz, first version)

We assume that the vector field $f : I \times \Omega \to \mathbb{R}^d$ is continuous and **locally lipschitz** *with respect to the state variable.*
For all Cauchy data $(t_0,x_0) \in I \times \Omega$, there **exists** *a $\delta > 0$ (depending on the data !) such that there exists a* **unique** *solution x of the Cauchy problem (I.1) defined on $J_\delta := (t_0-\delta, t_0+\delta) \cap I$.*

Proof :

The idea of the proof is to apply the global Cauchy-Lipschitz theorem by using an extension/restriction argument.

Let's start by giving ourselves an $\varepsilon > 0$ such that the set $K_\varepsilon = ([t_0 - \varepsilon, t_0 + \varepsilon] \cap I) \times \overline{B}(x_0, \varepsilon)$ is contained in $I \times \Omega$.

In addition, we can choose ε small enough so that K_ε is compact. If $t_0 \in \overset{\circ}{I}$, it is enough to choose ε so that $[t_0 - \varepsilon, t_0 + \varepsilon] \subset I$. If $t_0 \in \partial I$ then we choose ε so that $[t_0 - \varepsilon, t_0 + \varepsilon] \cap I = [t_0, t_0 + \varepsilon]$ or $[t_0 - \varepsilon, t_0 + \varepsilon] \cap I = [t_0 - \varepsilon, t_0]$.

Let's now introduce the set $U_\varepsilon = ((t_0 - \varepsilon, t_0 + \varepsilon) \cap I) \times B(x_0, \varepsilon)$ that is open in $I \times \Omega$.

According to Corollary I.19, there exists a continuous vector field $f_\varepsilon : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ that is globally lipschitz with respect to the state variable and that coincides with f on K_ε. We can now apply the global Cauchy-Lipschitz theorem for f_ε and deduce the existence and uniqueness of a global solution $(\mathbb{R}, x_\varepsilon)$ to the following Cauchy problem

$$\begin{cases} x_\varepsilon'(t) &= f_\varepsilon(t, x_\varepsilon(t)), \\ x_\varepsilon(t_0) &= x_0. \end{cases}$$

Since $x_\varepsilon(t_0) = x_0$ and $t \mapsto (t, x_\varepsilon(t))$ is a continuous function, there exists $0 < \delta < \varepsilon$ such that

$$(t, x_\varepsilon(t)) \in U_\varepsilon, \quad \forall t \in I \cap (t_0 - \delta, t_0 + \delta).$$

Since f_ε and f coincide on U_ε, we can see that $(I \cap (t_0 - \delta, t_0 + \delta), x_\varepsilon)$ is indeed a solution to the initial Cauchy problem. On the other hand, every other solution to the Cauchy problem will coincide with the solution x_ε on the open interval centered at the point t_0. ■

Corollary I.45 (Uniqueness principle)

Let (J_1, x_1) and (J_2, x_2) be two solutions of the equation $x' = f(t, x)$ such that $J_1 \cap J_2 \neq \emptyset$.
The following propositions are equivalent:

1. *There exists $t^* \in J_1 \cap J_2$ such that $x_1(t^*) = x_2(t^*)$.*
2. *The two functions x_1 and x_2 coincide on $J_1 \cap J_2$.*

Proof :

2. $\Rightarrow$ 1. Immediate.

1. $\Rightarrow$ 2. We define the set

$$J := \{t \in J_1 \cap J_2, \quad x_1(t) = x_2(t)\}.$$

- J is non empty since we assumed that $t^* \in J$.
- J is closed in $J_1 \cap J_2$: because x_1 and x_2 are continuous.
- J is open in $J_1 \cap J_2$:
  Indeed, if we take $t_0 \in J$ then (J_1, x_1) and (J_2, x_2) are two solutions of the **same** Cauchy problem defined on the interval $J_1 \cap J_2$:

$$\begin{cases} y'(t) &= f(t, y), \\ y(t_0) &= y_0. \end{cases} \tag{I.17}$$

  According to the local Cauchy-Lipschitz theorem, there exists $\delta > 0$ s.t. (I.17) has a unique solution on the interval $(J_1 \cap J_2) \cap (t_0 - \delta, t_0 + \delta)$. In particular, $x_1 = x_2$ on this interval therefore

$$(J_1 \cap J_2) \cap (t_0 - \delta, t_0 + \delta) \subset J.$$

Since the interval $J_1 \cap J_2$ is connected, we deduce that

$$J = J_1 \cap J_2,$$

therefore x_1 and x_2 coincide on $J_1 \cap J_2$.

■

Theorem I.46 (Local Cauchy-Lipschitz, second version)

Under the hypotheses of Theorem I.44, for all given Cauchy data there exists a unique maximal solution (J, x) to the Cauchy problem (I.1). In addition, J is an open set in I.
Finally, every other solution $(\tilde{J}, \tilde{x})$ is the restriction of x on $\tilde{J}$.

Proof :

Let's define the set $\mathcal{J}$ by

$$\mathcal{J} := \bigcup_{\substack{(J,x) \\ \text{solution of (I.1)}}} J.$$

According to the first version of the local Cauchy-Lipschitz theorem (Theorem I.44), we know that $\mathcal{J}$ is non empty.

Let $t \in \mathcal{J}$. Let's consider two solutions (J_1, x_1) and (J_2, x_2) of the Cauchy problem (I.1) such that $t \in J_1$ and $t \in J_2$. By definition, we know that $[t_0, t] \subset J_1 \cap J_2$ (where $[t, t_0]$ if $t < t_0$ ...) and from the uniqueness property given by the Corollary I.45, we know that x_1 and x_2 coincide on $[t_0, t]$ (and $[t, t_0]$ ...) so in particular $x_1(t) = x_2(t)$.

Consequently, we can now rightly define the quantity

$$y(t) = \text{the value at the point } t \text{ of any solution } (J, x) \text{ s.t. } t \in J.$$

Clearly, $(\mathcal{J}, y)$ is a solution of the Cauchy problem (I.1), and even more so, by definition of $\mathcal{J}$ it's the unique maximal solution of the problem !

Let's now show that $\mathcal{J}$ is open in I. To do this, we take any $t^* \in \mathcal{J}$ and define $y^* := y(t^*)$. We then consider the auxiliary Cauchy problem,

$$\begin{cases} z'(t) &= f(t, z) \\ z(t^*) &= y^*, \end{cases} \tag{I.18}$$

associated to the initial time t^*.

From the first version of the local Cauchy-Lipschitz theorem (Theorem I.44), we get the existence of $\delta > 0$ such that the problem (I.18) has a unique solution on the interval $(t^* - \delta, t^* + \delta) \cap I$.

So we can now define $J = \mathcal{J} \cup ((t^* - \delta, t^* + \delta) \cap I)$ and

$$\tilde{y}(t) = \begin{cases} y(t), & \text{if } t \in \mathcal{J}, \\ z(t), & \text{if } t \in (t^* - \delta, t^* + \delta) \cap I. \end{cases}$$

This definition is consistent with the uniqueness principle and we have constructed a new solution $(J, \tilde{y})$ of the initial Cauchy problem.

By definition of $\mathcal{J}$ we have $J \subset \mathcal{J}$ which shows that

$$(t^* - \delta, t^* + \delta) \cap I \ \subset \ \mathcal{J}.$$

So we conclude that $\mathcal{J}$ is open in I.

■

IV.2 Criteria for globality

The local Cauchy-Lipschitz theorem (Theorem I.46) gives us the existence and uniqueness of a maximal solution to a Cauchy problem but it does not say much *a priori* on the interval J on which this solution is defined. We can, for instance, wonder whether the interval J is equal to the interval I or not, in other words, whether the maximal solution is global or not. Of course, this is true if the vector field is globally lipschitz with respect to the state variable (thanks to the global Cauchy-Lipschitz theorem) but in general, we need some criteria to determine the globality of a given solution.

Let's start with a result where the state space is the whole $\mathbb{R}^d$, we will later give a more general result.

Theorem I.47 (Finite time blow-up theorem)

Let $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a continuous vector field that is locally lipschitz with respect to the state variable and let (t_0, x_0) be an associated Cauchy data.
Let (J, x) be the unique maximal solution of the Cauchy problem. Remember that J is always open; we denote it by $J = (\alpha, \beta)$ where $\beta \in (t_0, +\infty]$ and $\alpha \in [-\infty, t_0)$.

- *If $\beta < +\infty$, then $\sup_{t \in [t_0, \beta)} \|x(t)\| = +\infty$.*
- *If $\alpha > -\infty$, then $\sup_{t \in (\alpha, t_0]} \|x(t)\| = +\infty$.*

Remark I.48

- *Since x is a continuous function, it is bounded on all compact intervals and the above conclusions show that x is not bounded on a neighborhood of β (resp. α).*
- *We can actually show a stronger property by replacing the conclusions with $\lim_{t\to\beta} \|x(t)\| = +\infty$ (resp. $\lim_{t\to\alpha} \|x(t)\| = +\infty$), it is a good exercise and the reader is invited to try it.*

  The above version is, in general, sufficient for applications.

Proof :

Suppose that $\beta < +\infty$ and $R := \sup_{t\in[t_0,\beta)} \|x(t)\| < +\infty$.

By restriction and extension, we can construct a continuous vector field $f_R : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ that is globally lipschitz with respect to the state variable and that coincides with f on $[t_0, \beta+1] \times \bar{B}(0, R+1)$. The associated Cauchy problem to f_R and the data (t_0, x_0) admits a unique global solution such that the restriction to $J \cup \{\beta\}$ is a solution of the initial problem, which contradicts the maximality of the solution (J, x). ■

Exercise I.1

Study the set of solutions of the equation $x' = x(1-x)$ (without computing them).

The following is a more general version of the previous result, in the case where the vector field is only defined on an open set of $\mathbb{R}^d$.

Theorem I.49

Let I be a non empty open interval, $f : I \times \Omega \to \mathbb{R}^d$ a continuous vector field that is locally lipschitz with respect to the state variable and (t_0, x_0) an associated Cauchy data.
Let (J, x) be the unique maximal solution of the Cauchy problem; we denote $J = (\alpha, \beta)$ where $\beta > t_0$ and $\alpha < t_0$.

- *If $\beta \in I$, then for all compact $K \subset \Omega$, we have*

$$x([t_0, \beta)) \not\subset K.$$

- *If $\alpha \in I$, then for all compact $K \subset \Omega$, we have*

$$x((\alpha, t_0]) \not\subset K.$$

Example I.50

The following scalar equation is a good illustration of this phenomena

$$x' = \frac{1}{x(1-x)}.$$

The same remarks can be made as for the previous result.

Remark I.51

Let's consider the first case where $\beta \in I$, since similar remarks can be made for the other cases.

- *Since x is a continuous function, we can establish that for all $t^* \in J$, we have $x([t^*, \beta)) \not\subset K$ so x exits K no matter how close we are to the extremity β.*
- *We can actually show that there exists $t^* \in J$ such that $x([t^*, \beta)) \subset \Omega \setminus K$, which is of course stronger.*

  The above version is, in general, sufficient for applications.

Proof :

Let's concentrate on the first case. Suppose that $\beta \in I$ and, for a certain compact $K \subset \Omega$, we have $x([t_0, \beta)) \subset K$.

Construct a bounded open set U such that $K \subset U$ and $\overline{U} \subset \Omega$[5]. In addition, since $\beta \in I$, we can find $\delta > 0$ such that $[t_0, \beta+\delta] \subset I$. By applying an argument of restriction extension (Corollary I.19) on $\mathcal{K} = [t_0, \beta+\delta] \times \overline{U}$, we get a vector field $\tilde{f}$ that is continuous and globally lipschitz on $\mathbb{R} \times \mathbb{R}^d$, which coincides with f on $\mathcal{K}$.

Therefore, there exists a unique global solution $\tilde{x} : \mathbb{R} \to \mathbb{R}^d$ to the Cauchy problem

$$\begin{cases} \tilde{x}'(t) = \tilde{f}(t, \tilde{x}(t)), \quad \forall t \in \mathbb{R}, \\ \tilde{x}(t_0) = x_0, \end{cases} \tag{I.19}$$

- Since $\tilde{f}$ and f coincide on $[t_0, \beta) \times K$ and $x([t_0, \beta)) \subset K$, we can see that $([t_0, \beta), x)$ is also a solution of (I.19). By the uniqueness principle, we can deduce that $x = \tilde{x}$ on $[t_0, \beta)$.
- Since the function $\tilde{x}$ is continuous, there exists $\varepsilon > 0$ (that we can choose so that $\varepsilon \leq \delta$) such that $\tilde{x}([t_0, \beta+\varepsilon)) \subset U$. Since f and $\tilde{f}$ coincide on $[t_0, \beta+\varepsilon) \times U$, we can observe that $\tilde{x}$ verifies

  $$\tilde{x}'(t) = \tilde{f}(t, \tilde{x}(t)) = f(t, \tilde{x}(t)), \quad \forall t \in [t_0, \beta+\varepsilon).$$

  so $([t_0, \beta+\varepsilon), \tilde{x})$ is a solution of the initial Cauchy problem (which coincides with the solution x on $[t_0, \beta)$).
- The function $\bar{x}$ defined by

  $$\bar{x}(t) = \begin{cases} x(t), & \forall t \in (\alpha, t_0], \\ \tilde{x}(t), & \forall t \in [t_0, \beta+\varepsilon), \end{cases}$$

  is then a solution to the initial Cauchy problem defined on a strictly bigger interval than $J = (\alpha, \beta)$, which contradicts the maximality of (J, x).

■

V Equilibrium. Stability. Asymptotic stability

In this section, we will consider an **autonomous** vector field $f : \Omega \to \mathbb{R}^d$ of class $\mathcal{C}^1$ (hence locally lipschitz)[6]. In this case, the initial time for the Cauchy data has no importance and can be arbitrarily chosen as $t_0 = 0$.

Definition I.52 (Equilibrium)

We call an **Equilibrium of the system** *(or a critical point of the vector field f) any point $x^* \in \Omega$ such that*

$$f(x^*) = 0.$$

This implies that the constant map $t \in \mathbb{R} \mapsto x^*$ is a particular solution of the ordinary differential equation $x' = f(x)$.

In the sequel, we will be interested in the stability property (when $t \to +\infty$) of this solution, which is formalized in the following way.

Definition I.53

Let x^* be an equilibrium of the differential equation $x' = f(x)$.

- *We say that x^* is a* **stable** *equilibrium if : for all $\varepsilon > 0$, there exists $\delta > 0$ such that, for all initial data x_0 at a distance of at most δ from x^*, the solution of the equation starting from x_0 is, on the one hand, defined on the interval $[0, +\infty)$ and, on the other hand, stays close to x^* at a distance of ε in the course of time. See figure I.6. By using the notion of flow, this property can be written as*

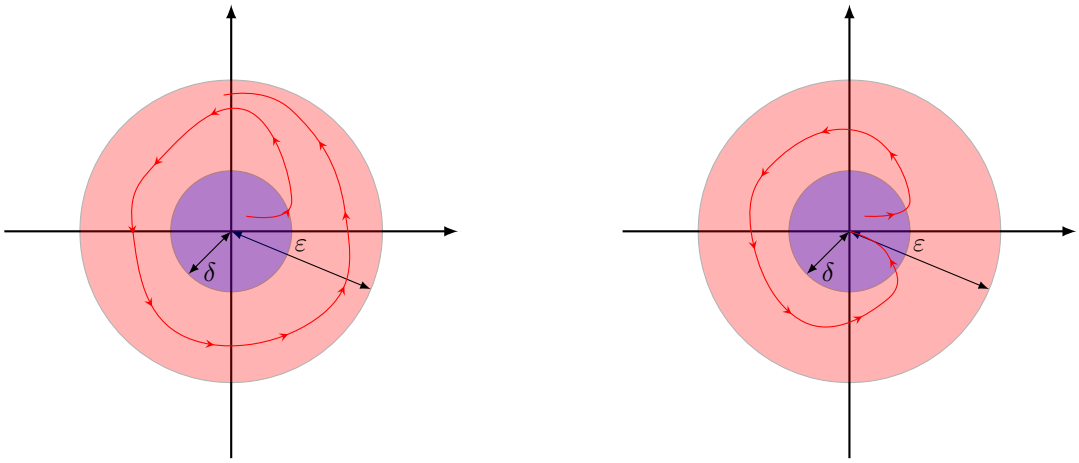
  $$\forall \varepsilon > 0, \exists \delta > 0, \quad \Phi(t)(B(x^*, \delta)) \subset B(x^*, \varepsilon), \quad \forall t \geq 0.$$

- *We say that x^* is an* **asymptotically stable** *equilibrium if it is stable and if on top of that, for any given initial data x_0 sufficiently close to x^*, the associated solution converges to x^* as t goes to infinity. By using the notion of flow, this can be expressed in the following way: there exists $\delta > 0$ such that*

  $$\forall x_0 \in B(x^*, \delta), \quad \varphi(t, x_0) \xrightarrow[t \to \infty]{} x^*.$$

[5] Do you know how to construct such an open set ?

[6] We can refine the regularity hypothesis but that will not be our main interest here



(a) The origin point is stable

(b) The origin is asymptotically stable

Figure I.6: Notions of stability of an equilibrium point

One should carefully note that in the definition of stability, the radius δ can be smaller that the radius ε: in particular, one cannot say that the balls are invariant under the action of the flow ! This is illustrated on the figure I.7.

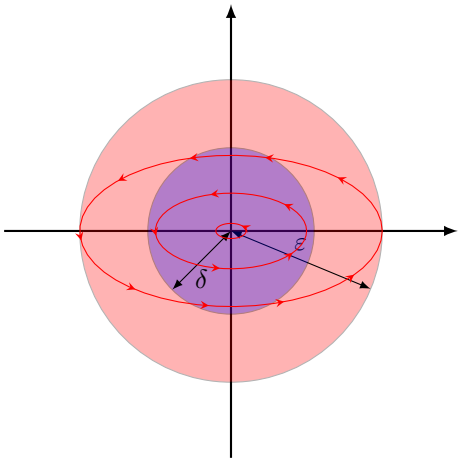


Figure I.7: An example of a stable system that is not asymptotically stable

V.1 Linear case

First, let us try to understand the linear case $x' = Ax$ where $A \in M_d(\mathbb{R})$.

V.1.a In the 2d plane

For what will follow, it is important to understand the geometry of solutions in the case where $d = 2$. Three cases can occur (one can go check, for instance, [5, page 290] for a more detailled discription):

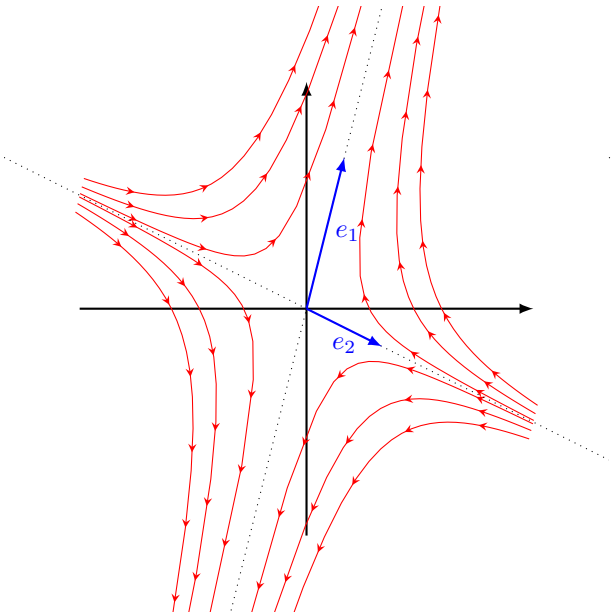
- First case: A can be diagonalized in $\mathbb{R}$.

If λ_1 and λ_2 are the two eigenvalues and e_1, e_2 the associated eigenvectors, the solutions to the equation are given by

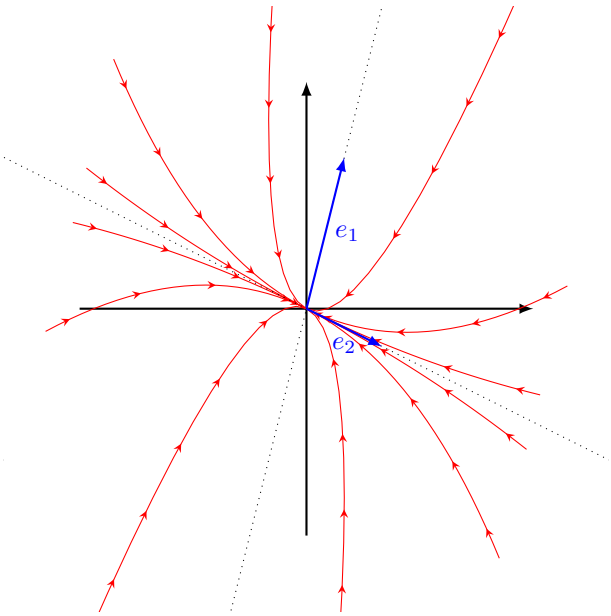
$$y(t) = \alpha_1 e^{\lambda_1 t} e_1 + \alpha_2 e^{\lambda_2 t} e_2.$$

Depending on the signs of the eigenvalues, one can draw the different trajectories, see Figure I.8.

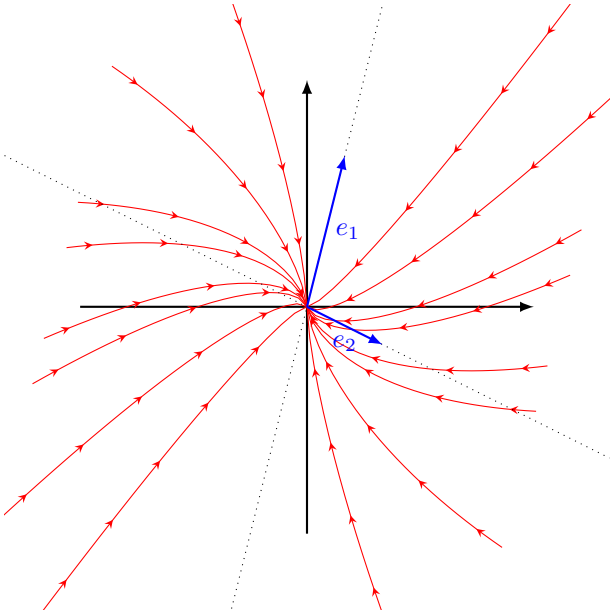
We can observe that the origin is stable (resp. asymptotically stable) if and only if the two eigenvalues are non positive (resp. negative).



(a) The case where $\lambda_1 > 0 > \lambda_2$: saddle point



(b) The case where $0 > \lambda_2 > \lambda_1$: stable node



(c) The case where $0 > \lambda_2 = \lambda_1$ cannot be diagonalized : degenerated stable node

Figure I.8: Three typical phase portraits in the case where the two eigenvalues are real

- Second case : A has real eigenvalues but cannot be diagonalized.

  In this case, the eigenvalue is necessarily unique and not semi-simple. From Jordan's theorem (or Dunford's decomposition theorem, see below Proposition I.55), we can deduce that the matrix A can be written as

$$P^{-1}AP = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix},$$

and in this case, we can see that the exponential term can be written as

$$P^{-1}e^{tA}P = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix}.$$

If we denote e_1 and e_2 the two columns of P (i.e. an eigenvector and a generalized eigenvector of A), we can find

$$y(t) = (\alpha_1 + \alpha_2 t)e^{\lambda t}e_1 + \alpha_2 e^{\lambda t}e_2.$$

The stability (and the asymptotic stability) of the equilibrium point is therefore equivalent to the negativity of the eigenvalue.

- Third case : A has two distinct complex conjugate eigenvalues $\lambda_\pm = a \pm bi$, where $b \neq 0$.
  In this case, we can verify that the exponential of tA is given by a formula of the following type

$$P^{-1}e^{tA}P = e^{ta}\begin{pmatrix} \cos(bt) & \sin(bt) \\ -\sin(bt) & \cos(bt) \end{pmatrix}.$$

The stability of the equilibrium is therefore given by the value of a : if $a = 0$ the trajectories are periodic, if $a > 0$ then the equilibrium point is unstable and if $a < 0$ the point is stable. Qualitatively, the trajectories go around the origin. The direction of rotation is given by the analysis of the eigenvectors of A, see Figure I.9.

V.1.b Stability of the origin for linear autonomous systems of differential equations

We can now state the general theorem. But before that, we need to remind a definition from linear algebra and a useful property.

Definition I.54

Let $A \in M_n(\mathbb{C})$ be any matrix and $\lambda \in \mathrm{Sp}(A)$. We say that λ is semi-simple if it is a simple root of the minimal polynomial of A, or even (equivalently) if Ker $(A-\lambda I) =$ Ker $(A-\lambda I)^2$.

Proposition I.55 (Dunford decomposition theorem)

Every matrix $A \in M_n(\mathbb{C})$ can be written in a unique way as

$$A = D + N,$$

where D is diagonalizable in $\mathbb{C}$, N is nilpotent and $DN = ND$.
In addition, A and D have the same eigenvalues and we have the following characterization:

$$\lambda \text{ is a semi-simple e.v. of } A \iff \text{Ker } (D-\lambda) \subset \text{Ker } N,$$

which means that N is null on the eigenspaces of D for the eigenvalue in question.

Proof :

For the existence and uniqueness of the decomposition, the reader is redirected to a standard course on linear algebra[7]. We will only prove the characterization property of the semi-simple eigenvalues.

- Let λ be a non semi-simple eigenvalue. There exists $v \in \mathbb{C}^n$, such that $(A-\lambda)^2 v = 0$ and $(A-\lambda)v \neq 0$. By using Newton's formula (which is valid since D and N commute) we get

$$(D-\lambda)^{n+1} = \sum_{k=0}^{n+1} C_{n+1}^k (-N)^k (A-\lambda)^{n+1-k} = \sum_{k=0}^{n-1} C_{n+1}^k (-N)^k (A-\lambda)^{n+1-k},$$

by using the fact that N is is nilpotent to eliminate the terms indexed by $k \geq n$. We can observe that the power on $(A-\lambda)$ in the formula is at least 2 and by multiplying the equality by the vector v on the right we get

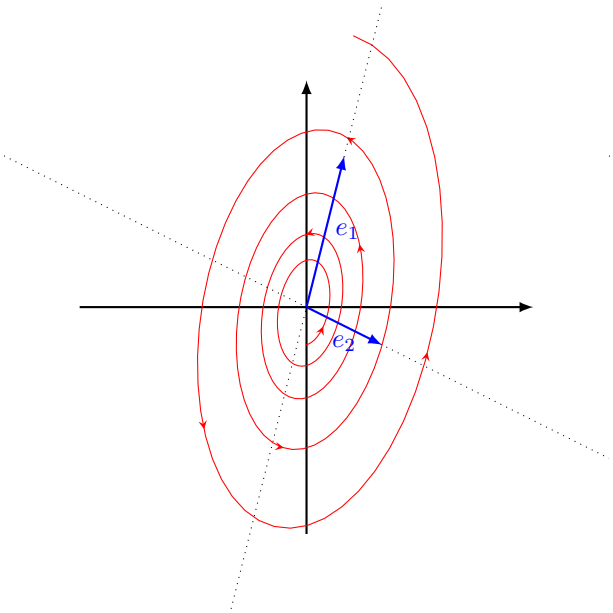
$$(D-\lambda)^{n+1}v = 0.$$

Since D is diagonalizable, this implies that v is an eigenvector of D for the eigenvalue of λ. By writing that $(A-\lambda) = (D-\lambda) + N$ and by multiplying by v we get that

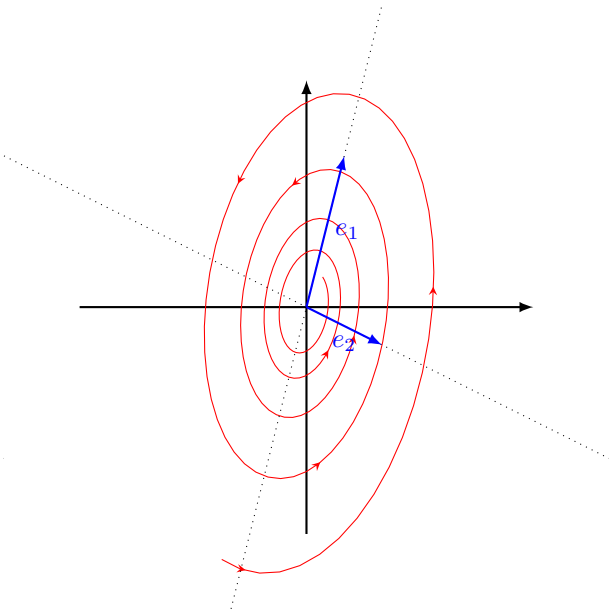
$$Nv = (A-\lambda)v \neq 0, \quad \text{by assumption.}$$

So we have found an element of Ker $(D-\lambda)$ that is not in Ker N.

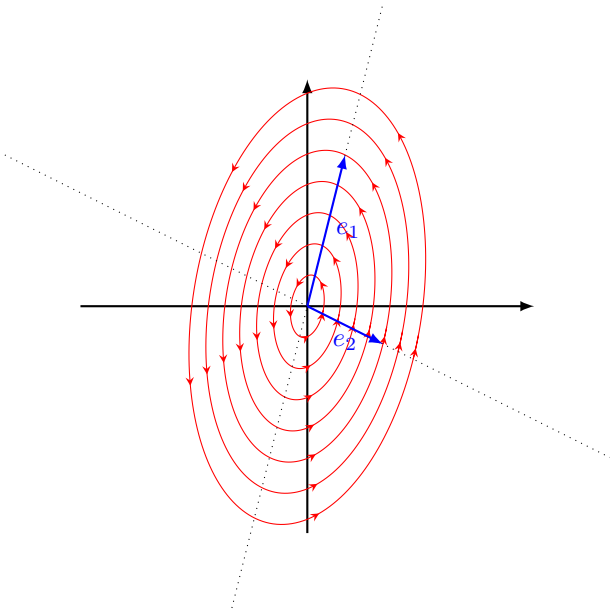
[7] see also the exercise sheets !



(a) The case where $\lambda_\pm = \alpha \pm i\beta$, and $\alpha > 0$: unstable focus



(b) The case where $\lambda_\pm = \alpha \pm i\beta$, and $\alpha < 0$: stable focus



(c) The case where $\lambda_\pm = \pm i\beta$: center

Figure I.9: Typical phase portraits in the case of two complex conjugated eigenvalues (the direction of rotation depends on the sign of β and on the relative position of the eigenvectors associated to λ_+ and λ_-)

- If now λ is a semi-simple eigenvalue, we start by observing that the hypothesis implies, by simple induction, that

$$\text{Ker } (A-\lambda)^k = \text{Ker } (A-\lambda)^{k+1}, \ \ \forall k \geq 0,$$

so

$$\text{Ker } (A-\lambda)^k = \text{Ker } (A-\lambda), \ \ \forall k \geq 0. \tag{I.20}$$

Once again, by using Newton's formula we can write

$$(A-\lambda)^n = \sum_{k=0}^{n} C_n^k N^k (D-\lambda)^{n-k} = \sum_{k=0}^{n-1} C_n^k N^k (D-\lambda)^{n-k},$$

by using the fact that N is nilpotent. In this formula, we observe that the powers of $(D-\lambda)$ are all positive and therefore if we take $v \in \text{Ker } (D-\lambda)$, then the previous formula shows that

$$(A-\lambda)^{n+1} v = 0,$$

and by using (I.20), we conclude that $(A-\lambda)v = 0$.

Thanks to the equality $A - \lambda = (D-\lambda) + N$, we can conclude that $Nv = 0$ and the result is thus proved.

■

Theorem I.56

Consider the linear and autonomous system of differential equations $x' = Ax$.

- *The equilibrium state $x^* = 0$ is asymptotically stable if and only if the real part of all the eigenvalues of A are negative.*

  In this case, the equilibrium is exponentially stable : there exists $C, \gamma > 0$ such that

$$\|e^{tA}\| \leq Ce^{-\gamma t}, \quad \forall t > 0.$$

- *The equilibrium state $x^* = 0$ is stable if and only if the real part of all the eigenvalues of A are non positive and those whose real part is zero are semi-simple.*

  In this case, the flow is uniformly bounded (for $t \geq 0$) : there exists $C > 0$ such that

$$\|e^{tA}\| \leq C, \quad \forall t > 0.$$

Note that the same result is valid for all other equilibrium points x^* (meaning for all x^* in the kernel of A).

Proof :

- First, suppose that A admits an eigenvalue $\lambda = \alpha + i\beta$ whose real part is positive. Let us show that the equilibrium point cannot be stable.

  Let $v = v_1 + iv_2$ be an eigenvector of A for this eigenvalue. Suppose that $v_1 \neq 0$ (which we can always do even if it means multiplying v by i).

  We choose a number L such that $\beta L \in 2\pi\mathbb{Z}$ and observe that, if we define $t_n = nL$, then we get

$$e^{t_n A} v = e^{t_n(\alpha+i\beta)} v = e^{t_n \alpha} e^{inL\beta} v = e^{t_n \alpha} v.$$

  Since the matrix A is real, we can take the real part of this equality and get that

$$e^{t_n A} v_1 = e^{t_n \alpha} v_1,$$

  which proves that, since $\alpha > 0$, $(e^{t_n A} v_1)_n$ is not bounded and therefore the origin is an unstable equilibrium point of the system.

- The same computation shows that if A admits a purely imaginary eigenvalue $i\beta$, then there exists $v_1 \neq 0$ and a sequence $(t_n)_n$ that goes to infinity such that

$$e^{t_n A} v_1 = v_1,$$

  and therefore the equilibrium cannot be asymptotically stable.

- Now, let us suppose that A admits a purely imaginary eigenvalue $\lambda = i\beta$ that is not semi-simple. We will show that the equilibrium point is not stable (but it should be noted that the instability is less violent than the previous case).

  From Proposition I.55, there exists an eigenvector v of D for the eigenvalue λ such that $Nv \neq 0$.

  Let us compute the solution of the equation of the given initial data v:

  $$e^{tA}v = e^{tN}e^{tD}v = e^{t\lambda}e^{tN}v = e^{it\beta}\sum_{k=0}^{n} C_n^k t^k N^k v.$$

  Since the factor $e^{it\beta}$ has an absolute value of 1, we see that $e^{tA}v$ is bounded if and only if the polynomial in t

  $$\sum_{k=0}^{n} C_n^k t^k N^k v,$$

  is bounded, which is not possible since the term of order 1 is not zero since $Nv \neq 0$.

- Let us now study the case where we have stability. First, we will diagonalize D (which is possible by definition of Dunford's decomposition). So there exists $P \in GL_d(\mathbb{C})$ such that

  $$P^{-1}DP = \Lambda, \text{ where } \Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_d),$$

  and since $\tilde{N} = P^{-1}NP$ is also nilpotent (and commutes with Λ) we see that under those conditions, we can easily compute

  $$e^{tA} = Pe^{t\Lambda}e^{t\tilde{N}}P^{-1}.$$

  The matrix P does not play a role in the (asymptotic) stability of the system and therefore we can concentrate on the study of the stability of the following flow

  $$t \mapsto e^{t\Lambda}e^{t\tilde{N}}.$$

  Considering that

  $$e^{t\Lambda} = \operatorname{diag}(e^{t\lambda_1}, ..., e^{t\lambda_d}),$$

  and that

  $$|e^{t\lambda_i}| = e^{t\Re\lambda_i},$$

  we can have the following cases:

  – If the real part of all the eigenvalues are negative, we have

    $$\|e^{t\Lambda}\| \leq e^{-\gamma^* t},$$

    where

    $$\gamma^* := \inf\{(-\Re\lambda_i), 1 \leq i \leq d\},$$

    which is, by assumption, a positive number.

    Since $\tilde{N}$ is nilpotent, $e^{t\tilde{N}}$ is polynomial in time and therefore we have

    $$\|e^{tA}\| \leq \|P\|\|P^{-1}\|\|e^{t\tilde{N}}\|e^{-\gamma^* t},$$

    and, since the exponential term dominates the polynomial terms, we get, for all $0 < \gamma < \gamma^*$, the existence of $C_\gamma > 0$ such that

    $$\|e^{tA}\| \leq C_\gamma e^{-\gamma t}, \ \ \forall t \geq 0,$$

    which implies the asymptotic stability of the equilibrium point.

  – If there exist eigenvalues whose real parts are zero, we can no longer reason in the same way since $\|e^{t\Lambda}\| = 1$ and the polynomial terms in time might be unpleasant to work with.

    We decompose the space $\mathbb{C}^n$ in two parts $\mathbb{C}^n = E_- \oplus E_0$ where E_- is the sum of the eigenspaces of Λ associated to the eigenvalues with a negative real part and E_0 the sum of the eigenspaces associated to eigenvalues that are purely imaginary. Since $\tilde{N}$ commutes with Λ, the spaces E_- and E_0 are stable under $\tilde{N}$. On the other hand, the semi-simplicity hypothesis assures that $\tilde{N}(E_0) = 0$.

    By studying seperately what happens on E_- and E_0, we can reach the desired conclusion.

■

V.2 Non-linear case

V.2.a The theorem to know

Theorem I.57 (Spectral criterion for stability)

Let $x^* \in \mathbb{R}^d$ such that $f(x^*) = 0$. We denote $A = D_x f(x^*)$ the Jacobian matrix of f at the point x^*.

1. *If all the eigenvalues of A have a* **negative** *real part, then x^* is an asymptotically stable equilibrium point of the equation $x' = f(x)$.*
2. *If all the eigenvalues of A have a* **positive** *real part, then x^* is an unstable equilibrium point of the equation $x' = f(x)$.*
3. *If all the eigenvalues of A have a non positive real part and at least one of them has a zero real part, then we cannot conclude with this criterion.*

Proof :

Let us start by assuming that $x^* = 0$ without loss of generality (since we can always change f into $f - f(x^*)$).

1. From the hypothesis on the eigenvalues of A and the computations that were made in the linear case, we know that there exists $\gamma > 0$ and $C_1 > 0$ such that

$$\|e^{tA}\| \leq C_1 e^{-\gamma t}, \ \forall t \geq 0.$$

Now, we define $g(x) := f(x) - Ax$. By definition of the Jacobian at 0, we know that $\lim_{x\to 0} g(x)/\|x\| = 0$. In particular, there exists $\delta > 0$ such that

$$\|g(x)\| \leq \frac{\gamma}{2C_1}\|x\|, \ \forall x \in \bar{B}(0,\delta).$$

So we define

$$g_\delta(x) := g(P_\delta x),$$

where P_δ is the orthogonal projection on $\bar{B}(0,\delta)$. It is clear that g_δ is globally lipschitz and that it verifies

$$\|g_\delta(x)\| \leq \frac{\gamma}{2C_1}\|x\|, \ \forall x \in \mathbb{R}^d.$$

So we define $f_\delta(x) := Ax + g_\delta(x)$ which is a globally lipschitz map on $\mathbb{R}^d$ and coincides with f on $\bar{B}(0,\delta)$.
Now, let x_δ be the global solution of the system $x' = f_\delta(x)$ for a fixed initial data x_0. We write the equation as

$$x'_\delta = Ax_\delta + g_\delta(x_\delta),$$

and use the Duhamel formula to write

$$x_\delta(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A} g_\delta(x_\delta(s))\, ds,$$

and by taking the norm, for $t \geq 0$, we get

$$\|x_\delta(t)\| \leq C_1 e^{-\gamma t}\|x_0\| + C_1 \frac{\gamma}{2C_1}\int_0^t e^{-\gamma(t-s)}\|x_\delta(s)\|\, ds.$$

By rewriting this inequality as follows

$$e^{\gamma t}\|x_\delta(t)\| \leq C_1\|x_0\| + \frac{\gamma}{2}\int_0^t e^{\gamma s}\|x_\delta(s)\|\, ds, \ \forall t \geq 0,$$

we can apply Gronwall's lemma to the function $t \mapsto e^{\gamma t}\|x_\delta(t)\|$ and obtain the following inequality

$$e^{\gamma t}\|x_\delta(t)\| \leq C_1\|x_0\| e^{\gamma t/2}, \ \forall t \geq 0,$$

and immediately deduce that

$$\|x_\delta(t)\| \leq C_1\|x_0\| e^{-\gamma t/2}.$$

- The first consequence of this inequality is that

$$\|x_\delta(t)\| \leq C_1\|x_0\|, \ \ \forall t \geq 0,$$

so if we choose an initial state x_0 that verifies

$$\|x_0\| \leq \frac{\delta}{C_1},$$

then the solution x_δ of the modified system verifies

$$\|x_\delta(t)\| \leq \delta, \ \ \forall t \geq 0,$$

therefore, since f_δ and f coincide on the ball centered at zero and with radius δ, we have

$$x'_\delta(t) = f_\delta(x_\delta(t)) = f(x_\delta(t)).$$

So $t \in [0,\infty) \mapsto x_\delta(t)$ is non other than the unique maximal solution in positive time of the initial Cauchy problem $x' = f(x)$. Which proves the existence of global solutions.

- Once again, take an initial data that verifies $\|x_0\| \leq \frac{\delta}{C_1}$, we have finally obtained that

$$\|x(t)\| \leq C_1\|x_0\|e^{-\gamma t/2},$$

and that $\lim_{t\to\infty} x(t) = 0$, which proves the asymptotic stability of the equilibrium point 0 for the system $x' = f(x)$.

Actually, we have shown a much stronger property (the exponential stability) which shows that the flow φ of the equation verifies, for all $r > 0$ small enough,

$$\varphi(t, B(0,r)) \subset e^{-\gamma t/2} B\left(0, rC_1\right).$$

2. We denote $(t,x) \mapsto \varphi(t,x) = \varphi(t,0,x)$ the flow of the system that only depends on one variable of time since the system is autonomous (see the remark I.36). We immediately observe that $\psi(t,x) = \varphi(-t,x)$ is the flow of the system (backward in time)

$$y' = -f(y),$$

which has the same equilibrium x^*. On the other hand, we know that $\psi(t,\cdot) = \varphi(t,\cdot)^{-1}$.

