Contents

Chapter 1

Reminder on differential calculus

What you should know or be able to do after this chapter

- Know the definition of the differential, and be able to use it.
- Be able to compute the differential or partial derivatives of a function, when given an explicit expression.
- Be able to convert between the different expressions of the differential (linear map $\leftrightarrow$ Jacobian matrix $\leftrightarrow$ partial derivatives).
- Know that a differentiable function has partial derivatives, but be able to give an example of a function which has partial derivatives, and no differential.
- Prove the classical result on the differentiability of a composition of differentiable functions.
- Be able to apply this result to an explicit example (with no error on the point at which each differential must be computed!).
- Know the definition of the gradient and Hessian.
- Know the definitions of homeomorphism and diffeomorphism.
- When you want to prove that a function is locally invertible, think to the local inversion theorem, and be able to apply it correctly.
- When you want to parametrize a set defined by an equation, think to the implicit function theorem, and be able to apply it correctly.
- Propose examples which show that the assumption "$\partial_y f(x_0, y_0)$ is bijective" is necessary.
- Know the definition of an immersion and a submersion.
- Be able to apply the normal form theorems on explicit examples.
- When you want to upper bound the values of a differentiable function, or the difference between its values, think to the mean value inequality, and be able to apply it.

1.1 Definition of differentiability

Let $(E, ||.||_E)$, $(F, ||.||_F)$, and $(G, ||.||_G)$ be normed vector spaces. We denote the set of continuous linear mappings from E to F by $\mathcal{L}(E, F)$ [1].

[1] Recall that when E is of finite dimension, all linear mappings from E to F are continuous. This is no longer true if E is of infinite dimension.

Definition 1.1 : differentiability at a point

Let $U \subset E$ be an open set, and $f : U \to F$ be a function.
If x is a point in U, we say that f is *differentiable at* x if there exists $L \in \mathcal{L}(E, F)$ such that

$$\frac{||f(x+h) - f(x) - L(h)||_F}{||h||_E} \to 0 \quad \text{as } ||h||_E \to 0,$$

(or, equivalently, $f(x+h) = f(x) + L(h) + o(||h||_E)$).
We then call L the *differential of* f *at* x and denote it $df(x)$.

Remark

If $(E, ||.||_E) = (\mathbb{R}, |.|)$, then the differential, when it exists, takes the form

$$h \in \mathbb{R} \quad \to \quad hz_x \in F,$$

for a certain element z_x in F. In this case, we write

$$f'(x) = z_x.$$

We then recover the well-known formula:

$$f(x+h) = f(x) + f'(x)h + o(h) \quad \text{as } h \to 0.$$

Definition 1.2 : functions of class C^n

Let $U \subset E$ be an open set, and $f : U \to F$ a function.
The function f is said to be *differentiable on* U if it is differentiable at every point of U.
It is *of class* C^1 if it is differentiable and $df : U \to \mathcal{L}(E, F)$ is a continuous mapping.
More generally, for any $n \geq 1$, it is *of class* C^n if it is differentiable and df is of class C^{n-1}.
It is *of class* C^∞ if it is of class C^n for every $n \geq 1$.

We won't revisit the basic properties related to differentiability (e.g., the sum of differentiable functions is differentiable, etc.), except for the one concerning composite functions.

Theorem 1.3 : composite of differentiable functions

Let $U \subset E, V \subset F$ be open sets. Let $f : U \to V$ and $g : V \to G$ be two functions. Let $x \in U$.
If f is differentiable at x and g is differentiable at $f(x)$, then

- $g \circ f$ is differentiable at x;
- $d(g \circ f)(x) = dg(f(x)) \circ df(x)$.

1.2 Partial derivatives

In differential geometry, it is common to perform explicit calculations involving differentials of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. For this purpose, it is useful to represent differentials as matrices of size $m \times n$ (or vectors if $m = 1$) whose coordinates can be computed. The concept of *partial derivatives* allows us to achieve this.

Definition 1.4 : partial derivative

Let $n \in \mathbb{N}^*$. Let U be an open subset of $\mathbb{R}^n$, and $f : U \to \mathbb{R}$ a function.

Let $x = (x_1, \ldots, x_n) \in U$. For any $i = 1, \ldots, n$, we say that f is *differentiable with respect to its i-th variable* at x if the function

$$y \to f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots)$$

is differentiable at x_i. We then denote the derivative as $\partial_i f(x)$, $\partial_{x_i} f(x)$, or $\frac{\partial f}{\partial x_i}(x)$.

Remark

If f is differentiable at x, then it is also differentiable at x with respect to each of its variables. The converse is not necessarily true.

Remark

More generally, if $E_1, \ldots, E_n, F$ are normed vector spaces, U is an open subset of $E_1 \times \cdots \times E_n$, and $f : U \to F$ is a function, we can define, for all $x = (x_1, \ldots, x_n) \in U$ and $i = 1, \ldots, n$, the partial derivative of f with respect to x_i,

$$\partial_{x_i} f(x) \in \mathcal{L}(E_i, F).$$

Now let $n, m \in \mathbb{N}^*$, U be an open subset of $\mathbb{R}^n$, and $f : U \to \mathbb{R}^m$ be a differentiable function. For any x, $df(x)$ is a linear mapping from $\mathbb{R}^n \to \mathbb{R}^m$; we denote $Jf(x)$ its matrix representation in the canonical bases. If we identify $\mathbb{R}^n$ (respectively $\mathbb{R}^m$) with the set of column vectors of size n (respectively m), then

$$\forall u \in \mathbb{R}^n, \quad df(x)(u) = Jf(x) \times u.$$

The matrix $Jf(x)$ is called the *Jacobian matrix* of f at the point x.

Proposition 1.5

Let $f_1, \ldots, f_m : U \to \mathbb{R}$ be the components of f. Then, for any x,

$$Jf(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \cdots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \cdots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \frac{\partial f_m}{\partial x_2}(x) & \cdots & \frac{\partial f_m}{\partial x_n}(x) \end{pmatrix}.$$

Proof. Fix $x = (x_1, \ldots, x_n) \in U$. Let $\nu \in 1, \ldots, n$. Denote e_ν as the ν-th vector of the canonical basis of $\mathbb{R}^n$ (i.e., the vector whose coordinates are all 0 except the ν-th one, which is 1).

According to the definition of the differential,

$$\begin{aligned} f(x_1, \ldots, x_{\nu-1}, y, x_{\nu+1}, \ldots) &= f(x + (y - x_\nu)e_\nu) \\ &= f(x) + (y - x_\nu)df(x)(e_\nu) + o(y - x_\nu) \\ &\qquad \text{as } y \to x_\nu. \end{aligned}$$

For any $\mu \in 1, \ldots, m$, we have

$$f_\mu(x_1, \ldots, x_{\nu-1}, y, x_{\nu+1}, \ldots) = f_\mu(x) + (y - x_\nu)(df(x)(e_\nu))_\mu + o(y - x_\nu) \qquad \text{as } y \to x_\nu.$$

Thus, according to the definition of the partial derivative,

$$\begin{aligned} \partial_\nu f_\mu(x) &= \lim_{y \to x_\nu} \frac{f_\mu(x_1, \ldots, x_{\nu-1}, y, x_{\nu+1}, \ldots) - f_\mu(x)}{y - x_\nu} \\ &= (df(x)(e_\nu))_\mu. \end{aligned}$$

By the definition of the Jacobian matrix, $(Jf(x))_{\mu,\nu} = (df(x)(e_\nu))_\mu$, so

$$(Jf(x))_{\mu,\nu} = \partial_\nu f_\mu(x).$$

□

Example 1.6

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be such that, for every $(x_1, x_2) \in \mathbb{R}^2$,

$$f(x_1, x_2) = (x_1 x_2, x_1 + x_2).$$

It is differentiable. Its Jacobian matrix is

$$\forall (x_1, x_2) \in \mathbb{R}^2, \quad Jf(x_1, x_2) = \begin{pmatrix} x_2 & x_1 \\ 1 & 1 \end{pmatrix}$$

and its differential is

$$\forall (x_1, x_2), (h_1, h_2) \in \mathbb{R}^2, \quad df(x_1, x_2)(h_1, h_2) = (h_1 x_2 + h_2 x_1, h_1 + h_2).$$

In the particular case where $m = 1$, the Jacobian matrix has a single row:

$$\forall x \in U, \quad Jf(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) & \frac{\partial f}{\partial x_2}(x) & \dots & \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

Its transpose is then called the *gradient*:

$$\forall x \in U, \quad \nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

For all $x \in U, h = (h_1, \dots, h_n) \in \mathbb{R}^n$,

$$df(x)(h) = Jf(x) \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x) h_i = \langle \nabla f(x), h \rangle,$$

where the notation "$\langle ., . \rangle$" denotes the usual scalar product in $\mathbb{R}^n$.

Still assuming $m = 1$, let us consider the case where f is twice differentiable. Its second differential can also be represented by a matrix. Indeed, for any x, $d^2 f(x) = d(df)(x)$ belongs to $\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}))$. The application

$$(h, l) \in \mathbb{R}^n \times \mathbb{R}^n \quad \to d^2 f(x)(h)(l) \tag{1.1}$$

is therefore bilinear. As stated in the following property, it is even a quadratic form (i.e., it is symmetric), and the matrix associated with it in the canonical basis has a simple expression in terms of the partial derivatives of f.

Proposition 1.7 : Hessian matrix

Let $x \in U$. The application defined in (1.1) is a symmetric bilinear form. The matrix representing it in the canonical basis is

$$H(f)(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix}.$$

It is called the *Hessian matrix* of f at point x.

1.3 Local inversion

Definition 1.8 : homeomorphism

Let U, V be two topological spaces[a]. An application $\phi : U \to V$ is a *homeomorphism* from U to V if it satisfies the following three properties:

1. ϕ is a bijection from U to V;
2. ϕ is continuous on U;
3. ϕ^{-1} is continuous on V.

[a] Readers not familiar with the concept of "topological space" can limit themselves to the case where U and V are two metric spaces, or even to the case where U and V are subsets, respectively, of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ for $n_1, n_2 \in \mathbb{N}$.

Definition 1.9 : diffeomorphism

Let $n \in \mathbb{N}^*$ be an integer, $U, V \subset \mathbb{R}^n$ be two open sets. An application $\phi : U \to V$ is a *diffeomorphism* if it satisfies the following three properties:

1. ϕ is a bijection from U to V;
2. ϕ is C^1 on U;
3. ϕ^{-1} is C^1 on V.

If, moreover, ϕ and ϕ^{-1} are C^k for an integer $k \in \mathbb{N}^*$, we say that ϕ is a C^k*-diffeomorphism.*

Theorem 1.10 : local inversion

Let $n, k \in \mathbb{N}^*$ be integers, $U, V \subset \mathbb{R}^n$ be two open sets, and $x_0 \in U$. Let $\phi : U \to V$ be a C^k application. If $d\phi(x_0) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is bijective, then there exist $U_{x_0} \subset U$ an open neighborhood of x_0 and $V_{\phi(x_0)} \subset V$ an open neighborhood of $\phi(x_0)$ such that ϕ is a C^k-diffeomorphism from U_{x_0} to $V_{\phi(x_0)}$.

For the proof of this result, one can refer to [Paulin, 2009, p. 250].

An important consequence of the local inversion theorem is the implicit functions theorem, which allows to parameterize the set of solutions of an equation.

Theorem 1.11 : implicit functions

Let $n, m \in \mathbb{N}^*$. Let $U \subset \mathbb{R}^n \times \mathbb{R}^m$ be an open set, $f : U \to \mathbb{R}^m$ be a C^k application for an integer $k \in \mathbb{N}^*$, and (x_0, y_0) be a point in U such that

$$f(x_0, y_0) = 0.$$

If $\partial_y f(x_0, y_0) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ is bijective, then there exist

- an open neighborhood $U_{(x_0,y_0)} \subset U$ of (x_0, y_0),
- an open neighborhood $V_{x_0} \subset \mathbb{R}^n$ of x_0,
- an application $g : V_{x_0} \to \mathbb{R}^m$ of class C^k

such that, for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\big((x, y) \in U_{(x_0,y_0)} \text{ and } f(x, y) = 0\big) \quad \Longleftrightarrow \quad (x \in V_{x_0} \text{ and } y = g(x)).$$

In this theorem, the expression "$f(x, y) = 0$" should be interpreted as an equation depending on a parameter x, whose unknown is y. The theorem states that, in the neighborhood of (x_0, y_0), the equation has, for each value of the parameter x, a unique solution (which is $g(x)$) and that this solution is C^k relatively to x.

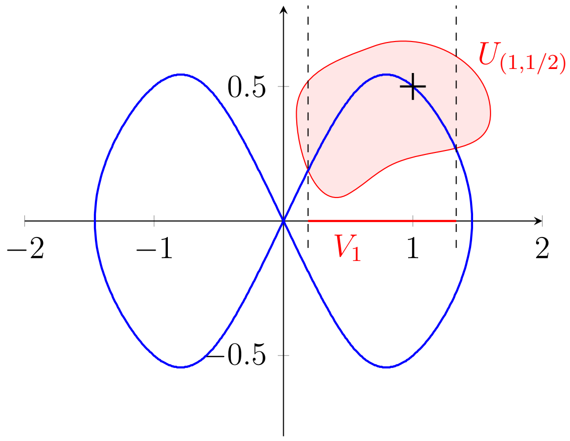


Figure 1.1: In blue, $\{(x,y) \in \mathbb{R}^2, \cos(\pi x) - \cos(\pi y) + 3x^2y^2 + \frac{x^4}{4} = 0\}$. This set is not the graph of a function. However, the part of the set inside $U_{(1,1/2)}$ coincides with the graph of a function $g : V_1 \to \mathbb{R}$.

Example 1.12

There exists an open neighborhood $U_{(1,1/2)} \subset \mathbb{R}^2$ of $(1, 1/2)$ and an open neighborhood $U_1 \subset \mathbb{R}$ of 1 such that the solutions of the equation

$$\cos(\pi x) - \cos(\pi y) + 3x^2y^2 + \frac{x^4}{4} = 0$$

for $(x,y) \in U_{(1,1/2)}$ are exactly the points of the set $\{(x, g(x))\}$ for a certain function $g : U_1 \to \mathbb{R}$ of class C^∞.
This is proven by applying the implicit functions theorem to

$$f : (x,y) \in \mathbb{R} \times \mathbb{R} \quad \to \quad \cos(\pi x) - \cos(\pi y) + 3x^2y^2 + \frac{x^4}{4} \in \mathbb{R}.$$

The bijectivity assumption of $\partial_y f(1, 1/2)$ is indeed satisfied:

$$\partial_y f(1, 1/2) = \pi + 3 \neq 0.$$

The set of solutions to the equation is represented in Figure 1.1.

Proof. Let us define

$$\begin{array}{rccc} \phi & : & U & \to & \mathbb{R}^n \times \mathbb{R}^m \\ & & (x,y) & \to & (x, f(x,y)). \end{array}$$

This is a C^k function, and for all $(h,l) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\begin{aligned} d\phi(x_0,y_0)(h,l) &= (h, df(x_0,y_0)(h,l)) \\ &= (h, \partial_x f(x_0,y_0)(h) + \partial_y f(x_0,y_0)(l)). \end{aligned}$$

The map $d\phi(x_0,y_0)$ is injective. Indeed, for all $(h,l) \in \mathbb{R}^n \times \mathbb{R}^m$, we must have, if $d\phi(x_0,y_0)(h,l) = 0$,

$$h = 0 \text{ and } \partial_y f(x_0,y_0)(l) = 0.$$

Since $\partial_y f(x_0,y_0)$ is bijective, this implies $l = 0$.

Thus, $d\phi(x_0,y_0)$ is an injective map from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^n \times \mathbb{R}^m$. Therefore, it is bijective (its domain and codomain have the same dimension).

We apply the local inversion theorem at (x_0,y_0). There exists an open neighborhood $U_{(x_0,y_0)}$ of (x_0,y_0), an open neighborhood V of $\phi(x_0,y_0) = (x_0, 0)$ such that ϕ is a C^k-diffomorphism from $U_{(x_0,y_0)}$ to V. Let

$$\psi : V \to U_{(x_0,y_0)}$$

be its inverse.

For all $(x,y) \in V$, we write $\psi(x,y) = (\psi_1(x,y), \psi_2(x,y)) \in \mathbb{R}^n \times \mathbb{R}^m$. For all $(x,y) \in V$,

$$\begin{aligned}(x,y) &= \phi \circ \psi(x,y) \\ &= \phi(\psi_1(x,y), \psi_2(x,y)) \\ &= (\psi_1(x,y), f(\psi_1(x,y), \psi_2(x,y))).\end{aligned}$$

Therefore,

$$\psi_1(x,y) = x.$$

We set

$$\begin{aligned}V_{x_0} &= \{x \in \mathbb{R}^n, (x,0) \in V\}; \\ g &: x \in V_{x_0} \to \psi_2(x,0) \in \mathbb{R}^m.\end{aligned}$$

As required, V_{x_0} is an open neighborhood of x_0 and g has class C^k. For all $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\begin{aligned}&\big((x,y) \in U_{(x_0,y_0)} \text{ and } f(x,y) = 0\big) \\ &\quad\iff \big((x,y) \in U_{(x_0,y_0)} \text{ and } \phi(x,y) = (x,0)\big) \\ &\quad\iff \big((x,y) \in U_{(x_0,y_0)} \text{ and } (x,0) \in V \text{ et } (x,y) = \psi(x,0)\big) \\ &\quad\iff ((x,0) \in V \text{ and } (x,y) = \psi(x,0) = (x, \psi_2(x,0))) \\ &\quad\iff (x \in V_{x_0} \text{ and } y = g(x))\,.\end{aligned}$$

□

1.4 Immersions and submersions

We now introduce two particular categories of differentiable functions, namely *immersions* and *submersions*, which will play a crucial role in the remainder of the course. Let $n, m \in \mathbb{N}^*$ be integers. Let $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$ be a C^k function (for some $k \geq 1$), with U an open set.

Definition 1.13 : immersions and submersions

For any point $x \in U$, we say that f is an *immersion at x* if $df(x) : \mathbb{R}^n \to \mathbb{R}^m$ is injective. We say that f is an *immersion* if it is an immersion at every point $x \in U$.
For any point $x \in U$, we say that f is a *submersion at x* if $df(x) : \mathbb{R}^n \to \mathbb{R}^m$ is surjective. We say that f is a *submersion* if it is a submersion at every point $x \in U$.

Remark

The function f can only be an immersion if $n \leq m$ and a submersion if $n \geq m$.

If f is an immersion at a point x, it is injective in a neighborhood of x (a consequence of Theorem 1.14). However, being an immersion is a significantly stronger property than local injectivity. Similarly, a submersion is locally surjective, but not all locally surjective functions are submersions.

When $n \leq m$, the simplest immersion from $\mathbb{R}^n$ to $\mathbb{R}^m$ is the function

$$(x_1, \dots, x_n) \in \mathbb{R}^n \quad \to \quad (x_1, \dots, x_n, 0, \dots, 0) \in \mathbb{R}^m.$$

The following theorem asserts that, up to a change of coordinates in the codomain (i.e., a transformation of the codomain by a diffeomorphism), all immersions are locally equal to this one.

Theorem 1.14 : normal form of immersions

Suppose that $0_{\mathbb{R}^n} \in U$ and $f(0_{\mathbb{R}^n}) = 0_{\mathbb{R}^m}$.
If f is an immersion at $0_{\mathbb{R}^n}$, there exists a neighborhood U' of $0_{\mathbb{R}^n}$ and a C^k-diffeomorphism ψ from a neighborhood of $0_{\mathbb{R}^m}$ to a neighborhood of $0_{\mathbb{R}^m}$ such that

$$\forall (x_1, \ldots, x_n) \in U', \quad \psi \circ f(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0).$$

Proof. Suppose that f is an immersion at $0_{\mathbb{R}^n}$.

Let $e_1, \ldots, e_n$ be the vectors of the canonical basis of $\mathbb{R}^n$, and $\epsilon_1, \ldots, \epsilon_m$ be those of the canonical basis of $\mathbb{R}^m$. Start by assuming that

$$\forall r \in \{1, \ldots, n\}, \quad df(0_{\mathbb{R}^n})(e_r) = \epsilon_r.$$

Define

$$\begin{array}{rccc} \phi & : & \mathbb{R}^m & \to & \mathbb{R}^m \\ & & (x_1, \ldots, x_m) & \to & f(x_1, \ldots, x_n) + (0, \ldots, 0, x_{n+1}, \ldots, x_m). \end{array}$$

We have $\phi(0) = 0$. Moreover, ϕ is a C^k function, and for any $h = (h_1, \ldots, h_m) \in \mathbb{R}^m$,

$$\phi(0_{\mathbb{R}^m})(h) = df(0_{\mathbb{R}^n})(h_1, \ldots, h_n) + (0, \ldots, 0, h_{n+1}, \ldots, h_m).$$

From this formula, it can be verified that $d\phi(0)(\epsilon_r) = \epsilon_r$ for all $r = 1, \ldots, m$, meaning that $d\phi(0) = \mathrm{Id}_{\mathbb{R}^m}$. In particular, $d\phi(0)$ is bijective.

According to the inverse function theorem, there exist open neighborhoods V_1, V_2 of $0_{\mathbb{R}^m}$ such that ϕ is a C^k-diffeomorphism between them. Let $\psi : V_2 \to V_1$ be its inverse. For any $x = (x_1, \ldots, x_n) \in U' \overset{def}{=} f^{-1}(V_2)$,

$$f(x_1, \ldots, x_n) = \phi(x_1, \ldots, x_n, 0, \ldots, 0),$$

so

$$\psi \circ f(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0).$$

This completes the proof of the theorem under the assumption that $df(0)(e_r) = \epsilon_r$ for all $r = 1, \ldots, n$.

Now, let's drop this assumption. For any $r \in \{1, \ldots, n\}$, denote $v_r = df(0_{\mathbb{R}^n})(e_r)$. As $df(0_{\mathbb{R}^n})$ is injective, the family $(v_1, \ldots, v_n)$ is linearly independent; it can be completed to a basis of $\mathbb{R}^m$, denoted by $(v_1, \ldots, v_m)$. Let $L \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ be such that

$$\forall r \in \{1, \ldots, m\}, \quad L(v_r) = \epsilon_r.$$

It is a bijection since it sends a basis to a basis.

Let $\tilde{f} = L \circ f$. We have $\tilde{f}(0_{\mathbb{R}^n}) = 0_{\mathbb{R}^m}$ and $d\tilde{f}(0_{\mathbb{R}^n}) = L \circ df(0_{\mathbb{R}^n})$. In particular, $\tilde{f}(0_{\mathbb{R}^n})$ is an immersion at 0. For any $r \in \{1, \ldots, n\}$,

$$d\tilde{f}(0_{\mathbb{R}^n})(e_r) = L(df(0_{\mathbb{R}^n})(e_r)) = L(v_r) = \epsilon_r.$$

Thus, the function $\tilde{f}$ satisfies our previous assumption. Consequently, there exist U' an open neighborhood of $0_{\mathbb{R}^n}$ and $\tilde{\psi}$ a diffeomorphism between two neighborhoods of $0_{\mathbb{R}^m}$ such that, for all $(x_1, \ldots, x_n) \in U'$,

$$\tilde{\psi} \circ \tilde{f}(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0),$$
$$\text{meaning } (\tilde{\psi} \circ L) \circ f(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0).$$

We set $\psi = \tilde{\psi} \circ L$ to conclude. □

A similar result holds for submersions and has a similar proof. When $n \geq m$, the simplest submersion from $\mathbb{R}^n$ to $\mathbb{R}^m$ is the projection onto the first m coordinates:

$$(x_1, \ldots, x_n) \in \mathbb{R}^n \quad \to \quad (x_1, \ldots, x_m) \in \mathbb{R}^m.$$

Subject to a change of coordinates in the domain, all submersions are locally equal to this one.

Theorem 1.15 : normal form of submersions

Suppose that $0_{\mathbb{R}^n} \in U$ and $f(0_{\mathbb{R}^n}) = 0_{\mathbb{R}^m}$.
If f is a submersion at $0_{\mathbb{R}^n}$, there exist U_1, U_2 open neighborhoods of $0_{\mathbb{R}^n}$ and a C^k diffeomorphism $\phi : U_1 \to U_2$ such that

$$\forall (x_1, \ldots, x_n) \in U_1, \quad f \circ \phi(x_1, \ldots, x_n) = (x_1, \ldots, x_m).$$

1.5 Mean value inequality

Let's conclude this chapter with a useful inequality, the mean value inequality.

Let $(E, ||.||_E)$ and $(F, ||.||_F)$ be normed vector spaces. We equip $\mathcal{L}(E, F)$ with the uniform norm: for any $u \in \mathcal{L}(E, F)$,

$$||u||_{\mathcal{L}(E,F)} = \sup_{x \in E \setminus \{0\}} \frac{||u(x)||_F}{||x||_E}.$$

Theorem 1.16 : mean value inequality

Let $U \subset E$ be a convex open set, and $f : U \to F$ a differentiable function.
Suppose there exists $M \in \mathbb{R}^+$ such that

$$\forall x \in U, \quad ||df(x)||_{\mathcal{L}(E,F)} \leq M.$$

Then,

$$\forall x, y \in U, \quad ||f(x) - f(y)||_F \leq M ||x - y||_E.$$

For the proof of this result, one can refer to [Paulin, 2009, p. 237].

Remark

Be careful not to forget the convexity assumption. The theorem may be false if it is not satisfied.
For example, the function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f(x) = -1$ for all $x < 0$ and $f(x) = 1$ for all $x > 0$ satisfies

$$|f'(x)| \leq 0 \quad \text{for all } x \in \mathbb{R} \setminus \{0\}$$

(as its derivative is zero).
However, it is not true that $|f(x) - f(y)| = 0$ for all $x, y \in \mathbb{R} \setminus \{0\}$.

Exercise 1 : classical application of the mean value inequality

Let $n, m \in \mathbb{N}^*$ be integers. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function such that, for any $x \in \mathbb{R}^n$,

$$||df(x)||_{\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)} \leq 1.$$

Show that, for any $x \in \mathbb{R}^n$,

$$||f(x)|| \leq ||f(0)|| + ||x||.$$

Chapter 2

Submanifolds of $\mathbb{R}^n$

What you should know or be able to do after this chapter

- Have an intuition of what is a submanifold of $\mathbb{R}^n$. In particular, from a drawing of a subset of $\mathbb{R}^2$ or $\mathbb{R}^3$, be able to guess with confidence whether it represents a submanifold or not.
- Know the four definitions of a submanifold of $\mathbb{R}^n$.
- When given the explicit expression of a set, be able to prove that it is a submanifold of $\mathbb{R}^n$, choosing the most appropriate of the four definitions.
- Know the definition of $\mathbb{S}^{n-1}$.
- Be able to prove that a set is a submanifold using the fact that it is a product of submanifolds.
- Understand the proof that $O_n(\mathbb{R})$ is a submanifold (i.e. be able to do it again alone, given only the definition of $\tilde{g}$).
- Be able to use the submersion definition of submanifolds to prove that sets are not submanifolds.
- Propose a definition of the tangent space to a submanifold, then remember the "true" one.
- Given a picture of a submanifold of $\mathbb{R}^2$ or $\mathbb{R}^3$, be able to draw (a plausible version of) the tangent space at any point.
- Given the explicit expression of a submanifold, be able to compute its tangent space, choosing the most appropriate of the four formulas.
- Know the tangent space to the sphere.
- Know that the tangent space of a product submanifold is the product of the tangent spaces.
- Be able to use the tangent space to prove that sets are not submanifolds (when possible).
- Guess a possible definition for the notion of "differentiable function" on a manifold, then remember the "true" one.
- Be able to show that a map between submanifolds is C^r, using the facts that compositions of C^r maps are C^r and that, on a C^k-submanifold, projections onto a coordinate are C^k.

In the whole chapter, let $k, n \in \mathbb{N}^*$ be fixed integers.

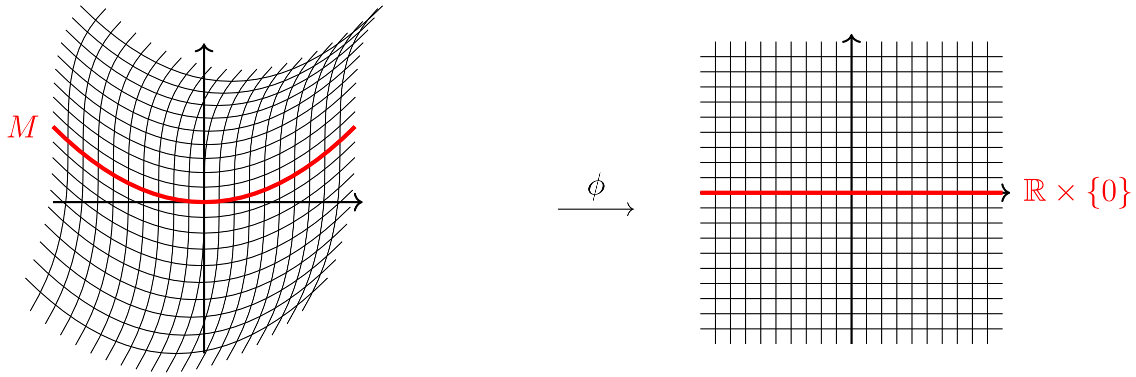


Figure 2.1: Illustration of property 1 in definition 2.1: there exists a local diffeomorphism from $\mathbb{R}^2$ to $\mathbb{R}^2$ that maps the set M onto $\mathbb{R} \times \{0\}$.

2.1 Definition

The simplest example of a submanifold of $\mathbb{R}^n$ is

$$\mathbb{R}^d \times \{0\}^{n-d} = \{(x_1, \dots, x_d, 0, \dots, 0) | x_1, \dots, x_d \in \mathbb{R}\},$$

where d is any integer between 0 and n. The concept of a *submanifold of* $\mathbb{R}^n$ generalizes this example: a set is a submanifold if it is locally the image of $\mathbb{R}^d \times \{0\}^{n-d}$ under a diffeomorphism from $\mathbb{R}^n$ to $\mathbb{R}^n$. Let's formalize this definition and provide other equivalent definitions.

Definition 2.1 : submanifolds

Let $d \in \{0, 1 \dots, n\}$.
Let $M \subset \mathbb{R}^n$. We say that the set M is a *submanifold of* $\mathbb{R}^n$ *of dimension* d *and class* C^k if it satisfies one of the following properties.

1. (Definition by diffeomorphism)

   For every $x \in M$, there exists a neighborhood $U \subset \mathbb{R}^n$ of x, a neighborhood $V \subset \mathbb{R}^n$ of 0, and a C^k-diffeomorphism $\phi : U \to V$ such that

   $$\phi(M \cap U) = (\mathbb{R}^d \times \{0\}^{n-d}) \cap V.$$

2. (Definition by immersion)

   For every $x \in M$, there exists a neighborhood $U \subset \mathbb{R}^n$ of x, an open set V in $\mathbb{R}^d$, a C^k function $f : V \to \mathbb{R}^n$ such that f is a homeomorphism between V and $f(V)$,

   $$M \cap U = f(V)$$

   and, denoting a as the unique pre-image of x under f, f is an immersion at a.

3. (Definition by submersion)

   For every $x \in M$, there exists a neighborhood $U \subset \mathbb{R}^n$ of x, a C^k function $g : U \to \mathbb{R}^{n-d}$ that is a submersion at x such that

   $$M \cap U = g^{-1}(\{0\})$$

4. (Definition by graph)

   For every $x \in M$, there exists a neighborhood $U \subset \mathbb{R}^n$ of x, an open set V in $\mathbb{R}^d$, a C^k function $h : V \to \mathbb{R}^{n-d}$, and a coordinate system[a] in which

   $$\begin{aligned} M \cap U &= \operatorname{graph}(h) \\ &\overset{def}{=} \{(x_1, \dots, x_d, h(x_1, \dots, x_d)), (x_1, \dots, x_d) \in V\}. \end{aligned}$$

[a] A *coordinate system* is the specification of a basis $(e_1, \dots, e_n)$ for $\mathbb{R}^n$. In this system, the notation $(x_1, \dots, x_n)$ denotes the point $x_1 e_1 + \dots + x_n e_n$.

Theorem 2.2

The four properties in Definition 2.1 are equivalent.

Among the four equivalent definitions in the theorem, the definition by diffeomorphism (property 1, illustrated in figure 2.1) is the one that most clearly reveals the connection between a general submanifold and the "model" submanifold $\mathbb{R}^d \times \{0\}^{n-d}$. However, it is not the most convenient to manipulate: when proving that a given set is a submanifold, the definitions by immersion, submersion, or graph are generally more convenient, as we will see in Section 2.2.

Remark

Pay attention to the fact that, in the definition by submersion (property 3), the function g maps into $\mathbb{R}^{n-d}$ and not into $\mathbb{R}^d$.
In a very informal way, in this definition, a submanifold is defined as the set of points in $\mathbb{R}^n$ that satisfy a set of scalar equations

$$g(x)_1 = 0, g(x)_2 = 0, ...$$

Intuitively, we expect the set of solutions to have $n-e$ "degrees of freedom", where e is the number of equations. For the submanifold defined in this way to be of dimension d, we need to have $e = n-d$, meaning that g maps into $\mathbb{R}^{n-d}$.

We advise the reader to study the examples in Section 2.2 before reading the proof of Theorem 2.2.

Proof of Theorem 2.2.
$\boxed{1 \Rightarrow 3}$: Suppose that M satisfies Property 1. We show that it satisfies Property 3.

Let $x \in M$. Consider U a neighborhood of x in $\mathbb{R}^n$, V a neighborhood of 0 in $\mathbb{R}^n$, and $\phi : U \to V$ a C^k-diffeomorphism such that

$$\phi(M \cap U) = (\mathbb{R}^d \times \{0\}^{n-d}) \cap V.$$

Denote $\mathrm{pr}_2 : \mathbb{R}^n \to \mathbb{R}^{n-d}$ the projection onto the last $n-d$ coordinates and define

$$g = \mathrm{pr}_2 \circ \phi : U \to \mathbb{R}^{n-d}.$$

It is a submersion at x because $dg(x)(\mathbb{R}^n) = \mathrm{pr}_2(d\phi(x)(\mathbb{R}^n)) = \mathrm{pr}_2(\mathbb{R}^n) = \mathbb{R}^{n-d}$ (recall that ϕ is a diffeomorphism, and thus, $d\phi(x)$ is bijective, meaning $d\phi(x)(\mathbb{R}^n) = \mathbb{R}^n$).

We verify that $M \cap U = g^{-1}(\{0\})$.

For every $x' \in M \cap U$, $\phi(x') \in \phi(M \cap U) = (\mathbb{R}^d \times \{0\}^{n-d}) \cap V \subset \mathbb{R}^d \times \{0\}^{n-d}$, so $\mathrm{pr}_2 \circ \phi(x') = 0$, i.e., $g(x') = 0$.

On the other hand, if $x' \in g^{-1}(\{0\})$, then $\mathrm{pr}_2(\phi(x')) = 0$, so $\phi(x') \in \mathbb{R}^d \times \{0\}^{n-d}$. Since $x' \in U$, $\phi(x') \in V$, and thus, $\phi(x') \in (\mathbb{R}^d \times \{0\}^{n-d}) \cap V = \phi(M \cap U)$, implying $x' \in M \cap U$.

$\boxed{3 \Rightarrow 4}$: Suppose that M satisfies Property 3. We show that it satisfies Property 4.

Let $x \in M$. Consider U a neighborhood of x in $\mathbb{R}^n$, and $g : U \to \mathbb{R}^{n-d}$ a C^k function, submersive at x, such that

$$M \cap U = g^{-1}(\{0\}).$$

Let $(e_1, \dots, e_n)$ be an orthonormal basis of $\mathbb{R}^n$ such that

$$\mathrm{Vect}\{dg(x)(e_{d+1}), \dots, dg(x)(e_n)\} = \mathbb{R}^{n-d}. \tag{2.1}$$

(Such a basis exists because $dg(x) : \mathbb{R}^n \to \mathbb{R}^{n-d}$ is surjective.) We now use the coordinate system defined by this basis. In this system, we denote

$$x = (x_1, \dots, x_n).$$

According to Equation (2.1), the derivative of g with respect to $(x_{d+1}, \dots, x_n)$ is surjective from $\mathbb{R}^{n-d}$ to $\mathbb{R}^{n-d}$, hence bijective. Thus, by the implicit function theorem (Theorem 1.11), there exist $U' \subset U$ a neighborhood of x, V a neighborhood of $(x_1, \dots, x_d)$, and $h : V \to \mathbb{R}^{n-d}$ of class C^k such that

$$U' \cap g^{-1}(\{0\}) = \{(t, h(t)), t \in V\}.$$

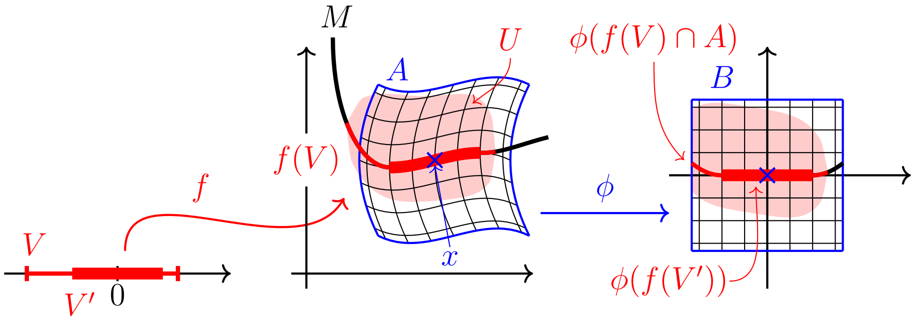


Figure 2.2: Illustration of the objects used in the proof of the implication $\boxed{2 \Rightarrow 1}$ of Theorem 2.2

Hence we have $M \cap U' = U' \cap g^{-1}(\{0\}) = \text{graph}(h)$.

$\boxed{4 \Rightarrow 2}$: Let's assume that M satisfies Property 4, and show that it satisfies Property 2.

Let $x \in M$. Without loss of generality, we can assume $x = 0$ to simplify notation. Let U be a neighborhood of $x = 0$ in $\mathbb{R}^n$, V an open set in $\mathbb{R}^d$, and $h : V \to \mathbb{R}^{n-d}$ be a C^k function such that, in a suitably chosen coordinate system,

$$M \cap U = \text{graph}(h) = \{(t, h(t)) \mid t \in V\}.$$

Note that $0 \in V$ and $h(0) = 0$, since $x = 0$ belongs to $M \cap U$.

Define

$$\begin{array}{cccc} f & : & V & \to & \mathbb{R}^n \\ & & t & \to & (t, h(t)). \end{array}$$

This is a C^k map. It is an immersion at 0 because, for any $t \in \mathbb{R}^d$, $df(0)(t)$ is given by

$$(t_1, \dots, t_d, dh(0)(t)),$$

which can only be zero if $t = 0$.

We have $f(0) = 0 = x$ and f is a homeomorphism between V and $f(V)$ (its inverse being the projection onto the first d coordinates). Furthermore,

$$M \cap U = \text{graph}(h) = f(V).$$

$\boxed{2 \Rightarrow 1}$: Let's assume that M satisfies Property 2, and show that it satisfies Property 1.

Let $x \in M$. Let U, V be neighborhoods of x and 0 in $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively, and let $f : V \to \mathbb{R}^n$ be a C^k function realizing a homeomorphism from V to $f(V)$, such that

$$M \cap U = f(V)$$

and f is immersive at a, where a is the unique preimage of x under f. Without loss of generality, we can assume, for simplicity, that $a = 0$, i.e., $f(0) = x$.

According to the local immersion theorem (1.14), there exists a neighborhood $V' \subset V$ of $0_{\mathbb{R}^d}$ and a C^k diffeomorphism $\phi : A \to B$ between a neighborhood A of x and a neighborhood B of $0_{\mathbb{R}^n}$ such that

$$\forall (t_1, \dots, t_d) \in V', \quad \phi \circ f(t_1, \dots, t_d) = (t_1, \dots, t_d, 0, \dots, 0). \tag{2.2}$$

An illustration of the various definitions in this proof is given in Figure 2.2.

Let $E \subset A \cap U$ be a neighborhood of x such that

- $f^{-1}(f(V) \cap E) \subset V'$ (such a neighborhood exists because f is a homeomorphism onto its image, so f^{-1} is well-defined and continuous on $f(V)$);
- $\phi(E) \subset V' \times \mathbb{R}^{n-d}$ (it also exists because ϕ is continuous, $V' \times \mathbb{R}^{n-d}$ is open and $\phi(x) = \phi \circ f(0) = 0 \in V' \times \mathbb{R}^{n-d}$).

Let $F = \phi(E)$.

The map ϕ is a C^k-diffeomorphism from E to F. Let's show that

$$\phi(M \cap E) = (\mathbb{R}^d \times \{0\}^{n-d}) \cap F. \tag{2.3}$$

For any $x' \in M \cap E$, we have $x' \in M \cap U = f(V)$, so $x' = f(t)$ for some $t \in V$. As $x' \in f(V) \cap E$, t is an element of V' according to the definition of E. Thus, by Equation (2.2), $\phi(x') = \phi(f(t)) \in \mathbb{R}^d \times \{0\}^{n-d}$. Moreover, $\phi(x') \in \phi(E) = F$. Therefore, $\phi(x') \in (\mathbb{R}^d \times \{0\}^{n-d}) \cap F$, which shows

$$\phi(M \cap E) \subset (\mathbb{R}^d \times \{0\}^{n-d}) \cap F.$$

Conversely, if $(t_1, \dots, t_d, 0, \dots, 0) \in (\mathbb{R}^d \times \{0\}^{n-d}) \cap F$, then $t \stackrel{\text{def}}{=} (t_1, \dots, t_d)$ is an element of V' (because $F = \phi(E) \subset V' \times \mathbb{R}^{n-d}$). Therefore, according to Equation (2.2),

$$(t_1, \dots, t_d, 0, \dots, 0) = \phi(f(t)).$$

As $f(t) \in f(V) \subset M$ and $f(t) \in \phi^{-1}(F) = E$, this shows that

$$(t_1, \dots, t_d, 0, \dots, 0) \in \phi(M \cap E).$$

Hence the inclusion $\phi(M \cap E) \supset (\mathbb{R}^d \times \{0\}^{n-d}) \cap F$, which completes the proof of Equation (2.3). □

2.2 Examples and counterexamples

As seen in the previous section, for any $d \in 0, \dots, n$,

$$\mathbb{R}^d \times \{0\}^{n-d}$$

is a submanifold of $\mathbb{R}^n$ (of class C^∞ and of dimension d).

Open sets provide another simple example of submanifolds: any non-empty open set in $\mathbb{R}^n$ is a submanifold of dimension n of $\mathbb{R}^n$.

2.2.1 Sphere

Definition 2.3

The *unit sphere in* $\mathbb{R}^n$ is the set

$$\mathbb{S}^{n-1} = \{(x_1, \dots, x_n) \in \mathbb{R}^n | x_1^2 + \cdots + x_n^2 = 1\}.$$

Proposition 2.4

The set $\mathbb{S}^{n-1}$ is a submanifold of $\mathbb{R}^n$, of class C^∞, and of dimension $n - 1$[a].

[a] It is precisely denoted $\mathbb{S}^{n-1}$ instead of $\mathbb{S}^n$ because its dimension is $n - 1$.

Proof. We will use the definition by submersion (Property 3 of Definition 2.1).

Let $x \in \mathbb{S}^{n-1}$. Consider $g : (t_1, \dots, t_n) \in \mathbb{R}^n \to t_1^2 + \cdots + t_n^2 - 1 \in \mathbb{R}$. This is a C^∞ function. It is a submersion at x. Indeed, $dg(x)$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}$, so it is either the zero map or a surjective map. Now,

$$\forall t = (t_1, \dots, t_n) \in \mathbb{R}^n, \quad dg(x)(t_1, \dots, t_n) = 2(x_1 t_1 + \cdots + x_n t_n).$$

Since $x_1^2 + \cdots + x_n^2 = 1$, x is not the zero vector, so $dg(x)$ is not the zero map; it is surjective.

Moreover, the definition of g implies that

$$\mathbb{S}^{n-1} = g^{-1}(\{0\}).$$

Property 3 of Definition 2.1 is therefore satisfied (with $U = \mathbb{R}^n$). □

2.2.2 Product of submanifolds

Proposition 2.5

Let $n_1, n_2 \in \mathbb{N}^*, d_1 \in \{0, \dots, n_1\}, d_2 \in \{0, \dots, n_2\}$. If M_1 is a submanifold of $\mathbb{R}^{n_1}$ of class C^k and dimension d_1, and M_2 is a submanifold of $\mathbb{R}^{n_2}$ of class C^k and dimension d_2, then

$$M_1 \times M_2 \overset{def}{=} \{(x_1, x_2), x_1 \in M_1, x_2 \in M_2\}$$

is a submanifold of $\mathbb{R}^{n_1+n_2}$ of dimension $d_1 + d_2$.

Proof. We use the definition by immersion (Property 2 of Definition 2.1). Let $x = (x_1, x_2) \in M$.

As M_1 is a submanifold, there exists a neighborhood U_1 of x_1, an open set V_1 in $\mathbb{R}^{d_1}$, and a function $f_1 : V_1 \to \mathbb{R}^{n_1}$ of class C^k, which is a homeomorphism onto its image, such that

$$M_1 \cap U_1 = f_1(V_1)$$

and f_1 is immersive at $f_1^{-1}(x_1)$.

Define similarly U_2, V_2, and $f_2 : V_2 \to \mathbb{R}^{n_2}$.

The function $f : (t_1, t_2) \in V_1 \times V_2 \to (f_1(t_1), f_2(t_2)) \in \mathbb{R}^{n_1+n_2}$ is of class C^k. It is a homeomorphism onto its image (with inverse $(z_1, z_2) \in f(V_1 \times V_2) \to (f_1^{-1}(z_1), f_2^{-1}(z_2))$, if we denote f_1^{-1} and f_2^{-1} the respective inverses of f_1 and f_2). Furthermore,

$$\begin{aligned}(M_1 \times M_2) \cap (U_1 \times U_2) &= (M_1 \cap U_1) \times (M_2 \cap U_2) \\ &= f_1(V_1) \times f_2(V_2) \\ &= f(V_1 \times V_2).\end{aligned}$$

Finally, f is immersive at $f^{-1}(x) = (f_1^{-1}(x_1), f_2^{-1}(x_2))$. Indeed, for any $t = (t_1, t_2) \in \mathbb{R}^{n_1+n_2}$,

$$df(f^{-1}(x_1), f^{-1}(x_2))(t_1, t_2) = (df_1(f_1^{-1}(x_1))(t_1), df_2(f_2^{-1}(x_2))(t_2)),$$

which equals 0 only if $t_1 = 0$ and $t_2 = 0$, since $df_1(f_1^{-1}(x_1))$ and $df_2(f_2^{-1}(x_2))$ are injective.

Thus, the set $M_1 \times M_2$ satisfies Property 2 of Definition 2.1. □

Example 2.6 : torus

The set $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$ is a submanifold of $\mathbb{R}^4$, of dimension 2. It is called a *torus of dimension* 2.