Since the Jacobian of $-f$ is equal to $-A$ and that it has eigenvalues with negative real part, we know that x^* is exponentially stable for this new system and so, for all $r > 0$ small enough we have

$$\psi(t, B(0,r)) \subset e^{-\gamma t/2} B\left(0, rC_1\right) = B\left(0, rC_1 e^{-\gamma t/2}\right).$$

By applying $\varphi(t,.)$ to this formula we find

$$B(0,r) \subset \varphi\left(t, B\left(0, rC_1 e^{-\gamma t/2}\right)\right), \ \ \forall t \geq 0.$$

Now, we fix $r > 0$ for which this property is true and also give ourselves any $\delta > 0$. By defining

$$t_\delta = \frac{2}{\gamma}\log\left(\frac{rC_1}{\delta}\right),$$

we can see that the previous inclusion gives us that

$$B(0,r) \subset \varphi(t_\delta, B\left(0,\delta\right)).$$

In particular, for all x such that $\|x\| = r/2$, there exists x_0 such that $\|x_0\| < \delta$ and

$$\varphi(t_\delta, x_0) = x,$$

so in particular

$$\|\varphi(t_\delta, x_0)\| = r/2,$$

which proves that, no matter how small δ is, the trajectory will always find a way to exit the ball of radius $r/2$ so the equilibrium point cannot be stable.

Of course, we remark that $t_\delta \to +\infty$ when $\delta \to 0$, in other words : the smaller $\delta > 0$ is, the longer it will take to get out of the ball of radius $r/2$, which is natural.

3. A very simple example is the following scalar problem:

$$x' = \alpha x^3,$$

where $\alpha \in \{-1, 0, 1\}$. Note that this equation can be solved explicitly without much difficulty[8] but in what will follow we will be using more generic arguments.

The "Jacobian" of f is the derivative $f'(0) = 0$ which has of course no real part.

- For $\alpha = 0$, the system is reduced to $x' = 0$ and the stability of the equilibrium is clear.
- For $\alpha = 1$, we have $x' = x^3$. If $x_0 > 0$, x stays positive so x is increasing.
  As long as the solution exists, we have

$$x'(t) \geq x_0^3, \ \ \forall t \geq 0, t \in J,$$

  so

$$x(t) \geq x_0 + x_0^3 t, \ \ \forall t \geq 0, t \in J.$$

  For any choice of $x_0 > 0$, even very small, we have that $\lim_{t \to \sup J} x(t) = +\infty$ so the equilibrium is unstable.

  In a similar way, we can show that $\lim_{t \to \inf J} x(t) = -\infty$ for $x_0 < 0$.
- For $\alpha = -1$, we have $x' = -x^3$. If $x_0 > 0$, the solution stays non negative and non increasing. On the other hand, it is bounded by 0 from below. The finite time blow-up theorem shows that this solution is well defined on $[0, +\infty)$ and from the previous properties we know that there exists $\bar{x} \in [0, x_0)$ such that $\lim_{t \to +\infty} x(t) = \bar{x}$.

  So from the differential equation we get that

$$\lim_{t \to +\infty} x'(t) = -\bar{x}^3.$$

  If we now use the lemma I.58 presented below, we can deduce that

$$-\bar{x}^3 = 0,$$

  so

$$\bar{x} = 0.$$

  So we have shown that, for any positive initial data the solution converges to 0, the equilibrium point. We can reason in a similar way for the negative datas and conclude that this point is asymptotically stable.

■

Lemma I.58

Let $x : [0, +\infty) \to \mathbb{R}$ be a differentiable function. If $\lim_{t \to +\infty} x(t)$ and $\lim_{t \to +\infty} x'(t)$ exists, then we have

$$\lim_{t \to +\infty} x'(t) = 0.$$

Remark I.59

If one does not assume that the limit of x' exists, then the result is false. One can consider for instance the function $x(t) = \frac{\sin(t^2)}{t}$.

Proof :

Let us denote $\alpha = \lim_{t \to +\infty} x'(t)$ and assume that $\alpha \neq 0$. We can assume that $\alpha > 0$, since we can always change x into $-x$. By definition of limits, there exists $A \geq 0$ such that

$$x'(t) \geq \frac{\alpha}{2}, \ \ \forall t \geq A.$$

By integrating this between the bounds A to $t \geq A$, we obtain that

$$x(t) - x(A) \geq \frac{\alpha}{2}(t - A), \ \ \forall t \geq A,$$

so

$$x(t) \geq x(A) + \frac{\alpha}{2}(t - A) \xrightarrow[t \to +\infty]{} +\infty,$$

which implies that $\lim_{t \to +\infty} x(t) = +\infty$. This contradicts the assumptions of the lemma. ■

[8] Do it as an exercise !

V.2.b A more precise theorem but out of scope of this lecture

The theorem of the last section only allows us to conclude when all the eigenvalues of the Jacobian matrix have a negative real part or a positive real part. Obviously, it is not possible to understand the behavior of the solutions in most cases with such a criterion. The following theorem allows us to understand this behavior in a more general case (and, once again, this does not treat all the possible cases).

Theorem I.60 (Hartmann-Grobman)

Let $x' = f(x)$ be an autonomous system of differential equations and 0 an equilibrium of this system. Let $A = Df(0)$ be the Jacobian matrix.
Suppose that 0 is a **hyperbolic** *equilibrium (which means that all the eigenvalues of A have a* **non zero** *real part).*
Then, in a neighborhood of 0, the flow φ of the equation is conjugated to the flow of the linear system $y' = Ay$ (denoted ψ).
This means that there exist $\delta > 0$, two open sets U and $V \subset \mathbb{R}^d$ containing 0, and a homeomorphism $h : U \to V$ such that

$$\forall t \in (-\delta, \delta), \quad \varphi(t, .) = h^{-1} \circ \psi(t, .) \circ h, \quad \text{in } U.$$

In other words,

$$\forall t \in (-\delta, \delta), \forall x \in U, \quad \varphi(t, x) = h^{-1}(e^{tA} h(x)).$$

In particular, for a hyperbolic equilibrium, it is sufficient to have at least one of the eigenvalues with a positive real part for the equilibrium to be unstable, which is more precise that the second point stated in theorem I.57.

It should be noted that this result still says nothing about the case where some eigenvalues have a zero real part and, actually, the theorem is false in this case as we have seen previously on the system $x' = \alpha x^3$. In this case, the linearized system is $y' = 0$ for which all the solutions are constant. If the associated (trivial) flow ψ was conjugated to the flow φ of the non linear equation, we would deduce that, in a neighborhood of 0, all the solutions to the non linear equation are constant, which is obviously not the case.

To pursue this study even further, some new tools are necessary like the notion of central (resp. stable, unstable) manifold, or the notion of normal form of a system in a neighborhood of an equilibrium point. This is totally outside the scope of this course but it is good to know that, in case of need, this theory exists and has been very well developed.

V.2.c Conserved quantities. Invariant sets

Definition I.61

A differentiable function $E : \Omega \to \mathbb{R}$ is called a **conserved quantity** *of the vector field f if E is constant on all the trajectories of the system :*

$$\frac{d}{dt} E(\varphi(t, x)) = 0, \quad \forall t \in I, \forall x \in \Omega.$$

Lemma I.62

The function E is a conserved quantity of f if and only if we have

$$(f \cdot \nabla E)(x) = 0, \quad \forall x \in \Omega,$$

in other words if, at every point, the vector $f(x)$ and the gradient of E at x are orthogonal.

Definition I.63

For any function $E : \Omega \to \mathbb{R}$ and any value α, we call the **level set of** *E for the level α the set E_α defined by*

$$E_\alpha := \{x \in \Omega, \quad E(x) = \alpha\} = E^{-1}(\{\alpha\}).$$

We call **sub-level set** *(resp.* **strict sub-level set***) of E for the level α, the set $E_{\leq \alpha}$ (resp. $E_{<\alpha}$) defined by*

$$E_{\leq \alpha} := \{x \in \Omega, \quad E(x) \leq \alpha\} = E^{-1}((-\infty, \alpha]),$$

$$E_{<\alpha} := \{x \in \Omega, \quad E(x) < \alpha\} = E^{-1}((-\infty, \alpha)).$$

If such a function exists, we can deduce that all trajectories are completely contained in the level set E_α.

If the function E is constant, this does not mean much but if the function E has some good properties, the set E_α will be a hypersurface (i.e. a surface in dimension $d = 3$, a curve in dimension $d = 2$) and this can give us some valuable information. Let us consider two examples :

- *The pendulum equation:*

  Let us consider the pendulum equation (where the physical constants are taken to be 1):

  $$\theta'' + \sin(\theta) = 0.$$

  We observe that the (positive) quantity

  $$E(\theta, \theta') = \frac{1}{2}(\theta')^2 + (1 - \cos(\theta)),$$

  is conserved in the course of time, so it is indeed a conserved quantity of the system.

  Moreover, for $\alpha \in [0, 2)$ the level curve E_α are unions of closed curves. The trajectories of the system corresponding to a Cauchy data verifying $E(\theta_0, \theta_0') < 2$ are always contained in one of these curves.

- *Lotka-Voltera* : it is a population dynamics model that takes the following form

  $$\begin{cases} x' = ax - bxy, \\ y' = -cy + dxy, \end{cases} \tag{I.21}$$

  where the constants a, b, c and d are all positive.

  This system was introduced to model the evolution of the population size of preys (say, for example, sardines whose population size at time t is $x(t)$) and predators (say sharks whose population size at time t is $y(t)$). The number $a > 0$ represents the growth rate of sardines (they have infinite resources at their disposal, the plankton), the number $-c < 0$ represents the death rate of sharks in the absence of food. The terms $-bxy$ and dxy represent the impact of a shark/sardine meeting on the two populations : the sharks eat the sardines so the number of sardines decreases and the sharks regain their health, then reproduce, etc ...

  We can show that this system is globally well posed for all positive initial data x_0, y_0 (see Section VI for a more detailed study of a similar example).

  Furthermore, we can observe that there is a unique equilibrium point in the area of interest given by $x^* = c/d$ and $y^* = a/b$. Moreover, we can verify that the system admits a conserved quantity given by

  $$E(x, y) = by + dx - a \log y - c \log x.$$

  An elementary study shows that the level curves of E are closed curves that can be easily drawn and thus we can obtain a phase portrait of the system (see Figure I.10)

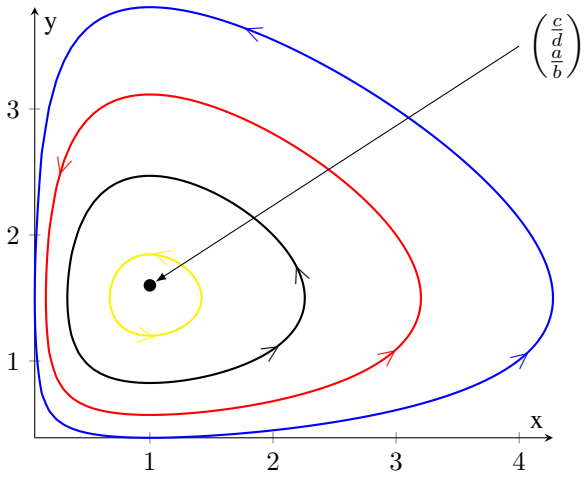


Figure I.10: Phase portrait of the Lotka-Volterra model

It may happen that, even if the function E is not constant on the trajectories of the system, we can get interesting information about these trajectories by using the following argument.

Proposition I.64 (Invariant strict sub-level sets)

Let E be a $\mathcal{C}^1$ function and $\alpha \in \mathbb{R}$ a given real number.
Suppose that

$$\forall x \in E_\alpha, \ \ (f \cdot \nabla E)(x) < 0,$$

then the strict sub-level set $E_{<\alpha}$ is invariant by the flow (in positive time), meaning that

$$\Phi(t)(E_{<\alpha}) \subset E_{<\alpha}, \ \ \forall t \geq 0.$$

Example I.65

Let us consider the following system in dimension 2

$$\begin{cases} x' = x(1 - ax^2 - by^2), \\ y' = y(1 - cx^2 - dy^2), \end{cases}$$

where $a, b, c, d > 0$. Consider the function $E(x, y) = x^2 + y^2$.
By performing an elementary computation, we get

$$(f \cdot \nabla E)(x, y) = 2x^2(1 - ax^2 - by^2) + 2y^2(1 - cx^2 - dy^2).$$

Moreover, there exists $\alpha_0 > 0$ large enough such that

$$x^2 + y^2 = \alpha_0 \Rightarrow ax^2 + by^2 > 1,$$

$$x^2 + y^2 = \alpha_0 \Rightarrow cx^2 + dy^2 > 1.$$

For any value of $\alpha \geq \alpha_0$, we have $(f \cdot \nabla E) < 0$ on E_α. Therefore the strict sub-level sets $E_{<\alpha}$ (that are none other than some sufficiently big open balls) are invariant under the flow. We deduce that, in particular, all the maximal solutions of the system are bounded and therefore are well defined on $[0, +\infty)$, thanks to the finite time blow-up theorem.
Note, however, that according to the spectral criterion (Theorem I.57), the origin is an unstable equilibrium point of the system.

V.2.d A few words on Lyapunov's theory

Consider a vector field $f : \Omega \to \mathbb{R}^d$ that is locally lipschitz.

Let $V : \Omega \to \mathbb{R}$ be a function of class $\mathcal{C}^1$ that verifies the following properties[9]:

- **Monotony/dissipation relative to f:**

$$f(x) \cdot \nabla V(x) \leq 0, \ \ \forall x \in \Omega,$$

- **Structure of critical points of V relative to f:**

  Define

$$C := \{x \in \Omega \ \text{ s.t. } f(x) \cdot \nabla V(x) = 0\},$$

  the set of critical points of V relative to f. Suppose that for any $\alpha \in \mathbb{R}$, the intersection of C with the level set of V of level α defined by

$$C_\alpha := C \cap V_\alpha,$$

  is locally finite.

- **Compactness:**

  The sub-level sets $V_{\leq \alpha}$ have compact connected components.

[9] other set of hypothesis can be considered and we don't claim to have an optimal result here

Theorem I.66 (Liapounov's Theorem)

Under the previous assumptions, for any given initial data x_0, the associated solution $t \mapsto \varphi(t, x_0)$ converges as $t \to +\infty$ to an equilibrium point of the system.
In particular, all the connected components of the $V_{\leq \alpha}$ contains one equilibrium point.

In practice, we often use this theorem through the following corollary that can be deduced immediately.

Corollary I.67

Let x^* be an equilibrium point of the system. If there exists $\alpha > V(x^*)$ such that x^* is the unique equilibrium point of the system in its connected component in $V_{\leq \alpha}$ then x^* is asymptotically stable.
In particular, every strict local minima of V is an asymptotically stable equilibrium point of the system $x' = f(x)$.

Proof (of Theorem I.66):
We denote $(t, x) \mapsto \varphi(t, x)$ the flow of the system (that only depends on one single time variable since the system is autonomous).

- Step 1 : *Compactness of the trajectories.*

  From our hypotheses, we have that for any $x \in \Omega$,

  $$\frac{d}{dt} V(\varphi(t, x)) = \partial_t \varphi(t, x) \cdot \nabla V(\varphi(t, x)) = (f \cdot \nabla V)(\varphi(t, x)) \leq 0.$$

  Therefore the function V decreases along the trajectories, in particular $\varphi(t, x_0)$ is an element of the sub-level set $V_{\leq \alpha_0}$ where $\alpha_0 := V(x_0)$ for all $t \geq 0$ for which the solution is defined. Since the trajectories depend continuously on time, it is even contained in one of the connected components of $V_{\leq \alpha_0}$.

  Since these connected components are compact by assumption, Theorem I.49 shows that the solution in question is defined on $[0, +\infty)$, and this is true for all given initial data.

  Furthermore, since the half-trajectory $(\varphi(t, x_0))_{t \geq 0}$ is contained in a compact set, it has some accumulation points. We will now show that the half-trajectory actually has only one accumulation point and that it converges to this one.

- Step 2 : *Characterization of accumulation points.*

  Let $a = \lim_{n \to \infty} \varphi(t_n, x_0)$ be such an accumulation point, where $t_n \to \infty$. By monotony and continuity, we obtain that

  $$V(\varphi(t_n, x_0)) \xrightarrow[n \to \infty]{} V(a), \quad \text{while decreasing.}$$

  Since $t \mapsto V(\varphi(t, x_0))$ is a non increasing function with an accumulation point, it converges to this point. In other words, we have obtained that

  $$\lim_{t \to \infty} V(\varphi(t, x_0)) = V(a). \tag{I.22}$$

  For all $s \geq 0$, by using the group property and the continuity of the flow and also the continuity of V

  $$V(\varphi(t_n + s, x_0)) = V(\varphi(s, \varphi(t_n, x_0))) \xrightarrow[n \to \infty]{} V(\varphi(s, a)),$$

  therefore, by comparing this to (I.22), we deduce

  $$V(\varphi(s, a)) = V(a), \quad \text{for all } s.$$

  The trajectory $s \mapsto \varphi(s, a)$ is thus contained in the level set V_α, for $\alpha := V(a)$. Moreover, by differentiating this equality with respect to s, we obtain that $\varphi(s, a) \in C$, for all s. So, we have

  $$\varphi(s, a) \in C_\alpha, \quad \forall s \geq 0.$$

  But since we assumed that C_α is locally finite, we obtain $\varphi(s, a) = a$ for all s, in other words, that a is an equilibrium of the system, i.e. $f(a) = 0$.

- Step 3 : *Convergence to the accumulation point a and conclusion.*

  We will reason by contradiction and suppose that $\varphi(t, x_0)$ does not converge to a as $t \to \infty$. This means that there exists $\varepsilon_0 > 0$ and an increasing sequence $(\tau_n)_n$ tending towards infinity such that

  $$\|\varphi(\tau_n, x_0) - a\| \geq \varepsilon_0.$$

  We can assume that

  $$\|\varphi(t_n, x_0) - a\| \leq \varepsilon_0/2,$$

  and

  $$t_n < \tau_n < t_{n+1}, \ \forall n \geq 0,$$

  otherwise, we can replace the sequences with convenient sub-sequences. Fix $\varepsilon_0/2 < \varepsilon < \varepsilon_0$. By construction, we have

  $$\|\varphi(t_n, x_0) - a\| < \varepsilon < \|\varphi(\tau_n, x_0) - a\|, \ \ \forall n \geq 0.$$

  By the intermediate value theorem, we can deduce that there exists $s_n \in (t_n, \tau_n)$ verifying

  $$\|\varphi(s_n, x_0) - a\| = \varepsilon.$$

  Since the sequence $(\varphi(s_n, x_0))_n$ is contained in a compact, it admits an accumulation point, denoted c_ε, which is, with the help of the above reasoning, a point of C_α and that verifies $\|a - c_\varepsilon\| = \varepsilon$. Since this is true for all ε, this contradicts the fact that C_α is locally finite which concludes the proof by contradiction. So, under the given assumptions we have shown the convergence of the trajectory to the point a. The proof is thus complete.

■

With the help of Liapounov's method, we obtain the asymptotic stability criterion for linear systems given in the first part of the theorem I.57, and here it will be presented as a corollary. Note that Liapounov's method is more precise because it can allow us to evaluate the size of the neighborhood of the equilibrium for which we get convergence. In some cases, it even allows us to prove the global asymptotic stability of an equilibrium (i.e. : for any given initial data, the solution converges to x^*), which we can never do with a criterion of linear nature that only sees what happens when we are close to the equilibrium.

Corollary I.68

Let x^* an equilibrium of the system $x' = f(x)$ and A the Jacobian matrix of f at the point x^*. If all the eigenvalues of A have a negative real part, then there exists $R > 0$ such that the system admits a Liapounov function on the ball $\Omega = B(0, R)$ and, in particular, this equilibrium is asymptotically stable.

Proof :

We have already seen that, under the theorem's hypotheses, there exist $C, \gamma > 0$ such that

$$\|e^{tA}\| \leq Ce^{-t\gamma}, \ \ \forall t \geq 0.$$

We can therefore define the following function

$$V(x) = \int_0^{+\infty} \|e^{sA}x\|^2 \, ds,$$

that verifies, for a well chosen $\alpha_{min}, \alpha_{max} > 0$,

$$\alpha_{min}\|x\|^2 \leq V(x) \leq \alpha_{max}\|x\|^2, \ \ \forall x \in \mathbb{R}^d. \tag{I.23}$$

A simple computation shows that

$$\begin{aligned}
\frac{d}{dt}V(\varphi(t,x)) &= 2\int_0^{+\infty} \left(e^{sA}f(\varphi(t,x)), e^{sA}\varphi(t,x)\right) ds \\
&= 2\int_0^{+\infty} \left(e^{sA}A\varphi(t,x), e^{sA}\varphi(t,x)\right) ds + 2\int_0^{+\infty} \left(e^{sA}g(\varphi(t,x)), e^{sA}\varphi(t,x)\right) ds \\
&= \int_0^{+\infty} \frac{d}{ds}\|e^{sA}\varphi(t,x)\|^2 \, ds + 2\sqrt{V(g(\varphi(t,x)))}\sqrt{V(\varphi(t,x))} \\
&= -\|\varphi(t,x)\|^2 + 2C_2\|g(\varphi(t,x))\|\|\varphi(t,x)\|.
\end{aligned}$$

By evaluating at $t = 0$, we get

$$(f \cdot \nabla V)(x) \leq -\|x\|^2 + 2C_2\|g(x)\|\|x\|.$$

Let $\delta > 0$ such that

$$\|g(y)\| \leq \frac{1}{4C_2}\|y\|, \ \ \forall y \in \bar{B}(0,\delta),$$

then we define $\Omega = B(0,\delta)$. From the previous computation, for any $x \in \Omega$, we have

$$(f \cdot \nabla V)(x) \leq -\|x\|^2 + \frac{1}{2}\|x\|^2 = -\frac{1}{2}\|x\|^2.$$

This shows the monotony of V relative to f. The set of critical points of V relative to f is simply the point 0 and, by definition of V, the sub-level sets of V are contained in some balls and are therefore compact.

Liapounov's theorem applies on the domain Ω and shows that 0 is an asymptotically stable equilibrium point.

It should be noted that the above inequality, with the help of the properties of V (I.23), can be precised by the following

$$(f \cdot \nabla V)(x) \leq -\frac{1}{2}\|x\|^2 \leq -\frac{1}{2\alpha_{min}}V(x).$$

This implies that V will verify

$$\frac{d}{dt}V(\varphi(t,x)) \leq -\frac{1}{2\alpha_{min}}V(\varphi(t,x)),$$

along the trajectories. This is a differential inequality and the comparison lemma on differential inequalities I.22 allows us to deduce

$$V(\varphi(t,x)) \leq V(x)e^{-\frac{1}{2\alpha_{min}}t}, \ \ \forall t \geq 0,$$

and therefore we obtain the exponential convergence of the trajectories thanks to (I.23).

■

Let us give two examples.

1. *Gradient flows:*

   The first case of a differential system for which the existence of a Liapounov function is clear is in the case of **gradient flows**. Given a function $V : \Omega \to \mathbb{R}$ of class $\mathcal{C}^2$, we can consider the following differential system

   $$x' = -\nabla V(x).$$

   In other words, the vector field that defines the problem is exactly the opposite of the gradient of a *potential* V. These types of systems come up a lot in physics.

   We can easily verify that if V is coercive with a finite number of critical points, then it is a Liapounov function and the stable equilibriums of the system are the local minimums of V.

2. *Damped pendulum system:*

   Consider the equation

   $$\theta'' + \mu\theta' + \sin\theta = 0,$$

   that models the evolution of a damped pendulum (θ being the angle between the pendulum and the vertical line), where the coefficient $\mu > 0$ measures the damping related to, for instance, the friction exerted on the system at the attachment point of the pendulum.

   We can show that all the equilibrium points of the form $(\theta, \theta') = (2k\pi, 0)$ are asymptotically stable by doing a spectral analysis of the Jacobian.

   We can be more precise by considering a Liapounov function, which is non other than the energy functional of the system, defined by

   $$V(\theta, \theta') = \frac{1}{2}|\theta'|^2 + (1 - \cos\theta).$$

   Its evolution over time along the solutions verifies

   $$\frac{d}{dt}V(\theta, \theta') = -\mu|\theta'|^2,$$

   which proves the monotony of V relative to the vector field of the dampened pendulum.

   Furthermore, the relative critical points of V are exactly the points of the phase plane verifying $\theta' = 0$, that is points of the form $(\theta, 0)$. For any $\alpha \in \mathbb{R}$ the intersection of this set with the level curve $\{V = \alpha\}$ is a discrete set (it is the set $(\theta, 0)$ where $\cos(\theta) = 1 - \alpha$) so the structure properties of the critical points of V relative to the vector field is

verified. At last, the compactness property of the connected components of sub-level sets $\{V \leq \alpha\}$ is also verified (at least for α smaller than 2).

From this, we can deduce that equilibrium points of the form $X_k = (2k\pi, 0)$, that verifies $V(X_k) = 0$ are asymptotically stable and that for any given initial data verifying $V(\theta_0, \theta_0') < 2$, the solution of the system will converge to the unique equilibrium belonging to the same connected components of sub-level sets of V. The other equilibrium points $Y_k = ((2k+1)\pi, 0)$ verify $V(Y_k) = 2$ and therefore do not fit in the previous study and in fact, we observe that the Jacobian of the field at these points is

$$A = \begin{pmatrix} 0 & 1 \\ 1 & -\mu \end{pmatrix}.$$

We can see that it has two real eigenvalues of opposite signs, and in particular, one of them is positive which shows that this equilibrium is not stable.

V.2.e Barriers

Let us conclude this discussion by presenting a tool that might be useful to describe a little more precisely the behavior of solutions of **scalar** differential equations. It is a result that allows the comparisons of functions that verify certain differential inequalities. It can be seen as a generalization of Proposition I.22 that treats the linear case.

Theorem I.69

Let $f : I \times \mathbb{R} \to \mathbb{R}$ be a continuous vector field that is locally lipschitz with respect to the state variable. Suppose there are two differentiable functions $\alpha, \beta : J \to \mathbb{R}$, where $J \subset I$, that verifies

$$\beta'(t) \geq f(t, \beta(t)), \quad \forall t \in J,$$

$$\alpha'(t) \leq f(t, \alpha(t)), \quad \forall t \in J.$$

For $t_0 \in J$, we have the following properties

$$\alpha(t_0) \leq \beta(t_0) \Longrightarrow \alpha(t) \leq \beta(t), \quad \forall t \in J, t \geq t_0,$$
$$\alpha(t_0) \geq \beta(t_0) \Longrightarrow \alpha(t) \geq \beta(t), \quad \forall t \in J, t \leq t_0.$$

Proof :

Let us assume $\alpha(t_0) \leq \beta(t_0)$, the other case being similar.

- If one of the inequalities in the hypotheses is strict, then the proof becomes easier. Indeed, we have in this case, for all $t \in J$

  $$\beta'(t) - \alpha'(t) > f(t, \beta(t)) - f(t, \alpha(t)),$$

  so at each point t where α and β coincide, we know that $(\beta - \alpha)'(t) > 0$.

  Let us start by proving the existence of an $\varepsilon > 0$ such that $\alpha < \beta$ on $(t_0, t_0 + \varepsilon)$:

  - If $\beta(t_0) > \alpha(t_0)$, then the expected result can be obtained by using the continuity of these two functions.
  - If $\beta(t_0) = \alpha(t_0)$, we can use the above property (that was deduced from the hypotheses) to obtain that $\beta - \alpha$ is increasing in a neighborhood of t_0 and obtain the expected result.

  Now, let us introduce the following set

  $$T = \{t \in (t_0, +\infty) \cap J, \ \ \beta(t) = \alpha(t)\}.$$

  Let us prove by contradiction that this set is empty which will prove that $\alpha - \beta$ cannot change signs therefore must stay positive on $[t_0, +\infty)$. So, suppose that T is non empty. In this case, since we know that this set is bounded by ε from below, we would have

  $$t^* := \inf T \in (\varepsilon, +\infty).$$

  By continuity, we know that $t^* \in T$ so by using the preliminary remark we have $(\beta - \alpha)'(t^*) > 0$, therefore $\beta - \alpha$ is negative on an interval of the form $(t^* - \delta, t^*)$. By the intermediate value theorem, we can deduce the existence of a root $\bar{t}$ of $\beta - \alpha$ in (ε, t^*), which is of course an element of T. This contradicts the minimality of t^*.

- The general case is more delicate and we need Proposition I.10.

  For this, we introduce the following function[10]

  $$z(t) := (\alpha(t) - \beta(t)) + |\alpha(t) - \beta(t)|.$$

  This function is non negative and verifies the property : $z(t) = 0$ if and only if $\alpha(t) \leq \beta(t)$.

  Let us write the equation that z verifies (thanks to Proposition I.10)

  $$z(t) = z(t_0) + \int_{t_0}^{t} (1 + \mathrm{sgn}(\alpha(s) - \beta(s)))(\alpha - \beta)'(s)\, ds.$$

  Since $1 + \mathrm{sgn}$ is non-negative, since $\alpha' \leq f(t, \alpha(t))$ and since $\beta' \geq f(t, \beta(t))$ we obtain

  $$z(t) \leq z(t_0) + \int_{t_0}^{t} (1 + \mathrm{sgn}(\alpha(s) - \beta(s)))(f(s, \alpha(s)) - f(s, \beta(s)))\, ds.$$

  Since f is lipschitz, we deduce

  $$z(t) \leq z(t_0) + \int_{t_0}^{t} L(s)(1 + \mathrm{sgn}(\alpha(s) - \beta(s)))|\alpha(s) - \beta(s)|\, ds.$$

  On the other hand, we have

  $$(1 + \mathrm{sgn}(\alpha(s) - \beta(s)))|\alpha(s) - \beta(s)| = z(s),$$

  which gives us the following inequality

  $$z(t) \leq z(t_0) + \int_{t_0}^{t} L(s) z(s)\, ds.$$

  By using Gronwall's lemma, we obtain

  $$z(t) \leq z(t_0) e^{\int_{t_0}^{t} L(s)\, ds}.$$

  So, if $\alpha(t_0) \leq \beta(t_0)$ then $z(t_0) = 0$ and the above inequality shows that $z(t) = 0$ for all t, in other words $\alpha(t) \leq \beta(t)$, which is what we wanted to prove.

■

VI A detailed study of an example : An epidemic propagation model

Preliminary remarks. The numerical results that are presented here are obtained from a Python program, whose notebook is available at the following address

https://gist.github.com/FranckBoyer/df2f2f839aef35f60e68aa7878e69a33

Presentation of the model. Here, we consider the following system of ordinary differential equations

$$\begin{cases} S'(t) = -\beta I(t) S(t) + \gamma R(t), \\ R'(t) = \nu I(t) - \gamma R(t), \\ I'(t) = -\nu I(t) + \beta I(t) S(t), \end{cases} \tag{I.24}$$

with initial conditions

$$\begin{cases} S(0) = S_0, \\ R(0) = R_0, \\ I(0) = I_0. \end{cases} \tag{I.25}$$

In this system, β, γ and ν are three positive parameters. This system describes the evolution over time (the time unit being *a day*) of a population subject to a contagious disease. The following are the hypotheses and notations that we will be using:

- The total population stays constant over time : in this model, we will not take into account the effects of birth and death rates, we assume that they compensate each other exactly.

[10] this quantity is nothing but the double of the positive part of the difference $\alpha - \beta$

- $I(t) \in [0,1]$ represents the proportion over time of individuals that are *Infected* by the disease. These individuals are contagious and are the vector for the spread of the disease.
- $R(t) \in [0,1]$ represents the proportion at time t of individuals that are *Resistant*, meaning, who have contracted the disease, recovered from it and are now immune to it.
- $S(t) \in [0,1]$ represents the proportion at time t of individuals that are *Susceptible* to contracting the disease. This includes all individuals who are not sick and are not immune.
- The following is what each coefficient represents:
  - $1/\nu$ is the average time of recovery of an infected patient.
  - β is the contamination factor, meaning the probability that, if an *Infected* individual is in contact with a *Susceptible* individual for 1 unit of time, then the latter will contract the disease.
  - $1/\gamma$ is the average time during which an individual who has contracted the disease will be immune to it.

VI.1 Existence and uniqueness of a global solution in positive time

Theorem I.70

For any given initial data $S_0, R_0, I_0 \geq 0$, there exists a unique solution of the system (I.24)- (I.25) defined on $[0, +\infty)$. Moreover, this solution verifies

$$S(t) \geq 0, R(t) \geq 0, I(t) \geq 0, \quad \forall t \geq 0,$$

and

$$S(t) + R(t) + I(t) = S_0 + R_0 + I_0, \quad \forall t \geq 0.$$

Proof :

- The (autonomous) vector field f that defines the equation is polynomial in all its variables, so it is $\mathcal{C}^1$ and therefore locally lipschitz. So, the local Cauchy-Lipschitz theorem applies and we get the existence and uniqueness of a maximal solution $(J, (S, R, I))$ of the Cauchy problem.
- If we add up the three equations, we immediately see that

  $$S'(t) + R'(t) + I'(t) = 0, \ \ \forall t \in J,$$

  and therefore the sum of the three unknowns is constant over time.
- Let us now show that the solutions stay positive on J.
  - We will start by studying the case of the variable I:
    We first observe that it is solution to the following linear and autonomous differential equation

    $$I'(t) = (-\nu + \beta S(t))I(t),$$

    so

    $$I(t) = I_0 e^{-\int_0^t (-\nu + \beta S(s))\, ds}.$$

    We can see that the solution is indeed non-negative and that if $I_0 > 0$, then we have $I(t) > 0$ for all $t \in J$.
    If $I = 0$, the system is reduced to

    $$S' = \gamma R, R' = -\gamma R,$$

    whence the explicit resolution

    $$R(t) = R_0 e^{-\gamma t},$$

    $$S(t) = S_0 + R_0 \left(1 - e^{-\gamma t}\right),$$

    which satisfies the stated property.
    From now on, we will suppose that $I(t) > 0$ for all $t \geq 0$.

– Let us now take the equation on R and use that I is non negative to deduce that

$$R'(t) \geq -\gamma R(t), \ \ \forall t \in J,$$

and therefore by Proposition I.22, we deduce that for $t \in J, t \geq 0$ we have the inequality

$$R(t) \geq R_0 e^{-\gamma t},$$

which proves, in particular, that R is non-negative on $[0, +\infty) \cap J$.

– And finally, let us take the equation on S and by using the sign of R, we obtain

$$S'(t) \geq -\beta I(t) S(t),$$

and therefore, once again by Proposition I.22, we get

$$S(t) \geq S_0 e^{-\beta \int_0^t I(s)\, ds},$$

so S is non negative on $[0, +\infty) \cap J$.

– So we have shown that R, S and I are non negative on $[0, +\infty) \cap J$ and that, moreover, their sum is constant. It follows that R, S and I are bounded on $[0, +\infty) \cap J$. The finite time blow-up theorem allows us to conclude to the existence of global solutions in positive time, which means $[0, +\infty) \subset J$.

■

VI.2 Equilibrium state

We will now assume that $S_0 + R_0 + I_0 = 1$ because they are supposed to represent proportions[11].

Since $S + R + I = 1$ over time, we can in fact study the following system of dimension 2

$$\begin{cases} S' = -\beta IS + \gamma(1 - S - I), \\ I' = -\nu I + \beta IS. \end{cases} \tag{I.26}$$

- The first equilibrium point that we find is given by

$$I = 0, S = 1.$$

The Jacobian matrix of the vector field at this point is given by

$$A = \begin{pmatrix} -\gamma & -\beta - \gamma \\ 0 & \beta - \nu \end{pmatrix}.$$

This matrix has one negative eigenvalue and another whose sign depends on the parameters:

– If $\beta > \nu$, then A has a positive eigenvalue therefore the system is unstable (via Hartman Grobman), in this case an epidemic can develop from a very small number of infected individuals.

– If $\beta < \nu$, then the two e.v. of A are negative and therefore the equilibrium in question is asymptotically stable. This situation corresponds to the impossibility of developing an epidemic.

– In the case $\beta = \nu$, 0 is an eigenvalue so we cannot *a priori* conclude. It is worth studying the particular system thus obtained, which we write in the new variables $\tilde{S} = 1 - S$ and $\tilde{I} = I$. It should be noted that this choice of new variables implies that $\tilde{I} \geq 0$ and $\tilde{S} \geq 0$,

$$\begin{cases} \tilde{S}' = (\beta + \gamma)\tilde{I} - \beta \tilde{I}\tilde{S} - \gamma \tilde{S}, \\ \tilde{I}' = -\beta \tilde{I}\tilde{S}. \end{cases}$$

Consider the following function

$$V(\tilde{S}, \tilde{I}) := \beta \tilde{S}^2 + 2(\beta + \gamma)\tilde{I}.$$

This function verifies

$$\frac{d}{dt} V(\tilde{S}, \tilde{I}) = -2\beta^2 \tilde{I}\tilde{S}^2 - 2\beta\gamma \tilde{S}^2 \leq 0.$$

[11] However, the following analysis would work in a similar way if we replace the value 1 by any other non zero value.

This function is therefore non increasing along the trajectories and the set C of critical points of V relative to the considered vector field is of the form

$$C := \{(0, \tilde{I}),\ \ \tilde{I} \geq 0\}.$$

On this set, we have $V(0, \tilde{I}) = 2(\beta + \gamma)\tilde{I}$ so the intersection of C with the level lines of V are reduced to a point.

Therefore, the hypotheses of the Liapounov's theorem are verified and all the trajectories of the reduced system converge to the unique equilibrium $(0, 0)$, which proves the global asymptotic stability of the unique equilibrium point of the initial system.

Remark I.71 (A proof without Liapounov, or almost ...)

One can prove the **global** *asymptotic stability of the equilibrium $I^* = 0$, $S^* = 1$ under the hypothesis that $\beta \leq \nu$ by doing a direct analysis without using a general Liapounov type theorem.*
Reconsider the variables $\tilde{S}, \tilde{I}$ (without assuming $\beta = \nu$)

$$\begin{cases} \tilde{S}' = (\beta + \gamma)\tilde{I} - \beta\tilde{I}\tilde{S} - \gamma\tilde{S}, \\ \tilde{I}' = -\beta\tilde{I}\tilde{S} - (\nu - \beta)\tilde{I}. \end{cases}$$

and recalculate the derivative of $V(\tilde{S}, \tilde{I})$ along the trajectory. This gives us

$$\frac{d}{dt}V(\tilde{S}, \tilde{I}) = -2\beta^2\tilde{I}\tilde{S}^2 - 2\beta\gamma\tilde{S}^2 - 2(\beta + \gamma)(\nu - \beta)\tilde{I}. \tag{I.27}$$

- *If $\nu > \beta$, we obtain the following inequality for a certain $\delta > 0$*

  $$\frac{d}{dt}V(\tilde{S}, \tilde{I}) \leq -\delta V(\tilde{S}, \tilde{I}),$$

  and the Proposition I.22 gives us

  $$V(\tilde{S}(t), \tilde{I}(t)) \leq V_0 e^{-\delta t}, \quad \forall t \geq 0.$$

  In the end, we obtain $\lim_{t\to+\infty} V(\tilde{S}(t), \tilde{I}(t)) = 0$ which shows that $\tilde{I}$ and $\tilde{S}$ go to 0 at $+\infty$.

- *If $\nu = \beta$, we start by noting that $t \mapsto V(\tilde{S}(t), \tilde{I}(t))$ is non increasing and non negative, therefore it converges towards a limit that will be denoted $V_\infty \geq 0$. It follows, by integrating (I.27) between the bounds 0 and $+\infty$, that*

  $$\int_0^{+\infty} \left(2\beta\gamma\tilde{S}^2 + 2\beta^2\tilde{I}\tilde{S}^2\right) dt = V(S_0, I_0) - V_\infty < +\infty.$$

  Moreover, the equation on $\tilde{I}$ immediately gives us that $t \mapsto \tilde{I}$ is also non increasing and non negative, and therefore convergent towards a limit that will be denoted I_∞. Since $V(\tilde{S}, \tilde{I})$ also has a limit, we deduce that $\tilde{S}^2$ also has a limit as $t \to +\infty$ and since $\tilde{S}^2$ is integrable at $+\infty$, this limit can only be equal to 0. At last, we have proven that $\tilde{S} \to 0$ when $t \to \infty$.

  From the equation on $\tilde{S}$, we have

  $$\lim_{t\to+\infty} \tilde{S}'(t) = (\beta + \gamma)I_\infty.$$

  According to Lemma I.58, since this limit exists and that $\tilde{S}$ also has a limit, we must necessarily have $I_\infty = 0$. We have thus shown that the solution to the reduced system verifies $(\tilde{S}, \tilde{I}) \to (0, 0)$, which is what we wanted to prove.

  Note that, here, we did not prove the exponential convergence of the solution. One can be convinced that this property is actually false in this case. See the illustrations below.

- If $\beta \neq \nu$, there is another non trivial equilibrium

  $$S^* = \frac{\nu}{\beta},$$

  $$I^* = \frac{\gamma(\beta - \nu)}{\beta(\nu + \gamma)}.$$

they are positive (the only acceptable cases for the model) if and only if $\nu < \beta$ (if $\nu = \beta$, we find the previous case), which is exactly the case where the trivial equilibrium is unstable.

From now on, suppose that $\nu < \beta$. The Jacobian matrix of the system at this equilibrium point is given by

$$\begin{aligned} A &= \begin{pmatrix} -\beta I^* - \gamma & -\beta S^* - \gamma \\ \beta I^* & -\nu + \beta S^* \end{pmatrix} \\ &= \begin{pmatrix} \gamma\dfrac{\nu-\beta}{\nu+\gamma} - \gamma & -\nu - \gamma \\ -\gamma\dfrac{\nu-\beta}{\nu+\gamma} & 0 \end{pmatrix} \\ &= -\begin{pmatrix} \gamma\dfrac{\gamma+\beta}{\nu+\gamma} & \nu+\gamma \\ \gamma\dfrac{\nu-\beta}{\nu+\gamma} & 0 \end{pmatrix}. \end{aligned}$$

The determinant of A is given by

$$\det A = \gamma(\beta - \nu),$$

and it is positive from the hypothesis on the parameters.

Furthermore, the trace of A is given by $\operatorname{Tr} A = -\frac{\gamma(\gamma+\beta)}{\nu+\gamma}$ which is a negative value.

So, we can deduce that the matrix A has two eigenvalues whose real parts are negative, we are therefore dealing with an asymptotically stable equilibrium point.

VI.3 Some examples of trajectories

Here, we show some examples of trajectories of the system. The figures presented here have been obtained thanks to a jupyter/python notebook available at the address `https://git.io/fhH9L`. In this notebook, you can vary the parameters interactively to visualize the different possible behaviors of the solutions.

The left side of the figures shows the evolution of the three quantities S, R and I over time, and the right side shows the trajectory in the phase plane (I, S). The blue dot is the initial point and the large red dots are the equilibriums of the system.

VI.3.a The extinction of an epidemic of moderate intensity

Suppose that :

- Half of the population is infected at the initial time : $I_0 = S_0 = 1/2$, $R_0 = 0$.
- The average duration of an infection is 2 days : $\nu = 1/2$
- The probability of contagion is $\beta = 1/30$
- Immunity disappears after about ten days : $\gamma = 1/10$

In these conditions, we obtain the solution of the system shown in Figure I.11. We observe that the epidemic does not spread (this corresponds to the asymptotic stability of the equilibrium $(S = 1, I = 0)$ of the system) and that it is eradicated after a few weeks.

VI.3.b An outbreak of a flu epidemic in New York in the 60's

This highly virulent flu started with 10 infected individuals on a total population size of 8 million inhabitants, which gives $I_0 = 1.25 \times 10^{-6}$.

The average duration of infection was observed to be 3 days and the probability of contagion was $1/2$. We assume that the acquired immunity is permanent, i.e. $\gamma = 0$ (this changes a little the study of the system compared to what we have done above : in particular all states of the form $(I = 0, S = S^*)$ are equilibriums of the system).

We observe that it took almost 50 days for the epidemic to develop and that it lasted 150 days in total. Since the obtained immunity is permanent, the limit at infinity of R in this example gives us the total proportion of the population that have been infected at some point or another by the virus. We therefore observe that after about 6 months, almost 60% of New York's population has been infected.

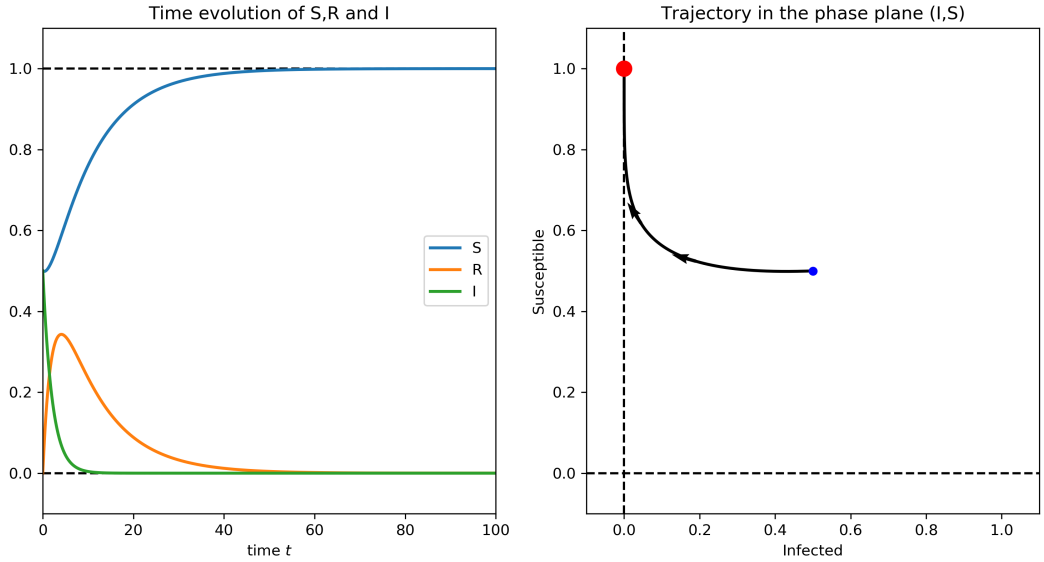


Figure I.11: An epidemic of moderate force

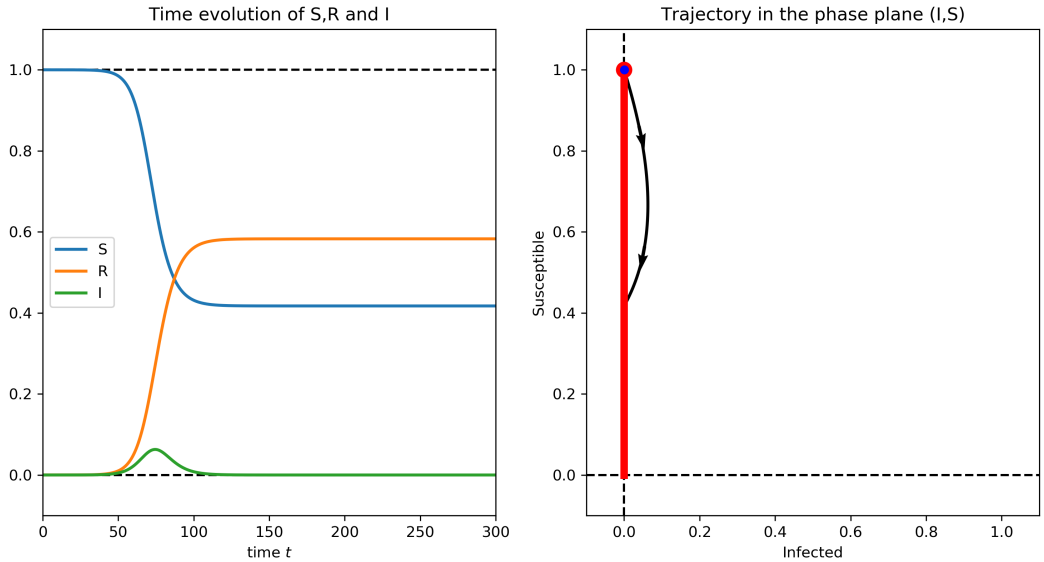


Figure I.12: A virulent epidemic in New York

VI.3.c And what if the virus mutates ?

Let us reconsider the previous example by assuming that this time, the virus is capable of slowly mutating. We take this into account in the model by assuming that the acquired immunity is not permanent. We consider the same parameters as in the previous case and assume that this time we have $\gamma = 1/365$ (we estimate that it takes one year for the virus to mutate). The obtained results are illustrated in the figure I.13.

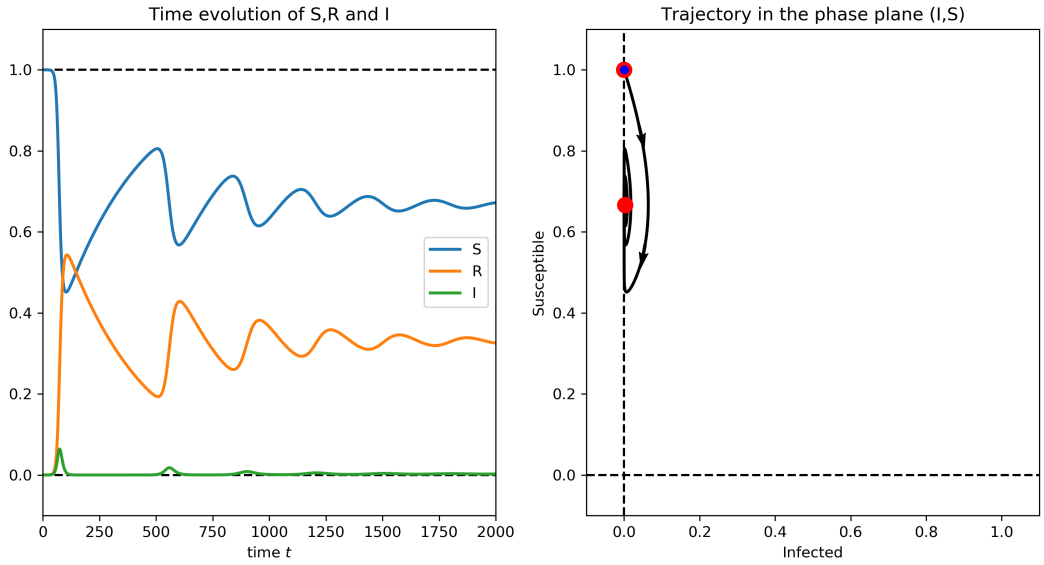


Figure I.13: A virulent epidemic in New York with mutation

The beginning of the dynamics is essentially the same, but we then see that the population is less and less immune and thus allows the epidemic to have a new peak (however a lower peak than the previous one) and so on for several years. At last, the solution converges to the expected equilibrium (for which the proportion of infected individuals is small but not null, here approximately 0.25%).

VI.3.d The case $\beta = \nu$

In this case, we have seen that the global asymptotic stability of the equilibrium $(S^*, I^*) = (1, 0)$ is still true but is not exponential. It is what we observe in the figures I.14 and I.15 where the speed of return to equilibrium is of the form

$$I(t) \sim_{t\to+\infty} \frac{C}{t}, \quad 1 - S(t) \sim_{t\to+\infty} \frac{C'}{t}.$$

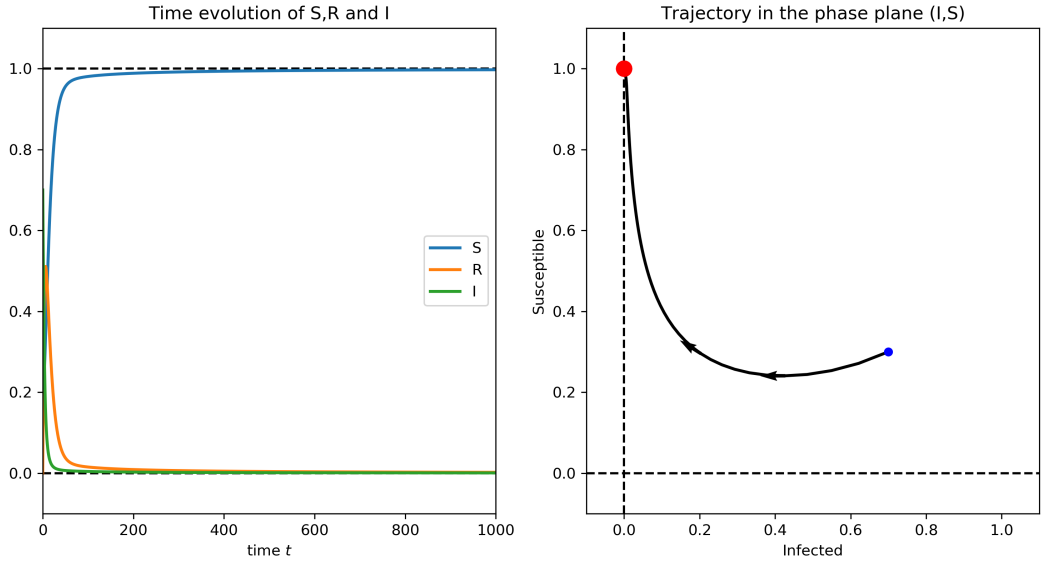


Figure I.14: An epidemic in the case $\beta = \nu$

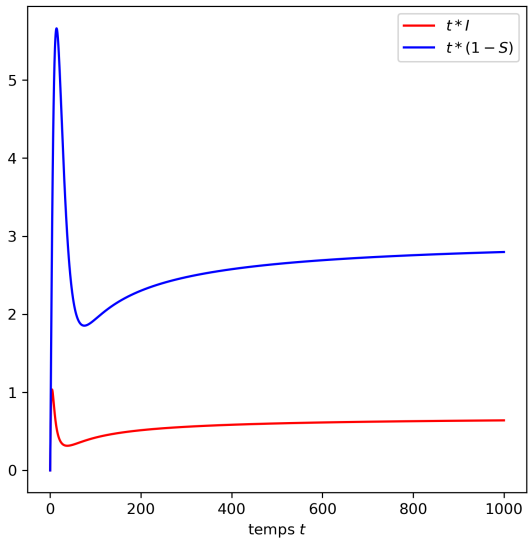


Figure I.15: An epidemic in the case $\beta = \nu$. Speed of return to equilibrium

What should be retained from this chapter

In priority

- The main definitions (vector fields and their regularity, Cauchy problems, local solutions, maximal solutions, global solutions, flow, resolvent, stability, asymptotic stability ...)
- The statement and proof of Gronwall's lemma
- The resolution of linear and autonomous differential equations. Duhamel's formula.
- To be capable of drawing (at least the shape of) the phase portrait of an autonomous and linear 2×2 system.
- The statements of the local and global Cauchy-Lipschitz theorems, the finite time blow-up theorem and the theorem I.49.

  Knowing how to use them properly to prove qualitative properties of solutions to ODEs (positivity, etc ...)
- Properties of the flow.
- The statement of the stability/asymptotic stability in the linear autonomous case.

  To know the proof in the case where all the eigenvalues have a negative real part.
- The statement of the spectral criterion for asymptotic stability for a non linear and autonomous ODE.

  Knowing how to use it and understand the cases where we cannot conclude.

To go even further

- Work on the proof of the main theorems. Observe that many of them use Gronwall's lemma in a crucial way.
- The statement of the Hartmann-Grobman theorem.
- Liapounov's theory and barriers theorem.

Chapter II

Transport equations

In this chapter and the following one, we will be studying partial differential equations. We will be using interchangeably the following notations for partial derivatives

$$\partial_t, \frac{\partial}{\partial t}, \partial_x, \frac{\partial}{\partial x}, \partial_{x_i}, \partial_i, ...$$

depending on the context in order to alleviate the formulas as much as possible.

Moreover, in this chapter we will be using some classical differential operators (essentially the divergence and the gradient) whose definitions and basic properties are recalled at the beginning of the appendix A. We will also be using (not a lot though) the formalism of the theory of distributions that is also recalled in the same appendix, and one can consult it in case of need.

Usually, while studying PDEs, unless explicitly mentioned otherwise, the operators ∇, div (and also Δ that we will encounter in the next chapter) only act on the space variables and not on the time variable. The derivatives with respect to time are always explicitly given by the operators ∂_t, ∂_t^2, ... So, for example, if $(t,x) \in \mathbb{R} \times \mathbb{R}^d \mapsto f(t,x) \in \mathbb{R}$ is a regular function, we will denote

$$\nabla f(t,x) = \begin{pmatrix} \frac{\partial f}{\partial_{x_1}}(t,x) \\ \vdots \\ \frac{\partial f}{\partial_{x_d}}(t,x) \end{pmatrix}.$$

I Transport model in 1D

I.1 Road traffic

We are interested in the mathematical modeling of road traffic. It is a vast subject that we will barely touch upon in order to motivate the study of "transport" type equations. One of the challenges of these models (in their more sophisticated version) is, for instance, the understanding of the formation of traffic jams as we can see in the following videos that show an experiment and the corresponding numerical simulation

Experiment : https://www.youtube.com/watch?v=7wm-pZp_mi0
Simulation : https://www.youtube.com/watch?v=Q78Kb4uLAdA

- **Modeling hypotheses :** We model a straight national road, with only one lane of traffic. The road is assumed to be infinite (in other words, we will not bother with boundary problems) and that there are no overruns. Moreover, we will not consider cars that enter or exit the road.

  We denote $\rho(t,x)$ the density of vehicles (average number of vehicles per unit length in a neighborhood of the position x at time t) and $v(t,x)$ the average velocity of the vehicles at time t and at point x.

- **Trajectory of one particular vehicle :** Suppose that we know ρ and v at each point and at each time.

  Consider a vehicle that is located at position x_0 and at time t_0. We denote $X(t,t_0,x_0)$ its position at another time t. By definition of the velocity field, we have the following relation

  $$\partial_t X(t,t_0,x_0) = v(t, X(t,t_0,x_0)), \quad \forall t \geq t_0, \tag{II.1}$$

  and of course

  $$X(t_0,t_0,x_0) = x_0.$$

  We can see that, if we know v, the trajectory of a vehicle is given by the solution of the differential equation (II.1) associated to the initial data (t_0,x_0). In other words, X is the associated flow of the vector field v.

Throughout this chapter, we will assume that v is sufficiently regular so that we can apply the global Cauchy-Lipschitz theorem and thus ensure that the trajectories X(=the flow) are defined globally without having to worry about their domain of definition.

- **The law of conservation of vehicles :**

Proposition II.1

If they are regular, the functions ρ and v are related by the following partial differential equation

$$\partial_t \rho + \partial_x(\rho v) = 0. \tag{II.2}$$

Proof :

We will prove the result in several ways (they are of course totally equivalent but it is rather *the idea behind the computations* that is different).

– *Physicists' proof.*
We will evaluate the evolution of the number of vehicles between the points a and b and between the times t and $t+\delta t$, where δt is assumed to be small. This assessment is therefore written *in first order* which means that we can neglect all terms that are small in front of δt.

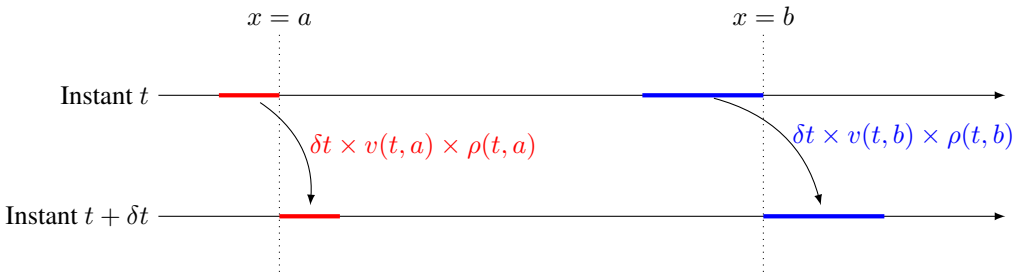


Figure II.1: Physicists' proof

$$\int_a^b \rho(t+\delta t, x)\, dx \approx \int_a^b \rho(t,x)\, dx + \delta t\, v(t,a)\rho(t,a) - \delta t\, v(t,b)\rho(t,b).$$

On the other hand, we write that $\rho(t+\delta t, x) \approx \rho(t,x) + \delta t \partial_t \rho(t,x)$ and observe that the contribution of the last two terms can also be written as $-\delta t \int_a^b \partial_x(\rho v)(t,x)\, dx$. So we have obtained

$$\delta t \int_a^b \partial_t \rho(t,x)\, dx \approx -\delta t \int_a^b \partial_x(\rho v)(t,x)\, dx,$$

therefore, by simplifying the δt's, we get

$$\int_a^b \left[\partial_t \rho + \partial_x(\rho v)\right] dx \approx 0.$$

Since this "equality" has to be true for all t, all a, and all b, we can differentiate with respect to b and conclude that

$$\partial_t \rho + \partial_x(\rho v) = 0,$$

which gives us the desired result.

– *Mathematicians' proof.*
We first write the following equality

$$\int_a^b \rho(t,x)\, dx = \int_{X(t+s,t,a)}^{X(t+s,t,b)} \rho(t+s,x)\, dx, \quad \forall s \in \mathbb{R},$$

which describes the fact that the number of vehicles between two given vehicles (those that are at a and b at time t) remains constant over time. So we use the Lemma II.2 found below to differentiate this quantity with respect to s and obtain (by evaluating the result at $s = 0$)

$$0 = \int_a^b \partial_t \rho(t,x)\, dx + \partial_t X(t,t,b)\rho(t, X(t,t,b)) - \partial_t X(t,t,a)\rho(t, X(t,t,a)).$$

By using the definition of trajectories (II.1), we obtain that

$$0 = \int_a^b \partial_t \rho(t,x)\, dx + v(t,b)\rho(t,b) - v(t,a)\rho(t,a) = \int_a^b \partial_t \rho + \partial_x(\rho v)\, dx.$$

The end of the proof is identical to the previous case.

■

In the above proof, we have used the following lemma.

Lemma II.2

Let $f : (t,x) \in \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^1$ function and $\alpha, \beta : \mathbb{R} \to \mathbb{R}$ be two other $\mathcal{C}^1$ functions. We have the following differentiation formula

$$\frac{d}{dt}\int_{\alpha(t)}^{\beta(t)} f(t,x)\, dx = \int_{\alpha(t)}^{\beta(t)} \partial_t f(t,x)\, dx + \beta'(t) f(t,\beta(t)) - \alpha'(t) f(t,\alpha(t)).$$

Proof :

Define

$$\Phi : (a,b,t) \in \mathbb{R}^3 \mapsto \int_a^b f(t,x)\, dx.$$

The usual theorems from calculus show[1] that Φ has the following partial derivatives :

$$\partial_a \Phi(a,b,t) = -f(t,a),$$
$$\partial_b \Phi(a,b,t) = f(t,b),$$
$$\partial_t \Phi(a,b,t) = \int_a^b \partial_t f(t,x)\, dx.$$

These partial derivatives are continuous functions of the three variables (a,b,t) therefore the function Φ is $\mathcal{C}^1$. The result of the lemma just comes from the differentiation of the function $t \mapsto \Phi(\alpha(t),\beta(t),t)$. ■

One should note that the conservation law (II.2) is valid in all generality under the assumptions made at the beginning.

- **Modeling the behavior of vehicles :** To *close*[2] the system, one should find a relation between the density ρ and the speed v.

  1. First model : all drivers go at the same speed (the maximum speed limit) regardless of traffic conditions :

     $$v(t,x) = V_{max} = cst.$$

     In these conditions, v is of course no longer unknown and thus we get the equation

     $$0 = \partial_t \rho + \partial_x(\rho V_{max}) = \partial_t \rho + V_{max}\partial_x \rho, \tag{II.3}$$

     which is called the transport equation with constant speed V_{max}.
  2. Second model : The maximum speed limit varies over the length of the road according to a given law $V_{max}(x)$, so the equation becomes

     $$0 = \underbrace{\partial_t \rho + \partial_x(\rho V_{max}(x))}_{\text{conservative form}} = \underbrace{\partial_t \rho + V_{max}(x)\partial_x \rho + V'_{max}(x)\rho}_{\text{non conservative form}}. \tag{II.4}$$

  3. Third model : the drivers adapt their speed to the traffic. The denser the traffic is, the lower the speed is, which gives, for example, a law such as

     $$v = V_{max}(1-\rho),$$

     and therefore the **non linear** partial differential equation becomes

     $$0 = \partial_t \rho + V_{max}\partial_x(\rho(1-\rho)). \tag{II.5}$$

     We can also write it in terms of the speed

     $$0 = \partial_t v + \partial_x(v(v - V_{max})).$$

Depending on the chosen model, the vehicle density will therefore verify one of the three equations (II.3), (II.4) or (II.5).

[1] Make sure you know how to do it !

[2] That is, obtain as many equations as there are unknown variables

I.2 Simplified gas dynamics

Let us study a simple context of fluid mechanics: the evolution of a perfect gas in one dimension (in a straight pipe with a small section for example). The variables which describe the fluid in these conditions are: the density of the gas $\rho(t,x)$ (i.e. the average density of the gas at time t and in a neighborhood of x), the velocity field $v(t,x)$ (i.e. the **average** velocity of the gas particles at time t and at position x) and the pressure $p(t,x)$. We neglect the effects of gravity and we will assume that these quantities are sufficiently regular to justify the calculations that will follow.

The first equation that governs the evolution of the system is the equation for the conservation of mass which is obtained exactly as for the equation for the conservation of vehicles in the road traffic model

$$\partial_t \rho + \partial_x(\rho v) = 0.$$

To obtain the second equation, the behavior of the gas particles must be modeled. This behavior is given by Newton's law and it tells us that the rate of change of any mechanical system's momentum is equal to the sum of external forces applied to it. The momentum at time t_0 of the portion of the gas between the abscissas a and b is given by

$$\int_a^b \rho(t_0,x)v(t_0,x)\,dx.$$

In the course of evolution, the portion of the gas in question is between the abscissas $X(t,t_0,a)$ and $X(t,t_0,b)$ and the total momentum at time t is therefore given by

$$\int_{X(t,t_0,a)}^{X(t,t_0,b)} \rho(t,x)v(t,x)\,dx.$$

This portion of gas is only subject pressure forces (exerted by the gas particles outside the portion in question) so in this context Newton's law can be written as

$$\frac{d}{dt}\left(\int_{X(t,t_0,a)}^{X(t,t_0,b)} \rho(t,x)v(t,x)\,dx\right) = \text{The sum of forces} = -p(t,X(t,t_0,b)) + p(t,X(t,t_0,a)).$$

By using once again Lemma II.2 and the definition of characteristics, we can compute the derivative of the quantity on the left hand side of the equation

$$\begin{aligned}
\frac{d}{dt}\left(\int_{X(t,t_0,a)}^{X(t,t_0,b)} \rho(t,x)v(t,x)\,dx\right) =& (\rho v)(t,X(t,t_0,b))\frac{d}{dt}X(t,t_0,b) - (\rho v)(t,X(t,t_0,a))\frac{d}{dt}X(t,t_0,a)\\
&+ \int_{X(t,t_0,a)}^{X(t,t_0,b)} \partial_t(\rho v)(t,x)\,dx\\
=& (\rho v)(t,X(t,t_0,b))v(t,X(t,t_0,b)) - (\rho v)(t,X(t,t_0,a))v(t,X(t,t_0,a))\\
&+ \int_{X(t,t_0,a)}^{X(t,t_0,b)} \partial_t(\rho v)(t,x)\,dx\\
=& \int_{X(t,t_0,a)}^{X(t,t_0,b)} \big[\partial_t(\rho v) + \partial_x(\rho v^2)\big](t,x)\,dx.
\end{aligned}$$

For the right hand side, we can write it as an integral of the derivative of the pressure and we get

$$\int_{X(t,t_0,a)}^{X(t,t_0,b)} \big[\partial_t(\rho v) + \partial_x(\rho v^2) + \partial_x p\big](t,x)\,dx = 0.$$

For $t = t_0$, we get

$$\int_a^b \big[\partial_t(\rho v) + \partial_x(\rho v^2) + \partial_x p\big](t,x)\,dx = 0.$$

Since this is true for all real numbers $a < b$, and since the function inside the integral is assumed to be continuous with respect to x, we can differentiate this equality, for any fixed t, with respect to b and obtain the following partial differential equation

$$\partial_t(\rho v) + \partial_x(\rho v^2) + \partial_x p = 0, \forall t > 0, \forall x \in \mathbb{R}.$$

This last argument is specific to a 1 dimensional space. More generally, we can obtain the same conclusion by using the Du Bois-Reymond lemma (or more exactly its variant Lemma A.5).

The obtained partial differential equation is often written in the more compact form

$$\partial_t(\rho v) + \partial_x(\rho v^2 + p) = 0.$$

The obtained system of two partial differential equations

$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + p) = 0, \end{cases}$$

is called : **Euler's equations**. It remains to model the thermodynamic behavior of the gas since we only have two equations and three unknown variables (density, speed, pressure).

1. First case : gas without pressure, or, to be more precise, with constant pressure. This is a very simplifying assumption, and therefore we get the system
$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2) = 0 \end{cases}$$
which is formally equivalent (for regular solutions such that $\rho \neq 0$) to the scalar equation
$$\partial_t v + v\partial_x v = 0,$$
coupled with the equation of conservation of mass.

2. Second case : Consider a simple *isentropic* setting. In this condition, thermodynamics[3] tells us that the pressure is related to the density by
$$p = p_0 \rho^\gamma,$$
where p_0 and γ are some given positive constants (for example $\gamma = 1.4$ for a diatomic gas O_2, H_2, ...).
The system of equations thus becomes the following system of isentropic Euler equations :
$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + p_0 \rho^\gamma) = 0. \end{cases}$$
This system is nonlinear and highly non-trivial to understand and solve.

3. Third case : Assume for now that the flow is isothermal[4]. In the case of perfect gases, thermodynamics[5] tells us that the law relating the pressure and the density is linear of the form $p = c^2\rho$. We thus find a complete system of the form
$$\begin{cases} \partial_t \rho + \partial_x(\rho v) = 0, \\ \partial_t(\rho v) + \partial_x(\rho v^2 + c^2 \rho) = 0. \end{cases}$$
We can rewrite this system by defining $\eta = \log \rho$ as a new variable, which gives
$$\begin{cases} \partial_t \eta + v\partial_x \eta + \partial_x v = 0, \\ \partial_t v + v\partial_x v + c^2 \partial_x \eta = 0. \end{cases}$$
If we look at small speed ranges (we say that we are linearizing around $v = 0$), we can consider the approximate system
$$\begin{cases} \partial_t \eta + \partial_x v = 0, \\ \partial_t v + c^2 \partial_x \eta = 0. \end{cases}$$
Therefore η and v verify the same wave equation
$$\partial_t^2 \eta - c^2 \partial_x^2 \eta = 0,$$
$$\partial_t^2 v - c^2 \partial_x^2 v = 0.$$
The number c is the wave speed, in this case it is the sound speed.

These systems are called *hyperbolic* which means, *roughly speaking*, that locally, we can always find some change of variables thanks to which the system resembles equations of transport type.

[3] which we will admit here of course...
[4] the temperature remains constant during the evolution
[5] still taken for granted !

II Transport model in any dimension

We will now work on $\mathbb{R}^d$. We suppose that we are given a regular and bounded vector field $v : (t,x) \in \mathbb{R} \times \mathbb{R}^d \mapsto v(t,x) \in \mathbb{R}^d$.

In the study of transport equations, it is common to use another name for the flow that is associated to a vector field, one that we have encountered in the definition I.34.

Definition II.3

Throughout this chapter, the trajectories that are associated to the vector field v will be called **characteristic curves** *(or simply* **characteristics***) and are denoted with the letter X. They are, by definition, solutions of*

$$\begin{cases} \dfrac{\partial}{\partial t} X(t,t_0,x_0) = v(t, X(t,t_0,x_0)), \\ X(t_0,t_0,x_0) = x_0. \end{cases}$$

II.1 Liouville's theorem

We will start by proving an important theorem, originating from physics[6]. It is about the evolution over time of the determinant of the Jacobian matrix of X. This quantity measures the change of volumes under the action of the flow.

For all $t \in \mathbb{R}, y \in \mathbb{R}^d$, we will denote $J(t,y)$ the determinant of the Jacobian $y \mapsto X(t,0,y)$ between the times 0 and t, at point y.

Theorem II.4 (Liouville)

The determinant of the Jacobian J verifies the following linear scalar ordinary differential equation

$$\partial_t J(t,y) = (\operatorname{div} v)(t, X(t,0,y))\ J(t,y). \tag{II.6}$$

Note that y plays the role of a parameter.

Proof :

We use the flow differentiability theorem which gives us the equation (I.14) satisfied by the Jacobian matrix

$$\partial_t (D_y X) = (D_x v)(t, X(t,0,y)).(D_y X).$$

Now, we remind that the determinant map $\det : GL_d(\mathbb{R}) \to \mathbb{R}$ is differentiable and its differential is given by

$$(D\det)(M).H = (\det M)\operatorname{Tr}(M^{-1}.H), \quad \forall M \in GL_d(\mathbb{R}),\ \forall H \in M_d(\mathbb{R}).$$

We can compute the derivative of J since it is the composition of $D_y X$ with $\det$

$$\begin{aligned} \partial_t J(t,y) &= \partial_t(\det(D_y X)(t,0,y)) \\ &= (\det(D_y X)(t,0,y))\operatorname{Tr}((D_y X(t,0,y))^{-1}.\partial_t(D_y X)(t,0,y)) \\ &= J(t,y)\operatorname{Tr}((D_y X(t,0,y))^{-1}.(D_x v)(t,X(t,0,y)).(D_y X(t,0,y))) \\ &= J(t,y)\operatorname{Tr}((D_x v)(t,X(t,0,y))) \\ &= J(t,y)(\operatorname{div} v)(t,X(t,0,y)). \end{aligned}$$

■

Corollary II.5 (Divergence-free fields)

If v is a divergence-free vector field i.e. $\operatorname{div} v = 0$, then the associated flow preserves volumes over time.

Proof :

Liouville's theorem shows that $J(t,y)$ does not depend on t and since we have $X(0,.) = \mathrm{Id}$ at initial time, whose determinant is 1, we obtain that

$$J(t,y) = 1, \quad \forall t \in \mathbb{R}, \forall y \in \mathbb{R}^d.$$

The change of variables theorem shows that for any Borel set $A \subset \mathbb{R}^d$, and any $t \in \mathbb{R}$, the Borel set $X(t,0,A)$ has the same volume (i.e. the same measure) than A. By the group property of the flow, we immediately have that $X(t,s,A)$ has the same volume than A for all times s and t. ■

[6] It can be found in different forms, in Hamiltonian mechanics for instance

II.2 Reynolds's theorem

For any bounded and regular open set $\Omega \subset \mathbb{R}^d$, we denote

$$\Omega_t := X(t, 0, \Omega), \quad \forall t \in \mathbb{R},$$

the image of Ω by the flow X between the times 0 and t. The domains of $\mathbb{R}^d$ that evolve over time by *following* the flow as illustrated in Figure II.2.

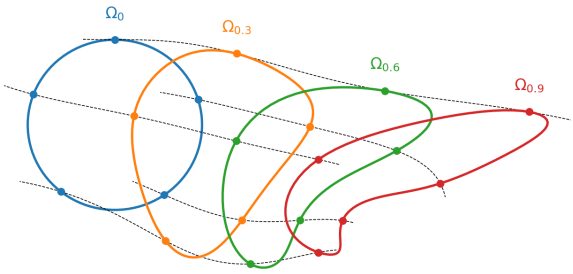


Figure II.2: Drawing at times $t = 0, 0.3, 0.6, 0.9$ of a domain Ω_t evolving with the flow. Some trajectories are drawn in dotted lines.

Let us now consider a $\mathcal{C}^1$ function $f : (t, x) \in \mathbb{R} \times \mathbb{R}^d \mapsto f(t, x) \in \mathbb{R}$. We wish to express the evolution over time of a quantity such as

$$t \mapsto \int_{\Omega_t} f(t, x)\, dx.$$

To do this, we can compute its derivative using the following theorem

Theorem II.6 (Reynolds)

Under the previous hypotheses, for all t we have

$$\frac{d}{dt}\left(\int_{\Omega_t} f(t, x)\, dx\right) = \left(\int_{\Omega_t} \big[\partial_t f + \operatorname{div}(fv)\big](t, x)\, dx\right).$$

Proof :

Let us first give a name to the quantity of interest

$$F(t) := \int_{\Omega_t} f(t, x)\, dx.$$

We will prove this theorem by carrying out the change of variables $x = X(t, 0, y)$ in the integral that we wish to differentiate, which gives

$$F(t) = \int_{\Omega} f(t, X(t, 0, y)) J(t, y)\, dy,$$

where J is the determinant of the Jacobian defined above. Note that we have used the fact that $J > 0$!

Now we differentiate the terms in the integral (all the quantities are sufficiently regular, so this computation is legal). So we have

$$F'(t) = \int_{\Omega} \Big[\partial_t f(t, X(t, 0, y)) + (D_x f)(t, X(t, 0, y)).\partial_t X(t, 0, y)\Big] J(t, y) + f(t, X(t, 0, y))\partial_t J(t, y)\, dy$$

now by using Liouville's theorem and (II.1)

$$= \int_{\Omega} \Big[\partial_t f(t, X(t, 0, y)) + v(t, X(t, 0, y)) \cdot (\nabla f)(t, X(t, 0, y)) + f(t, X(t, 0, y))(\operatorname{div} v)(t, X(t, 0, y))\Big] J(t, y)\, dy.$$

And now by performing the inverse change of variables[7], we come back to the variable x and obtain

$$F'(t) = \int_{\Omega_t} \Big[\partial_t f(t,x) + v(t,x) \cdot (\nabla f)(t,x) + f(t,x)(\operatorname{div} v)(t,x) \Big] \, dx.$$

Thus, by using the formula $\operatorname{div}(fv) = f(\operatorname{div} v) + v \cdot \nabla f$ (see (A.2) in Appendix A) we get the desired result. ■

It is interesting to observe that, thanks to the Stokes formula (Theorem B.1 of Appendix B), we can also write

$$\frac{d}{dt} \left(\int_{\Omega_t} f(t,x) \, dx \right) = \int_{\Omega_t} \partial_t f \, dx + \int_{\partial\Omega_t} f(v \cdot n) \, dx, \tag{II.7}$$

where n is the outward unitary normal vector of the boundary of Ω_t.

This shows that the rate of change of $t \mapsto F(t)$ is due to two phenomena ; the rate of change over time of the function f itself and the fact that the boundary of the domain of integration Ω_t will move around over time according to the field v. In particular, we see that this rate of change only depends on the values of $v \cdot n$ at the boundary of Ω_t, which is quite natural[8].

In the form (II.7), the formula is very similar to the one in the Lemma (II.2).

Remark II.7

If we apply Theorem II.6 to the constant function 1*, we get an equation for the rate of change over time of the volume (=of the Lebesgue measure) of the domain* Ω_t

$$\frac{d}{dt} |\Omega_t| = \int_{\Omega_t} (\operatorname{div} v) \, dx.$$

This of course gives back Corollary II.5 but in a way that is perhaps more convenient to handle. We also see that if $\operatorname{div} v \geq 0$ *then the volume of* Ω_t *increases over time (expansion) and if* $\operatorname{div} v \leq 0$ *then the volume of* Ω_t *decreases over time (compression).*

II.3 Practical example of the establishment of a conservation law

Consider a system consisting in a large number of particles in a macroscopic description[9] (just like we did in Section I). This could be fluid particles, pedestrians in the modeling of a crowd movement, etc.

Let $\rho(t,x)$ be the mass density of particles (=mass of particles per unit volume at time t and in a neighborhood of x) and $v(t,x)$ be the velocity field of the particles. Suppose that ρ and v are unknown functions that are sufficiently regular.

Theorem II.8 (The mass conservation law)

We assume that mass is neither created nor destroyed in the course of evolution, therefore ρ *and* v *are related by the following partial differential equation*

$$\partial_t \rho + \operatorname{div}(\rho v) = 0.$$

This equation is called the **continuity equation** *or the* **mass conservation equation**.

If we also assume that the system has two types of particles (red and blue for example), we can therefore define $\alpha(t,x) \in [0,1]$ the mass fraction of the first type at time t and at point x.

Theorem II.9 (The conservation law for each species)

If the total density ρ *is positive, then* α *and* v *are related by the following partial differential equation*

$$\partial_t \alpha + v \cdot \nabla \alpha = 0.$$

This equation is called the **transport equation** *or the* **convection equation**.

Proof (of Theorem II.8):

[7] Observe that J is now factored in the integral

[8] Imagine, for instance, what $v \cdot n = 0$ means and why it is clear that only the first term must remain in the equality

[9] This means that we are at a scale where particles cannot be studied individually and are rather considered as a continuous medium

For any initial open set Ω, the total quantity of matter in the domain Ω_t stays constant over time because the domain of integration evolves according to the field v and therefore contains exactly the same particles. So we have

$$0 = \frac{d}{dt}\int_{\Omega_t} \rho(t,x)\,dx = \int_{\Omega_t} \partial_t\rho + \mathrm{div}(\rho v)\,dx.$$

Since this is true for any t and any set Ω, we deduce that

$$\int_U (\partial_t\rho + \mathrm{div}(\rho v))\,dx = 0, \quad \forall t \in \mathbb{R}, \quad \forall U \text{ open set.}$$

From the Du Bois-Reymond lemma (or rather its variant - Lemma A.5), we deduce that the following equation is verified

$$\partial_t\rho + \mathrm{div}(\rho v) = 0.$$

■

Proof (of Theorem II.9):

We can reason the same way for the mass conservation for the first species and get

$$\partial_t(\alpha\rho) + \mathrm{div}(\alpha\rho v) = 0.$$

By using the formula for product differentiation, we obtain

$$0 = \alpha(\partial_t\rho + \mathrm{div}(\rho v)) + \rho(\partial_t\alpha + v\cdot\nabla\alpha)$$

and recognize the mass conservation equation in the first term that is zero. Since we assumed that ρ is non zero everywhere, we have finally obtained the desired equation for α. ■

This type of reasoning is at the basis for obtaining many models in which a *transport* phenomenon of certain quantities by a velocity field is involved.

III Classical solutions of transport equations

We will now solve the equations obtained in the previous sections.

III.1 General case of the convection equation

Consider a velocity field $v : \mathbb{R}\times\mathbb{R}^d \mapsto v(t,x) \in \mathbb{R}^d$ that we assume to be $\mathcal{C}^1$ and bounded.

For $u_0 \in \mathcal{C}^1(\mathbb{R}^d)$, we want to solve the following Cauchy problem

$$\begin{cases} \partial_t u + v(t,x)\cdot\nabla u = 0, \quad \forall t > 0, \forall x \in \mathbb{R}^d, \\ u(t=0,x) = u_0(x), \quad \forall x \in \mathbb{R}^d, \end{cases} \tag{II.8}$$

meaning that we are looking for a function $u \in \mathcal{C}^1([0,+\infty)\times\mathbb{R}^d)$ that is a solution of these equations.

Remark II.10 (Beware !)

Although the problem (II.8) has the name **Cauchy problem** *due to the fact that it is an evolution equation in time to which we add an initial data, one should be careful not to rely on any Cauchy-Lipschitz theorem. The equation is not an ordinary differential equation because if we write the equation abstractly as*

$$\partial_t u = \mathcal{F}(t,u),$$

then $\mathcal{F}$ is not a map in the usual sense but it is rather a differential operator which involves derivatives (with respect to x) of the unknown function.
Therefore, we cannot apply the results that we have obtained in Chapter I to solve these partial differential equations.

Theorem II.11

For any regular and bounded field v and for any given initial data $u_0 \in \mathcal{C}^1(\mathbb{R}^d)$, there exists a unique solution u of (II.8) given by the formula

$$u(t,x) = u_0(X(0,t,x)), \quad \forall t > 0, \forall x \in \mathbb{R}^d, \tag{II.9}$$

where X represents the characteristics associated to the field v.

In the context of the study of transport equations, the integral curves of the velocity field, that we will denote by $t \mapsto X(t, 0, x_0)$, are called the **characteristic curves** of the problem. The above theorem consists in showing that the solution is constant along the characteristic curves, see Figure II.3 (traditionally, in this domain of mathematics, time is always represented on the y-axis and position on the x-axis).

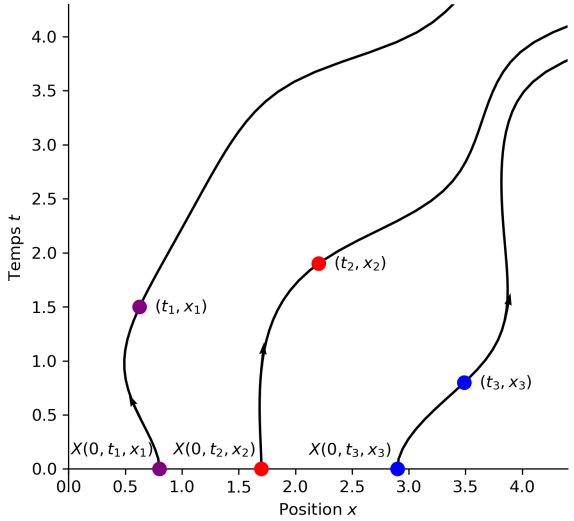


Figure II.3: Illustration of the characteristics method in dimension $d = 1$

Proof :

- Let us start by showing that, if a solution exists, it necessarily verifies the explicit formula (II.9). This will also show the uniqueness of the solution.