2.2.3 $O_n(\mathbb{R})$

Let $\mathbb{R}^{n\times n}$ denote the set of $n \times n$ matrices with real coefficients. If we reindex the coordinates, this set can also be viewed as $\mathbb{R}^{n^2}$. Several important subsets of $\mathbb{R}^{n\times n}$ have a submanifold structure. Here, we focus on the orthogonal group.

Definition 2.7 : orthogonal group

The *orthogonal group* is defined as

$$O_n(\mathbb{R}) = \{A \in \mathbb{R}^{n\times n}, I_n = {}^tAA\}.$$

Proposition 2.8

The set $O_n(\mathbb{R})$ is a submanifold of $\mathbb{R}^{n\times n}$, of class C^∞ and of dimension $\frac{n(n-1)}{2}$.

Proof. We will use the definition by submersion. Let $G \in O_n(\mathbb{R})$. We must express $O_n(\mathbb{R})$ as $g^{-1}(\{0\})$, where g is a C^∞ function, submersive at G.

A first idea is to define

$$g : A \in \mathbb{R}^{n\times n} \to {}^tAA - I_n \in \mathbb{R}^{n\times n}.$$

The definition of the orthogonal group implies that $O_n(\mathbb{R}) = g^{-1}(\{0\})$. However, this function is not a submersion at G. Indeed,

$$\forall A \in \mathbb{R}^{n\times n}, \quad dg(G)(A) = {}^tGA + {}^tAG,$$

so $dg(G)(\mathbb{R}^{n\times n})$ is contained in Sym_n, the set of symmetric matrices of size $n \times n$. We even have $dg(G)(\mathbb{R}^n) = \mathrm{Sym}_n$ because, for any $S \in \mathrm{Sym}_n$,

$$dg(G)\left(\frac{GS}{2}\right) = \frac{{}^tGGS + {}^tS{}^tGG}{2} = \frac{S + {}^tS}{2} = S.$$

In particular, $dg(G)(\mathbb{R}^{n\times n}) \neq \mathbb{R}^{n\times n}$.

Therefore, we define instead

$$\tilde{g} = \mathrm{Tri} \circ g : \mathbb{R}^{n\times n} \to \mathbb{R}^{\frac{n(n+1)}{2}},$$

where Tri is the function that extracts the upper triangular part of an $n \times n$ matrix:

$$\forall A \in \mathbb{R}^{n\times n}, \quad \mathrm{Tri}(A) = (A_{ij})_{i\leq j} \in \mathbb{R}^{\frac{n(n+1)}{2}}.$$

The function $\tilde{g}$ is C^∞. It is a submersion at G:

$$\begin{aligned} d\tilde{g}(G)(\mathbb{R}^{n\times n}) &= (\mathrm{Tri} \circ dg(G))(\mathbb{R}^{n\times n}) \\ &= \mathrm{Tri}(dg(G)(\mathbb{R}^{n\times n})) \\ &= \mathrm{Tri}(\mathrm{Sym}_n) \\ &= \mathbb{R}^{\frac{n(n+1)}{2}}. \end{aligned}$$

Furthermore, for any matrix $A \in \mathbb{R}^{n\times n}$, ${}^tAA = I_n$ if and only if ${}^tAA - I_n = 0$, which is equivalent to $\mathrm{Tri}({}^tAA - I_n) = 0$, since ${}^tAA - I_n$ is a symmetric matrix. Thus,

$$O_n(\mathbb{R}) = \tilde{g}^{-1}(\{0\}),$$

so $O_n(\mathbb{R})$ indeed satisfies Property 3, with $U = \mathbb{R}^{n\times n}$ and $d = n - \frac{n(n+1)}{2} = \frac{n(n-1)}{2}$. □

2.2.4 Equation solutions and function images

Proposition 2.9

Let $d \in \{0, \ldots, n\}$. Let U be an open subset of $\mathbb{R}^n$, and

$$g : U \to \mathbb{R}^{n-d}$$

a C^k function. Assume that g is a submersion over $g^{-1}(\{0\})$ (meaning that g is a submersion at x for all $x \in g^{-1}(\{0\})$).
Then $g^{-1}(\{0\})$ is a submanifold of $\mathbb{R}^n$, of class C^k and dimension d.

Proof. This is a direct application of Definition 2.1, "submersion" version. □

We have already seen two examples of submanifolds defined as in Proposition 2.9:

- the sphere $\mathbb{S}^{n-1}$ is equal to $g^{-1}(\{0\})$ for the function $g : x \in \mathbb{R}^n \to ||x||^2 - 1 \in \mathbb{R}$;
- the orthogonal group $O_n(\mathbb{R})$ is equal to $g^{-1}(\{0\})$ for the function $g : A \in \mathbb{R}^{n\times n} \to \mathrm{Tri}({}^tAA - I_n)$.

Proposition 2.10

Let $d \in \{0, \ldots, n\}$. Let U be an open subset of $\mathbb{R}^d$, and $f : U \to \mathbb{R}^n$ be a C^k function. Assume that f is an immersion, and is a homeomorphism from U to $f(U)$.
Then $f(U)$ is a submanifold of $\mathbb{R}^n$, of class C^k and dimension d.

Proof. This is a direct application of Definition 2.1, "immersion" version. □

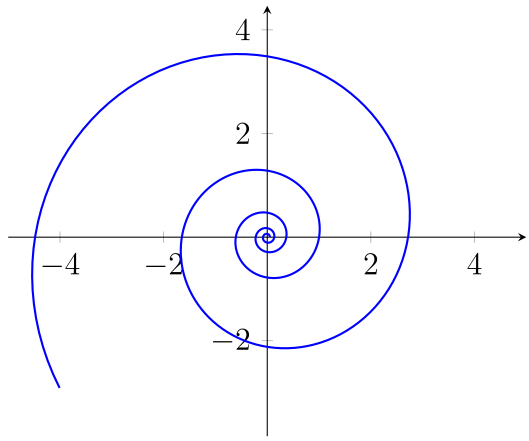


Figure 2.3: Image of the function f defined in Example 2.11

Example 2.11 : spiral

Let's define

$$\begin{array}{cccc} f & : & \mathbb{R} & \to & \mathbb{R}^2 \\ & & \theta & \to & \left(e^\theta\cos(2\pi\theta), e^\theta\sin(2\pi\theta)\right). \end{array}$$

Its image $f(\mathbb{R})$ is a submanifold. It is represented in Figure 2.3.
Indeed, for any $\theta\in\mathbb{R}$,

$$f'(\theta) = e^\theta\left((\cos(2\pi\theta),\sin(2\pi\theta)) + 2\pi\left(-\sin(2\pi\theta),\cos(2\pi\theta)\right)\right),$$

which never vanishes (we observe, for example, that $\langle f'(\theta),(\cos(2\pi\theta),\sin(2\pi\theta))\rangle = e^\theta \neq 0$ for any $\theta\in\mathbb{R}$). Thus, the function f is an immersion. Moreover, it is a homeomorphism from $\mathbb{R}$ to $f(\mathbb{R})$. Indeed, it is continuous, injective[a] and therefore bijective onto $f(\mathbb{R})$. For any $\theta\in\mathbb{R}$,

$$e^{2\theta} = ||f(\theta)||^2,$$

so $\theta = \frac{1}{2}\log\left(||f(\theta)||^2\right)$. As a consequence, the inverse of f is given by the following explicit expression:

$$\begin{array}{cccc} f^{-1} & : & f(\mathbb{R}) & \to & \mathbb{R} \\ & & (x,y) & \to & \frac{1}{2}\log(x^2+y^2). \end{array}$$

From this expression, we see that f^{-1} is the restriction to $f(\mathbb{R})$ of a continuous function on $\mathbb{R}^2\setminus(0,0)$, so f^{-1} is continuous.

[a] For any θ_1,θ_2, if $f(\theta_1)=f(\theta_2)$, then $e^{2\theta_1} = ||f(\theta_1)||^2 = ||f(\theta_2)||^2 = e^{2\theta_2}$, so $\theta_1=\theta_2$.

2.2.5 Submanifolds of dimension 0 and n

Proposition 2.12

Let M be any subset of $\mathbb{R}^n$. The following properties are equivalent:

1. M is a C^k-submanifold of $\mathbb{R}^n$ with dimension n ;
2. M is an open subset of $\mathbb{R}^n$.

Proof. $\boxed{1\Rightarrow 2}$: We assume that M is a C^k-submanifold with dimension n, and show that it is an open set.

Let x be any point of M. We use the "diffeomorphism" definition of submanifolds: let $U\subset\mathbb{R}^n$ be a neighborhood of x, $V\subset\mathbb{R}^n$ a neighborhood of 0, and $\phi: U\to V$ a C^k-diffeomorphism such that

$$\phi(M\cap U) = (\mathbb{R}^n\times\{0\}^{n-n})\cap V = V.$$

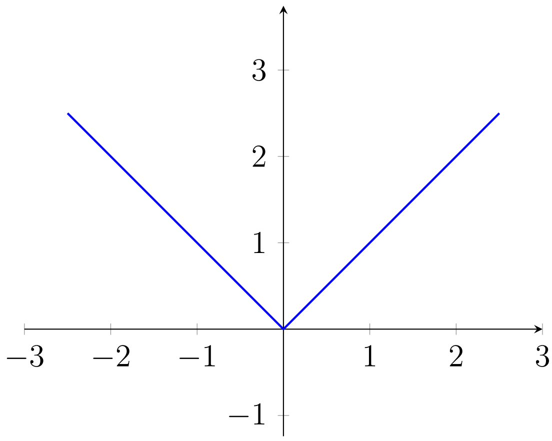


Figure 2.4: The graph of the absolute value is not a submanifold of $\mathbb{R}^2$.

Since ϕ is a bijection from U to V, this equality implies that $M \cap U = U$. Therefore, M contains U, a neighborhood of x. Since this property is true at any point x, M is an open set.

$\boxed{2 \Rightarrow 1}$: We assume that M is an open set, and show that it is a submanifold with dimension n.

Let x be a point in M. We show that M satisfies the "diffeomorphism" definition of submanifolds. We set $U = B(x, r)$, for $r > 0$ small enough so that $U \subset M$. We also set $V = B(0, r)$ and $\phi : y \in U \to y - x \in V$. This map is a diffeomorphism (with reciprocal $(y \in V \to y + x \in U)$). It holds

$$\phi(M \cap U) = \phi(U) = V = (\mathbb{R}^n \times \{0\}^{n-n}) \cap V.$$

□

Proposition 2.13

Let M be any subset of $\mathbb{R}^n$. The following properties are equivalent:

1. M is a C^k-submanifold of $\mathbb{R}^n$ with dimension 0 ;
2. M is a discrete set.[a]

[a] that is, for any $x \in M$, there exists $U \subset \mathbb{R}^n$ a neighborhood of x such that $M \cap U = \{x\}$.

Proof. $\boxed{1 \Rightarrow 2}$: We assume that M is a C^k-submanifold with dimension 0, and show that it is a discrete set.

Let x be any point of M. Let us show that there exists U a neighborhood of x such that $M \cap U = \{x\}$.

We use the "diffeomorphism" definition of submanifolds: let $U \subset \mathbb{R}^n$ be a neighborhood of x, $V \subset \mathbb{R}^n$ a neighborhood of $(0, \dots, 0)$ and $\phi : U \to V$ a C^k-diffeomorphism such that

$$\phi(M \cap U) = (\mathbb{R}^0 \times \{0\}^n) \cap V = \{(0, \dots, 0)\}.$$

As ϕ is injective and $\phi(M \cap U)$ contains only one point, $M \cap U$ itself must be a singleton. Since it contains x, $M \cap U = \{x\}$.

$\boxed{2 \Rightarrow 1}$: We assume that M is a discrete set, and show that it is a submanifold of $\mathbb{R}^n$, of dimension 0.

Let x be any point in M. We show that M satisfies the "diffeomorphism" definition of submanifolds in the neighborhood of x.

Let $U \subset \mathbb{R}^n$ be a neighborhood of x such that $M \cap U = \{x\}$. Let us set $V = \{u - x, u \in U\}$ (the translation of U by $-x$) and $\phi : y \in U \to y - x \in V$. This is a C^∞-diffeomorphism (with reciprocal $(y \in V \to y + x \in U)$). It holds

$$\phi(M \cap U) = \phi(\{x\}) = \{\phi(x)\} = \{(0, \dots, 0)\} = (\mathbb{R}^0 \times \{0\}^n) \cap V.$$

□

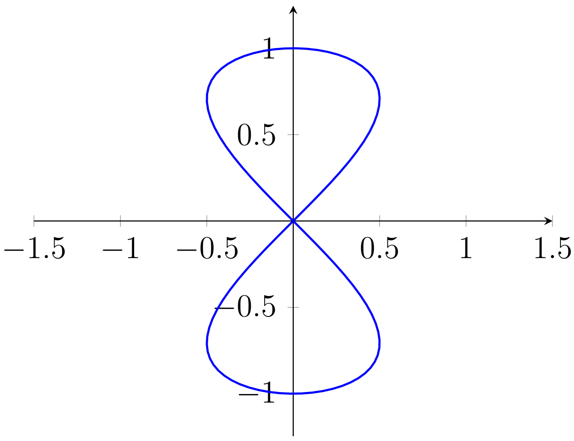


Figure 2.5: The "eight" is not a submanifold of $\mathbb{R}^2$.

2.2.6 Two counterexamples

The graph of the absolute value (Figure 2.4) is not a submanifold of $\mathbb{R}^2$. Intuitively, the reason is that this graph has a “non-regular” point at $(0,0)$.

To prove this rigorously, the simplest way is to proceed by contradiction. Suppose that it is a submanifold and denote its dimension by d. Then, according to the "submersion" definition of submanifolds (Property 3 of Definition 2.1), there exists $U \subset \mathbb{R}^2$ a neighborhood of $(0,0)$ and $g : U \to \mathbb{R}^{2-d}$ a function, at least C^1, submersive at $(0,0)$, such that

$$\{(t,|t|), t \in \mathbb{R}\} \cap U = g^{-1}(\{0\}). \tag{2.4}$$

Such an application g must satisfy, for all t close enough to 0,

$$\begin{aligned} \text{if } t \leq 0, \quad & 0 = g(t,|t|) = g(t,-t), \\ \text{if } t \geq 0, \quad & 0 = g(t,|t|) = g(t,t). \end{aligned}$$

Differentiating these two equalities, we get:

$$\begin{aligned} \partial_1 g(0,0) - \partial_2 g(0,0) = 0; \\ \partial_1 g(0,0) + \partial_2 g(0,0) = 0. \end{aligned}$$

This implies that $\partial_1 g(0,0) = \partial_2 g(0,0) = 0$, i.e., $dg(0,0) = 0$. As $dg(0,0)$ is surjective, this is impossible, unless $\mathbb{R}^{2-d} = \{0\}$, i.e., $d = 2$. But if $d = 2$, then $g^{-1}(\{0\}) = U$, so Equality (2.4) implies that the graph of the absolute value contains a neighborhood of $(0,0)$ in $\mathbb{R}^2$, which is not true. Thus, we reach a contradiction.

The "eight" (Figure 2.5) is also not a submanifold of $\mathbb{R}^2$. Here, the reason is that the eight is a regular curve but with a point of "self-intersection" at zero. This can be rigorously demonstrated using the same method as before.

Remark

This example highlights the importance of the property "f is a homeomorphism onto its image" in the "immersion" definition of submanifolds (Property 2 of Definition 2.1), as well as in Proposition 2.10. Indeed, the eight is equal to $f(]-\pi;\pi[)$, where f is the application

$$\begin{array}{cccc} f & : &]-\pi;\pi[& \to & \mathbb{R}^2 \\ & & \theta & \to & (\sin(\theta)\cos(\theta), \sin(\theta)), \end{array}$$

which is an immersion, and a bijection between $]-\pi;\pi[$ and $f(]-\pi;\pi[)$, but not a homeomorphism (its inverse is not continuous).

2.3 Tangent spaces

2.3.1 Definition

Intuitively, the tangent space to a submanifold M at a point x is the set of directions an ant could take while moving on the surface of M starting from the point x. More formally, the definition is as follows.

Definition 2.14 : tangent space

Let M be a submanifold of $\mathbb{R}^n$, and x a point on M.
The tangent space to M at x, denoted T_xM, is the set of vectors $v \in \mathbb{R}^n$ such that there exists an open interval I containing 0 and $c : I \to \mathbb{R}^n$ a C^1 function satisfying

- $c(t) \in M$ for all $t \in I$;
- $c(0) = x$;
- $c'(0) = v$.

Proposition 2.15

Keeping the notations from the previous definition, the set T_xM is a vector subspace of $\mathbb{R}^n$, with the same dimension as M.

Proof. This is a consequence of the following theorem. □

The four equivalent properties that define the notion of submanifold (Definition 2.1) each provide a way to explicitly compute the tangent space.

Theorem 2.16 : computing the tangent space

Let M be a submanifold of $\mathbb{R}^n$, and x a point on M. Let d be the dimension of M.

1. (Computation by diffeomorphism)

   If U and V are neighborhoods of x and 0 in $\mathbb{R}^n$, respectively, and $\phi : U \to V$ is a C^k-diffeomorphism such that $\phi(x) = 0$ and $\phi(M \cap U) = (\mathbb{R}^d \times \{0\}^{n-d}) \cap V$, then
   $$T_xM = d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d}).$$

2. (Computation by immersion)

   If U is a neighborhood of x in $\mathbb{R}^n$, V an open set in $\mathbb{R}^d$, and $f : V \to \mathbb{R}^n$ a C^k map, which is a homeomorphism between V and $f(V)$, such that $M \cap U = f(V)$ and f is an immersion at $z_0 \overset{def}{=} f^{-1}(x)$, then
   $$T_xM = df(z_0)(\mathbb{R}^d)(= \mathrm{Im}(df(z_0)))$$

3. (Computation by submersion)

   If U is a neighborhood of x and $g : U \to \mathbb{R}^{n-d}$ a C^k map surjective at x such that $M \cap U = g^{-1}(\{0\})$, then
   $$T_xM = \mathrm{Ker}(dg(x)).$$

4. (Computation by graph)

   If U is a neighborhood of x, V an open set in $\mathbb{R}^d$, and $h : V \to \mathbb{R}^{n-d}$ is a C^k map such that, in a well-chosen coordinate system, $M \cap U = \mathrm{graph}(h)$, then
   $$T_xM = \{(t_1, \dots, t_d, dh(x_1, \dots, x_d)(t_1, \dots, t_d)), t_1, \dots, t_d \in \mathbb{R}\}.$$

Proof. Let's begin with Property 1. Let U, V, and ϕ be as stated in the property.

First, let's prove the inclusion $T_xM \subset d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d})$. Let v be an arbitrary element in T_xM; we will show that it belongs to $d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d})$.

Let c be as in the definition of the tangent space, i.e., c is a C^1 function from an open interval I in $\mathbb{R}$ containing 0 to $\mathbb{R}^n$, with images in M, such that $c(0) = x$ and $c'(0) = v$.

For any t close enough to 0, $c(t)$ belongs to U, so $\phi(c(t))$ is well-defined. Moreover, since $\phi(M \cap U) \subset \mathbb{R}^d \times \{0\}^{n-d}$, we must have

$$0 = \phi(c(t))_{d+1} = \cdots = \phi(c(t))_n.$$

Differentiating these equalities at $t = 0$ gives:

$$\begin{aligned}
0 &= d\phi(c(0))(c'(0))_{d+1} = d\phi(x)(v)_{d+1}, \\
&\ \dots \\
0 &= d\phi(x)(v)_n.
\end{aligned}$$

Therefore, $d\phi(x)(v) \in \mathbb{R}^d \times \{0\}^{n-d}$, i.e., $v \in d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d})$.

Now, let's prove the other inclusion: $d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d}) \subset T_xM$. Let $v \in d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d})$; we will show that $v \in T_xM$.

Denote

$$w = d\phi(x)(v) \in \mathbb{R}^d \times \{0\}^{n-d}.$$

We must find a function c as in the definition of the tangent space. We will define it as the preimage by ϕ of a function γ with images in $\mathbb{R}^n$ such that $\gamma(0) = 0$ and $\gamma'(0) = w$.

Choose an open interval I containing 0 small enough, and define

$$\begin{array}{cccc}
\gamma & : & I & \to & \mathbb{R}^n \\
 & & t & \to & tw.
\end{array}$$

This is a C^∞ function satisfying

$$\gamma(0) = 0 \quad \text{and} \quad \gamma'(0) = w.$$

If I is small enough, $\gamma(I) \subset V$. Thus, we can define

$$c = \phi^{-1} \circ \gamma : I \to \mathbb{R}^n.$$

This is a C^k function. It takes values in M because $\gamma(t) \in \mathbb{R}^d \times \{0\}^{n-d}$ for all $t \in I$ (since $w \in \mathbb{R}^d \times \{0\}^{n-d}$). Therefore,

$$c(t) \in \phi^{-1}\left(\left(\mathbb{R}^d \times \{0\}^{n-d}\right) \cap V\right) = M \cap U.$$

Moreover,

$$c(0) = \phi^{-1}(\gamma(0)) = \phi^{-1}(0) = x$$

and

$$\begin{aligned}
w &= \gamma'(0) \\
&= (\phi \circ c)'(0) \\
&= d\phi(c(0))(c'(0)) \\
&= d\phi(x)(c'(0)).
\end{aligned}$$

Therefore,

$$c'(0) = d\phi(x)^{-1}(w) = v.$$

The function c satisfies the required properties in the definition of the tangent space. Therefore,

$$v \in T_xM.$$

This completes the proof of the equality

$$T_xM = d\phi(x)^{-1}(\mathbb{R}^d \times \{0\}^{n-d}).$$

Before proving the remaining three properties of the theorem, let's observe that the equality we have just obtained already shows that T_xM is a vector subspace of $\mathbb{R}^n$ of dimension d. Indeed, it is the image of a vector subspace of dimension d of $\mathbb{R}^n$ ($\mathbb{R}^d \times \{0\}^{n-d}$) under a linear isomorphism ($d\phi(x)^{-1}$).

This observation simplifies the proof of properties 2, 3, and 4. Indeed, the sets

$$\begin{gathered} df(z_0)(\mathbb{R}^d), \mathrm{Ker}(dg(x)) \\ \text{and } \{(t_1, \dots, t_d, dh(x_1, \dots, x_d)(t_1, \dots, t_d)), t_1, \dots, t_d \in \mathbb{R}\}, \end{gathered}$$

which appear in these properties, are vector subspaces of $\mathbb{R}^n$ of dimension d (the first is the image of $\mathbb{R}^d$ by an injective linear map, the second is the kernel of a surjective linear map from $\mathbb{R}^n$ to $\mathbb{R}^{n-d}$, and the third is generated by the following free family of d elements:

$$\begin{gathered} (1, 0, \dots, 0, dh(x_1, \dots, x_d)(1, 0, \dots, 0)), \\ \dots, \\ (0, \dots, 0, 1, dh(x_1, \dots, x_d)(0, \dots, 0, 1))). \end{gathered}$$

To show that they are equal to T_xM, it is therefore sufficient to prove either

- that they contain T_xM,
- or that they are included in T_xM.

Let's prove Property 2. Let U, V, and f be as in the statement of the property. We will show that

$$df(z_0)(\mathbb{R}^d) \subset T_xM. \tag{2.5}$$

Let $v \in df(z_0)(\mathbb{R}^d)$ be arbitrary; let's show that $v \in T_xM$. Let $a \in \mathbb{R}^d$ be such that $df(z_0)(a) = v$. Choose an interval $I \subset \mathbb{R}$ containing 0, small enough, and define

$$\begin{array}{rccc} c & : & I & \to & \mathbb{R}^n \\ & & t & \to & f(z_0 + ta). \end{array}$$

The function c is well-defined if I is small enough, as $z_0 + ta \in V$ for all $t \in I$. It is a C^k (thus C^1) function. For all $t \in I$, $c(t) \in f(V) \subset M$. Moreover,

$$c(0) = f(z_0) = x$$

and

$$c'(0) = df(z_0)(a) = v.$$

This shows that $v \in T_xM$. Thus, Equation (2.5) is true.

Now let's prove Property 3. Let U and g be as in the statement of the property. We will show that

$$T_xM \subset \mathrm{Ker}(dg(x)).$$

Let $v \in T_xM$ be arbitrary. Let us show that v is in $\mathrm{Ker}(dg(x))$. Let I be an interval in $\mathbb{R}$ containing 0, and let $c : I \to \mathbb{R}^n$ be as in the definition of the tangent space.

For any t close enough to 0, $c(t)$ is an element of U; it is also an element of M. Since $M \cap U = g^{-1}(\{0\})$,

$$0 = g(c(t)).$$

Differentiating this equality at 0,

$$0 = dg(c(0))(c'(0)) = dg(x)(v).$$

Therefore, $v \in \mathrm{Ker}(dg(x))$.

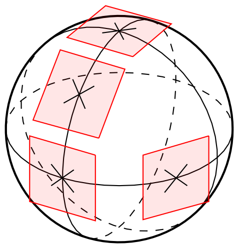


Figure 2.6: The sphere $\mathbb{S}^2$ and its affine tangent space at a few points.

Finally, let's prove Property 4. Let U, V, and h be as in the statement of this property. Let

$$E = \{(t_1, \dots, t_d, dh(x_1, \dots, x_d)(t_1, \dots, t_d)), t_1, \dots, t_d \in \mathbb{R}\}$$

We show that

$$E \subset T_x M.$$

Let $(t, dh(x_1, \dots, x_d)(t)) \in E$, with $t \in \mathbb{R}^d$. Let us show that this is an element of $T_x M$.

Choose an interval I in $\mathbb{R}$ containing 0 small enough, and define

$$\begin{array}{rcccc} c & : & I & \to & \mathbb{R}^n \\ & & s & \to & ((x_1, \dots, x_d) + st, h((x_1, \dots, x_d) + st)). \end{array}$$

This function is well-defined if I is small enough, as $(x_1, \dots, x_d) + st$ belongs to V for all $s \in I$ (since V contains $(x_1, \dots, x_d)$ and is open). It is of class C^k (thus C^1). It is in the graph of h, and therefore in M. Moreover,

$$c(0) = (x_1, \dots, x_d, h(x_1, \dots, x_d)) = x$$

and

$$c'(0) = (t, dh(x_1, \dots, x_d)(t)).$$

This shows that $(t, dh(x_1, \dots, x_d)(t)) \in T_x M$.

□

To finish with the definitions, let's introduce the affine tangent space, which is simply the tangent space, translated so that it goes through the point x. This is not a notion that we will heavily use in the rest of the course, except in the figures: it is much more natural to draw tangent spaces that actually touch[1] the submanifold they are associated with than tangent spaces which all contain 0.

Definition 2.17

If M is a submanifold of $\mathbb{R}^n$ and $x \in M$, the *affine tangent space to* M *at* x is defined as the set

$$x + T_x M.$$

2.3.2 Examples

In this paragraph, we go back to the examples of submanifolds from Section 2.2 and compute their tangent spaces.

Proposition 2.18 : tangent space of the sphere

For any $x \in \mathbb{S}^{n-1}$,

$$T_x \mathbb{S}^{n-1} = \{x\}^\perp = \{t \in \mathbb{R}^n, \langle t, x \rangle = 0\}.$$

[1] The word "tangent" comes from the Latin verb *tangere*, which means "to touch".

Proof. Let's define, as in Subsection 2.2.1,

$$\begin{array}{rccc} g & : & \mathbb{R}^n & \to & \mathbb{R} \\ & & (t_1, \dots, t_n) & \to & t_1^2 + \dots + t_n^2 - 1. \end{array}$$

It satisfies $\mathbb{S}^{n-1} = g^{-1}(\{0\})$ and is a submersion at x. According to Property 3 of Theorem 2.16,

$$T_x\mathbb{S}^{n-1} = \text{Ker}(dg(x)).$$

Now, for any $t \in \mathbb{R}^n$, $dg(x)(t) = 2\langle x, t\rangle$. Therefore,

$$T_x\mathbb{S}^{n-1} = \{x\}^\perp.$$

□

Proposition 2.19 : tangent space of a product submanifold

Let $n_1, n_2 \in \mathbb{N}^*$. Suppose M_1 is a submanifold of $\mathbb{R}^{n_1}$ and M_2 is a submanifold of $\mathbb{R}^{n_2}$. For any $x = (x_1, x_2) \in M_1 \times M_2$,

$$\begin{aligned} T_x(M_1 \times M_2) &= T_{x_1}M_1 \times T_{x_2}M_2 \\ &= \{(t_1, t_2), t_1 \in T_{x_1}M_1, t_2 \in T_{x_2}M_2\}. \end{aligned}$$

Proof. Let $x = (x_1, x_2) \in M_1 \times M_2$.

We will use the expression for the tangent space associated with the "immersion" definition of submanifolds (Property 2 of Theorem 2.16).

Let d_1 be the dimension of M_1. Suppose U_1 is a neighborhood of x_1 in $\mathbb{R}^{n_1}$, V_1 is a neighborhood of 0 in $\mathbb{R}^{d_1}$, and $f_1 : V_1 \to \mathbb{R}^{n_1}$ is a function which is a homeomorphism onto its image, such that

$$M_1 \cap U_1 = f_1(V_1)$$

and f_1 is immersive at z_1, where $z_1 \in V_1$ is the point such that $f_1(z_1) = x_1$.

Define similarly $d_2, U_2, V_2, f_2 : V_2 \to \mathbb{R}^{n_2}$ and z_2.

According to Property 2 of Theorem 2.16, we have

$$T_{x_1}M_1 = df_1(z_1)(\mathbb{R}^{d_1}) \quad \text{and} \quad T_{x_2}M_2 = df_2(z_2)(\mathbb{R}^{d_2}).$$

Moreover, as shown in the proof of Proposition 2.5, the function $f : (t_1, t_2) \in V_1 \times V_2 \to (f_1(t_1), f_2(t_2)) \in \mathbb{R}^{n_1+n_2}$ is a homeomorphism onto its image, satisfies

$$f(V_1 \times V_2) = (M_1 \times M_2) \cap (U_1 \times U_2)$$

and is immersive at $(z_1, z_2) = f^{-1}(x)$. From Property 2 of Theorem 2.16, we have

$$\begin{aligned} T_x(M_1 \times M_2) &= df(z_1, z_2)(\mathbb{R}^{d_1+d_2}) \\ &= \{df(z_1, z_2)(t_1, t_2), t_1 \in \mathbb{R}^{d_1}, t_2 \in \mathbb{R}^{d_2}\} \\ &= \{(df_1(z_1)(t_1), df_2(z_2)(t_2)), t_1 \in \mathbb{R}^{d_1}, t_2 \in \mathbb{R}^{d_2}\} \\ &= df_1(z_1)(\mathbb{R}^{d_1}) \times df_2(z_2)(\mathbb{R}^{d_2}) \\ &= T_{x_1}M_1 \times T_{x_2}M_2. \end{aligned}$$

□

Example 2.20 : tangent space of the torus

For any $(x_1, x_2) \in \mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$,

$$T_{(x_1,x_2)}\mathbb{T}^2 = T_{x_1}\mathbb{S}^1 \times T_{x_2}\mathbb{S}^1 = \{x_1\}^\perp \times \{x_2\}^\perp.$$

If we fix θ_1, θ_2 such that $x_1 = (\cos(\theta_1), \sin(\theta_1)), x_2 = (\cos(\theta_2), \sin(\theta_2))$, we have

$$\begin{aligned}\{x_1\}^\perp &= (\sin(\theta_1), -\cos(\theta_1))\mathbb{R} \\ &= \{(t_1 \sin(\theta_1), -t_1 \cos(\theta_1)), t_1 \in \mathbb{R}\}\end{aligned}$$

and similarly for x_2. This allows us to write the previous expression for the tangent to the torus in a slightly more explicit way:

$$T_{(x_1,x_2)}\mathbb{T}^2 = \{(t_1 \sin(\theta_1), -t_1 \cos(\theta_1), t_2 \sin(\theta_2), -t_2 \cos(\theta_2)), t_1, t_2 \in \mathbb{R}\}.$$

Proposition 2.21 : tangent space of the orthogonal group

For any $G \in O_n(\mathbb{R})$,

$$T_G O_n(\mathbb{R}) = \{GR, R \in \mathbb{R}^{n\times n} \text{ is antisymmetric}\}.$$

Proof. Let $G \in O_n(\mathbb{R})$.

As shown in the proof of Proposition 2.8, $O_n(\mathbb{R})$ is equal to $\tilde{g}^{-1}(\{0\})$, where $\tilde{g}$ is defined at

$$\begin{array}{rccc}\tilde{g} & : & \mathbb{R}^{n\times n} & \to & \mathbb{R}^{\frac{n(n+1)}{2}} \\ & & A & \to & \mathrm{Tri}({}^tAA - I_n),\end{array}$$

is a submersion at G, with differential

$$d\tilde{g}(G) : A \in \mathbb{R}^{n\times n} \to \mathrm{Tri}({}^tGA + {}^tAG) \in \mathbb{R}^{\frac{n(n+1)}{2}}.$$

According to Property 3 of Theorem 2.16,

$$T_G O_n(\mathbb{R}) = \mathrm{Ker}(d\tilde{g}(G)) = \left\{A \in \mathbb{R}^{n\times n}, \mathrm{Tri}({}^tGA + {}^tAG) = 0\right\}.$$

Now, for any A,

$$\begin{aligned}\mathrm{Tri}({}^tGA + {}^tAG) = 0 &\iff {}^tGA + {}^tAG = 0 \\ &\qquad \text{(because } {}^tGA + {}^tAG \text{ is symmetric)} \\ &\iff ({}^tGA) + {}^t({}^tGA) = 0 \\ &\iff {}^tGA = R \text{ for some antisymmetric } R \\ &\iff A = GR \text{ for some antisymmetric } R \\ &\qquad \text{(because } G{}^tG = I_n).\end{aligned}$$

Therefore,

$$T_G O_n(\mathbb{R}) = \{GR, R \in \mathbb{R}^{n\times n} \text{ is antisymmetric}\}.$$

□

Proposition 2.22

Let $d \in \{0, \ldots, n\}$. Let U be an open set in $\mathbb{R}^n$, and $g : U \to \mathbb{R}^{n-d}$ be a C^k function. Assume that g is a submersion on $g^{-1}(\{0\})$.

For any $x \in g^{-1}(\{0\})$,

$$T_x(g^{-1}(\{0\})) = \mathrm{Ker}(dg(x)).$$

Proof. This is a direct application of Property 3 of Theorem 2.16. □

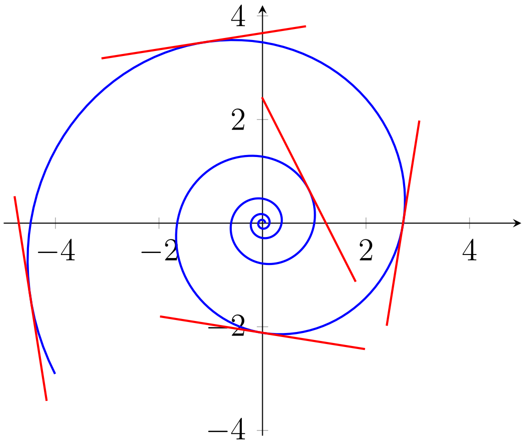


Figure 2.7: The spiral from Example 2.24 and its affine tangent space at a few points.

Proposition 2.23

Let $d \in \{0, \dots, n\}$. Let U be an open set in $\mathbb{R}^d$, and $f : U \to \mathbb{R}^n$ be an immersion, which is a homeomorphism from U to $f(U)$.
For any $x \in f(U)$,

$$T_x f(U) = df(z)(\mathbb{R}^d),$$

where z is the element of U such that $x = f(z)$.

Proof. This is a direct application of Property 2 of Theorem 2.16. □

Example 2.24 : tangent space of the spiral

Consider the function from Example 2.11:

$$\begin{array}{cccc} f & : & \mathbb{R} & \to & \mathbb{R}^2 \\ & & \theta & \to & \left(e^\theta \cos(2\pi\theta), e^\theta \sin(2\pi\theta)\right). \end{array}$$

Let $(x, y) \in f(\mathbb{R})$. Denote $\theta \in \mathbb{R}$ the real number such that $(x, y) = f(\theta)$. According to Proposition 2.23:

$$\begin{aligned} T_{(x,y)} f(\mathbb{R}) &= f'(\theta)\mathbb{R} \\ &= e^\theta((\cos(2\pi\theta), \sin(2\pi\theta)) + 2\pi(-\sin(2\pi\theta), \cos(2\pi\theta)))\mathbb{R} \\ &= (x - 2\pi y, y + 2\pi x)\mathbb{R} \\ &= \{((x - 2\pi y)t, (y + 2\pi x)t), t \in \mathbb{R}\}. \end{aligned}$$

An illustration is shown on Figure 2.7.

2.3.3 Application: proof that a set is not a submanifold

The "eight" Let us go back to the second set considered in Subsection 2.2.6, the "eight", represented on Figure 2.5. This set is

$$M \overset{def}{=} \{f(\theta), \theta \in]-\pi;\pi[\}.$$

where f is defined as

$$\begin{array}{rccc} f & : &]-\pi;\pi[& \to & \mathbb{R}^2 \\ & & \theta & \to & (\sin(\theta)\cos(\theta), \sin(\theta)). \end{array}$$

Here, we prove that M is not a submanifold of $\mathbb{R}^2$, using a different technique from Subsection 2.2.6.

By contradiction, let us assume that it is a submanifold. We compute its tangent space at $(0,0)$.

First, we define

$$c_1 = f :]-\pi;\pi[\to \mathbb{R}^2.$$

It holds $c_1(t) \in M$ for all $t \in]-\pi;\pi[$, $c_1(0) = (0,0)$ and c_1 is C^1. Therefore,

$$(1,1) = c_1'(0) \in T_{(0,0)}M. \tag{2.6}$$

Second, we define

$$\begin{array}{rccc} c_2 & : &]-\pi;\pi[& \to & \mathbb{R}^2 \\ & & \theta & \to & (\sin(\theta)\cos(\theta), -\sin(\theta)). \end{array}$$

It holds $c_2(t) \in M$ for all $t \in]-\pi;\pi[$. Indeed, for any $t \in]-\pi;0[$, $c_2(t) = f(t+\pi) \in M$; $c_2(0) = f(0) \in M$ and, for any $t \in]0;\pi[$, $c_2(t) = f(t-\pi) \in M$. In addition, $c_2(0) = (0,0)$ and c_2 is C^1. Therefore,

$$(1,-1) = c_2'(0) \in T_{(0,0)}M. \tag{2.7}$$

As $T_{(0,0)}M$ is a vector subspace of $\mathbb{R}^2$, Equations (2.6) and (2.7) together imply that

$$T_{(0,0)}M = \mathbb{R}^2.$$

In particular, since the dimension of the tangent space is the same as the dimension of the submanifold, $\dim M = 2$. In virtue of Proposition 2.12, M must thus be an open set of $\mathbb{R}^2$. As this is not true (because, for instance, M contains no element of the form $(t,0)$, except $(0,0)$ itself, so it does not contain a neighborhood of $(0,0)$), we have reached a contradiction.

Graph of the absolute value The proof we have presented for the "eight" does not apply to the graph of the absolute value (try it!). However, used differently, the notion of tangent space also allows to prove that the graph is not a submanifold of $\mathbb{R}^2$.[2]

The graph of the absolute value (Figure 2.4) is the set

$$M = \{(x, |x|), x \in \mathbb{R}\}.$$

By contradiction, let us assume that it is a submanifold of $\mathbb{R}^2$.

We show that $T_{(0,0)}M = \{0\}$. Let v be any element of $T_{(0,0)}M$. From the definition of the tangent space, there exist I a open interval containing 0 and $c : I \to \mathbb{R}^2$ a C^1 map such that $c(I) \subset M$, $c(0) = (0,0)$ and $c'(0) = v$. Let us fix such I, c.

Let us denote (c_1, c_2) the components of c. It holds, for any $t \in I$,

$$c_2(t) = |c_1(t)|.$$

In particular, $c_2'(0)t + o(t) = |c_1'(0)t + o(t)| = |c_1'(0)||t| + o(t)$. This means that, for $t \in I \cap \mathbb{R}^+$,

$$c_2'(0)t + o(t) = |c_1'(0)|t + o(t)$$

and, for $t \in I \cap \mathbb{R}^-$,

$$c_2'(0)t + o(t) = -|c_1'(0)|t + o(t).$$

[2] But the proof is not very different from what we could have done using the "immersion" definition of submanifolds.

As Taylor series are unique, we must have

$$c_2'(0) = |c_1'(0)| = -|c_1'(0)|,$$

which implies that $c_1'(0) = c_2'(0) = 0$ and, therefore,

$$v = c'(0) = (0, 0).$$

This shows that $T_{(0,0)}M$ contains no other vector than 0, hence $T_{(0,0)}M = \{0\}$.

Since each tangent space to M is a vector space with the same dimension as M, the dimension of M must be zero. From Proposition 2.13, M must be discrete. This is not true: we have reached a contradiction.

2.4 Maps between submanifolds

In this section, we consider functions between two submanifolds $M \subset \mathbb{R}^{n_1}$ and $N \subset \mathbb{R}^{n_2}$:

$$f : M \to N.$$

If $M = \mathbb{R}^{d_1} \times \{0\}^{n_1-d_1}$ and $N = \mathbb{R}^{d_2} \times \{0\}^{n_2-d_2}$, f is essentially a function from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$. The notions of "differentiability" and "differential" are then well-defined for f, in accordance with Chapter 1.

However, if M is not a vector subspace of $\mathbb{R}^{n_1}$, this is no longer the case: Definition 1.1 involves linear maps between the domain and codomain, which do not exist if the sets are not vector spaces.

To give a meaning to the notion of "differentiability" for f, one can use the fact that M and N are identifiable with open sets in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ through diffeomorphisms. We say that f is differentiable if, when composed with these diffeomorphisms, it is a differentiable map from an open set in $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$. This is, in a slightly different form, the content of the following definition.

Definition 2.25 : C^1 map from a submanifold to $\mathbb{R}^m$

Let $m \in \mathbb{N}$.
Consider M a C^k submanifold of $\mathbb{R}^n$, and a function

$$f : M \to \mathbb{R}^m.$$

We say that f is *of class* C^1 if, for any integer $s \in \mathbb{N}^*$, any open set V in $\mathbb{R}^s$, and any C^1 function $\phi : V \to \mathbb{R}^n$ such that $\phi(V) \subset M$, the function

$$f \circ \phi : V \to \mathbb{R}^m$$

is of class C^1.

Remark

Similarly, one can define the notion of *function of class* C^r from M to $\mathbb{R}^m$, for any $r = 1, \dots, k$. Simply replace "C^1" with "C^r" in the above definition.
It can be shown that a function of class C^r is necessarily of class $C^{r'}$ for any $r' \leq r$.

Example 2.26 : projection onto a coordinate

Let $M \subset \mathbb{R}^n$ be a C^k-submanifold. For any $r = 1, \dots, n$, we define the projection onto the r-th coordinate

$$\begin{array}{cccc} \pi_r & : & M & \to & \mathbb{R} \\ & & (x_1, \dots, x_n) & \to & x_r. \end{array}$$

This is a C^k map.

Proof. Let $r \in \{1, \ldots, n\}$. Take $s \in \mathbb{N}^*$, V an open set in $\mathbb{R}^s$, and $\phi : V \to \mathbb{R}^n$ of class C^k such that $\phi(V) \subset M$. For any $x \in \mathbb{R}^s$, denote $\phi(x) = (\phi_1(x), \ldots, \phi_n(x))$. The functions $\phi_1, \ldots, \phi_n$ are C^k. Hence, $\pi_r \circ \phi = \phi_r$ is C^k. □

Definition 2.27 : C^1 function between two submanifolds

Let M, N be two C^k submanifolds, respectively of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$. Consider a function

$$f : M \to N.$$

Since $N \subset \mathbb{R}^{n_2}$, we can view f as a map from M to $\mathbb{R}^{n_2}$ rather than from M to N. We say that f is *of class* C^1 (more generally, C^r, for $r \in \{1, \ldots, k\}$) between M and N if it is of class C^1 (more generally, C^r) when viewed as a map from M to $\mathbb{R}^{n_2}$.

Example 2.28 : projection on a product submanifold

Let A, B be two C^k-submanifolds, respectively of $\mathbb{R}^a$ and $\mathbb{R}^b$. Recall that $A \times B$ is a submanifold of $\mathbb{R}^{a+b}$ (Proposition 2.5).
We define the projection onto A as

$$\begin{array}{cccl} \pi_A & : & A \times B & \to A \\ & & (x_A, x_B) & \to x_A. \end{array}$$

This is a C^k function.
Similarly, the projection onto B is C^k.

Proof. Consider π_A as a function from $A \times B$ to $\mathbb{R}^a$ and show that this function is C^k. Take $s \in \mathbb{N}^*$, V an open set in $\mathbb{R}^s$, and $\phi : V \to \mathbb{R}^{a+b}$ a C^k map such that $\phi(V) \subset A \times B$.

For any $x \in \mathbb{R}^s$, denote $\phi(x) = (\phi_1(x), \ldots, \phi_{a+b}(x))$. The functions $\phi_1, \ldots, \phi_{a+b}$ are C^k. The function $\pi_A \circ \phi$ is given by

$$\begin{aligned} \forall x \in \mathbb{R}^s, \quad \pi_A \circ \phi(x) &= \pi_A(\underbrace{\phi_1(x), \ldots, \phi_a(x)}_{\text{element of } A}, \underbrace{\phi_{a+1}(x), \ldots, \phi_{a+b}(x)}_{\text{element of } B}) \\ &= (\phi_1(x), \ldots, \phi_a(x)). \end{aligned}$$

Thus, $\pi_A \circ \phi$ is equal to $(\phi_1, \ldots, \phi_a)$, which is C^k, and consequently, $\pi_A \circ \phi$ is C^k. □

Definitions 2.25 and 2.27 are more abstract than the definition of differentiability for a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. However, one must not be intimidated. In practice, one rarely needs to resort to these definitions to show that a function is of class C^1 (or, more generally, C^r). Indeed, as is the case for functions from $\mathbb{R}^n \to \mathbb{R}^m$, basic operations on functions between manifolds preserve differentiability. For example, if M is a submanifold and m an integer, the sum of two C^r functions from M to $\mathbb{R}^m$ is also C^r. Similarly, the product of two C^r functions from M to $\mathbb{R}$ is also C^r. We will not state each of these properties here; let us focus on the one related to function composition.

Proposition 2.29 : composition of C^1 functions

Let M, N, P be three C^k submanifolds of, respectively, $\mathbb{R}^{n_M}$, $\mathbb{R}^{n_N}$, and $\mathbb{R}^{n_P}$. Consider two functions

$$f_1 : M \to N \quad \text{and} \quad f_2 : N \to P.$$

If f_1 and f_2 are of class C^r, for some $r \in \{1, \ldots, k\}$, then

$$f_2 \circ f_1 : M \to P$$

is also of class C^r.