  So let u be a solution of (II.8). Fix a point $x_0 \in \mathbb{R}^d$. Now we will study the values of u along the characteristic curves $t \mapsto (t, X(t, 0, x_0))$ in the plane (t, x). More precisely, we define

  $$\varphi(t) := u(t, X(t, 0, x_0)), \quad \forall t \geq 0,$$

  and observe that

  $$\varphi(0) = u(0, x_0) = u_0(x_0),$$

  and that (considering the definition of X)

  $$\begin{aligned}\varphi'(t) &= \partial_t u(t, X(t, 0, x_0)) + (\partial_t X(t, 0, x_0)) \cdot \nabla u(t, X(t, 0, x_0)) \\ &= \partial_t u(t, X(t, 0, x_0)) + v(t, X(t, 0, x_0)) \cdot \nabla u(t, X(t, 0, x_0)) \\ &= (\partial_t u + v \cdot \nabla u)(t, X(t, 0, x_0)) \\ &= 0.\end{aligned}$$

  So, the function φ is constant, which gives

  $$u(t, X(t, 0, x_0)) = u_0(x_0).$$

  For any $x \in \mathbb{R}^d$, we can inverse the characteristic, thanks to the group property of the flow, and define $x_0 = X(0, t, x)$, to obtain

  $$u(t, x) = u_0(X(0, t, x)),$$

  which is indeed the desired formula.

- Conversely, if we define u by the formula (II.9), then we know that u is $\mathcal{C}^1$ thanks to the regularity of the flow and of u_0. Furthermore, we know that for any t

  $$u(t, X(t, 0, x_0)) = u_0(X(0, t, X(t, 0, x_0))) = u_0(x_0),$$

and therefore the function $u(t, X(t, 0, x_0))$ does not depend on time. By performing a computation identical to the previous one, we can show that, for any t and any x_0, we have

$$(\partial_t u + v \cdot \nabla u)(t, X(t, 0, x_0)) = 0.$$

For all $x \in \mathbb{R}^d$, we can apply this formula to $x_0 = X(0, t, x)$ and thus obtain

$$(\partial_t u + v \cdot \nabla u)(t, x) = 0, \quad \forall t \geq 0, \forall x \in \mathbb{R}^d,$$

which is indeed the desired transport equation.

■

Remark II.12 (The main ingredient)

The essence of the above proof lies in the use of the chain rule

$$\frac{d}{dt}\big(u(t, X(t, 0, x_0))\big) = \big(\partial_t u + v \cdot \nabla u\big)(t, X(t, 0, x_0)), \tag{II.10}$$

that is valid for any differentiable function u and for any regular vector field v where we denoted the flow by X. We will encounter this computation several times in the sequel.

III.2 Important special cases

There are some special cases where we can explicitly compute the characteristics of a vector field, this means that we have an explicit formula for solutions of the transport equation.

- **Constant velocity :** If $v(t, x) = v$ is independent of t and x, then the characteristics can be easily calculated

$$X(t, s, x_0) = x_0 + (t - s)v,$$

and therefore the solution of the transport problem can be written as

$$u(t, x) = u_0(x - tv), \quad \forall t \in \mathbb{R}, \forall x \in \mathbb{R}^d.$$

Figure II.4 shows an example of a time evolution of the solution to a 1D transport problem with constant velocity $v = 2$ for a Gaussian initial data.

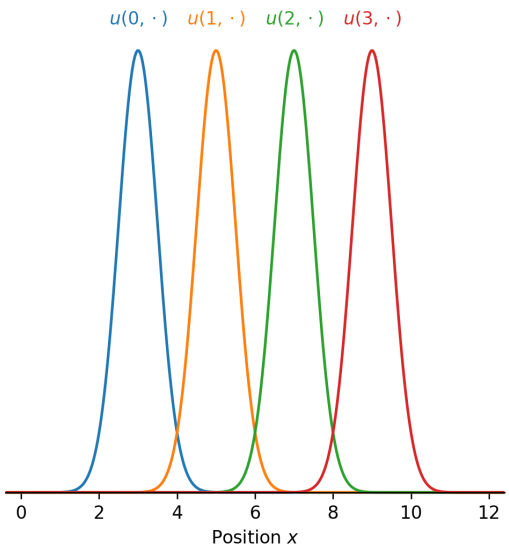


Figure II.4: Solution of a transport problem with constant velocity at times $t = 0, 1, 2, 3$

- **Uniform velocity in space :** Suppose that $v(t,x) = v(t)$ does not depend on the space variable. Therefore, the characteristics are given by

$$X(t,s,x_0) = x_0 + \int_s^t v(\tau)\,d\tau.$$

So we find

$$u(t,x) = u_0\left(x + \int_t^0 v(\tau)\,d\tau\right) = u_0\left(x - \int_0^t v(\tau)\,d\tau\right).$$

Figure II.5 shows an example of a time evolution of the solution to a 1D transport problem with uniform velocity but depending on time, for the same initial data as above. We observe that the dynamics of the solution is different but this solution is not deformed over time.

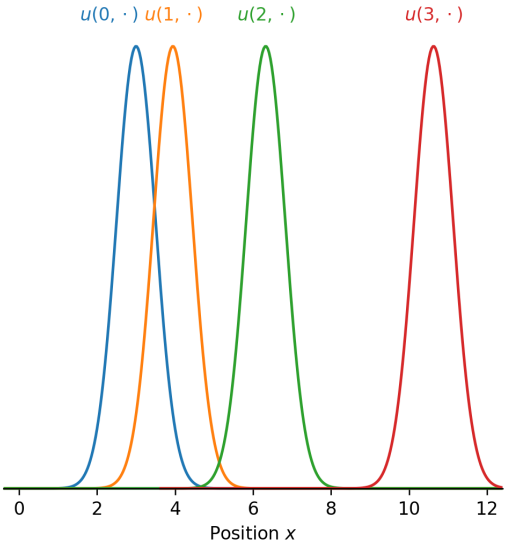


Figure II.5: Solution of a transport problem with a non constant uniform velocity $v(t) = 0.7 + 0.5t$ at times $t = 0, 1, 2, 3$

For comparison, the shape of the solution when the velocity field is no longer uniform with respect to x is given in Figure II.6. This time, we observe that the solution is deformed over time under the influence of the transport phenomenon.

- **A rotating field in 2D :** In dimension 2, suppose that there exists $\omega \in \mathbb{R}$ such that

$$v(x) = \omega \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}, \ \forall x \in \mathbb{R}^2.$$

The associated characteristics are therefore concentric circles given by

$$X(t,s,x_0) = \begin{pmatrix} \cos(\omega(t-s)) & -\sin(\omega(t-s)) \\ \sin(\omega(t-s)) & \cos(\omega(t-s)) \end{pmatrix} x_0.$$

The solution of the equation is therefore given by

$$u(t,x) = u_0(R_{-\omega t}x),$$

where R_θ is a rotation matrix of angle θ on the plan oriented in the counter clockwise direction.

- **A rotating field in 3D :** In dimension 3, suppose that there exists a non zero vector $\omega \in \mathbb{R}^3$ such that

$$v(t,x) = \omega \wedge x, \ \ \forall t \in \mathbb{R}, \forall x \in \mathbb{R}^3.$$

We denote P the plane that is orthogonal to ω ($P = \omega^\perp$). We observe that the ω component of the characteristics is constant

$$X(t,s,x_0) \cdot \omega = x_0 \cdot \omega, \ \ \forall t,s \in \mathbb{R}, \forall x_0 \in \mathbb{R}^3.$$

Moreover, the projection on P of the characteristic $X(t,s,x_0)$ verifies the same equation than the one from the previous case modulo the judicious choice of orientation. So, the trajectories are circles going around the ω axis and with angular speed $|\omega|$.

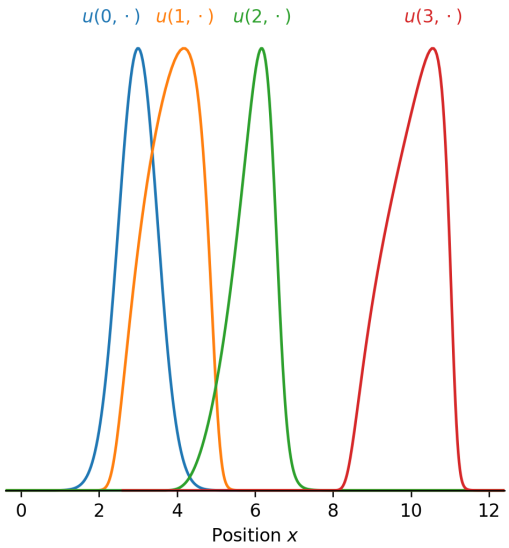


Figure II.6: Solution of a transport problem with varying velocity $v(t,x) = 0.7 + 0.5t + 0.5\sin(2x)$ at times $t = 0, 1, 2, 3$

III.3 Other applications of the characteristics method

The method presented above, which is mostly based on the formula (II.10), allows us to solve many other equations whose principal part[10] is of *transport* type. We will be giving here some examples.

Consider a bounded vector field v of class $\mathcal{C}^1$ and X the characteristics associated to it.

III.3.a Source terms

Theorem II.13

For all $u_0 \in \mathcal{C}^1(\mathbb{R}^d)$ and $f \in \mathcal{C}^0(\mathbb{R}\times\mathbb{R}^d)$, there exists a unique solution $u \in \mathcal{C}^1([0,+\infty)\times\mathbb{R}^d)$ to the following Cauchy problem

$$\begin{cases} \partial_t u + v(t,x)\cdot\nabla u = f(t,x), \quad \forall t>0, \forall x\in\mathbb{R}, \\ u(t=0,x) = u_0(x), \quad \forall x\in\mathbb{R}. \end{cases} \tag{II.11}$$

Proof :

We proceed just like previously. First, suppose that a solution u exists. We fix $x_0 \in \mathbb{R}^d$ and study the evolution of the solution u along the characteristic starting from the Cauchy data $(0, x_0)$. So we define

$$\varphi(t) := u(t, X(t,0,x_0)), \quad \forall t\in\mathbb{R}.$$

We always have $\varphi(0) = u(0, X(0,0,x_0)) = u(0,x_0) = u_0(x_0)$, and now we can compute the derivative of φ just like in (II.10)

$$\varphi'(t) = (\partial_t u + v\cdot\nabla u)(t, X(t,0,x_0)).$$

This time, the quantity is not null but we can use the equation that u verifies to obtain the relation

$$\varphi'(t) = f(t, X(t,0,x_0)),$$

which only depends on the given of the problem and so it allows us to determine φ by integration

$$\varphi(t) = u_0(x_0) + \int_0^t f(s, X(s,0,x_0))\,ds.$$

By coming back to the definition of φ, we see that we have obtained

$$u(t, X(t,0,x_0)) = u_0(x_0) + \int_0^t f(s, X(s,0,x_0))\,ds, \ \forall t\in\mathbb{R}, \ \forall x_0\in\mathbb{R}^d.$$

[10] meaning the terms of the equations that contains some derivatives of the unknown function

It remains to invert the characteristics: fix t and x and define $x_0 = X(0,t,x)$ to obtain, thanks to the group property of the flow,

$$u(t,x) = u_0(X(0,t,x)) + \int_0^t f(s, X(s,t,x))\, ds.$$

This shows the uniqueness of the solution u of the problem and provides a (relatively) explicit formula for the solution. We can now redo all the computations in reverse to establish that the above formula indeed defines a $\mathcal{C}^1$ function which is a solution of (II.11). ■

III.3.b Reaction terms. Continuity equation

Once again, we assume that the field v is regular and bounded. Consider the equation

$$\partial_t \rho + \operatorname{div}(\rho v) = 0,$$

that we discussed above. This equation is not a usual transport equation since we have a term $\operatorname{div}(\rho v)$ instead of $v \cdot \nabla \rho$. Nevertheless, we can use the product differentiation formulas (formula (A.2) of the appendix A) to write the equation as

$$\partial_t \rho + v \cdot \nabla \rho + (\operatorname{div} v)\rho = 0.$$

Since v is a given of the problem, $\operatorname{div} v$ is a function that we know so the main part is indeed of type *transport*. In fact, observe that if $\operatorname{div} v = 0$ we exactly get a transport equation. In general, however, this equation is a special case of a more general problem

$$\partial_t u + v \cdot \nabla u + au = 0,$$

where $a : (t,x) \in \mathbb{R} \times \mathbb{R}^d \mapsto a(t,x) \in \mathbb{R}$ is a given continuous function. The term au in this equation is sometimes called the **reaction** term.

Theorem II.14

For all $u_0 \in \mathcal{C}^1(\mathbb{R}^d)$ and all $a \in \mathcal{C}^0(\mathbb{R} \times \mathbb{R}^d)$, there exists a unique solution $u \in \mathcal{C}^1([0,+\infty) \times \mathbb{R}^d)$ to the following Cauchy problem

$$\begin{cases} \partial_t u + v(t,x) \cdot \nabla u + a(t,x)u = 0, \quad \forall t > 0, \forall x \in \mathbb{R}, \\ u(t=0,x) = u_0(x), \quad \forall x \in \mathbb{R}. \end{cases} \tag{II.12}$$

Furthermore, if $u_0(x) \geq 0$ for all x, then $u(t,x) \geq 0$ for all t, x.

Proof :

We take the usual approach : determine the evolution of u along the characteristics of the field v. So we define

$$\varphi(t) := u(t, X(t,0,x_0)),$$

where u is a possible solution. We have the initial data $\varphi(0) = u_0(x_0)$, and also

$$\varphi'(t) = (\partial_t u + v \cdot \nabla u)(t, X(t,0,x_0)) = -a(t, X(t,0,x_0))u(t, X(t,0,x_0)).$$

On the right hand side we recognize the definition of φ, this gives

$$\varphi'(t) = -a(t, X(t,0,x_0))\varphi(t).$$

This time, we do not get an explicit formula for φ, but a first order linear differential equation that we can solve

$$\varphi(t) = u_0(x_0) \exp\left(- \int_0^t a(s, X(s,0,x_0))\, ds \right).$$

Therefore

$$u(t, X(t,0,x_0)) = u_0(x_0) \exp\left(- \int_0^t a(s, X(s,0,x_0))\, ds \right),$$

which allows us, just like earlier, to get u by inverting the characteristic

$$u(t,x) = u_0(X(0,t,x)) \exp\left(- \int_0^t a(s, X(s,t,x))\, ds \right).$$

Once again, we have a formula of u, which proves the uniqueness of a potential solution. It remains to redo the computations in the reverse order to prove that the formula indeed defines a solution of the problem.

Clearly, by observing the obtained formula, if u_0 is non negative, then u is also non negative since it is the product of some value of u_0 by an exponential term. ■

IV Weak solutions to the transport equation

We will try to define a more general notion of solution for transport equations. There are many reasons for doing this (some of which, and not the least, are beyond the scope of this course), but one can for instance cite the fact that the initial and/or boundary data for the problem may not be differentiable (and even not continuous): one can think of the modeling of the start at a red light in the context of road traffic. When the light turns green, the density is zero downstream and non-zero upstream of the light, so there is a discontinuity in the initial data.

This objective is all the more reasonable since we can see that, for example, the formula

$$u(t,x) = u_0(X(0,t,x)),$$

obtained in Theorem II.11 for the general solution of a transport equation is perfectly well defined even if u_0 is only a continuous function or even a measurable function ...

Moreover, distribution theory allows us to give a broader meaning to the notion of solution of a PDE. We will try here to make a connection between all these concepts.

IV.1 Definition of weak solutions

Definition II.15

Let v be a $\mathcal{C}^1$ bounded vector field. For any initial data $u_0 \in L^1_{loc}(\mathbb{R}^d)$ we say that a function $u \in L^1_{loc}((0,+\infty)\times\mathbb{R}^d)$ is a weak solution of the Cauchy problem (II.8) if, for any test function $\varphi \in \mathcal{C}^\infty_c(\mathbb{R}\times\mathbb{R}^d)$, we have the following

$$\int_0^{+\infty}\int_{\mathbb{R}^d} u(t,x)\big[\partial_t\varphi + \mathrm{div}(\varphi v)\big](t,x)\,dx\,dt + \int_{\mathbb{R}^d} u_0(x)\varphi(0,x)\,dx = 0. \tag{II.13}$$

Note that the local integrability of u and u_0 is sufficient to make sense of all the terms in this equality because φ is compactly supported.

Remark II.16

It is quite clear in the definition that the values of $\varphi(t,x)$ for $t<0$ play no role in the definition. Any weak solution u verifies the formula (II.13) once $\varphi \in \mathcal{C}^\infty(\mathbb{R}\times\mathbb{R}^d)$ verifies

$$\mathrm{Supp}(\varphi) \cap \Big([0,+\infty)\times\mathbb{R}^d\Big) \quad \textit{is compact.}$$

Let us start by making the connection with the other notions of solution that we know and thus justify the relevance of this definition.

Proposition II.17

- *If $u \in \mathcal{C}^1([0,+\infty)\times\mathbb{R}^d)$ is a solution in the usual sense of the transport problem (II.8), then u is also a weak solution of the problem.*
- *If $u \in L^1_{loc}((0,+\infty)\times\mathbb{R}^d)$ is a weak solution of the transport problem, then u verifies the equation*

$$\partial_t u + \mathrm{div}(uv) - (\mathrm{div}\, v)u = 0, \quad \textit{in } \mathcal{D}'((0,+\infty)\times\mathbb{R}^d), \tag{II.14}$$

in the sense of distributions.

But if we know beforehand that $u \in \mathcal{C}^1((0,+\infty)\times\mathbb{R}^d)$, then u verifies the transport equation in the usual sense

$$\partial_t u + v\cdot\nabla u = 0.$$

Notice the particular form of the equation (II.14): one must be careful not to put it in the form $\partial_t u + v\cdot\nabla u = 0$ because the product $v\cdot\nabla u$ is *a priori* not defined if u is a distribution and v a vector field that is not $\mathcal{C}^\infty$.

Proof :

- The proof is a simple integration by parts. Let us take a test function φ just like in the definition, multiply by φ the equation that u verifies, integrate on $(0,+\infty)\times\mathbb{R}^d$, then perform an integration by parts by using the formulas given in the appendix

$$\begin{aligned} 0 &= \int_0^{+\infty} \int_{\mathbb{R}^d} \varphi(\partial_t u + v \cdot \nabla u)\, dx\, dt \\ &= \int_{\mathbb{R}^d} \left(\int_0^{+\infty} \varphi \partial_t u\, dt \right) dx + \int_0^{+\infty} \left(\int_{\mathbb{R}^d} \varphi v \cdot \nabla u\, dx \right) dt \\ &= \int_{\mathbb{R}^d} \left(-\int_0^{+\infty} \partial_t \varphi u\, dt - \varphi(0,x) u(0,x) \right) dx - \int_0^{+\infty} \left(\int_{\mathbb{R}^d} \operatorname{div}(\varphi v) u\, dx \right) dt. \end{aligned}$$

By using the fact that $u(0,x) = u_0(x)$ and rearranging the other terms, we get the desired equation.

- Let us take a test function $\varphi \in \mathcal{D}((0,+\infty) \times \mathbb{R}^d)$. If we extend it by 0 everywhere else, we obtain a function (denoted φ again) that is in $\mathcal{D}(\mathbb{R} \times \mathbb{R}^d)$ which allows us to apply the definition, while observing that $\varphi(0,x) = 0$ for all x. We get

$$0 = \int_0^{+\infty} \int_{\mathbb{R}^d} u(t,x) \big[\partial_t \varphi + \operatorname{div}(\varphi v) \big](t,x)\, dx\, dt,$$

and if we expand the term $\operatorname{div}(\varphi v)$ we get

$$0 = \int_0^{+\infty} \int_{\mathbb{R}^d} u \partial_t \varphi + uv \nabla \varphi + u(\operatorname{div} v)\varphi\, dx\, dt,$$

so

$$0 = \langle u, \partial_t \varphi \rangle_{\mathcal{D}',\mathcal{D}} + \langle uv, \nabla \varphi \rangle_{\mathcal{D}',\mathcal{D}} + \langle u(\operatorname{div} v), \varphi \rangle_{\mathcal{D}',\mathcal{D}}\,.$$

By definition of derivatives in the sense of distributions (or more exactly, of the divergence in the sense of distributions) we have

$$0 = \langle -\partial_t u, \varphi \rangle_{\mathcal{D}',\mathcal{D}} - \langle \operatorname{div}(uv), \varphi \rangle_{\mathcal{D}',\mathcal{D}} + \langle u(\operatorname{div} v), \varphi \rangle_{\mathcal{D}',\mathcal{D}}\,,$$

that is

$$0 = \langle -\partial_t u - \operatorname{div}(uv) + u(\operatorname{div} v), \varphi \rangle_{\mathcal{D}',\mathcal{D}}\,,$$

which gives the desired equation.

If u is a regular function, then the derivatives in the sense of distributions coincide with the usual derivatives (Proposition A.18) so we have

$$\partial_t u + \operatorname{div}(uv) - u(\operatorname{div} v) = 0,$$

and we recognize the transport equation thanks to the formula (A.2).

■

In dimension $d = 1$, we can determine weak solutions that are piecewise regular in the following way.

Proposition II.18

Let us give ourselves a number $a \in \mathbb{R}$, and define the two sub-domains

$$\Omega_l = \{(t,x),\ \ x < X(t,0,a)\},\ \text{ and }\ \Omega_r = \{(t,x),\ \ x > X(t,0,a)\},$$

delimited by an interface Γ which is nothing but the characteristic curve starting from a,

$$\Gamma = \{(t, X(t,0,a)), t > 0\}.$$

Let u be a measurable function defined on $(0,+\infty) \times \mathbb{R}$ and piecewise $\mathcal{C}^1$ on the domains Ω_l and Ω_r. We remind that this means that there exist $u_l, u_r \in \mathcal{C}^1(\mathbb{R} \times \mathbb{R})$ two functions such that

$$u(t,x) = \begin{cases} u_l(t,x), & \text{if } (t,x) \in \Omega_l, \\ u_r(t,x), & \text{if } (t,x) \in \Omega_r. \end{cases}$$

We do not impose in advance any relations on u_l and u_r so u can be (the class of) a discontinuous function.

If u_l (resp. u_r) is a solution, in the usual sense, of the transport equation in Ω_l (resp. in Ω_r), then u is a weak solution of the transport equation in the whole domain $\mathbb{R} \times \mathbb{R}$.

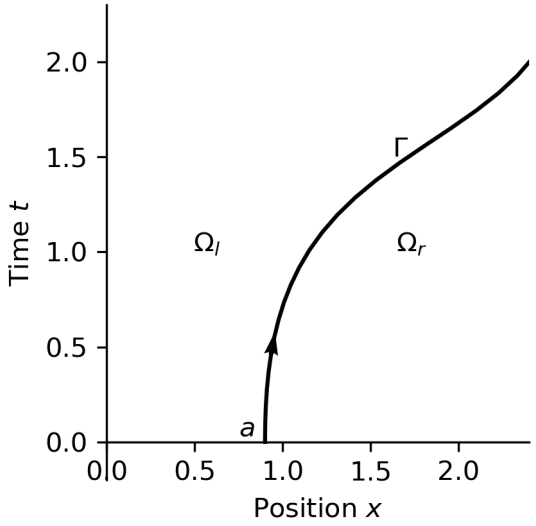


Figure II.7: Weak solutions that are piecewise regular

The situation described here is illustrated in Figure II.7.
For example, the piecewise constant function defined by

$$u(t,x) = \begin{cases} u_l, & \text{if } (t,x) \in \Omega_l, \\ u_r, & \text{if } (t,x) \in \Omega_r, \end{cases}$$

where $u_l, u_r \in \mathbb{R}$ are real numbers, is a weak solution of the Cauchy problem (II.8) for the initial data

$$u_0(x) = \begin{cases} u_l, & \text{if } x < a, \\ u_r, & \text{if } x > a. \end{cases}$$

Proof :
Since $u = u_l$ on Ω_l and since it verifies the equation on Ω_l, we can write

$$\begin{aligned} 0 &= \int_{\Omega_l} \varphi(\partial_t u + v\partial_x u)\,dx\,dt \\ 0 &= \int_{\Omega_l} \varphi(\partial_t u_l + v\partial_x u_l)\,dx\,dt \\ &= \int_0^{+\infty} \int_{-\infty}^{X(t,0,a)} \varphi(\partial_t u_l + v\partial_x u_l)\,dx\,dt \\ &= \int_0^{+\infty} \left(\int_{-\infty}^{X(t,0,a)} \varphi \partial_t u_l\,dx \right) dt + \int_0^{+\infty} \left(\int_{-\infty}^{X(t,0,a)} \varphi v \partial_x u_l\,dx \right) dt. \end{aligned} \tag{II.15}$$

Now, we will perform an integration by parts in time for the first term and in space for the second one. Let us treat the two cases separately:

- *Integration by parts in time.*

  We will use Lemma II.2 to perform the following computation

$$\begin{aligned} \frac{d}{dt}\left(\int_{-\infty}^{X(t,0,a)} u_l \varphi\,dx \right) &= \int_{-\infty}^{X(t,0,a)} (\partial_t u_l)\varphi\,dx + \int_{-\infty}^{X(t,0,a)} u_l(\partial_t \varphi)\,dx \\ &\quad + \left(\frac{d}{dt} X(t,0,a) \right) u_l(t, X(t,0,a))\varphi(t, X(t,0,a)) \\ &= \int_{-\infty}^{X(t,0,a)} (\partial_t u_l)\varphi\,dx + \int_{-\infty}^{X(t,0,a)} u_l(\partial_t \varphi)\,dx \\ &\quad + (\varphi u_l v)(t, X(t,0,a)). \end{aligned}$$

Now, we integrate this equality with respect to time between the bounds 0 and $+\infty$ to obtain

$$-\int_{-\infty}^{a} u_l(0,x)\varphi(0,x)\,dx = \int_0^{+\infty}\int_{-\infty}^{X(t,0,a)} (\partial_t u_l)\varphi\,dx\,dt + \int_0^{+\infty}\int_{-\infty}^{X(t,0,a)} u_l(\partial_t\varphi)\,dx\,dt + \int_0^{+\infty} (\varphi u_l v)(t,X(t,0,a))\,dt.$$

By rearranging the terms, we can express the first term of the right hand side of (II.15) as

$$\int_0^{+\infty}\int_{-\infty}^{X(t,0,a)} \varphi(\partial_t u_l)\,dx\,dt = -\int_0^{+\infty}\int_{-\infty}^{X(t,0,a)} u_l(\partial_t\varphi)\,dx\,dt - \int_{-\infty}^{a} u_l(0,x)\varphi(0,x)\,dx - \int_0^{+\infty} (\varphi u_l v)(t,X(t,0,a))\,dt. \quad \text{(II.16)}$$

- *Integration by parts in space.*

  This computation is simpler and only uses the fact that φ is compactly supported in space

$$\int_0^{+\infty}\left(\int_{-\infty}^{X(t,0,a)} \varphi v\partial_x u_l\,dx\right)dt = -\int_0^{+\infty}\left(\int_{-\infty}^{X(t,0,a)} \partial_x(\varphi v)u_l\,dx\right)dt + \int_0^{+\infty}(\varphi u_l v)(t,X(t,0,a))\,dt. \quad \text{(II.17)}$$

If we add up (II.16) and (II.17), we observe that the last two terms (supported on the interface Γ) cancel out. By using (II.15), we finally obtain

$$0 = -\int_{\Omega_l} u_l(\partial_t\varphi + \partial_x(\varphi v))\,dx\,dt - \int_{-\infty}^{a} u_l(0,x)\varphi(0,x)\,dx.$$

A similar computation in the domain Ω_r gives

$$0 = -\int_{\Omega_r} u_r(\partial_t\varphi + \partial_x(\varphi v))\,dx\,dt - \int_{a}^{+\infty} u_r(0,x)\varphi(0,x)\,dx.$$

Finally, since $u = u_l$ on Ω_l and $u = u_r$ on Ω_r, we can add up these equations and obtain (by removing all $-$ signs)

$$0 = \int_0^{+\infty}\int_{\mathbb{R}} u(\partial_t\varphi + \partial_x(\varphi v))\,dx\,dt + \int_{-\infty}^{a} u_l(0,x)\varphi(0,x)\,dx + \int_a^{+\infty} u_r(0,x)\varphi(0,x)\,dx.$$

Thus, we have shown that u is a weak solution (II.13) with

$$u_0(x) = \begin{cases} u_l(0,x), & \text{if } x < a, \\ u_r(0,x), & \text{if } x > a. \end{cases}$$

as the initial data. ∎

IV.2 Validity of the representation formula using characteristics

The goal of this section is to prove that Formula (II.9) indeed defines a weak solution of the transport equation for any given initial data u_0 with little regularity, say $L^1_{loc}(\mathbb{R}^d)$.

Case of constant velocity Let us start by the simpler case where the velocity is constant.

Consider any $u_0 \in L^1_{loc}(\mathbb{R}^d)$ and a constant vector $v \in \mathbb{R}^d$. Define

$$u(t,x) := u_0(x - tv), \quad \forall t > 0, \forall x \in \mathbb{R}^d.$$

Observe that this definition is consistent with that of the equivalence classes of functions that are equal almost everywhere (i.e. the class of u does not depend on the choice of u_0 among all the other elements in the same class).

Let us take a test function $\varphi \in \mathcal{D}(\mathbb{R}\times\mathbb{R}^d)$ and perform the following computation (by using the change of variables $(t,x) \longmapsto (t, y = x - tv)$)

$$\begin{aligned}\int_0^{+\infty}\int_{\mathbb{R}^d} u(t,x)(\partial_t\varphi + v\cdot\nabla\varphi)(t,x)\,dx\,dt &= \int_0^{+\infty}\int_{\mathbb{R}^d} u_0(x-tv)(\partial_t\varphi + v\cdot\nabla\varphi)(t,x)\,dx\,dt \\ &= \int_0^{+\infty}\int_{\mathbb{R}^d} u_0(y)(\partial_t\varphi + v\cdot\nabla\varphi)(t,y+tv)\,dy\,dt\end{aligned}$$

then by the computation (II.10) of the derivative of the function $t \mapsto \varphi(y+tv)$

$$= \int_0^{+\infty} \int_{\mathbb{R}^d} u_0(y) \frac{\partial}{\partial t} \big[\varphi(t, y+tv)\big] \, dy \, dt$$

finally we use Fubini's theorem

$$= \int_{\mathbb{R}^d} u_0(y) \left(\int_0^{+\infty} \frac{\partial}{\partial t} \big[\varphi(t, y+tv)\big] \, dt \right) dy$$

then the fact that $\varphi(t, .)$ is zero for t large enough

$$= - \int_{\mathbb{R}^d} u_0(y) \varphi(0, y) \, dy,$$

which gives the desired formula.

General case In the case of a more general vector field, the computation is similar but a little more complicated because one should take into account the determinant of the Jacobian of the change of variables $(t, x) \mapsto (t, y = X(0, t, x))$. More precisely, Liouville's theorem (Theorem II.4) can be used to describe the evolution of the determinant of the Jacobian $J(t, y) = (\det D_y X)(t, 0, y)$. We now define

$$u(t, x) := u_0(X(0, t, x))$$

and carry out the following computation

$$\begin{aligned}
&\int_0^{+\infty} \int_{\mathbb{R}^d} u(t,x)(\partial_t \varphi + \operatorname{div}(\varphi v))(t,x) \, dx \, dt \\
&\qquad = \int_0^{+\infty} \int_{\mathbb{R}^d} u_0(X(0,t,x))(\partial_t \varphi + \operatorname{div}(\varphi v))(t,x) \, dx \, dt \\
&\qquad = \int_0^{+\infty} \int_{\mathbb{R}^d} u_0(y) \big(\partial_t \varphi + v \cdot \nabla \varphi + (\operatorname{div} v) \varphi\big)(t, X(t,0,y)) J(t,y) \, dy \, dt
\end{aligned}$$

by using (II.10), we get

$$= \int_0^{+\infty} \int_{\mathbb{R}^d} u_0(y) \left(\frac{\partial}{\partial t} \big[\varphi(t, X(t,0,y))\big] J(t,y) + (\operatorname{div} v)(t, X(t,0,y)) J(t,y) \varphi(t, X(t,0,y)) \right) dy \, dt$$

and thanks to Liouville's theorem (Theorem II.4), we get

$$\begin{aligned}
&= \int_0^{+\infty} \int_{\mathbb{R}^d} u_0(y) \left(\frac{\partial}{\partial t} \big[\varphi(t, X(t,0,y))\big] J(t,y) + \big[\partial_t J(t,y)\big] \varphi(t, X(t,0,y)) \right) dy \, dt \\
&= \int_0^{+\infty} \int_{\mathbb{R}^d} u_0(y) \left(\frac{\partial}{\partial t} \big[J(t,y) \varphi(t, X(t,0,y))\big] \right) dy \, dt
\end{aligned}$$

finally we apply Fubini's theorem

$$\begin{aligned}
&= \int_{\mathbb{R}^d} u_0(y) \left(\int_0^{+\infty} \frac{\partial}{\partial t} \big[J(t,y) \varphi(t, X(t,0,y))\big] \, dt \right) dy \\
&= - \int_{\mathbb{R}^d} u_0(y) J(0,y) \varphi(0,y) \, dy
\end{aligned}$$

and since $J(0, y) = 1$ because $X(0, 0, .) = \mathrm{Id}$,

$$= - \int_{\mathbb{R}^d} u_0(y) \varphi(0, y) \, dy.$$

The result is thus proved.

IV.3 Uniqueness

As soon as we weaken the notion of solution to a given equation, we enlarge the set of possible solutions and in this case, we must make sure that we did not create undesirable new solutions in order for the theory to still be usable.

Theorem II.19

Suppose that the velocity field is $\mathcal{C}^1$. For any given initial data $u_0 \in L^1_{loc}(\mathbb{R}^d)$, the formula (II.9) yields the unique weak solution of the Cauchy problem (II.8).

We have already established that this formula gives a weak solution. It remains to show that it is the only one (more precisely, its equivalence class). Since we are dealing with a linear problem, it is sufficient to show[11] that, if $u_0(x) = 0$ for almost all x, then the only weak solution is $u = 0$.

Proof (Case of constant velocity):

Let $u \in L^1_{loc}((0,+\infty)\times\mathbb{R}^d)$ be a weak solution, with $u_0 = 0$, meaning that

$$\int_0^{+\infty}\int_{\mathbb{R}^d} u(t,x)\big(\partial_t\varphi + v\cdot\nabla\varphi\big)\,dx\,dt = 0, \tag{II.18}$$

for any test function $\varphi \in \mathcal{D}(\mathbb{R}\times\mathbb{R}^d)$. The idea of the proof is to show that $u = 0$ with the help of the Du Bois-Reymond Lemma and to do this, given a function $\psi \in \mathcal{D}((0,+\infty)\times\mathbb{R}^d)$, we will look for a function $\varphi \in \mathcal{C}^\infty(\mathbb{R}\times\mathbb{R}^d)$ with compact support in $[0,+\infty)\times\mathbb{R}^d$ such that

$$\partial_t\varphi + v\cdot\nabla\varphi = \psi, \quad \text{in } (0,+\infty)\times\mathbb{R}^d.$$

We have already solved this problem in Theorem II.13 and we got

$$\varphi(t,x) = \varphi_0(x-tv) + \int_0^t \psi(s, x+(s-t)v)\,ds,$$

but for the moment φ_0 is not determined.

In particular, we have

$$\varphi(t,x+tv) = \varphi_0(x) + \int_0^t \psi(s,x+sv)\,ds,$$

therefore, since we want φ to have compact support in $[0,+\infty)\times\mathbb{R}^d$, the limit of the above quantity as $t\to+\infty$ has to be zero, which gives necessarily

$$\varphi_0(x) = -\int_0^{+\infty} \psi(s,x+sv)\,ds.$$

In the end, we proved that the right choice of φ is given by

$$\varphi(t,x) = -\int_t^{+\infty} \psi(s,x+(s-t)v)\,ds.$$

This function verifies the appropriate properties and can be chosen as a test function in (II.18) (see Remark II.16), which gives

$$\int_0^{+\infty}\int_{\mathbb{R}^d} u(t,x)\psi(t,x)\,dx\,dt = 0,$$

for any function ψ, and now we can apply the Du Bois Reymond lemma (Lemma A.4) to conclude. ∎

For the general case, the proof is very similar but with slightly more heavy computations since the characteristics of the field are not explicit. We leave this proof as an exercise to the reader.

IV.4 The boundary-value problem

Consider once a again a regular and bounded vector field v on $\mathbb{R}\times\mathbb{R}^d$ (its characteristics will always be denoted by X).

Now, let Ω be a domain of $\mathbb{R}^d$ (a regular open set in the sense of Definition B.8) and let $u_0 \in \mathcal{C}^1(\Omega)$ be a given initial data. We will consider the Cauchy problem this time defined on the domain Ω :

$$\begin{cases} \partial_t u + v(t,x)\cdot\nabla u = 0, \quad \forall t>0, \forall x\in\Omega, \\ u(t=0,x) = u_0(x), \quad \forall x\in\Omega, \end{cases} \tag{II.19}$$

[11] you should be able to convince yourself of this!

IV.4.a Tangent field

We start by assuming that the field v is tangent to the boundary of the domain. This means that

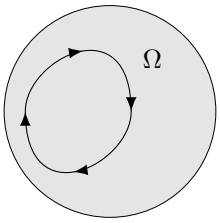
$$v(t,x)\cdot n(x)=0,\ \ \forall t\in\mathbb{R},\forall x\in\partial\Omega.$$

We will now prove that the characteristics that start in Ω will always stay in Ω.

Proposition II.20

Under the previous assumptions, we have

$$X(t,s,x_0)\in\Omega,\ \ \forall t,s\in\mathbb{R},\ \ \forall x_0\in\Omega.$$



This allows us to solve the Cauchy problem stated above.

Corollary II.21

For a velocity field v that is tangent to the boundary of the domain, the problem (II.19) admits a unique solution that is regular, and once again given by the formula

$$u(t,x)=u_0(X(0,t,x)),\ \forall t>0,\ \forall x\in\Omega.$$

Proof (of Proposition II.20):

We will start by assuming (only to simplify the proof a little, it is not absolutely useful ...) that the open set Ω is bounded. In these conditions, its boundary $\partial\Omega$ is compact and by using the regularity property of Ω, we can show that there exists $\varepsilon>0$ such that the map

$$\Phi:(s,y)\in(-\varepsilon,\varepsilon)\times\partial\Omega\times\longmapsto y-\varepsilon n(y)\in\mathbb{R}^d,$$

is a $\mathcal{C}^1$ diffeomorphism on its image U_ε given by

$$U_\varepsilon:=\{x\in\mathbb{R}^d,d(x,\partial\Omega)<\varepsilon\}.$$

This set is called a tubular ε-neighborhood of $\partial\Omega$, and the notations are described in Figure II.8. Furthermore, we can show that the inverse map of Φ has the form

$$\Phi^{-1}(x)=\big(\delta(x),P(x)\big)\in(-\varepsilon,\varepsilon)\times\partial\Omega,$$

where $\delta(x)$ is the **signed distance** of a point x to the boundary of Ω defined as follows

$$\delta(x):=\begin{cases}d(x,\partial\Omega), & \text{if } x\in\overline{\Omega},\\ -d(x,\partial\Omega), & \text{if } x\in\overline{\Omega}^c,\end{cases}$$

and $P(x)$ is the projection of x on $\partial\Omega$, meaning the unique[12] point of the compact $\partial\Omega$ that verifies

$$\|x-P(x)\|=d(x,\partial\Omega).$$

Finally, we can find

$$\nabla\delta(x)=-n(P(x)).$$

Now that we have acquired these somewhat technical but fairly intuitive preliminaries, we can proceed to the proof of the announced property.

[12]The uniqueness comes from the fact that Ω is a regular set and ε is small, ... in general, the projection of a point on an arbitrary compact set is not necessarily unique !

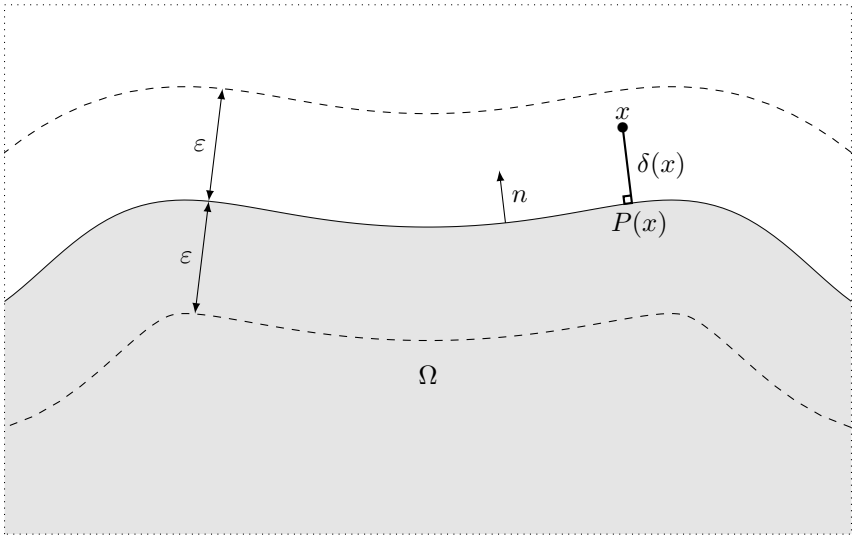


Figure II.8: Notations in the tubular neighborhood of the boundary of Ω

For any point $x \in U_\varepsilon$, since v and δ are regular, the mean value theorem gives us

$$\big|(v\cdot\nabla\delta)(t,x) - (v\cdot\nabla\delta)(t,P(x))\big| \leq C\|x-P(x)\| = C|\delta(x)|, \quad \forall t\in[-T,T], \forall x\in U_\varepsilon,$$

the constant C depending on a fixed $T>0$. On the other hand, we know, from our assumption, that on the boundary of our domain we have

$$(v\cdot\nabla\delta)(t,y) = (v\cdot n)(t,y) = 0, \quad \forall t, \forall y\in\partial\Omega.$$

By combining the above two results, we get

$$\big|(v\cdot\nabla\delta)(t,x)\big| \leq C|\delta(x)|, \forall t\in[-T,T], \forall x\in U_\varepsilon. \tag{II.20}$$

Now, let us take a $x_0 \in \partial\Omega$ and study the evolution of the quantity $\delta(X(t,0,x_0))$, which is the distance of the characteristic, that starts at x_0, to the boundary of the domain. We carry out the following simple computations

$$\begin{aligned}\frac{d}{dt}\delta(X(t,0,x_0)) &= \nabla\delta(X(t,0,x_0))\cdot\frac{d}{dt}X(t,0,x_0)\\ &= \nabla\delta(X(t,0,x_0))\cdot v(t,X(t,0,x_0)).\end{aligned}$$

By integration, and by also using (II.20), we get that, as long as the characteristic stays in U_ε,

$$|\delta(X(t,0,x_0)| \leq |\delta(x_0)| + C\left|\int_0^t |\delta(X(s,0,x_0))|\,ds\right|, \quad \forall t\in[-T,T].$$

Gronwall's lemma tells us that

$$|\delta(X(t,0,x_0))| \leq |\delta(x_0)|e^{C|t|}, \quad \forall t\in[-T,T].$$

Since we started from a point on the boundary of Ω, we have that $\delta(x_0)=0$ and finally deduce that

$$\delta(X(t,0,x_0)) = 0, \quad \forall t\in[-T,T].$$

Since this is true for all T, this means we have shown that the characteristic $t\mapsto X(t,0,x_0)$ is completely contained in the boundary of Ω.