Proof. We view $f_2 \circ f_1$ as a function from M to $\mathbb{R}^{n_P}$ and show that this function is C^r. Let $s \in \mathbb{N}^*$ be an integer, V an open set in $\mathbb{R}^s$ and $\phi : V \to \mathbb{R}^{n_M}$ a C^r function such that $\phi(V) \subset M$. We must show that $f_2 \circ f_1 \circ \phi$ is of class C^r on V.

Since $f_1 : M \to N$ is of class C^r, it is also C^r when viewed as a function from M to $\mathbb{R}^{n_N}$. From Definition 2.25, $f_1 \circ \phi : V \to \mathbb{R}^{n_N}$ is C^r. Moreover, $(f_1 \circ \phi)(V) \subset f_1(M) \subset N$. As $f_2 : N \to P \subset \mathbb{R}^{n_P}$ is C^r, the function $f_2 \circ (f_1 \circ \phi)$ is C^r, also from Definition 2.25.

Since $f_2 \circ f_1 \circ \phi = f_2 \circ (f_1 \circ \phi)$, this proves that $f_2 \circ f_1 \circ \phi$ is C^r. ☐

Exercise 2

Show that the map

$$f : \begin{array}{ccc} \mathbb{S}^1 & \to & \mathbb{S}^1 \\ (x_1, x_2) & \to & (x_1^2, x_2\sqrt{1+x_1^2}) \end{array}$$

is well-defined and C^∞.

Solution. Showing that f is well-defined consists in showing that $f(x_1, x_2)$ indeed belongs to $\mathbb{S}^1$ for all $(x_1, x_2) \in \mathbb{S}^1$. Let us consider any $(x_1, x_2) \in \mathbb{S}^1$. It holds

$$\begin{aligned} \left(x_1^2\right)^2 + \left(x_2\sqrt{1+x_1^2}\right)^2 &= x_1^4 + x_2^2(1+x_1^2) \\ &= x_1^2(x_1^2 + x_2^2) + x_2^2 \\ &= x_1^2 + x_2^2 \\ &= 1. \end{aligned}$$

Therefore, $f(x_1, x_2) \in \mathbb{S}^1$.

Let us now show that f is C^∞. From Definition 2.27, we must show that

$$\tilde{f} : \begin{array}{ccc} \mathbb{S}^1 & \to & \mathbb{R}^2 \\ (x_1, x_2) & \to & (x_1^2, x_2\sqrt{1+x_1^2}) \end{array}$$

is C^∞. From Example 2.26, we know that

$$\pi_1 \times \pi_2 : \begin{array}{ccc} \mathbb{S}^1 & \to & \mathbb{R}^2 \\ (x_1, x_2) & \to & (x_1, x_2) \end{array}$$

is C^∞. As $\tilde{f}$ is the composition of $\pi_1 \times \pi_2$ with the map

$$g : \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R}^2 \\ (x_1, x_2) & \to & (x_1^2, x_2\sqrt{1+x_1^2}), \end{array}$$

which is C^∞ (it is a composition of $\sqrt{\cdot} : \mathbb{R}_+^* \to \mathbb{R}$, which is C^∞ on this domain, and polynomial functions). From Proposition 2.29, $\tilde{f}$ is C^∞. ☐

Note that, unlike the case where the functions considered go from $\mathbb{R}^n$ to $\mathbb{R}^m$, we have defined the notion of *differentiable function* between manifolds without introducing the notion of *differential*. Nevertheless, one can still define this notion; this is the aim of the following definition.

Definition 2.30 : differential on manifolds

Let M, N be two C^k submanifolds of, respectively, $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$. Let

$$f : M \to N$$

be a C^r function, where $r \in \{1, \dots, k\}$.

Let $x \in M$. For any $v \in T_xM$, fix I_v an open interval in $\mathbb{R}$ containing 0 and $c_v : I \to \mathbb{R}^{n_1}$ as in the definition of the tangent space (2.14), i.e., a C^1 function with values in M such that $c_v(0) = x$ and $c_v'(0) = v$.

The *differential of f at x*, denoted $df(x)$, is the following map:

$$\begin{array}{cccc} df(x) & : & T_xM & \to & T_{f(x)}N \\ & & v & \to & (f\circ c_v)'(0). \end{array}$$

The map $df(x)$ is well-defined: $f\circ c_v : I_v \to \mathbb{R}^{n_2}$ is a C^1 function, with values in N, such that $f\circ c_v(0) = f(x)$, so $(f\circ c_v)'(0)$ is indeed an element of $T_{f(x)}N$.

Remark

If M is an open subset of $\mathbb{R}^{n_1}$, then f, viewed as a function from this open subset of $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_2}$, is differentiable in the usual sense, and the differentials defined in Definitions 1.1 and 2.30 coincide, as in that case, denoting $df(x)$ the usual differential,

$$(f\circ c_v)'(0) = df(c_v(0))(c_v'(0)) = df(x)(v).$$

Theorem 2.31

We keep the notations from Definition 2.30.
The map $df(x)$ does not depend on the choice of intervals I_v and functions c_v.
Moreover, it is linear.

Proof. Let $v \in T_xM$. Show that $df(x)(v) = (f\circ c_v)'(0)$ does not depend on the choice of I_v and c_v. To do this, we will give an alternative expression for $df(x)(v)$ that does not involve I_v or c_v.

Let d_1 and d_2 be the dimensions of M and N. Let $U_M, V_M \subset \mathbb{R}^{n_1}$ be neighborhoods of x and 0, respectively, and $\phi_M : U_M \to V_M$ be a C^k-diffeomorphism such that $\phi_M(x) = 0$ and

$$\phi_M(M\cap U_M) = (\mathbb{R}^{d_1}\times\{0\}^{n_1-d_1})\cap V_M.$$

Denote $\phi_{M,0}^{-1}$ the restriction of ϕ_M^{-1} to $(\mathbb{R}^{d_1}\times\{0\}^{n_1-d_1})\cap V_M$. We have

$$\begin{aligned} df(x)(v) &= (f\circ c_v)'(0) \\ &= (f\circ\phi_{M,0}^{-1}\circ\phi_M\circ c_v)'(0) \\ &= ((f\circ\phi_{M,0}^{-1})\circ\phi_M\circ c_v)'(0). \end{aligned}$$

The function $f\circ\phi_{M,0}^{-1}$ is defined on an open subset of $\mathbb{R}^{d_1}$ (actually, on $(\mathbb{R}^{d_1}\times\{0\}^{n_1-d_1})\cap V_M$, but this is exactly an open set of $\mathbb{R}^{d_1}$ if one ignores the (n_1-d_1) zeros). It is of class C^r on this subset, since it is the composition of two C^r functions. Thus, the functions $f\circ\phi_{M,0}^{-1}, \phi_M$ and c_v are defined on open subsets of $\mathbb{R}^n$ (for different values of n) and differentiable in the usual sense. The usual theorem on the composition of differentials then gives

$$\begin{aligned} df(x)(v) &= (d(f\circ\phi_{M,0}^{-1})(\phi_M\circ c_v(0))\circ d\phi_M(c_v(0)))(c_v'(0)) \\ &= d(f\circ\phi_{M,0}^{-1})(0)\circ d\phi_M(x)(v). \end{aligned}$$

As announced, this expression does not depend on c_v, which completes the first part of the proof.

The linearity of $df(x)$ follows from the same argument. Indeed, the reasoning we have just done has shown that

$$df(x) = d(f\circ\phi_{M,0}^{-1})(0)\circ d\phi_M(x),$$

i.e., $df(x)$ is the composition of two linear maps. Therefore, it is linear. $\square$

As before, the notion of differential for functions between manifolds is governed by almost the same rules as for functions between $\mathbb{R}^m$ and $\mathbb{R}^n$. Let's state, for example, the rule of composition of differentials.

Proposition 2.32

Let M, N, P be three C^k submanifolds of $\mathbb{R}^{n_M}$, $\mathbb{R}^{n_N}$, and $\mathbb{R}^{n_P}$, respectively. Consider two C^r maps, for $r \in \{1, \dots, k\}$,

$$f_1 : M \to N \quad \text{and} \quad f_2 : N \to P.$$

For any $x \in M$,

$$d(f_2 \circ f_1)(x) = df_2(f_1(x)) \circ df_1(x).$$

Proof. Let $v \in T_x M$. Show that

$$d(f_2 \circ f_1)(x)(v) = df_2(f_1(x)) \circ df_1(x)(v).$$

Let I_v be an open interval in $\mathbb{R}$ containing 0, and let $c_v : I_v \to \mathbb{R}^{n_M}$ be a C^1 function such that $c_v(I_v) \subset M$, $c_v(0) = x$, and $c_v'(0) = v$. The definition of the differential gives

$$d(f_2 \circ f_1)(x)(v) = (f_2 \circ f_1 \circ c_v)'(0).$$

Let $w = (f_1 \circ c_v)'(0) = df_1(x)(v) \in \mathbb{R}^{n_N}$. The function $f_1 \circ c_v : I_v \to \mathbb{R}^{n_N}$ is C^1 and $f_1 \circ c_v(I_v) \subset N$. It satisfies $f_1 \circ c_v(0) = f_1(x)$ and, by definition of w, $(f_1 \circ c_v)'(0) = w$. The definition of the differential for f_2 then gives

$$df_2(f_1(x))(w) = (f_2 \circ f_1 \circ c_v)'(0).$$

Thus,

$$\begin{aligned} d(f_2 \circ f_1)(x)(v) &= df_2(f_1(x))(w) \\ &= df_2(f_1(x))(df_1(x)(v)) \\ &= \left[df_2(f_1(x)) \circ df_1(x)\right](v). \end{aligned}$$

□

Beyond the rules for differentiability and the usual operations, many notions and results of classical differential calculus naturally extend to differential calculus on manifolds. Below, we give two examples: the concept of *diffeomorphism* and the local inversion theorem.

Definition 2.33 : diffeomorphism between manifolds

Let M, N be two C^k submanifolds of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively. Consider a map

$$\phi : M \to N.$$

For any $r \in \{1, \dots, k\}$, we say that ϕ is a C^r-diffeomorphism between M and N if it satisfies the following three properties:

1. ϕ is a bijection from M to N;
2. ϕ is of class C^r on M;
3. ϕ^{-1} is of class C^r on N.

Theorem 2.34 : local inversion on submanifolds

Let M, N be two C^k submanifolds of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively. Let $x_0 \in M$. For $r \in \{1, \dots, k\}$, consider a C^r map,

$$f : M \to N.$$

If $df(x_0) : T_{x_0} M \to T_{f(x_0)} N$ is bijective, then there exist U_{x_0} an open neighborhood of x_0 in M and $V_{f(x_0)}$ an open neighborhood of $f(x_0)$ in N such that f is a C^r-diffeomorphism from U_{x_0} to $V_{f(x_0)}$.

Proof. Let d be the dimension of M. Note that N has the same dimension d: $df(x_0)$ is a bijective linear map between $T_{x_0}M$ and $T_{f(x_0)}N$, so

$$\dim T_{f(x_0)}N = \dim T_{x_0}M = d.$$

Let $U_M, V_M \subset \mathbb{R}^{n_1}$ be open neighborhoods of x_0 and 0, respectively, and $\phi_M : U_M \to V_M$ a C^k-diffeomorphism such that

$$\phi_M(M \cap U_M) = (\mathbb{R}^d \times \{0\}^{n_1-d}) \cap V_M.$$

Without loss of generality, assume that $\phi_M(x_0) = 0$.

Similarly, let $U_N, V_N \subset \mathbb{R}^{n_2}$ be open neighborhoods of $f(x_0)$ and 0, and $\phi_N : U_N \to V_N$ a C^k-diffeomorphism such that

$$\phi_N(N \cap U_N) = (\mathbb{R}^d \times \{0\}^{n_2-d}) \cap V_N.$$

Assume that $\phi_N(f(x_0)) = 0$.

The idea of the proof is to go back to the case where f is defined on an open subset of $\mathbb{R}^d$ and then apply the classical local inversion theorem. To do this, we "transfer" f to a map from $\mathbb{R}^d \times \{0\}^{n_1-d}$ to $\mathbb{R}^d \times \{0\}^{n_2-d}$ by composing it with the diffeomorphisms ϕ_M and ϕ_N.

More precisely, let $\phi_{M,0}^{-1}$ be the restriction of ϕ_M^{-1} to $(\mathbb{R}^d \times \{0\}^{n_1-d}) \cap V_M$. Define

$$g \overset{def}{=} \phi_N \circ f \circ \phi_{M,0}^{-1} : (\mathbb{R}^d \times \{0\}^{n_1-d}) \cap V_M \to (\mathbb{R}^d \times \{0\}^{n_2-d}) \cap V_N.$$

This definition is valid (if we reduce U_M, V_M a bit, so that $f(U_M) \subset U_N$). The function g is of class C^r and its differential at 0 is injective: it is the composition of $d\phi_N(f(x_0))$, $df(x_0)$, and $d\phi_{M,0}^{-1}(0)$, all of which are injective. Since it goes from $\mathbb{R}^d$ to $\mathbb{R}^d$, it is bijective[3].

According to the classical local inversion theorem (Theorem 1.10), there exist E_M, E_N open neighborhoods of 0 in $\mathbb{R}^d$ such that g is a C^r-diffeomorphism from $E_M \times \{0\}^{n_1-d}$ to $E_N \times \{0\}^{n_2-d}$. Then f is a C^r-diffeomorphism from $U_{x_0} \overset{def}{=} \phi_M^{-1}(E_M \times \{0\}^{n_1-d})$ to $V_{f(x_0)} \overset{def}{=} \phi_N^{-1}(E_N \times \{0\}^{n_2-d})$: on these sets,

$$f = \phi_N^{-1} \circ g \circ \phi_M.$$

Since ϕ_M is a diffeomorphism (of class C^k hence also of class C^r) from U_{x_0} to $E_M \times \{0\}^{n_1-d}$, g is a C^r-diffeomorphism from $E_M \times \{0\}^{n_1-d}$ to $E_N \times \{0\}^{n_2-d}$, and ϕ_N^{-1} is a diffeomorphism (C^k hence also C^r) from $E_N \times \{0\}^{n_2-d}$ to $V_{f(x_0)}$, the map f is a composition of C^r-diffeomorphisms, hence a C^r-diffeomorphism. □

[3]We can see $\phi_N \circ f \circ \phi_{M,0}^{-1}$ as a map between two open subsets of $\mathbb{R}^d$.

Chapter 3

Riemannian geometry

What you should know or be able to do after this chapter

- Know the definition of curves and parametrized curves.
- Given a curve, introduce a convenient parametrization of it,
  - either a local one as in Proposition 3.4,
  - or a global one, as in Corollary 3.7.
- Know that a connected curve is diffeomorphic to either $\mathbb{S}^1$ or $\mathbb{R}$.
- Be able to manipulate the length of a curve (e.g. compute it, when possible, or upper bound it otherwise).
- In general dimension, propose a definition of distance intrinsic to a manifold, and remember the "standard" one.
- Understand (i.e. be able to reexplain) the intuition of why minimizing paths satisfy the geodesic equation.
- Know the explicit description of geodesics on the sphere.
- Know the relation between minimizing paths and geodesics (a minimizing path is a geodesic, and a geodesic is locally a minimizing path).

Let $k, n \in \mathbb{N}^*$ be fixed.

In the previous chapter, we introduced the concept of differentiability for maps between submanifolds. This concept allows us to study the *topological* properties of submanifolds: one may wonder which submanifolds are diffeomorphic to each other and what properties characterize whether or not they are diffeomorphic. Informally speaking, one can ask questions like: "Is a donut diffeomorphic to a balloon?"[1]

In this chapter, we delve into finer properties of submanifolds, specifically *metric* properties involving notions of length, angle, etc. We will introduce a notion of isometry, which is more restrictive than that of diffeomorphism (in the sense that two isometric manifolds are necessarily diffeomorphic, whereas the converse is not true).

As the formal definitions of these properties are subtle, and since the objective here is only to provide an overview rather than a complete description, we will mainly focus on the simplest case, one-dimensional submanifolds. Submanifolds of general dimension will be discussed only towards the end of the chapter.

3.1 Submanifolds of dimension 1

Definition 3.1 : curve

A *curve* is a submanifold of $\mathbb{R}^n$ of dimension 1.

[1] Answer: no.

Figure 3.1: The image of the parametrized curve $\gamma : t \in \mathbb{R} \to (t(t+1)^2, t^2(t+1))$ (left figure) is not a submanifold of $\mathbb{R}^2$ because $(0,0)$ is a multiple point. However, $\gamma(]-\epsilon;\epsilon[)$ is a submanifold of $\mathbb{R}^2$ for any sufficiently small ϵ (right figure).

3.1.1 Parametrized curves

Curves, in comparison to higher-dimensional manifolds, have the particularity that they admit a simple parametrization. In essence, they can be seen as the image of an open set of $\mathbb{R}$ through a C^1 function. This parametrization allows for a convenient definition of metric quantities, as we will see later in this section.

Definition 3.2 : parametrized curve

A *parametrized curve* of class C^k is a pair (I, γ), where I is an interval in $\mathbb{R}$ and $\gamma : I \to \mathbb{R}^n$ is a C^k function.

The image of a parametrized curve is not necessarily a submanifold of $\mathbb{R}^n$, especially because the curve can intersect itself (we say that it has a *multiple point*). However, the following proposition shows that the image of a parametrized curve (I, γ) locally defines a submanifold, in the vicinity of points where γ' does not vanish. This result is illustrated in Figure 3.1.

Proposition 3.3

Let (I, γ) be a parametrized curve. For $t \in \mathring{I}$ and $x = \gamma(t)$, we say that x is a *regular point* if $\gamma'(t) \neq 0$. In this case, there exists $\epsilon > 0$ such that $]t-\epsilon; t+\epsilon[\subset I$, and the set

$$C \overset{def}{=} \gamma(]t-\epsilon; t+\epsilon[)$$

is a curve. Moreover,

$$T_x C = \mathbb{R}\gamma'(t).$$

Proof. Suppose x is regular, i.e., γ is an immersion at t. If we can show that, for $\epsilon > 0$ sufficiently small, γ induces a homeomorphism from $]t-\epsilon; t+\epsilon[$ to its image, the theorem is proved. Indeed, we can then choose $\epsilon > 0$ small enough so that γ' does not vanish (i.e., γ is immersive) over the entire interval $]t-\epsilon; t+\epsilon[$. Proposition 2.10 then ensures that

$$C \overset{def}{=} \gamma(]t-\epsilon; t+\epsilon[)$$

is a submanifold of $\mathbb{R}^n$ of dimension 1, i.e., a curve, and Property 2 of Theorem 2.16 tells us that

$$T_x C = \mathrm{Im}(d\gamma(t)) = \mathbb{R}\gamma'(t).$$

To show that γ induces a homeomorphism from $]t-\epsilon; t+\epsilon[$ to its image if $\epsilon > 0$ is sufficiently small, we use the normal form theorem for immersions (Theorem 1.14). Let ψ be a diffeomorphism from a neighborhood of x to a neighborhood of $0_{\mathbb{R}^n}$ and $\epsilon > 0$ be such that

$$\forall t' \in]t-\epsilon; t+\epsilon[, \quad \psi \circ \gamma(t') = (t', 0, \ldots, 0).$$

Defining $\pi_1 : \mathbb{R}^n \to \mathbb{R}$ as the projection onto the first coordinate, we have

$$\forall t' \in]t-\epsilon; t+\epsilon[, \quad \pi_1 \circ \psi \circ \gamma(t') = t'.$$

Consequently, γ is injective on $]t-\epsilon; t+\epsilon[$. It is therefore a bijection from $]t-\epsilon; t+\epsilon[$ to its image. From the previous equation, its reciprocal is $\pi_1 \circ \psi$ (more precisely, the restriction of $\pi_1 \circ \psi$ to $\gamma(]t-\epsilon; t+\epsilon[)$), which is continuous, so γ is a homeomorphism between $]t-\epsilon; t+\epsilon[$ and $\gamma(]t-\epsilon; t+\epsilon[)$. □

Conversely, any curve is locally the image of a parametrized curve.

Proposition 3.4

Let $C \subset \mathbb{R}^n$ be a C^k curve. For any $x \in C$, there exists a neighborhood V of x in $\mathbb{R}^n$ and a parametrized curve (I, γ) of class C^k such that
$$C \cap V = \gamma(I).$$

Proof. Let x be in C. From the "immersion" definition of submanifolds, there exists a neighborhood V of x, an open set $U \subset \mathbb{R}$ and a C^k map $f : U \to \mathbb{R}^n$, which is a homeomorphism onto its image, such that
$$C \cap V = f(U). \tag{3.1}$$
Let $t_0 \in U$ be the preimage of x by f (that is, $f(t_0) = x$). The set U may not be an interval but, if we replace V with a smaller set, we can replace U with the connected component of t_0[2], while keeping Equality (3.1) true. We can then set $I = U$ and $\gamma = f$. □

Actually, any connected curve[3] is the image of a parametrized curve (globally, not locally as in the previous proposition). This is a consequence of the following theorems.

Theorem 3.5 : compact curves

Let $M \subset \mathbb{R}^n$ be a compact and connected curve of class C^k. It is C^k-diffeomorphic to the circle $\mathbb{S}^1$.

Theorem 3.6 : non-compact curves

Let $M \subset \mathbb{R}^n$ be a connected non-compact curve of class C^k. It is C^k-diffeomorphic to $\mathbb{R}$.

The proof of these theorems is difficult. We will limit ourselves to the proof of the first one, which will be given in subsection 3.1.2. The proof of the second one uses partly the same strategy but requires additional ideas.

Corollary 3.7 : global parametrization of connected curves

Let $M \subset \mathbb{R}^n$ be a connected curve of class C^k.

- If M is non-compact, there exists a parametrized curve (I, γ) of class C^k such that
  - I is an open interval;
  - $\gamma(I) = M$;
  - γ is a difféomorphism between I and M.
- If M is compact, then, for any $a, b \in \mathbb{R}$ such that $a < b$, there exists a parametrized curve $([a; b[, \gamma)$ of class C^k such that
  - $\gamma([a; b[) = M$;
  - γ is a bijection between $[a; b[$ and M;
  - $\lim_b \gamma^{(r)} = \gamma^{(r)}(a)$ for any $r \in \{0, \dots, k\}$.

In both cases, we call such parametrized curve a *global parametrization* of M.

[2] To give more detail: let U' be the connected component of t_0 in U. Then $f(U')$ is an open set of $f(U)$, because U' is open and f is a homeomorphism between U and $f(U)$. Therefore, there exists $V' \subset \mathbb{R}^n$ an open set such that $f(U') = f(U) \cap V'$. Then, $f(U') = f(U) \cap V' = C \cap (V \cap V')$. We can therefore replace U by U', V by $V \cap V'$, and Equality (3.1) is stil true.

[3] Some reminders on connectedness can be found in Appendix A.

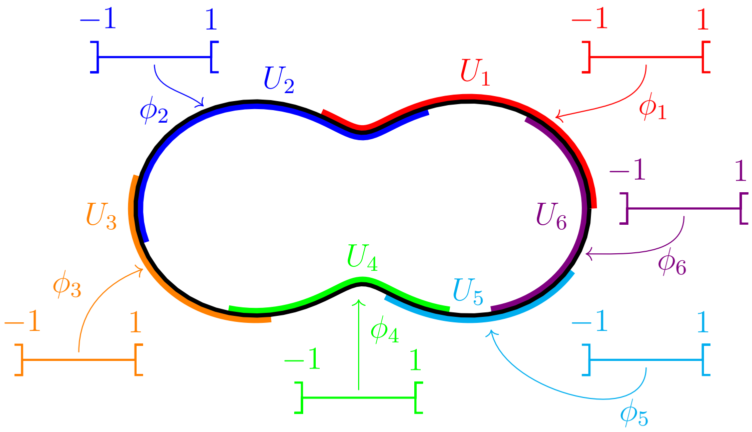


Figure 3.2: Illustration of Lemma 3.8: the curve M (the black line) and its covering by the open sets U_s.

Proof. First, if M is non-compact, from Theorem 3.6, there exists $\phi : \mathbb{R} \to M$ a C^k-diffeomorphism. We can set $I = \mathbb{R}$ and $\gamma = \phi$.

Let us now assume that M is compact. Let $\phi : \mathbb{S}^1 \to M$ be a C^k-diffeomorphism as in Theorem 3.5. We define

$$\begin{array}{rccc} \sigma : & [a;b[& \to & \mathbb{S}^1 \\ & t & \to & \left(\cos\left(2\pi\frac{t-a}{b-a}\right), \sin\left(2\pi\frac{t-a}{b-a}\right)\right). \end{array}$$

and set $\gamma = \phi \circ \sigma : [a;b[\to M$. It defines a parametrized curve of class C^k. Since σ is a bijection between $[a;b[$ and $\mathbb{S}^1$, and ϕ a bijection between $\mathbb{S}^1$ and M, γ is a bijection between $[a;b[$ and M. In addition, as σ (hence also γ) is the restriction to $[a;b[$ of a $(b-a)$-periodic C^k function, it holds, for all $r \in \{0,\dots,k\}$,

$$\gamma^{(r)}(t) \xrightarrow{t\to b} \gamma^{(r)}(a).$$

$\square$

3.1.2 Proof of Theorem 3.5

The proof is intricate. Students are not expected to read it, but can do so if they are curious. In this case, they are encouraged to focus on the following two things first:

- understand the statements of Lemmas 3.8 to 3.11, and why these lemmas imply the theorem (roughly this page and the next two);
- in a second time, read the proof of Lemma 3.10, focusing on understanding the definitions of the various objects and Figure 3.3 rather than the precise technical details.

The proof relies on several intermediate lemmas, the proofs of which will be given later.

The first lemma, whose proof is based solely on the definition of submanifolds and the compactness of M, asserts that M can be covered by a finite number of open sets diffeomorphic to $]-1;1[$.

Lemma 3.8

There exists a finite number of open sets in M, denoted $U_1,\dots,U_S$, such that

1. $M = U_1 \cup \dots \cup U_S$;
2. for every $s \le S$, U_s is C^k-diffeomorphic to $]-1;1[$.

The principle of the proof is to consider a finite covering as in the previous lemma and to construct, step by step, a progressively smaller covering by gradually merging the open sets of the covering. Let $(U_1,\dots,U_S)$ be a covering as in Lemma 3.8. For every s, let

$$\phi_s :]-1;1[\to U_s$$

be a C^k-diffeomorphism.

We will now judiciously choose two open sets U_{s_1}, U_{s_2} and merge them to obtain, according to the properties of $U_{s_1} \cap U_{s_2}$,

- either directly that M is C^k-diffeomorphic to $\mathbb{S}^1$;
- or that there exists a covering as in Lemma 3.8, with size $S-1$ instead of S.

In the first case, the proof is complete. In the second case, the procedure will be iteratively reapplied to obtain a covering with a decreasing number of elements.

The following lemma indicates what $U_{s_1} \cap U_{s_2}$ might look like.

Lemma 3.9

For all $s_1, s_2 \leq S$ distinct, the intersection $U_{s_1} \cap U_{s_2}$ satisfies one of the following properties:

1. $U_{s_1} \cap U_{s_2}$ is empty.
2. $U_{s_1} \cap U_{s_2}$ has a single connected component. In this case, we are in one of the following situations:
   (a) $U_{s_1} \subset U_{s_2}$ or $U_{s_2} \subset U_{s_1}$;
   (b) $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})$ and $\phi_{s_2}^{-1}(U_{s_1} \cap U_{s_2})$ are intervals of the form $]-1;\alpha[$ or $]\alpha;1[$, with $\alpha \in]-1;1[$.
3. $U_{s_1} \cap U_{s_2}$ has two connected components. In this case, $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})$ and $\phi_{s_2}^{-1}(U_{s_1} \cap U_{s_2})$ are of the form $]-1;\alpha[\cup]\beta;1[$, with $\alpha, \beta \in]-1;1[, \alpha < \beta$.

We can show that there exist $s_1, s_2 \in \{1, \dots, S\}$ distinct such that $U_{s_1} \cap U_{s_2} \neq \emptyset$. Indeed, let's proceed by contradiction and suppose there are no $s_1 \neq s_2$ such that $U_{s_1} \cap U_{s_2} \neq \emptyset$. Then we are in one of the following situations:

1. $S = 1$;
2. $S > 1$ and $U_{s_1} \cap U_{s_2} = \emptyset$ for all $s_1 \neq s_2$.

In the first case, we must have $M = U_1$. Since U_1 is C^k-diffeomorphic to $]-1;1[$, M is also. This is impossible: a compact set (here, M) cannot be diffeomorphic to a non-compact set (here, $]-1;1[$). In the second case,

$$U_1 \text{ and } U_2 \cup \dots \cup U_S$$

are non-empty, disjoint open sets whose union is M. So M is not connected: again, this leads to an impossibility. Therefore, we can choose $s_1, s_2 \in \{1, \dots, S\}$ distinct such that $U_{s_1} \cap U_{s_2} \neq \emptyset$.

Since the intersection $U_{s_1} \cap U_{s_2}$ is non-empty, we are in situation 2 or 3 of Lemma 3.9. If we are in situation 3, the following lemma directly concludes the proof of the theorem.

Lemma 3.10 : two connected components

If U_{s_1}, U_{s_2} satisfy Property 3 of Lemma 3.9, then M is C^k-diffeomorphic to $\mathbb{S}^1$.

If, on the contrary, we are in Situation 2, another lemma must be used.

Lemma 3.11 : one connected component

If U_{s_1}, U_{s_2} satisfy Property 2 of Lemma 3.9, then $U_{s_1} \cup U_{s_2}$ is C^k-diffeomorphic to $]-1;1[$.

In this case, we obtain that $\{U_s, s \neq s_1, s_2\} \cup \{U_{s_1} \cup U_{s_2}\}$ is a collection of open sets, C^k-diffeomorphic to $]-1;1[$, whose union is the entire M. Thus, we have found a set $\tilde{U}_1, \dots, \tilde{U}_{S-1}$ of open sets satisfying the properties of Lemma 3.8 but with cardinality strictly less than S.

We can then reapply the same reasoning: there exist $\tilde{s}_1 \neq \tilde{s}_2$ such that $\tilde{U}_{\tilde{s}_1} \cap \tilde{U}_{\tilde{s}_2} \neq \emptyset$. If the intersection has two connected components, then M is C^k-diffeomorphic to $\mathbb{S}^1$, which concludes the proof. If it has only one connected component, then we can find a set of $S-2$ open sets satisfying the properties of Lemma 3.8. And so on.

The reasoning cannot be applied more than S times (otherwise, we would find a covering of M by a negative number of open sets). Therefore, there must come a time when the intersection has two connected components, which implies that M is C^k-diffeomorphic to $\mathbb{S}^1$ and concludes.

Proof of Lemma 3.8. First, consider any $x \in M$. Let V be an open neighborhood of x in $\mathbb{R}^n$, I an open neighborhood of 0 in $\mathbb{R}$, and $f : I \to V$ a C^k map which is a homeomorphism onto its image, such that

$$f(I) = V \cap M$$

and f is immersive at $z_0 = f^{-1}(x)$. (This is the "immersion" definition of a submanifold of dimension 1 - Property 2 of Definition 2.1.)

By reducing I and V slightly, we can assume that I is a bounded open interval and that f is immersive over the entire I. We set

$$U(x) = f(I) = V \cap M.$$

It is an open subset of M. Moreover, it is C^k-diffeomorphic to I (indeed, it is homeomorphic to I by hypothesis on f; for any x', $df(x')$ is injective, hence bijective, from $T_{x'}I$ to $T_{f(x')}M$; according to the local inversion theorem 2.34, f is then a local C^k-diffeomorphism, implying that f^{-1} is C^k). Since any non-empty open interval in $\mathbb{R}$ is C^k-diffeomorphic to $]-1;1[$, $U(x)$ is C^k-diffeomorphic to $]-1;1[$.

Now we no longer consider a fixed x.

For any $x \in M$, $x \in U(x) \subset \cup_{x' \in M} U(x')$. Thus,

$$M \subset \bigcup_{x' \in M} U(x'),$$

meaning that the $U(x')$, for all $x' \in M$, form a covering of M by open sets. Since M is compact, we can extract a finite sub-covering: there exist $x_1, \dots, x_S$ such that

$$M = U(x_1) \cup \dots \cup U(x_S).$$

As we have seen that $U(x_s)$ is diffeomorphic to $]-1;1[$ for every s, the result is proved. ☐

Proof of Lemma 3.9. The set $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})$ is an open subset of $]-1;1[$. Therefore, it can be expressed as a union of disjoint open intervals in $]-1;1[$ (see Example A.5):

$$\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) = \bigcup_{l \in E}]a_l; b_l[,$$

where E is an index set (which can be finite or infinite).

Let's start by assuming that there exists $k \in E$ such that $-1 < a_k < b_k < 1$. We will show that in this case, $U_{s_2} \subset U_{s_1}$.

The function $\phi_{s_2}^{-1} \circ \phi_{s_1} : \phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) \to]-1;1[$ is continuous and injective (being the composition of two continuous and injective functions). Hence, it is monotonic on each interval contained in $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})$. Let's assume, for example, that it is increasing on $]a_k; b_k[$ (a similar reasoning can be applied if it is decreasing).

Set

$$B_k = \lim_{t \to b_k^-} \phi_{s_2}^{-1} \circ \phi_{s_1}(t).$$

(Note that the limit exists: $\phi_{s_2}^{-1} \circ \phi_{s_1}$ is an increasing and bounded function, as its values are between -1 and 1; therefore, it converges in b_k^- to a value in $]-1;1]$.)

It is impossible that $B_k < 1$. Indeed, if $B_k < 1$, then $\phi_{s_2}(B_k)$ is well-defined and, by the continuity of ϕ_{s_2},

$$\begin{aligned}\phi_{s_2}(B_k) &= \phi_{s_2}\big(\lim_{t \to b_k^-} \phi_{s_2}^{-1} \circ \phi_{s_1}(t)\big) \\ &= \lim_{t \to b_k^-} \phi_{s_1}(t)\end{aligned}$$

$$= \phi_{s_1}(b_k).$$

Thus, $\phi_{s_1}(b_k) \in \phi_{s_1}(]-1;1[) \cap \phi_{s_2}(]-1;1[) = U_{s_1} \cap U_{s_2}$, implying

$$b_k \in \phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) = \bigcup_{l \in E}]a_l; b_l[.$$

Therefore, $b_k \in]a_l; b_l[$ for some $l \in E$ such that $l \neq k$, and for this l, we must have $]a_k; b_k[\cap]a_l; b_l[\neq \emptyset$, contradicting the fact that the intervals $]a_l; b_l[$ are disjoint. Thus, $B_k = 1$.

Similarly, we define

$$A_k = \lim_{t \to a_k^+} \phi_{s_2}^{-1} \circ \phi_{s_1}(t)$$

and the same reasoning shows that $A_k = -1$.

The image of $]a_k; b_k[$ under $\phi_{s_2}^{-1} \circ \phi_{s_1}$ is an interval (it is the image of an interval under a continuous function); it is included in $]-1;1[$, and we have just seen that

$$\phi_{s_2}^{-1} \circ \phi_{s_1}(t) \overset{t \to b_k^-}{\longrightarrow} 1 \quad \text{and} \quad \phi_{s_2}^{-1} \circ \phi_{s_1}(t) \overset{t \to a_k^+}{\longrightarrow} -1.$$

Thus,

$$\begin{aligned} & \phi_{s_2}^{-1} \circ \phi_{s_1}(]a_k; b_k[) =]-1;1[\\ \Rightarrow \quad & U_{s_2} = \phi_{s_2}(]-1;1[) = \phi_{s_2}(\phi_{s_2}^{-1} \circ \phi_{s_1}(]a_k; b_k[)) = \phi_{s_1}(]a_k; b_k[) \subset U_{s_1}. \end{aligned}$$

Thus, we have shown that if there exists $k \in E$ such that $-1 < a_k < b_k < 1$, then $U_{s_2} \subset U_{s_1}$, placing us in Case 2a of the lemma's statement. Now, suppose that there is no $k \in E$ such that $-1 < a_k < b_k < 1$. This means that for every $l \in E$, $a_l = -1$ or $b_l = 1$ (or both). Considering the fact that the intervals $]a_l; b_l[$ are disjoint, we have five possibilities:

(i) $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) = \emptyset$;

(ii) $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) =]-1;1[$;

(iii) $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) =]-1;\alpha[$ for some $\alpha \in]-1;1[$;

(iv) $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) =]\alpha;1[$ for some $\alpha \in]-1;1[$;

(v) $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) =]-1;\alpha[\cup]\beta;1[$, with $\alpha, \beta \in]-1;1[, \alpha < \beta$.

In Case (i), we must have $U_{s_1} \cap U_{s_2} = \emptyset$ (since ϕ_{s_1} is surjective onto U_{s_1}); thus, we are in Case 1 of the lemma's statement.

In Case (ii), we have

$$U_{s_1} = \phi_{s_1}(]-1;1[) = \phi_{s_1}(\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})) = U_{s_1} \cap U_{s_2},$$

so $U_{s_1} \subset U_{s_2}$; we are in Case 2a of the lemma's statement.

In Case (iii) or (iv), $U_{s_1} \cap U_{s_2}$ has exactly one connected component (see Proposition A.7); in Case (v), $U_{s_1} \cap U_{s_2}$ has two connected components. Therefore, we are in Case 2b or 3 of the lemma's statement, respectively. (Note that the reasoning we have done for $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})$ is also valid for $\phi_{s_2}^{-1}(U_{s_1} \cap U_{s_2})$: this set is also of the form $]-1;\alpha[$ or $]\alpha;1[$ if $U_{s_1} \cap U_{s_2}$ has a single connected component and $U_{s_1} \not\subset U_{s_2}, U_{s_2} \not\subset U_{s_1}$, and of the form $]-1;\alpha[\cup]\alpha;\beta[$ if $U_{s_1} \cap U_{s_2}$ has two connected components.) $\square$

Proof of Lemma 3.10.

First step: Let's begin by assuming that $U_{s_1} \cup U_{s_2}$ is C^k-diffeomorphic to $\mathbb{S}^1$. Then $U_{s_1} \cup U_{s_2}$ is an open and closed subset of M (open because it's a union of open sets, closed because it's homeomorphic to a compact set, hence compact). As M is connected and $U_{s_1} \cup U_{s_2}$ is non-empty, we must have (according to Proposition A.2)

$$M = U_{s_1} \cup U_{s_2}.$$

Thus, M is C^k-diffeomorphic to $\mathbb{S}^1$.

Second step: Let's show that $U_{s_1} \cup U_{s_2}$ is C^k-diffeomorphic to $\mathbb{S}^1$.

Let C_1, C_2 be the two connected components of $U_{s_1} \cap U_{s_2}$. Since we are in Case 3 of Lemma 3.9, there exist α_1, β_1 such that

$$\begin{aligned} & \phi_{s_1}^{-1}(C_1) =]-1;\alpha_1[\quad \text{and} \quad \phi_{s_1}^{-1}(C_2) =]\beta_1;1[\\ \text{or} \quad & \phi_{s_1}^{-1}(C_1) =]\beta_1;1[\quad \text{and} \quad \phi_{s_1}^{-1}(C_2) =]-1;\alpha_1[. \end{aligned} \tag{3.2}$$

By exchanging C_1 and C_2 if necessary, we can assume that Equation (3.2) is true. Similarly, there exist α_2, β_2 such that

$$\begin{aligned} & \phi_{s_2}^{-1}(C_1) =]-1;\alpha_2[\quad \text{and} \quad \phi_{s_2}^{-1}(C_2) =]\beta_2;1[\\ \text{or} \quad & \phi_{s_2}^{-1}(C_1) =]\beta_2;1[\quad \text{and} \quad \phi_{s_2}^{-1}(C_2) =]-1;\alpha_2[. \end{aligned} \tag{3.3}$$

By replacing ϕ_{s_2} with $\tilde{\phi}_{s_2} : t \in]-1;1[\to \phi_{s_2}(-t)$ (which is also a C^k-diffeomorphism from $]-1;1[$ to U_{s_2}), we can assume that Equation (3.3) is true.

Proposition 3.12

The map $\phi_{s_2}^{-1} \circ \phi_{s_1}$ is a decreasing C^k-diffeomorphism from $]-1;\alpha_1[$ to $]-1;\alpha_2[$ and from $]\beta_1;1[$ to $]\beta_2;1[$.

Proof. Let's prove it for the intervals $]-1;\alpha_1[$ and $]-1;\alpha_2[$; the proof is identical for $]\beta_1;1[$ and $]\beta_2;1[$.

Since ϕ_{s_1} is a C^k-diffeomorphism from $]-1;\alpha_1[$ to C_1, and $\phi_{s_2}^{-1}$ is a C^k-diffeomorphism from C_1 to $]-1;\alpha_2[$, the map $\phi_{s_2}^{-1} \circ \phi_{s_1}$ is a C^k-diffeomorphism from $]-1;\alpha_1[$ to $]-1;\alpha_2[$. Let's show that it is decreasing.

As a diffeomorphism between two intervals is always strictly monotonic, it suffices to show that it is not increasing. Suppose, by contradiction, that it is increasing. Then

$$\phi_{s_2}^{-1} \circ \phi_{s_1}(t) \xrightarrow{t \to \alpha_1} \alpha_2,$$

which implies

$$\phi_{s_2}(\alpha_2) = \phi_{s_2}\left(\lim_{t \to \alpha_1} \phi_{s_2}^{-1} \circ \phi_{s_1}(t)\right) = \lim_{t \to \alpha_1} \phi_{s_1}(t) = \phi_{s_1}(\alpha_1)$$

and therefore

$$\phi_{s_1}(\alpha_1) \in U_{s_1} \cap U_{s_2},$$

which contradicts the fact that $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2}) =]-1;\alpha_1[\cup]\beta_1;1[$ and does not contain α_1. Therefore, it is impossible for $\phi_{s_2}^{-1} \circ \phi_{s_1}$ to be increasing. $\square$

Fix four real numbers c_1, c_2, c_3, c_4 such that $-1 < c_1 < c_2 < \alpha_1$ and $\beta_1 < c_3 < c_4 < 1$ (see Figure 3.3 for an illustration of the notations). For all $k = 1, 2, 3, 4$, denote

$$P_k = \phi_{s_1}(c_k) \quad \text{and} \quad d_k = \phi_{s_2}^{-1}(P_k) = \phi_{s_2}^{-1}(\phi_{s_1}(c_k))$$

Since c_1, c_2 belong to $]-1;\alpha_1[$ and $c_1 < c_2$, Proposition 3.12 implies that d_1, d_2 belong to $]-1;\alpha_2[$ and $d_2 < d_1$. Similarly, d_3, d_4 belong to $]\beta_2;1[$ and $d_4 < d_3$. Note (this will be useful later) that, again due to Proposition 3.12:

$$\begin{gathered} \phi_{s_1}(]-1;c_1]) = \phi_{s_2}(\phi_{s_2}^{-1} \circ \phi_{s_1}(]-1;c_1])) = \phi_{s_2}([d_1;\alpha_2[), \\ \phi_{s_1}([c_1;c_2]) = \phi_{s_2}([d_2;d_1]), \\ \phi_{s_1}([c_2;\alpha_1[) = \phi_{s_2}(]-1;d_2]), \\ \phi_{s_1}(]\beta_1;c_3]) = \phi_{s_2}([d_3;1[), \\ \phi_{s_1}([c_3;c_4]) = \phi_{s_2}([d_4;d_3]), \\ \phi_{s_1}([c_4;1[) = \phi_{s_2}(]\beta_2;d_4]). \end{gathered} \tag{3.4}$$

Now, let's construct a C^k-diffeomorphism $\psi : \mathbb{S}^1 \to M$. We will impose, as shown in Figure 3.3,

$$\psi\left(e^{i\frac{\pi}{4}}\right) = P_3, \quad \psi\left(e^{i\frac{3\pi}{4}}\right) = P_2, \quad \psi\left(e^{i\frac{5\pi}{4}}\right) = P_1, \quad \psi\left(e^{i\frac{7\pi}{4}}\right) = P_4. \tag{3.5}$$

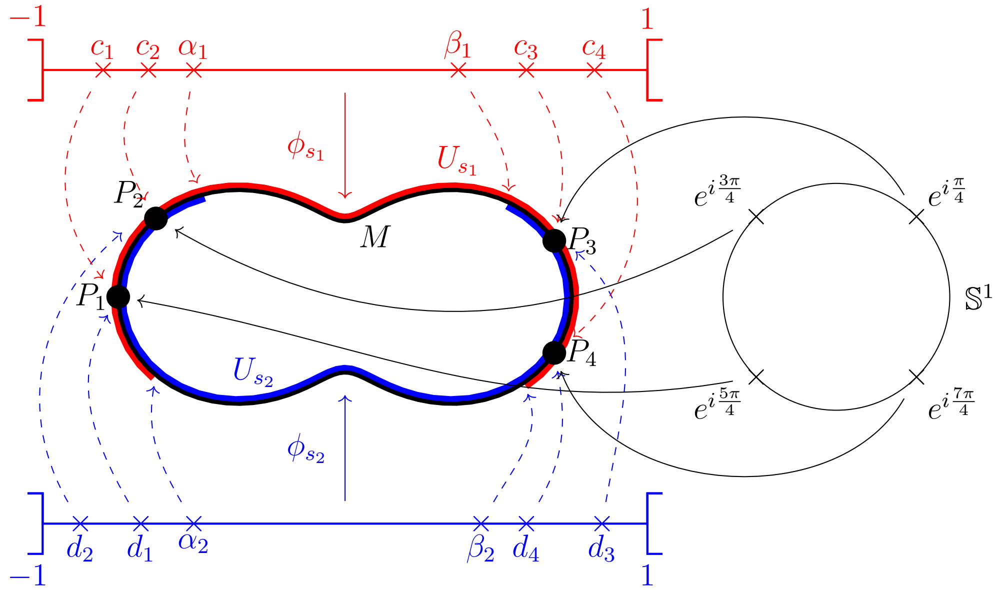


Figure 3.3: Illustration of the notation in Lemma 3.10 and schematic representation of the diffeomorphism from $\mathbb{S}^1$ to M ($e^{i\frac{\pi}{4}}$ is mapped to P_3, etc.).

We will define ψ piecewise as follows:

$$\psi(e^{i\theta}) = \phi_{s_1} \circ \delta_{s_1}(\theta) \text{ for all } \theta \in \left[-\frac{\pi}{4}; \frac{5\pi}{4}\right]; \tag{3.6a}$$

$$\psi(e^{i\theta}) = \phi_{s_2} \circ \delta_{s_2}(\theta) \text{ for all } \theta \in \left[\frac{3\pi}{4}; \frac{9\pi}{4}\right], \tag{3.6b}$$

with $\delta_{s_1} : \left[-\frac{\pi}{4}; \frac{5\pi}{4}\right] \to]-1;1[$ and $\delta_{s_2} : \left[\frac{3\pi}{4}; \frac{9\pi}{4}\right] \to]-1;1[$ appropriately chosen functions.

We start by choosing δ_{s_1}. Let δ_{s_1} be a C^∞-diffeomorphism from $\left[-\frac{\pi}{4}; \frac{5\pi}{4}\right]$ to $[c_1; c_4]$ such that

$$\delta_{s_1}\left(-\frac{\pi}{4}\right) = c_4, \quad \delta_{s_1}\left(\frac{\pi}{4}\right) = c_3, \quad \delta_{s_1}\left(\frac{3\pi}{4}\right) = c_2, \quad \delta_{s_1}\left(\frac{5\pi}{4}\right) = c_1. \tag{3.7}$$

(Such a diffeomorphism exists, see Proposition B.3 in the appendix).

Now, let's define δ_{s_2}. The definitions in Equations (3.6a) and (3.6b) must coincide at the points where they both give a value to ψ. Thus, for all $\theta \in \left[\frac{3\pi}{4}; \frac{5\pi}{4}\right]$,

$$\phi_{s_1}(\delta_{s_1}(\theta)) = \phi_{s_2}(\delta_{s_2}(\theta))$$

and, for all $\theta \in \left[\frac{7\pi}{4}; \frac{9\pi}{4}\right]$,

$$\phi_{s_1}(\delta_{s_1}(\theta - 2\pi)) = \phi_{s_2}(\delta_{s_2}(\theta)).$$

Define

$$\delta_{s_2}(\theta) = \phi_{s_2}^{-1}(\phi_{s_1}(\delta_{s_1}(\theta))) \text{ for all } \theta \in \left[\frac{3\pi}{4}; \frac{5\pi}{4}\right], \tag{3.8a}$$

$$\delta_{s_2}(\theta) = \phi_{s_2}^{-1}(\phi_{s_1}(\delta_{s_1}(\theta - 2\pi))) \text{ for all } \theta \in \left[\frac{7\pi}{4}; \frac{9\pi}{4}\right]. \tag{3.8b}$$

It can be verified that the quantities above are well-defined, thanks to the equalities in Equation (3.7), which imply that $\delta_{s_1}(\theta)$ and $\delta_{s_1}(\theta - 2\pi)$ belong to $]-1;\alpha_1[\cup]\beta_1;1[$ for all $\theta \in \left[\frac{3\pi}{4}; \frac{5\pi}{4}\right] \cup \left[\frac{7\pi}{4}; \frac{9\pi}{4}\right]$. With these definitions, δ_{s_2} is alreadya C^∞-diffeomorphism between $\left[\frac{3\pi}{4}; \frac{5\pi}{4}\right]$ and

$$\left[\phi_{s_2}^{-1}\left(\phi_{s_1}\left(\delta_{s_1}\left(\frac{3\pi}{4}\right)\right)\right); \phi_{s_2}^{-1}\left(\phi_{s_1}\left(\delta_{s_1}\left(\frac{5\pi}{4}\right)\right)\right)\right] = [d_2; d_1]$$

and between $\left[\frac{7\pi}{4}; \frac{9\pi}{4}\right]$ and

$$\left[\phi_{s_2}^{-1}\left(\phi_{s_1}\left(\delta_{s_1}\left(-\frac{\pi}{4}\right)\right)\right); \phi_{s_2}^{-1}\left(\phi_{s_1}\left(\delta_{s_1}\left(\frac{\pi}{4}\right)\right)\right)\right] = [d_4; d_3].$$

On $\left[\frac{5\pi}{4}; \frac{7\pi}{4}\right]$, let's define δ_{s_2} as any C^∞-increasing diffeomorphism from $\left[\frac{5\pi}{4}; \frac{7\pi}{4}\right]$ to $[d_1; d_4]$ whose derivatives up to order k at the endpoints of the interval are compatible with those of the definitions (3.8a) and (3.8b): for all $k' = 1, \dots, k$,

$$\begin{aligned}
\delta_{s_2}^{(k')}\left(\frac{5\pi}{4}\right) &= (\phi_{s_2}^{-1} \circ \phi_{s_1} \circ \delta_{s_1})^{(k')}\left(\frac{5\pi}{4}\right), \\
\delta_{s_2}^{(k')}\left(\frac{7\pi}{4}\right) &= (\phi_{s_2}^{-1} \circ \phi_{s_1} \circ \delta_{s_1})^{(k')}\left(-\frac{\pi}{4}\right).
\end{aligned}$$

Such a diffeomorphism exists (see Proposition B.4 in the appendix). With these definitions, δ_{s_2} is a C^k-diffeomorphism from $\left[\frac{3\pi}{4}; \frac{9\pi}{4}\right]$ to $[d_2; d_3]$.

Now, we have finished defining ψ, in accordance with Equations (3.6a) and (3.6b). Let's verify that this definition indeed makes it a C^k-diffeomorphism from $\mathbb{S}^1$ to $U_{s_1} \cup U_{s_2}$. First, it is a C^k function: it is C^k on $\left\{e^{i\theta}, \theta \in \left]-\frac{\pi}{4}; \frac{5\pi}{4}\right[\right\}$ since $\phi_{s_1} \circ \delta_{s_1}$ is, and it is C^k on $\left\{e^{i\theta}, \theta \in \left]\frac{3\pi}{4}; \frac{9\pi}{4}\right[\right\}$ since $\phi_{s_2} \circ \delta_{s_2}$ is. Thus, it is C^k on the union of these two sets, which is the entire $\mathbb{S}^1$.

Proposition 3.13

The map ψ establishes a bijection from $\mathbb{S}^1$ to $U_{s_1} \cup U_{s_2}$, and its inverse is given by:

$$\begin{aligned}
\zeta(x) &= e^{i\delta_{s_1}^{-1}(\phi_{s_1}^{-1}(x))} \quad \text{for all } x \in \phi_{s_1}([c_1; c_4]), \\
&= e^{i\delta_{s_2}^{-1}(\phi_{s_2}^{-1}(x))} \quad \text{for all } x \in \phi_{s_2}([d_2; d_3]).
\end{aligned}$$

Proof. The map ψ is surjective onto $U_{s_1} \cup U_{s_2}$. Indeed, according to its definition (Equations (3.6a) and (3.6b)),

$$\begin{aligned}
\psi(\mathbb{S}^1) &= \phi_{s_1}\left(\delta_{s_1}\left(\left[-\frac{\pi}{4}; \frac{5\pi}{4}\right]\right)\right) \cup \phi_{s_2}\left(\delta_{s_2}\left(\left[\frac{3\pi}{4}; \frac{9\pi}{4}\right]\right)\right) \\
&= \phi_{s_1}([c_1; c_4]) \cup \phi_{s_2}([d_2; d_3])
\end{aligned}$$

Now,

$$\begin{aligned}
U_{s_1} \cup U_{s_2} &= \phi_{s_1}(]-1; 1[) \cup \phi_{s_2}(]-1; 1[) \\
&= \phi_{s_1}(]-1; c_1]) \cup \phi_{s_1}(]c_1; c_4[) \cup \phi_{s_1}([c_4; 1[) \\
&\qquad \cup \phi_{s_2}(]-1; d_2]) \cup \phi_{s_2}(]d_2; d_3[) \cup \phi_{s_2}([d_3; 1[) \\
&= \phi_{s_2}([d_1; \alpha_2[) \cup \phi_{s_1}(]c_1; c_4[) \cup \phi_{s_2}(]\beta_2; d_4]) \\
&\qquad \cup \phi_{s_1}([c_2; \alpha_2[) \cup \phi_{s_2}(]d_2; d_3[) \cup \phi_{s_1}(]\beta_1; c_3]) \\
&\quad \text{(by Equation (3.4))} \\
&\subset \phi_{s_1}(]c_1; c_4[) \cup \phi_{s_2}(]d_2; d_3[) \\
&\subset U_{s_1} \cup U_{s_2},
\end{aligned}$$

which implies $\phi_{s_1}([c_1; c_4]) \cup \phi_{s_2}([d_2; d_3]) = U_{s_1} \cup U_{s_2}$.

On the other hand, ψ is injective. To show this, suppose $\theta, \theta' \in \mathbb{R}$ such that

$$\psi(e^{i\theta}) = \psi(e^{i\theta'}),$$

and prove that $e^{i\theta} = e^{i\theta'}$. First, if both θ and θ' belong to $\left[-\frac{\pi}{4}; \frac{5\pi}{4}\right]$ (modulo 2π), then, according to the definition (3.6a) and the injectivity of ϕ_{s_1} and δ_{s_1},

$$\theta \equiv \theta'[2\pi] \quad \Rightarrow \quad e^{i\theta} = e^{i\theta'}.$$

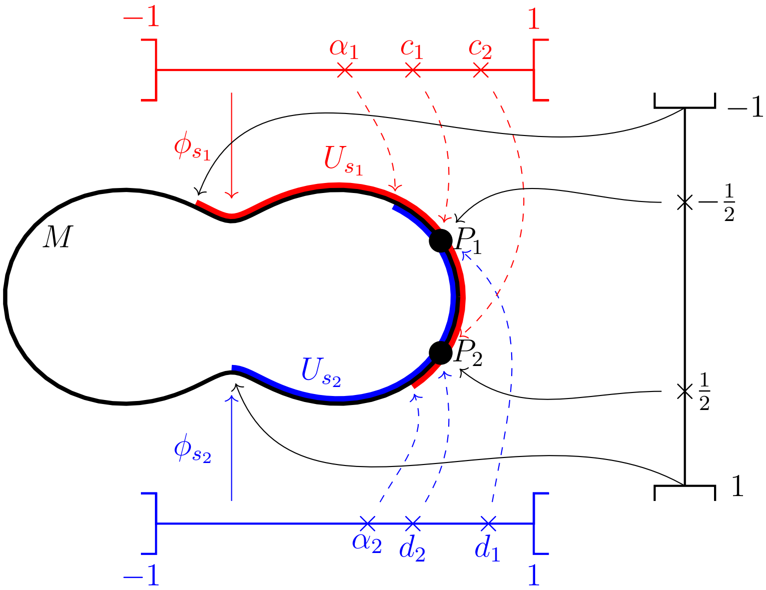


Figure 3.4: Illustration of the notation of Lemma 3.11 and a schematic representation of the diffeomorphism from $]-1;1[$ to $U_{s_1} \cup U_{s_2}$.

Similarly, if both θ and θ' belong to $\left[\frac{3\pi}{4};\frac{9\pi}{4}\right]$ (modulo 2π), then $e^{i\theta} = e^{i\theta'}$. Now, assume that neither of these situations holds, for example, that θ belongs to $\left[-\frac{\pi}{4};\frac{5\pi}{4}\right]$ but not to $\left[\frac{3\pi}{4};\frac{9\pi}{4}\right]$ (meaning θ belongs to $\left]\frac{\pi}{4};\frac{3\pi}{4}\right[$) and θ' belongs to $\left[\frac{3\pi}{4};\frac{9\pi}{4}\right]$ but not to $\left[-\frac{\pi}{4};\frac{5\pi}{4}\right]$ (meaning θ' belongs to $\left]\frac{5\pi}{4};\frac{7\pi}{4}\right[$). Then

$$\psi(e^{i\theta}) \in \phi_{s_1}\left(\delta_{s_1}\left(\left]\frac{\pi}{4};\frac{3\pi}{4}\right[\right)\right) = \phi_{s_1}(]c_2;c_3[)$$
$$\psi(e^{i\theta'}) \in \phi_{s_2}\left(\delta_{s_2}\left(\left]\frac{5\pi}{4};\frac{7\pi}{4}\right[\right)\right) = \phi_{s_2}(]d_1;d_4[).$$

However, $\phi_{s_1}(]c_2;c_3[)$ and $\phi_{s_2}(]d_1;d_4[)$ have an empty intersection (see Figure 3.3; this is verified with Equation (3.4)). Therefore, we cannot have $\psi(e^{i\theta}) = \psi(e^{i\theta'})$: this case is impossible. This completes the proof of injectivity.

Thus, we have shown that ψ is a bijection. The formula for the inverse follows from the definition of ψ in Equations (3.6a) and (3.6b). □

Finally, since $\psi^{-1} = \zeta$ is of class C^k (the functions $\delta_{s_1}, \delta_{s_2}, \phi_{s_1}, \phi_{s_2}$ are C^k), ψ is a C^k-diffeomorphism. □

Proof of Lemma 3.11. The proof is quite similar to that of Lemma 3.10, and only the main ideas will be outlined here.

We assume that U_{s_1}, U_{s_2} satisfy Property 2 of Lemma 3.9. If $U_{s_1} \subset U_{s_2}$, then $U_{s_1} \cup U_{s_2} = U_{s_2}$ is C^k-diffeomorphic to $]-1;1[$, according to our assumptions on U_{s_2}. The same holds if $U_{s_2} \subset U_{s_1}$.

We can therefore assume that the sub-property 2b is true: $\phi_{s_1}^{-1}(U_{s_1} \cap U_{s_2})$ and $\phi_{s_2}^{-1}(U_{s_1} \cap U_{s_2})$ are of the form $]-1;\alpha[$ or $]\alpha;1[$. We can assume that they are respectively equal to $]\alpha_1;1[$ and $]\alpha_2;1[$ for real numbers $\alpha_1, \alpha_2 \in]-1;1[$ (see Figure 3.4 for an illustration of the notations).

Let $c_1, c_2 \in]\alpha_1;1[$ such that $c_1 < c_2$. We denote

$$P_1 = \phi_{s_1}(c_1), \quad P_2 = \phi_{s_1}(c_2),$$
$$d_1 = \phi_{s_2}^{-1}(P_1), \quad d_2 = \phi_{s_2}^{-1}(P_2).$$

Since $\phi_{s_2}^{-1} \circ \phi_{s_1}$ is a decreasing C^k-diffeomorphism from $]\alpha_1;1[$ to $]\alpha_2;1[$ (for the same reasons as in Proposition 3.12), we have $\alpha_2 < d_2 < d_1 < 1$.

We define $\psi :]-1;1[\to U_{s_1} \cup U_{s_2}$ as follows:

$$\psi(x) = \phi_{s_1}(\delta_{s_1}(x)) \quad \text{for all } x \in \left]-1;\frac{1}{2}\right] \tag{3.9a}$$

$$\psi(x) = \phi_{s_2}(\delta_{s_2}(x)) \quad \text{for all } x \in \left[-\frac{1}{2}; 1\right[\tag{3.9b}$$

where δ_{s_1} is a C^∞-diffeomorphism from $]-1; \frac{1}{2}]$ to $]-1; c_2]$ such that

$$\delta_{s_1}\left(-\frac{1}{2}\right) = c_1, \quad \delta_{s_1}\left(\frac{1}{2}\right) = c_2,$$

and δ_{s_2} is a decreasing C^k-diffeomorphism from $[-\frac{1}{2}; 1[$ to $]-1; d_1]$ such that, on $[-\frac{1}{2}; \frac{1}{2}]$,

$$\delta_{s_2} = \phi_{s_2}^{-1} \circ \phi_{s_1} \circ \delta_{s_1}$$

and, on $[\frac{1}{2}; 1[$, δ_{s_2} is any decreasing C^k-diffeomorphism from $[\frac{1}{2}; 1[$ to $]-1; d_2]$ such that, for all $k' = 1, \ldots, k$,

$$\delta_{s_2}^{(k')}\left(\frac{1}{2}\right) = \left(\phi_{s_2}^{-1} \circ \phi_{s_1} \circ \delta_{s_1}\right)^{(k')}\left(\frac{1}{2}\right).$$

The existence of $\delta_{s_1}, \delta_{s_2}$ is guaranteed by Propositions B.3 and B.4. With these definitions for $\delta_{s_1}, \delta_{s_2}$, the definition of ψ in Equations (3.9a) and (3.9b) is valid. Moreover, the function ψ is of class C^k.

The same reasoning as in Proposition 3.13 can be used to show that ψ is a bijection between $]-1; 1[$ and $U_{s_1} \cup U_{s_2}$. Its inverse is given by

$$\begin{aligned} \zeta(x) &= \delta_{s_1}^{-1}(\phi_{s_1}^{-1}(x)) \quad \text{for all } x \in \phi_{s_1}(]-1; c_2]), \\ &= \delta_{s_2}^{-1}(\phi_{s_2}^{-1}(x)) \quad \text{for all } x \in \phi_{s_2}(]-1; d_1]). \end{aligned}$$

Since this inverse is C^k, ψ is a C^k-diffeomorphism between $]-1; 1[$ and $U_{s_1} \cup U_{s_2}$.

□

3.1.3 Length and arc length parametrization

We will now define the *length* of a curve. Intuitively, what is it? Let (I, γ) be a global parameterization of the curve, and imagine an ant walking along the curve: at time t, it is at point $\gamma(t)$. The length of the arc is the total distance covered by the ant over time. As, at time t, its absolute velocity is $||\gamma'(t)||_2$, the length should be defined as the integral over I of $||\gamma'||_2$.

Definition 3.14 : length of a curve

Let M be a connected curve. Let (I, γ) be a global parameterization of M. The *length* of M is defined as

$$\ell(M) = \int_I ||\gamma'(t)||_2 dt.$$

Proposition 3.15

The length is well-defined: if (I, γ) and (J, δ) are two global parameterizations of M, then

$$\int_I ||\gamma'(t)||_2 dt = \int_J ||\delta'(t)||_2 dt.$$

Proof. Let's consider the case where M is non-compact. Then γ and δ are diffeomorphisms from (respectively) I and J to M. Let

$$\theta = \gamma^{-1} \circ \delta : J \to I.$$

It is a diffeomorphism from J to I, and we have $\delta = \gamma \circ \theta$. Then

$$\int_J ||\delta'(t)||_2 \, dt = \int_J ||(\gamma \circ \theta)'(t)||_2 \, dt$$

$$= \int_J |\theta'(t)| \, ||\gamma' \circ \theta(t)||_2 \, dt$$
$$= \int_I ||\gamma'(t)||_2 \, dt.$$

The last equality is obtained by the change of variable formula applied to the function $||\gamma'||$, with change of variable given by θ.

We omit the case where M is compact. The principle is the same, with a subtlety related to the fact that γ and δ are not exactly diffeomorphisms from their domain to M.[4] □

Definition 3.16 : arc length

A global parametrization (I, γ) of a connected curve M is called an *arc length parametrization* if

$$||\gamma'(t)||_2 = 1, \quad \forall t \in I.$$

It is worth noting that if (I, γ) is an arc length parametrization of M, then the length of M is equal to the length of I:

$$\ell(M) = \int_I 1 \, dt = \sup I - \inf I.$$

Theorem 3.17 : existence of an arc length parametrization

For every connected curve M, there exists an arc length parametrization.

Proof. Let's consider the case where M is not compact (the compact case is similar with slightly different notation). Let $\phi : \mathbb{R} \to M$ be a C^k-diffeomorphism. We seek an arc length parametrization in the form $(I, \phi \circ \theta)$ where I is an open interval containing 0 and $\theta : I \to \mathbb{R}$ is an increasing C^k-diffeomorphism such that $\theta(0) = 0$.

For $(I, \phi \circ \theta)$ to be an arc length parametrization, it must satisfy, for all $t \in I \cap \mathbb{R}_0^+$,

$$\begin{aligned} t &= \ell(\phi \circ \theta(]0; t[)) \\ &= \ell(\phi(]\theta(0); \theta(t)[)) \\ &= \int_0^{\theta(t)} ||\phi'(s)||_2 \, ds. \end{aligned} \tag{3.10}$$

A similar equation holds for $t \in I \cap \mathbb{R}_0^-$.

Let's define

$$\begin{array}{rccc} L: & \mathbb{R} & \to & \mathbb{R} \\ & T & \to & \int_0^T ||\phi'(s)||_2 \, ds. \end{array}$$

This is a C^k-smooth map whose derivative does not vanish: it is a C^k-diffeomorphism between $\mathbb{R}$ and its image, which is an open interval. Let I be this image. Define, as required by Equation (3.10),

$$\theta = L^{-1} : I \to \mathbb{R}.$$

With this definition, $(I, \phi \circ \theta)$ is a global parametrization of M. For all $t \in I$,

$$\begin{aligned} (\phi \circ \theta)'(t) &= \theta'(t)\phi'(\theta(t)) \\ &= (L^{-1})'(t)\phi'(\theta(t)) \end{aligned}$$

[4] For particularly curious readers, here's how to resolve this difficulty. Let a, b, c, d be real numbers such that $I = [a; b[$ and $J = [c; d[$. Let $\alpha \in [0; d - c[$ be such that $\gamma(a) = \delta(c + \alpha)$. By replacing (J, δ) with $(\tilde{J}, \tilde{\delta})$, where $\tilde{J} = [c + \alpha; d + \alpha[$ and $\tilde{\delta} = \delta$ on $[c + \alpha; d[$ and $\tilde{\delta} = \delta(. - (d - c))$ elsewhere (which does not change the integral of $||\delta'||$), we can assume that $\gamma(a) = \delta(c)$. Then γ and δ are diffeomorphisms from $]a; b[$ and $]c; d[$ to $M - \{\gamma(a)\}$. We can define, as in the non-compact case,

$$\theta = \gamma^{-1} \circ \delta :]c; d[\to]a; b[$$

and proceed in the same way as before.

$$= \frac{\phi'(\theta(t))}{L'(L^{-1}(t))}$$
$$= \frac{\phi'(\theta(t))}{L'(\theta(t))}$$
$$= \frac{\phi'(\theta(t))}{||\phi'(\theta(t))||_2}.$$

This vector always has norm 1: $(I, \phi \circ \theta)$ is an arc length parametrization. □

The concept of arc length parametrization allows for the straightforward definition of several quantities that describe the "local shape" of curves. We do not have time to present them in detail in this course, but for general culture, here are some examples. If (I, γ) is an arc length parametrization, the vector

$$\gamma'(t)$$

is called the *unit tangent vector* at the point $\gamma(t)$. If γ is of class C^2, the vector

$$\frac{\gamma''(t)}{||\gamma''(t)||_2}$$

is called the *principal unit normal vector* at $\gamma(t)$ (which is well-defined only if $\gamma''(t) \neq 0$), and

$$||\gamma''(t)||_2$$

is the *curvature* at $\gamma(t)$ (which can be assigned a sign, positive or negative, when the curve is a submanifold of $\mathbb{R}^2$). Informally, curvature characterizes how quickly the curve "turns" in the vicinity of $\gamma(t)$.

3.2 Submanifolds of any dimension

In this section, several proofs are deferred to the appendix to make reading easier.

3.2.1 Distance and geodesics

We will now use the notion of length introduced in Definition 3.14 to define a distance on any connected submanifold M of $\mathbb{R}^n$: the distance between two points x_1, x_2 is the infimum of the lengths of paths connecting these points.

In this section, we call a *path* connecting two points x_1 and x_2 any function $\gamma : [0; A] \to M$, for some $A \in \mathbb{R}^+$, such that

- γ is continuous;
- γ is piecewise C^1;
- $\gamma(0) = x_1$ and $\gamma(A) = x_2$.

We can extend Definition 3.14 from curves to paths: the *length* of a path γ is

$$\ell(\gamma) = \int_0^A ||\gamma'(t)||_2 dt.$$

Definition 3.18 : distance on a submanifold

Let M be a connected submanifold of $\mathbb{R}^n$. We define a distance on M as follows: for all $x_1, x_2 \in M$,

$$\text{dist}_M(x_1, x_2) = \inf\{\ell(\gamma), \gamma \text{ is a path connecting } x_1 \text{ and } x_2\}.$$

Proposition 3.19

The function dist_M is well-defined: for all x_1, x_2, there exists a path connecting x_1 and x_2.

Proof. See section C.1. □

Proposition 3.20

The function dist_M is indeed a distance.

Proof.

- Symmetry: let $x_1, x_2 \in M$. Consider a sequence $(\gamma_n)_{n\in\mathbb{N}}$ of paths connecting x_1 to x_2 such that

$$\ell(\gamma_n) \xrightarrow{n\to+\infty} \text{dist}_M(x_1, x_2).$$

For each n, let $[0; A_n]$ be the domain of γ_n, and define

$$\begin{array}{rccc} \delta_n : & [0; A_n] & \to & M \\ & t & \to & \gamma_n(A_n - t). \end{array}$$

This is a path connecting x_2 to x_1. Moreover, for every n,

$$\ell(\delta_n) = \int_0^{A_n} ||-\gamma_n'(A_n - t)||_2 dt = \int_0^{A_n} ||\gamma_n'(t)||_2 dt = \ell(\gamma_n),$$

so that $\text{dist}_M(x_2, x_1) \leq \ell(\delta_n) = \ell(\gamma_n)$. By taking the limit as $n \to +\infty$, we deduce

$$\text{dist}_M(x_2, x_1) \leq \text{dist}_M(x_1, x_2).$$

The reasoning we just presented remains true if we exchange x_1 and x_2. Therefore, we also have

$$\text{dist}_M(x_1, x_2) \leq \text{dist}_M(x_2, x_1).$$

Hence, $\text{dist}_M(x_1, x_2) = \text{dist}_M(x_2, x_1)$.

- Triangle inequality: let $x_1, x_2, x_3 \in M$. Let's prove the inequality

$$\text{dist}_M(x_1, x_3) \leq \text{dist}_M(x_1, x_2) + \text{dist}_M(x_2, x_3).$$

Consider $(\gamma_n : [0; A_n] \to M)_{n\in\mathbb{N}}$ and $(\delta_n : [0; B_n] \to M)_{n\in\mathbb{N}}$ two sequences of paths connecting, respectively, x_1 to x_2 and x_2 to x_3, such that

$$\ell(\gamma_n) \xrightarrow{n\to+\infty} \text{dist}_M(x_1, x_2);$$
$$\ell(\delta_n) \xrightarrow{n\to+\infty} \text{dist}_M(x_2, x_3).$$

For each n, define

$$\begin{array}{rcccl} \zeta_n : & [0; A_n + B_n] & \to & M & \\ & t & \to & \gamma_n(t) & \text{if } t \leq A_n \\ & & & \delta_n(t - A_n) & \text{if } A_n < t. \end{array}$$

For each n, we have $\zeta_n(0) = x_1$ and $\zeta_n(A_n + B_n) = x_3$. As γ_n and δ_n are continuous, ζ_n is continuous on $[0; A_n[$ and on $]A_n; A_n + B_n]$. It is also continuous at A_n since it has left and right limits at this point, which are identical:

$$\zeta_n(t) \xrightarrow{t\to A_n^-} \gamma_n(A_n) = x_2 = \delta_n(0) \xleftarrow{t\to A_n^+} \zeta_n(t).$$

Therefore, the function ζ_n is continuous. Moreover, it is piecewise C^1 since γ_n and δ_n are piecewise C^1, so it is a path. Its length is

$$\ell(\zeta_n) = \int_0^{A_n+B_n} ||\zeta_n'(t)||_2 dt$$

$$
\begin{aligned}
&= \int_0^{A_n} ||\gamma_n'(t)||_2 dt + \int_{A_n}^{A_n+B_n} ||\delta_n'(t-A_n)||_2 dt \\
&= \int_0^{A_n} ||\gamma_n'(t)||_2 dt + \int_0^{B_n} ||\delta_n'(t)||_2 dt \\
&= \ell(\gamma_n) + \ell(\delta_n).
\end{aligned}
$$

Thus, for every n, $\mathrm{dist}_M(x_1, x_3) \leq \ell(\gamma_n) + \ell(\delta_n)$, implying, in the limit,

$$\mathrm{dist}_M(x_1, x_3) \leq \mathrm{dist}_M(x_1, x_2) + \mathrm{dist}_M(x_2, x_3).$$

- Separation: for any $x \in M$, $\mathrm{dist}_M(x, x) = 0$: by choosing a constant path γ with value x, we have $\mathrm{dist}_M(x, x) \leq \ell(\gamma) = 0$.

  Let's prove the converse. For all $x_1, x_2 \in M$ and any path γ connecting x_1 to x_2,

$$
\begin{aligned}
\ell(\gamma) &= \int_0^A ||\gamma'(t)||_2 dt \\
&\geq \left|\left| \int_0^A \gamma'(t) dt \right|\right|_2 \quad \text{(by triangle inequality)} \\
&= \left|\left| [\gamma(t)]_0^A \right|\right|_2 \\
&= ||x_2 - x_1||_2.
\end{aligned}
$$

  Consequently,

$$\mathrm{dist}_M(x_1, x_2) \geq ||x_2 - x_1||_2.$$

  In particular, if $\mathrm{dist}_M(x_1, x_2) = 0$, then $||x_2 - x_1||_2 = 0$, implying $x_1 = x_2$.

□

Theorem 3.21 : existence of minimizing paths

Let M be, again, a connected submanifold of $\mathbb{R}^n$, of class C^k. Additionally, suppose that

- $k \geq 2$;
- M is closed in $\mathbb{R}^n$.

Then, for all $x_1, x_2 \in M$, the infimum in Definition 3.18 is a minimum: there exists a path γ connecting x_1 to x_2 such that

$$\ell(\gamma) = \mathrm{dist}_M(x_1, x_2).$$

If γ is a minimizing path, as in the previous theorem, there exists a reparametrization $\tilde{\gamma} \overset{def}{=} \gamma \circ \phi$ of constant speed: for some c,

$$||\tilde{\gamma}'(t)||_2 = c \text{ for all } t.$$

(The argument is the same as for Theorem 3.17; one can even impose $c = 1$ if desired.)

These minimizing paths traversed with constant speed are characterized by a simple differential equation, given in a new theorem.

Theorem 3.22 : geodesic equation

Keep the same notations and assumptions as in the previous theorem. Let $\gamma : [0; A] \to M$ be a path connecting x_1 to x_2, with constant speed, such that $\ell(\gamma) = \mathrm{dist}_M(x_1, x_2)$. Then, γ is C^2, and

$$\gamma''(t) \in (T_{\gamma(t)} M)^\perp, \quad \forall t \in [0; A]. \tag{3.11}$$

Simultaneous proof of Theorems 3.21 and 3.22. Fix x_1, x_2. To simplify notation, let $D = \mathrm{dist}_M(x_1, x_2)$.

Let $(\gamma_N)_{N\in\mathbb{N}}$ be any sequence of paths connecting x_1 to x_2 such that

$$\ell(\gamma_N) \overset{n\to+\infty}{\longrightarrow} D.$$

Without loss of generality, we can assume that the γ_N have constant speed $c > 0$ (for some arbitrary c):

$$||\gamma_N'(t)||_2 = c, \quad \forall N \in \mathbb{N}, \forall t.$$

After this reparametrization, the domain of γ_N is $[0; \ell(\gamma_N)/c]$.

A compactness argument allows us to assume that $(\gamma_N)_{N\in\mathbb{N}}$ converges uniformly to a certain limit. This is stated in the following proposition, proved in section C.2.

Proposition 3.23

There exists a function $\delta : [0; D/c] \to M$ and an extraction $\rho : \mathbb{N} \to \mathbb{N}$ such that

$$||\gamma_{\rho(N)} - \delta||_\infty \overset{n\to+\infty}{\longrightarrow} 0.$$

Moreover,

- δ is c-Lipschitz;
- $\delta(0) = x_1$ and $\delta(D/c) = x_2$.

Define δ and ρ as in the proposition. We replace $(\gamma_N)_{N\in\mathbb{N}}$ with the subsequence $(\gamma_{\rho(N)})_{n\in\mathbb{N}}$. We will show that δ is a path connecting x_1 to x_2 (we only need to show that it is piecewise C^1), of class C^2, satisfying Equation (3.11), and such that

$$\ell(\delta) = D.$$

This will directly prove Theorem 3.21 and will also imply Theorem 3.22 (because for any path γ such that $\ell(\gamma) = D$, we can apply the reasoning we just did to the constant sequence $\gamma_N = \gamma, \forall N \in \mathbb{N}$; the only accumulation point of this sequence is $\delta \overset{def}{=} \gamma$, so if δ satisfies Equation (3.11), then γ satisfies it as well).

This proof must still be completed. It will hopefully be available in a few days or weeks.

□

Remark

Theorem 3.21, which guarantees the existence of a path with minimal length between arbitrary points, may no longer be true if the considered submanifold is not closed. For example, in the submanifold $M \overset{def}{=} \mathbb{R}^2 \setminus \{(0,0)\}$, there is no minimizing path between $(-1,0)$ and $(1,0)$.
However, when the submanifold M is not closed, it can be shown (and the proof is very similar to the preceding one) that any point $x_1 \in M$ has a neighborhood V such that, for any $x_2 \in V$, there exists a path of minimal length between x_1 and x_2.
Theorem 3.22, on the other hand, remains true if the considered submanifold is not closed.

Curves satisfying Equation (3.11), whether or not they are paths of minimal length between two points, are called *geodesics*.

Definition 3.24 : geodesics

Let M be a submanifold of $\mathbb{R}^n$ of class C^k with $k \geq 2$. We call a *geodesic* any map $\gamma : I \to M$ (for I a non-empty interval of $\mathbb{R}$) of class C^2 such that, for all $t \in I$,

$$\gamma''(t) \in (T_{\gamma(t)}M)^\perp.$$

$$\begin{tikzcd}
\boxed{\text{a geodesic}} \arrow[r, shift right, "\text{is locally}"'] & \boxed{\begin{array}{c}\text{a path of minimal length,}\\ \text{parametrized at constant speed}\end{array}} \arrow[l, shift right, "\text{is}"']
\end{tikzcd}$$

Figure 3.5: Relations between geodesics and a path of minimal length

Proposition 3.25

A geodesic γ always has constant speed: $||\gamma'(t)||_2$ is independent of t.

Proof. Let $\gamma : I \to M$ be a geodesic in a certain submanifold M. Define

$$N : t \in I \to ||\gamma'(t)||_2^2.$$

This function is differentiable and, for all t,

$$N'(t) = 2 \left\langle \gamma'(t), \gamma''(t) \right\rangle.$$

Now, for all t, $\gamma'(t) \in T_{\gamma(t)}M$, and since γ is a geodesic, $\gamma''(t) \in (T_{\gamma(t)}M)^\perp$. So, for all t,

$$N'(t) = 0,$$

which means that N, and thus also $||\gamma'||_2$, is a constant function. □

As summarized on Figure 3.5, a path of minimal length, parametrized at constant speed, is always a geodesic (from Theorem 3.22). The converse may not be true (an example will be provided in Subsection 3.2.2). However, it is *locally* true, as stated in the following proposition.

Proposition 3.26 : geodesics are locally minimizing

Let M be a submanifold of $\mathbb{R}^n$, of class C^k with $k \geq 2$. Let I be a non-empty interval and $\gamma : I \to M$ a geodesic.
For all $t \in I$, there exists $\epsilon > 0$ such that, for all $t' \in [t - \epsilon; t + \epsilon]$,

$$\gamma_{|[t;t']} \text{ is a path with minimal length between } \gamma(t) \text{ and } \gamma(t').$$

Unfortunately, the proof of this proposition requires tools from differential equations, which will only be introduced in the next chapter. Depending on the time I have, I may or may not provide a proof for it at some point in the semester.

Exercise 3 : geodesics on product submanifolds

Let $n_1, n_2 \in \mathbb{N}^*$ be integers. Let $M_1 \subset \mathbb{R}^{n_1}$ and $M_2 \subset \mathbb{R}^{n_2}$ be submanifolds of class C^2.
Let $I \subset \mathbb{R}$ be a non-empty interval and $\gamma : I \to M_1 \times M_2$ be a map. We denote $\gamma_1 : I \to M_1, \gamma_2 : I \to M_2$ its components.

1. Show that γ is a geodesic in M if and only if γ_1 is a geodesic in M_1 and γ_2 is a geodesic in M_2.
2. In this question, we assume that M_1, M_2 are closed. We also assume that γ is a path, joining two points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ in $M_1 \times M_2$.
   a) Show that, if γ_1 and γ_2 have constant speed, then
   $$\ell(\gamma) = \sqrt{\ell(\gamma_1)^2 + \ell(\gamma_2)^2}.$$
   b) Show that, if γ has constant speed and $\ell(\gamma) = \text{dist}_{M_1 \times M_2}(x, y)$, then γ_1 and γ_2 have constant speed.
   [Hint: use Theorem 3.22, Question 1. and Proposition 3.25.]

c) Deduce from the previous question that

$$\mathrm{dist}_{M_1\times M_2}(x,y) \geq \sqrt{\mathrm{dist}_{M_1}(x_1,y_1)^2 + \mathrm{dist}_{M_2}(x_2,y_2)^2}.$$

d) Show that

$$\mathrm{dist}_{M_1\times M_2}(x,y) = \sqrt{\mathrm{dist}_{M_1}(x_1,y_1)^2 + \mathrm{dist}_{M_2}(x_2,y_2)^2}.$$

e) Show that γ is a path with minimal length connecting x to y, with constant speed, if and only if γ_1 is a path with minimal length connecting x_1 to y_1, with constant speed, and γ_2 is a path with minimal length connecting x_2 to y_2, with constant speed.

f) For $n_1 = n_2 = 1$ and $M_1 = M_2 = \mathbb{R}$, give an example of paths γ_1, γ_2 connecting 0 to 1, with minimal length (but non-constant speed) such that $\gamma \overset{def}{=} (\gamma_1, \gamma_2)$ is not a path with minimal length connecting $(0,0)$ to $(1,1)$.

3.2.2 Examples: the model submanifold and the sphere

Exercise 4 : model submanifold

For any $n \in \mathbb{N}^*$ and $d \in \{1,\dots,n\}$, we define $M \overset{def}{=} \mathbb{R}^d \times \{0\}^{n-d}$. Give a simple description of the geodesics in M.
(The solution is provided in Example 3.27, but do not read it before spending some time on the exercise!)

Example 3.27 : model submanifold

Let $n \in \mathbb{N}^*$ and $d \in \{1,\dots,n\}$. The geodesics of the "model" submanifold $M = \mathbb{R}^d \times \{0\}^{n-d}$ are the maps $\gamma : I \to \mathbb{R}^n$ of class C^2 such that

1. $\gamma_{d+1}(t) = \dots = \gamma_n(t) = 0$ for all $t \in I$ (since $\gamma(t) \in M$);
2. $\gamma_1''(t) = \dots = \gamma_d''(t) = 0$ for all $t \in I$ (since $\gamma''(t) \in (T_{\gamma(t)}M)^\perp = \{0\}^d \times \mathbb{R}^{n-d}$).

These are the maps whose last $n-d$ components are zero, and the first d components are affine. Geodesics are therefore exactly the maps of the form

$$\gamma : t \in I \to x_0 + tv,$$

for any $x_0, v \in \mathbb{R}^d \times \{0\}^{n-d}$.
More geometrically, we can say that geodesics are maps which parametrize lines in $\mathbb{R}^d \times \{0\}$ at constant speed.

Exercise 5 : geodesics on $\mathbb{S}^{n-1}$

Let $n \in \mathbb{N}^*$ be fixed. We want to compute the geodesics of $\mathbb{S}^{n-1}$.

1. Let us consider a geodesic γ, defined over an interval I. We know that it has constant speed. Let $c \in \mathbb{R}^+$ be this speed.
   a) Show that, for all $t \in I$, $\langle \gamma(t), \gamma'(t) \rangle = 0$.
   b) Differentiate the previous equality, and show that, for all $t \in I$,
   $$\big\langle \gamma(t), \gamma''(t) \big\rangle + c^2 = 0.$$
   c) Show that, for all $t \in I$, $\gamma''(t) = -c^2\gamma(t)$.
   d) Deduce from the previous equation that there exist $e_1, e_2 \in \mathbb{R}^n$ such that
   $$\gamma(t) = \cos(ct)e_1 + \sin(ct)e_2, \forall t \in I.$$

(e) Show that $\langle e_1, e_2 \rangle = 0$ and $||e_1||_2 = ||e_2||_2 = 1$.

2. Read and prove Proposition 3.28.

Proposition 3.28 : geodesics on $\mathbb{S}^{n-1}$

Let $n \geq 2$.
The geodesics on $\mathbb{S}^{n-1}$ are all maps of the form

$$\begin{array}{rrcc} \gamma : & I & \to & \mathbb{S}^{n-1} \\ & t & \to & \cos(ct)e_1 + \sin(ct)e_2, \end{array}$$

for any non-empty interval I, any real $c > 0$, and any vectors $e_1, e_2 \in \mathbb{R}^n$ such that

$$||e_1||_2 = ||e_2||_2 = 1 \quad \text{and} \quad \langle e_1, e_2 \rangle = 0.$$

Remark

This means that the geodesics on the sphere are parametrizations with constant speed of a "great circle"

$$\{\cos(s)e_1 + \sin(s)e_2, s \in \mathbb{R}\},$$

or an arc of it.

Proof of Proposition 3.28. First, let γ be a map of the specified form. Let's check that it is a geodesic. For any t,

$$(T_{\gamma(t)}\mathbb{S}^{n-1})^\perp = \left(\{\gamma(t)\}^\perp\right)^\perp = \mathrm{Vect}\{\gamma(t)\}.$$

Now, for any $t \in I$,

$$\gamma'(t) = c\,(-\sin(ct)e_1 + \cos(ct)e_2)\,;$$
$$\gamma''(t) = -c^2\,(\cos(ct)e_1 + \sin(ct)e_2) = -c^2\gamma(t) \in \mathrm{Vect}\{\gamma(t)\}.$$

Therefore, the geodesic equation is satisfied.

Conversely, let γ be a geodesic defined on an interval I. Let c be its speed (i.e., the positive real number such that $||\gamma'(t)||_2 = c$ for all t; recall that γ has constant speed according to Proposition 3.25). If $c = 0$, γ is constant, so γ is of the desired form (with $e_1 = \gamma(t_0)$ and any e_2). Let us now assume $c > 0$.

For any $t \in I$, $\gamma'(t) \in T_{\gamma(t)}\mathbb{S}^{n-1} = \{\gamma(t)\}^\perp$, so

$$0 = \left\langle \gamma(t), \gamma'(t) \right\rangle.$$

We differentiate this equality: for any t,

$$\begin{aligned} 0 &= \left\langle \gamma(t), \gamma''(t) \right\rangle + \left\langle \gamma'(t), \gamma'(t) \right\rangle \\ &= \left\langle \gamma(t), \gamma''(t) \right\rangle + c^2. \end{aligned}$$

Thus, $\langle \gamma(t), \gamma''(t) \rangle = -c^2$. As $\gamma''(t) \in (T_{\gamma(t)}\mathbb{S}^{n-1})^\perp = \mathrm{Vect}\{\gamma(t)\}$ and $\gamma(t)$ is a unit vector, we must have

$$\gamma''(t) = -c^2\gamma(t).$$

We know that any solution to this differential equation is of the form

$$\gamma : t \in I \to \cos(ct)e_1 + \sin(ct)e_2.$$

Fix e_1, e_2 so that γ has this expression. It remains to verify that $||e_1||_2 = ||e_2||_2 = 1$ and $\langle e_1, e_2 \rangle = 0$.

For this, fix any $t_0 \in I$. Let

$$v_1 = \gamma(t_0) \text{ and } v_2 = \frac{\gamma'(t_0)}{c}.$$

These are two unit vectors orthogonal to each other. Express e_1, e_2 in terms of v_1, v_2:

$$v_1 = \gamma(t_0) = \cos(ct_0)e_1 + \sin(ct_0)e_2;$$
$$v_2 = \frac{\gamma'(t_0)}{c} = -\sin(ct_0)e_1 + \cos(ct_0)e_2.$$

We deduce

$$e_1 = \cos(ct_0)v_1 - \sin(ct_0)v_2 \text{ and } e_2 = \sin(ct_0)v_1 + \cos(ct_0)v_2.$$

So, $||e_1||_2^2 = \cos^2(ct_0)||v_1||_2^2 - 2\cos(ct_0)\sin(ct_0)\langle v_1, v_2\rangle + \sin^2(ct_0)||v_2||_2^2 = 1$ and, similarly, $||e_2||_2^2 = 1$, $\langle e_1, e_2\rangle = 0$. □

Remark

The example of the sphere shows that geodesics are not always paths with minimal length between their endpoints. Indeed, for any e_1, e_2, the geodesic

$$\gamma : t \in [0; 2\pi] \to \cos(t)e_1 + \sin(t)e_2$$

joins e_1 to itself. However, the length of γ is non-zero.

Remark

The example of the sphere also shows that there can be multiple paths γ between two points x_1 and x_2 such that

$$\ell(\gamma) = \mathrm{dist}_M(x_1, x_2)$$

which are different even after reparameterization.
For instance, for any vectors e_1, e_2 with norm 1 and orthogonal to each other, the geodesics

$$\gamma_1 : t \in [0; \pi] \to \cos(t)e_1 + \sin(t)e_2,$$
$$\gamma_2 : t \in [0; \pi] \to \cos(t)e_1 - \sin(t)e_2$$

are paths of minimal length between e_1 and $-e_1$, but they are not equal even after reparameterization. However, it can be shown that paths of minimal length are "locally unique".

Corollary 3.29 : distance on $\mathbb{S}^{n-1}$

Let $n \geq 2$. Let $x_1, x_2 \in \mathbb{S}^{n-1}$. Then

$$\mathrm{dist}_{\mathbb{S}^{n-1}} = \arccos(\langle x_1, x_2\rangle).$$

Proof. According to Theorems 3.21 and 3.22, there exists at least one path γ connecting x_1 and x_2 such that

$$\ell(\gamma) = \mathrm{dist}_{\mathbb{S}^{n-1}}(x_1, x_2)$$

and such a path, if reparameterized at constant speed, is a geodesic.

Hence,

$$\mathrm{dist}_{\mathbb{S}^{n-1}}(x_1, x_2) = \min\{\ell(\gamma), \gamma \text{ geodesic connecting } x_1 \text{ and } x_2\}.$$

Let us compute this minimum.

Let γ be any geodesic connecting x_1 to x_2. We determine the possible values for its length. We can be assume that it is defined on an interval of the form $[0; A]$. Let c, e_1, e_2 be such that, for all $t \in [0; A]$,

$$\gamma(t) = \cos(ct)e_1 + \sin(ct)e_2.$$

It must hold that $x_1 = \gamma(0) = e_1$ and

$$x_2 = \gamma(A) = \cos(cA)e_1 + \sin(cA)e_2.$$

In particular, $\langle x_1, x_2\rangle = \langle e_1, x_2\rangle = \cos(cA)$, so

$$\begin{aligned} cA &= \arccos(\langle x_1, x_2\rangle) + 2k\pi \\ \text{or } cA &= (2\pi - \arccos(\langle x_1, x_2\rangle)) + 2k\pi, \end{aligned}$$

for some $k \in \mathbb{Z}$ (in fact, $k \in \mathbb{N}$ since $cA \geq 0$). As $\ell(\gamma) = cA$, it follows that the length of γ is at least

$$\min\left(\arccos(\langle x_1, x_2\rangle), 2\pi - \arccos(\langle x_1, x_2\rangle)\right) = \arccos(\langle x_1, x_2\rangle).$$

Thus,

$$\mathrm{dist}_{\mathbb{S}^{n-1}}(x_1, x_2) \geq \arccos(\langle x_1, x_2\rangle).$$

To show that the inequality is an equality, we observe that, if $e_2 = \frac{x_2 - \langle x_1, x_2\rangle x_1}{\sqrt{1 - \langle x_1, x_2\rangle^2}}$, the geodesic

$$\begin{array}{rccc} \gamma: & [0; \arccos(\langle x_1, x_2\rangle)] & \to & \mathbb{S}^{n-1} \\ & t & \to & \cos(t)x_1 + \sin(t)e_2 \end{array}$$

connects x_1 to x_2 and has length $\arccos(\langle x_1, x_2\rangle)$.