Now, if $x_0\in\Omega$ then $t\mapsto X(t,0,x_0)$ is entirely contained in Ω. In fact, if that was not the case, this trajectory should intersect the boundary of Ω and therefore, from the reasoning that was previously made, should be completely contained in the boundary, which is not the case since $X(0,0,x_0)=x_0$ is in the open set Ω.

■

IV.4.b Non tangent field

If the vector field v is no longer tangent to the boundary of the domain, Proposition II.20 is no longer valid and the characteristics method shows that, to solve the problem, we lack one information about what happens on the part of the boundary where the field enters.

To simplify, let us assume that v is independent of time and introduce the following subset of the boundary

$$\Gamma_{\text{ent}} = \big\{x \in \partial\Omega,\ v(x)\cdot n(x) < 0\big\}.$$

The Cauchy problem associated to the transport equation in Ω has to be completed by a boundary condition in the following way : given $u_0 \in \mathcal{C}^1(\Omega)$ and $u_{\text{ent}} \in \mathcal{C}^1(\mathbb{R}\times\Gamma_{\text{ent}})$, we want to find a solution u of

$$\begin{cases} \partial_t u + v(t,x)\cdot\nabla u = 0, & \forall t>0, \forall x\in\Omega, \\ u(t=0,x) = u_0(x), & \forall x\in\Omega, \\ u(t,x) = u_{\text{ent}}(t,x), & \forall t>0, \forall x\in\Gamma_{\text{ent}}. \end{cases} \tag{II.21}$$

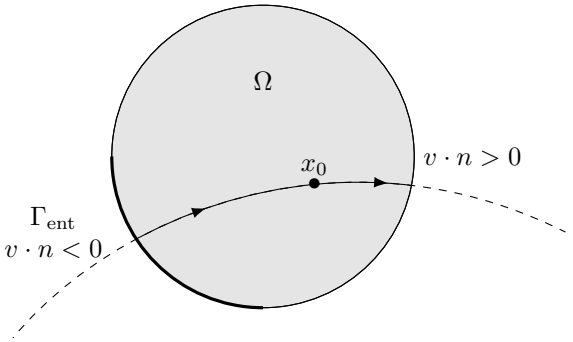


Figure II.9: The characteristics associated to a vector field that is not tangent to the boundary of Ω

We give the following result without proof:

Theorem II.22

For any given data u_0, u_{ent} just like above, there exists a unique weak solution u of Problem (II.21).

Even if proving this result rigorously is not simple, we can easily understand the structure of the solution by *backtracking the characteristics* just like we did previously, see Figure II.9. For fixed values of $t>0$ and $x\in\Omega$, we will study the characteristic $s\mapsto X(s,t,x)$ starting from $s=t$ and going back in time until we get to $s=0$. Two cases can arise:

- If $X(s,t,x)\in\Omega$ for all $s\in[0,t]$, then the solution at point (t,x) will be given by

  $$u(t,x) = u_0(X(0,t,x)),$$

  and does not depend on the boundary data.

- In the opposite case, if we denote $\tau(t,x)$ the largest instant in $[0,t]$ for which $X(\tau(t,x),t,x)$ belongs to the boundary of Ω, then the solution at point (t,x) will be given by

  $$u(t,x) = u_{\text{ent}}(\tau(t,x), X(\tau(t,x),t,x)),$$

  and, this time, it does not depend on the initial data.

IV.5 An example of a population dynamics model

Derivation of the equations Consider an age-structured cell division model (see [7]). The "space" variable x is the age of each cell so, in particular, it's a positive quantity. The density of the cells with age x at time t, denoted $n(t,x)$, is the unknown.

In absence of any other phenomenon, the natural aging of cells (whose age increases at a rate of 1 as a function of time[13]) is governed by the free transport equation

$$\partial_t n + \partial_x n = 0,\ \ \forall t>0, \forall x>0.$$

[13] As for everyone else ...

We now wish to model cell division. Suppose that it takes place with a rate $k(x)$ that depends on the age of the cells[14].

Let us fix two ages $0 < a < b$ and look at the amount of cells whose age at time t is between $a + t$ and $b + t$

$$F(t) := \int_{a+t}^{b+t} n(t,x)\,dx.$$

A change in this quantity is only due to cell division. More precisely, among the cells considered in $F(t)$, the number of those that will divide per unit of time is given by

$$\int_{a+t}^{b+t} k(x)n(t,x)\,dx.$$

As the cells get divided, they disappear and give birth to two newborn cells of age 0. So we deduce that

$$F'(t) = -\int_{a+t}^{b+t} k(x)n(t,x)\,dx.$$

After calculating the derivative of F by using Lemma II.2, and the usual du Bois-Reymond type argument gives us the following partial differential equation

$$\partial_t n + \partial_x n + k(x)n = 0.$$

But we must also take into account the fact that, at each time, a certain amount of cells of age 0 are created. There are twice as many cells that are created than those that got divided, and this gives us the relation

$$n(t,0) = 2\int_0^{+\infty} k(x)n(t,x)\,dx.$$

If we put together these two equations, we end up with the following final model

$$\begin{cases} \partial_t n + \partial_x n + k(x)n = 0,\ \forall t > 0, \forall x > 0, \\ \qquad n(t,0) = 2\displaystyle\int_0^{+\infty} k(y)n(t,y)\,dy,\ \forall t > 0, \\ \qquad n(0,x) = n_0(x),\ \forall x > 0. \end{cases} \tag{II.22}$$

Resolution

Theorem II.23

Suppose that n_0 has compact support and that k is positive and bounded. In this case, there exists a unique weak solution of (II.22).

Proof :

Define the anti-derivative derivative K of k by

$$K(x) = \int_0^x k(y)\,dy,\ \ \forall x \geq 0.$$

Then perform the change of variables

$$N(t,x) = e^{K(x)}n(t,x),$$

so that N verifies the new system

$$\begin{cases} \partial_t N + \partial_x N = 0,\ \forall t > 0, \forall x > 0, \\ \qquad N(t,0) = 2\displaystyle\int_0^{+\infty} k(y)e^{-K(y)}N(t,y)\,dy,\ \forall t > 0, \\ \qquad N(0,x) = N_0(x),\ \forall x > 0, \end{cases} \tag{II.23}$$

where N_0 is defined by

$$N_0(x) = e^{K(x)}n_0(x).$$

Therefore N verifies an equation of transport type with constant velocity (equal to 1) with a known initial data and an unknown boundary condition data $t \mapsto N(t,0)$. We denote it by $\rho(t) := N(t,0)$ so that the problem is reduced to the determination of ρ. In fact, once we have ρ, we can fully determine N by using the characteristics method and by taking into account the boundary just like in Section IV.4.b, by distinguishing two zones:

[14] typically the younger cells do not divide or divide very little

- If $x > t$, we immediately obtain
$$N(t,x) = N(0, x-t) = N_0(x-t).$$
- If $x < t$, we obtain
$$N(t,x) = N(t-x, 0) = \rho(t-x).$$

Now let us interpret the boundary condition for all $t > 0$

$$\rho(t) = 2\int_0^{+\infty} k(y)e^{-K(y)}N(t,y)\,dy.$$

Let's decompose the integral into two parts

$$\rho(t) = 2\int_0^{t} k(y)e^{-K(y)}N(t,y)\,dy + 2\int_t^{+\infty} k(y)e^{-K(y)}N(t,y)\,dy,$$

and use the results obtained above to express N in the two integrals

$$\rho(t) = 2\int_0^{t} k(y)e^{-K(y)}\rho(t-y)\,dy + 2\int_t^{+\infty} k(y)e^{-K(y)}N_0(y-t)\,dy.$$

Observe that the second term is a function of t that is completely determined by the data of the problem, we will denote it by ρ_0. The problem is therefore reduced to finding a function ρ that verifies the following integral problem,

$$\rho(t) = 2\int_0^{t} k(y)e^{-K(y)}\rho(t-y)\,dy + \rho_0(t), \quad \forall t > 0. \tag{II.24}$$

We will use a fixed point method, in a similar way to what we did in the proof of the global Cauchy-Lipschitz theorem in Chapter I. Define $\gamma = 4\|k\|_\infty$ and the space

$$E = \left\{ z \in \mathcal{C}^0([0,+\infty),\mathbb{R}), \quad \|z\|_E := \sup_{t\geq 0} e^{-\gamma t}|z(t)| < +\infty \right\},$$

which is a Banach space.

Then, define the map $\Theta : E \to E$ by

$$\Theta : z \in E \mapsto \Theta(z) : \left(t \in [0,+\infty) \mapsto 2\int_0^t k(y)e^{-K(y)}z(t-y)\,dy + \rho_0(t) \right) \in E.$$

One should verify that Θ sends elements of E into itself by using Gronwall's lemma and the fact that ρ_0 is bounded.

We will now show that Θ is contracting. To do this, take $z_1, z_2 \in E$ and, for all $t \geq 0$, we compute

$$\begin{aligned}
|\Theta(z_1)(t) - \Theta(z_2)(t)| &\leq 2\int_0^t k(y)e^{-K(y)}|z_1(t-y) - z_2(t-y)|\,dy \\
&\leq 2\|k\|_\infty \int_0^t |z_1(s) - z_2(s)|\,ds \\
&\leq 2\|k\|_\infty \int_0^t e^{\gamma s}e^{-\gamma s}|z_1(s) - z_2(s)|\,ds \\
&\leq 2\|k\|_\infty \left(\int_0^t e^{\gamma s}\,ds \right) \|z_1 - z_2\|_E \\
&\leq \frac{2\|k\|_\infty}{\gamma} e^{\gamma t}\|z_1 - z_2\|_E.
\end{aligned}$$

By rearranging the term $e^{\gamma t}$, and by using the definition of γ, we obtain

$$\|\Theta(z_1) - \Theta(z_2)\|_E \leq \frac{1}{2}\|z_1 - z_2\|_E.$$

Therefore, Θ admits a unique fixed point in E that we will denote by ρ which is indeed the unique solution to the equation (II.24). Observe that since ρ_0 is positive, it is clear as to why ρ is also positive.

Since we now know ρ, the unique weak solution N of our equation is completely determined. It is $\mathcal{C}^1$ in both domains $\{t < x\}$ and $\{t > x\}$ but possibly discontinuous at the interface (See Proposition II.18).

■

In Figure II.10, we show an example of the dynamics of this system. The initial data n_0 is Gaussian centered at $x = 1.5$ and the function k is a piecewise constant function which is equal to 0 for $x \leq 3$ and 1 for $x \geq 3$.

We can simultaneously observe the aging and the cell division phenomenon that has, as a consequence, the progressive disappearance of cells above the age of 3 and the concomitant appearance of newborn cells of age 0. Globally, the size of the population is increasing (which, by the way, can be directly verified on the equation). We also observe that the solution is continuous but its derivative has some discontinuity at point $x = 3$.

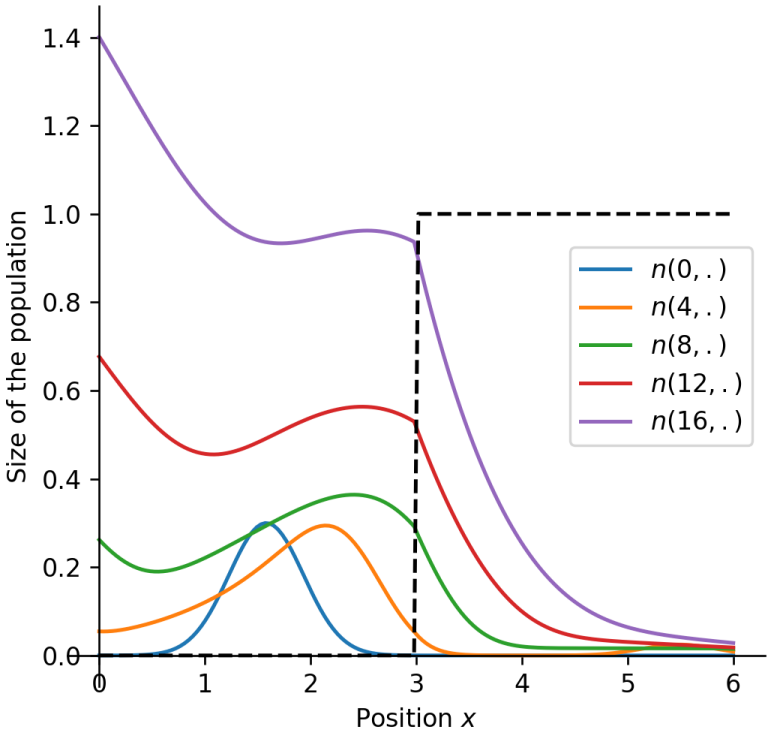


Figure II.10: An example of the time evolution of the cell division model

What should be retained from this chapter

In priority

- The differentiation of integrals with one parameter just like in Lemma II.2.
- To recognize some modeling situations (primarily in dimension 1 of space) that leads to an equation of transport type (or continuity).
- To know the characteristics method for solving the Cauchy problem for transport equations (and its variations) with a given regular data : in particular the 1D case has to be mastered perfectly (see exercises).
- The explicit solution of these problems must be known in the simple cases where the characteristics are simple to compute.

To go even further

- Reynolds's theorem.
- Proofs of the existence and uniqueness of weak solutions. Weak solutions that are piecewise regular.
- The boundary conditions problem of transport equations.

Chapter III

Variational formulation of boundary value problems

The purpose of this chapter is to present, through a simple (and relatively concrete) example, some notions of calculus of variations. This will lead us to the notion of *variational* formulation of boundary value problems (meaning an elliptic PDE + boundary conditions). We will also see why it is natural, starting from the 1D case, to introduce new function spaces that are well adapted to this type of approach.

Later we will generalize these concepts to deal with more complex problems, in particular in any dimension.

I The problem of the elastic string/membrane at equilibrium

I.1 Presentation

Consider an elastic membrane which, at rest, is represented by a compact region $\overline{\Omega}$ of the horizontal plane $\mathbb{R}^2$ (we consider it to be at an altitude of 0), where Ω is a **connected** open set of $\mathbb{R}^2$ (see Figure III.1)

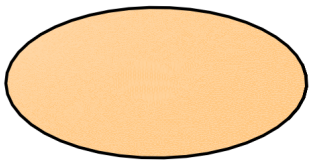


Figure III.1: A circular and plane membrane at rest

We apply a vertical force on the membrane, this will lead to its deformation. This force is exerted at each point of the membrane, and more precisely we denote $f(x)$ the force density that will be exerted at each point of the membrane. In other words, an elementary surface dx around the point x will be subjected to a force of magnitude $f(x)\,dx$ which will imply that the total force F_A exerted on a region A of the membrane will be given by the integral

$$F_A = \int_A f(x)\,dx.$$

Under the action of the applied force, for any $x \in \overline{\Omega}$, the initial point of the physical space $(x, 0) \in \mathbb{R}^3$ is moved to a new point of $\mathbb{R}^3$ denoted $\tilde{u}(x) = (x, u(x))$, where u is a function going from $\overline{\Omega}$ to $\mathbb{R}$, which represents the vertical displacement that we will try to determine. So the question is : among the infinite number of possible equilibrium positions of the membrane (see the three examples in Figure III.2) which one will really be observed for a given force f? To answer this question we will have to write in mathematical terms a fundamental principle from physics : under the action of an exterior force, the system will *choose* the position u that will minimize the total energy of the membrane. This energy is the sum of two potential energies : one related to the work of the exerted forces and the other related to the elasticity of the material of which the membrane is made out of (when you stretch a rubber band you give it a certain amount of energy - which you can release by letting go of the band!).

We will only consider small displacements (i.e. f and u are assumed to be **small** and we will see later what this

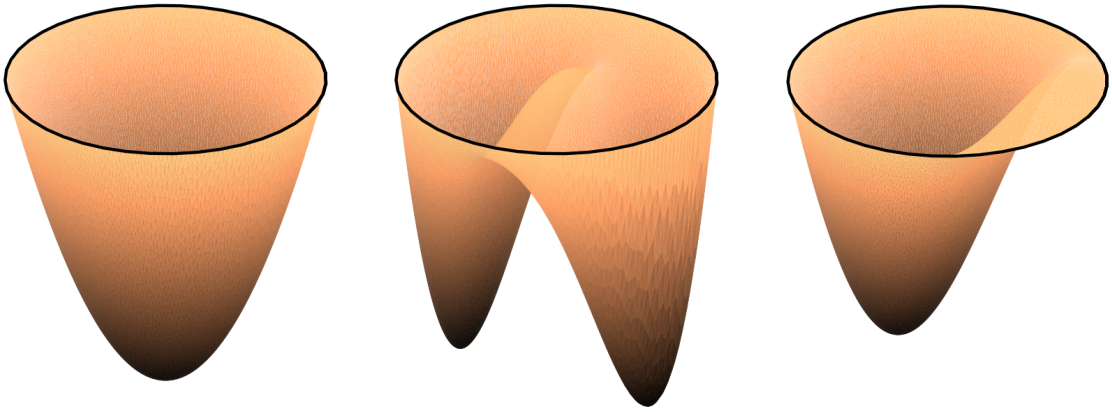


Figure III.2: Different possible positions for the membrane

implies). Furthermore, we will study the situation in which the boundary of the membrane is attached to a fixed reference (think of the skin of a drum). This implies that we will look for a function u that is zero on the boundary $\partial\Omega$.

Now, let us do an assessment of the potential energy of the system :

- The system acquires a virtual potential energy that is equal to the opposite of the work done by the exterior force f in relation to u, i.e.
$$E_1(u) := -\int_\Omega u(x)f(x)\,dx.$$
The presence of the sign $-$ is natural : if the force is oriented downwards ($f \leq 0$) then the areas of the membrane with high potential energy are those that have high altitudes (so for large $u \geq 0$) (like the field of gravity for instance).

- Moreover, the system has an elastic potential energy due to the stretching of the membrane. We admit that this energy is proportional to the change in surface area of the membrane[1], meaning of the form
$$k(\text{Stretched surface area} - \text{Surface area at rest}) = k(|\tilde{u}(\Omega)| - |\Omega|),$$
where $k > 0$ is the *stiffness* of the considered material.
By a change of variables, we get a new way of expressing the energy
$$k\int_\Omega \left(\sqrt{1+|\nabla u|^2} - 1\right) dx.$$
In the assumption of small displacements, u is small as well as its derivatives. By using the usual Taylor expansion, we can approximate this expression by the following *quadratic* quantity:
$$E_2(u) := \frac{k}{2}\int_\Omega |\nabla u|^2\,dx.$$

The total potential energy of the system for a possible position u is given by
$$E(u) := E_1(u) + E_2(u) = \frac{k}{2}\int_\Omega |\nabla u|^2\,dx - \int_\Omega fu\,dx.$$
In this formula, note that $u : \Omega \to \mathbb{R}$ is a function and $E(u)$ is a number.

The fundamental principle of Lagrangian mechanics tells us that, under the action of the force field f, the membrane will be stretched according to a displacement u that will minimize the total potential energy (that we will call *the action* according to the *ad hoc* vocabulary of mechanics). So, we will study the following problem : find $u \in X$ such that
$$E(u) = \inf_{v\in X} E(v) = \inf_{v\in X}\left(\frac{k}{2}\int_\Omega |\nabla u|^2\,dx - \int_\Omega fu\,dx\right), \tag{III.1}$$

[1] This can be "proved", for instance, by assimilating the membrane to a network of small springs

where X is the following function space (we have represented three elements of X in Figure III.2; the black boundary of the membrane is the fixed part):

$$X = \{v : \overline{\Omega} \mapsto \mathbb{R},\ \text{differentiable},\ \ v = 0,\ \ \text{on } \partial\Omega\},$$

constituted of the set of **admissible** positions of the membrane.

Remark III.1

Note that the model has been established under one condition that cannot be found in the formulation that we presented here; so, in some applications of this model, it is relevant to ask oneself a posteriori if, after resolution of the problem, the obtained solution is valid. If the solution turns out to be too big (compared to what ?) then it is possible that the simplifications of the model that were used are not, in fact, relevant. We had already encountered this problem when discussing the validity of the linearized pendulum equation.

We will later see that the space X defined above is not necessarily the best choice to make the mathematical methods, that we will present here, work. Actually, a portion of this chapter will consist of understanding and finding the best function space to consider for X.

I.2 The mathematical questions that we would like to solve

In what will follow, we will try to answer the following questions:

1. Does the problem (III.1) admit a unique solution ? In particular, is the infimum of E finite on X ?
2. If such a solution exists, is it unique ?
3. If such a solution exists, can we characterize it with the help of a "simple" equation that we can eventually solve ?

I.2.a Rewriting in the case of dimension 1

From now on, and until Section IV, we will focus on the easier case of dimension 1. Therefore the model now corresponds to an elastic string (it is no longer a membrane). To simplify the problem even more, we will assume that $\Omega = (0, 1)$ so that the problem can be formulated in the following way : find $u : [0, 1] \mapsto \mathbb{R}$ differentiable, that is equal to 0 at $x = 0$ and $x = 1$ and such that

$$E(u) = \inf_{v \in X} E(v), \tag{III.2}$$

where

$$X = \{v : [0, 1] \mapsto \mathbb{R}, \text{differentiable and s.t. } v(0) = v(1) = 0\},$$

and the energy is written as

$$E(v) = \frac{k}{2} \int_0^1 |v'(x)|^2 \, dx - \int_0^1 f(x)v(x) \, dx.$$

Additionally, f is assumed to be at least integrable.

I.2.b Uniqueness

We will start by proving the uniqueness, which is actually simpler. This property is a natural consequence of the **strict convexity** of E (and also the convexity of the set X which is a vector space).

Suppose we have two given functions $u_1, u_2 \in X$ that are solutions to the problem (III.2). This means that, in particular, the infimum of E is finite. We will denote its value by

$$I_E = \inf_{v \in X} E(v),$$

and therefore we have, by assumption, that $E(u_1) = E(u_2) = I_E$.

We define $u = \frac{u_1+u_2}{2}$, which is indeed in X since this set is convex, and we calculate its energy by using the parallelogram law[2]

$$\begin{aligned} E(u) &= \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)+u_2'(x)}{2}\right|^2 dx - \int_0^1 f(x)\frac{u_1(x)+u_2(x)}{2}\,dx \\ &= \frac{k}{4}\int_0^1 |u_1'(x)|^2\,dx + \frac{k}{4}\int_0^1 |u_2'(x)|^2\,dx - \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx \\ &\quad - \frac{1}{2}\int_0^1 f(x)u_1(x)\,dx - \frac{1}{2}\int_0^1 f(x)u_2(x)\,dx \\ &= \frac{1}{2}E(u_1) + \frac{1}{2}E(u_2) - \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx. \end{aligned}$$

But, by assumption, u_1 and u_2 are solutions to the problem (III.2) so, we end up

$$E(u) = I_E - \frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx.$$

Moreover, by definition of the infimum, we have $E(u) \geq I_E$ (here we used the fact that $u \in X$). And since the stiffness constant k is positive, we deduce that we necessarily have

$$\frac{k}{2}\int_0^1 \left|\frac{u_1'(x)-u_2'(x)}{2}\right|^2 dx = 0.$$

Since the integrand function is non negative, we immediately obtain

$$\forall x \in [0,1],\ \ u_1'(x) = u_2'(x).$$

This implies that $u_1 - u_2$ is a constant function, but since $u_1(0) = u_2(0) = 0$ (by definition of the boundary conditions in X), we have at last shown that

$$\forall x \in [0,1],\ \ u_1(x) = u_2(x),$$

which proves the uniqueness of a *potential* solution of the problem (III.2).

I.2.c Characterization of the solution

In this paragraph, we will keep on assuming that the solution $u \in X$ to the problem (III.2) exists. We will show, under some hypotheses, that we can in fact characterize it by a partial differential equation (in 1D it will only concern one variable).

The method that will be presented here is a standard one in optimization and calculus of variations. It consists of establishing the **Euler-Lagrange** equations associated to the minimization problem (III.2).

The proof of this result is, ultimately, the infinite dimensional version ($\dim X = \infty$) of the following elementary yet fundamental result:

Lemma III.2

Let $\varphi : \mathbb{R} \mapsto \mathbb{R}$ be a differentiable function. Suppose that there exists $t^* \in \mathbb{R}$ such that

$$\varphi(t^*) = \inf_{t\in\mathbb{R}} \varphi(t), \tag{III.3}$$

therefore we have

$$\varphi'(t^*) = 0.$$

It does not seem useless to recall the proof of this lemma to understand how the hypothesis intervenes.

Proof :

From (III.3), for all $h > 0$ (the sign of h is crucial here !), we have

$$\varphi(t^* + h) \geq \varphi(t^*).$$

Since $h > 0$, we can deduce that

$$\frac{\varphi(t^*+h) - \varphi(t^*)}{h} \geq 0.$$

[2] Reminder : in a Hilbert space H, for any $a, b \in H$ we have $\|\frac{a+b}{2}\|^2 + \|\frac{a-b}{2}\|^2 = \frac{\|a\|^2}{2} + \frac{\|b\|^2}{2}$.

As $h \to 0^+$ in this equality, by definition of derivatives we get that

$$\varphi'(t^*) \geq 0.$$

If we do the same computation but this time with $h < 0$, we get

$$\varphi(t^* + h) \geq \varphi(t^*),$$

but this time (since $h < 0$!) we get

$$\frac{\varphi(t^* + h) - \varphi(t^*)}{h} \leq 0.$$

And as $h \to 0^-$, we obtain

$$\varphi'(t^*) \leq 0,$$

which gives the desired result. ■

Let us come back to our elastic string problem. Let us assume that a solution u of (III.2) exists. Let v be any function in X, in such a way that, for all $t \in \mathbb{R}$, we have $u + tv \in X$ (because X is a vector space !). So, by definition of the infimum, we have

$$\forall t \in \mathbb{R}, \quad E(u + tv) \geq E(u),$$

which shows that the function $\varphi_v : \mathbb{R} \mapsto \mathbb{R}$ defined by $\varphi_v(t) = E(u + tv)$ admits a minimum at point $t^* = 0$. Moreover, this function is differentiable (in fact we will see that it is a polynomial of degree 2 in the variable t). From the previous lemma, we deduce that $\varphi_v'(0) = 0$.

It remains to compute $\varphi_v'(0)$. To do this, we write φ_v in the following form

$$\varphi_v(t) = E(u) + t\left[k\int_0^1 u'(x)v'(x)\,dx - \int_0^1 f(x)v(x)\,dx\right] + \frac{t^2}{2}k\int_0^1 |v'(x)|^2\,dx, \tag{III.4}$$

and therefore, the relation $\varphi_v'(0) = 0$ amounts to writing that the coefficient of t in this polynomial is zero, meaning

$$k\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx.$$

Since this is true for any function $v \in X$, we have shown that if the solution of the problem (III.2) exists, it also verifies the following Euler-Lagrange equations:

$$\forall v \in X, \quad k\int_0^1 u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx. \tag{III.5}$$

Coming back to (III.4), we observe that **in this particular example** the converse is also true : if $u \in X$ verifies (III.5), then u is a solution of the problem (III.2) because the term in t^2 in (III.4) is always non negative.

BEWARE : this is not true in general because we know that the converse of Lemma III.2 is not true : if $\varphi'(t^*) = 0$, then t^* is not necessarily a local extremum of φ.

To summarize, we have shown the following result

Proposition III.3

A function $u \in X$ verifies (III.5) if and only if u is a solution of the problem (III.2).

Up to now, we did not need any particular hypothesis on the solution u of our problem. If we assume that the solution is a little bit more regular, then we can go further in the analysis.

Theorem III.4

We suppose that the problem (III.2) admits a solution $u \in X$.
Additionally, if we assume that $u \in \mathcal{C}^2([0,1])$ and $f \in \mathcal{C}^0([0,1])$, then u verifies the following boundary value problem :

$$\begin{cases} -k\partial_x^2 u = f, \ \text{in } (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{III.6}$$

Note that the converse is also true : any solution of (III.6) verifies (III.5) and therefore (III.2) also.

The problem (III.6) is called the *Poisson problem with homogeneous Dirichlet condition*.

Proof :

To prove the converse, it is sufficient to multiply the equation in (III.6) by a test function v and perform an integration by parts. The boundary terms are null since v is zero on the boundary.

To prove the direct implication, we carry out a simple integration by parts. For any $v \in X$, since u is of class $\mathcal{C}^2$, we can in fact integrate by parts the equation (III.5) and obtain

$$\int_0^1 (k\partial_x^2 u(x) + f(x))v(x)\,dx - \big[k(\partial_x u)v\big]_0^1 = 0.$$

We do not have any information on the values of $\partial_x u$ on $x = 0$ and $x = 1$, however we have $v(0) = v(1) = 0$, which shows that the last term in this formula is zero.

If we define $G(x) = k\partial_x^2 u(x) + f(x)$, we obtain

$$\forall v \in X, \quad \int_0^1 G(x)v(x)\,dx = 0.$$

Since X contains all $\mathcal{C}^\infty$ functions with compact support, the Du Bois-Reymond lemma (Lemma A.4) allows us to draw the conclusion $G = 0$. It is important to mention that we cannot take $v = G$ in the above integral because G is not in the space X (in particular, there is no reason for G to be zero on the boundary). Thus, (III.6) is proved. ■

Vocabulary

- (III.6) is a boundary value problem (a PDE + boundary conditions). Even if we are working with functions with one variable here, this problem **is not** a Cauchy problem !
- (III.5) is a variational formulation of this boundary value problem where u is the solution.
- The function v is called test function. In general, it belongs to the same space as u.

The previous computations we did in dimension 1 can also be carried out in any dimension d (i.e. in the case of the membrane). The formal boundary value problem that we get instead of (III.6)

$$\begin{cases} -k\Delta u = f, \text{ in } \Omega, \\ \qquad u = 0, \text{ on } \partial\Omega, \end{cases} \tag{III.7}$$

and the associated varational formulation is

$$\forall v \in X, \quad k\int_\Omega \nabla u(x)\cdot\nabla v(x)\,dx = \int_\Omega f(x)v(x)\,dx. \tag{III.8}$$

We will study this case more closely in Section IV.

Remark III.5

If we repeat the previous analysis assuming that the stiffness k of the string/membrane depends on the point x, then we obtain the following equations

$$-\partial_x(k(x)\partial_x u) = f(x), \quad \textit{in } (0,1),$$

in dimension 1, *and*

$$-\operatorname{div}(k(x)\nabla u) = f, \quad \textit{in } \Omega,$$

in dimension d. This is always accompanied by the boundary conditions that we did not rewrite here.

For the time being, we have shown that if the minimization problem of E on X has a solution, then the latter is unique and additionally, if it is regular then it is also a solution of the Poisson equation.

Conversely, any regular solution of the Poisson equation is a solution of the variational formulation and minimizes the energy on X.

We still have to show the existence of a minimizer, and this is where the choice of the function space X will be crucial.

I.3 How to prove the existence of a minimizer ?

The general principle for proving the existence of a minimizer for a functional E on a set X is the following :

1. We first need to prove that $\inf_X E > -\infty$. Otherwise, there is no need to start looking for a minimizer. Therefore, it is an essential step.
2. Then we consider a *minimizing sequence*, meaning a sequence $(u_n)_n \subset X$ that verifies $\lim_{n\to\infty} E(u_n) = \inf_X E$. Such a sequence **always** exists by simply using the definition of the infimum.
3. We will try to prove that this sequence (or one of its subsequences) converges (in a sense that we will try to specify: ideally for the natural norm on the space X). We denote by u the obtained limit.
4. We will try to prove that u is in fact the sought out minimum of E. In general, to do this, we will try to show that $E(u_n) \to E(u)$ and this will give us the desired result thanks to point 2. So we need some form of continuity of the functional E in X.

It should be noted that the first two points only depend on X as a set (and on E also) but the last two require a topology on it, since they deal with the convergence of sequences, the continuity of functions, etc ... So, to make all of this work, it is crucial to choose a good function space for X and a convenient topology. In fact, since the space X is, in general, of infinite dimension, the choice for a topology (even a n.v.s. topology) is not trivial at all.

Generally, to carry out the third step of the above list, we will try to show that the minimizing sequence is Cauchy, therefore convergent, **provided that the space X is complete**. This is why the completeness of the space is an important notion. Our first idea is to modify a little bit the space X defined above

$$X = \{v : [0,1] \to \mathbb{R}, \text{ of class } \mathcal{C}^1 \text{ and s.t. } v(0) = v(1) = 0\},$$

endowed with the norm $\|v\|_X = \|v\|_{L^\infty} + \|v'\|_{L^\infty}$.

We know that this space is a Banach, therefore, it might potentially be convenient for our needs.

Finiteness of the infimum Let us start by proving that the infimum is finite. To do this, we will use the following result

Lemma III.6

If $v \in X$, then we have

$$\|v\|_{L^\infty} \le \|v'\|_{L^2}.$$

Proof :

For $x \in [0,1]$, we write

$$v(x) = \underbrace{v(0)}_{=0, \text{since } v \in X} + \int_0^x v'(t)\,dt,$$

then we bound the integral by using Cauchy-Schwarz inequality. ∎

Therefore, for all $v \in X$,

$$E(v) = \frac{k}{2}\|v'\|_{L^2}^2 - \int_0^1 fv\,dx \ge \frac{k}{2}\|v'\|_{L^2}^2 - \|f\|_{L^1}\|v\|_{L^\infty} \ge \frac{k}{2}\|v'\|_{L^2}^2 - \|f\|_{L^1}\|v'\|_{L^2}.$$

But the (polynomial) function $y \mapsto \frac{k}{2}y^2 - \|f\|_{L^1}y$ is bounded from below on $\mathbb{R}$ by $-\frac{\|f\|_{L^1}^2}{2k}$.

The result is thus proved

$$\inf_X E \ge -\frac{\|f\|_{L^1}^2}{2k}.$$

Study of a minimizing sequence How can we show the convergence of a minimizing sequence in this space X ?

One of the few things that we know about the minimizing sequence, is that the sequence of real numbers $(E(u_n))_n$ is bounded and converges to the infimum. Is this enough to deduce that the sequence $(u_n)_n$ is bounded in X ?

In general, the answer to this question is no ! We can, for instance, easily construct a sequence $(u_n)_n$ of functions in X that is not bounded but such that $(E(u_n))_n$ is bounded (see TD).

We see that it is not obvious to extract a useful information on the sequence $(u_n)_n$ from the knowledge that we have on the sequence of energies $(E(u_n))_n$.

In optimization theory, we say that the functional E is not **coercive** on $(X, \|\cdot\|_X)$.

One could consider changing the norm on X to better fit the problem being studied. So, if we endow X with the following norm

$$\|u\|_{H^1} = \sqrt{\|u\|_{L^2}^2 + \|u'\|_{L^2}^2},$$

then it is obvious that E is coercive on $(X, \|\cdot\|_{H^1})$.

Proposition III.7

If $(u_n)_n$ is a sequence of elements of X such that $(E(u_n))_n$ is bounded, then $(u_n)_n$ is bounded in $(X, \|\cdot\|_{H^1})$.

Proof :

Define $C = \sup_n E(u_n) < +\infty$. By using the definition of the energy, we have

$$\frac{k}{2}\|u_n'\|_{L^2}^2 = E(u_n) + \int_0^1 f(x)u_n(x)\,dx \leq C + \|f\|_{L^1}\|u_n\|_{L^\infty} \leq C + \|f\|_{L^1}\|u_n'\|_{L^2}.$$

We now use the following Young's inequality[3]

$$\forall \varepsilon > 0, \forall a, b \in \mathbb{R}, \quad ab \leq \varepsilon a^2 + \frac{1}{4\varepsilon}b^2,$$

to bound the last term in the following way

$$\frac{k}{2}\|u_n'\|_{L^2}^2 \leq C + \underbrace{\frac{k}{4}}_{=\varepsilon}\|u_n'\|_{L^2}^2 + \frac{\|f\|_{L^1}^2}{k}.$$

Subsequently, we deduce that

$$\frac{k}{4}\|u_n'\|_{L^2}^2 \leq C + \frac{\|f\|_{L^1}^2}{k},$$

which gives a bound on $\|u_n'\|_{L^2}$. We now need a bound on $\|u_n\|_{L^2}$, but this can be obtained easily by using Lemma III.6

$$\|u_n\|_{L^2}^2 = \int_0^1 |u_n|^2\,dx \leq \|u_n\|_{L^\infty}^2 \leq \|u_n'\|_{L^2}^2.$$

■

So, this norm on X lets us prove that E is coercive and therefore deduce that all minimizing sequences are bounded for the norm $\|\cdot\|_{H^1}$. This is a first step towards the convergence of the sequence. We can even prove that this sequence is Cauchy for the norm $\|\cdot\|_{H^1}$ (see the proof below).

Unfortunately, by changing the norm on X, we have lost one important property ; the completeness. Therefore, there isn't much to deduce form the fact that the minimizing sequence is Cauchy.

Summary : Replacing X with its **completion** space for the norm $\|\cdot\|_{H^1}$ is a good choice for which the approach will work. This space is, *a priori*, an abstract space. We will see later in this chapter that we can construct it in a relatively explicit way.

II Sobolev spaces in dimension 1

II.1 The space $H^1(a, b)$

Although we are focusing for the moment on the one-dimensional case in order to understand the implemented ideas, it is important to remember that all the methods generalize to the multi-dimensional case, as we will see in the rest of the course. Besides, it is in spaces with multiple, even infinite, dimensions that these tools from function analysis are really useful and effective.

Still BEWARE : theoretical results on Sobolev spaces can vary depending on the dimension of the space.

[3] Do you know how to prove it ? it is a really good exercise and it only takes a couple of lines ...

From now on, we will only consider the dimension 1 and therefore work on a bounded and open interval $I = (a, b)$.

Definition III.8

$H^1(I)$ is called a **Sobolev space** *and is defined as the set of $L^2(I)$ functions u whose derivative in the sense of distributions is also a function in $L^2(I)$.*
We remind what this means : there exists a function $g \in L^2(I)$ (necessarily unique almost everywhere, by the du Bois-Reymond lemma) such that

$$\text{For any test function } \varphi \in \mathcal{C}_c^\infty(I), \text{ we have } \int_a^b u(x)\varphi'(x)\,dx = -\int_a^b g(x)\varphi(x)\,dx. \tag{III.9}$$

From now on, the function g is denoted by u' (or sometimes $\partial_x u$ or ∇u, analogous to the case of multiple dimensions).
We endow the vector space $H^1(I)$ with the norm

$$\|u\|_{H^1} = \sqrt{\|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2}.$$

The following "integration by parts" formula is crucial and should be remembered, it is in fact the **definition** of a weak derivative

$$\forall \varphi \in \mathcal{C}_c^\infty(I), \text{ we have } \int_a^b u(x)\varphi'(x)\,dx = -\int_a^b u'(x)\varphi(x)\,dx.$$

We will now establish that this definition generalizes the usual notion of derivatives.

Proposition III.9

If u is a $\mathcal{C}^1$ function on $\overline{I} = [a, b]$, then $u \in H^1(I)$, and its weak derivative u' coincide almost everywhere with the usual derivative (which justifies the notation).
Therefore, we get the following algebraic inclusion

$$\mathcal{C}^1(\overline{I}) \subset H^1(I),$$

but this injection is actually continuous, more precisely this means that

$$\forall u \in \mathcal{C}^1(\overline{I}),\ \ \|u\|_{H^1} \leq \sqrt{2(b-a)}\|u\|_{\mathcal{C}^1}.$$

Proof :

If u is a $\mathcal{C}^1$ function, the formula (III.9) is true if we take $g = u'$ (the usual derivative) by performing a simple integration by parts, the boundary terms will be null since φ has compact support.

To compare the norms, we simply use the fact that for any continuous function f on $[a, b]$, we have $\|f\|_{L^2} \leq \sqrt{b-a}\|f\|_{L^\infty}$. ∎

Theorem III.10 (Main properties of the space $H^1(I)$)

1. *The space $(H^1(I), \|.\|_{H^1})$ introduced in Definition III.8 is a Hilbert space.*
2. *Any function $u \in H^1(I)$ admits a representative (i.e. that is equal almost everywhere) that is continuous on $\overline{I}$, that we will always denote by u, and we have*

$$u(\beta) - u(\alpha) = \int_\alpha^\beta \partial_x u\,dx,\ \ \forall \alpha, \beta \in \overline{I}.$$

3. *The set $\mathcal{C}^\infty(\overline{I})$ is dense in $H^1(I)$.*

Note that the property 2 (the continuity of $H^1(I)$ functions) is not true in higher dimension.

Proof :

1. It is clear that the norm $\|\cdot\|_{H^1}$ is the norm associated to the H^1 scalar product defined by

$$(u, v)_{H^1} = (u, v)_{L^2} + (\partial_x u, \partial_x v)_{L^2} = \int_I uv\,dx + \int_I \partial_x u \partial_x v\,dx.$$

It remains to show the completeness of the space $H^1(I)$. So, let $(u_n)_n$ be a Cauchy sequence in $H^1(I)$. By definition of the H^1 norm, we deduce that $(u_n)_n$ and $(\partial_x u_n)_n$ are Cauchy sequences in the space $L^2(I)$. And since $L^2(I)$ is complete, there exist $u, g \in L^2(I)$ such that

$$u_n \to u, \text{ and } \partial_x u_n \to g.$$

For any test function φ, and for all n, we have

$$\int_I u_n(x)\varphi'(x)\,dx = -\int_I \partial_x u_n(x)\varphi(x)\,dx,$$

And it is clear that the convergence of the sequences that has been established above allows us to pass to the limit in this formula (φ is a fixed test function !). This proves that $u \in H^1(I)$ and that $\partial_x u = g$.