$\square$

Chapter 4

Differential equations: existence and uniqueness

What you should know or be able to do after this chapter

- Identify a Cauchy problem.
- Know the Cauchy-Lipschitz theorem; be able to apply it to particular situations.
- In the Cauchy-Lipschitz theorem, understand why the local Lipschitz continuity assumption is necessary. When possible, use the fact that the function is C^1 to show that this hypothesis is verified.
- Know what a maximal solution is.
- When true, show that the maximal solution exists and is unique, using Proposition 4.4.
- When an upper bound on the norm of the maximal solution is available, combine it with the théorème des bouts to show that the maximal solution is global (as in Example 4.9).

4.1 Cauchy-Lipschitz theorem

A *Cauchy problem* is a differential equation where the unknown is a function of one variable (often denoted as t), together with an initial condition. It is thus a problem of the following form:

$$\boxed{\begin{cases} u' &= f(t,u), \\ u(t_0) &= u_0. \end{cases}} \tag{Cauchy}$$

Here,

- $f : I \times U \to \mathbb{R}^n$ is a fixed function, with I an open interval of $\mathbb{R}$ and U an open set of $\mathbb{R}^n$ (for some $n \in \mathbb{N}^*$);
- t_0 is an element of I and u_0 an element of U;
- u is the unknown function, which must be defined on an interval J such that $t_0 \in J \subset I$, take values in U and be differentiable.

The equality "$u' = f(t,u)$" is a shortened notation for "$u'(t) = f(t,u(t))$": u is indeed a *function*, which depends on a variable, here called t.

Remark

In Problem (Cauchy), we impose the differential equation to be of order 1 (meaning it contains only one derivative). This is not a restriction. Indeed, a Cauchy problem containing a differential equation of any order $N \geq 1$ can be reformulated as a Cauchy problem of order 1. Precisely, consider a problem of the form

$$u^{(N)} = g\left(t, u, u', \dots, u^{(N-1)}\right)$$
$$u(t_0) = u_{0,0}, \quad u'(t_0) = u_{0,1}, \quad \dots, \quad u^{(N-1)}(t_0) = u_{0,N-1}.$$

If we denote $v_0 = u, v_1 = u', \dots, v_{N-1} = u^{(N-1)}$, it is equivalent to

$$\begin{aligned} v_0' &= v_1 \\ &\dots \\ v_{N-2}' &= v_{N-1} \\ v_{N-1}' &= g(t, v_0, v_1, \dots, v_{N-1}) \end{aligned}$$
$$v_0(t_0) = u_{0,0}, \quad v_1(t_0) = u_{0,1}, \quad \dots, \quad v_{N-1}(t_0) = u_{0,N-1},$$

which is a first-order problem on the unknown function $\begin{pmatrix} v_0 \\ \vdots \\ v_{N-1} \end{pmatrix}$.

Exercise 6

Show that a map $u : J \to U$ is a solution of Problem (Cauchy) if and only if the map

$$\begin{array}{rccc} \tilde{u}: & J & \to & J \times U \\ & t & \to & (t, u(t)) \end{array}$$

is a solution to another Cauchy problem, where the initial condition u_0 is replaced with (t_0, u_0) and f is replaced with a map $\tilde{f} : \mathbb{R} \times (I \times U) \to \mathbb{R}^{n+1}$ whose definition you will provide, which does not depend on its first argument.

The starting point of the theory of differential equations is the Cauchy-Lipschitz theorem, which, under regularity assumptions on f, guarantees that Problem (Cauchy) has a unique solution in the vicinity of t_0.

Theorem 4.1 : Cauchy-Lipschitz

Suppose f is continuous and there exists a neighborhood $H \subset I \times U$ of (t_0, u_0) where it is Lipschitz continuous in its second variable:

$$\begin{gathered} \forall t, u, v \text{ such that } (t,u), (t,v) \in H, \\ ||f(t,u) - f(t,v)||_2 \leq C||u - v||_2, \end{gathered} \tag{4.1}$$

for some constant $C > 0$ (which should not depend on t).
Then we have the following conclusions:

- (Existence)
  There exists an interval $J \subset I$ whose interior contains t_0 and a function $u : J \to U$ of class C^1 which is a solution of Problem (Cauchy).
- (Local Uniqueness)
  If u_1, u_2 are two C^1 maps solving Problem (Cauchy), defined on intervals J_1, J_2 containing t_0 (in their interior or on the boundary), then
  $$u_1 = u_2 \text{ on } J_1 \cap J_2 \cap [t_0 - \epsilon; t_0 + \epsilon]$$
  for any sufficiently small $\epsilon > 0$.

The most classical proof of this theorem uses (implicitly or explicitly) the *Picard fixed-point theorem.* Interested readers can find it, for example, in [Benzoni-Gavage, 2010, p. 142].

The Lipschitz continuity condition around (t_0, u_0) (Equation (4.1)) is automatically satisfied whenever f is C^1. Indeed, in this case, we can take $H = \overline{B}((t_0, u_0), \epsilon)$, for any $\epsilon > 0$ sufficiently small. Equation (4.1) then follows from the mean value inequality (Theorem 1.16), with

$$C = \max_{(t,u)\in\overline{B}((t_0,u_0),\epsilon)} ||df(t,u)||_{\mathcal{L}(\mathbb{R}^{n+1},\mathbb{R}^n)}.$$

The "existence" part of the theorem holds even without the Lipschitz condition (it suffices for f to be continuous; this is the *Peano theorem*). However, the "uniqueness" part may be false without this condition. To provide an example of possible non-uniqueness, consider the Cauchy problem

$$u' = \sqrt{u},$$
$$u(0) = 0.$$

It can be verified that the maps

$$\begin{array}{rccll} u_1 : & \mathbb{R} & \to & \mathbb{R} & \\ & t & \to & \frac{t^2}{4} & \text{if } t \geq 0, \\ & & & 0 & \text{if } t < 0, \end{array}$$

$$\begin{array}{rccl} u_2 : & \mathbb{R} & \to & \mathbb{R} \\ & t & \to & 0, \end{array}$$

are both solutions to this problem. However, they are not identical.

Let's conclude this section with a simple but useful property concerning the regularity of solutions to a Cauchy problem.

Proposition 4.2

If f is of class C^r for some $r \in \mathbb{N}$, any solution u of Problem (Cauchy) is of class C^{r+1}. In particular, if f is C^∞, every solution is C^∞.

Proof. We prove the result by induction on r. For $r = 0$, it is true: if u is a solution, it is differentiable by definition. In particular, it is continuous. Its derivative is

$$u' = f(t, u).$$

Since f and u are continuous, u' is also continuous, meaning u is C^1.

Let us assume that the result holds for some $r \in \mathbb{N}$ and prove it for $r + 1$. Suppose f is of class C^{r+1} and let u be a solution. Since f is also of class C^r, the induction hypothesis tells us u is C^{r+1}. Therefore,

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

is a composition of C^{r+1} functions. Thus, it is C^{r+1}, meaning u is C^{r+2}. □

Remark : extension to Banach spaces

Here, we limit ourselves to differential equations in finite dimension, meaning that the function u of Problem (Cauchy) takes values in $\mathbb{R}^n$. More generally, one can consider equations where the unknown function takes values in a Banach space[a], and everything said in this section remains true, except for Peano's theorem.

[a] that is, a complete normed vector space

4.2 Maximal solutions

Definition 4.3 : maximal solutions

Let $u : J \to U$ be a solution of a problem of the form (Cauchy). We say that it is a *maximal solution* of the problem if it cannot be extended to a larger interval: for any other solution $\tilde{u} : \tilde{J} \to U$ such that $J \subset \tilde{J}$ and $\tilde{u}_{|J} = u$, we have

$$\tilde{J} = J \quad \text{and} \quad \tilde{u} = u.$$

Proposition 4.4 : existence of a unique maximal solution

If the map f of Problem (Cauchy) is continuous, and Lipschitz continuous in its second variable around every point, then the problem has a unique maximal solution.

Moreover, if we denote by $u : J \to U$ this maximal solution, the set of solutions of Problem (Cauchy) is

$$\left\{ u_{|\tilde{J}} : \tilde{J} \to U \text{ with } \tilde{J} \text{ interval such that } t_0 \in \tilde{J} \subset J \right\}. \tag{4.2}$$

Proof. We start with a proposition (whose proof follows this one) which establishes a uniqueness result for solutions of Problem (Cauchy). This result is very similar to the one from the Cauchy-Lipschitz theorem, but it is global, while the Cauchy-Lipschitz theorem provides local guarantees only (uniqueness holds in a neighborhood of t_0). Here, we have a global uniqueness guarantee because f is Lipschitz in its second variable *around every point*, not just around (t_0, u_0).

Proposition 4.5

If $u_1 : J_1 \to U$ and $u_2 : J_2 \to U$ are two solutions of Problem (Cauchy), then

$$u_1 = u_2 \quad \text{on } J_1 \cap J_2.$$

Moreover, the function $u : J_1 \cup J_2 \to U$ which coincides with u_1 on J_1 and u_2 on J_2 is a solution of Problem (Cauchy).

From this proposition, we can already deduce that the maximal solution, if it exists, is unique and that the set of solutions of Problem (Cauchy) is indeed the one given in Equation (4.2).

Indeed, suppose there exists a maximal solution u, defined on an interval J. For any interval $\tilde{J}$ such that $t_0 \in \tilde{J} \subset J$, $u_{|\tilde{J}}$ is a solution of Problem (Cauchy). Conversely, if $v : \tilde{J} \to U$ is a solution of the problem, there exists (from the previous proposition) a solution defined on $J \cup \tilde{J}$, equal to u on J and v on $\tilde{J}$. Since u is maximal, we must have $J \cup \tilde{J} = J$, i.e., $\tilde{J} \subset J$, and $v = u$ on $\tilde{J} \cap J = \tilde{J}$. Therefore,

$$v = u_{|\tilde{J}}.$$

This proves Equation (4.2).

Equation (4.2), in turn, implies that the maximal solution is unique: every solution is of the form $u_{|\tilde{J}}$ for some $\tilde{J} \subset J$. Therefore, every solution $u_{|\tilde{J}}$ can be extended to the larger interval J, except u itself.

To conclude, let's show existence. Let us define

$$J = \{t \in \mathbb{R},\ \text{Problem (Cauchy) has a solution defined on } [t_0; t]\}.$$

For any $t \in J$, let v_t be a solution of Problem (Cauchy) defined on $[t_0; t]$[1] and define

$$u(t) = v_t(t).$$

This defines a function $u : J \to U$.

First, let's show that u is a solution of Problem (Cauchy). Its domain J is an interval: for any $t, t' \in J$ and any $t'' \in [t; t']$, we have that either $[t_0; t]$ or $[t_0; t']$ (or both) contains $[t_0; t'']$. Thus, the restriction of v_t or $v_{t'}$ to $[t_0; t'']$ is well-defined and it is a solution of (Cauchy). Therefore, $t'' \in I$.

[1] We denote the interval "$[t_0; t]$" for simplicity, but of course, if $t < t_0$, we actually consider the interval "$[t; t_0]$".

The function u satisfies the initial condition: $u(t_0) = v_{t_0}(t_0)$, and since v_{t_0} is a solution of the problem, we have $v_{t_0}(t_0) = u_0$, hence

$$u(t_0) = u_0.$$

We then show that for any $t \in J$, u is differentiable at t and satisfies the equation

$$u'(t) = f(t, u(t)). \tag{4.3}$$

Let's fix any $t \in J$ arbitrarily. To simplify notation, let's assume $t > t_0$ (we can do the exact same reasoning if $t < t_0$ and a very similar one if $t = t_0$) and distinguish two cases.

- First case: $t < \sup J$. In this case, let $t' \in]t; \sup J[$. The function u coincides with $v_{t'}$ on $[t_0; t']$. Indeed, for any $t'' \in [t_0; t']$, according to Proposition 4.5,

  $$v_{t'} = v_{t''} \quad \text{on } [t; t'] \cap [t; t''] = [t; t''].$$

  So $u(t'') = v_{t''}(t'') = v_{t'}(t'')$.

  Since $v_{t'}$ is differentiable and a solution of the Cauchy problem, the equality $u = v_{t'}$ on $[t_0; t']$ implies that u is also differentiable on $]t_0; t'[$, in particular, differentiable at t, and satisfies Equation (4.3).

- Second case: $t = \sup J$. In this case, J is of the form $[\alpha; t]$ or $]\alpha; t]$, for some $\alpha \in [-\infty; t_0]$.

  Following the same reasoning as in the first case, we see that u coincides with v_t on $[t_0; t]$. This implies that u is differentiable on $]t_0; t]$, which is a neighborhood of t in J, and that Equation (4.3) is satisfied.

This ends the proof that u is a solution of Problem (Cauchy).

Finally, let's show that this solution is maximal. Let $\tilde{u} : \tilde{J} \to U$ be a solution extending u (i.e., $J \subset \tilde{J}$ and $\tilde{u}_{|J} = u$). For any $t \in \tilde{J}$, $\tilde{u}_{|[t_0;t]}$ is a solution of Problem (Cauchy), so t belongs to J. Hence, $\tilde{J} \subset J$. Therefore, $\tilde{J} = J$ and $\tilde{u} = u$. □

Proof of Proposition 4.5. Let $u_1 : J_1 \to U$ and $u_2 : J_2 \to U$ be two solutions of Problem (Cauchy). Let

$$H = \{t \in J_1 \cap J_2 \text{ such that } u_1(t) = u_2(t)\}.$$

The set H is non-empty (it contains t_0) and closed in $J_1 \cap J_2$ (because u_1 and u_2 are continuous). If we manage to show that it is open in $J_1 \cap J_2$, then $H = J_1 \cap J_2$ (as $J_1 \cap J_2$ is an intersection of intervals, hence a connected set) and therefore that

$$u_1 = u_2 \text{ on } H = J_1 \cap J_2.$$

Let's show that it is open. Take any $t_1 \in H$. Consider the modified Cauchy problem.

$$\begin{cases} u' &= f(t, u), \\ u(t_1) &= u_1(t_1). \end{cases} \tag{Cauchy t_1}$$

Both u_1 and u_2 are solutions of this problem since they are solutions of (Cauchy) and $u_1(t_1) = u_2(t_1)$ according to the definition of H.

We can apply the Cauchy-Lipschitz theorem to (Cauchy t_1): f is continuous and Lipschitz with respect to its second variable in a neighborhood of $(t_1, u_1(t_1))$. According to the local uniqueness result of this theorem, there exists $\epsilon > 0$ such that

$$u_1 = u_2 \quad \text{on } J_1 \cap J_2 \cap [t_1 - \epsilon; t_1 + \epsilon].$$

This implies that $J_1 \cap J_2 \cap [t_1 - \epsilon; t_1 + \epsilon] \subset H$ and thus that H contains a neighborhood of t_1 in $J_1 \cap J_2$. This shows that H is open in $J_1 \cap J_2$.

To conclude, let $u : J_1 \cup J_2 \to U$ be the function which coincides with u_1 on J_1 and u_2 on J_2. Let's verify that it is a solution of Problem (Cauchy).

It satisfies the condition $u(t_0) = u_0$ (because u_1 and u_2 satisfy it). Let's show that it is differentiable and satisfies the equation

$$u' = f(t, u). \tag{4.4}$$

We verify (using basic properties of intervals) that $(J_1 \cup J_2) \cap [t_0; +\infty[$ is included in J_1 or J_2. Therefore, u is differentiable on this interval (it coincides with u_1 or u_2, which is differentiable) and satisfies Equation (4.4) (because u_1 and u_2 satisfy it). The same holds on $(J_1 \cup J_2) \cap]-\infty; t_0]$. This implies that u is differentiable and satisfies (4.4) on $(J_1 \cup J_2) \setminus t_0$. Moreover, it has left and right derivatives at t_0, which also satisfy (4.4). Due to this equality, the left and right derivatives coincide (they are equal to $f(t_0, u_0)$) so u is differentiable at t_0 and satisfies (4.4) at this point as well.

□

4.3 Maximal solutions leave compact sets

In this section, we consider a Cauchy problem and assume that f is continuous and Lipschitz with respect to its second variable in the vicinity of every point. This allows us to apply the results from the previous section: there exists a unique maximal solution $u : J \to U$.

Proposition 4.6

The definition set J of the maximal solution u is an open interval in $\mathbb{R}$.

Proof. We know that J is an interval. We must show that it is open.

Let $T \in J$ be arbitrary. According to the Cauchy-Lipschitz theorem, the Cauchy problem

$$\begin{aligned} v' &= f(t, v), \\ v(T) &= u(T) \end{aligned}$$

has a solution v defined on an interval whose interior contains T. Let H be this interval.

According to Proposition 4.5, since both v and u are solutions to this Cauchy problem, the function $w : J \cup H \to U$ which coincides with u on J and v on H is also a solution. This function w is also a solution to the original problem (Cauchy) (since $w(t_0) = u(t_0) = u_0$).

Since u is a maximal solution, we must have $J \cup H \subset J$, which means $H \subset J$. Thus, J contains a neighborhood of T.

This is true for any $T \in J$, so J is open. □

An important question regarding the maximal solution is to determine its definition set. In particular, is this maximal solution global, i.e., is it defined on the same interval I as the function f? The following theorem provides a criterion which, in some cases, answers this question.[2]

Theorem 4.7 : théorème des bouts

We still assume that $f : I \times U \to \mathbb{R}^n$ is continuous and Lipschitz with respect to its second variable in the neighborhood of every point. We still denote $u : J \to U$ the maximal solution of Problem (Cauchy). One of the following two properties is necessarily true.

1. $\sup J = \sup I$;
2. u "leaves any compact set of U" in the neighborhood of $\sup J$: for any compact $K \subset U$, there exists $\eta < \sup J$ such that, for any $t \in]\eta; \sup J[$,

$$u(t) \in U \setminus K.$$

A similar result holds for $\inf J$.

Proof. Let's proceed by contradiction and assume that both properties are false. In particular, $\sup J < \sup I$, so $\sup J \in I$. Let $K \subset U$ be a compact set which u does not leave: for any $\eta < \sup J$, there exists $t \in]\eta; \sup J[$ such that $u(t) \in K$.

[2] As it does not seem to have a well-established name in English, we will stick to the French terminology, « théorème des bouts ».

Then, there exists (and we fix one for the rest of the proof) a sequence $(t_n)_{n\in\mathbb{N}}$ of elements of J such that

$$t_n \overset{n\to+\infty}{\longrightarrow} \sup J;\ u(t_n) \in K, \quad \forall n \in \mathbb{N}.$$

Since K is compact, we can assume, replacing t with a subsequence if necessary, that $(u(t_n))_{n\in\mathbb{N}}$ converges to some $u_{\lim} \in K$.

The proof will be in two steps:

1. we show that $u(t) \to u_{\lim}$ as $t \to \sup J$;
2. we deduce that u can be extended to a solution of Problem (Cauchy) defined on $J \cup \{\sup J\}$, which contradicts the maximality of u.

First step: since f is continuous, it is bounded in a neighborhood of $(u_{\lim}, \sup J)$. So, let $M \in \mathbb{R}$ and $\epsilon > 0$ be such that

$$\forall (t,v) \in]\sup J - \epsilon; \sup J + \epsilon[\times B(u_{\lim}, \epsilon), \quad ||f(t,v)||_2 \leq M.$$

Intuitively, this inequality implies that if, for some n, t_n is close to $\sup J$ and $u(t_n)$ is close to $u_{\lim}$, then $u' = f(t,u)$ is bounded by M close to t_n; in particular, $||u(t) - u(t_n)||_2 \leq M|t - t_n|$ for any t in a neighborhood of t_n whose size we can estimate. This is formalized by the following proposition (the proof of which is given at the end of the theorem's proof).

Proposition 4.8

Let n be any integer such that

$$|t_n - \sup J| < \frac{\epsilon}{2} \quad \text{and} \quad ||u(t_n) - u_{\lim}||_2 < \frac{\epsilon}{2}. \tag{4.5}$$

For any $t \in \left]t_n - \frac{\epsilon}{2\max(M,1)}; t_n + \frac{\epsilon}{2\max(M,1)}\right[\cap J$,

$$||u(t) - u(t_n)||_2 \leq M|t - t_n|.$$

Since $(t_n, u(t_n)) \overset{n\to+\infty}{\to} (\sup J, u_{\lim})$, for any n large enough, we have

$$|t_n - \sup J| < \frac{\epsilon}{2\max(M,1)} \quad \text{and} \quad ||u(t_n) - u_{\lim}||_2 < \frac{\epsilon}{2}.$$

For such values of n, the hypothesis (4.5) is satisfied, thus

$$||u(t) - u(t_n)||_2 \leq M|t - t_n|, \quad \forall t \in \left]t_n - \frac{\epsilon}{2\max(M,1)}; t_n + \frac{\epsilon}{2\max(M,1)}\right[\cap J.$$

Since $t_n + \frac{\epsilon}{2\max(M,1)} > \sup J$, this implies that, for any $t \in [t_n; \sup J[$,

$$\begin{aligned}
||u(t) - u_{\lim}||_2 &\leq ||u(t) - u(t_n)||_2 + ||u(t_n) - u_{\lim}||_2 \\
&\leq M|t - t_n| + ||u(t_n) - u_{\lim}||_2 \\
&\leq M|t_n - \sup J| + ||u(t_n) - u_{\lim}||_2 \\
&\to 0 \text{ as } n \to +\infty.
\end{aligned}$$

So $u(t) \to u_{\lim}$ as $t \to \sup J$.

Second step: let's extend u continuously to $J \cup \{\sup J\}$, that is, let's define

$$\begin{array}{rccll}
\bar{u}: & J \cup \sup J & \to & U & \\
 & t & \to & u(t) & \text{if } t < \sup J \\
 & & & u_{\lim} & \text{otherwise.}
\end{array}$$

This is a continuous function. It is differentiable on J and

$$u'(t) = f(t, u(t)) \overset{t\to\sup J}{\longrightarrow} f(\sup J, u_{\lim}),$$

which shows that u is also differentiable at $\sup J$, with derivative $f(\sup J, u_{\lim})$.

Therefore, the function $\bar{u}$ is a solution of Problem (Cauchy), extending u but not equal to u. This contradicts the maximality of u. ☐

Proof of Proposition 4.8. We first show that for any $t \in \left[t_n; t_n + \frac{\epsilon}{2\max(M,1)}\right[\cap J$, $||u(t) - u_{\lim}||_2 < \epsilon$. We can assume that the set

$$\{t \in J, t \geq t_n, ||u(t) - u_{\lim}||_2 \geq \epsilon\}$$

is non-empty, otherwise the property is necessarily true. Let's define

$$T = \inf\{t \in J, t \geq t_n, ||u(t) - u_{\lim}||_2 \geq \epsilon\}$$

and show that $T \geq t_n + \frac{\epsilon}{2\max(M,1)}$. Let's assume by contradiction that this is not the case.

By continuity of u, we must have $||u(T) - u_{\lim}||_2 \geq \epsilon$. For all $t \in [t_n; T[$, we have

$$||u(t) - u_{\lim}||_2 < \epsilon$$

and, since $|t - \sup J| \leq |t_n - \sup J| < \epsilon$,

$$\left|\left|u'(t)\right|\right|_2 = ||f(t, u(t))||_2 \leq M.$$

This is also true at $t = T$ due to the continuity of u'. Therefore, u is M-Lipschitz on $[t_n; T]$ and

$$\begin{aligned}
||u(T) - u_{\lim}||_2 &\leq ||u(T) - u(t_n)||_2 + ||u(t_n) - u_{\lim}||_2 \\
&\leq M|T - t_n| + ||u(t_n) - u_{\lim}||_2 \\
&< M\frac{\epsilon}{2\max(M,1)} + \frac{\epsilon}{2} \\
&\leq \epsilon.
\end{aligned}$$

This contradicts the inequality $||u(T) - u_{\lim}||_2 \geq \epsilon$.

We have thus shown that for any $t \in \left[t_n; t_n + \frac{\epsilon}{2\max(M,1)}\right[\cap J$, $||u(t) - u_{\lim}||_2 < \epsilon$. Similarly, we can show that for any $t \in \left]t_n - \frac{\epsilon}{2\max(M,1)}; t_n\right[$, $||u(t) - u_{\lim}||_2 < \epsilon$.

Consequently, for any $t \in \left]t_n - \frac{\epsilon}{2\max(M,1)}; t_n + \frac{\epsilon}{2\max(M,1)}\right[\cap J$,

$$\left|\left|u'(t)\right|\right|_2 = ||f(t, u(t))||_2 \leq M.$$

This implies that u is M-Lipschitz on the considered interval. In particular, for all $t \in \left]t_n - \frac{\epsilon}{2\max(M,1)}; t_n + \frac{\epsilon}{2\max(M,1)}\right[\cap J$,

$$||u(t) - u(t_n)||_2 \leq M|t - t_n|.$$

☐

The following example shows how the théorème des bouts allows to prove that a maximal solution of a differential equation is global.

Example 4.9

Consider the problem (Cauchy), for a function $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Suppose that f is continuous, Lipschitz with respect to its second variable in the neighborhood of every point, and satisfies the inequality

$$||f(t, u)||_2 \leq ||u||_2, \quad \forall (t, u) \in \mathbb{R} \times \mathbb{R}^n. \tag{4.6}$$

Its maximal solution is global (i.e. defined on $\mathbb{R}$).

Proof. Let $u : J \to \mathbb{R}^n$ be this maximal solution and show that $J = \mathbb{R}$. We will prove that $\sup J = +\infty$; a similar reasoning shows that $\inf J = -\infty$.

Let's proceed by contradiction and assume that $\sup J < +\infty$. According to the théorème des bouts, u leaves any compact set in the neighborhood of $\sup J$. We will obtain a contradiction by showing that u is actually bounded in the neighborhood of $\sup J$.

Consider the function $N : t \in J \to ||u(t)||_2^2 \in \mathbb{R}$. It is differentiable and, for all $t \in J$:

$$\begin{aligned} |N'(t)| &= \left|2\left\langle u(t), u'(t)\right\rangle\right| \\ &= 2\left|\langle u(t), f(t, u(t))\rangle\right| \\ &\leq 2||u(t)||_2||f(t, u(t))||_2 \\ &\leq 2||u(t)||_2^2 \\ &= 2N(t). \end{aligned}$$

From this point on, it is possible to show that N (hence u) is bounded by using Gronwall's lemma (Lemma D.1 in the appendix). In the next lines, we propose an argument which does not explicitely invoke this lemma, but reaches the same conclusion.

We define $N_2 : t \in J \to N(t)e^{-2t}$. For all t,

$$N_2'(t) = (N'(t) - 2N(t))e^{-2t} \leq 0,$$

thus N_2 is decreasing and, for all $t \in]t_0; \sup J[$, $N_2(t) \leq N_2(t_0) = ||u_0||_2^2 e^{-2t_0}$, which implies

$$N(t) \leq \left(||u_0||_2 e^{t-t_0}\right)^2.$$

Consequently, for all $t \in]t_0; \sup J[$,

$$||u(t)||_2 \leq ||u_0||_2 e^{t-t_0} \leq ||u_0||_2 e^{\sup J - t_0}.$$

If we set $M = ||u_0||_2 e^{\sup J - t_0}$, we obtain that u does not leave the compact set $\bar{B}(0, M)$. We have reached a contradiction. □

The result stated in the example remains valid if we replace the bound (4.6) by a more general linear upper bound

$$||f(t, u)||_2 \leq C_1||u||_2 + C_2, \quad \forall (t, u) \in \mathbb{R} \times \mathbb{R}^n,$$

for constants $C_1, C_2 > 0$.

However, it is no longer valid if we replace the bound "$||u||_2$" with "$||u||_2^\alpha$" for a power $\alpha > 1$. To convince ourselves of this, we can consider the following Cauchy problem, where the unknown u takes values in $\mathbb{R}$:

$$\begin{aligned} u' &= |u|^\alpha, \\ u(0) &= 1. \end{aligned}$$

We can check that its maximal solution is

$$u : \begin{array}{ccc} \left]-\infty; \frac{1}{\alpha-1}\right[& \to & \mathbb{R} \\ t & \to & \frac{1}{(1-(\alpha-1)t)^{\frac{1}{\alpha-1}}}, \end{array}$$

which is not defined on $\mathbb{R}$ as a whole.

Exercise 7

Let $f : \mathbb{R} \to \mathbb{R}$ be a C^1 map such that

$$\begin{aligned} f(0) &= 0; \\ f(t) &\geq t^2, \quad \forall t \in \mathbb{R}. \end{aligned}$$

Fo fixed $t_0, u_0 \in \mathbb{R}$, we consider the Cauchy problem

$$\begin{cases} u'(t) &= f(u(t)), \\ u(t_0) &= u_0. \end{cases}$$

1. Show that this problem has a unique maximal solution.

Let J be the domain of this maximal solution, and u be the solution.

2. a) Show that, if $u_0 = 0$, then $J = \mathbb{R}$ and $u(t) = 0, \forall t \in \mathbb{R}$.
   b) Show that, for any $t_1 \in J$, u is solution of the Cauchy problem, where the initial condition (t_0, u_0) is replaced with $(t_1, u(t_1))$.
   c) Deduce that, if $u(t_1) = 0$ for some $t_1 \in J$, then $J = \mathbb{R}$ and $u(t) = 0, \forall t \in \mathbb{R}$.

Let us now assume that $u_0 > 0$.

3. a) Show that, for all $t \in]-\infty; t_0] \cap J$, $u(t) \in]0; u_0]$.
   b) Deduce from the previous question that $]-\infty; t_0] \subset J$.
   c) Show that $u(t) \to 0$ when $t \to -\infty$.
4. a) Show that $-\frac{1}{u}$ is well-defined and negative over J.
   b) Show that, for all $t \in [t_0; +\infty[\cap J$,

$$-\frac{1}{u(t)} \geq -\frac{1}{u(t_0)} + (t - t_0).$$

   c) Show that $\sup J < +\infty$.
   d) Show that $u(t) \to +\infty$ when $t \to \sup J$.

4.4 Regularity in the initial condition

The content of this section is crucial for the theory of differential equations. We will need the results it contains in a later chapter. However, by lack of time, it will not be covered in class.

In this section, we look at the pair (t_0, u_0), which is the initial condition of Problem (Cauchy), and let it vary. This defines a family of solutions to the differential equation "$u' = f(t, u)$". When f is C^2, this family of solutions is differentiable with respect to (t_0, u_0). Furthermore, its partial derivatives can be described as solutions to another Cauchy problem.

To simplify notation, we first state this result in the case where t_0 is fixed and only u_0 varies. The general case is given afterwards.

Theorem 4.10 : regularity in the initial condition

Let I be a non-empty open interval of $\mathbb{R}$, U an open set in $\mathbb{R}^n$, and $f : I \times U \to \mathbb{R}^n$ be a function. We assume that f is C^2.

Let us fix $t_0 \in I$. For every $u_0 \in U$, let $u_{u_0} : J_{u_0} \to U$ be the maximal solution of the Cauchy problem

$$\begin{cases} u'_{u_0} &= f(t, u_{u_0}), \\ u_{u_0}(t_0) &= u_0. \end{cases} \qquad \text{(Cauchy } u_0\text{)}$$

The set $\Omega = \{(u_0, t), u_0 \in U, t \in J_{u_0}\}$ is an open subset of $U \times I$ and the map

$$\begin{array}{cccc} V: & \Omega & \to & U \\ & (u_0, t) & \to & u_{u_0}(t) \end{array}$$

is C^1.

Moreover, for every u_0, $\frac{dV}{du_0}(u_0, .) : J_{u_0} \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is a solution of the following Cauchy problem:

$$\begin{cases} \frac{d}{dt}\left(\frac{dV}{du_0}\right) &= \frac{df}{du}(t, V(u_0, t)) \circ \frac{dV}{du_0}(u_0, t), \\ \frac{dV}{du_0}(u_0, t_0) &= \mathrm{Id}_{\mathbb{R}^n}. \end{cases} \qquad \text{(Cauchy } \tfrac{dV}{du_0}\text{)}$$

Remark

It is not necessary to memorize by heart Problem (Cauchy $\frac{dV}{du_0}$) for which $\frac{dV}{du_0}$ is a solution. It suffices to remember that V is C^1. Then, (Cauchy $\frac{dV}{du_0}$) can be obtained by differentiating (Cauchy u_0). Indeed, (Cauchy u_0) can be rewritten in terms of V as

$$\begin{cases} \frac{dV}{dt}(u_0,t) & = f(t, V(u_0,t)), \\ V(u_0,t) & = u_0. \end{cases}$$

Differentiating with respect to u_0 both sides of each of the two equalities yields exactly (Cauchy $\frac{dV}{du_0}$).

Proof of Theorem 4.10. To simplify a bit, let's assume that f does not depend on t. We can make this assumption thanks to the lemma that follows (the proof of which is found in the appendix D.2). We thus denote "$f(u)$" instead of "$f(t,u)$", and use interchangeably the notation "$\frac{df}{du}$" of "df" for the differential.

Lemma 4.11

If the theorem holds for all maps f independent of t, it holds for all maps f.

The following lemma further simplifies the problem by showing that it suffices to establish the regularity of V in a neighborhood of each u_0, for times t close to t_0. It is proven in the appendix D.3.

Lemma 4.12

Suppose that

for every $u_0 \in U$, Ω contains a neighborhood of (u_0,t_0), on which V is C^1 and satisfies the equations (Cauchy $\frac{dV}{du_0}$). (4.7)

Then Ω is open, V is C^1 on Ω and satisfies the equations (Cauchy $\frac{dV}{du_0}$).

It remains to show that Property (4.7) is true. Let $u_0 \in U$.

First step: V is defined in a neighborhood of (u_0,t_0).

Let $M_1, \epsilon > 0$ be such that $\overline{B}(u_0,\epsilon) \subset U$ and

$$\forall v \in B(u_0,\epsilon), \quad ||f(v)||_2 \leq M_1.$$

The following proposition, proven in the appendix D.4, shows that Ω contains $B\left(u_0,\frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$.

Proposition 4.13

For every $v \in B\left(u_0,\frac{\epsilon}{2}\right)$,

$$\left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[\subset J_v.$$

Furthermore, for every $t \in \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$,

$$u_v(t) \in B(u_0,\epsilon).$$

Second step: V is Lipschitz on this neighborhood.

For all $v,t \in B\left(u_0,\frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$,

$$u_v'(t) = f(u_v(t)) \quad \Rightarrow \quad ||u_v'(t)||_2 \leq M_1.$$

Therefore, for all $v \in B\left(u_0,\frac{\epsilon}{2}\right)$, u_v is M_1-Lipschitz on $\left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$, meaning that V is M_1-Lipschitz with respect to its second variable.

Let $M_2 > 0$ be such that

$$\forall v \in \bar{B}(u_0, \epsilon), \quad ||df(v)||_{\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)} \leq M_2.$$

(Recall that f is C^2. In particular, its differential is continuous on U, hence bounded on $\bar{B}(u_0, \epsilon)$.)

The function f is M_2-Lipschitz on $B(u_0, \epsilon)$ by the mean value inequality. Thus, for all $v_1, v_2 \in B(u_0, \frac{\epsilon}{2}), t \in \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$,

$$\begin{aligned} ||u'_{v_1}(t) - u'_{v_2}(t)||_2 &= ||f(u_{v_1}(t)) - f(u_{v_2}(t))||_2 \\ &\leq M_2 ||u_{v_1}(t) - u_{v_2}(t)||_2. \end{aligned}$$

We integrate and use the triangular inequality: for all $t \in \left[t_0; t_0 + \frac{\epsilon}{2M_1}\right[$,

$$\begin{aligned} ||u_{v_1}(t) - u_{v_2}(t)||_2 &= \left|\left| u_{v_1}(t_0) - u_{v_2}(t_0) + \int_{t_0}^{t} \left(u'_{v_1}(s) - u'_{v_2}(s)\right) ds \right|\right|_2 \\ &\leq ||u_{v_1}(t_0) - u_{v_2}(t_0)||_2 + \int_{t_0}^{t} ||u'_{v_1}(s) - u'_{v_2}(s)||_2 ds \\ &\leq ||u_{v_1}(t_0) - u_{v_2}(t_0)||_2 + \int_{t_0}^{t} M_2 ||u_{v_1}(s) - u_{v_2}(s)||_2 ds. \end{aligned}$$

Thus, according to Gronwall's lemma (Lemma D.1 in the appendix), for all $t \in \left[t_0; t_0 + \frac{\epsilon}{2M_1}\right[$,

$$\begin{aligned} ||u_{v_1}(t) - u_{v_2}(t)||_2 &\leq ||u_{v_1}(t_0) - u_{v_2}(t_0)||_2 e^{M_2(t - t_0)} \\ &= ||v_1 - v_2||_2 e^{M_2(t - t_0)} \\ &\leq ||v_1 - v_2||_2 e^{\frac{\epsilon M_2}{2M_1}}. \end{aligned}$$

Symmetrically, the inequality is also valid for $t \in \left]t_0 - \frac{\epsilon}{2M_1}; t_0\right]$, which shows that V is $e^{\frac{\epsilon M_2}{2M_1}}$-Lipschitz with respect to its first variable on $B\left(u_0, \frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$. Hence, V is globally Lipschitz (and therefore continuous) on this open set.

<u>Third step: differentiability of V with respect to t.</u>

According to its definition, V is differentiable with respect to its second variable, and for all v, t,

$$\frac{dV}{dt}(v, t) = u'_v(t) = f(V(v, t)).$$

Since f is continuous on U and V is continuous on $B\left(u_0, \frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[$, the function $\frac{dV}{dt}$ is also continuous on this latter set.

<u>Fourth step: differentiability of V with respect to u_0</u>

Let's show that V has a partial derivative with respect to its first variable, which is continuous and satisfies the Problem (Cauchy $\frac{dV}{du_0}$). We will proceed "backwards": we consider the solution to Problem (Cauchy $\frac{dV}{du_0}$) and show that it is continuous and is the partial derivative of V with respect to u_0. For any $v \in B\left(u_0, \frac{\epsilon}{2}\right)$, let $w_v : \tilde{I}_v \subset \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[\to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ be the maximal solution to the problem

$$\begin{aligned} w'_v(t) &= \frac{df}{du}(V(v, t)) \circ w_v(t) \\ w_v(t_0) &= \mathrm{Id}_{\mathbb{R}^n}. \end{aligned}$$

The maximal solution exists and is unique because, for any v, the map

$$(t, x) \in \left]t_0 - \frac{\epsilon}{2M_1}; t_0 + \frac{\epsilon}{2M_1}\right[\times \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \quad \to \quad \frac{df}{du}(V(v, t)) \circ x$$

is M_2-Lipschitz with respect to x, hence Cauchy-Lipschitz theorem applies.

The same reasoning as we did for u_v in the second step shows that there exists a constant $M_3 \geq M_1$ such that, for any $v \in B\left(u_0, \frac{\epsilon}{2}\right)$, the domain of definition of w_v contains

$$\left]t_0 - \frac{\epsilon}{2M_3}; t_0 + \frac{\epsilon}{2M_3}\right[$$

and the mapping $(v,t) \to w_v(t)$ is Lipschitz and therefore continuous on $B\left(u_0, \frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_3}; t_0 + \frac{\epsilon}{2M_3}\right[$ (this is the point of the proof that uses the hypothesis that f is C^2).

Finally, let's show that V is differentiable with respect to its first variable, and, for all $v, t \in B\left(u_0, \frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_3}; t_0 + \frac{\epsilon}{2M_3}\right[$,

$$\frac{dV}{du_0}(v,t) = w_v(t).$$

To do this, we will perform a kind of first-order Taylor expansion of Problem (Cauchy u_0) in u_0.

Let $v, h \in \mathbb{R}^n$ be such that $v, v+h \in B\left(u_0, \frac{\epsilon}{2}\right)$. Consider the function

$$\Delta : t \in \left]t_0 - \frac{\epsilon}{2M_3}; t_0 + \frac{\epsilon}{2M_3}\right[\to u_{v+h}(t) - u_v(t) - w_v(t)(h).$$

We have

$$\Delta(t_0) = (v+h) - v - \mathrm{Id}_{\mathbb{R}^n}(h) = 0.$$

Moreover, for any t,

$$\begin{aligned}
\Delta'(t) &= u'_{v+h}(t) - u'_v(t) - w'_v(t)(h) \\
&= f(u_{v+h}(t)) - f(u_v(t)) - \frac{df}{du}(u_v(t)) \circ w_v(t)(h) \\
&= \frac{df}{du}(u_v(t))(u_{v+h}(t) - u_v(t)) - \frac{df}{du}(u_v(t)) \circ w_v(t)(h) + E(t) \\
&= \frac{df}{du}(u_v(t))(\Delta(t)) + E(t)
\end{aligned}$$

with $E(t) = f(u_{v+h}(t)) - f(u_v(t)) - \frac{df}{du}(u_v(t))(u_{v+h}(t) - u_v(t))$ and thus, by one of the Taylor inequalities,

$$||E(t)||_2 \leq \frac{1}{2}\left(\sup_{\tilde{v} \in \bar{B}(u_0,\epsilon)} \left|\left|\frac{d^2 f}{du^2}(\tilde{v})\right|\right|_{\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n))}\right) ||u_{v+h}(t) - u_v(t)||_2^2.$$

Let $C_1 = \frac{1}{2}\sup_{\tilde{v} \in \bar{B}(u_0,\epsilon)} \left|\left|\frac{d^2 f}{du^2}(\tilde{v})\right|\right|_{\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n))}$ and C_2 be the Lipschitz constant of V with respect to its first variable (whose existence we proved a few paragraphs ago). With these notations, for any t,

$$||E(t)||_2 \leq C_1 C_2 ||h||_2^2$$

and thus

$$\left|\left|\Delta'(t) - \frac{df}{du}(u_v(t))(\Delta(t))\right|\right| 2 \leq C_1 C_2 ||h||_2^2.$$

Denoting $C_3 = \sup_{\tilde{v} \in \bar{B}(u_0,\epsilon)} \left|\left|\frac{df}{du}(\tilde{v})\right|\right|_{\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)}$, we deduce

$$\left|\left|\Delta'(t)\right|\right|_2 \leq C_1 C_2 ||h||_2^2 + C_3 ||\Delta(t)||_2.$$

Therefore, for any $t \in \left[t_0; t_0 + \frac{\epsilon}{2M_3}\right[$,

$$||\Delta(t)||_2 = \left|\left|\Delta(t_0) + \int_{t_0}^{t} \Delta'(s) ds\right|\right|_2$$

$$\begin{aligned}&= \left|\left| \int_{t_0}^{t} \Delta'(s) ds \right|\right|_2 \\ &\leq \int_{t_0}^{t} \left|\left|\Delta'(s)\right|\right|_2 ds \\ &\leq \int_{t_0}^{t} \left(C_1 C_2 ||h||_2^2 + C_3 \left|\left|\Delta(s)\right|\right|_2\right) ds \\ &= C_1 C_2 ||h||_2^2 (t-t_0) + \int_{t_0}^{t} C_3 \left|\left|\Delta(s)\right|\right|_2 ds.\end{aligned}$$

From Gronwall's lemma, for any $t \in \left[t_0; t_0 + \frac{\epsilon}{2M_3}\right[$,

$$\begin{aligned}||\Delta(t)||_2 &\leq C_1 C_2 ||h||_2^2 (t-t_0) + C_1 C_2 C_3 ||h||_2^2 \int_{t_0}^{t} e^{C_3(t-s)}(s-t_0) ds \\ &= \frac{C_1 C_2}{C_3} ||h||_2^2 \left(e^{C_3(t-t_0)} - 1\right).\end{aligned}$$

Symmetrically, the inequality is also valid if $t \in \left]t_0 - \frac{\epsilon}{2M_3}; t_0\right]$, provided that we replace "$e^{C_3(t-t_0)}$" with "$e^{C_3|t-t_0|}$" on the right-hand side.

If we set $C_4 = \frac{C_1 C_2}{C_3}\left(e^{\frac{C_3 \epsilon}{2M_3}} - 1\right)$, we have thus shown that, for any v, h such that $v, v+h \in B\left(u_0, \frac{\epsilon}{2}\right)$ and for any $t \in \left]t_0 - \frac{\epsilon}{2M_3}; t_0 + \frac{\epsilon}{2M_3}\right[$,

$$||V(v+h,t) - V(v,t) - w_v(t)(h)||_2 = ||\Delta(t)||_2 \leq C_4 ||h||_2^2.$$

Therefore, V is differentiable with respect to its first variable, and for any v, t in the considered open set,

$$\frac{dV}{du_0}(v,t) = w_v(t).$$

Conclusion.

We have seen that V is continuous on $B\left(u_0, \frac{\epsilon}{2}\right) \times \left]t_0 - \frac{\epsilon}{2M_3}; t_0 + \frac{\epsilon}{2M_3}\right[$, has partial derivatives with respect to each of its two variables on this open set, and that these partial derivatives are continuous. Therefore, V is C^1 on this open set. In the fourth step, we have also shown that the partial derivative $\frac{dV}{du_0}$ is a solution of Problem (Cauchy $\frac{dV}{du_0}$). Hence, Property (4.7) is true.

□

Theorem 4.14 : Regularity, general case

We keep the notation from the previous theorem; f is still C^2.
For any pair $(t_0, u_0) \in I \times U$, let $u_{t_0,u_0} : J_{t_0,u_0} \to U$ be the maximal solution of the Cauchy problem

$$\begin{cases} u'_{t_0,u_0} &= f(t, u_{t_0,u_0}), \\ u_{t_0,u_0}(t_0) &= u_0. \end{cases} \qquad \text{(Cauchy } (t_0,u_0))$$

The set $\Omega = \{(t_0, u_0, t), t_0 \in I, u_0 \in U, t \in J_{t_0,u_0}\} \subset I \times U \times I$ is open and the map

$$V : \begin{array}{ccc} \Omega & \to & U \\ (t_0, u_0, t) & \to & u_{t_0,u_0}(t) \end{array}$$

is of class C^1.
Moreover, the partial derivatives of V are solutions of the following Cauchy problems:

$$\frac{d}{dt}\left(\frac{dV}{du_0}\right) = \frac{df}{du}(t, V(t_0, u_0, t)) \circ \frac{dV}{du_0}(t_0, u_0, t),$$

$$\frac{dV}{du_0}(t_0, u_0, t_0) = \mathrm{Id}_{\mathbb{R}^n}.$$

$$\frac{d}{dt}\left(\frac{dV}{dt_0}\right) = \frac{df}{du}(t, V(t_0, u_0, t))\left(\frac{dV}{dt_0}(t_0, u_0, t)\right),$$
$$\frac{dV}{dt_0}(t_0, u_0, t_0) = -f(t_0, u_0).$$

This theorem can be derived from the previous one as in the proof of Lemma 4.11.

Remark

An even more general theorem holds: we can assume that f is a function of three variables instead of two, yielding a Cauchy problem of the form

$$u' = f(t, u, a),$$
$$u(t_0) = u_0.$$

If f is C^2, the maximal solutions of this problem are C^1 in (t_0, u_0, a).