Therefore, the convergence of $(u_n)_n$ to u for the H^1 norm is immediate.

2. Let v be the function defined on I by the formula

$$v(x) = \int_a^x \partial_x u(t)\,dt, \quad \forall x \in I.$$

By the dominated convergence theorem, this function is continuous on $\overline{I}$. Let us verify that $v \in H^1(I)$ and $\partial_x v = \partial_x u$. To do this, we choose φ a test function and make the following computations by using Fubini's theorem

$$\int_a^b \left(\int_a^x \partial_x u(t)\,dt\right)\varphi'(x)\,dx = \int_a^b \left(\int_t^b \varphi'(x)\,dx\right)\partial_x u(t)\,dt = -\int_a^b \varphi(t)\partial_x u(t)\,dt.$$

So, $u - v$ is a function of $H^1(I)$ with a null derivative in the sense of distributions. From Proposition A.20, we can deduce the existence of a constant C such that $u = C + v$ almost everywhere.

This shows that $C + v$ is a continuous representative of u. By identifying u to this continuous representative and by using the formula that defines v, we obtain the desired result.

3. This result, that is quite technical, is accepted without proof. The proof is similar to the proof of the density of regular functions in $L^2(I)$: by extension then convolution.

■

Of course the construction of this space is justified by the fact (among others) that it allows to take into account more functions than the functions of class $\mathcal{C}^1$ while containing the differentiable functions in the usual sense.

Proposition III.11 (The relations between weak derivatives and classical derivatives)

Let $u \in H^1(I)$ and $J \subset I$ be an open interval.

1. *If the restriction $u_{|J}$ is a $\mathcal{C}^1$ function then the weak derivative of u on J coincide almost everywhere with the usual derivative of $u_{|J}$.*
2. *If the restriction of the weak derivative of u on J coincide almost everywhere with a $\mathcal{C}^k$ function ($k \geq 0$), then the restriction of u to the interval J is a $\mathcal{C}^{k+1}$ function whose usual derivative coincide with the weak derivative of u.*

Proof :

1. Take φ a test function whose support is compact and contained in J, and since $u_{|J}$ is $\mathcal{C}^1$, we can perform an integration by parts on J (without boundary terms)

$$\int_I u\varphi' = \int_J u\varphi' = -\int_J (u_{|J})'\varphi = -\int_I (u_{|J})'\varphi.$$

But, by definition of the weak derivative of u we also have

$$\int_I u\varphi' = -\int_I (\partial_x u)\varphi.$$

So we have established that

$$\int_I (u'_{|J} - \partial_x u)\varphi = 0,$$

for any test function whose support is contained in J. From Lemma A.4, we deduce that $\partial_x u = u'_{|J}$ almost everywhere on J.

2. Fix any point $y_0 \in J$. By using the third property of Theorem III.10, we can write

$$u(y) = u(y_0) + \int_{y_0}^{y} (\partial_x u)\, dt.$$

From our assumption, $\partial_x u$ is $\mathcal{C}^k$ on J and u is one of its antiderivatives on J. The result is thus proved.

■

Examples : Take $I = (-1, 1)$.

- The function $x \mapsto u(x) = |x|$ is in $H^1(I)$ and $\partial_x u$ is the Heaviside function $\mathcal{H}$ defined by

$$\mathcal{H}(x) = \begin{cases} -1, & \text{on } (-1, 0), \\ 1, & \text{on } (0, 1). \end{cases}$$

- The Heaviside function $\mathcal{H}$ does not belong in $H^1(I)$ since it does not admit a continuous representative (see Theorem III.10). In fact, the derivative of $\mathcal{H}$ in the sense of distributions is a Dirac delta function

$$\partial_x \mathcal{H} = 2\delta_0.$$

Indeed, a simple computation shows

$$\int_{-1}^{1} H(x)\varphi'(x)\, dx = -2\varphi(0) = -2\langle \delta_0, \varphi \rangle_{\mathcal{D}',\mathcal{D}}.$$

It is important to convince oneself that the distribution δ_0 cannot be written as an element of the space $L^2(I)$: in other words, there is no function $g \in L^2(I)$ that verifies the property

$$\varphi(0) = \int_{-1}^{1} g\varphi\, dx, \quad \forall \varphi \in \mathcal{C}_c^\infty((-1, 1)).$$

- The function $x \mapsto u(x) = |x|^\alpha$ is in $H^1(I)$ if and only if $\alpha > 1/2$ and in that case we have $\partial_x u(x) = \alpha |x|^{\alpha-2} x$, for $x \neq 0$.

Corollary III.12

The canonical injection map (i.e. the identity map !) from $H^1(I)$ to $\mathcal{C}^0(\overline{I})$ is continuous and more precisely, we have

$$\|u\|_{L^\infty} \leq \left(\frac{1}{|b-a|} + |b-a| \right) \|u\|_{H^1}, \quad \forall u \in H^1(I).$$

Subsequently : if $(u_n)_n \subset H^1(I)$ verifies $u_n \xrightarrow[n\to\infty]{} u$ in $H^1(I)$, then $(u_n)_n$ uniformly converges towards u on $\overline{I}$.

Proof :

From the Cauchy-Schwarz inequality, for all x, y we have

$$|u(y)| \leq |u(x)| + |b-a|^{\frac{1}{2}} \|\partial_x u\|_{L^2}.$$

If we integrate this with respect to x, we find

$$|b-a||u(y)| \leq \int_I |u(x)|\, dx + |b-a|^{\frac{3}{2}} \|\partial_x u\|_{L^2}.$$

Once again, by using Cauchy-Schwarz we get

$$|b-a||u(y)| \leq |b-a|^{\frac{1}{2}} \|u\|_{L^2} + |b-a|^{\frac{3}{2}} \|\partial_x u\|_{L^2}.$$

Finally, by taking the sup with respect to y, we find

$$\|u\|_{\mathcal{C}^0(\overline{I})} \leq |b-a|^{-\frac{1}{2}} \|u\|_{L^2} + |b-a|^{\frac{1}{2}} \|\partial_x u\|_{L^2},$$

which allows us to conclude. ■

For example, we can prove the following properties (see exercise sheets).

Proposition III.13

- *The product of two functions u, v in $H^1(I)$ is still in $H^1(I)$ and we have the following equality almost everywhere*
$$\partial_x(uv) = u(\partial_x v) + (\partial_x u)v.$$
- *More generally, if $u \in H^1(I)$ and $F \in \mathcal{C}^1(\mathbb{R})$, then $F(u) \in H^1(I)$ and we have the following equality almost everywhere*
$$\partial_x(F(u)) = F'(u)\partial_x u.$$

II.2 The space $H^1_0(I)$

Definition III.14

We denote by $H^1_0(I)$, the closed subspace of $H^1(I)$ of functions that are zero on the boundary of I.

This definition makes sense because we have seen that the functions in $H^1(I)$ have a unique continuous representative on $\overline{I}$, which legitimizes the notion of "boundary value" for functions *a priori* defined only almost everywhere.

This space is closed since it is the kernel of the **continuous** map (by using Corollary III.12) defined by

$$u \in H^1(I) \longmapsto \begin{pmatrix} u(a) \\ u(b) \end{pmatrix} \in \mathbb{R}^2.$$

The following is a fundamental result of this theory. It is related to Lemma III.6.

Proposition III.15 (Poincaré inequality)

For any $u \in H^1_0(I)$, we have the following inequality
$$\|u\|_{L^2} \leq |b-a| \|\partial_x u\|_{L^2}.$$

Corollary III.16

The map $u \mapsto \|\partial_x u\|_{L^2}$ is a norm on $H^1_0(I)$ that is equivalent to the norm of $H^1(I)$. The corresponding Hilbert structure is equivalent to to the structure inherited from H^1.

Notation

For $u \in H^1(I)$, we denote
$$|u|_{H^1} \stackrel{\text{def}}{=} \|\partial_x u\|_{L^2},$$
and we call it the H^1 **semi-norm**. The previous corollary says that, on $H^1_0(I)$, this semi-norm is in fact a norm that is equivalent to the standard norm of H^1.

The proof of the corollary is immediate. Let us prove the Poincaré inequality:

Proof :

Since $u(a) = u(b) = 0$, we can apply the third property of Theorem III.10, and obtain that for all $x \in I$, we have

$$|u(x)| \leq \int_a^x |\partial_x u| \, dt \leq |b-a|^{\frac{1}{2}} \|\partial_x u\|_{L^2}.$$

By squaring each side of the inequality, and then by integrating on the interval I, we get

$$\int_a^b |u(x)|^2 \, dx \leq |b-a|^2 \|\partial_x u\|^2_{L^2},$$

which gives the desired result. Note that we only used the fact that $u(a) = 0$. ■

Remark III.17

The constant $|b-a|$ that appears in the Poincaré inequality above is not optimal. We can show that the optimal value (i.e. the smallest) for this constant is $\frac{|b-a|}{\pi}$.

Finally, we get the following result that we will take for granted.

Theorem III.18

The set of functions $\mathcal{C}_c^\infty(I)$ is dense in $H_0^1(I)$.

This result proves that $H_0^1(I)$ is indeed **the** completion space, for the H^1 norm, of $X = \{v \in \mathcal{C}^1([0,1],\mathbb{R}), v(0) = v(1)\}$ that we introduced above. Indeed, we have the inclusions

$$\mathcal{C}_c^\infty(I) \subset X \subset H_0^1(I),$$

and therefore the above proposition establishes the density of the space X in $H_0^1(I)$ for the H^1 norm. Since this space is complete, it is the unique completion (up to an isomorphism) of X for this norm.

Therefore, the space $H_0^1(I)$ seems to be a good candidate for our analysis.

II.3 Solving the variational problem for the elastic string

Let us come back to the problem that has motivated the theory of Sobolev spaces. We now understand that it is better to work on the $H_0^1(I)$ space endowed with its natural norm than on the (smaller) space X that was introduced at the beginning. Therefore, we can reformulate the problem in the following way:

Let $f \in L^2(I)$ (or $f \in \mathcal{C}^0(\bar{I})$), find a function $u \in H_0^1(I)$ that verifies

$$E(u) = \inf_{v \in H_0^1(I)} E(v).$$

$H_0^1(I)$ is a vector space and in these conditions, the proof of the uniqueness of a potential minimum that we have already done is still valid.

Let us now prove the existence of a minimum. To do this, we will start by proving that E is bounded from below on the space $H_0^1(I)$. In fact, for all $v \in H_0^1(I)$, we have

$$\begin{aligned}\int_I fv\,dx &\leq \|f\|_{L^2}\|v\|_{L^2}, \quad \text{by Cauchy-Schwarz inequality}\\ &\leq |b-a|\|f\|_{L^2}\|\partial_x v\|_{L^2}, \quad \text{by Poincaré inequality, since } v \in H_0^1(I)\\ &\leq \frac{1}{2k}|b-a|^2\|f\|_{L^2}^2 + \frac{k}{2}\|\partial_x v\|_{L^2}^2, \quad \text{by Young's inequality.}\end{aligned}$$

All of this shows that

$$E(v) \geq -\frac{|b-a|^2}{2k}\|f\|_{L^2}^2.$$

Since E is bounded from below, its infimum is finite and by definition, there exists a minimizing sequence, meaning a sequence of functions $(u_n)_n$ in $H_0^1(I)$ such that the sequence of real numbers $(E(u_n))_n$ converges towards $\inf_{H_0^1} E$.

Now, we must show that the sequence $(u_n)_n$ converges. To do this, we will exploit the completeness of $H_0^1(I)$ (and finally harvest the fruits of our labor !) to directly show that $(u_n)_n$ is a Cauchy sequence in $H_0^1(I)$.

We denote $I_E = \inf_{v \in H_0^1(I)} E(v)$. In the proof of the uniqueness property (by using the parallelogram law), we have shown

$$I_E \leq E\left(\frac{u_n + u_{n+p}}{2}\right) = \frac{1}{2}E(u_n) + \frac{1}{2}E(u_{n+p}) - \frac{k}{8}\|\partial_x u_n - \partial_x u_{n+p}\|_{L^2}^2,$$

from this we deduce

$$\frac{k}{8}\|\partial_x u_n - \partial_x u_{n+p}\|_{L^2}^2 \leq \frac{1}{2}E(u_n) + \frac{1}{2}E(u_{n+p}) - I_E. \tag{III.10}$$

Let $\varepsilon > 0$, since $E(u_n)$ tends towards I_E on $H_0^1(I)$, there exists $n_0 \geq 0$ such that

$$\forall n \geq n_0, \ \ |E(u_n) - I_E| \leq \varepsilon.$$

So, for $n \geq n_0$ and $p \geq 0$, the inequality (III.10) gives us

$$\frac{k}{8}\|\partial_x u_n - \partial_x u_{n+p}\|_{L^2}^2 \leq \varepsilon.$$

Alternatively, this can be written as

$$|u_n - u_{n+p}|_{H^1}^2 \leq \frac{8\varepsilon}{k}, \qquad \forall n \geq n_0, \forall p \geq 0.$$

Since the H^1 semi-norm is a norm on $H_0^1(I)$ that is equivalent to the H^1 norm, this shows that $(u_n)_n$ is a Cauchy sequence in $H_0^1(I)$, so by completeness of this space, it converges to a certain function $u \in H_0^1(I)$.

Then, we prove that $E(u) = \lim_{n\to\infty} E(u_n)$ by passing to the limit in all the terms of $E(u_n)$ (for the first one, it is a simple consequence of the convergence in H^1 and for the second one, it is because of the convergence in L^2). In other words, E is continuous on H_0^1.

Finally, since $(u_n)_n$ is a minimizing sequence, we have $E(u) = I_E$.

We can now reformulate the Euler-Lagrange equations associated to this minimizing problem in the exact same way we did in Section I.2.c and obtain

$$u \in H_0^1(I) \text{ and verify } \forall v \in H_0^1(I), \quad \int_I k\partial_x u \partial_x v \, dx = \int_I f v \, dx. \tag{III.11}$$

Note that this variational formulation admits a unique solution. In fact, if u_1 and u_2 are two solutions, then their difference $\bar{u} \in H_0^1(I)$ satisfies the same formulation without source terms. By taking $v = \bar{u}$ in the formulation, we obtain

$$k\int_I |\partial_x \bar{u}|^2 \, dx = 0,$$

which proves that $\bar{u}$ is a constant that can only be 0 since $\bar{u}$ is equal to zero on the boundary.

Regularity of the solution: In the case of dimension 1, it is now easy to prove that the solution $u \in H_0^1(I)$ is sufficiently regular and verifies the Poisson problem in the usual sense.

Indeed, since $\mathcal{C}_c^\infty(I) \subset H_0^1(I)$, we can consider all elements of $\mathcal{C}_c^\infty(I)$ as test functions in the Euler-Lagrange equations (III.11) to obtain

$$\forall \varphi \in \mathcal{C}_c^\infty(I), \quad \int_0^1 (\partial_x u)\partial_x \varphi \, dx = \int_0^1 \frac{f}{k}\varphi \, dx.$$

By definition, this shows that $\partial_x u$ is itself an element of the Sobolev space $H^1(I)$ that has $-\frac{1}{k}f$ as a weak derivative in L^2. Therefore we have

$$\partial_x^2 u = \partial_x(\partial_x u) = -\frac{f}{k},$$

which shows that u is a weak solution of the partial differential equation $-k\partial_x^2 u = f$. The boundary conditions are also verified (u is in $H_0^1(I)$ and therefore vanishes on the boundary).

We can say much more on the regularity of u. Indeed, since $\partial_x u \in H^1(I)$, Theorem III.10 tells us that $\partial_x u \in \mathcal{C}^0(\overline{I})$ (or more precisely, that it admits a continuous representative). So, Proposition III.11 shows that u is a $\mathcal{C}^1$ function on $\overline{I}$.

Remark III.19

The solution u thus obtained belongs to the space X defined at the beginning, and since $X \subset H_0^1(I)$, the function u is also the minimum of the functional E on the space X:

$$E(u) = \inf_{v\in X} E(v).$$

However, although u also realizes the minimum of E on X, one must understand that this property could not have been established directly by working on the space X and that the introduction of Sobolev spaces was absolutely crucial to solve this problem !

Up until now, we only used the fact that $f \in L^2(I)$. If we now make the additional assumption that f is a continuous function on $\overline{I}$, we can go further in our analysis

$$\partial_x^2 u = -\frac{f}{k} \in \mathcal{C}^0(\overline{I}).$$

Once again, Proposition III.11 shows that $\partial_x u$ is a $\mathcal{C}^1$ function and therefore u is a $\mathcal{C}^2$ function. So, the function u solves the following partial differential equation, in the classical sense:

$$\begin{cases} -k\partial_x^2 u = f, & \text{in } I \\ u(0) = u(1) = 0. \end{cases}$$

III Variational formulation of a linear boundary value problem. Lax-Milgram theorem.

We have seen on the example of the elastic string that, starting from an optimization problem, we can write the Euler-Lagrange equations which characterize the critical points of the functional being studied. If these equations have a solution u, then we can try to deduce the partial differential equation verified by the solution u and the boundary conditions. This allows us to construct a solution of a boundary value problem.

In a more general way, we will now try to reverse the process and establish the existence and uniqueness of solutions of boundary value problems via a "variational formulation" even if we will see later that it does not necessarily come from a calculus of variations problem.

III.1 General principle

Consider the following **linear** boundary value problem

$$\begin{cases} Au = f, \ \text{ on a domain } \Omega \text{ of } \mathbb{R}^d, \\ \quad + \text{boundary conditions on } u \text{ and its derivatives ...,} \end{cases} \tag{III.12}$$

where A is a linear differential operator (i.e. Au is a map from Ω to $\mathbb{R}$ that is made up of some derivatives of u), f is a given source term.

Let X be a function space in which we will look for a solution u. Also, let V be a space of test functions. We multiply the equation $Au = f$ by an element of v and then integrate it on the domain Ω. So, any potential solution $u \in X$ will verify

$$\int_\Omega Au.v\, dx = \int_\Omega fv\, dx, \ \ \forall v \in V. \tag{III.13}$$

We have seen that, if V is sufficiently large (if it contains $\mathcal{C}_c^\infty(I)$ functions for example) then the du Bois-Reymond lemma (Lemma A.4) allows us to go back to the formulation (III.12) from (III.13).

Now, we perform an integration by parts a certain number of times on the left hand side of the equation (or not, many choices are possible here) to put it in the following form

$$\int_\Omega (B_1 u).(B_2 v)\, dx = \int_\Omega fv\, dx + \cdots \text{potential boundary terms}, \ \ \forall v \in V. \tag{III.14}$$

The test function v will carry some of the derivatives and the rest will stay on the unknown function u. At this stage, everything is formal: we do not yet ask ourselves whether u exists or not, nor whether it is regular... in short, all computations are authorized here.

But of course, one is not advised to do whatever it wants to since the goal is to make sure that the equations we obtain after these manipulations have a unique solution and subsequently show that the initial problem has been solved. Therefore, the following objects must be well chosen :

- The space of solutions X.
- The space of test functions V.
- The exact form of the integration by parts, meaning the operators B_1 and B_2 that have been obtained from A.

Note that a boundary value problem can have different variational formulations of the form (III.14). But, later we will see that there is often a *natural* choice.

To choose these objects, it is necessary to hold on to an abstract result that allows us to ensure the existence of solutions to the variational problem (III.14). We observe that the equations (III.14) have two types of terms : bilinear terms in (u, v), they are the main terms and they are usually put on the left hand side of the equation, and linear terms in v that do not contain u, they correspond to the source terms' contribution and potential boundary terms, and they are usually put on the right hand side of the equation.

Here is the basic result of this theory. You have surely encountered it in your Functional Analysis course.

Theorem III.20 (Lax-Milgram)

Let H be a real Hilbert space. Let a be a bilinear form on H and L a linear form on H. Suppose that

1. *a is continuous :*
$$|a(u,v)| \leq \|a\|\|u\|_H\|v\|_H, \ \ \forall u,v \in H.$$
2. *a is coercive : there exists $\alpha > 0$ such that*
$$a(u,u) \geq \alpha\|u\|_H^2, \ \ \forall u \in H.$$
3. *L is continuous :*
$$|L(u)| \leq \|L\|\|u\|_H, \ \ \forall u \in H.$$

Then, there **exists** *a* **unique** *element u in H that verifies*

$$a(u,v) = L(v), \ \ \forall v \in H, \tag{III.15}$$

and

$$\|u\|_H \leq \frac{\|L\|}{\alpha}.$$

Moreover, if a is symmetric, then u is also the unique element of H that minimizes the functional

$$J(v) = \frac{1}{2}a(v,v) - L(v).$$

This theorem gives sufficient conditions for solving a set of equations of the form (III.15). But let us insist on the fact that these are not sufficient conditions. There is a result of the same type giving necessary and sufficient conditions of existence and uniqueness (including in Banach spaces) but it is out of the scope of this course.

If we want to apply the Lax-Milgram theorem to the abstract framework previously introduced, then we must :

- Choose the same spaces $X = V$ and make sure it is a Hilbert space, we will denote it by H from now on.
- Define (let us forget for the moment the boundary terms, which, depending on the cases, can either be in a or in L)
$$a(u,v) = \int_\Omega (B_1u).(B_2v)\,dx, \ \ \forall u,v \in H,$$
$$L(v) = \int_\Omega fv\,dx, \ \ \forall u,v \in H.$$
- To make sure that these are well defined and continuous on H, typically B_1 and B_2 must be well defined and continuous on H with values in $L^2(\Omega)$. If B_1 and B_2 are first order differential operators, then H should contain at least the Sobolev space H^1.
- To ensure the coercivity of a, we must have
$$a(u,u) = \int_\Omega (B_1u).(B_2u)\,dx \geq \alpha\|u\|_H^2,$$
this shows that the norm on the space H must be bounded by the quantity $a(u,u)$ so this norm must not contain more derivatives than B_1 and B_2 contain.

In summary, if we restrict ourselves to second order boundary value problems in dimension 1 for the moment, we see that we can systematically choose for H a closed subspace of the space $H^1(I)$. This one will depend on the chosen boundary conditions. The fact that it is closed is necessary for it to be a Hilbert space.

III.2 Examples in dimension 1

III.2.a Poisson problem with Dirichlet conditions

Let $I = [0,1]$. Let $k : I \to \mathbb{R}$ be a measurable, non negative and bounded function such that $\alpha = \inf_I k > 0$. And let $f \in L^2(I)$ be a source term. Consider the following boundary value problem

$$\begin{cases} -\partial_x(k(x)\partial_x u) = f(x), & x \in I, \\ u(0) = u(1) = 0. \end{cases} \tag{III.16}$$

It is non other than the elastic string problem that we have studied before, we will study it once more in light of what we have discovered.

The adequate functional framework is $H = H_0^1(I)$. We now define

$$a(u,v) = \int_I k(x)(\partial_x u)(\partial_x v)\,dx,$$

and

$$L(v) = \int_I f(x)v\,dx.$$

The space H is complete, the bilinear form a is continuous on H since k is bounded and the Cauchy-Schwarz inequality gives us

$$|a(u,v)| \leq \|k\|_\infty \|\partial_x u\|_{L^2}\|\partial_x v\|_{L^2} \leq \|k\|_\infty \|u\|_H \|v\|_H.$$

Similarly, we also have

$$|L(v)| \leq \|f\|_{L^2}\|v\|_{L^2} \leq \|f\|_{L^2}\|v\|_H.$$

To prove that a is coercive, we use the assumptions we made on k to obtain

$$a(u,u) = \int_I k|\partial_x u|^2\,dx \geq \alpha\|\partial_x u\|_{L^2}^2,$$

and by using Poincaré inequality (Proposition III.15 and its Corollary III.16), we get

$$a(u,u) \geq C\|u\|_H^2.$$

We are in the correct framework to apply the Lax-Milgram theorem, therefore we deduce the existence and uniqueness of a solution $u \in H_0^1(I)$ of the variational formulation of (III.16) given by

$$\int_I k(x)(\partial_x u)(\partial_x v)\,dx = \int_I f(x)v\,dx, \quad \forall v \in H_0^1(I).$$

Non homogeneous boundary conditions Let us now try to solve the same problem but this time with non zero boundary conditions. The boundary conditions are of the same nature as the previous ones (Dirichlet) but this time they are non homogeneous. Therefore, let $u_0, u_1 \in \mathbb{R}$ and consider the problem

$$\begin{cases} -\partial_x(k(x)\partial_x u) = f(x), & x \in I, \\ u(0) = u_0, \\ u(1) = u_1. \end{cases} \tag{III.17}$$

This time, we should not be looking for a solution $u \in H_0^1(I)$ since it would mean that u is zero on the boundary. We will therefore *lift* the boundary conditions in the following way:

Take any sufficiently regular function $R : I \to \mathbb{R}$ (as a *lift*), that verifies

$$R(0) = u_0, \;\; R(1) = u_1.$$

Now the idea is to look for u in the form $u = R + w$ where w is zero on the boundary, thus w satisfies the following equation

$$\begin{cases} -\partial_x(k(x)\partial_x w) = f + \partial_x(k(x)\partial_x R), & x \in I, \\ w(0) = w(1) = 0. \end{cases}$$

It is therefore a Poisson problem with homogeneous conditions and a new source term

$$\tilde{f} = f + \partial_x(k(x)\partial_x R).$$

Since R is regular, $\partial_x R$ is also regular.

- If k is regular, the function $\tilde{f}$ is well defined and belongs to the L^2 space therefore the existence and uniqueness of a solution $w \in H_0^1$ of the above problem is a consequence of the analysis of the homogeneous case. So, we can now deduce the existence and uniqueness of a solution $u \in H^1$ of the initial problem.
- But if we have no regularity on k (if k is discontinuous for example), the function $\tilde{f}$ might not be well defined ... so the idea is to integrate by parts the inconvenient term in the definition of the linear form L.

More precisely, instead of $L(v) = \int_I \tilde{f} v \, dx$, we will rather define L by

$$L(v) = \int_I f v \, dx - \int_I k(x) \partial_x R \partial_x v \, dx.$$

We can easily verify that L is linear and continuous on H so we can apply the Lax-Milgram theorem once again and obtain the existence and uniqueness of a solution $w \in H_0^1$ of the problem

$$\int_I k(x) \partial_x w \partial_x v \, dx = \int_I f v \, dx - \int_I k(x) \partial_x R \partial_x v \, dx, \ \ \forall v \in H_0^1,$$

which can also be written as

$$\int_I k(x) \partial_x (w + R) \partial_x v \, dx = \int_I f v \, dx, \ \ \forall v \in H_0^1,$$

which gives the desired solution $u = w + R$.

III.2.b Adding a linear reaction term

Under the same assumptions as the previous case, we now consider the following problem

$$\begin{cases} -\partial_x (k(x) \partial_x u) + \gamma u = f(x), & x \in I, \\ u(0) = u(1) = 0, \end{cases} \tag{III.18}$$

where $\gamma \in \mathbb{R}$. We can reproduce the previous formalism by defining $H = H_0^1(I)$,

$$a(u, v) = \int_I k(x) (\partial_x u)(\partial_x v) \, dx + \gamma \int_I uv \, dx,$$

$$L(v) = \int_I f v \, dx.$$

All the hypotheses of the Lax-Milgram theorem can easily be verified, except maybe the coercivity hypothesis on a. In fact, we have

$$a(u, u) = \int_I k |\partial_x u|^2 \, dx + \gamma \int_I |u|^2 \, dx.$$

- If $\gamma \geq 0$, then the second term is non negative so we can bound $a(u, u)$ from below by $\|u\|_H^2$ up to a multiplicative constant.
- If $\gamma < 0$, then we don't have much control on the second term.

  But if γ is "too" negative then a will not be coercive. In fact, for any non zero $\bar{u} \in H_0^1(I)$, we can always find $\gamma < 0$ such that

  $$a(\bar{u}, \bar{u}) = \int_I k |\partial_x \bar{u}|^2 \, dx + \gamma \int_I |\bar{u}|^2 \, dx < 0,$$

  which implies that a is not coercive.

  There is a real obstacle here, in fact we can prove that[4], for instance, there is no solution to the following boundary value problem

  $$\begin{cases} -\partial_x^2 u - \pi^2 u = \sin(\pi x), & x \in (0, 1), \\ u(0) = u(1) = 0. \end{cases}$$

III.2.c Convection-diffusion problem with Dirichlet conditions

Consider now the following problem

$$\begin{cases} -\partial_x^2 u + \partial_x u = f(x), & x \in I, \\ u(0) = u(1) = 0. \end{cases} \tag{III.19}$$

Compared to the Poisson problem that we have seen before, we have added here a first order term known as the *convection* or *transport* term. But since its order is inferior to that of the principal term, we can still work on the functional space $H = H_0^1(I)$ and adopt the following weak formulation

$$a(u, v) = \int_0^1 (\partial_x u)(\partial_x v) \, dx + \int_0^1 (\partial_x u) v \, dx,$$

[4] which is not too hard to do by the way ...

$$L(v) = \int_0^1 f v \, dx.$$

Once again, the continuity of a and L on H is easy to demonstrate; so let us now prove that a is coercive. For $u \in H$, we have

$$a(u,u) = \int_0^1 |\partial_x u|^2 \, dx + \int_0^1 (\partial_x u) u \, dx.$$

We cannot say much beforehand on the sign of the second term, which can cause problems. But in fact, we will see that it is zero. In fact, we know (see TD) that $u^2 \in H_0^1(I)$ and that $\partial_x(u^2) = 2u\partial_x u$ in such a way that we get

$$\int_0^1 (\partial_x u) u = \frac{1}{2}\int_0^1 \partial_x(u^2) \, dx = \frac{1}{2}u^2(1) - \frac{1}{2}u^2(0) = 0.$$

It follows that a is coercive, we can therefore apply the Lax-Milgram theorem. The rest is the same as in the case of Dirichlet's problem: take $v = \varphi \in \mathcal{C}_c^\infty(I)$ in the weak formulation

$$\int_0^1 \partial_x u \partial_x \varphi \, dx = \int_0^1 (f - \partial_x u)\varphi \, dx.$$

Since $f \in L^2$ and $\partial_x u \in L^2$, we can deduce that $\partial_x u$ is indeed in $H^1(I)$ and that it verifies the following in the weak sense

$$\partial_x(\partial_x u) = -(f - \partial_x u).$$

If we now assume that $f \in \mathcal{C}^0(\overline{I})$, knowing that $\partial_x u \in \mathcal{C}^0(\overline{I})$ by using Corollary III.12, then we obtain that $\partial_x(\partial_x u)$ is continuous therefore $\partial_x u \in \mathcal{C}^1(\overline{I})$ and $u \in \mathcal{C}^2(\overline{I})$. So the function u thus constructed is indeed a solution of the problem (III.19) in the usual sense.

Remark III.21

In this case, the bilinear form a is not symmetric on H (see for yourself why this is true !) and therefore, although we are still talking about variational methods, it is important to understand that this problem does not arise from a minimization problem of a functional !

In the previous analysis, we took advantage of the cancellation of the contribution of the new term in the coercivity estimate. Unfortunately, we will see that this is not always true. For example, consider the following problem

$$\begin{cases} -\partial_x^2 u + b(x)\partial_x u = f(x), \quad x \in I, \\ u(0) = u(1) = 0, \end{cases} \tag{III.20}$$

where $b : \overline{I} \to \mathbb{R}$ is a given regular function. The bilinear form associated to the problem can be written as (the space H and the linear form L stay unchanged)

$$a(u,v) = \int_0^1 (\partial_x u)(\partial_x v) \, dx + \int_0^1 b(x)(\partial_x u) v \, dx,$$

and the coercivity computation becomes

$$a(u,u) = \int_0^1 |\partial_x u|^2 \, dx + \int_0^1 b(x)(\partial_x u) u \, dx.$$

We can, for instance, integrate by parts the second term and obtain

$$\int_0^1 b(x)(\partial_x u) u \, dx = \frac{1}{2}\int_0^1 b(x)\partial_x(u^2) \, dx = -\frac{1}{2}\int_0^1 b'(x)u^2 \, dx.$$

We observe that this term is no longer zero which can cause problems. Let us look at some examples:

- If b is non increasing, then the term in question is non negative and therefore we have

$$a(u,u) \geq \int_0^1 |\partial_x u|^2 \, dx \geq C\|u\|_H^2,$$

so we can apply the Lax-Milgram theorem.

- If b is of the form $b(x) = -\gamma x$ then we have

$$a(u,u) = \int_0^1 |\partial_x u|^2 \, dx + \frac{\gamma}{2} \int_0^1 u^2 \, dx.$$

  Just like we saw before, if γ is sufficiently negative, then this quantity can become negative and therefore we cannot apply the Lax-Milgram theorem.

- Suppose now that b is "small" in a sense that will later be specified. By using Cauchy-Schwarz inequality, we can write

$$\left| \int_0^1 b(x)(\partial_x u)u \, dx \right| \leq \|b\|_\infty \|\partial_x u\|_{L^2} \|u\|_{L^2}$$

  and then by using Poincaré inequality (Proposition III.15)

$$\left| \int_0^1 b(x)(\partial_x u)u \, dx \right| \leq \|b\|_\infty \|\partial_x u\|_{L^2}^2,$$

  in such a way that

$$a(u,u) \geq (1 - \|b\|_\infty)\|\partial_x u\|_{L^2}^2.$$

  So, we have just showed that if $\|b\|_\infty < 1$ then the bilinear form a is coercive and therefore the rest follows as usual.

- Note that the two cases in which we were able to apply the Lax-Milgram theorem are completely different:
  - Case 1 : b is non increasing but can be as large as desired.
  - Case 2 : b is small but with no assumption on its monotony.

  Let us conclude this discussion by mentioning that, although the Lax-Milgram theorem does not always apply for solving this problem, it is possible to show the existence of a solution for any function b but this uses other techniques which are beyond the scope of this course. The only thing to remember is that the Lax-Milgram theorem, while very useful, does not work every time.

III.2.d Neumann boundary conditions

For Neumann conditions, we impose the values on the derivatives of the solution on the boundary of the domain, instead of imposing the values on the solution itself. So in the elastic string model for instance, if we impose $\partial_x u(0) = 0$, it means that the tension forces at the left end of the string are zero, or even that the string is fixed on a vertical axis on which it can slide freely (without friction).

Mixed Dirichlet-Neumann case Let us now try to solve the following problem

$$\begin{cases} -\partial_x^2 u = f(x), \quad x \in I, \\ \partial_x u(0) = 0 \\ u(1) = 0. \end{cases} \tag{III.21}$$

Let us assume that there exists a solution u sufficiently regular and let v be a test function (also sufficiently regular but let us not specify for the moment the space in which it lives). We multiply the equation by v, integrate it and try to put it in a "Lax-Milgram" form.

$$\begin{aligned} \int_0^1 f v \, dx &= \int_0^1 (-\partial_x^2 u) v \, dx \\ &= [-(\partial_x u)v]_0^1 + \int_0^1 (\partial_x u)(\partial_x v) \, dx \\ &= (\partial_x u)(0)v(0) - (\partial_x u)(1)v(1) + \int_0^1 (\partial_x u)(\partial_x v) \, dx. \end{aligned}$$

Let us observe the boundary terms. The first one must be zero because we demand that $\partial_x u(0) = 0$ (so the value of $v(0)$ has no importance in this computation). The second one has no reason to be zero. On the other hand, remember that, in order to make the Lax-Milgram theorem work, the solution u and the test functions v must live in the same space. Thus, since we want $u(1) = 0$ (and this condition will be directly taken into account in the definition of the function space) we will also restrict ourselves to test functions verifying $v(1) = 0$, which will therefore make the second boundary term *disappear*.

To summarize: we will consider

$$H = \{v \in H^1(I),\ \ v(1) = 0\},$$
$$a(u,v) = \int_0^1 (\partial_x u)(\partial_x v)\, dx,\ \ \forall u, v \in H,$$
$$L(v) = \int_0^1 f v\, dx.$$

Beware : It might seem that a and L are the same (bi-)linear forms as the ones in for the Dirichlet conditions but this is not the case because they are not defined on the same spaces.

We observe that H is a closed subspace of H^1 (hence a Hilbert), and that a and L are continuous on H. Once again it remains to study the coercivity of a. As before we have

$$a(u,u) = \int_0^1 |\partial_x u|^2\, dx,$$

but since we are no longer on the space $H_0^1(I)$, we don't know if we can use Poincaré inequality to control the H^1 norm by this term. In fact, we can because if we use the same proof as in Proposition III.15, we can show that

$$\|u\|_{L^2} \leq \|\partial_x u\|_{L^2},\ \ \forall u \in H.$$

In other words, Poincaré inequality is also true for functions that are zero at only one of the two points of the boundary.

Finally, we have shown that a is coercive, and the Lax-Milgram theorem gives us the existence and uniqueness of a solution $u \in H$ of the problem

$$a(u,v) = L(v),\ \ \forall v \in H.$$

What did we solve ? For the moment we know that $u \in H^1(I)$ and that $u(1) = 0$ by definition of H. It remains to show that u verifies the PDE and the second boundary condition. We still have $\mathcal{C}_c^\infty(I) \subset H$ and so we can take $v = \varphi \in \mathcal{C}_c^\infty(I)$ as a test function in the variational formulation. This is exactly the same computation as before. It shows us that $\partial_x u$ is in fact a function of $H^1(I)$ whose weak derivative is f. In particular $\partial_x u$ is continuous.

Now, let us take any $v \in H$ and carry out the following integration by parts

$$\begin{aligned} -\int_0^1 f v\, dx &= \int_0^1 \partial_x(\partial_x u) v\, dx \\ &= [(\partial_x u)v]_0^1 - \int_0^1 \partial_x u \partial_x v\, dx. \end{aligned}$$

Since $v(1) = 0$, the above formula can be rewritten as

$$a(u,v) - L(v) = (\partial_x u)(0) v(0),$$

and must be verified for all $v \in H$. But we know that, by definition of the variational formulation that u verifies, the left hand side is null for all $v \in H$. Therefore, we have shown that

$$(\partial_x u)(0) v(0) = 0,\ \ \forall v \in H.$$

Since H contains functions that are non zero at the point 0 (for example $v(x) = 1 - x$) we conclude that

$$\partial_x u(0) = 0,$$

hence the initial problem is solved.

Neumann everywhere case What happens if we impose a Neumann condition on the boundary of the interval.

$$\begin{cases} -\partial_x^2 u = f(x), & x \in I, \\ \partial_x u(0) = \partial_x u(1) = 0. \end{cases}$$

Let us make two important remarks first:

- The solution intervenes only through its derivative in the problem, which shows that if u is a solution, then $u + c$ is also a solution for all $c \in \mathbb{R}$. So we cannot hope for uniqueness of solutions.

  To "fix" the constant c, we often have to impose an additional condition, for example the fact that u must have zero integral (i.e. zero mean).

- If there is a solution, we can integrate the equation on the interval I and we find

$$\int_0^1 f(x)\,dx = \int_0^1 (-\partial_x^2 u)\,dx = \partial_x u(0) - \partial_x u(1) = 0.$$

  Therefore, there can only be a solution for source terms with zero mean. This is called the compatibility condition between the source term and the boundary conditions.

It is now time to be more precise. So let us consider the following complete problem

$$\begin{cases} -\partial_x^2 u = f(x), & x \in I, \\ \partial_x u(0) = \partial_x u(1) = 0, \\ \displaystyle\int_0^1 u\,dx = 0, \end{cases} \tag{III.22}$$

under the assumption that

$$\int_0^1 f\,dx = 0. \tag{III.23}$$

Our experience from the previous examples shows us that the Neumann boundary conditions should not appear in the definition of the function space[5]. On the other hand, it seems natural to impose the constraint of zero mean.

So we will introduce the space $H = H^1_m(I)$ of functions with zero mean where

$$H^1_m(I) = \{v \in H^1(I),\ m(v) = 0\},\ \text{ where we denoted } m(v) = \int_0^1 v\,dx.$$

It is a closed subspace of $H^1(I)$ and therefore a Hilbert space.

Let us define

$$a(u,v) = \int_0^1 (\partial_x u)(\partial_x v)\,dx, \quad L(v) = \int_0^1 fv\,dx.$$

Once again, it is the coercivity of a that can be challenging to prove. So

$$a(u,u) = \int_0^1 |\partial_x u|^2\,dx,$$

and so we need a Poincaré type inequality. It turns out that this inequality is true:

Proposition III.22 (inégalité de Poincaré-moyenne ou de Poincaré-Wirtinger)

There exists $C > 0$ such that

$$\|u\|_{L^2} \le C\|\partial_x u\|_{L^2},\ \ \forall u \in H^1_m.$$

Proof :

For $s, t \in I$, we write

$$u(t) = u(s) + \int_s^t \partial_x u(x)\,dx,$$

and integrate this equality with respect to s to be able to use the zero mean condition on u.

$$u(t) = \underbrace{\int_0^1 u(s)\,ds}_{=m(u)=0} + \int_0^1 \left(\int_s^t \partial_x u\,dx\right) ds.$$

So for any t, we have

$$|u(t)| \le \int_0^1 \left|\int_s^t |\partial_x u|\,dx\right| ds \le \int_0^1 |\partial_x u|\,dx \le \|\partial_x u\|_{L^2}.$$

We square this equality and integrate with respect to t to obtain the desired result. ∎

From this inequality, we can deduce that a is coercive on H^1_m and Lax-Milgram theorem allows us to conclude on the existence and uniqueness of a solution $u \in H^1_m$ of

$$a(u,v) = L(v),\ \ \forall v \in H^1_m.$$

[5] Besides, for any function u of H^1, the value of $\partial_x u(0)$ is not well defined...

For the moment, we know that $u \in H^1$ and that it has zero mean. it remains to show that it verifies the PDE and the boundary conditions.

Here, we face a new difficulty because elements of $\mathcal{C}_c^\infty(I)$ are not all in H^1_m because they do not all have zero mean ! So we cannot take $v = \varphi$ in the variational formulation without being careful. On the other hand, we can take $v = \varphi - m(\varphi)$ which is indeed an element of H^1_m and thus write

$$a(u, \varphi - m(\varphi)) = L(\varphi - m(\varphi)), \ \ \forall \varphi \in \mathcal{C}_c^\infty(I).$$

Since the weak derivative of constant functions are zero, we have

$$\int_0^1 \partial_x u \partial_x \varphi \, dx = \int_0^1 f(\varphi - m(\varphi)) \, dx.$$

The second term can also be written as

$$\int_0^1 f\varphi \, dx - m(\varphi) \int_0^1 f \, dx = \int_0^1 f\varphi \, dx,$$

thanks to the compatibility hypothesis (III.23). At last, we have established that

$$\int_0^1 \partial_x u \partial_x \varphi \, dx = \int_0^1 f\varphi \, dx, \ \ \forall \varphi \in \mathcal{C}_c^\infty(I).$$

This shows, in a fashion that we are now getting used to, that u is a solution of the PDE in the weak sense. Let us now take any test function $v \in H^1_m$ and test it against the equation. By using the variational formulation that we just solved above, we get at last

$$(\partial_x u)(0)v(0) - (\partial_x u)(1)v(1) = 0, \ \ \forall v \in H^1_m.$$

We can choose any function v such that $v(0) = 0$, $v(1) = 1$ and $\int_0^1 v \, dx = 0$ to deduce that $\partial_x u(1) = 0$ and any function v such that $v(0) = 1$, $v(1) = 0$ and $\int_0^1 v \, dx = 0$ to obtain $\partial_x u(0) = 0$.

The initial Neumann problem is thus solved.

III.3 Proof of the Lax-Milgram theorem

The aim here is to give the proof of Theorem III.20 that you have already seen in the functional analysis course.

The symmetric case This case can be treated in two different ways (that are very close in reality):

Proof 1: we directly introduce the functional J from the statement of the theorem and we do the same analysis we did for the elastic string problem:

1. Show that J is bounded from below (by using the continuity of L and the coercivity of a).
2. Take a minimizing sequence $(u_n)_n$.
3. Show that $(u_n)_n$ is a Cauchy sequence by using the polarization identity for the bilinear form a as well as the coercivity of a.
4. Deduce that $(u_n)_n$ converges towards a limit $u \in H$ (by using the fact that H is complete).
5. We show that $J(u) = \inf_H J$, by using the fact that L and a are continuous.
6. At last, write $J(u + tv) \geq J(u)$ for all $v \in H$ and $t \in \mathbb{R}$ and deduce the associated Euler-Lagrange equations of the problem which are exactly the sought out equations.

Proof 2: Since a is symmetric and coercive (in particular positive definite), it is a scalar product on H, so the latter has two scalar products defined on it. By continuity and coercivity, the scalar product a induces a norm that is equivalent to the initial norm. So L is also continuous on H for this new scalar product.

Therefore, we can apply the Riesz representation theorem on this new space and immediately obtain the desired result.

The non symmetric case This case is more delicate and we cannot resort to any *variational* type method. In fact, if we assume that a quadratic functional J of the form $J(v) = \frac{1}{2}a(v,v) - L(v)$ has a minimizer in H, then we can show that u verifies the following formulation

$$\frac{1}{2}(a(u,v) + a(v,u)) = L(v), \ \ \forall v \in H,$$

meaning, we have $\tilde{a}(u,v) = L(v)$ for a certain **symmetric** bilinear form $\tilde{a}$.

In order to prove the theorem, we must therefore introduce a reformulation of the problem in the form of an operator for which we must show that an inverse exists.

Let us note that for any $u \in H$, the linear form $v \in H \mapsto a(u,v)$ is continuous and therefore there exists a unique $Au \in H$ such that $a(u,v) = (Au,v)_H$, by applying the Riesz representation theorem. It is clear that A constructed as such is a linear operator.

From our assumption on a, we have that $\forall u \in H, \|Au\|_H \leq |||a|||\|u\|_H$ hence A is a continuous operator. On the other hand, we have

$$(Au,u)_H = a(u,u) \geq \alpha\|u\|_H^2, \ \ \forall u \in H,$$

so

$$\|Au\|_H \geq \alpha\|u\|_H, \ \forall u \in H.$$

In particular, this implies that A is injective. From Riesz's theorem, we can represent L by an element $l \in H$, and we are now led to prove the existence of an element $u \in H$ such that $Au = l$.

Let $\rho > 0$, we introduce the map $T : u \in H \mapsto Tu = u - \rho(Au - l)$ and we see that solving the problem comes down to finding a fixed point for the map T. In a Hilbert space, the result will be proved if we show that T is a contraction. To do this, we carry out the following computations

$$\begin{aligned}\|Tu - Tv\|_H^2 &= \|u-v\|_H^2 + \rho^2\|Au - Av\|_H^2 - 2\rho(u-v, A(u-v))_H \\ &\leq \|u-v\|_H^2 + \rho^2|||a|||^2\|u-v\|_H^2 - 2\rho\alpha\|u-v\|_H^2 \\ &= (1 - 2\rho\alpha + \rho^2|||a|||^2)\|u-v\|_H^2.\end{aligned}$$

By choosing a sufficiently small ρ, and as $\alpha > 0$, we see that T is indeed a contraction, hence the result is proved.

IV Sobolev spaces and elliptic problems on a domain of $\mathbb{R}^d$

IV.1 Sobolev spaces on a domain of $\mathbb{R}^d$

By using the formalism of distributions recalled in the appendix, we can now introduce Sobolev spaces in any dimension in the following way.

Definition III.23 (The space $H^1(\Omega)$)

Let Ω be an open set of $\mathbb{R}^d$. We denote $H^1(\Omega)$ the subspace of $L^2(\Omega)$ that contains functions whose first order derivatives in the sense of distributions are elements of $L^2(\Omega)$, which gives

$$H^1(\Omega) = \big\{u \in L^2(\Omega), \ \ \textit{such that } \partial_{x_i}u \in L^2(\Omega), \ \forall i \in \{1,...,d\}\big\}.$$

The norm

$$\|u\|_{H^1(\Omega)} = \sqrt{\|u\|_{L^2}^2 + \sum_{i=1}^{d}\|\partial_{x_i}u\|_{L^2}^2}$$

gives the space a Hilbert structure.

It is easy to see that, in dimension $d = 1$, this definition coincides with the one introduced earlier in the chapter and what

we called *the weak derivative* is none other than the derivative in the sense of distributions.

Definition III.24 (The space $H^k(\Omega)$)

Let Ω be an open set of $\mathbb{R}^d$ and $k \geq 1$ an integer. The space $H^k(\Omega)$ is defined by

$$H^k(\Omega) = \big\{ u \in L^2(\Omega), \ \text{such that } \forall \alpha \in \mathbb{N}^d, \text{ with } |\alpha| \leq k, \ \partial^\alpha u \in L^2(\Omega) \big\}.$$

The norm

$$\|u\|_{H^k} = \sqrt{\sum_{\substack{\alpha \in \mathbb{N}^d \\ |\alpha| \leq k}} \|\partial^\alpha u\|_{L^2}^2},$$

gives the space a Hilbert structure.

Theorem III.25 (Density of regular functions)

1. *Let Ω be any bounded open set of $\mathbb{R}^d$. Then*

$$\mathcal{C}^\infty(\Omega) \cap H^k(\Omega)$$

   is dense in $H^k(\Omega)$.

2. *Let Ω be a bounded regular ($\mathcal{C}^1$) domain. Then*

$$\mathcal{C}^\infty(\overline{\Omega})$$

   is dense in $H^k(\Omega)$.

Remark III.26

For $d \geq 2$, unlike the one dimensional case, we have

$$H^1(\Omega) \not\subset \mathcal{C}^0(\overline{\Omega}).$$

For instance, in dimension $d = 2$, we define $\Omega = B(0,1)$ and

$$u(x) = \log|\log r|, \ \ r = |x|.$$

Clearly, this function is not continuous (it is not bounded on any neighborhood of 0).
But, by using polar coordinates, we have

$$\int_\Omega |u(x)|^2 \, dx = \int_\Omega |\log|\log r||^2 \, dx = 2\pi \int_0^1 r |\log|\log r||^2 \, dr < +\infty,$$

and

$$\int_\Omega |\nabla u(x)|^2 \, dx = \int_\Omega \left| \frac{1}{r \log r} \right|^2 dx = 2\pi \int_0^1 \left| \frac{1}{r \log r} \right|^2 r \, dr = 2\pi \int_0^1 \frac{1}{r |\log r|^2} \, dr < +\infty,$$

by using standard integration results.
So, the function u is an element of $H^1(\Omega)$ without being continuous on its domain of definition.

Nevertheless, we have the following weaker result.

Theorem III.27 (Sobolev embeddings)

Let Ω be a bounded regular domain of $\mathbb{R}^d$, $d \geq 2$. We first define

$$2^* = \begin{cases} \frac{2d}{d-2}, & \text{if } d \geq 3, \\ q, & \text{for any } q < +\infty, \text{ if } d = 2. \end{cases}$$

Then we have

$$u \in L^{2^*}(\Omega) \text{ and } \|u\|_{L^{2^*}(\Omega)} \leq C\|u\|_{H^1(\Omega)}, \ \ \forall u \in H^1(\Omega),$$

for a constant $C > 0$ that only depends on Ω.
These embeddings are optimal.

Nonetheless, we will try to define the trace on the boundary of the domain Ω of a function in a Sobolev space. To do that, we start by defining the space of functions whose square is integrable on the boundary of Ω. Assuming that (U, γ) is a parametrization of $\partial\Omega$ (possibly local ...) this space is written as

$$L^2(\partial\Omega) = \{f : \partial\Omega \to \mathbb{R}, \ \text{ such that } f \circ \gamma \in L^2(U)\},$$

that we endow with the norm

$$\|f\|^2_{L^2(\partial\Omega)} = \int_{\partial\Omega} |f|^2 \, d\sigma.$$

The following results will be stated without any proof.

Theorem III.28 (Traces)

Let Ω be a bounded regular domain (at least $\mathcal{C}^1$). The map

$$\gamma_0 : u \in \mathcal{C}^1(\overline{\Omega}) \mapsto u_{|\partial\Omega} \in \mathcal{C}^0(\partial\Omega),$$

verifies, for a constant $C > 0$ that only depends on Ω, the following inequality

$$\|\gamma_0 u\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}, \ \ \forall u \in \mathcal{C}^1(\overline{\Omega}). \tag{III.24}$$

So γ_0 can be extended in a unique manner to a linear and continuous operator

$$\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega),$$

called the **trace operator** *associated to the space $H^1(\Omega)$.*

Proposition III.29 (Stokes formula in Sobolev spaces)

If $u \in H^1(\Omega)$ and $V \in (H^1(\Omega))^d$ then we have

$$\int_\Omega u(\operatorname{div} V) \, dx = -\int_\Omega (\nabla u) \cdot V \, dx + \int_{\partial\Omega} (\gamma_0 u)(\gamma_0 V) \cdot n \, d\sigma.$$

Theorem III.30 (The trace space and lifting)

The trace operator γ_0 **is not surjectif**. *Conventionally, its image will be denoted $H^{\frac{1}{2}}(\partial\Omega)$*

$$H^{\frac{1}{2}}(\partial\Omega) = \gamma_0(H^1(\Omega)).$$

Endowing it with the quotient norm

$$\|v\|_{H^{\frac{1}{2}}} = \inf_{\substack{u\in H^1(\Omega)\\ \gamma_0(u)=v}} \|u\|_{H^1(\Omega)},$$

gives it a Hilbert structure.
The space $H^{\frac{1}{2}}(\partial\Omega)$ thus obtained contains $\mathcal{C}^1(\partial\Omega)$ and is dense in $L^2(\partial\Omega)$.
And finally, there exists a linear and continuous operator (that is not unique) $R_0 : H^{\frac{1}{2}}(\partial\Omega) \to H^1(\Omega)$ (called the **lifting operator***) that verifies*

$$\gamma_0\big(R_0 v\big) = v, \ \ \forall v \in H^{\frac{1}{2}}(\partial\Omega).$$

Remark III.31

1. *In this theorem, we have mentioned the space $\mathcal{C}^1(\partial\Omega)$ of $\mathcal{C}^1$ functions on the boundary of Ω. If (U,γ) is a parametrization (possibly local) of $\Gamma = \partial\Omega$, we will say that a function $f : \Gamma \to \mathbb{R}$ is $\mathcal{C}^1$ if $f\circ\gamma \in \mathcal{C}^1(U,\mathbb{R})$.*

   Under the hypothesis of the theorem, we can show that $f : \partial\Omega \to \mathbb{R}$ is a $\mathcal{C}^1$ function if and only if there exists a $\mathcal{C}^1$ function $F : \mathbb{R}^d \to \mathbb{R}$ such that

   $$F_{|\Gamma} = f.$$

2. *It is quite difficult, and out of the scope of this course, to explain why this space is denoted $H^{1/2}(\partial\Omega)$. Roughly speaking, this means that functions in this space possess* half derivatives *.... This can also be understood by refining the inequality (III.24). In fact, we can show that there exists a constant $C > 0$ such that*

   $$\|\gamma_0 u\|_{L^2(\partial\Omega)} \leq C\|u\|^{\frac{1}{2}}_{L^2(\Omega)}\|u\|^{\frac{1}{2}}_{H^1(\Omega)}.$$

   Therefore, the L^2 norm of the trace of u can be estimated with only the square root of the H^1 norm (the one that contains derivatives of u in Ω).

Definition III.32 (The space $H^1_0(\Omega)$)

The space $H^1_0(\Omega)$ is defined as the **closed** *subspace of $H^1(\Omega)$ made of functions whose trace is zero at the boundary. In other words,*

$$H^1_0(\Omega) = \mathit{Ker}\,\gamma_0.$$

Proposition III.33

Let Ω be a regular and bounded domain of $\mathbb{R}^d$. The space $\mathcal{D}(\Omega)$ is dense in $H^1_0(\Omega)$.

Theorem III.34 (Poincaré inequality)

Let Ω be a regular and bounded domain of $\mathbb{R}^d$. There exists a constant $C > 0$ depending only on Ω that verifies

$$\|u\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}, \ \ \forall u \in H^1_0(\Omega).$$

Theorem III.35 (Poincaré-Wirtinger inequality)

Let Ω be a **connected***, bounded and regular domain of $\mathbb{R}^d$. There exists a constant $C > 0$ depending only on Ω that verifies*

$$\|u\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}, \ \ \forall u \in H^1_m(\Omega),$$

where $H^1_m(\Omega)$ is the subset of functions in H^1 that have zero mean.

IV.2 Elliptic boundary value problems

Here, we will reconsider some of the examples that we have seen previously in this chapter but in higher dimension. In what will follow, Ω will denote a bounded and regular domain of $\mathbb{R}^d$.

IV.2.a Poisson problem with Dirichlet conditions

Let $k : \Omega \to \mathbb{R}$ be a measurable and bounded function such that $\alpha = \inf_\Omega k > 0$ and $f \in L^2(\Omega)$. We denote $H = H^1_0(\Omega)$. Now we define

$$a(u,v) = \int_\Omega k(x)(\nabla u)\cdot(\nabla v)\,dx, \quad L(v) = \int_\Omega fv\,dx.$$

Obviously, a and L are continuous. But to show that a is coercive, we can use the Poincaré inequality and the assumption made on k.

Thus, we can apply the Lax-Milgram theorem and obtain a function $u \in H^1_0(\Omega)$ that verifies

$$\int_\Omega k(x)\nabla u\cdot\nabla v\,dx = \int_\Omega fv\,dx, \ \ \forall v \in H^1_0(\Omega).$$

By taking $v = \varphi \in \mathcal{D}(\Omega)$, this formula becomes

$$\sum_{i=1}^d \langle k\partial_{x_i}u, \partial_{x_i}\varphi\rangle_{\mathcal{D}',\mathcal{D}} = \langle f,\varphi\rangle_{\mathcal{D}',\mathcal{D}},$$

so we have solved the problem

$$-\operatorname{div}(k\nabla u) = f, \ \text{ in } \mathcal{D}'(\Omega),$$

with the homogeneous Dirichlet condition on the boundary.

Remark III.36

If k is constant, we have solved the following problem

$$-k\Delta u = f,$$

with homogeneous Dirichlet conditions on the boundary.

In dimension 1, we were able to deduce that the second derivative u was an element of L^2 but here we must admit the following (difficult) result

Theorem III.37 (Elliptic regularity)

If Ω is sufficiently regular and k is a $\mathcal{C}^1$ function on $\overline{\Omega}$, then the solution to the previous problem verifies

$$u \in H^2(\Omega), \text{ and } \|u\|_{H^2} \leq C\|f\|_{L^2},$$

for a constant $C > 0$ that only depends on Ω.

For the non-homogeneous case, we search for a function $u \in H^1(\Omega)$ that verifies

$$\begin{cases} -\operatorname{div}(k(x)\nabla u) = f, \ \text{ in } \Omega, \\ \qquad\qquad\quad u = u_b, \ \text{ on } \partial\Omega. \end{cases}$$

According to the trace theorem, this requires that $u_b \in H^{\frac{1}{2}}(\partial\Omega)$. In fact, if we make this assumption, the lifting theorem tells us that there exists a function $R_0u_b \in H^1(\Omega)$ whose trace is equal to u_b.

We then modify the linear form L as follows

$$\tilde{L}(v) = \int_\Omega fv\,dx - \int_\Omega k(x)\nabla(R_0u_b)\cdot\nabla v\,dx.$$

The new term is always continuous for the $H^1(\Omega)$ norm. We then solve the problem

$$w \in H_0^1(\Omega), \quad a(w,v) = \tilde{L}(v), \ \ \forall v \in H_0^1(\Omega),$$

in such a way that $u = w + R_0u_b$ verifies the Poisson equation $-\operatorname{div}(k\nabla u) = f$ with the initially prescribed boundary condition $\gamma_0 u = u_b$.

IV.2.b Advection-diffusion problem

Let $b : \Omega \to \mathbb{R}^d$ be a regular vector field. We are interested in solving the following problem

$$\begin{cases} -\Delta u + b\cdot\nabla u = f, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases}$$

We propose the following variational formulation

$$\int_\Omega \nabla u\cdot\nabla v\,dx + \int_\Omega (b\cdot\nabla u)v\,dx = \int_\Omega fv\,dx, \ \ \forall v \in H_0^1(\Omega).$$

Like in dimension 1, the main difficulty is to establish the coercivity of the associated bilinear form, and here this means that we must estimate the advection term

$$\int_\Omega (b\cdot\nabla u)u\,dx,$$

which can be done easily if $\|b\|_\infty$ is small enough. We can also carry out an integration by parts on the advection term to obtain

$$-\frac{1}{2}\int_\Omega (\operatorname{div} b)u^2\,dx,$$

which, for example, makes it possible to bound it from below if $\operatorname{div} b \le 0$.

Again this is an example of the use of the Lax-Milgram theorem in a non-symmetric setting.

IV.2.c Neumann boundary conditions

Suppose that Ω is connected. Consider $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ such that

$$\int_\Omega f\,dx + \int_{\partial\Omega} g\,d\sigma = 0. \tag{III.25}$$

We define the space $H_m^1(\Omega)$ by

$$H_m^1(\Omega) = \left\{ u \in H^1(\Omega), \ \ \int_\Omega u\,dx = 0 \right\},$$

for which we admit that the following Poincaré-Wirtinger inequality still holds

$$\|u\|_{L^2(\Omega)} \le C\|\nabla u\|_{L^2(\Omega)}, \ \ \forall u \in H_m^1(\Omega).$$

We then define the following linear and bilinear forms

$$a(u,v) = \int_\Omega \nabla u\cdot\nabla v\,dx, \ \ L(v) = \int_\Omega fv\,dx + \int_{\partial\Omega} g\gamma_0(v)\,d\sigma, \ \ \forall v \in H_m^1(\Omega).$$

- Clearly, a is continuous and its coercivity comes from the Poincaré-Wirtinger inequality.
- The continuity of L comes from the Cauchy-Schwarz inequality and the trace theorem

$$|L(v)| \le \|f\|_{L^2}\|v\|_{L^2} + \|g\|_{L^2(\partial\Omega)}\|\gamma_0(v)\|_{L^2(\partial\Omega)} \le (\|f\|_{L^2(\Omega)} + C\|g\|_{L^2(\partial\Omega)})\|v\|_{H^1}.$$

The Lax-Milgram theorem applies and we get a unique function $u \in H^1_m(\Omega)$ that verifies

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\partial\Omega} g\gamma_0(v) \, d\sigma, \ \ \forall v \in H^1_m(\Omega).$$

For $v \in H^1(\Omega)$, we observe that $v - \frac{1}{|\Omega|}\int_\Omega v \, dx$ is an element of $H^1_m(\Omega)$. We can therefore apply the above property and use the compatibility condition (III.25) to obtain

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\partial\Omega} g\gamma_0(v) \, d\sigma, \ \ \forall v \in H^1(\Omega). \tag{III.26}$$

Now we can take $v = \varphi \in \mathcal{D}(\Omega)$ in (III.26) and obtain

$$\int_\Omega \nabla u \cdot \nabla \varphi \, dx = \int_\Omega f \varphi \, dx.$$

This shows that

$$-\Delta u = -\operatorname{div}(\nabla u) = f, \ \ \text{in the sense of distributions.}$$

If we assume that $u \in H^2(\Omega)$ (which can be proved if the given data are regular enough) then we can use the integration by parts formula given in Proposition III.29 to obtain, for all $v \in H^1_m(\Omega)$

$$\int_\Omega f v \, dx = \int_\Omega -\operatorname{div}(\nabla u) v \, dx = \int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \gamma_0(v)\gamma_0(\nabla u) \cdot n \, d\sigma.$$

By comparing this equality to (III.26), we get

$$\int_{\partial\Omega} (g - \gamma_0(\nabla u) \cdot n)\gamma_0(v) \, d\sigma = 0, \ \ \forall v \in H^1(\Omega).$$

But since the image of the trace operator (the space $H^{\frac{1}{2}}(\partial\Omega)$) is dense in $L^2(\partial\Omega)$, we deduce that

$$\gamma_0(\nabla u) \cdot n = g.$$

Thus, we have solved the Poisson problem with the above non-homogeneous Neumman condition.

What should be retained from this chapter

In priority

- The definition and properties of Sobolev spaces in dimension 1 : $H^1(I)$ and $H^1_0(I)$.
- The Lax-Milgram theorem. The proof in the symmetric case by minimizing a functional.
- Formulating a linear boundary value problem in a variational form and solving it using the Lax-Milgram theorem.
- Proving the regularity of the weak solution and afterwards interpreting the solved equation and the boundary conditions.

To go even further

- Sobolev spaces and variational formulations in higher dimensions.

Appendix A

Basic facts on distribution theory

The goal of this appendix is to remind the principal elements of distribution theory that allow us to comfortably work with the notion of weak derivatives. We will of course not treat the whole theory here but only the main definitions and properties that are essential to the study of partial differential equations, including their numerical analysis.

I Integration by parts in dimension d : the case of functions with compact support

We wish to have an integration by parts formula for functions with several variables that is similar to the one we know for functions with one variable.

Let us start by reminding what makes the computation work in dimension 1 for functions defined on the whole set $\mathbb{R}$ but with compact support.

1. We know that for any function $f \in \mathcal{C}^1_c(\mathbb{R})$ we have
$$\int_{\mathbb{R}} f'(x)\,dx = 0. \tag{A.1}$$
2. We use the fact that for any $u, v \in \mathcal{C}^1(\mathbb{R})$ we have
$$(uv)' = u'v + uv'.$$
3. We apply the first formula to $f = uv$ assuming that one of the two functions u or v (or maybe both) has compact support. So f has compact support and we have
$$0 = \int_{\mathbb{R}} f'(x)\,dx = \int_{\mathbb{R}} u'(x)v(x) + u(x)v'(x)\,dx,$$
which gives
$$\int_{\mathbb{R}} u'(x)v(x)\,dx = -\int_{\mathbb{R}} u(x)v'(x)\,dx,$$
which is indeed an integration by parts formula for functions with compact support.

Let us try to do the same for functions with several variables. Let Ω an open set of $\mathbb{R}^d$.

Definition A.1

For any scalar function $u \in \mathcal{C}^1(\Omega, \mathbb{R})$ and any vector field $F \in \mathcal{C}^1(\Omega, \mathbb{R}^d)$ we define

$$\nabla u = \begin{pmatrix} \partial_{x_1} u \\ \vdots \\ \partial_{x_d} u \end{pmatrix},$$

$$\Delta u = \sum_{i=1}^{d} \partial^2_{x_i} u,$$

$$\operatorname{div} F = \sum_{i=1}^{d} \partial_{x_i} F_i.$$

Proposition A.2 (A few useful formulas)

$$\begin{aligned} \operatorname{div}(uF) &= F\cdot(\nabla u) + (\operatorname{div} u)F, \\ \Delta u &= \operatorname{div}\big(\nabla u\big). \end{aligned} \tag{A.2}$$

Proposition A.3 (Integration by parts for functions with compact support)

1. *Let $F \in \mathcal{C}^1_c(\Omega,\mathbb{R}^d)$ be a vector field with compact support. We have*
$$\int_\Omega \operatorname{div} F\, dx = 0.$$
2. *Let $u \in \mathcal{C}^1(\Omega,\mathbb{R})$ and $V \in \mathcal{C}^1(\Omega,\mathbb{R}^d)$ such that u or V has compact support. We have*
$$\int_\Omega u(\operatorname{div} V)\, dx = -\int_\Omega (\nabla u)\cdot V\, dx.$$
3. *Let $u,v \in \mathcal{C}^2(\Omega,\mathbb{R})$ such that u or v has compact support. We have*
$$\int_\Omega u(-\Delta v)\, dx = \int_\Omega (\nabla u)\cdot(\nabla v)\, dx = \int_\Omega (-\Delta u)v\, dx.$$

Proof :

We will only do the proof in the case $\Omega = \mathbb{R}^d$. Indeed in the general case, at least one of the functions in play has compact support in Ω so we can extend it by 0 on the whole space without changing the fact that it is $\mathcal{C}^1$ and therefore the general case can be reduced to this one.

1. By definition we have $\operatorname{div} F = \sum_{i=1}^d \partial_{x_i} F_i$, we can therefore focus on any of the terms and see that the Fubini theorem gives us
$$\int_{\mathbb{R}^d} \partial_{x_i} F_i\, dx = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \partial_{x_i} F_i\, dx_1 \cdots dx_d = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \partial_{x_i} F_i\, dx_i \right) dx_1 \cdots \widehat{dx_i} \cdots dx_d,$$
and from (A.1) the integral with respect to x_i is null for all $x_1, \cdots, \widehat{x_i}, \cdots, x_d$.
2. We can apply the above result to the product $F = uV$ and use the formula (A.2) to deduce
$$0 = \int_{\mathbb{R}^d} \operatorname{div}(uV)\, dx = \int_{\mathbb{R}^d} u(\operatorname{div} V)\, dx + \int_{\mathbb{R}^d} (\nabla u)\cdot V\, dx.$$
3. It is sufficient to apply the above formula with $V = \nabla v$ for example and then observe that u and v play symmetric roles.

■

In Appendix B, we have detailed what needs to be done to be able to "integrate by parts" in the most general case where the functions of interest are no longer compactly supported. But the main result that should be retained is the following

$$\int_\Omega \operatorname{div} F\, dx = \int_{\partial\Omega} F\cdot n\, d\sigma,$$

where n corresponds to the outward unit normal to Ω and $\int_{\partial\Omega}$ is the surface integral. These objects are defined in this appendix and require a little more regularity for the boundary of Ω, which therefore cannot be any open set of $\mathbb{R}^d$.

II An important lemma from integration theory

In the following, we will be needing the following lemma

Lemma A.4 (of Du Bois-Reymond)

Let $\Omega \subset \mathbb{R}^d$ be a non empty open set and $f \in L^1_{loc}(\Omega)$ such that

$$\int_\Omega f\varphi\, dx = 0, \quad \forall \varphi \in \mathcal{C}^\infty_c(\Omega), \tag{A.3}$$

then $f = 0$.

Proof (of Lemma A.4):

We will give different proofs as per the regularity of f. Only the last one is the most general.

- First, let us study the simplest case where we suppose that the function f is continuous.

  By contradiction, suppose that f is non identically zero. Since f is continuous, and since we can always exchange f for $-f$, we can suppose that there exists a constant $C > 0$ and a non trivial ball $B(\alpha, r) \subset \Omega$ such that

  $$\forall x \in B(\alpha, r), f(x) \geq C.$$

  We therefore construct a function v of $\mathcal{C}^\infty$ class, positive, identically zero outside of $B(\alpha, r)$ and such that $\int_{B(\alpha,r)} v(x)\, dx > 0$. The construction of such a function is well known but one does not need to know the exact formula for this proof.

  From (A.3) and the properties of f and v, we have

  $$0 = \int_\Omega f(x)v(x)\, dx = \int_{B(\alpha,r)} f(x)v(x)\, dx \geq C \int_{B(\alpha,r)} v(x)\, dx > 0,$$

  which is obviously a contradiction.

- Now let us consider the case where $f \in L^2(\Omega)$.

  We will use the following property (from integration theory or functional analysis):

  $$\text{the set of functions } \mathcal{C}^\infty \text{ with compact support in } \Omega \text{ is dense in } L^2(\Omega).$$

  Since $f \in L^2$, this means that there exists a sequence $(\varphi_n)_n$ of elements of $\mathcal{C}^\infty_c$ such that $\|\varphi_n - f\|_{L^2} \to 0$.

  We therefore have

  $$\int_\Omega |f(x)|^2\, dx = \int_\Omega f(x)(f(x) - \varphi_n(x))\, dx + \int_\Omega f(x)\varphi_n(x)\, dx.$$

  Since $\varphi_n \in \mathcal{C}^\infty_c$, the second term is zero from the hypothesis (A.3), and we can bound the first term from above as follows

  $$\int_\Omega |f(x)|^2\, dx \leq \|f\|_{L^2}\|f - \varphi_n\|_{L^2} \xrightarrow[n\to\infty]{} 0.$$

  So, $|f|^2$ is a positive function with zero integral, therefore it is zero almost everywhere.

- Let us now do the proof in the general case. Consider any open and bounded set $U \subset \Omega$ and the function $g = \operatorname{sgn}(f)1_U$ which is measurable, bounded and integrable.

  By density of $\mathcal{C}^\infty_c(\Omega)$ in $L^1(\Omega)$, there exists a sequence $(\varphi_n)_n$ of elements of $\mathcal{C}^\infty_c(\Omega)$ such that $\|g - \varphi_n\|_{L^1} \to 0$. We would like to use these functions φ_n in the hypothesis but passing to the limit in the integral will be difficult since we do not have a uniform bound for the products $f\varphi_n$. So we will slightly transform the φ_n to guaranty a uniform bound without modifying their convergence to g.

  So, consider a function $\beta : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^\infty$ such that

  $$\begin{aligned} \beta(s) &= s, \text{ for } s \in [-1, 1], \\ |\beta(s)| &\leq 2, \text{ for } s \in \mathbb{R}. \end{aligned}$$

  Now define $\psi_n = \beta \circ \varphi_n$. By construction, we have $\psi_n \in \mathcal{C}^\infty_c(\Omega)$ and $\|\psi_n\|_{L^\infty} \leq 2$.

  We can also suppose that ψ_n converges almost everywhere towards g, if not we can extract a subsequence with such property.

  We observe that

  $$|f(x)\psi_n(x)| \leq |f(x)|\|\psi_n\|_{L^\infty} = 2|f(x)|, \text{ for almost all } x,$$

which allows us to use the dominated convergence theorem to establish

$$0 = \int_\Omega f(x)\psi_n(x)\,dx \xrightarrow[n\to\infty]{} \int_\Omega f(x)g(x)\,dx = \int_U f(x)\mathrm{sgn(f)}(x)\,dx = \int_U |f(x)|\,dx.$$

This proves that $\int_U |f| = 0$ and therefore $f = 0$ on U. Since this is valid for any open and bounded set U, we have shown that $f = 0$.

■

The density of the set of test functions we are using in the space $L^1(\Omega)$ is crucial for the proof. There exists other versions of this lemma, with a different set of test functions, that can be useful in other situations like the following result for instance.

Lemma A.5

If $f \in L^1_{loc}(\Omega)$ verifies

$$\int_B f(x)\,dx = 0, \quad \textit{for any ball } B \textit{ contained in } \Omega,$$

then $f = 0$ almost everywhere.

It should be noted that it is formally the same statement than the previous lemma but instead of taking test functions that are regular and with compact support, we take test functions of the form $\varphi = 1_B$ in (A.3).

Proof :

- Once again, the proof is significantly simpler if f is continuous. For now, let us do the proof in this case.

  Consider $x_0 \in \Omega$. We apply the hypothesis to $B = B(x_0, \varepsilon)$, where $\varepsilon > 0$ is small enough so that $B(x_0, \varepsilon) \subset \Omega$. It follows that

$$\begin{aligned} f(x_0) &= f(x_0) - \frac{1}{|B(x_0,\varepsilon)|}\underbrace{\left(\int_{B(x_0,\varepsilon)} f(x)\,dx\right)}_{=0}, \\ &= \frac{1}{|B(x_0,\varepsilon)|}\int_{B(x_0,\varepsilon)} (f(x_0) - f(x))\,dx, \end{aligned}$$

  whence we deduce

$$|f(x_0)| \le \frac{1}{|B(x_0,\varepsilon)|}\int_{B(x_0,\varepsilon)} |f(x_0) - f(x)|\,dx.$$

  As a result, it follows that

$$|f(x_0)| \le \sup_{x\in B(x_0,\varepsilon)} |f(x_0) - f(x)| \xrightarrow[\varepsilon\to 0]{} 0,$$

  since f is continuous at x_0. The proof is thus complete.

- Now, let us study the general case.

  The proof consists in noticing that, by linearity, the equality

$$\int_U f(x)\,dx = 0$$

  is valid for all U that can be written as a finite union of disjoint balls that are in Ω, then, by passing to the limit (for example, the dominated convergence theorem can be used), for any open and bounded set U contained in Ω[1].

  Now, consider any bounded Borel set A and define the following open set

$$A_\varepsilon = \bigcup_{x\in A} B(x,\varepsilon),$$

  that contains A by construction and that verifies

$$0 = \int_{A_\varepsilon} f(x)\,dx = \int_\Omega 1_{A_\varepsilon}(x)f(x)\,dx.$$

[1] such an open set can in fact be written as a countable union of balls

By the dominated convergence theorem, the convergence when $\varepsilon \to 0$ in the integral is justified[2] and, finally, we obtain

$$0 = \int_A f(x)\,dx.$$

This being true for any bounded Borel set A, in particular we can take

$$A^+ = \{x \in \Omega, \text{s.t.} \|x\| \leq R, f(x) \geq 0\},$$

$$A^- = \{x \in \Omega, \text{s.t.} \|x\| \leq R, f(x) < 0\},$$

to obtain

$$\int_{\Omega \cap B(0,R)} |f(x)|\,dx = \int_{A^+} f(x)\,dx + \int_{A^-} (-f(x))\,dx = 0 + 0 = 0.$$

This shows that f is zero on $\Omega \cap B(0,R)$ for any $R > 0$ therefore $f = 0$ on Ω.

■

III The space of test functions. The space of distributions.

Notations

- A d-tuple of positive integers $\alpha = (\alpha_1, ..., \alpha_d) \in \mathbb{N}^d$ is called a multi-index.
- The integer $|\alpha| = \alpha_1 + \cdots + \alpha_d$ is the length of the multi-index.
- For any $\alpha \in \mathbb{N}^d$ and any function f that is sufficiently differentiable, we define the partial derivative
$$\partial^\alpha f = \frac{\partial^{|\alpha|}}{\partial_{x_1}^{\alpha_1} \cdots \partial_{x_d}^{\alpha_d}} f.$$

These notations can often allow us to write heavy formulas in a more compact way[3].

III.1 Definitions, examples

We have already used the set of functions of class $\mathcal{C}^\infty$ with compact support. This set will play a major role in what will come in the following, and from now on, for any open set Ω of $\mathbb{R}^d$, we will denote

$$\mathcal{D}(\Omega) = \mathcal{C}_c^\infty(\Omega) = \{f \in \mathcal{C}^\infty(\Omega), \text{ there exists a compact set } K \subset \Omega \text{ such that } f = 0 \text{ on } K^c\}.$$

Be aware that the compact K which intervenes in this definition depends of course on the considered function f. So we define

$$\text{Supp}\,\varphi = \text{The smallest compact set } K \text{ such that } \varphi = 0 \text{ on } K^c.$$

[2] Make sure you know how to do it !
[3] See for example : https://en.m.wikipedia.org/wiki/Multi-index_notation

It is useful to know that such functions exist and that we can easily construct them with good properties.

Lemma A.6 (Construction of functions of class $\mathcal{C}^\infty$ with compact support)

1. *The function defined by*
$$\eta : s \in \mathbb{R} \mapsto \begin{cases} e^{-1/s^2}, & \text{for } s > 0, \\ 0, & \text{for } s < 0, \end{cases}$$
is $\mathcal{C}^\infty$ on $\mathbb{R}$.

2. *Let $a < b$. The function $\xi_{a,b}$ defined by*
$$\xi_{a,b}(s) = \frac{\eta(s-a)}{\eta(s-a) + \eta(b-s)}, \quad \forall s \in \mathbb{R}$$
is $\mathcal{C}^\infty$ on $\mathbb{R}$ and verifies
$$\xi_{a,b}(s) \begin{cases} = 0, & \text{for } s \le a, \\ \in (0,1), & \text{for } a < s < b, \\ = 1, & \text{for } s \ge b. \end{cases}$$

3. *Let $a < \alpha < \beta < b$. The function $\varphi_{a,\alpha,\beta,b}$ defined by*
$$\varphi_{a,\alpha,\beta,b}(s) = \eta_{a,\alpha}(s)(1 - \eta_{\beta,b}(s)), \quad \forall s \in \mathbb{R},$$
is $\mathcal{C}^\infty$ on $\mathbb{R}$ and verifies
$$\varphi_{a,\alpha,\beta,b} = 0, \text{ outside of } [a,b],$$
$$\varphi_{a,\alpha,\beta,b} = 1, \text{ in } [\alpha,\beta].$$

4. *Let U, V be non empty open sets of $\mathbb{R}^d$ that verifies $\overline{U} \subset V$ and such that U is bounded. Then, there exists φ, a $\mathcal{C}^\infty$ function on $\mathbb{R}^d$, such that*
$$\varphi = 1, \text{ in } U, \text{ and } \varphi = 0, \text{ outside of } V.$$

Proof :

1. Clearly, η is $\mathcal{C}^\infty$ on $(-\infty, 0)$ (any order derivative is zero !) and also on $(0, +\infty)$. It remains to show that all the right derivatives at 0 are equal to the left derivatives (in other words : are all zero).

   To do this, we establish the following induction property
$$\forall k \ge 0, \exists P_k \in \mathbb{R}[X], \text{ such that } \eta^{(k)}(s) = P_k(1/s)\eta(s), \quad \forall s > 0. \tag{A.4}$$
   The case $k = 0$ is immediate, and the case $k = 1$ gives
$$\eta'(s) = \frac{2}{s^3} e^{-1/s^2} = P_1(1/s)\eta(s),$$
   with $P_1(X) = 2X^3$.

   Now let us assume that the result is true for k and compute the $(k+1)$th derivative
$$\begin{aligned} \eta^{(k+1)}(s) &= (\eta^{(k)})'(s) \\ &= \frac{d}{ds}\left(P_k(1/s)\eta(s)\right) \\ &= \left(-\frac{1}{s^2} P_k'(1/s) + P_k(1/s)P_1(1/s)\right)\eta(s), \end{aligned}$$
   which gives the desired result if we consider the polynomial $P_{k+1}(X) = -X^2 P_k'(X) + P_K(X)P_1(X)$.

   From (A.4), and by using a well known result on limits, we deduce that
$$\lim_{s \to 0^+} \eta^{(k)}(s) = \lim_{x \to +\infty} \left(P_k(x)e^{-x^2}\right) = 0,$$

2. We observe that the denominator is never zero, which assures us of the $\mathcal{C}^\infty$ regularity of the function. The rest is immediate : one should use the fact that $\eta = 0$ on $\mathbb{R}^-$.

3. Same as the in previous case.

4. For any point $x \in \overline{U}$, there exists $r_x > 0$ such that $B(x, r_x) \subset V$. Then we write

$$\overline{U} \subset \bigcup_{x \in \overline{U}} B\left(x, \frac{r_x}{2}\right),$$

from which a finite covering can be extracted, by using the compactness of $\overline{U}$. So we have obtained, for $i = 1, ..., N$, $x_i \in \overline{U}$ and $r_i > 0$, such that

$$\overline{U} \subset \bigcup_{i=1}^{N} B\left(x_i, \frac{r_i}{2}\right).$$

For any $i \in \{1, ..., N\}$, we define

$$\psi_i(x) = \xi_{r_i/2, r_i}(\|x - x_i\|),$$

in such a way that ψ_i is $\mathcal{C}^\infty$ (there is actually no problem with the singularity of the norm function at 0, do you see why ?) and verifies

$$\psi_i = 0, \text{ in } B(x_i, r_i/2),$$

$$\psi_i = 1, \text{ outside of } B(x_i, r_i).$$

We now define

$$\varphi(x) = 1 - \prod_{i=1}^{N} \psi_i(x),$$

which is indeed a $\mathcal{C}^\infty$ function that verifies the desired properties.

■

We will now need to define a notion of convergence for sequences of elements of $\mathcal{D}(\Omega)$ but we will not treat in detail the rather complex topology of this space (in particular, it is not a normed vector space ...).

Definition A.7 (Convergence in $\mathcal{D}(\Omega)$)

We say that a sequence $(\varphi_n)_n \subset \mathcal{D}(\Omega)$ converges, in the sense of $\mathcal{D}(\Omega)$, to another function $\varphi \in \mathcal{D}(\Omega)$ if and only if

1. *There exists a compact set $K \subset \Omega$ such that $\mathrm{Supp}\,(\varphi_n) \subset K$ for any n and $\mathrm{Supp}\,(\varphi) \subset K$.*
2. *For* **any** *multi-index $\alpha \in \mathbb{N}^d$, the sequence of derivatives $(\partial^\alpha \varphi_n)_n$ uniformly converges to $\partial^\alpha \varphi$ in Ω, i.e.*

$$\|\partial^\alpha \varphi_n - \partial^\alpha \varphi\|_{L^\infty} \xrightarrow[n \to \infty]{} 0.$$

We will denote this convergence by $\varphi_n \xrightarrow[n \to \infty]{} \varphi$ in $\mathcal{D}(\Omega)$.