Chapter 5

Explicit solutions in particular situations

What you should know or be able to do after this chapter

- Solve an autonomous scalar equation.
- Solve a linear scalar equation.
- Identify a linear equation.
- Know that the solution of a linear differential equation is global.
- Admitting that the resolvent of a linear equation is C^1, write the Cauchy problem of which it is a solution.
- Use this Cauchy problem to show that a given map is the resolvent of a Cauchy problem.
- Remember that, for all t_1, t_2, t_3, $R(t_3, t_2)R(t_2, t_1) = R(t_3, t_1)$ and that, for all t_1, t_2, $R(t_2, t_1)^{-1} = R(t_1, t_2)$.
- Write the solution(s) of a linear equation in terms of the resolvent (with or without source term, with or without an initial condition).
- Recall (= be able to find it again by yourself) the explicit expression of the resolvent when the equation has constant coefficients.
- Compute the exponential of a matrix when the Dunford decomposition is given.

5.1 Autonomous scalar equations

In this section, we consider a *scalar* equation (the images of u are in $U \subset \mathbb{R}$ and not in $\mathbb{R}^n$ for some $n > 1$) and *autonomous* (the map f does not depend on time). Thus, we have an equality of the form

$$u' = f(u), \tag{5.1}$$

for some $f : U \to \mathbb{R}$, with U a non-empty open subset of $\mathbb{R}$. Throughout this section, we assume that f is locally Lipschitz, so that the Cauchy-Lipschitz theorem applies. We will describe the maximal solutions of Equation (5.1).

Let's start with the simplest solutions: the constants.

Proposition 5.1

We assume that f is locally Lipschitz.
For any $u_0 \in U$, the constant function $u : t \in \mathbb{R} \to u_0$ is a maximal solution of the differential equation (5.1) if and only if $f(u_0) = 0$.

Proof. Let $u_0 \in U$. Let $u : t \in \mathbb{R} \to u_0$. Its derivative is zero. Thus, it is a solution of the differential equation (5.1) if and only if

$$0 = f(u_0).$$

When it is, it is a *maximal* solution as it is defined on $\mathbb{R}$ and can thus not be extended. □

Now, let's describe the non-constant solutions, using the primitives of $\frac{1}{f}$. Consider $u : J \to \mathbb{R}$ a maximal solution whose derivative is not identically zero. Let $t_0 \in J$ be such that $u'(t_0) \neq 0$. For simplicity, assume $f(u(t_0)) = u'(t_0) > 0$; a very similar reasoning is possible if $f(u(t_0)) < 0$.

Let $]\alpha;\beta[$ be the maximal interval containing $u(t_0)$ on which f is strictly positive (with possibly $\alpha = -\infty$ and $\beta = +\infty$).

Proposition 5.2

For any $t \in J$, $u(t) \in]\alpha;\beta[$.

Proof. Let's argue by contradiction and assume it is not true. Since $u(t_0) \in]\alpha;\beta[$, the continuity of u and the intermediate value theorem imply that there exists $t_1 \in J$ such that $u(t_1) = \alpha$ or $u(t_1) = \beta$. Let us for instance assume $u(t_1) = \alpha$.

Then u is a solution of the following Cauchy problem:

$$\begin{cases} u' &= f(u), \\ u(t_1) &= \alpha. \end{cases}$$

The constant function $\tilde{u} : t \in \mathbb{R} \to \alpha$ is a maximal solution of this problem (indeed, $f(\alpha) = 0$, because $]\alpha;\beta[$ is a maximal interval on which f is strictly positive). Since the maximal solution of the problem is unique, as f is locally Lipschitz, $u = \tilde{u}$, which means u is constant. This is a contradiction. □

Let $\Phi :]\alpha;\beta[\to \mathbb{R}$ be a primitive of $\frac{1}{f}$: for any arbitrary constant C, we define

$$\Phi(v) = C + \int_{u(t_0)}^{v} \frac{1}{f(s)} ds, \quad \forall v \in]\alpha;\beta[.$$

This is a continuous function with strictly positive derivative. Hence, it induces a diffeomorphism onto its image, which is an open interval, denoted $]\gamma;\delta[$.

We observe that, for any $t \in J$,

$$(\Phi \circ u)'(t) = \Phi'(u(t))u'(t) = \frac{u'(t)}{f(u(t))} = 1.$$

Thus, for any $t \in J$,

$$\Phi \circ u(t) = \Phi \circ u(t_0) + (t - t_0) = t - t_0 + C.$$

Therefore, for any $t \in J$, $u(t) = \Phi^{-1}(t - t_0 + C)$.

Proposition 5.3

The interval J is equal to $]\gamma + t_0 - C; \delta + t_0 - C[$.

Proof. For any $t \in J$, since $\phi \circ u(t) = t - t_0 + C$, we must have $t - t_0 + C \in]\gamma;\delta[$, thus $t \in]\gamma + t_0 - C; \delta + t_0 - C[$. This shows that $J \subset]\gamma + t_0 - C; \delta + t_0 - C[$.

As u is a maximal solution, it is defined on the whole $]\gamma + t_0 - C; \delta + t_0 - C[$. Indeed, if it were not the case, the map $\tilde{u} : t \in]\gamma + t_0 - C; \delta + t_0 - C[\to \Phi^{-1}(t - t_0 + C) \in U$ would be a solution of Equation (5.1) that strictly extends it. □

This leads to the following theorem.

Theorem 5.4

The non-constant maximal solutions of Equation (5.1) are all maps of the form

$$t \in]\gamma + D; \delta + D[\quad \to \quad \Phi^{-1}(t - D),$$

where Φ is a primitive of $\frac{1}{f}$, defined on a maximal interval where f does not vanish, $]\gamma;\delta[$ is the image of Φ, and $D \in \mathbb{R}$ is an arbitrary constant.

Proof. The reasoning we just did shows that all non-constant maximal solutions have this form (where D corresponds to the previous $t_0 - C$). Conversely, any map of this form is a solution of Equation (5.1), since, for all t,

$$\begin{aligned}\left(\Phi^{-1}\right)'(t-D) &= \frac{1}{\Phi'\left(\Phi^{-1}(t-D)\right)}\\ &= f\left(\Phi^{-1}(t-D)\right).\end{aligned}$$

It is maximal because, when $t \to \gamma + D$, $\Phi^{-1}(t-D) \to \alpha$ or β, hence $\Phi'\left(\Phi^{-1}(t-D)\right) \to 0$, which means that $\left(\Phi^{-1}\right)'(t-D)$ diverges, hence $\Phi(.-D)$ cannot be extended into a differentiable map in $\gamma + D$. The same reasoning holds for $\delta + D$. □

Example 5.5

Let's find all maximal solutions of the differential equation

$$u' = -u^3.$$

The map $x \to -x^3$ is locally Lipschitz (it is C^1). It vanishes only at 0. Thus, the only constant solution is $u \equiv 0$.
Now let's search for non-constant solutions. The maximal intervals where $x \to -x^3$ does not vanish are $]-\infty;0[$ and $]0;+\infty[$. On these intervals, primitives of $x \to \frac{1}{-x^3}$ are

$$\Phi_1 : x \in]-\infty;0[\to \frac{1}{2x^2}, \quad \Phi_2 : x \in]0;+\infty[\to \frac{1}{2x^2}.$$

The first one is a bijection between $]-\infty;0[$ and $]0;+\infty[$, with inverse

$$\Phi_1^{-1} : x \in]0;+\infty[\to -\frac{1}{\sqrt{2x}} \in]-\infty;0[$$

and the second one is a bijection between $]0;+\infty[$ and $]0;+\infty[$, with inverse

$$\Phi_2^{-1} : x \in]0;+\infty[\to \frac{1}{\sqrt{2x}} \in]0;+\infty[.$$

Thus, maximal solutions are all maps of the form

$$\begin{aligned} &u : t \in]D;+\infty[\to -\frac{1}{\sqrt{2(x-D)}}\\ \text{and} \quad &u : t \in]D;+\infty[\to \frac{1}{\sqrt{2(x-D)}}\end{aligned}$$

for any real number D.

Exercise 8

Let $u_0 \in \mathbb{R}_+^*$ be fixed. Compute the maximal solution of the following Cauchy problem:

$$\begin{cases} u'(t) &= \frac{e^{-u(t)^2}}{2u(t)},\\ u(0) &= u_0.\end{cases}$$

5.2 Scalar linear equations

A scalar linear differential equation is an equation of the form

$$u'(t) = a(t)u(t) + b(t), \tag{5.2}$$

where a, b are continuous maps on an interval $I \subset \mathbb{R}$. The function b is sometimes called the "source term".

Let's first solve this equation in the case where b is zero.

Proposition 5.6 : with no source term

Let $a : I \to \mathbb{R}$ be a continuous map, for some open interval I. Let $A : I \to \mathbb{R}$ be a primitive of a. The maximal solutions of the differential equation

$$u'(t) = a(t)u(t)$$

are all maps of the form $u : t \in I \to Ce^{A(t)}$, where C is an arbitrary real number.

Proof. A map of the form $t \to Ce^{A(t)}$ is necessarily a solution of the equation. It is maximal because it is defined on I.

Conversely, if $u : J \to \mathbb{R}$ is a maximal solution, we define $v : t \in J \to u(t)e^{-A(t)} \in \mathbb{R}$. This map is differentiable and, for any $t \in J$,

$$v'(t) = (u'(t) - A'(t)u(t))e^{-A(t)} = (u'(t) - a(t)u(t))e^{-A(t)} = 0.$$

This means that v is constant. Let us denote C its value. For any $t \in J$, $u(t) = Ce^{A(t)}$. Since u is maximal, we must have $J = I$; hence, the map is of the desired form. □

Now let's consider the general Equation (5.2), without assuming that b is zero. To solve it, we use the method called *variation of constants*[1]. Let's again denote $A : I \to \mathbb{R}$ a primitive of a. For a differentiable map $u : J \to \mathbb{R}$ with J a subinterval of I, we write u in the form

$$u(t) = v(t)e^{A(t)}$$

(by setting $v(t) = u(t)e^{-A(t)}$ for all t).

The map u is a solution of the equation if and only if, for all $t \in J$,

$$\begin{aligned}(v'(t) + a(t)v(t))e^{A(t)} &= u'(t)\\ &= a(t)u(t) + b(t) = a(t)v(t)e^{A(t)} + b(t),\end{aligned}$$

which is equivalent to, for all t,

$$v'(t) = b(t)e^{-A(t)}.$$

We denote B an arbitrary primitive of $t \to b(t)e^{-A(t)}$. The previous equation holds if and only if there exists a real number C such that

$$v = C + B.$$

This is equivalent to the existence of $C \in \mathbb{R}$ such that, for all $t \in J$,

$$u(t) = Ce^{A(t)} + B(t)e^{A(t)}.$$

From this reasoning, we can deduce the following theorem.

Theorem 5.7 : solution of scalar linear equations

For any u_0, the maximal solution of the Cauchy problem

$$\begin{cases} u'(t) &= a(t)u(t) + b(t),\\ u(t_0) &= u_0,\end{cases}$$

[1] "variation de la constante" in French

where a, b are continuous maps on an open interval I and u_0 is a real number, is given by

$$u : t \in I \quad \to \quad u_0 e^{\int_{t_0}^t a(s)ds} + \int_{t_0}^t b(s) e^{\int_s^t a(\tau)d\tau} ds.$$

5.3 Linear equations in general dimension

In this section, we consider a linear differential equation of dimension $n \in \mathbb{N}^*$, that is, an equation of the form

$$u'(t) = A(t)u(t) + b(t), \tag{5.3}$$

where $A \in C^0(I, \mathbb{R}^{n\times n})$ and $b \in C^0(I, \mathbb{R}^n)$, with I an interval of $\mathbb{R}$.

Proposition 5.8

The maximal solutions of Equation (5.3) are global (i.e., defined on the entire interval I).

Proof. The proof relies on the théorème des bouts (Theorem 4.7); it is very similar to that of Example 4.9.

Let $u : J \to \mathbb{R}^n$ be a maximal solution. Let's argue by contradiction and assume that $J \neq I$. For example, we assume that $\sup J < \sup I$. Let $\epsilon > 0$ be such that $[\sup J - \epsilon; \sup J + \epsilon] \subset I$. We set $t_0 = \sup J - \epsilon$.

First step: we establish an inequality relating $||u||_2$ and its primitive.

Let $C > 0$ be such that, for all $t \in [\sup J - \epsilon; \sup J + \epsilon]$,

$$||A(t)||_{\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)} \leq C \text{ and } ||b(t)||_2 \leq C.$$

Such a constant exists because A and b are continuous.

We deduce that, for all t sufficiently close to $\sup J$,

$$||u'(t)||_2 \leq C(||u(t)||_2 + 1).$$

For all $t \in [t_0; \sup J[$,

$$\begin{aligned}
||u(t)||_2 &= \left|\left| u(t_0) + \int_{t_0}^t u'(s)ds \right|\right|_2 \\
&\leq ||u(t_0)||_2 + \int_{t_0}^t ||u'(s)||_2 ds \\
&\leq ||u(t_0)||_2 + \int_{t_0}^t C(||u(s)||_2 + 1) ds \\
&= ||u(t_0)||_2 + C(t - t_0) + \int_{t_0}^t C||u(s)||_2 ds.
\end{aligned}$$

Second step: we upper bound $||u||_2$ using Gronwall's lemma.

Gronwall's lemma (Lemma D.1 in the appendix) then implies that, for all $t \in [t_0; \sup J[$,

$$||u(t)||_2 \leq (||u(t_0)||_2 + 1)\, e^{C(t-t_0)} - 1 \leq (||u(t_0)||_2 + 1)\, e^{C\epsilon} - 1.$$

Conclusion: u is bounded in the neighborhood of $\sup J$, meaning that it stays within a compact subset of $\mathbb{R}^n$. This contradicts the théorème des bouts.

□

5.3.1 Without source term

Let's first consider the equation without a source term:

$$u'(t) = A(t)u(t), \tag{5.4}$$

with $A \in C^0(I, \mathbb{R}^{n\times n})$.

Remark

Since the equation is linear in u, a linear combination of solutions is also a solution: if $u_1, u_2 : I \to \mathbb{R}^n$ are two solutions and λ, μ are arbitrary real numbers, $\lambda u_1 + \mu u_2$ is also a solution.

Let us fix any $t_0 \in I$. We denote u_{u_0} the maximal solution of the following Cauchy problem:

$$\begin{cases} u'(t) & = A(t)u(t), \\ u(t_0) & = u_0, \end{cases}$$

For any $t \in I$, from the previous remark, $u_0 \in \mathbb{R}^n \to u_{u_0}(t) \in \mathbb{R}^n$ is a linear map. It can therefore be represented by some matrix $R(t, t_0) \in \mathbb{R}^{n\times n}$: for all u_0,

$$u_{u_0}(t) = R(t, t_0)u_0. \tag{5.5}$$

We call R the *resolvent* of Equation (5.4).

If we can compute the resolvent, then we have access (according to Equation (5.5)) to all maximal solutions of our differential equation (5.4). Unfortunately, in general, we cannot compute an explicit expression of R. However, we can characterize R as the solution to a certain Cauchy problem.

Theorem 5.9

For any $t_0 \in I$, $R(., t_0) : I \to \mathbb{R}^{n\times n}$ is the maximal solution of the Cauchy problem

$$\begin{cases} \frac{dR}{dt}(t, t_0) & = A(t)R(t, t_0), \\ R(t_0, t_0) & = \mathrm{Id}_n. \end{cases}$$

Proof. Let $t_0 \in I$ be fixed. Let $M : I \to \mathbb{R}^{n\times n}$ be the maximal solution of the Cauchy problem:

$$\begin{cases} M'(t) & = A(t)M(t), \\ M(t_0) & = \mathrm{Id}_n. \end{cases}$$

It is defined on the entire interval I according to Proposition 5.8. Let's show that, for all $t \in I$, $M(t) = R(t, t_0)$.

According to the definition of R (Equation (5.5)), we must show that, for all $u_0 \in \mathbb{R}^n$ and all $t \in I$, $u_{u_0}(t) = M(t)u_0$. Let us fix $u_0 \in \mathbb{R}^n$ and define $v : t \in I \to M(t)u_0$. This is a differentiable map, solution of the Cauchy problem

$$\begin{cases} v'(t) & = M'(t)u_0 & = A(t)M(t)u_0 = A(t)v(t), \\ v(t_0) & = M(t_0)u_0 & = u_0. \end{cases}$$

Therefore, $v = u_{u_0}$ and we indeed have, for all t, $u_{u_0}(t) = v(t) = M(t)u_0$. □

Remark

It is tempting to say, by analogy with the scalar case, that the solution to the problem

$$\begin{cases} M'(t) & = A(t)M(t), \\ M(t_0) & = \mathrm{Id}_n \end{cases}$$

is the map $t \in I \to \exp\left(\int_{t_0}^t A(s)ds\right)$. Unfortunately, this is not true (unless the matrices $A(s)$ are pairwise commuting), because, in general, for $X, H \in \mathbb{R}^{n\times n}$, $d\exp(X)(H) \neq H \exp(X)$.

Exercise 9

Let us assume that $n = 1$ (that is, A is real-valued). Given an explicit expression for the resolvent of Equation (5.4).
(The solution is given in a remark of the following subsection.)

Before moving on to linear equations with a source term, here is a classical property of the resolvent.

Proposition 5.10

For all $t_1, t_2, t_3 \in I$, $R(t_3, t_2)R(t_2, t_1) = R(t_3, t_1)$.

Proof. Let $t_1, t_2, t_3 \in I$ be fixed. We fix any $u_1 \in \mathbb{R}^n$, and show that

$$R(t_3, t_2)R(t_2, t_1)u_1 = R(t_3, t_1)u_1.$$

Let $u_{u_1} : I \to \mathbb{R}^n$ be the maximal solution of the Cauchy problem

$$\begin{cases} u'_{u_1}(t) &= A(t)u_{u_1}(t), \\ u_{u_1}(t_1) &= u_1. \end{cases}$$

According to the definition of R, $R(t_3, t_1)u_1 = u_{u_1}(t_3)$ and $R(t_2, t_1)u_1 = u_{u_1}(t_2)$.

Let $u_2 = R(t_2, t_1)u_1$ and $u_{u_2} : I \to \mathbb{R}^n$ be the maximal solution of the Cauchy problem

$$\begin{cases} u'_{u_2}(t) &= A(t)u_{u_2}(t), \\ u_{u_2}(t_2) &= u_2. \end{cases}$$

According to the definition of R, $R(t_3, t_2)R(t_2, t_1)u_1 = R(t_3, t_2)u_2 = u_{u_2}(t_3)$.

Now, u_{u_1} is a solution of the Cauchy problem that defines u_{u_2}. Indeed, $u_{u_1}(t_2) = R(t_2, t_1)u_1 = u_2$. Therefore, $u_{u_1} = u_{u_2}$, and

$$R(t_3, t_2)R(t_2, t_1)u_1 = u_{u_2}(t_3) = u_{u_1}(t_3) = R(t_3, t_1)u_1.$$

□

Corollary 5.11

For all $t_1, t_2 \in I$, $R(t_1, t_2)R(t_2, t_1) = R(t_1, t_1) = \mathrm{Id}_n$, hence $R(t_2, t_1)$ is invertible, with inverse $R(t_1, t_2)$.

5.3.2 With a source term

We now return to the general Equation (5.3) with a source term:

$$u'(t) = A(t)u(t) + b(t). \tag{5.3}$$

As in the scalar case, the method of variation of constants allows us to compute its solutions. Let $u : I \to \mathbb{R}^n$ be any map. Let $t_0 \in I$ and $v : I \to \mathbb{R}^n$ be such that, for all t,

$$u(t) = R(t, t_0)v(t)$$

(i.e., we set $v(t) = R(t_0, t)u(t)$). The map u is a solution of Equation (5.3) if and only if, for all t,

$$\begin{aligned} A(t)R(t, t_0)v(t) + R(t, t_0)v'(t) &= \frac{dR}{dt}(t, t_0)v(t) + R(t, t_0)v'(t) \\ &= u'(t) \\ &= A(t)u(t) + b(t) \\ &= A(t)R(t, t_0)v(t) + b(t). \end{aligned}$$

This is equivalent to stating that, for all t, $R(t, t_0)v'(t) = b(t)$, i.e., v is a primitive of $t \to R(t_0, t)b(t)$. Therefore, u is a solution if and only if there exists $v_0 \in \mathbb{R}^n$ such that, for all $t \in I$,

$$v(t) = v_0 + \int_{t_0}^{t} R(t_0, s)b(s)ds,$$

which is equivalent to

$$\begin{aligned} u(t) &= R(t, t_0)v_0 + \int_{t_0}^{t} R(t, t_0)R(t_0, s)b(s)ds \\ &= R(t, t_0)v_0 + \int_{t_0}^{t} R(t, s)b(s)ds. \end{aligned}$$

This leads us to the following theorem.

Theorem 5.12 : Duhamel's formula

Let I be an open interval, $A \in C^0(I, \mathbb{R}^{n\times n}), b \in C^0(I, \mathbb{R}^n)$.
The maximal solutions of Equation (5.3) are all maps of the form

$$u : t \in I \quad \rightarrow \quad R(t, t_0)v_0 + \int_{t_0}^{t} R(t, s)b(s)ds,$$

for some $v_0 \in \mathbb{R}^n$.

Corollary 5.13

Let I be an open interval, $A \in C^0(I, \mathbb{R}^{n\times n}), b \in C^0(I, \mathbb{R}^n)$, and $u_0 \in \mathbb{R}^n$.
The maximal solution of the Cauchy problem

$$\begin{cases} u'(t) & = A(t)u(t) + b(t), \\ u(t_0) & = u_0 \end{cases}$$

is

$$u : t \in I \quad \rightarrow \quad R(t, t_0)u_0 + \int_{t_0}^{t} R(t, s)b(s)ds.$$

Remark

If $n = 1$, the resolvent has an explicit expression. Indeed, for any t_0, $R(., t_0)$ is the maximal solution of the Cauchy problem

$$\begin{cases} \frac{dR}{dt}(t, t_0) & = A(t)R(t, t_0), \\ R(t_0, t_0) & = \mathrm{Id}_1 = 1. \end{cases}$$

(Note that if $n = 1$, A is a real-valued map.) Therefore, for any t,

$$R(t, t_0) = \exp\left(\int_{t_0}^{t} A(s)ds\right).$$

If we replace R by its value in Duhamel's formula, we recover, as expected, Theorem 5.7.

Exercise 10

We consider the following differential equation:

$$u'(t) = A(t)u(t) + b(t),$$

with

$$A(t) = \begin{pmatrix} t^3 + 2t & t^4 + 3t^2 \\ -t^2 - 1 & -t^3 - 2t \end{pmatrix} \text{ and } b(t) = \begin{pmatrix} -2t^4 - 3t^2 + 3 \\ 2t^3 + t \end{pmatrix}.$$

Let us denote R its resolvent.

1. a) Write the Cauchy problem of which $R(., 0)$ is solution.
   b) Show that, for all $t \in \mathbb{R}$,
   $$R(t, 0) = \begin{pmatrix} 1 + t^2 & t^3 \\ -t & 1 - t^2 \end{pmatrix}.$$
   c) For all $t \in \mathbb{R}$, compute $R(0, t)$.
2. Find all maximal solutions of the differential equation.
3. What is the maximal solution of the following Cauchy problem?
   $$\begin{cases} u'(t) & = A(t)u(t) + b(t), \\ u(1) & = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right). \end{cases}$$

5.3.3 Constant coefficients

Matrix exponential When A is a constant map, the resolvent has an explicit expression. To provide it, it is necessary to recall the definition and main properties of the matrix exponential. The exponential is defined identically for matrices with real or complex coefficients. Here, we state the definition and properties in the general case of complex coefficients.

Definition 5.14 : matrix exponential

For any matrix $A \in \mathbb{C}^{n\times n}$, we define

$$\exp(A) = \sum_{k=0}^{+\infty} \frac{A^k}{k!} \in \mathbb{C}^{n\times n}.$$

This definition is correct, in the sense that the series $\sum_{k=0}^{+\infty} \frac{A^k}{k!}$ converges in $\mathbb{C}^{n\times n}$.

Proposition 5.15

1. For any matrix $A \in \mathbb{C}^{n\times n}$, if the coefficients of A are real, then the coefficients of $\exp(A)$ are also real.
2. For all $A, B \in \mathbb{C}^{n\times n}$, if A and B commute (i.e., $AB = BA$), then
$$\exp(A + B) = \exp(A)\exp(B) = \exp(B)\exp(A).$$
3. For all $A, G \in \mathbb{C}^{n\times n}$ such that G is invertible,
$$\exp(GAG^{-1}) = G\exp(A)G^{-1}.$$
4. For any $A \in \mathbb{C}^{n\times n}$, the map $h : t \in \mathbb{R} \to \exp(tA)$ is differentiable and
$$h'(t) = A\exp(tA) = \exp(tA)A, \quad \forall t \in \mathbb{R}.$$

Corollary 5.16 : exponential of a diagonalizable matrix

Let $A \in \mathbb{C}^{n\times n}$. We assume that there exist $G \in GL(n, \mathbb{C})$ and $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ such that

$$A = G \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} G^{-1}.$$

Then

$$\exp(A) = G \begin{pmatrix} e^{\lambda_1} & 0 & \ldots & 0 \\ 0 & e^{\lambda_2} & & \\ \vdots & & \ddots & \\ 0 & & & e^{\lambda_n} \end{pmatrix} G^{-1}.$$

Proof. According to Property 3 of Proposition 5.15,

$$A = G\exp\left[\begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}\right] G^{-1}.$$

Moreover, for any $k \in \mathbb{N}$,

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}^k = \begin{pmatrix} \lambda_1^k & 0 & \dots & 0 \\ 0 & \lambda_2^k & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n^k \end{pmatrix},$$

which implies that

$$\begin{aligned} \exp\left[\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}\right] &= \sum_{k=0}^{+\infty} \frac{1}{k!} \begin{pmatrix} \lambda_1^k & 0 & \dots & 0 \\ 0 & \lambda_2^k & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n^k \end{pmatrix} \\ &= \begin{pmatrix} \sum_{k=0}^{+\infty} \frac{\lambda_1^k}{k!} & 0 & \dots & 0 \\ 0 & \sum_{k=0}^{+\infty} \frac{\lambda_2^k}{k!} & & \\ \vdots & & \ddots & \\ 0 & & & \sum_{k=0}^{+\infty} \frac{\lambda_n^k}{k!} \end{pmatrix} \\ &= \begin{pmatrix} e^{\lambda_1} & 0 & \dots & 0 \\ 0 & e^{\lambda_2} & & \\ \vdots & & \ddots & \\ 0 & & & e^{\lambda_n} \end{pmatrix}. \end{aligned}$$

□

This corollary allows for the computation of the exponential of any diagonalizable matrix. For matrices that are not diagonalizable, the exponential can be computed using the *Dunford decomposition*. Let's briefly outline the main steps of the computation.

Let $A \in \mathbb{C}^{n \times n}$ be any matrix. The starting point of the method is to write A in the following form:

$$A = G(D + N)G^{-1},$$

where $G, D, N \in \mathbb{C}^{n \times n}$ are matrices such that

- G is invertible;
- D is diagonal;
- N is nilpotent (i.e., there exists $K \in \mathbb{N}^*$ such that $N^K = 0$);
- N and D commute.

This form is called the *Dunford decomposition*. The matrices G, D, N can be explicitely computed from the *characteristic subspaces* of A, but this goes beyond the scope of this course.

Assuming we have found G, D, N, Property 5.15 allows us to write

$$\exp(A) = G \exp(D + N)G^{-1} = G \exp(D) \exp(N)G^{-1}.$$

The exponential of D is given by Corollary 5.16. To compute $\exp(N)$, we directly use the definition: since N is nilpotent, the infinite sum in the definition is actually finite. Denoting K the smallest integer such that $N^K = 0$, we have

$$\exp(N) = \sum_{k=0}^{+\infty} \frac{N^k}{k!} = \sum_{k=0}^{K-1} \frac{N^k}{k!}.$$

Constant coefficients Consider the following Cauchy problem, with constant coefficients:

$$\begin{cases} u'(t) &= Au(t) + b, \\ u(t_0) &= u_0. \end{cases} \tag{5.6}$$

where $A \in \mathbb{R}^{n \times n}, b, u_0 \in \mathbb{R}^n$.

Proposition 5.17

For any $t_0 \in \mathbb{R}$, the resolvent of Equation (5.6) satisfies

$$R(t, t_0) = \exp((t - t_0)A), \quad \forall t \in \mathbb{R}.$$

Proof. For any t_0, according to Theorem 5.9, $R(., t_0)$ is the maximal solution of

$$\begin{cases} \frac{dR}{dt}(t, t_0) & = AR(t, t_0), \\ R(t_0, t_0) & = \mathrm{Id}_n. \end{cases}$$

It suffices to check that $(t \in \mathbb{R} \to \exp((t - t_0)A))$ is this maximal solution. In fact, it suffices to check that $(t \in \mathbb{R} \to \exp((t - t_0)A))$ is a solution of the Cauchy problem: if it is, it is necessarily maximal since it is defined over $\mathbb{R}$.

It satisfies the initial condition: $\exp((t_0 - t_0)A) = \exp(0_{n\times n}) = \mathrm{Id}_n$.

Moreover, according to Property 4 of Proposition 5.15, this map is differentiable and, for all $t \in \mathbb{R}$, its derivative is

$$A \exp((t - t_0)A),$$

so it satisfies the first equation of the Cauchy problem. □

This expression for the resolvent, combined with Duhamel's formula, provides an explicit value for the solution of the Cauchy problem (5.6).

Corollary 5.18

The maximal solution of the problem (5.6) is

$$u : t \in \mathbb{R} \quad \to e^{(t-t_0)A} u_0 + \int_{t_0}^{t} e^{(s-t_0)A} b, ds.$$

When A is invertible, this simplifies to

$$u : t \in \mathbb{R} \quad \to e^{(t-t_0)A} u_0 + \left(e^{(t-t_0)A} - \mathrm{Id}_n \right) A^{-1} b.$$

Chapter 6

Equilibria of autonomous equations

What you should know or be able to do after this chapter

- Know the definition of the flow $(\phi_t)_{t\in\mathbb{R}}$ of an autonomous equation (including the correct domain of each ϕ_t).
- Be able to express the maximal solution of a Cauchy problem in terms of the flow.
- Draw the phase portrait of a two-dimensional differential equation, when it is possible to explicitely compute a solution.
- Draw the vector field associated to a two-dimensional equation (don't forget that it must be tangent to the orbits!).
- Be able to prove that, if A is diagonal with (real) eigenvalues $\lambda_1, \dots, \lambda_n$, an equilibrium of $u' = Au + b$ is
  - *stable* if and only if $\lambda_k \leq 0$ for all $k \in \{1, \dots, n\}$;
  - *asymptotically stable* if and only if $\lambda_k < 0$ for all $k \in \{1, \dots, n\}$.
- Know that an equilibrium u_0 of an equation $u' = f(u)$ is
  - *asymptotically stable* if (but not only if) $\mathrm{Re}(\lambda_k) < 0$ for all $k \in \{1, \dots, n\}$;
  - *unstable* if (but not only if) there exists k such that $\mathrm{Re}(\lambda_k) > 0$,

  where $\lambda_1, \dots, \lambda_n$ are the (complex) eigenvalues of $Jf(u_0)$.
- When you know a first integral of a differential equation in $\mathbb{R}^2$, and the form of its level lines, be able to draw the phase portrait.

6.1 Definitions

The notion of "equilibrium" is mainly meaningful for *autonomous* problems, i.e., for problems of the form (Cauchy) where f does not depend on t. Therefore, in this chapter, we consider a map $f : U \to \mathbb{R}^n$, and, for any $u_0 \in U$, the associated Cauchy problem

$$\begin{cases} u' &= f(u), \\ u(t_0) &= u_0. \end{cases} \tag{Autonomous}$$

We assume that f is locally Lipschitz, so that the Cauchy-Lipschitz theorem is valid.

6.1.1 Flow

Definition 6.1 : Flow of Equation (Autonomous)

For any $u_0 \in U$, let $u_{u_0} : I_{u_0} \to U$ be the maximal solution of Problem (Autonomous) with $t_0 = 0$. For any $t \in I_{u_0}$, we define

$$\phi_t(u_0) = u_{u_0}(t).$$

We call $(\phi_t)_{t\in\mathbb{R}}$ the *flow* of the differential equation.

Remark

The domain of ϕ_t depends on t. For any t, it is given by

$$\{u_0 \in U, t \in I_{u_0}\}.$$

The most intuitive way to understand the flow is as follows. Let's imagine that u represents some physical quantity (such as the position or orientation of an object, for example), and the differential equation $u' = f(u)$ describes its evolution. For any $t \in \mathbb{R}$, ϕ_t represents the action of the evolution on the physical quantity u for t units of time: in our example, if an object is at position u_0 at a reference time 0, it will be at position $\phi_t(u_0)$ at time t.

When f is of class C^2, the map ϕ_t is, for any t, defined on an open set and of class C^1. It is a consequence of the results from Section 4.4 (where the notation was different: the flow was essentially the map V).

Let us remark that, since we consider autonomous equations only, defining the flow using $t_0 = 0$ as the reference point is not a limitation: as the following proposition shows, the solution of Problem (Autonomous) can be expressed in terms of $(\phi_t)_{t\in\mathbb{R}}$ even when $t_0 \neq 0$.

Proposition 6.2

For all $t_0 \in \mathbb{R}, u_0 \in U$, the maximal solution of Problem (Autonomous) is

$$\begin{array}{ccc} I_{u_0} + t_0 & \to & U \\ t & \to & \phi_{t-t_0}(u_0) = u_{u_0}(t - t_0). \end{array}$$

Proof. Let $v_1 : I_1 \to U$ be the maximal solution of (Autonomous) and

$$v_2 : t \in I_2 \stackrel{def}{=} I_{u_0} + t_0 \to u_{u_0}(t - t_0).$$

We must show that these maps are equal.

First step: We show that v_1 is an extension of v_2.

For any t, $v_2'(t) = u_{u_0}'(t - t_0) = f(v_2(t))$ and $v_2(t_0) = u_{u_0}(0) = u_0$. Thus, v_2 is a solution of (Autonomous), so, from Proposition 4.4,

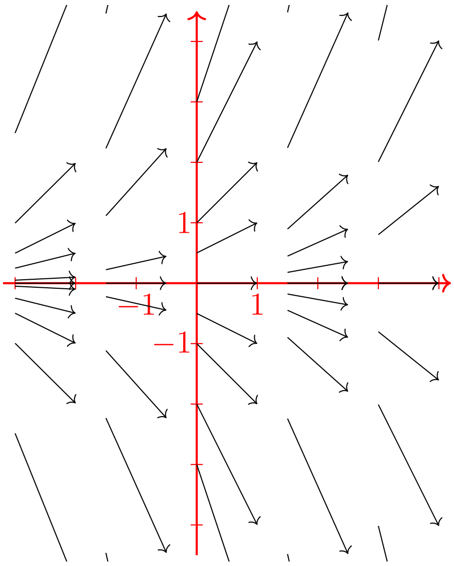
$$I_2 \subset I_1 \text{ and } v_2 = v_1 \text{ on } I_2.$$

Second step: We show that $I_1 = I_2$.

Similarly, the map $t \in I_1 - t_0 \to v_1(t + t_0)$ is a solution of Equation (Autonomous) when t_0 is replaced by 0. Since $u_{u_0} = v_2(. + t_0)$ is the maximal solution of this equation,

$$I_1 - t_0 \subset I_{u_0} = I_2 - t_0.$$

This implies $I_1 = I_2$ and $v_1 = v_2$. $\square$



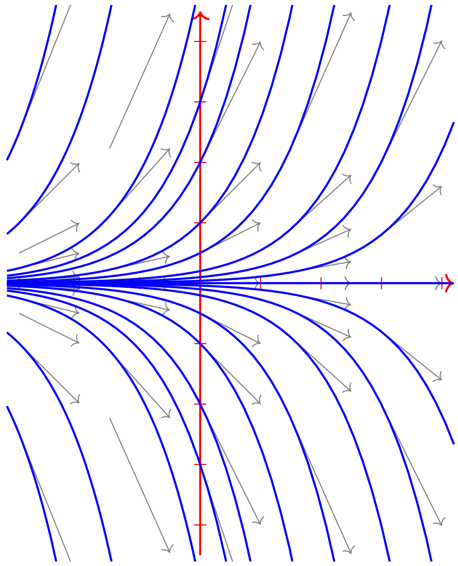


Figure 6.1: On the left, the vector field $f(x,y) = (1,y)$; each arrow links a point (x,y) to $(x,y)+f(x,y)$. On the right, the phase portrait (that is, a few representative orbits).

6.1.2 Phase portrait

Definition 6.3 : orbits

The set
$$\mathcal{O}_{u_0} \stackrel{def}{=} \{\phi_t(u_0), t \in I_{u_0}\}.$$
is called the *orbit* of a point $u_0 \in U$ by the flow $(\phi_t)_{t\in\mathbb{R}}$ of Equation (Autonomous).

The set of orbits forms a "partition" of U, meaning that every point belongs to an orbit (every point belongs at least to its own orbit), and any two orbits are either disjoint (having no common points) or identical.[1] This partition is called the *phase portrait* of Equation (Autonomous).

Example 6.4

Consider the function
$$f: \begin{array}{ccc} \mathbb{R}^2 & \to & \mathbb{R}^2 \\ (x,y) & \to & (1,y) \end{array}$$
and the associated autonomous equation:
$$\begin{cases} x' &= 1, \\ y' &= y, \\ (x(0),y(0)) &= (x_0,y_0). \end{cases}$$
For any $x_0, y_0 \in \mathbb{R}$, the maximal solution is
$$u_{(x_0,y_0)} \quad : \quad \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R}^2 \\ t & \to & (x_0+t, y_0e^t), \end{array}$$
which means that the orbit is
$$\mathcal{O}_{(x_0,y_0)} = \{(x_0+t, y_0e^t), t \in \mathbb{R}\}.$$

[1]Indeed, if for two points $u_0, u_1 \in U$, $\mathcal{O}_{u_0} \cap \mathcal{O}_{u_1} \neq \emptyset$, it means that there exist $t_0 \in I_{u_0}, t_1 \in I_{u_1}$ such that $\phi_{t_0}(u_0) = \phi_{t_1}(u_1)$. With the same reasoning as in the proof of Proposition 6.2, we see that $I_{u_0}+t_1-t_0 = I_{u_1}$ and, for all $t \in I_{u_0}$, $\phi_t(u_0) = \phi_{t+t_1-t_0}(u_1)$, which implies $\mathcal{O}_{u_0} = \mathcal{O}_{u_1}$.

In order to draw the orbits, a useful observation is that this latter set is the graph of a simple map: for any $x_0, y_0 \in \mathbb{R}$,

$$\begin{aligned}\mathcal{O}_{(x_0,y_0)} &= \{(x, y_0 e^{x-x_0}), x \in \mathbb{R}\}\\ &= \{(x, (y_0 e^{-x_0}) e^x), x \in \mathbb{R}\}.\end{aligned}$$

Since $(x_0, y_0) \in \mathbb{R}^2 \to y_0 e^{-x_0} \in \mathbb{R}$ is a surjective map, the possible orbits are all sets of the form

$$\{(x, ce^x), x \in \mathbb{R}\},$$

for some constant $c \in \mathbb{R}$: the orbits are the graphs of all multiples of the exponential map.
The phase portrait is drawn on Figure 6.1. Observe that the vector field f is tangent to the orbits. Indeed, each orbit is the image of a map u such that $u' = f(u)$. Therefore, for each t such that $f(u(t)) \neq 0$, the orbit is a 1-dimensional submanifold in the neighborhood of $u(t)$, with tangent space $\mathrm{Vect}\{u'(t)\} = \mathrm{Vect}\{f(u(t))\}$, from Theorem 2.16.

Exercise 11

Consider the map

$$\begin{array}{cccc} f: & \mathbb{R}^2 & \to & \mathbb{R}^2 \\ & (x,y) & \to & (x(1-x), (1-2x)y). \end{array}$$

The goal is the exercise is to draw the phase portrait of the corresponding autonomous equation

$$u' = f(u). \tag{6.1}$$

Describing the orbits of an arbitrary equation may not be an easy task. However, in this case, as in the previous example, it is possible to explicitely compute them. This is the goal of the first question.

1. Let us fix any $(x_0, y_0) \in \mathbb{R}^2$. We consider the Cauchy problem

$$\begin{cases} x' &= x(1-x), \\ y' &= (1-2x)y, \\ (x(0), y(0)) &= (x_0, y_0). \end{cases}$$

   Let $(x, y) : I \to \mathbb{R}^2$ be the maximal solution of this problem.
   a) Let us assume that there exists $t \in I$ such that $x(t) = 0$. Compute (x, y) and I.
   b) Let us assume that there exists $t \in I$ such that $x(t) = 1$. Compute (x, y) and I.
   c) In this subquestion, and up to 1.f), we assume that $x(t) \notin \{0, 1\}$ for all $t \in I$. It is possible to explicitely compute (x, y) and I, and deduce the orbits from their expression. However, we will follow a different strategy.
      Show that
      - if $x_0 < 0$, x is a decreasing map, with values in $]-\infty; 0[$;
      - if $0 < x_0 < 1$, x is an increasing map, with values in $]0; 1[$;
      - if $x_0 > 1$, x is a decreasing map, with values in $]1; +\infty[$.
   d) Show that $\frac{y}{x(1-x)}$ is constant on I.
   e) Compute the value of y on I, in terms of x, x_0, y_0.
   f) Show that, if $x_0 < 0$, then $x \to 0$ at $\inf I$ and $x \to -\infty$ at $\sup I$.
      [Hint: use the monotonicity of x to show the existence of limits. Then, proceed by contradiction to show that the limits cannot belong to $]-\infty; 0[$.]
      With a similar reasoning, it is possible to show that
      - if $0 < x_0 < 1$, $x \to 0$ at $\inf I$ and $x \to 1$ at $\sup I$;
      - if $1 < x_0$, $x \to +\infty$ at $\inf I$ and $x \to 1$ at $\sup I$.
   g) Find an explicit expression for the orbit of (x_0, y_0).
2. Draw the phase portrait of Equation (6.1).

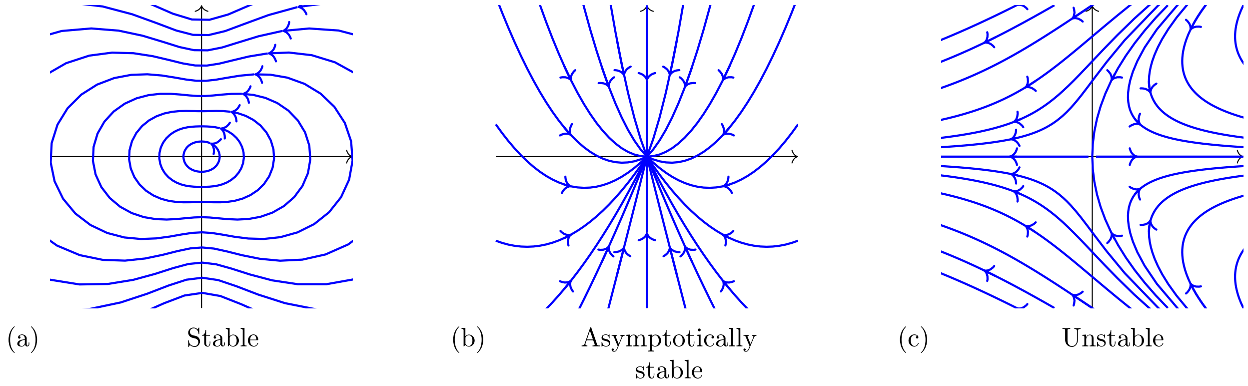


Figure 6.2: Trajectories of Equation (Autonomous), for three different maps $f : \mathbb{R}^2 \to \mathbb{R}^2$ such that $(0,0)$ is an equilibrium.

6.1.3 Equilibria

Definition 6.5 : equilibrium

A point $u_0 \in U$ is an *equilibrium* of the differential equation (Autonomous) if $f(u_0) = 0$ (in other words, if the constant function with value u_0 is a solution of (Autonomous)).

In this chapter, we will try to describe the behavior near equilibria of solutions to Equation (Autonomous). Informally, we will say that an equilibrium is *stable* if every solution starting close enough to the equilibrium remains close to it, and *asymptotically stable* if every trajectory starting close enough to the equilibrium converges to it

Definition 6.6 : stability

If $u_0 \in U$ is an equilibrium of Equation (Autonomous), we say that u_0 is *stable* if, for every neighborhood V_0 of u_0, there exists a neighborhood $V_1 \subset U$ of u_0 such that

- for every $u_1 \in V_1$, $\phi_t(u_1)$ is defined for every $t \in \mathbb{R}^+$ (meaning $\mathbb{R}^+$ is a subset of I_{u_1});
- for every $u_1 \in V_1$ and $t \in \mathbb{R}^+$, $\phi_t(u_1) \in V_0$.

We say that u_0 is *asymptotically stable* if it is stable and, furthermore, there exists a neighborhood $V_2 \subset U$ of u_0 such that, for every $u_2 \in V_2$,

$$\phi_t(u_1) \overset{t\to+\infty}{\longrightarrow} u_0.$$

If u_0 is not stable, we say it is *unstable*.

An illustration of these concepts can be found in Figure 6.2.

6.2 Linear equations

In this section, we study the stability of an equilibrium for a linear differential equation with constant coefficients:

$$u' = Au + b, \tag{6.2}$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

Let us assume that this equation has an equilibrium z_0. By translation,[2] we can assume $z_0 = 0$ and thus $0 = Az_0 + b = b$. The equation is then simply

$$u' = Au. \tag{6.3}$$

[2]"Translation" means considering the differential equation $v' = Av + b + Az_0$ instead of (6.2). Its solutions are the maps $u - z_0$, for all solutions u to (6.2). The point 0 is an equilibrium of the translated equation.

Recall that, according to Corollary 5.18, the flow of any $u_0 \in \mathbb{R}^n$ is

$$\phi_t(u_0) = \exp(tA)u_0, \quad \forall t \in \mathbb{R}.$$

Thus, it is necessary to study $\exp(tA)$.

6.2.1 Diagonalizable Case

First, consider the case where A is diagonalizable over $\mathbb{C}$: there exist complex numbers $\lambda_1, \dots, \lambda_n$ and an invertible matrix $G \in \mathbb{R}^{n\times n}$ such that

$$A = G \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{pmatrix} G^{-1}.$$

For any $t \in \mathbb{R}$, according to Corollary 5.16,

$$\exp(tA) = G \begin{pmatrix} e^{t\lambda_1} & 0 & \dots & 0 \\ 0 & e^{t\lambda_2} & & \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & e^{t\lambda_n} \end{pmatrix} G^{-1}.$$

Let us fix a vector $u_0 \in \mathbb{R}^n$. Denote

$$G^{-1}u_0 = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}.$$

For all t,

$$\phi_t(u_0) = \exp(tA)u_0 = G \begin{pmatrix} g_1 e^{t\lambda_1} \\ \vdots \\ g_n e^{t\lambda_n} \end{pmatrix}. \tag{6.4}$$

Theorem 6.7

The point 0 is a stable equilibrium of the equation (6.3) if and only if

$$\mathrm{Re}(\lambda_k) \le 0, \quad \forall k \in \{1, \dots, n\}.$$

It is an asymptotically stable equilibrium if and only if

$$\mathrm{Re}(\lambda_k) < 0, \quad \forall k \in \{1, \dots, n\}.$$

Proof. Let us first assume that

$$\mathrm{Re}(\lambda_k) \le 0, \quad \forall k \in \{1, \dots, n\}$$

and show that 0 is a stable equilibrium.