Remark A.8

An elementary remark : if $(\varphi_n)_n \subset \mathcal{D}(\Omega)$ converges to φ, then for any multi-index $\beta \in \mathbb{N}^d$, we have $\partial^\beta \varphi_n \xrightarrow[n \to \infty]{} \partial^\beta \varphi$.

We can now define the central object of this chapter.

Definition A.9 (The space of distributions)

We say that a linear form $T : \mathcal{D}(\Omega) \to \mathbb{R}$ *is a* **distribution on** Ω *if it is continuous in the following sense : for any sequence $(\varphi_n)_n \subset \mathcal{D}(\Omega)$ and $\varphi \in \mathcal{D}(\Omega)$ we have*

$$\left[\varphi_n \xrightarrow[n \to \infty]{} \varphi \text{ in } \mathcal{D}(\Omega)\right] \Longrightarrow T(\varphi_n) \xrightarrow[n \to \infty]{} T(\varphi).$$

We will denote $\mathcal{D}'(\Omega)$ the space of distributions on Ω. It is a real vector space.

Remark A.10

The space $\mathcal{D}'(\Omega)$ is therefore, in some sense, the topological dual of $\mathcal{D}(\Omega)$. In particular, we can often use the following notation

$$T(\varphi) = \langle T, \varphi \rangle_{\mathcal{D}',\mathcal{D}}, \quad \forall T \in \mathcal{D}'(\Omega), \forall \varphi \in \mathcal{D}(\Omega),$$

inspired by the standard notation of the scalar product in a Hilbert space.

III.1.a Functions of the space $L^1_{loc}(\Omega)$

Let us start by reminding the following

Definition A.11 (The $L^1_{loc}(\Omega)$ space)

Let Ω be an open set of $\mathbb{R}^d$. We denote

$L^1_{loc}(\Omega) = \{f : \Omega \to \mathbb{R}$, *the class of measurable functions such that $1_K f \in L^1(\Omega)$ for any compact set $K \subset \Omega\}$.*

We can endow this space with a distance that makes it a complete metric space and such that $(f_n)_n$ converges to f in $L^1_{loc}(\Omega)$ if and only if

$$\|1_K(f_n - f)\|_{L^1} \xrightarrow[n\to\infty]{} 0, \ \forall K \subset \Omega \text{ compact.}$$

Remark A.12

All the usual function spaces like $L^p(\Omega), 1 \le p \le \infty$, $\mathcal{C}^k(\Omega)$, $k \ge 0$, etc ... can be continuously embedded, in a natural way, into $L^1_{loc}(\Omega)$. So we can see it as a big space that contains all the others.

Proposition A.13

1. *For any function $f \in L^1_{loc}(\Omega)$, the map T_f defined by*

$$T_f : \varphi \in \mathcal{D}(\Omega) \mapsto \int_\Omega f\varphi \, dx,$$

   is a distribution on Ω.

2. *The map*

$$f \in L^1_{loc}(\Omega) \mapsto T_f \in \mathcal{D}'(\Omega),$$

   is injective.

Proof :

1. Let $(\varphi_n)_n$ be a sequence of $\mathcal{D}(\Omega)$ that converges to φ. Let us take a compact set K that contains all the support of the φ_n and of φ. We can now write

$$|T_f(\varphi_n) - T_f(\varphi)| \le \int_K |f||\varphi_n - \varphi| \, dx \le \|f\|_{L^1(K)} \|\varphi_n - \varphi\|_\infty \xrightarrow[n\to\infty]{} 0.$$

2. If we suppose that $T_f = 0$, this means

$$\int_\Omega f\varphi \, dx = 0, \ \ \forall \varphi \in \mathcal{D}(\Omega).$$

   Let U be any ball such that $\overline{U} \subset \Omega$. From the above, we can deduce that

$$\int_U f\varphi \, dx = 0, \ \ \forall \varphi \in \mathcal{C}^\infty_c(U),$$

   and since $f \in L^1(U)$, we can now apply Lemma A.4 to deduce that $f = 0$ on U. Since this is true for any ball U contained in Ω, we have shown that $f = 0$.

■

As a consequence of this proposition we will say, by a slight abuse of language, that the functions of the space $L^1_{loc}(\Omega)$ **are distributions** and we will systematically identify f with T_f. So, we will write

$$\langle f, \varphi \rangle_{\mathcal{D}',\mathcal{D}} = \int_\Omega f\varphi \, dx,$$

and

$$L^1_{loc}(\Omega) \subset \mathcal{D}'(\Omega).$$

For this reason it is not surprising that in some books, distributions are called **generalized functions**.

In particuliar, if $T \in \mathcal{D}'(\Omega)$ is an arbitrary distribution, and E a usual functional space (L^p, $\mathcal{C}^k$, etc ...), we will say that $T \in E$ if there exists $f \in E$ such that $T = T_f$ and therefore we will identify T to the function f in question (which is unique by the injectivity property).

III.1.b Dirac delta functions

Let $x_0 \in \Omega$. The Dirac delta function δ_{x_0} is a distribution on Ω defined by

$$\langle \delta_{x_0}, \varphi \rangle_{\mathcal{D}',\mathcal{D}} = \varphi(x_0), \ \forall \varphi \in \mathcal{D}(\Omega).$$

III.1.c Measures

More generally, any Borel measure μ that is locally finite on Ω is a distribution

$$\langle \mu, \varphi \rangle_{\mathcal{D}',\mathcal{D}} = \int_\Omega \varphi d\mu, \ \ \forall \varphi \in \mathcal{D}(\Omega).$$

III.1.d Other examples

The set of distributions contains even more general objects such as

$$\delta'_{x_0} : \varphi \in \mathcal{D}(\Omega) \mapsto \langle \delta'_{x_0}, \varphi \rangle_{\mathcal{D}',\mathcal{D}} = -\varphi'(x_0),$$

for $x_0 \in \Omega$,

$$\delta_\Gamma : \varphi \in \mathcal{D}(\Omega) \mapsto \langle \delta_\Gamma, \varphi \rangle_{\mathcal{D}',\mathcal{D}} = \int_\Gamma \varphi \, d\sigma,$$

where Γ is a regular hypersurface that is contained in Ω.

III.2 Convergence in the sense of distributions

Now, we are in need of a topology on the set of distributions. Once again, we will not enter into details so we will settle from a pragmatic level, a definition that will be sufficient for our needs.

Definition A.14 (Convergence in $\mathcal{D}'(\Omega)$)

We say that a sequence of distributions $(T_n)_n \subset \mathcal{D}'(\Omega)$ converges towards a distribution $T \in \mathcal{D}'(\Omega)$ if and only if

$$\langle T_n, \varphi \rangle_{\mathcal{D}',\mathcal{D}} \xrightarrow[n\to\infty]{} \langle T, \varphi \rangle_{\mathcal{D}',\mathcal{D}}, \ \ \forall \varphi \in \mathcal{D}(\Omega).$$

We will denote it by $T_n \to T$ in $\mathcal{D}'(\Omega)$. It is clear that, the limit of a sequence of distributions, if it exists, is unique.

This definition of convergence is quite easy to work with since it is a pointwise convergence. On the other hand, we will not be surprised to learn that this convergence is too weak to preserve some good properties. For instance, if $(f_n)_n$ is a sequence of continuous functions that converges to f in the sense of distributions, then there is no reason for f to be continuous. See some examples below.

Remark A.15

We can prove, but this is out of the scope of this course, that if for any test function $\varphi \in \mathcal{D}(\Omega)$ the sequences $(\langle T_n, \varphi \rangle_{\mathcal{D}',\mathcal{D}})_n$ are convergent, then there exists a distribution $T \in \mathcal{D}'(\Omega)$ such that $T_n \to T$ in $\mathcal{D}'(\Omega)$.

This definition of convergence in the sense of distributions is compatible with every reasonable notion of convergence we can think of. Let us give some fundamental examples.

- L^1_{loc} convergence and distributions

Proposition A.16

Let $(f_n)_n$ be a sequence of elements of $L^1_{loc}(\Omega)$ that converges, in this space, to a certain $f \in L^1_{loc}(\Omega)$. Then, we have

$$f_n \xrightarrow[n\to\infty]{} f, \ \text{ in } \mathcal{D}'(\Omega).$$

Proof :

Fix a $\varphi \in \mathcal{D}(\Omega)$ and introduce the compact set $K = \operatorname{Supp}\varphi$. We then write

$$|\langle f_n, \varphi\rangle_{\mathcal{D}',\mathcal{D}} - \langle f, \varphi\rangle_{\mathcal{D}',\mathcal{D}}| = \left|\int_\Omega (f_n - f)\varphi\, dx\right| = \left|\int_K (f_n - f)\varphi\, dx\right| \le \|\varphi\|_\infty \|f_n - f\|_{L^1(K)} \xrightarrow[n\to\infty]{} 0,$$

which proves the result. ∎

- Beware, it is possible (and this is one of the purposes of this theory) that the sequence $(f_n)_n$ converges to a distribution T that is not a function of the space $L^1_{loc}(\Omega)$. Let us study the following example: define $\Omega = B(0,1)$ the unit ball of $\mathbb{R}^d$ and fix a function $g \in L^1(\mathbb{R}^d)$ that verifies $\operatorname{Supp} g \subset \Omega$ and $\int_\Omega g\, dx = 1$.

  Then, we define $f_n(x) = n^d g(nx)$, which is a function of $L^1(\Omega)$. We will show that

  $$f_n \xrightarrow[n\to\infty]{} \delta_0, \ \text{ in the sense of distributions on } \Omega.$$

  To do this, we will choose $\varphi \in \mathcal{D}(\Omega)$ and show that

  $$\langle f_n, \varphi\rangle_{\mathcal{D}',\mathcal{D}} \xrightarrow[n\to\infty]{} \langle \delta_0, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \varphi(0).$$

  We observe that the support of f_n verifies $\operatorname{Supp} f_n \subset B(0, 1/n)$ and also, by a change of variables, that

  $$\int_\Omega f_n\, dx = \int_\Omega g\, dx = 1, \ \text{and} \ \int_\Omega |f_n|\, dx = \int_\Omega |g|\, dx = \|g\|_{L^1}.$$

  So, we can write

  $$\begin{aligned}
  |\langle f_n, \varphi\rangle_{\mathcal{D}',\mathcal{D}} - \varphi(0)| &= \left|\int_\Omega f_n\varphi\, dx - \varphi(0)\right| \\
  &= \left|\int_\Omega f_n(x)(\varphi(x) - \varphi(0))\, dx\right| \\
  &= \left|\int_{B(0,1/n)} f_n(x)(\varphi(x) - \varphi(0))\, dx\right| \\
  &\le \left(\sup_{x\in B(0,1/n)} |\varphi(x) - \varphi(0)|\right) \left|\int_\Omega |f_n|(x)\, dx\right| \\
  &= \|g\|_{L^1} \left(\sup_{x\in B(0,1/n)} |\varphi(x) - \varphi(0)|\right),
  \end{aligned}$$

  and this last quantity tends to 0 by continuity of φ at the point 0.

- We can also have $f_n \xrightarrow[n\to\infty]{} f$ in the sense of distributions without convergence of $(f_n)_n$ to f in $L^1_{loc}(\Omega)$.

  If we redo the previous computation, but with $\int_\Omega g\, dx = 0$ this time (and g is non identically zero), we can show that $(f_n)_n$ tends to 0 in the sense of distributions and yet we have

  $$\|f_n\|_{L^1} = \|g\|_{L^1} \neq 0, \ \ \forall n,$$

  which shows that it cannot converge to 0 in L^1.

- The convergence in the sense of distributions is not compatible with multiplication of functions ! Consider the following example:

  Let $f_n : \mathbb{R} \to \mathbb{R}$ a function defined by
  $$f_n(x) = \cos(nx), \ \ \forall x \in \mathbb{R}.$$
  Let us prove that
  $$f_n \xrightarrow[n\to\infty]{} 0, \text{ in } \mathcal{D}'(\mathbb{R}),$$
  $$f_n^2 \xrightarrow[n\to\infty]{} \frac{1}{2}, \text{ in } \mathcal{D}'(\mathbb{R}).$$
  Indeed, by fixing a test function $\varphi \in \mathcal{D}(\mathbb{R})$ and performing an integration by parts (by integrating f_n and by differentiating φ), we get
  $$\int_{\mathbb{R}} f_n \varphi \, dx = -\int_{\mathbb{R}} \frac{\sin(nx)}{n} \varphi'(x) \, dx.$$
  So
  $$\left| \int_{\mathbb{R}} f_n \varphi \, dx \right| \leq \frac{1}{n} \|\varphi'\|_{L^1} \xrightarrow[n\to\infty]{} 0.$$
  If now we compute f_n^2, thanks to some trigonometric formulas, we get
  $$f_n^2(x) = \frac{1}{2} + \frac{\cos(2nx)}{2} = \frac{1}{2}(1 + f_{2n}),$$
  and by using the above computation, we get the convergence of f_n^2 to $1/2$.

IV Differentiation in the sense of distributions.

Since distributions are *generalized functions* to help solve partial differential equations, it is necessary to know how to differentiate them.

One of the big advantages of this theory is that **every distribution is differentiable** which immensely simplifies the analysis, at least in this regard.

Definition and Proposition A.17 (Derivative of a distribution)

Let $T \in \mathcal{D}'(\Omega)$ be a distribution and $\alpha \in \mathbb{N}^d$ a multi-index. The map $\partial^\alpha T : \mathcal{D}(\Omega) \to \mathbb{R}$ defined by
$$\partial^\alpha T : \varphi \mapsto (-1)^{|\alpha|} \langle T, \partial^\alpha \varphi \rangle_{\mathcal{D}', \mathcal{D}},$$
is a distribution on Ω called the α^{th} derivative distribution of T.

Proof :

First, we observe that $\varphi \in \mathcal{D}(\Omega)$, so $\partial^\alpha \varphi$ is also in $\mathcal{D}(\Omega)$ and therefore the map $\partial^\alpha T$ is well defined.

It remains to show that it is continuous. To do this, we remark that if $\varphi_n \to \varphi$ in $\mathcal{D}(\Omega)$ then $\partial^\alpha \varphi_n \to \partial^\alpha \varphi$ in $\mathcal{D}(\Omega)$. ■

Of course, this notion is a generalization of the usual notion of derivatives according to the following proposition

Proposition A.18

If $f \in \mathcal{C}^k(\Omega, \mathbb{R})$, then, for any $\alpha \in \mathbb{N}^d$ such that $|\alpha| \leq k$, we have
$$\partial^\alpha (T_f) = T_{\partial^\alpha f}.$$
In other words, the notion of derivatives in the usual sense and the notion of derivatives in the sense of distributions coincide.

Proof :

One should simply use the integration by parts formula repeatedly for functions with compact support. ■

One last nice thing about distributions is the ability to switch differentiations and limits of a sequence "without second thoughts".

Theorem A.19

Let $(T_n)_n$ be a sequence of distributions that converges to a distribution T and $\alpha \in \mathbb{N}^d$. We have
$$\partial^\alpha T_n \xrightarrow[n\to\infty]{} \partial^\alpha T, \ \ \textit{in the sense of distributions}.$$

Proof :

By just writing the definitions, we get

$$\langle \partial^\alpha T_n, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = (-1)^{|\alpha|}\langle T_n, \partial^\alpha\varphi\rangle_{\mathcal{D}',\mathcal{D}} \xrightarrow[n\to\infty]{} (-1)^{|\alpha|}\langle T, \partial^\alpha\varphi\rangle_{\mathcal{D}',\mathcal{D}} = \langle \partial^\alpha T, \varphi\rangle_{\mathcal{D}',\mathcal{D}}.$$

■

Examples:

- The derivative of a Dirac delta function in 1D :

  We compute
  $$\langle \delta'_{x_0}, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = -\langle \delta_{x_0}, \varphi'\rangle_{\mathcal{D}',\mathcal{D}} = -\varphi'(x_0).$$

- Derivative of a piecewise $\mathcal{C}^1$ function in 1D :

  Let $f : [0,1] \to \mathbb{R}$ be a piecewise $\mathcal{C}^1$ function. We define $x_1 < \cdots < x_n$ the possible points of discontinuity of f and $x_0 = 0$, $x_{n+1} = 1$. By using integration by parts, we can prove the **jump formula**

  $$\begin{aligned}\langle \partial_x f, \varphi\rangle_{\mathcal{D}',\mathcal{D}} &= -\langle f, \partial_x\varphi\rangle_{\mathcal{D}',\mathcal{D}}\\ &= -\sum_{i=0}^{n}\int_{x_i}^{x_{i+1}} f\varphi'\,dx\\ &= -\sum_{i=0}^{n}\left[f(x_{i+1}^-)\varphi(x_{i+1}) - f(x_i^+)\varphi(x_i) - \int_{x_i}^{x_{i+1}} f'\varphi\,dx\right]\\ &= \sum_{i=1}^{n}\varphi(x_i)[f(x_i^+) - f(x_i^-)] + \sum_{i=0}^{n}\int_{x_i}^{x_{i+1}} f'\varphi\,dx.\end{aligned}$$

  Therefore, we have in the sense of distributions

  $$\partial_x f = f' + \sum_{i=1}^{n}[f(x_i^+) - f(x_i^-)]\delta_{x_i},$$

  where we denoted f' the piecewise continuous function that coincides with the derivative of f on each interval (x_i, x_{i+1}). The terms with the Dirac delta functions shows the influence of discontinuities of f in the derivative.

  We can see that $f \in H^1(0,1)$ if and only if there are no jumps, in other words, if f is continuous on $[0,1]$.

- The following is a classical trap : in general, the product of two distributions is not well defined (even the product of a distribution by a function that is not sufficiently regular).

  Therefore, if $u \in L^2(0,1)$, the quantity $u(\partial_x u)$ is not well defined in general (it will if $u \in H^1(0,1)$). In contrast, we know that (at least formally) we have $u(\partial_x u) = \frac{1}{2}\partial_x(u^2)$ and this is well defined since $u^2 \in L^1(0,1)$ is indeed a distribution that admits a derivative.

We conclude this appendix with a very natural and useful result that might not seem trivial at first sight.

Proposition A.20 (Distributions with zero gradient)

Let $T \in \mathcal{D}'(\Omega)$ be a distribution with zero gradient, meaning that

$$\partial_{x_i} T = 0, \quad \text{in } \mathcal{D}'(\Omega), \text{ for all } 1 \le i \le d.$$

In this case T is constant, meaning that there exists $\alpha \in \mathbb{R}$ such that

$$\langle T, \varphi\rangle_{\mathcal{D}',\mathcal{D}} = \alpha\int_\Omega \varphi\,dx, \quad \forall\varphi \in \mathcal{D}(\Omega).$$

Proof :

It will be enough for us to do the proof in dimension 1 where Ω is an interval $I = (a,b)$, since the general case is more delicate (see for example [3]).

Let $T \in \mathcal{D}'(I)$ such that for any test function $\varphi \in \mathcal{D}(I)$ we have:

$$\langle T, \varphi'\rangle_{\mathcal{D}',\mathcal{D}} = 0. \tag{A.5}$$

We fix once and for all a test function $\theta \in \mathcal{D}(I)$ such that $\int_I \theta(t)\,dt = 1$. Now, for any other test function $\psi \in \mathcal{D}(I)$,we define

$$\varphi(x) = \int_a^x \psi(t)\,dt - \left(\int_I \psi\right)\left(\int_a^x \theta(t)\,dt\right).$$

Of course, this function is $\mathcal{C}^\infty$, also we can verify that it has compact support. Moreover, we have

$$\varphi'(x) = \psi(x) - \left(\int_I \psi\right)\theta(x).$$

By applying (A.5) to the function φ thus constructed, we can find

$$\langle T, \psi\rangle_{\mathcal{D}',\mathcal{D}} = \left(\int_I \psi(x)\,dx\right)\underbrace{\langle T, \theta\rangle_{\mathcal{D}',\mathcal{D}}}_{\stackrel{\text{def}}{=} m_{T,\theta}}.$$

We have thus obtained that, for any test function ψ

$$\langle T, \psi\rangle_{\mathcal{D}',\mathcal{D}} = m_{T,\theta}\int_I \psi(x)\,dx = \langle m_{T,\theta}, \psi\rangle_{\mathcal{D}',\mathcal{D}},$$

so T is equal to the constant $m_{T,\theta}$. ■

Appendix B

Stokes formula

The purpose of this appendix is to understand what needs to be done to perform an "integration by parts" on any reasonable domain and with functions that are potentially non zero on the boundary. From our experience with integration by parts in 1D, boundary terms should necessarily appear in the formula.

Thus, our goal will be to prove the following result

Theorem B.1 (Stokes formula)

Let Ω be a bounded open set of $\mathbb{R}^d$ that satisfies some hypotheses that will be detailed later. Then, for any vector field $F \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$ we have

$$\int_\Omega \operatorname{div} F \, dx = \int_{\partial\Omega} F \cdot n \, d\sigma,$$

where n is the **outward unitary normal vector** *of the domain Ω.*

To prove this theorem, it is necessary to define the unitary outgoing normal vector of a domain Ω and also the integral of a scalar function on the boundary $\partial\Omega$ of the domain (it cannot be the usual Lebesgue integral because $\partial\Omega$ has zero measure in $\mathbb{R}^d$). This will be discussed in the following sections.

With the help of the above formula, we can actually generalize the results of Proposition A.3 for functions whose support is not necessarily compact in the following way

Proposition B.2

Let Ω be just like in the previous theorem.

1. *Let $u \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ and $V \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^d)$. We have*

$$\int_\Omega u(\operatorname{div} V) \, dx = -\int_\Omega (\nabla u) \cdot V \, dx + \int_{\partial\Omega} u(V \cdot n) \, d\sigma.$$

2. *Let $u, v \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$. We have*

$$\int_\Omega u(-\Delta v) \, dx = \int_\Omega (\nabla u) \cdot (\nabla v) \, dx - \int_{\partial\Omega} u(\nabla v \cdot n) \, d\sigma,$$

$$\int_\Omega (\Delta u) v \, dx - \int_\Omega u(\Delta v) \, dx = \int_{\partial\Omega} v(\nabla u \cdot n) \, d\sigma - \int_{\partial\Omega} u(\nabla v \cdot n) \, d\sigma.$$

I Hypersurfaces of $\mathbb{R}^d$. Surface integrals

I.1 Plane curves

Definition B.3

Let $J \subset \mathbb{R}$ be a compact interval of $\mathbb{R}$ and $\gamma : J \to \mathbb{R}^2$ a function of class $\mathcal{C}^1$ such that $\gamma'(t) \neq 0$ for all $t \in J$ and such that γ is injective. Therefore, we say that the image $\Gamma = \gamma(J) \subset \mathbb{R}^2$ is a simple curve of $\mathbb{R}^2$ and that the couple (J, γ) is a parametrization of Γ.

Definition B.4

For any $t \in J$, the vector $\gamma'(t) \in \mathbb{R}^2$ is a **tangent vector** *to the curve at point $\gamma(t)$.*
We denote $R_{\pi/2}$ the rotation map with angle $\pi/2$ (with an orientation being chosen beforehand). For any $t \in J$, the vector

$$n = \frac{R_{\pi/2}\gamma'(t)}{\|\gamma'(t)\|},$$

is a **normal unitary vector** *to the curve at point $\gamma(t)$. It does not depend on the parametrization, except perhaps for its orientation.*

Definition and Proposition B.5

Let $\Gamma = \gamma(J)$ be a simple curve of $\mathbb{R}^2$ and $f : \mathbb{R}^2 \to \mathbb{R}$ a continuous function. We define the quantity

$$I_\gamma(f) = \int_J f(\gamma(t))\|\gamma'(t)\|\, dt,$$

where $\|\cdot\|$ is the usual Euclidian norm on $\mathbb{R}^2$.
We then have the following:

- *$I_\gamma(f)$ only depends on the values of f on Γ.*
- *$I_\gamma(f)$ does not depend on the parametrization of Γ, meaning that if $(\tilde{J}, \tilde{\gamma})$ is another parametrization that verifies $\tilde{\gamma}(\tilde{J}) = \Gamma$, then we have*

$$I_{\tilde{\gamma}}(f) = I_\gamma(f).$$

Thus, the quantity $I_\gamma(f)$ will be called the integral of f on Γ and denoted by

$$\int_\Gamma f\, d\sigma, \quad \textit{ou} \quad \int_\Gamma f(x)\, d\sigma(x).$$

Proof :

We will not give the complete proof but we will quickly explain the main ideas, the key point being the change of variables theorem.

Let (J, γ) be a parametrization of Γ (we can suppose without loss of generality that $J = [0, T]$) and let $S : t \in [0, T] \mapsto [0, S(T)]$ be a bijective map of class $\mathcal{C}^1$ and such that $S'(t) > 0$ for any $t \in [0, T]$. Then, we can construct a new parametrization of Γ that is given by $(\tilde{J} = [0, S(T)], \tilde{\gamma})$ and

$$\tilde{\gamma}(s) = \gamma(S^{-1}(s)), \quad \forall s \in [0, S(T)].$$

Since S is bijective, this map is indeed well defined and we have $\gamma(J) = \tilde{\gamma}(\tilde{J}) = \Gamma$. At last, w get

$$\gamma(t) = \tilde{\gamma}(S(t)), \quad \forall t \in [0, T], \tag{B.1}$$

so

$$\gamma'(t) = S'(t)\tilde{\gamma}'(S(t)), \quad \forall t \in [0, T], \tag{B.2}$$

which proves in particular that $\tilde{\gamma}'$ cannot be zero. We have thus obtained a new parametrization of the same set Γ.

Let us now perform the following change of variables $s = S(t)$ in the integral that defines $I_\gamma(f)$

$$\begin{aligned} I_\gamma(f) &= \int_0^T f(\gamma(t))\|\gamma'(t)\|\,dt \\ &= \int_0^T f(\tilde\gamma(S(t)))\|S'(t)\tilde\gamma'(S(t))\|\,dt \\ &= \int_0^T f(\tilde\gamma(S(t)))\|\tilde\gamma'(S(t))\|\,S'(t)dt \\ &= \int_0^{S(T)} f(\tilde\gamma(s))\|\tilde\gamma'(s)\|\,ds \\ &= I_{\tilde\gamma}(f), \end{aligned}$$

which is indeed what we wanted to prove.

To finish the proof entirely, one should show that **all** parametrizations of Γ have indeed, more or less, the form $\tilde\gamma = \gamma \circ S^{-1}$ for a well chosen map S that verifies the above hypotheses. We leave this part as an exercise to the reader. ■

Remark B.6

With the definition, the **length** *of the curve* Γ *is the integral of the constant function* 1

$$|\Gamma| = \int_\Gamma 1\,d\sigma.$$

Let us give some examples:

- Integral on a line segment : Let $A = (a_1, a_2)$ and $B = (b_1, b_2)$ be two points of $\mathbb{R}^2$ and $\Gamma = [A, B]$ the line segment joining the points A and B. We will choose a natural parametrization of the segment

$$\gamma : t \in [0,1] \mapsto tA + (1-t)B = (ta_1 + (1-t)b_1, ta_2 + (1-t)b_2) \in \mathbb{R}^2.$$

  Then, for any function f, we have

$$\int_\Gamma f\,d\sigma = \int_0^1 f(tA + (1-t)B)\|A - B\|\,dt = \|A - B\| \int_0^1 f(tA + (1-t)B)\,dt.$$

  For $f = 1$, we retrieve the length of the segment which is $\|A - B\|$.

- Integral on a circle : We consider $\Gamma = C(0, R)$ the circle with center 0 and radius R and a parametrization given by

$$\gamma : t \in [0, 2\pi] \mapsto \gamma(t) = R(\cos(t), \sin(t)).$$

  Observe that here, $\gamma(0) = \gamma(2\pi)$ so γ is not injective but nevertheless, it is easy to convince oneself that the properties established above can be adapted without difficulty in the case of periodic parametrizations (closed curves).

  In these conditions, since $\|\gamma'(t)\| = R$ for all t, we have

$$\int_{C(0,R)} f\,d\sigma = R\int_0^{2\pi} f(R\cos(t), R\sin(t))\,dt,$$

  and for $f = 1$, once again we retrieve the circumference of the circle which is $2\pi R$.

- Integral on the graph of a function : Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function of class $\mathcal{C}^1$ and $[\alpha, \beta] \subset \mathbb{R}$. Consider the plane curve formed by a portion of the graph of φ defined by

$$\Gamma = \{(x, \varphi(x)),\ x \in [\alpha, \beta]\} \subset \mathbb{R}^2.$$

  This curve has a "natural" paramatrization that is given by

$$\gamma : x \in [\alpha, \beta] \mapsto (x, \varphi(x)) \in \Gamma.$$

  So for any function f, we have

$$\int_\Gamma f\,d\sigma = \int_\alpha^\beta f(x, \varphi(x))\sqrt{1 + \varphi'(x)^2}\,dx.$$

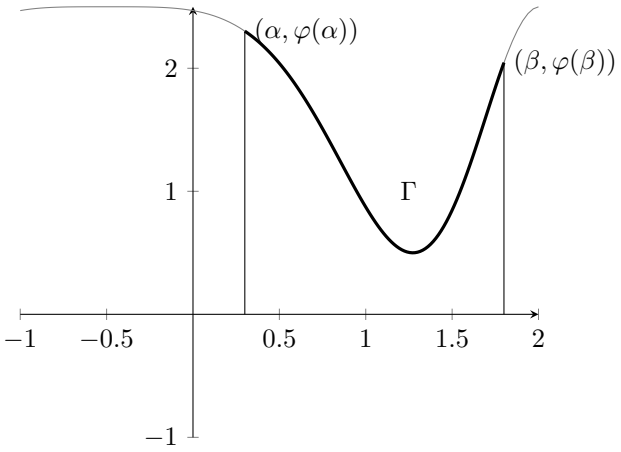


Figure B.1: Integral on the graph of a function

I.2 Integrals on hypersurfaces of $\mathbb{R}^d$

For simplicity, here we will only talk about the case $d = 3$ but the theory can be adapted for any dimension.

We will assume that we can define, like we did previously, the integral of any function (say continuous) on a hypersurface Γ of $\mathbb{R}^3$ by the formula

$$\int_\Gamma f \, d\sigma = \int_U f(\gamma(u)) \left\| \frac{\partial \gamma}{\partial u_1} \wedge \frac{\partial \gamma}{\partial u_2} \right\| du,$$

where

$$\gamma : u \in U \subset \mathbb{R}^2 \mapsto \gamma(u) \in \mathbb{R}^3,$$

is a parametrization of Γ (i.e. a bijection that is regular such that $\frac{\partial \gamma}{\partial u_1} \wedge \frac{\partial \gamma}{\partial u_2} \neq 0$ at each point u).

In practice, global parametrizations of Γ don't always exist and one is often forced to work with local parametrizations by using some other techniques but this will not be detailed here.

Definition B.7

For any $u \in U$, the vectors $\frac{\partial \gamma}{\partial u_1}(u), \frac{\partial \gamma}{\partial u_2}(u) \in \mathbb{R}^3$ are **tangent** *to the surface at the point $\gamma(u)$. Furthermore, they are linearly independent and the plane that they span does not depend on the chosen parametrization, and we call it* **the tangent plane** *to the surface at point $\gamma(u)$.*
For any $u \in J$, the vector

$$n = \frac{1}{\left\| \frac{\partial \gamma}{\partial u_1} \wedge \frac{\partial \gamma}{\partial u_2} \right\|} \frac{\partial \gamma}{\partial u_1} \wedge \frac{\partial \gamma}{\partial u_2},$$

is **a normal unitary vector** *to the surface at point $\gamma(u)$. It does not depend on the parametrization, except perhaps for its orientation.*

The case of graphs is always interesting to study so we will do that here more closely. Let $\varphi : U \subset \mathbb{R}^2 \to \mathbb{R}$ be a regular function, the graph of the function φ is defined by

$$\Gamma = \{(u, \varphi(u)), u \in U\} \subset \mathbb{R}^3,$$

and it is naturally parametrized by the following map

$$\gamma : u = (u_1, u_2) \in U \mapsto (u_1, u_2, \varphi(u)) \in \mathbb{R}^3.$$

A simple computation shows that

$$\frac{\partial \gamma}{\partial u_1} = \begin{pmatrix} 1 \\ 0 \\ \frac{\partial \varphi}{\partial u_1} \end{pmatrix}, \text{ and } \frac{\partial \gamma}{\partial u_2} = \begin{pmatrix} 0 \\ 1 \\ \frac{\partial \varphi}{\partial u_2} \end{pmatrix},$$

so that

$$\frac{\partial \gamma}{\partial u_1} \wedge \frac{\partial \gamma}{\partial u_2} = \begin{pmatrix} -\frac{\partial \varphi}{\partial u_1} \\ -\frac{\partial \varphi}{\partial u_2} \\ 1 \end{pmatrix},$$

and in particular, we have

$$\left\| \frac{\partial \gamma}{\partial u_1} \wedge \frac{\partial \gamma}{\partial u_2} \right\| = \sqrt{1 + \|\nabla \varphi(u)\|^2} > 0,$$

which proves that the parametrization is legitimate. The integral on Γ can therefore be expressed by

$$\int_\Gamma f \, d\sigma = \int_U f(u_1, u_2, \varphi(u_1, u_2)) \sqrt{1 + \|\nabla \varphi(u)\|^2} \, du. \tag{B.3}$$

Moreover, the normal unitary vector is given by

$$n = \frac{1}{\sqrt{1 + \|\nabla \varphi(u)\|^2}} \begin{pmatrix} -\frac{\partial \varphi}{\partial u_1} \\ -\frac{\partial \varphi}{\partial u_2} \\ 1 \end{pmatrix}, \tag{B.4}$$

and we observe that this vector is oriented “upwards" i.e. in the direction of increasing x_3's, and this independently of φ.

II Regular domains of $\mathbb{R}^d$

Definition B.8

We say that an open set Ω of $\mathbb{R}^d$ is a **regular domain** *(or* **a regular open set***, which is an abuse of language) if, locally, $\partial\Omega$ is a regular hypersurface of $\mathbb{R}^d$ and that Ω is located on only one side of $\partial\Omega$.*
So, at any point of $\partial\Omega$, we can define the unique unitary normal vector that is oriented from the interior to the exterior of Ω.

Some examples of open sets that **are not** regular domains:

- $\Omega = \mathbb{R}^d \setminus \{0\}$

  Indeed, the boundary of Ω is reduced to a point therefore it is not a hypersurface of $\mathbb{R}^d$.

- $\Omega = \mathbb{R}^d \setminus \left(\{0\} \times \mathbb{R}^{d-1}\right)$ is a union of two half-spaces.

  Its boundary is indeed a hypersurface of $\mathbb{R}^d$ but locally, Ω is located on both sides of its boundary, therefore we cannot define the *outgoing* normal vector.

  Observe that, in this situation, we can often (but not always) decompose Ω into two open regular domains.

- $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$

  Its boundary is not a regular hypersurface because of the square corners. In practice, we can still work with this type of domain if we weaken a little (but not too much) the assumptions of regularity on its boundary.

- Cusp domain :

  The open set

  $$\Omega = \{(x,y), \;\; 0 < x < 1, 0 < y < x^2\},$$

  drawn in Figure B.2 is not a regular domain because of the “sharp" singularity in the lower left corner. This type of domain presents a certain number of particularities that we will not detail further here.

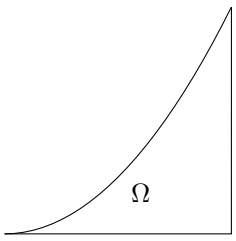


Figure B.2: Cusp domain

III Stokes Formula

We wish to prove Theorem B.1, at least in some particular cases.

III.1 The case of a half-space $\mathbb{R}^d_+$

Consider $\Omega = \mathbb{R}^d_+ = \mathbb{R}^{d-1} \times \mathbb{R}^+_*$. In the case $d = 3$, we have $\mathbb{R}^3_+ = \{(x_1, x_2, x_3),\ x_3 > 0\}$. The boundary of this open set is the plane $\Gamma = \partial\Omega = \mathbb{R}^2 \times \{0\}$.

Now, consider a vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ of class $\mathcal{C}^1$ and with compact support in $\mathbb{R}^d$. We integrate the divergence of F on Ω. The contribution of the terms $\partial_{x_1} F_1$ and $\partial_{x_2} F_2$ is always zero by Fubini's theorem. In contrast, the third term becomes

$$\int_\Omega \partial_{x_3} F_3 \, dx = \int_{\mathbb{R}} \int_{\mathbb{R}} \left(\int_0^{+\infty} \partial_{x_3} F_3(x_1, x_2, x_3) \, dx_3 \right) dx_1 dx_2 = - \int_{\mathbb{R}^2} F_3(x_1, x_2, 0) \, dx_1 dx_2.$$

By observing that the **outward** unitary normal vector to Ω on Γ is given by $n = {}^t(0, 0, -1)$, we have shown that

$$\int_\Omega \operatorname{div} F \, dx = \int_{\mathbb{R}^2} F \cdot n(x_1, x_2, 0) \, dx_1 dx_2.$$

The latter is indeed equal to the integral on the (flat) hypersurface $\partial\Omega$ that we have defined above.

III.2 The case of a half-space with a non planar boundary

Now, consider a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ and the following open set (called the epigraph of φ) defined by

$$\Omega = \{x = (x_1, x_2, x_3) \ \ x_3 > \varphi(x_1, x_2)\},$$

where the boundary $\Gamma = \partial\Omega$ is nothing but the graph of φ. The set Ω is therefore the set of all points lying *above* the graph Γ.

Consider once again a regular vector field F with compact support. We introduce the following change of variables

$$\Psi : \tilde{x} \in \mathbb{R}^3 \mapsto (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3 + \varphi(\tilde{x}_1, \tilde{x}_2)) \in \mathbb{R}^3.$$

We can easily verify that it is $\mathcal{C}^1$ and bijective from $\mathbb{R}^3$ to $\mathbb{R}^3$. Moreover, its Jacobian is given by

$$\operatorname{Jac}\Psi(\tilde{x}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \partial_{x_1}\varphi & \partial_{x_2}\varphi & 1 \end{pmatrix},$$

and its determinant is 1.

Finally, we observe that the image of the right half-space $\mathbb{R}^3_+$ by Ψ is exactly the domain Ω. We therefore define a new vector field $\tilde{F}$ with the formula

$$\tilde{F} = F \circ \Psi.$$

We immediately compute

$$(\operatorname{div} \tilde{F})(\tilde{x}) = (\operatorname{div} F)(\Psi(\tilde{x})) + \partial_{x_1}\varphi(\tilde{x})(\partial_{x_3} F_1)(\Psi(\tilde{x})) + \partial_{x_2}\varphi(\tilde{x}_1, \tilde{x}_2)(\partial_{x_3} F_2)(\Psi(\tilde{x})).$$

We observe that the last two terms can be written as derivatives with respect to $\tilde{x}_3$. In other words, if we define a new vector field

$$G(\tilde{x}) = \tilde{F}(\tilde{x}) + \begin{pmatrix} 0 \\ 0 \\ -(\partial_{x_1}\varphi)\tilde{F}_1 - (\partial_{x_2}\varphi)\tilde{F}_2 \end{pmatrix},$$

then, we have shown that

$$(\operatorname{div} G)(\tilde{x}) = (\operatorname{div} F)(\Psi(\tilde{x})).$$

We integrate this equality on the half-space $\mathbb{R}^3_+$ and obtain

$$\int_{\mathbb{R}^3_+} (\operatorname{div} G)(\tilde{x}) \, d\tilde{x} = \int_{\mathbb{R}^3_+} (\operatorname{div} F)(\Psi(\tilde{x})) \, d\tilde{x}.$$

Since the Jacobian of Ψ is 1 everywhere and that Ψ sends $\mathbb{R}^3_+$ on Ω, we can see that the second term is exactly $\int_\Omega \operatorname{div} F \, dx$. As for the first term, we can apply Stokes formula on the half-space that we have proved in the previous paragraph. We have finally established that

$$\int_\Omega \operatorname{div} F \, dx = - \int_{\mathbb{R}^2 \times \{0\}} G_3 \, d\sigma = - \int_{\mathbb{R}^2} G_3(\tilde{x}_1, \tilde{x_2}, 0) \, d\tilde{x}_1 d\tilde{x}_2.$$

By definition of G, this can be written as

$$\begin{aligned}\int_\Omega \operatorname{div} F\, dx &= \int_{\mathbb{R}^2} \Big[(\partial_{x_1}\varphi)\tilde{F}_1(\tilde{x}_1,\tilde{x}_2,0) + (\partial_{x_2}\varphi)\tilde{F}_2(\tilde{x}_1,\tilde{x}_2,0) - \tilde{F}_3(\tilde{x}_1,\tilde{x}_2,0)\Big]\, d\tilde{x}_1 d\tilde{x}_2 \\ &= \int_{\mathbb{R}^2} \Big[(\partial_{x_1}\varphi)F_1(\tilde{x}_1,\tilde{x}_2,\varphi(\tilde{x}_1,\tilde{x_2}) + (\partial_{x_2}\varphi)F_2(\tilde{x}_1,\tilde{x}_2,\varphi(\tilde{x}_1,\tilde{x}_2) - F_3(\tilde{x}_1,\tilde{x}_2,\varphi(\tilde{x}_1,\tilde{x_2})\Big]\, d\tilde{x}_1 d\tilde{x}_2.\end{aligned}$$

By the definitions recalled above (in particular the formulas (B.3), (B.4)), and by paying attention to the orientation of the normal vector, we have obtained that

$$\int_\Omega \operatorname{div} F\, dx = \int_{\partial\Omega} (F\cdot n)\, d\sigma.$$

Bibliography

[1] Sylvie Benzoni-Gavage. *Calcul différentiel et équations différentielles*. Dunod, 2010.

[2] Florent Berthelin. *Equations différentielles*. Cassini, 2017.

[3] Franck Boyer and Pierre Fabrie. *Mathematical Tools for the Study of the Incompressible Navier-Stokes Equations and Related Models*. Springer New York, 2013.

[4] Philippe G. Ciarlet. *Linear and nonlinear functional analysis with applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[5] Jean-Pierre Demailly. *Analyse Numérique et équations différentielles*. Collection Grenoble Sciences. Presses Universitaires de Grenoble, 1991.

[6] Françoise Demengel and Gilbert Demengel. *Espaces fonctionnels*. Savoirs Actuels (Les Ulis). [Current Scholarship (Les Ulis)]. EDP Sciences, Les Ulis; CNRS Éditions, Paris, 2007. Utilisation dans la résolution des équations aux dérivées partielles. [Application to the solution of partial differential equations].

[7] Benoît Perthame. *Transport Equations in Biology*. Birkhäuser Basel, 2007.