For all $t \ge 0$,

$$|e^{t\lambda_k}| = e^{t\mathrm{Re}(\lambda_k)}|e^{it\mathrm{Im}(\lambda_k)}| = e^{t\mathrm{Re}(\lambda_k)} \le 1, \quad \forall k \in \{1, \dots, n\}.$$

From Equation (6.4), we then have, for any u_0 and all $t \ge 0$,

$$\begin{aligned} ||\phi_t(u_0)||_2 &\le |||G^{-1}||| \left\| \begin{pmatrix} g_1 e^{t\lambda_1} \\ \vdots \\ g_n e^{t\lambda_n} \end{pmatrix} \right\|_2 \\ &\le |||G^{-1}||| \sqrt{|g_1|^2 + \dots + |g_n|^2} \end{aligned}$$

$$= |||G^{-1}||| \, ||Gu_0||_2$$
$$\leq |||G^{-1}||| \, |||G||| \, ||u_0||_2. \tag{6.5}$$

This proves that 0 is stable. Indeed, consider an arbitrary neighborhood $V_0 \subset \mathbb{R}^n$ of 0. Let $R > 0$ be such that $B(0, R) \subset V_0$. Define

$$V_1 = B\left(0, \frac{R}{|||G||| \, |||G^{-1}|||}\right).$$

From Equation (6.5), for any $u_0 \in V_1$, $\phi_t(u_0) \in B(0, R) \subset V_0$ for all $t \geq 0$, which establishes stability.

Let us now assume that

$$\text{Re}(\lambda_k) < 0, \quad \forall k \in \{1, \dots, n\}$$

and show that 0 is an asymptotically stable equilibrium. We have already shown that it is stable; let us show that there exists a neighborhood of 0 where all trajectories of the flow converge to 0. The reasoning is as before: for each k, since $\text{Re}(\lambda_k) < 0$,

$$|e^{t\lambda_k}| = e^{t\text{Re}(\lambda_k)} |e^{it\text{Im}(\lambda_k)}| = e^{t\text{Re}(\lambda_k)} \overset{t \to +\infty}{\longrightarrow} 0,$$

thus $e^{t\lambda_k} \overset{t \to +\infty}{\longrightarrow} 0$. Consequently, for any u_0,

$$g_k e^{t\lambda_k} \overset{t \to +\infty}{\longrightarrow} 0, \quad \forall k \in \{1, \dots, n\}.$$

Equation (6.4) therefore shows that $\phi_t(u_0) \overset{t \to +\infty}{\longrightarrow} 0$ for any initial point u_0. The equilibrium is asymptotically stable.

Now let's assume that there exists $k \in \{1, \dots, n\}$ such that

$$\text{Re}(\lambda_k) > 0$$

and let's show that 0 is an unstable equilibrium. For this, we will prove that every neighborhood of 0 contains a point u_0 such that $||\phi_t(u_0)|| \to +\infty$ as $t \to +\infty$. Let thus V be any neighborhood of 0.

Let $u_0 \in \mathbb{R}^n$ be such that $g_k \neq 0$. Such a vector u_0 exists: if not all coordinates of the k-th row of G^{-1} (denoted $(G^{-1})_{k,:}$) are pure imaginary numbers, we can take $u_0 = \text{Re}\left((G^{-1})_{k,:}\right)$ (because then $\text{Re}((G^{-1}u_0)_k) = \left|\left|\text{Re}\left((G^{-1})_{k,:}\right)\right|\right|^2 \neq 0$, hence $g_k \neq 0$). If, on the contrary, all coordinates are pure imaginary numbers, we can set $u_0 = \text{Im}\left((G^{-1})_{k,:}\right)$ (because then $\text{Im}((G^{-1}u_0)_k) = \left|\left|\text{Im}\left((G^{-1})_{k,:}\right)\right|\right|^2 \neq 0$, hence $g_k \neq 0$).

If we multiply u_0 by a sufficiently small constant, we can assume that $u_0 \in V$. According to Equation (6.4), $||G^{-1}\phi_t(u_0)||_2 \to +\infty$ as $t \to +\infty$. Indeed, the k-th coordinate of this vector is $g_k e^{t\lambda_k}$, and

$$\left|g_k e^{t\lambda_k}\right| = |g_k| e^{t\text{Re}(\lambda_k)} \overset{t \to +\infty}{\longrightarrow} +\infty.$$

Now, for any t, $||\phi_t(u_0)||_2 \geq \frac{||G^{-1}\phi_t(u_0)||_2}{|||G^{-1}|||}$. So $||\phi_t(u_0)||_2 \to +\infty$ as $t \to +\infty$, which concludes the proof of instability.

Similarly, let's assume that there exists $k \in \{1, \dots, n\}$ such that

$$\text{Re}(\lambda_k) \geq 0$$

and let's show that 0 is not asymptotically stable. Let's consider again an arbitrary neighborhood V of 0 and a point $u_0 \in V$ such that $g_k \neq 0$. Then

$$\left|g_k e^{t\lambda_k}\right| = |g_k| e^{t\text{Re}(\lambda_k)} \not\to 0 \quad \text{as } t \to +\infty,$$

thus $||\phi_t(u_0)||_2 \not\to 0$ as $t \to +\infty$, so there exists at least one point in V whose trajectory by the flow of Equation (6.3) does not go towards 0.

□

Exercise 12

Rewrite the previous proof, and simplify it as much as possible, in the case where A is a real diagonal matrix.

6.2.2 Non-diagonalizable case

By lack of time, the content of this subsection will not be covered in class. It is provided for curious readers only.

In this subsection, we extend the previous results to the case where A is not diagonalizable over $\mathbb{C}$. A classical result from linear algebra asserts that A is triangularizable and, more precisely, that A can be written in the form

$$A = G \begin{pmatrix} B_1 & 0 & \dots & 0 \\ 0 & B_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & B_K \end{pmatrix} G^{-1},$$

where, for every $k \in \{1, \dots, K\}$, B_k is a square matrix, of the form

$$B_k = \begin{pmatrix} \lambda_k & \star & \dots & \star \\ 0 & \lambda_k & \ddots & \vdots \\ \vdots & \ddots & \ddots & \star \\ 0 & \dots & 0 & \lambda_k \end{pmatrix},$$

for some $\lambda_k \in \mathbb{C}$. We denote $n_k \times n_k$ the dimension of B_k, and $N_k \in \mathbb{C}^{n_k \times n_k}$ the strictly upper triangular part of B_k, so that $B_k = \lambda_k \mathrm{Id}_{n_k} + N_k$.

For any vector $u_0 \in \mathbb{R}^n$, we write

$$G^{-1} u_0 = \begin{pmatrix} g_1 \\ \vdots \\ g_K \end{pmatrix},$$

where, this time, $g_1, \dots, g_K$ are vectors of lengths $n_1, n_2, \dots, n_K$. Analogously to Equation (6.4), Proposition 5.15 implies that, for any $t \geq 0$,

$$\phi_t(u_0) = \exp(tA) u_0 = G \begin{pmatrix} \exp(tB_1) g_1 \\ \vdots \\ \exp(tB_K) g_K \end{pmatrix}. \tag{6.6}$$

We need to compute $\exp(tB_1), \dots, \exp(tB_K)$. For any k, $B_k = \lambda_k \mathrm{Id}_{n_k} + N_k$ and, as $\lambda_k \mathrm{Id}_{n_k}$ and N_k commute,

$$\exp(tB_k) = \exp(t\lambda_k \mathrm{Id}_{n_k}) \exp(tN_k) = e^{t\lambda_k} \exp(tN_k).$$

Since N_k is nilpotent, $t \to \exp(tN_k)$ is a polynomial map, which is constant (equal to Id_{n_k}) if N_k is zero and non-constant otherwise.[3]

We now have all the necessary preliminary results to state and prove the following stability result.

Theorem 6.8

The point 0 is a stable equilibrium of Equation (6.3) if and only if, for every k,

$$(\mathrm{Re}(\lambda_k) < 0) \quad \text{or} \quad (\mathrm{Re}(\lambda_k) = 0 \text{ and } N_k = 0).$$

It is an asymptotically stable equilibrium if and only if, for every k,

$$\mathrm{Re}(\lambda_k) < 0.$$

[3] A quick justification of the fact that the map is not constant when $N_k \neq 0$ is that its derivative $t \to N_k \exp(tN_k)$ is non-zero if $N_k \neq 0$.

Proof. Suppose that, for every $k = 1, \dots, K$,

$$(\mathrm{Re}(\lambda_k) < 0) \quad \text{or} \quad (\mathrm{Re}(\lambda_k) = 0 \text{ and } N_k = 0)\,.$$

Let's show that 0 is a stable equilibrium. As in the proof of Theorem 6.7, it suffices to show the existence of a constant $C > 0$ such that, for every $u_0 \in \mathbb{R}^n$ and every $t \geq 0$,

$$||\phi_t(u_0)||_2 \leq C||u_0||. \tag{6.7}$$

For every k and every t, since $\exp(tB_k) = e^{t\lambda_k} \exp(tN_k)$,

$$|||\exp(tB_k)||| = |e^{t\lambda_k}|\, |||\exp(tN_k)||| = e^{t\mathrm{Re}(\lambda_k)}|||\exp(tN_k)|||.$$

For every k, if $\mathrm{Re}(\lambda_k) < 0$,

$$e^{t\mathrm{Re}(\lambda_k)}|||\exp(tN_k)||| \overset{t\to+\infty}{\longrightarrow} 0.$$

Indeed, the exponential term $e^{t\mathrm{Re}(\lambda_k)}$ goes to 0 while $||\exp(tN_k)||$ is bounded by a polynomial in t (and recall that the product of a polynomial and an exponential goes to 0 at $+\infty$ if the exponential goes to 0). Since $t \to e^{t\mathrm{Re}(\lambda_k)}|||\exp(tN_k)|||$ is continuous, its convergence to 0 at $+\infty$ implies that it is bounded over $\mathbb{R}^+$. Let M_k be an upper bound.

For every k, if $\mathrm{Re}(\lambda_k) = 0$ and $N_k = 0$, then for every t,

$$|||\exp(tB_k)||| = |e^{t\lambda_k}| = 1.$$

So we set $M_k = 1$.

Finally, we define $M = \max(M_1, \dots, M_K)$. From Equation (6.6), for every $u_0 \in \mathbb{R}^n$ and every $t \geq 0$,

$$\begin{aligned} ||\phi_t(u_0)||_2 &\leq M|||G|||\sqrt{||g_1||_2^2 + \cdots + ||g_K||^2} \\ &= M|||G|||\, ||G^{-1}u_0||_2 \\ &\leq M|||G|||\, |||G^{-1}|||\, ||u_0||_2. \end{aligned}$$

This proves Equation (6.7), and thus establishes stability.

The reasoning is similar, but simpler, to show asymptotic stability. Suppose that, for every $k \in 1, \dots, K$,

$$\mathrm{Re}(\lambda_k) < 0.$$

We have just shown that in this case, the equilibrium is stable. We have also seen that, for every k,

$$|||\exp(tB_k)||| \overset{t\to+\infty}{\longrightarrow} 0.$$

Thus, for every $u_0 \in \mathbb{R}^n$, according to Equation (6.6),

$$||\phi_t(u_0)||_2 \overset{t\to+\infty}{\longrightarrow} 0.$$

This shows asymptotic stability.

Now, suppose that it is not true that, for every $k \in 1, \dots, K$,

$$(\mathrm{Re}(\lambda_k) < 0) \quad \text{or} \quad (\mathrm{Re}(\lambda_k) = 0 \text{ and } N_k = 0)$$

and let's show that 0 is unstable. This assumption implies that, for some k,

$$(\mathrm{Re}(\lambda_k) > 0) \quad \text{or} \quad (\mathrm{Re}(\lambda_k) = 0 \text{ and } N_k \neq 0)\,.$$

Let's fix such a k. Let $V \subset \mathbb{R}^n$ be any neighborhood of 0.

Let's start by assuming that $\mathrm{Re}(\lambda_k) > 0$. Let $u_0 \in \mathbb{R}^n$ be such that $g_k \neq 0$. Without loss of generality, we can assume that $u_0 \in V$. For every sufficiently large t,

$$||\exp(tN_k)g_k||_2 \geq ||g_k||_2.$$

Indeed, $t \to \exp(tN_k)g_k$ is a polynomial function. Either it is non-constant, and then $||\exp(tN_k)g_k||_2 \to +\infty$ as $t \to +\infty$, or it is constant, and then for every t, $||\exp(tN_k)g_k||_2 = ||\exp(0N_k)g_k||_2 = ||g_k||_2$.

Thus,

$$||\exp(tB_k)g_k||_2 = e^{t\mathrm{Re}(\lambda_k)}||\exp(tN_k)g_k||_2 \stackrel{t\to+\infty}{\longrightarrow} +\infty.$$

According to Equation (6.6), $||\phi_t(u_0)||_2 \stackrel{t\to+\infty}{\longrightarrow} +\infty$, so $(\phi_t(u_0))_{t\in\mathbb{R}^+}$ does not remain in any neighborhood of 0: the equilibrium is unstable.

Now, let us assume that $\mathrm{Re}(\lambda_k) = 0$ and $N_k \neq 0$. Let $u_0 \in V$ be such that $N_k g_k \neq 0$ (such u_0 exists, by a similar argument as in the proof of Theorem 6.7). Then $t \to \exp(tN_k)g_k$ is a non-constant polynomial function (its derivative at 0 is $N_k g_k \neq 0$), so

$$||\exp(tN_k)g_k||_2 \stackrel{t\to+\infty}{\longrightarrow} +\infty.$$

Consequently, $||\exp(tB_k)g_k||_2 = ||\exp(tN_k)g_k||_2 \stackrel{t\to+\infty}{\longrightarrow} +\infty$, which leads to $||\phi_t(u_0)||_2 \stackrel{t\to+\infty}{\longrightarrow} +\infty$ and completes the proof of instability.

Finally, we assume that there exists $k \in \{1, \dots, K\}$ such that $\mathrm{Re}(\lambda_k) \geq 0$. Let us show that the equilibrium is not asymptotically stable. If $\mathrm{Re}(\lambda_k) > 0$ or $\mathrm{Re}(\lambda_k) = 0$ and $N_k \neq 0$, then the equilibrium is not stable, as we have just shown. The only remaining case we must consider is $\mathrm{Re}(\lambda_k) = 0$ and $N_k = 0$. Let $V \subset \mathbb{R}^n$ be any neighborhood of 0.

Let $u_0 \in V$ be such that $g_k \neq 0$. Then, for every $t \geq 0$,

$$||\exp(tB_k)g_k||_2 = ||e^{t\lambda_k}g_k||_2 = ||g_k||_2.$$

Thus, according to Equation (6.6), $||\phi_t(u_0)||_2 \not\to 0$ as $t \to +\infty$. The equilibrium is not asymptotically stable. □

In the case where the equilibrium is not stable, we can refine the previous reasoning to determine which trajectories of the flow tend toward 0. The resulting statement (which we will not prove) is most simply formulated when A is *hyperbolic*, as defined below.

Definition 6.9 : Hyperbolicity

We say that A is *hyperbolic* if all its complex eigenvalues have non-zero real parts:

$$\mathrm{Re}(\lambda_k) \neq 0, \quad \text{for all } k \in \{1, \dots, K\}.$$

Theorem 6.10 : Stable and unstable spaces

Let A be a hyperbolic matrix. Let us define

$$E^s = \{u_0 \in \mathbb{R}^n \text{ such that } g_k = 0 \text{ for all } k \text{ such that } \mathrm{Re}(\lambda_k) > 0\};$$
$$E^u = \{u_0 \in \mathbb{R}^n \text{ such that } g_k = 0 \text{ for all } k \text{ such that } \mathrm{Re}(\lambda_k) < 0\}.$$

(These sets are called the *stable* and *unstable* subspaces of A.)
Then

$$E^s = \{u_0 \in \mathbb{R}^n \text{ such that } \phi_t(u_0) \stackrel{t\to+\infty}{\longrightarrow} 0\},$$
$$E^u = \{u_0 \in \mathbb{R}^n \text{ such that } \phi_t(u_0) \stackrel{t\to-\infty}{\longrightarrow} 0\}.$$

Moreover, these spaces are complementary: $\mathbb{R}^n = E^s \oplus E^u$.

6.2.3 Graphical representation in dimension 2

In this subsection, we draw trajectories for several hyperbolic 2×2 matrices A. We distinguish three cases as follows:

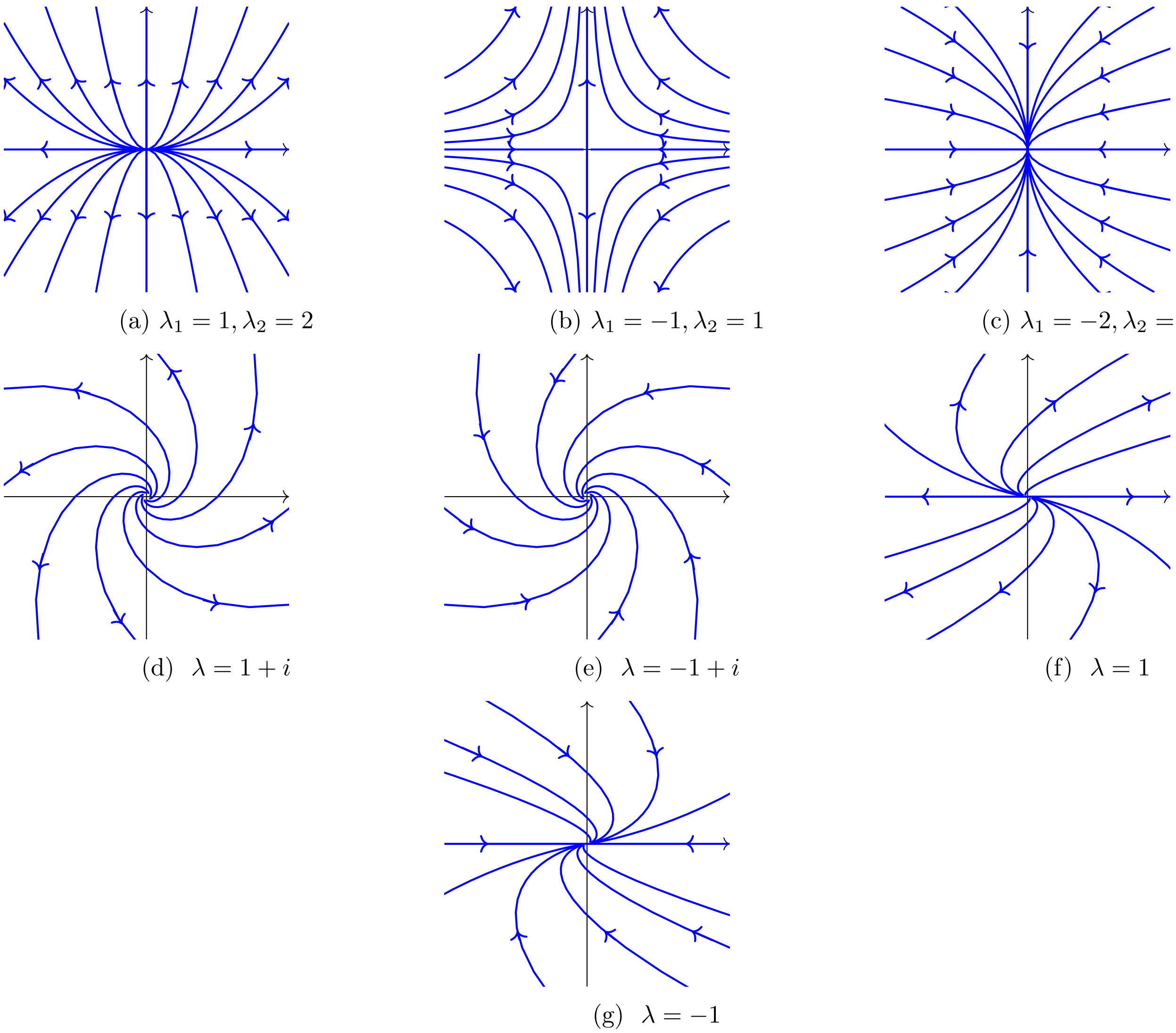
1. If A is diagonalizable with real eigenvalues, we can, after a change of basis (which may not necessarily be orthogonal and can therefore slightly distort the figure, without altering its main properties), assume that
$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$
where $\lambda_1 \leq \lambda_2$ are the eigenvalues. The eigenvalues are non 0 because A is hyperbolic. The flow of a point $u_0 = (x_0, y_0)$ is given by
$$\phi_t(u_0) = (x_0 e^{\lambda_1 t}, y_0 e^{\lambda_2 t}), \quad \forall t \in \mathbb{R}.$$
To draw the phase portrait, note that the orbit of u_0 is included in the graph of the map
$$x \in \mathbb{R} \to \frac{y_0}{|x_0|^{\lambda_2/\lambda_1}} |x|^{\lambda_2/\lambda_1} \in \mathbb{R}.$$
(Observe that λ_2/λ_1 can be positive or negative, depending on whether λ_1 and λ_2 have the same sign; this significantly affects the shape of the graph.)
   (a) $0 < \lambda_1 \leq \lambda_2$: see Figure 6.3a. All trajectories diverge (except the one that remains at 0).
   (b) $\lambda_1 < 0 < \lambda_2$: see Figure 6.3b. The stable space E^s is the x-axis and the unstable space E^u is the y-axis.
   (c) $\lambda_1 \leq \lambda_2 < 0$: see Figure 6.3c. This is an asymptotically stable case: all trajectories converge to 0.

2. If A is diagonalizable with non-real eigenvalues, let $\lambda \in \mathbb{C}$ be one of the eigenvalues. The other one is $\bar{\lambda}$. We can show that, after a suitable change of basis,
$$A = \begin{pmatrix} \mathrm{Re}(\lambda) & \mathrm{Im}(\lambda) \\ -\mathrm{Im}(\lambda) & \mathrm{Re}(\lambda) \end{pmatrix}.$$
We can check that, for any t,
$$\exp(tA) = e^{t\mathrm{Re}(\lambda)} \begin{pmatrix} \cos(t\mathrm{Im}(\lambda)) & \sin(t\mathrm{Im}(\lambda)) \\ -\sin(t\mathrm{Im}(\lambda)) & \cos(t\mathrm{Im}(\lambda)) \end{pmatrix},$$
which is the composition of a rotation with angle $t\mathrm{Im}(\lambda)$ and a homothety with ratio $\exp(t\mathrm{Re}(\lambda))$.
   (a) $\mathrm{Re}(\lambda) > 0$: see figure 6.3d. All trajectories diverge (except the one that remains at 0).
   (b) $\mathrm{Re}(\lambda) < 0$: see figure 6.3e. This is an asymptotically stable case: all trajectories converge to 0.

3. If A is not diagonalizable. In this case, A has only one eigenvalue (for any n, a matrix of size $n \times n$ with n distinct eigenvalues is diagonalizable). This eigenvalue is thus real (non-real eigenvalues can only appear in a pair, with their conjugate). Therefore, A is triangularizable over $\mathbb{R}$. In fact, after a suitable change of basis, we can assume that
$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix},$$
where λ is the eigenvalue. Then, for any t,
$$\exp(tA) = \exp(t\lambda \mathrm{Id}_2) \exp\left(t \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right) = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{pmatrix}.$$
The flow of a point $u_0 = (x_0, y_0)$ is
$$\phi_t(u_0) = ((x_0 + ty_0)e^{\lambda t}, y_0 e^{\lambda t}), \quad \forall t \in \mathbb{R}.$$
   (a) $\lambda > 0$: see Figure 6.3f. All trajectories diverge (except the one that remains at 0).
   (b) $\lambda < 0$: see Figure 6.3g. This is an asymptotically stable case: all trajectories converge to 0.

6.3 Non-linear equations

In this section, we return to Equation (Autonomous) in full generality, without assuming that f is linear. We state and partially prove a theorem that generalizes some of the results we have seen in the linear case.



(a) $\lambda_1 = 1, \lambda_2 = 2$

(b) $\lambda_1 = -1, \lambda_2 = 1$

(c) $\lambda_1 = -2, \lambda_2 = -1$

(d) $\lambda = 1 + i$

(e) $\lambda = -1 + i$

(f) $\lambda = 1$

(g) $\lambda = -1$

Figure 6.3: Flow of Equation (6.3) for various hyperbolic matrices.

Theorem 6.11

Suppose that the map f in Equation (Autonomous) is C^1. Let $u_0 \in U$ be an equilibrium.
If all eigenvalues (over $\mathbb{C}$) of the Jacobian matrix $Jf(u_0)$ have a strictly negative real part, then u_0 is asymptotically stable.
If one eigenvalue of the Jacobian matrix $Jf(u_0)$ has a strictly positive real part, then u_0 is unstable.

Partial Proof. We will only prove asymptotic stability in the case where all eigenvalues have a strictly negative real part.

Without loss of generality, we can assume $u_0 = 0$.

We assume that all eigenvalues of $Jf(0)$ have a strictly negative real part. Let us show that 0 is asymptotically stable. The principle of the proof is to exhibit what is called a *Lyapunov function* of the system, i.e., a map from U to $\mathbb{R}$ that decreases along the trajectories of Equation (Autonomous). This decrease ensures that the sublevel sets of the Lyapunov function are stable under the flow of the differential equation. If these sublevel sets form a "basis"[4] of neighborhoods of 0 (which will be the case), then the equilibrium is stable. By studying more precisely the decay rate of the Lyapunov function, we can even show asymptotic stability.

Our Lyapunov function will be quadratic, and it will be defined in terms of $Jf(0)$. Since $Jf(0)$ is triangularizable over $\mathbb{C}$, we fix $G \in GL(n, \mathbb{C})$ such that

$$Jf(0) = G(D+N)G^{-1},$$

with D a diagonal matrix (whose diagonal entries are the eigenvalues of $Jf(0)$) and N an upper triangular matrix.

Let $\mu = \max_{k=1,\dots,K} \mathrm{Re}(D_{k,k}) < 0$.

First, we show that we can assume $|||N||| < \frac{|\mu|}{2}$. Let us define, for ϵ small enough (we will specify how small ϵ should be later),

$$H = \begin{pmatrix} 1 & & & \\ & \epsilon^{-1} & & \\ & & \dots & \\ & & & \epsilon^{-n} \end{pmatrix}.$$

Then

$$Jf(0) = GH(H^{-1}DH + H^{-1}NH)H^{-1}G^{-1} = GH(D + H^{-1}NH)(GH)^{-1}$$

and, for all $i, j \in \{1, \dots, n\}$,

$$(H^{-1}NH)_{ij} = \frac{H_{jj}}{H_{ii}} N_{ij},$$

so that $(H^{-1}NH)_{ij} = 0$ if $i \geq j$ (i.e., $H^{-1}NH$ is strictly upper triangular) and, if $i > j$,

$$\left|(H^{-1}NH)_{ij}\right| \leq \epsilon |N_{ij}|.$$

Thus, for ϵ close enough to 0, $H^{-1}NH$ can be arbitrarily close to 0. If we replace G with GH and N with $H^{-1}NH$, we can assume that

$$|||N||| < \frac{|\mu|}{2}.$$

We will use as Lyapunov function the map $\left(u \in U \to ||G^{-1}u||_2^2\right)$. Along a trajectory $(\phi_t(u_0))$, following a computation which will be detailed later, its derivative at a point t is $2\mathrm{Re}\left\langle G^{-1}\phi_t(u_0), G^{-1}f(\phi_t(u_0))\right\rangle$. We must therefore bound $\mathrm{Re}\left\langle G^{-1}u, G^{-1}f(u)\right\rangle$, for $u \in U$:

$$\begin{aligned}
&\mathrm{Re}\left(\left\langle G^{-1}u, G^{-1}f(u)\right\rangle\right)\\
&\quad = \mathrm{Re}\left(\left\langle G^{-1}u, G^{-1}\left(f(0) + Jf(0)(u) + o(||u||)\right)\right\rangle\right)\\
&\quad = \mathrm{Re}\left(\left\langle G^{-1}u, G^{-1}G(D+N)G^{-1}u\right\rangle\right) + o\left(||u||^2\right)\\
&\quad = \mathrm{Re}\left(\left\langle G^{-1}u, (D+N)G^{-1}u\right\rangle\right) + o\left(||u||^2\right)
\end{aligned}$$

[4] that is, if any neighborhood of 0 contains a sublevel set

$$
\begin{aligned}
&= \sum_{k=1}^{K} \mathrm{Re}(D_{k,k})|(G^{-1}u)_k|^2 + \mathrm{Re}\left(\langle G^{-1}u, NG^{-1}u\rangle\right) + o\left(||u||^2\right) \\
&\leq \mu||G^{-1}u||_2^2 + |||N|||\,||G^{-1}u||_2^2 + o\left(||u||^2\right) \\
&\leq \frac{\mu}{2}||G^{-1}u||_2^2 + o\left(||u||^2\right) \\
&= \left(\frac{\mu}{2} + o(1)\right)||G^{-1}u||_2^2.
\end{aligned}
$$

Hence, there exists $\eta > 0$ such that, for all $u \in B(0,\eta)$,

$$
\mathrm{Re}\left(\langle G^{-1}u, G^{-1}f(u)\rangle\right) \leq \frac{\mu}{4}||G^{-1}u||_2^2. \tag{6.8}
$$

(Recall that μ is negative, so both terms in the inequality are negative.)

We can now prove asymptotic stability. Let's start with stability. Let $V \subset U$ be any neighborhood of 0. We show that there exists $W \subset U$ a neighborhood of 0 such that, for any $u_1 \in W$, $\phi_t(u_1)$ is well-defined and belongs to V for all $t \in \mathbb{R}^+$.

Let $W = \left\{u \in \mathbb{R}^n \text{ such that } ||G^{-1}u||_2 < \zeta\right\}$, with $\zeta > 0$ a number small enough so that $W \subset V \cap B(0,\eta)$ (the set W is called a *sublevel set* of $\left(u \in U \to ||G^{-1}u||_2^2\right)$). It is an open neighborhood of 0. Let $u_1 \in W$ be arbitrary. Then, for all $t \geq 0$,

$$
\begin{aligned}
\frac{d}{dt}||G^{-1}\phi_t(u_1)||_2^2 &= 2\mathrm{Re}\left(\left\langle G^{-1}\phi_t(u_1), \frac{d}{dt}\left[G^{-1}\phi_t(u_1)\right]\right\rangle\right) \\
&= 2\mathrm{Re}\left(\langle G^{-1}\phi_t(u_1), G^{-1}f\left(\phi_t(u_1)\right)\rangle\right).
\end{aligned}
$$

According to Equation (6.8), for all $t \geq 0$ such that $G^{-1}\phi_t(u_1) \in W$,

$$
\frac{d}{dt}||G^{-1}\phi_t(u_1)||_2^2 \leq \frac{\mu}{2}||G^{-1}\phi_t(u_1)||_2^2 \leq 0 \tag{6.9}
$$

(that is, $\left(u \in U \to ||G^{-1}u||_2^2\right)$ is a *Lyapunov function* on W).

Let $t_0 \in \mathbb{R}^+ \cup \{+\infty\}$ be the largest real number (possibly infinite) such that, for all $t \in [0; t_0[$, $\phi_t(u_1)$ is well-defined and belongs to W. Since W is bounded, $\phi_t(u_1)$ does not leave any compact set in the vicinity of t_0. Therefore, if $t_0 < +\infty$, $\phi_{t_0}(u_1)$ is well-defined (by the théorème des bouts). As we have just seen, our map $\left(t \to ||G^{-1}\phi_t(u_1)||_2^2\right)$ is decreasing on $]0; t_0[$. It is also continuous, so, if $t_0 < +\infty$, we must have

$$
||G^{-1}\phi_{t_0}(u_1)||_2 \leq ||G^{-1}\phi_0(u_1)||_2 < \zeta.
$$

Thus, $G^{-1}\phi_{t_0}(u_1) \in W$. Since W is open and the maximal solutions of (Autonomous) are defined on open sets, there exists $t_1 > t_0$ such that, for all $t \in [0; t_1[$, $\phi_t(u_1)$ is well-defined and belongs to W. This contradicts the definition of t_0. Therefore, it is impossible $t_0 < +\infty$. Hence $t_0 = +\infty$ and, for all $t \in \mathbb{R}^+$, $\phi_t(u_1)$ is well-defined and belongs to W (as well as to V, since $W \subset V$). This completes the proof of stability.

Asymptotic stability follows the same arguments. Let us define W as before (for an arbitrary neighborhood $V \subset U$ of 0) and consider again any arbitrary $u_1 \in W$. According to what we have just seen, the inequality (6.9) is true for all $t \geq 0$. Therefore, for all $t \geq 0$,

$$
\frac{d}{dt}\ln\left(||G^{-1}\phi_t(u_1)||_2^2\right) \leq \frac{\mu}{2},
$$

which implies that, for all $t \geq 0$,

$$
||G^{-1}\phi_t(u_1)||_2^2 \leq ||G^{-1}\phi_0(u_1)||_2^2 e^{-\frac{\mu}{2}t}.
$$

Thus $||G^{-1}\phi_t(u_1)||_2 \overset{t\to+\infty}{\longrightarrow} 0$ and, as a consequence, $||\phi_t(u_1)||_2 \overset{t\to+\infty}{\longrightarrow} 0$. This concludes the proof of asymptotic stability.

□

Exercise 13

We consider the following autonomous equation:

$$\left\{\begin{array}{ll} x' & = \frac{-\frac{x}{2}+y-x(x^2+y^2)}{1+x^2+y^2}, \\ y' & = \frac{-x-\frac{y}{2}-y(x^2+y^2)}{1+x^2+y^2}. \end{array}\right.$$

1. Show that $(0,0)$ is the only equilibrium of this system.
   [Hint: show that any equilibrium (x_0,y_0) is colinear to $(y_0,-x_0)$.]
2. Show that maximal solutions are global.
   [Hint: remember Example 4.9.]
3. Show that $(0,0)$ is an asymptotically stable equilibrium.
4. a) Show that (x,y) is a solution if and only if $(-y,x)$ is a solution.
   b) Which graphical property of the phase portrait can you deduce from the previous question?
5. Let (x,y) be a maximal solution. For any $t\in\mathbb{R}$, we define

   $$N(t)=x(t)^2+y(t)^2.$$

   a) Show that, for all $t\in\mathbb{R}$, $N'(t)\leq -N(t)$.
   b) Show that, for all t,

   $$\begin{aligned} N(t) &\leq N(0)e^{-t} \text{ if } t\geq 0 \\ &\geq N(0)e^{-t} \text{ otherwise.} \end{aligned}$$

   In particular, $N(t)\overset{t\to+\infty}{\longrightarrow} 0$ and, if $N(0)\neq 0$, $N(t)\overset{t\to-\infty}{\longrightarrow}+\infty$.
6. For any maximal solution (x,y), we define

   $$\begin{array}{rccc} S_{(x,y)} & : \mathbb{R} & \to & \mathbb{R}^2 \\ & t & \to & \begin{pmatrix} e^t x(t) \\ e^t y(t) \end{pmatrix} \end{array}$$

   a) Show that there exists a constant C such that, for any maximal solution (x,y) and any $t\in\mathbb{R}$,

   $$||S'_{(x,y)}(t)||_2\leq Ce^t.$$

   b) Let us now consider a fixed non-constant maximal solution (x,y). Show that, if $||(x(0),y(0))||_2>C$, then $S_{(x,y)}$ converges to a non-zero limit at $-\infty$ and, if we denote $L=(L_x,L_y)$ this limit,

   $$||S_{(x,y)}(t)-L||_2\leq Ce^t,\quad \forall t\in\mathbb{R}^-.$$

   c) Show that the result is also true if $||(x(0),y(0))||_2\leq C$.
   [Hint: consider any (x,y) such that $||(x(0),y(0))||_2\leq C$. Show that there exists $t_0<0$ such that $||(x(t_0),y(t_0))||_2>C$. Denote $x_{t_0}=x(.+t_0),y_{t_0}=y(.+t_0)$. Compute $S_{(x,y)}$ in terms of $S_{(x_{t_0},y_{t_0})}$ and apply the previous question to $S_{(x_{t_0},y_{t_0})}$.]
   d) Show that, when $t\to-\infty$,

   $$\begin{aligned} x(t) &= L_x e^{-t}+O(1); \\ y(t) &= L_y e^{-t}+O(1). \end{aligned}$$

   e) Show that there exists $M>0$ and $T<0$ such that, for all $t<T$,

   $$||S'_{(x,y)}(t)||_2\leq Me^{2t}.$$

   f) Show that $S_{(x,y)}(t)=L+O(e^{2t})$ when $t\to-\infty$, and deduce that, when $t\to-\infty$,

   $$\begin{aligned} x(t) &= L_x e^{-t}+O(e^t); \\ y(t) &= L_y e^{-t}+O(e^t). \end{aligned}$$

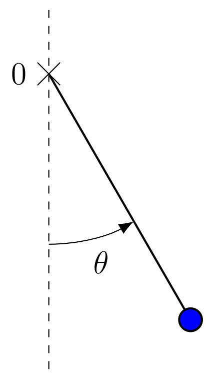


Figure 6.4: Schematic representation of the pendulum.

g) Show that the orbit $\mathcal{O}_{(x,y)}$ has the line $\mathbb{R}L$ as an asymptote.

7. For any maximal solution (x, y), we define

$$V_{(x,y)} \; : \; \begin{array}{ccc} \mathbb{R} & \to & \mathbb{R}^2 \\ t & \to & e^{\frac{t}{2}} R_t \begin{pmatrix} x(t) \\ y(t) \end{pmatrix}, \end{array}$$

where $R_t = \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix}$.

a) Show that there exists a constant $C > 0$ such that, for any maximal solution (x, y) and any $t \geq 0$,

$$||V'_{(x,y)}(t)||_2 \leq C||(x(0), y(0))||_2^3 e^{-t}.$$

[Hint: recall that, from Question 5., it holds for all $t \geq 0$ that $||(x(t), y(t))||_2 \leq e^{-\frac{t}{2}}||x(0), y(0)||_2$.]

b) For any (x, y), show that, if $||(x(0), y(0))||_2 < C^{-1/2}$, then $V_{(x,y)}$ converges to a non-zero limit at $+\infty$ and, if we denote $\lambda = (\lambda_x, \lambda_y)$ this limit,

$$V_{(x,y)}(t) = \lambda + O(e^{-t}) \quad \text{when } t \to +\infty.$$

c) Show that the result is also true if $||(x(0), y(0))||_2 \geq C^{-1/2}$.

d) Show that, when $t \to +\infty$,

$$x(t) = e^{-\frac{t}{2}} \left(\lambda_x \cos(t) + \lambda_y \sin(t)\right) + O\left(e^{-\frac{3t}{2}}\right);$$
$$y(t) = e^{-\frac{t}{2}} \left(-\lambda_x \sin(t) + \lambda_y \cos(t)\right) + O\left(e^{-\frac{3t}{2}}\right).$$

8. Draw a plausible phase portrait.

6.4 Example: the pendulum

In this final section, we study the phase portrait, the equilibria, and the stability of a particular differential equation, which models a pendulum.

6.4.1 Justification of the equation

Consider a pendulum, that is, a small mass, at the end of a rigid rod. The rod is attached to an axis around which it can rotate to the left or right (not forward or backward: the rod remains in a plane). For any $t \in \mathbb{R}$, let $\theta(t)$ denote the angle (positive or negative) between the rigid rod and the vertical at time t. This system is depicted in Figure 6.4.

Imagine that the pendulum is subject to two forces only: the tension of the rod (which ensures that the pendulum remains attached to the rod) and gravity. This is very simplistic: in reality, there would necessarily be frictional forces. Let m be the mass of the pendulum and R the length of the rod. If we take the point of contact between the axis and the rod as the origin, the coordinates of the pendulum in the plane where it moves are, at any instant $t \in \mathbb{R}$,

$$(R\sin(\theta(t)), -R\cos(\theta(t))).$$

The velocity is the derivative of the position,

$$v(t) \overset{def}{=} (R\theta'(t)\cos(\theta(t)), R\theta'(t)\sin(\theta(t))), \quad \text{for all } t \in \mathbb{R},$$

and the acceleration is the derivative of the velocity,

$$\begin{aligned} a(t) \overset{def}{=} (&-R(\theta'(t))^2\sin(\theta(t)) + R\theta''(t)\cos(\theta(t)),\\ &R(\theta'(t))^2\cos(\theta(t)) + R\theta''(t)\sin(\theta(t))), \quad \text{for all } t \in \mathbb{R}. \end{aligned}$$

The force due to gravity is represented by the vector

$$(0, -mg),$$

where g is the universal gravitational constant. The tension force does not have a direct explicit formula, but we know that its direction is the direction of the rod: for any t, there exists $k(t) \in \mathbb{R}$ such that this force is represented by the vector

$$(-k(t)\sin(\theta(t)), k(t)\cos(\theta(t))).$$

The second law of Newton allows us to write, for any t,

$$(0, -mg) + (-k(t)\sin(\theta(t)), k(t)\cos(\theta(t))) = ma(t).$$

Thus,

$$\begin{aligned} -k(t)\sin(\theta(t)) &= -R(\theta'(t))^2\sin(\theta(t)) + R\theta''(t)\cos(\theta(t));\\ -mg + k(t)\cos(\theta(t)) &= R(\theta'(t))^2\cos(\theta(t)) + R\theta''(t)\sin(\theta(t)). \end{aligned}$$

We multiply the first line by $\cos(\theta(t))$, the second line by $\sin(\theta(t))$, and then sum:

$$-mg\sin(\theta(t)) = R\theta''(t).$$

To simplify notation, we assume $mg = R$, which leads to the following equation:

$$\boxed{\theta''(t) = -\sin(\theta(t)).}$$

This is a second-order equation. To arrive at an equation of the form (Autonomous), we follow the remark before the Cauchy-Lipschitz theorem (Theorem 4.1): we introduce the map $u : t \in \mathbb{R} \to (\theta(t), \theta'(t)) \in \mathbb{R}^2$. It satisfies the equation

$$\boxed{u'(t) = f(u(t)),} \tag{Pendulum}$$

with $f : (u_1, u_2) \in \mathbb{R}^2 \to (u_2, -\sin(u_1))$.

It can already be noticed that the maximal solutions of (Pendulum) are defined on $\mathbb{R}$, by virtue of the property stated in Example 4.9.[5]

[5]Indeed, for any (u_1, u_2), since $|\sin(u_1)| \leq |u_1|$, we have $||f(u_1, u_2)||_2 \leq ||(u_1, u_2)||_2$.

6.4.2 Equilibria

The zeros of f (and thus the equilibria of the system (Pendulum)) are the points in $\mathbb{R}^2$ of the form

$$(k\pi, 0)$$

for all integers $k \in \mathbb{Z}$. When k is even, this corresponds to the "bottom" position of the pendulum; when k is odd, on the contrary, it corresponds to the "top" position.

Physical intuition tells us that the bottom position (k even) is stable (if the pendulum is at the bottom and is slightly moved, it will oscillate around the equilibrium position, and not move away from it), while the top position (k odd) is unstable (if the rod is vertical, with the pendulum above the axis, a small disturbance will rather cause the pendulum to fall than to return to this equilibrium position).

To prove this, we can attempt to apply Theorem 6.11. For any $k \in \mathbb{Z}$, the Jacobian matrix of f at $(k\pi, 0)$ is

$$Jf(k\pi, 0) = \begin{pmatrix} 0 & 1 \\ (-1)^{k+1} & 0 \end{pmatrix}.$$

We verify that the eigenvalues of this matrix are i and $-i$ if k is even, 1 and -1 if k is odd. Since $\mathrm{Re}(1) > 0$, the equilibrium $(k\pi, 0)$ must be unstable for all odd k.

However, if k is even, we cannot deduce anything from Theorem 6.11: i and $-i$ have zero real part.

6.4.3 First integral and phase portrait

The trajectories of Equation (Pendulum) do not have an explicit expression. However, they can be studied relatively accurately, and also the stability of the equilibria $(k\pi, 0)$ for even k, thanks to a very useful tool: a *first integral*. This is a map which stays constant along the trajectories of the system, so that the orbits are subsets of its level curves.

In our case, the most natural first integral is

$$F : (u_1, u_2) \in \mathbb{R}^2 \to -\cos(u_1) + \frac{u_2^2}{2}.$$

This is indeed a first integral because, if u is a solution of equation (Pendulum), then, for any t,

$$\begin{aligned}(F \circ u)'(t) &= u_1'(t)\sin(u_1(t)) + u_2'(t)u_2(t) \\ &= u_2(t)\sin(u_1(t)) - \sin(u_1(t))u_2(t) \\ &= 0,\end{aligned}$$

meaning that $F \circ u$ is constant.

What do the level curves of F look like? They are depicted in Figure 6.5.

- If $F_0 < -1$, $\{u, F(u) = F_0\} = \emptyset$, since $F(u_1, u_2) = -\cos(u_1) + \frac{u_2^2}{2} \geq -\cos(u_1) \geq -1$ for all $(u_1, u_2) \in \mathbb{R}^2$.
- If $F_0 = -1$, $\{u, F(u) = F_0\} = \{(2k\pi, 0), k \in \mathbb{Z}\}$; the level set is discrete.
- If $-1 < F_0 < 1$, $\{u, F(u) = F_0\}$ is a union of closed curves, identical to each other up to translation by a multiple of $(2\pi, 0)$.
- If $F_0 = 1$, $\{u, F(u) = F_0\}$ can be written as the union of two (regular) curves that intersect at points $(k\pi, 0)$ for odd k.
- If $F_0 > 1$, $\{u, F(u) = F_0\} = \{(u_1, u_2), u_2 = \pm\sqrt{2(F_0 + \cos(u_1))}\}$. This set has two connected components, both unbounded; one is included in the upper half-plane and the other one in the lower half-plane.

Knowing that the trajectories of Equation (Pendulum) are included in the level curves of F allows us to prove the following theorem.

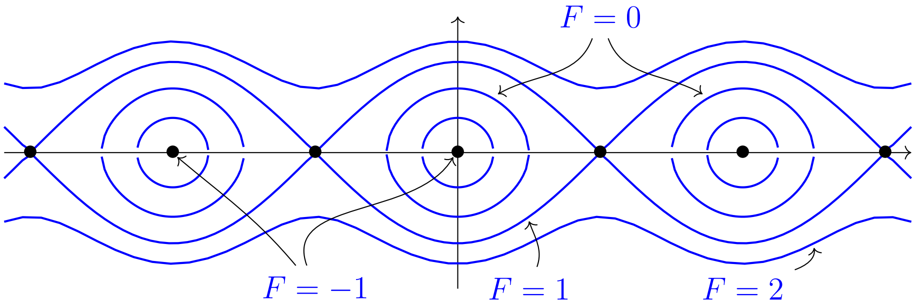


Figure 6.5: Level lines of F; black dots represent equilibria $(k\pi, 0), k \in \mathbb{Z}$.

Theorem 6.12

The constant maximal solutions of Equation (Pendulum) are the maps $(t \in \mathbb{R} \to (k\pi, 0))$ for all $k \in \mathbb{Z}$. Let $u = (u_1, u_2) : \mathbb{R} \to \mathbb{R}^2$ be a non-constant maximal solution.
We set $F_0 = F(u(0))$.

- If $F_0 < 1$, u is periodic. Moreover, there exists $k \in \mathbb{Z}$ an integer such that u_1 alternately increases from $2k\pi - \arccos(-F_0)$ to $2k\pi + \arccos(-F_0)$ and decreases from $2k\pi + \arccos(-F_0)$ to $2k\pi - \arccos(-F_0)$.

- If $F_0 > 1$, u is not periodic and u_1 diverges. However, there exists $T > 0$ such that
$$u(t+T) = u(t) + (2\pi, 0), \quad \text{for all } t \in \mathbb{R}$$
or
$$u(t+T) = u(t) - (2\pi, 0), \quad \text{for all } t \in \mathbb{R}.$$

- If $F_0 = 1$, there exists $k \in \mathbb{Z}$ an integer such that
$$u(t) \overset{t\to-\infty}{\longrightarrow} ((2k-1)\pi, 0) \text{ and } u(t) \overset{t\to+\infty}{\longrightarrow} ((2k+1)\pi, 0)$$
$$\text{or } u(t) \overset{t\to-\infty}{\longrightarrow} ((2k+1)\pi, 0) \text{ and } u(t) \overset{t\to+\infty}{\longrightarrow} ((2k-1)\pi, 0).$$

Before partly proving this theorem, let us discuss the physical meaning of the trajectories. The case $F_0 < 1$ corresponds to periodic oscillation movements around the “bottom” equilibrium position, between angles $-\arccos(-F_0)$ and $\arccos(-F_0)$. The case $F_0 > 1$ corresponds to rotational movements around the axis: starting (for example) from the bottom with a sufficiently high speed, the pendulum reaches the “top” equilibrium position, falls the other side, and repeats.

The case $F_0 = 1$ is quite special. These trajectories are “limits” between the two previous regimes: if the pendulum is launched with exactly the right impulse, it can theoretically go towards the “top” equilibrium position, with a speed that goes to 0 in such a way that the pendulum does not reach this top position in finite time but simply converges to it. These trajectories are never observed in reality.

Partial proof of the theorem. The assertion about constant solutions is due to the fact that the points $(k\pi, 0)$ for $k \in \mathbb{Z}$ are the only zeros of f.

For the rest, we will only prove the first point. The other ones follow from somewhat similar arguments.

Let us assume that $F_0 < 1$. In fact, $F_0 \in]-1; 1[$: F does not take values below -1, and we cannot have $F_0 = -1$, otherwise u would be constant (points reaching value -1 are equilibria).

Let $u = (u_1, u_2)$. The function u_2 is not constant (otherwise, $\sin(u_1) = -u_2'$ must be identically zero, so u_1 is also constant, meaning u is constant). Thus, there exists $t_0 \in \mathbb{R}$ such that $u_2(t_0) \neq 0$. Let's fix such a point. Let us for instance assume that $u_2(t_0) > 0$ (the same reasoning applies if $u_2(t_0) < 0$).

Firstly, we notice that u is bounded. Indeed, for any t,
$$-\cos(u_1(t)) + \frac{u_2(t)^2}{2} = F(u(t)) = F_0,$$

so $u_2(t)^2 \leq 2(F_0 + \cos(u_1)(t)) \leq 2(F_0 + 1)$. Moreover, $u_1(t)$ does not take any value of the form $k\pi$ with $k \in \mathbb{Z}$ odd (since, for any t, $\cos(u_1(t)) = -F_0 + \frac{u_2(t)^2}{2} \geq -F_0 > -1$). By the intermediate value theorem, u_1 thus remains in an interval of the form $]k\pi; (k+2)\pi[$ for some odd $k \in \mathbb{Z}$.

Let's first show that it is impossible that $u_2(t) > 0$ for all $t \geq t_0$. By contradiction, let's assume that $u_2(t) > 0$ for all $t \geq t_0$. Since $u_1'(t) = u_2(t)$ for all t, the function u_1 is increasing on $[t_0; +\infty[$. We have seen that it is bounded. Therefore, it has a limit $\theta_{+\infty}$ as t tends to $+\infty$.

If $\sin(\theta_{+\infty}) > 0$, then $u_2'(t) = -\sin(u_1(t)) < -\frac{1}{2}\sin(\theta_{+\infty})$ for all t large enough. Consequently, $u_2 \to -\infty$ as $t \to +\infty$, which contradicts the fact that u is bounded. Similarly, if $\sin(\theta_{+\infty}) < 0$, we arrive at a contradiction.

Therefore, we must have $\sin(\theta_{+\infty}) = 0$, i.e., $\theta_{+\infty} = k\pi$ for some $k \in \mathbb{Z}$. It is impossible for k to be odd (otherwise, $\cos(u_1(t)) \overset{t\to+\infty}{\longrightarrow} -1$, but we have already seen that $\cos(u_1(t)) \geq -F_0 > -1$ for all t). Thus, k is even, and

$$\cos(u_1(t)) \overset{t\to+\infty}{\longrightarrow} 1.$$

For all $t \geq t_0$, $u_2(t) = \sqrt{2(F_0 + \cos(u_1(t)))}$; consequently,

$$u_2(t) \overset{t\to+\infty}{\longrightarrow} \sqrt{2(F_0+1)}.$$

In particular, $u_2(t) > \sqrt{F_0+1}$ for all t large enough, which implies that $u_1'(t) = u_2(t) > \sqrt{F_0+1}$, so $u_1 \overset{+\infty}{\longrightarrow} +\infty$, which is again a contradiction (recall that we have said that u_1 is bounded).

We have thus shown that it is impossible that $u_2(t) > 0$ for all $t \geq t_0$. Similarly, it is impossible that $u_2(t) > 0$ for all $t \leq t_0$.

Let t_0^- be the largest real number below t_0 such that $u_2(t_0^-) = 0$ and t_0^+ be the smallest real number above t_0 such that $u_2(t_0^+) = 0$. We must have

$$-\cos(u_1(t_0^-)) = -\cos(u_1(t_0^+)) = F_0.$$

This means that $u_1(t_0^-)$ and $u_1(t_0^+)$ are of the form $2k\pi - \arccos(-F_0)$ or $2k\pi + \arccos(-F_0)$, for some $k \in \mathbb{Z}$ (which may not necessarily be the same for t_0^- and t_0^+).

For all t, $\cos(u_1(t)) = \frac{u_2(t)^2}{2} - F_0 \geq -F_0$. As u_1 is strictly increasing to the right of t_0^- (since $u_1' = u_2$), we cannot have $u_1(t_0^-) = 2k\pi + \arccos(-F_0)$ for some $k \in \mathbb{Z}$ (otherwise, as cos is strictly decreasing in the neighborhood of $2k\pi + \arccos(-F_0)$, we would have $\cos(u_1(t)) < -F_0$ for all t slightly greater than t_0^-). Therefore, there exists k_- such that

$$u_1(t_0^-) = 2k_-\pi - \arccos(-F_0).$$

A similar reasoning shows that there exists k_+ such that

$$u_1(t_0^+) = 2k_+\pi + \arccos(-F_0).$$

We must have $k_- \leq k_+$ because $u_1(t_0^-) < u_1(t_0^+)$. We cannot have $k_- < k_+$ otherwise, by the intermediate value theorem, there would exist t such that $u_1(t) = (2k_+ - 1)\pi$, and then $\cos(u_1(t)) = -1 < -F_0$. Thus, $k_- = k_+$. Let's denote this common value as k.

To conclude, we will show that, for all $t \in \mathbb{R}$,

$$\begin{aligned} u_1(t + t_0^+ - t_0^-) &= 4k\pi - u_1(t); \\ u_2(t + t_0^+ - t_0^-) &= -u_2(t). \end{aligned} \tag{6.10}$$

For now, let's assume that these relations hold true and deduce the result. For all t, we have

$$\begin{aligned} u_1(t + 2(t_0^+ - t_0^-)) &= 4k\pi - u_1(t + t_0^+ - t_0^-) = u_1(t); \\ u_2(t + 2(t_0^+ - t_0^-)) &= -u_2(t + t_0^+ - t_0^-) = u_2(t). \end{aligned}$$

This shows that u is $2(t_0^+ - t_0^-)$-periodic.

Additionally, we have seen that u_1 increases from $2k\pi - \arccos(-F_0)$ to $2k\pi + \arccos(-F_0)$ on $[t_0^-; t_0^+]$. The relation $u_1(t + t_0^+ - t_0^-) = 4k\pi - u_1(t)$ shows that it decreases from $2k\pi + \arccos(-F_0)$ to $2k\pi - \arccos(-F_0)$ on $[t_0^+; 2t_0^+ - t_0^-]$. Then, the $2(t_0^+ - t_0^-)$-periodicity shows that u_1 again increases from $2k\pi - \arccos(-F_0)$ to $2k\pi + \arccos(-F_0)$ on $[2t_0^+ - t_0^-; 3t_0^+ - 2t_0^-]$, and so on.

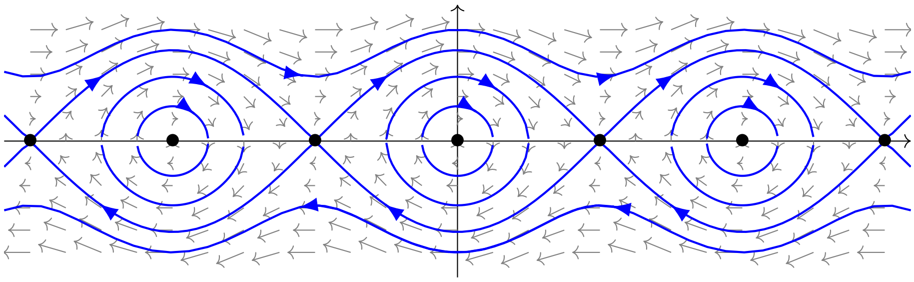


Figure 6.6: Phase portrait of Equation (Pendulum).

All that remains to prove is Equation (6.10). To do this, we define

$$\begin{array}{rccc} v: & \mathbb{R} & \to & \mathbb{R}^2 \\ & t & \to & (4k\pi - u_1(t), -u_2(t)). \end{array}$$

This is a solution of Equation (Pendulum): for all t,

$$\begin{aligned} v_1'(t) &= -u_1'(t) = -u_2(t) = v_2(t) \\ v_2'(t) &= -u_2'(t) = \sin(u_1(t)) = -\sin(4k\pi - u_1(t)) = -\sin(v_1(t)). \end{aligned}$$

The map $t \in \mathbb{R} \to u(t + (t_0^+ - t_0^-))$ is also a solution (as all translations of u). These two solutions are maximal, as they are defined on $\mathbb{R}$. They coincide at t_0^-:

$$\begin{aligned} v(t_0^-) &= (4k\pi - u_1(t_0^-), -u_2(t_0^-)) \\ &= (4k\pi - (2k\pi - \operatorname{arcos}(-F_0)), 0) \\ &= (2k\pi + \operatorname{arcos}(-F_0), 0) \\ &= u(t_0^+) \\ &= u(t_0^- + (t_0^+ - t_0^-)). \end{aligned}$$

By uniqueness of the maximal solution of a Cauchy problem under a locally Lipschitz assumption, we must have $v(t) = u(t + t_0^+ - t_0^-)$ for all $t \in \mathbb{R}$, which proves Equation (6.10).

□

The phase portrait is depicted in Figure 6.6. In this figure, we can clearly see the instability of the critical points $(k\pi, 0)$ for odd integers $k \in \mathbb{Z}$ (some trajectories move away from them even though they started extremely close). The figure also allows us to conjecture, in line with the physical intuition discussed earlier, that the critical points $(k\pi, 0)$ for even integers $k \in \mathbb{Z}$ are stable.

Theorem 6.13

For every even integer $k \in \mathbb{Z}$, $(k\pi, 0)$ is a stable equilibrium of the system.

Proof. Let us prove this for $k = 0$ (which simplifies the notation but does not modify the argument).

Let $V \subset \mathbb{R}^2$ be a neighborhood of $(0, 0)$. Choose $\eta > 0$ such that $]-\eta; \eta[^2 \subset V$. Define the neighborhood of 0 as

$$W = \left\{ u \in \mathbb{R}^2, F(u) < \min\left(-1 + \frac{\eta^2}{2}, -\cos(\eta)\right) \right\} \cap]\eta; -\eta[^2.$$

For any solution of (Pendulum) with $u(0) \in W$, we have $u(t) \in W \subset V$ for all $t \in \mathbb{R}$, and in particular for all $t \geq 0$.

Indeed, since $F \circ u$ is constant, we have for any $t \in \mathbb{R}$ that $F(u(t)) = F(u(0)) < \min\left(-1 + \frac{\eta^2}{2}, -\cos(\eta)\right)$. This implies that there exists no $t \in \mathbb{R}$ such that $u_1(t) = \pm\eta$ or $u_2(t) = \pm\eta$: if, for some t, $u_1(t) = \pm\eta$,

$$F(u(t)) \geq -\cos(u_1(t)) = -\cos(\eta) \geq \min\left(-1 + \frac{\eta^2}{2}, -\cos(\eta)\right)$$

and if $u_2(t) = \pm\eta$,

$$F(u(t)) \geq -1 + \frac{u_2(t)^2}{2} = -1 + \frac{\eta^2}{2} \geq \min\left(-1 + \frac{\eta^2}{2}, -\cos(\eta)\right).$$

In both cases, this is impossible. Since u is continuous, we must have $u(t) \in]-\eta;\eta[^2$ for all $t \in \mathbb{R}$, which completes the proof that $u(t) \in W$. $\square$

Chapter 7

Solutions of some exercises

7.1 Exercise 7

1. As f is C^1, it is locally Lipschitz. The Cauchy-Lipschitz theorem thus implies that the corresponding Cauchy problem has a unique maximal solution.
2. a) The zero map is a solution of the Cauchy problem. It is maximal, as it is defined on $\mathbb{R}$. Since the maximal solution is unique, the zero map is this solution.
   b) Since u is solution of the original problem, it holds $u'(t) = f(u(t))$ for all $t \in J$. In addition, the new initial condition reads $u(t_1) = u(t_1)$, so it is obviously satisfied by u.
   c) Let us assume that $u(t_1) = 0$ for some $t_1 \in J$. From Question 2.b), u is a solution to the Cauchy problem

   $$\begin{cases} u'(t) &= f(u(t)), \\ u(t_1) &= 0. \end{cases}$$

   From Question 2.a), the maximal solution of this problem is the zero map. From Proposition 4.4, u coincides with the maximal solution on its domain, meaning that $u(t) = 0$ for all $t \in J$. In particular, $u_0 = 0$ so we are in the configuration of Question 2.a), which implies that $J = \mathbb{R}$ and $u \equiv 0$.
3. a) As $f(t) \geq t^2 \geq 0$ for all $t \in \mathbb{R}$, u' is nonnegative, hence u is nondecreasing. Therefore, for any $t \in]-\infty; t_0] \cap J$,

   $$u(t) \leq u(t_0) = u_0.$$

   In addition, u is not the zero map (otherwise we would have $u_0 = u(t_0) = 0$). From Question 2., this means that $u(t) \neq 0$ for all $t \in J$. As u is continuous, it must therefore have constant sign. Since $u(t_0) > 0$, it must hold $u(t) > 0$ for all $t \in J$. Summing up, it holds for any $t \in]-\infty; t_0] \cap J$ that

   $$u(t) \in]0; u_0].$$

   b) The previous question implies that, in the neighborhood of $\inf J$, u stays within the compact set $[0; u_0]$. From the théorème des bouts, this implies that $\inf J = -\infty$.
   c) We have seen that u is nondecreasing and lower bounded by 0 on the interval $]-\infty; t_0]$. Consequently, it converges to some nonnegative limit, which we denote $u_{-\infty}$, in $-\infty$.
   By contradiction, we assume that $u_{-\infty} > 0$. Then, when $t \to -\infty$, as f is continuous,

   $$u'(t) = f(u(t)) \to f(u_{-\infty}).$$

   Since $f(u_{-\infty}) \geq u_{-\infty}^2 > 0$, the definition of the limit says that there exists $M \in J$ such that

   $$\forall t \in]-\infty; M], u'(t) \geq \frac{f(u_{-\infty})}{2}.$$

   Let us fix such a number M. For all $t \in]-\infty; M]$,

   $$u(M) - u(t) = \int_t^M u'(s)ds$$

$$\geq \int_t^M \frac{f(u_{-\infty})}{2} ds$$
$$= (M-t)\frac{f(u_{-\infty})}{2}.$$

Equivalently,

$$u(t) \leq u(M) + (t-M)\frac{f(u_{-\infty})}{2}.$$

As $u(M) + (t-M)\frac{f(u_{-\infty})}{2} \to -\infty$ when $t \to -\infty$, it must also hold that $u(t) \to -\infty$ when $t \to -\infty$, which contradicts the fact that u is nonnegative.
Therefore, $u_{-\infty} = 0$.

4. a) We have seen in Question 3.a) that $u(t) > 0$ for all $t \in J$. Therefore, $-\frac{1}{u}$ is well-defined and negative over J.
   b) By the theorem of composition of differentiable maps, $-\frac{1}{u}$ is differentiable over J and, for any $t \in J$,

$$\left(-\frac{1}{u}\right)'(t) = \frac{u'(t)}{u(t)^2}$$
$$= \frac{f(u(t))}{u(t)^2}$$
$$\geq 1.$$

   As a consequence, for any $t \in [t_0; +\infty[\cap J$,

$$-\frac{1}{u(t)} = -\frac{1}{u(t_0)} + \int_{t_0}^t \left(-\frac{1}{u}\right)'(s) ds$$
$$\geq -\frac{1}{u(t_0)} + \int_{t_0}^t 1 ds$$
$$= -\frac{1}{u(t_0)} + (t-t_0).$$

   c) By contradiction, if $\sup J = +\infty$, then, from the previous question, $-\frac{1}{u(t)} \to +\infty$ when $t \to +\infty$. This contradicts the fact that $-\frac{1}{u}$ is negative over J.
   d) We have already seen that u is nondecreasing. Therefore, either it goes to $+\infty$ in $\sup J$, or it stays bounded. It cannot stays bounded, otherwise this would contradict the théorème des bouts. Consequently, it goes to $+\infty$.

7.2 Exercise 10

1. a) This problem is

$$\begin{cases} \frac{dR}{dt}(t,0) & = A(t)R(t,0), \\ R(0,0) & = \mathrm{Id}_2. \end{cases}$$

   b) From the Cauchy-Lipschitz theorem, this problem has a unique maximal solution. If the map $F : t \to \begin{pmatrix} 1+t^2 & t^3 \\ -t & 1-t^2 \end{pmatrix}$ is a solution, it is a *maximal* solution (as its domain is $\mathbb{R}$), and it is therefore the only maximal solution.
   Let us check that F is a solution. It satisfies the initial condition: $F(0) = \mathrm{Id}_2$. Moreover, for all t,

$$\frac{dF}{dt}(t) = \begin{pmatrix} 2t & 3t^2 \\ -1 & -2t \end{pmatrix}$$

   and

$$A(t)F(t) = \begin{pmatrix} 2t & 3t^2 \\ -1 & -2t \end{pmatrix}.$$

c) For all $t \in \mathbb{R}$,

$$\begin{aligned}
R(0,t) &= R(t,0)^{-1} \\
&= \begin{pmatrix} 1+t^2 & t^3 \\ -t & 1-t^2 \end{pmatrix}^{-1} \\
&= \frac{1}{(1+t^2)(1-t^2)-(-t)t^3} \begin{pmatrix} 1-t^2 & -t^3 \\ t & 1+t^2 \end{pmatrix} \\
&= \begin{pmatrix} 1-t^2 & -t^3 \\ t & 1+t^2 \end{pmatrix}.
\end{aligned}$$

2. We use Duhamel's formula: the maximal solutions are all maps of the form

$$u : t \in \mathbb{R} \quad \to \quad R(t,0)u_0 + \int_0^t R(t,s)b(s)ds,$$

for some $u_0 \in \mathbb{R}^2$.
Let us compute $\int_0^t R(t,s)b(s)ds$ for all $t \in \mathbb{R}$. For all $t, s \in \mathbb{R}$,

$$\begin{aligned}
R(t,s) &= R(t,0)R(0,s)b(s) \\
&= R(t,0) \left(\begin{smallmatrix} 1-s^2 & -s^3 \\ s & 1+s^2 \end{smallmatrix}\right) \left(\begin{smallmatrix} -2s^4-3s^2+3 \\ 2s^3+s \end{smallmatrix}\right) \\
&= R(t,0) \left(\begin{smallmatrix} -6s^2+3 \\ 4s \end{smallmatrix}\right).
\end{aligned}$$

As a consequence, for all $t \in \mathbb{R}$,

$$\begin{aligned}
\int_0^t R(t,s)b(s)ds &= \int_0^t R(t,0) \left(\begin{smallmatrix} -6s^2+3 \\ 4s \end{smallmatrix}\right) ds \\
&= R(t,0) \int_0^t \left(\begin{smallmatrix} -6s^2+3 \\ 4s \end{smallmatrix}\right) ds \\
&= R(t,0) \left(\begin{smallmatrix} -2t^3+3t \\ 2t^2 \end{smallmatrix}\right) \\
&= \left(\begin{smallmatrix} t^3+3t \\ -t^2 \end{smallmatrix}\right).
\end{aligned}$$

The maximal solutions of the differential equations are all maps of the form

$$u : t \in \mathbb{R} \quad \to \quad R(t,0)u_0 + \left(\begin{smallmatrix} t^3+3t \\ -t^2 \end{smallmatrix}\right),$$

for some $u_0 \in \mathbb{R}^2$, which can equivalently be written as all maps of the form

$$u : t \in \mathbb{R} \quad \to \quad \begin{pmatrix} t^3+3t \\ -t^2 \end{pmatrix} + u_1 \begin{pmatrix} 1+t^2 \\ -t \end{pmatrix} + u_2 \begin{pmatrix} t^3 \\ 1-t^2 \end{pmatrix},$$

for some $u_1, u_2 \in \mathbb{R}$.

3. To solve the Cauchy problem, it suffices to find, among all maximal solutions, which one satisfies the equality $u(1) = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$. Let us compute for which u_1, u_2 (using the notation of the previous question) the equality holds. The equality is equivalent to

$$\begin{pmatrix} 4 \\ -1 \end{pmatrix} + u_1 \begin{pmatrix} 2 \\ -1 \end{pmatrix} + u_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This amounts to $u_1 = u_2 = -1$. The solution is therefore

$$u : t \in \mathbb{R} \quad \to \quad \begin{pmatrix} -t^2+3t-1 \\ t-1 \end{pmatrix}.$$

7.3 Exercise 11

1. a) Let $t_0 \in I$ be such that $x(t_0) = 0$. Then x is a solution of the Cauchy problem

$$\begin{cases} x' &= x(1-x) \\ x(t_0) &= 0 \end{cases}$$

From the Cauchy-Lipschitz theorem (which applies because $x \to x(1-x)$ is C^1 on $\mathbb{R}$), this problem has a unique maximal solution, and all solutions are restrictions of the maximal solution to a subinterval. Since the zero map is a maximal solution, it is the only maximal solution, and x is the restriction of this map to I, hence x is zero on I. In particular, it must hold that $x_0 = 0$.
For all $t \in I$, $y'(t) = (1-2x(t))y(t) = y(t)$. Therefore, y is a solution of the Cauchy problem

$$\begin{cases} y' &= y \\ y(0) &= y_0 \end{cases}$$

The maximal solution of this problem is $\left(t \in \mathbb{R} \to y_0 e^t\right)$. Therefore, $y(t) = y_0 e^t$ for all $t \in I$.
We have shown that, for all $t \in I$,

$$(x(t), y(t)) = (0, y_0 e^t).$$

Since (x, y) is a maximal solution, the interval I must be equal to $\mathbb{R}$.

b) The same reasoning as in the previous question shows that, if $x(t_0) = 1$ for some $t_0 \in I$, then $x \equiv 1$ on I. In particular, $x_0 = 1$.
The differential equation satisfied by y simplifies and we find that, for all $t \in I$,

$$y(t) = y_0 e^{-t}.$$

Finally, using the maximality of (x, y), we obtain that $I = \mathbb{R}$ and, for all $t \in \mathbb{R}$,

$$(x(t), y(t)) = (1, y_0 e^{-t}).$$

c) Since x is continuous on the interval I, the intermediate values theorem implies that, since $x(t) \notin \{0, 1\}$ for all t, either
- $x(t) < 0$ for all $t \in I$;
- or $0 < x(t) < 1$ for all $t \in I$;
- or $1 < x(t)$ for all $t \in I$.

In the first case, $x'(t) = x(t)(1-x(t)) < 0$ for all $t \in I$, hence x is decreasing. In the second case, $x'(t) = x(t)(1-x(t)) > 0$ for all $t \in I$, hence x is increasing. In the last case, $x'(t) = x(t)(1-x(t)) < 0$ for all $t \in I$, hence x is decreasing.

d) Let us define $F = \frac{y}{x(1-x)}$. It is well-defined on I, since $x(t) \notin \{0, 1\}$ for all $t \in I$. It is also differentiable, since y and x are, and

$$\begin{aligned} F' &= \frac{y'x(1-x) - yx'(1-2x)}{x^2(1-x)^2} \\ &= \frac{(1-2x)yx(1-x) - yx(1-x)(1-2x)}{x^2(1-x)^2} \\ &= 0. \end{aligned}$$

e) For any $t \in I$, $\frac{y(t)}{x(t)(1-x(t))} = \frac{y(0)}{x(0)(1-x(0))} = \frac{y_0}{x_0(1-x_0)}$. Consequently,

$$y(t) = \frac{y_0}{1-x_0} x(t)(1-x(t)).$$

f) Since x is decreasing over I, it must have limits at $\inf I$ and $\sup I$. In addition, since it takes its values in $]-\infty; 0[$,
- the limit at $\inf I$ is in $]-\infty; 0]$;

- the limit at $\sup I$ is in $\{-\infty\}\cup]-\infty;0[$.

We must show that none of these two limits is in $]-\infty;0[$.
By contradiction, let us assume that x converges to some $\ell\in]-\infty;0[$ at $\inf I$. Then y goes to $\frac{y_0}{1-x_0}\ell(1-\ell)$. This means that (x,y) is bounded in the neighborhood of $\inf I$. From the théorème des bouts, $\inf I=-\infty$. Since x is decreasing, $x(t)<\ell$ for all $t\in I$. This implies, for all $t\in I$, using the fact that $1-x(t)>1$, that

$$x'(t)=x(t)(1-x(t))<\ell(1-x(t))<\ell.$$

In particular, for all $t\in I$,

$$\begin{aligned}
x(t)&=x(M)-\int_t^M x'(s)ds\\
&\geq x(M)-\int_t^M \ell ds\\
&=x(M)-\ell(M-t)\\
&=\ell t+x(M)-\ell M\\
&\overset{t\to-\infty}{\longrightarrow}+\infty.
\end{aligned}$$

Therefore, x actually goes to $+\infty$ at $-\infty$, which is a contradiction. We have shown that x converges to 0 at $\inf I$.
By contradiction, let us assume that x converges to some $\ell\in]-\infty;0[$ at $\sup I$. In the same way as before, it must then hold $\sup I=+\infty$.
There exists $M\in I$ such that, for all $t\in[M;+\infty[$, $x(t)<\frac{\ell}{2}$. Then, for all $t\in[M;+\infty[$,

$$x'(t)=x(t)(1-x(t))<\frac{\ell}{2}.$$

As a consequence, for all $t\in[M;+\infty[$,

$$\begin{aligned}
x(t)&=x(M)+\int_M^t x'(s)ds\\
&\leq x(M)+\frac{\ell}{2}(t-M)\\
&\overset{t\to+\infty}{\longrightarrow}-\infty.
\end{aligned}$$

Therefore, $x(t)\overset{t\to+\infty}{\longrightarrow}-\infty$, which is a contradiction. This shows that x converges to $-\infty$ at $\sup I$.

g) From the Questions 1.a) and 1.b), if $x_0=0$ or $x_0=1$, the orbit is

$$\begin{aligned}
\mathcal{O}_{(x_0,y_0)}&=\{x_0\}\times\mathbb{R}_+^* && \text{if } y_0>0,\\
&=\{(x_0,0)\} && \text{if } y_0=0,\\
&=\{x_0\}\times\mathbb{R}_-^* && \text{if } y_0<0.
\end{aligned}$$

From Question 1.e), if $x_0\notin\{0,1\}$, then the orbit is a subset of

$$\left\{\left(x,\frac{y_0}{1-x_0}x(1-x)\right),x\in\mathbb{R}\right\}.$$

From Question 1.f), the orbit is

$$\begin{aligned}
\mathcal{O}_{(x_0,y_0)}&=\left\{\left(x,\tfrac{y_0}{1-x_0}x(1-x)\right),x\in\mathbb{R}_-^*\right\} && \text{if } x_0<0,\\
&=\left\{\left(x,\tfrac{y_0}{1-x_0}x(1-x)\right),x\in]0;1[\right\} && \text{if } 0<x_0<1,\\
&=\left\{\left(x,\tfrac{y_0}{1-x_0}x(1-x)\right),x\in\mathbb{R}_+^*\right\} && \text{if } 1<x_0.
\end{aligned}$$

2. The phase portrait is drawn on Figure 7.1.

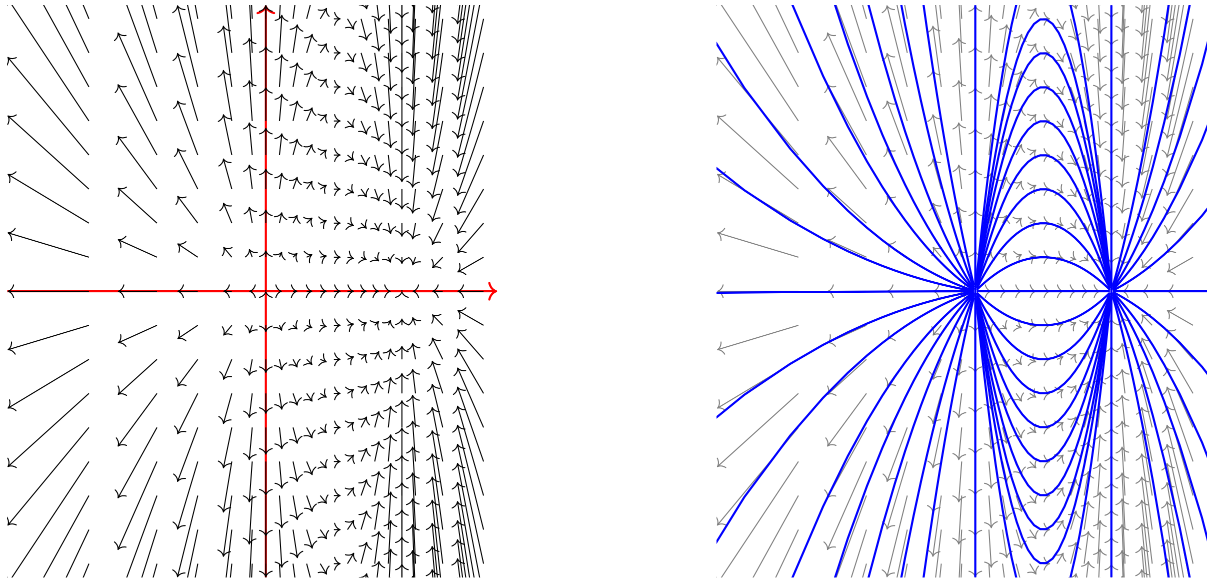


Figure 7.1: On the left, the vector field $f(x,y) = (x(1-x), (1-2x)y)$ (the length of each arrow has been divided by 5 for a better readability); on the right, the corresponding phase portrait.

7.4 Exercise 13

1. The point $(0,0)$ is an equilibrium because it cancels the right-hand side of the equation. Conversely, let (x_0, y_0) be an equilibrium. Then

$$-\frac{x_0}{2} + y_0 - x_0(x_0^2 + y_0^2) = 0;$$
$$-x_0 - \frac{y_0}{2} - y_0(x_0^2 + y_0^2) = 0,$$

which implies

$$\begin{pmatrix} y_0 \\ -x_0 \end{pmatrix} = \left(\frac{1}{2} + x_0^2 + y_0^2\right)\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

Therefore, $(y_0, -x_0)$ is colinear to (x_0, y_0). These vectors are orthogonal and have the same norm, hence this is only possible if $x_0 = y_0 = 0$.

2. For all $(x,y) \in \mathbb{R}^2$, we denote

$$f(x,y) = \begin{pmatrix} \frac{-\frac{x}{2}+y-x(x^2+y^2)}{1+x^2+y^2}, \\ \frac{-x-\frac{y}{2}-y(x^2+y^2)}{1+x^2+y^2} \end{pmatrix}.$$

It holds, for all (x,y),

$$\begin{aligned} ||f(x,y)||_2 &= \frac{\left|\left|-\left(\frac{1}{2}+x^2+y^2\right)\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right) + \left(\begin{smallmatrix} y \\ -x \end{smallmatrix}\right)\right|\right|_2}{1+x^2+y^2} \\ &\leq \frac{\frac{3}{2}+x^2+y^2}{1+x^2+y^2}\left|\left|\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)\right|\right|_2 \\ &\leq \frac{3}{2}\left|\left|\left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right)\right|\right|_2. \end{aligned}$$

Example 4.9 concludes.

3. The map f is C^∞. For $(x,y) \in \mathbb{R}^2$ close to zero,

$$f(x,y) = \begin{pmatrix} \frac{-\frac{x}{2}+y+O(||(x,y)||^3)}{1+O(||(x,y)||^2)} \\ \frac{-x-\frac{y}{2}+O(||(x,y)||^3)}{1+O(||(x,y)||^2)} \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{2} & 1 \\ -1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + O(||(x,y)||^3).$$

Therefore, the Jacobian at $(0,0)$ is

$$Jf(0,0) = \begin{pmatrix} -\frac{1}{2} & 1 \\ -1 & -\frac{1}{2} \end{pmatrix}.$$

This matrix has two eigenvalues, $-\frac{1}{2} + i$ and $-\frac{1}{2} - i$. Their real part is strictly negative, so $(0,0)$ is asymptotically stable, in virtue of Theorem 6.11.

4. a) Let (x,y) be a solution. Let us define $u = -y$ and $v = x$. It holds

$$u' = -y' = \frac{x + \frac{y}{2} + y(x^2+y^2)}{1+x^2+y^2} = \frac{-\frac{u}{2} + v - u(u^2+v^2)}{1+u^2+v^2};$$
$$v' = x' = \frac{-\frac{x}{2} + y - x(x^2+y^2)}{1+x^2+y^2} = \frac{-u - \frac{v}{2} - v(u^2+v^2)}{1+u^2+v^2},$$

so that $(u,v) = (-y,x)$ is also a solution of the equation. For the same reason, if $(-y,x)$ is a solution of the equation, then (x,y) is also a solution.

b) The phase portrait is invariant under a rotation of angle $\frac{\pi}{2}$.

5. a) For all $t \in \mathbb{R}$,

$$\begin{aligned} N'(t) &= 2(x(t)x'(t) + y(t)y'(t)) \\ &= -2(x(t)^2+y(t)^2)\frac{\frac{1}{2} + x(t)^2 + y(t)^2}{1+x(t)^2+y(t)^2} \\ &= -(x(t)^2+y(t)^2)\frac{1 + 2(x(t)^2+y(t)^2)}{1+x(t)^2+y(t)^2} \\ &\le -N(t). \end{aligned}$$

b) If (x,y) is the constant solution (the only one is the one which stays at $(0,0)$), then the result is true. Otherwise, N never vanishes, so we can consider $n \overset{def}{=} \ln(N)$. It is a C^∞ function and, for all t,

$$n'(t) = \frac{N'(t)}{N(t)} \le -1.$$

Consequently, for all $t \in \mathbb{R}$,

$$\begin{aligned} n(t) &\le n(0) - t \text{ if } t \ge 0, \\ &\ge n(0) - t \text{ if } t \le 0. \end{aligned}$$

This is equivalent to

$$\begin{aligned} N(t) &\le N(0)e^{-t} \text{ if } t \ge 0, \\ &\ge N(0)e^{-t} \text{ if } t \le 0. \end{aligned}$$

Therefore, by comparison, N goes to 0 at $+\infty$ and to $+\infty$ at $-\infty$.

6. a) For any maximal solution (x,y) and any $t \in \mathbb{R}$,

$$\begin{aligned} S'_{(x,y)}(t) &= \begin{pmatrix} e^t(x(t) + x'(t)) \\ e^t(y(t) + y'(t)) \end{pmatrix} \\ &= \frac{e^t}{1+x(t)^2+y(t)^2} \begin{pmatrix} \frac{x}{2} + y \\ -x + \frac{y}{2} \end{pmatrix}. \end{aligned}$$

Consequently,

$$\begin{aligned} ||S'_{(x,y)}(t)|| &\le \frac{3}{2}\frac{e^t}{1+x(t)^2+y(t)^2}||(x(t),y(t))||_2 \\ &\le \frac{3}{4}e^t. \end{aligned}$$

b) Let us assume that $||(x(0),y(0))||_2 > C$. It holds, for all $t \geq 0$,

$$S_{(x,y)}(t) = S_{(x,y)}(0) - \int_t^0 S'_{(x,y)}(s)ds.$$

Since $\int_{-\infty}^0 ||S'_{(x,y)}(s)||_2 ds \leq \int_{-\infty}^0 Ce^s ds = C < +\infty$, the integral is convergent, meaning that it has a limit when $t \to -\infty$. Therefore, $S_{(x,y)}$ also has a limit, which is

$$L \overset{def}{=} S_{(x,y)}(0) - \int_{-\infty}^0 S'_{(x,y)}(s)ds.$$

As $||S_{(x,y)}(0)||_2 = ||(x(0),y(0))||_2 > C$ and

$$\left|\left|\int_{-\infty}^0 S'_{(x,y)}(s)ds\right|\right|_2 \leq \int_{-\infty}^0 ||S'_{(x,y)}(s)||_2 ds \leq C,$$

the limit L must be non-zero.
For all $t \leq 0$,

$$\begin{aligned} \left|\left|S_{(x,y)}(t) - L\right|\right|_2 &= \left|\left|\int_{-\infty}^t S'_{(x,y)}(s)ds\right|\right|_2 \\ &\leq \int_{-\infty}^t Ce^s ds \\ &= Ce^t. \end{aligned}$$

c) We assume that $||(x(0),y(0))||_2 \leq C$. Let $t_0 < 0$ be such that $||(x(t_0),y(t_0))||_2 > C$; such a t_0 exist because, from Question 4.b), $||(x(t),y(t))||_2 \to +\infty$ when $t \to +\infty$.
Let us define $(\tilde{x},\tilde{y})$ the maximal solution such that

$$\begin{pmatrix} \tilde{x}(0) \\ \tilde{y}(0) \end{pmatrix} = \begin{pmatrix} x(t_0) \\ y(t_0) \end{pmatrix}.$$

Since the equation is autonomous, $x = \tilde{x}(. - t_0)$ and $y = \tilde{y}(. - t_0)$. In particular, for all $t \in \mathbb{R}$,

$$e^{t_0} S_{(\tilde{x},\tilde{y})}(t - t_0) = S_{(x,y)}(t). \tag{7.1}$$

From the previous subquestion, there exists $L \in \mathbb{R}^2 \setminus \{(0,0)\}$ such that $S_{(\tilde{x},\tilde{y})} \overset{-\infty}{\longrightarrow} L$ and, for all $t \in \mathbb{R}$,

$$||S_{(\tilde{x},\tilde{y})}(t) - L||_2 \leq Ce^t.$$

Using Equation (7.1), we get that $S_{(x,y)}$ goes to Le^{t_0} at $-\infty$ and, for all $t \in \mathbb{R}$,

$$||S_{(x,y)}(t) - e^{t_0}L||_2 \leq Ce^t.$$

d) When $t \to -\infty$,

$$\begin{aligned} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} &= e^{-t} S_{x,y}(t) \\ &= e^{-t}\left(L + O(e^t)\right) \\ &= Le^{-t} + O(1). \end{aligned}$$

e) Recall from Question 6.a) that, for all t,

$$\begin{aligned} ||S'_{(x,y)}(t)||_2 &\leq \frac{3}{2}\frac{e^t}{1 + x(t)^2 + y(t)^2} ||(x(t),y(t))||_2 \\ &\leq \frac{3}{2}\frac{e^t}{||(x(t),y(t))||_2}. \end{aligned}$$

Moreover, from the previous subquestion, there exists $a > 0$ such that, for all t small enough,

$$||(x(t), y(t))||_2 \geq a e^{-t}.$$

Then, for all t small enough,

$$||S'_{(x,y)}(t)||_2 \leq \frac{3}{2a} e^{2t}.$$

f) For all t small enough,

$$\begin{aligned} ||S_{(x,y)}(t) - L||_2 &= \left|\left| \int_{-\infty}^{t} S'_{(x,y)}(s) ds \right|\right|_2 \\ &\leq \int_{-\infty}^{t} M e^{2s} ds \\ &= \frac{M}{2} e^{2t}. \end{aligned}$$

This says that $S_{(x,y)}(t) = L + O(e^{2t})$. Consequently,

$$\begin{aligned} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} &= e^{-t} S_{x,y}(t) \\ &= e^{-t} \left(L + O(e^{2t}) \right) \\ &= L e^{-t} + O(e^{t}). \end{aligned}$$

g) For any t, the distance of $(x(t), y(t))$ to the line $\mathbb{R}L$ is at most

$$\left|\left| \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} - e^{-t} \begin{pmatrix} L_x \\ L_y \end{pmatrix} \right|\right|_2 .$$

From Question 6.f), this is of order $O(e^t)$, hence goes to 0.

7. a) Let (x, y) be a maximal solution. For any $t \geq 0$,

$$\begin{aligned} V'_{(x,y)}(t) &= e^{\frac{t}{2}} \left(R_t \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} + \frac{1}{2} R_t \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} + R'_t \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \right) \\ &= e^{\frac{t}{2}} \left(R_t \begin{pmatrix} x'(t) \\ y'(t) \end{pmatrix} + \frac{1}{2} R_t \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} + R_t \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \right) \\ &= e^{\frac{t}{2}} R_t \begin{pmatrix} x'(t) + \frac{x(t)}{2} - y(t) \\ y'(t) + \frac{y(t)}{2} + x(t) \end{pmatrix} \\ &= \frac{e^{\frac{t}{2}} (x(t)^2 + y(t)^2)}{1 + x(t)^2 + y(t)^2} R_t \begin{pmatrix} -\frac{x(t)}{2} - y(t) \\ x(t) - \frac{y(t)}{2} \end{pmatrix}. \end{aligned}$$

Therefore, for any $t \geq 0$,

$$\begin{aligned} ||V'_{(x,y)}(t)||_2 &\leq \frac{3}{2} e^{\frac{t}{2}} ||(x(t), y(t))||_2^3 \\ &\leq \frac{3}{2} e^{-t} ||(x(0), y(0))||_2^3. \end{aligned}$$

b) Let us assume that $||(x(0), y(0))||_2 < C^{-1/2}$. For all $t \geq 0$,

$$V_{(x,y)}(t) = V_{(x,y)}(0) + \int_0^t V'_{(x,y)}(s) ds.$$

The integral is convergent :

$$\int_0^{+\infty} ||V'_{(x,y)}(s)||_2 \leq C ||(x(0), y(0))||_2^3 \int_0^{+\infty} e^{-s} ds$$

$$= C||(x(0), y(0))||_2^3,$$

so this converges to

$$\lambda \stackrel{def}{=} V_{(x,y)}(0) + \int_0^{+\infty} V'_{(x,y)}(s)ds.$$

This limit is non-zero because

$$\begin{aligned}\left|\left|\int_0^{+\infty} V'_{(x,y)}(s)ds\right|\right|_2 &\leq C||(x(0), y(0))||_2^3 \\ &< ||(x(0), y(0))||_2 \\ &= ||V_{(x,y)}(0)||_2.\end{aligned}$$

For any t,

$$\begin{aligned}\left|\left|V_{(x,y)}(t) - \lambda\right|\right|_2 &= \left|\left|\int_t^{+\infty} V'_{(x,y)}(s)ds\right|\right|_2 \\ &\leq C||(x(0), y(0))||_2^3 \int_t^{+\infty} e^{-s}ds \\ &= C||(x(0), y(0))||_2^3 e^{-t},\end{aligned}$$

so that $V_{(x,y)}(t) = \lambda + O(e^{-t})$.

c) This is the same reasoning as in Question 6.c). Let $t_0 > 0$ be such that

$$||(x(t_0), y(t_0))||_2 < C^{-1/2}.$$

Let $(\tilde{x}, \tilde{y})$ be the solution whose value is $(x(t_0), y(t_0))$ at time 0. From the previous subquestion, $V_{(\tilde{x},\tilde{y})}$ satisfies, for some non-zero $\lambda \in \mathbb{R}^2$,

$$V_{(\tilde{x},\tilde{y})}(t) = \lambda + O(e^{-t}),$$

which implies

$$V_{(x,y)}(t) = e^{\frac{t_0}{2}} R_{t_0} V_{(\tilde{x},\tilde{y})}(t - t_0) = e^{\frac{t_0}{2}} R_{t_0}\lambda + O(e^{-t}).$$

d) For all $t \geq 0$,

$$\begin{aligned}\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} &= e^{-\frac{t}{2}} R_{-t} V_{(x,y)}(t) \\ &= e^{-\frac{t}{2}} \begin{pmatrix} \lambda_x \cos(t) + \lambda_y \sin(t) \\ -\lambda_x \sin(t) + \lambda_y \cos(t) \end{pmatrix} + O\left(e^{-\frac{3t}{2}}\right).\end{aligned}$$

8. The phase portrait is drawn in Figure 7.2. Observe the following properties:
   - the phase portrait is invariant under rotation by $\frac{\pi}{2}$;
   - all non-zero trajectories are asymptotic to a line going through zero at one end;
   - all non-zero trajectories go to $(0, 0)$ with a spiraling behavior (in the indirect sense) at the other end.

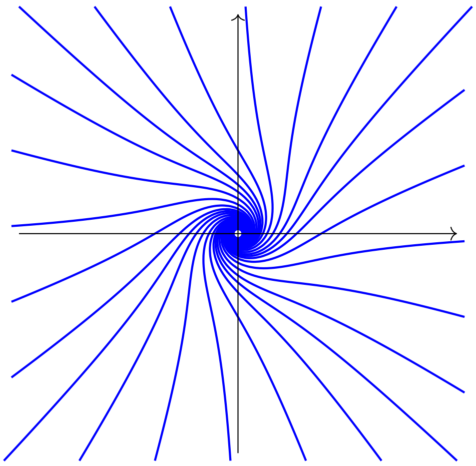


Figure 7.2: Phase portrait for the equation in Exercice 13.