Ordinary differential equations

Teo Banica

Department of Mathematics, University of Cergy-Pontoise, F-95000 Cergy-Pontoise, France. teo.banica@gmail.com

2010 *Mathematics Subject Classification.* 34A05

Key words and phrases. Differential equation, Dynamical system

Abstract. This is an introduction to the ordinary differential equations, with all needed preliminaries included. We first study the basic differential equations, of low order, notably with a detailed discussion regarding the classical orthogonal polynomials. Then we consider differential equations of arbitrary order, and we develop the standard theory here, using linearization ideas, and tools from linear algebra. As a continuation of this, we further discuss the geometric aspects, using tools from differential geometry. Finally, we restrict the attention to problems coming from classical and celestial mechanics, and we provide an introduction to the advanced theory here.

Preface

This is an introduction to the ordinary differential equations, with all needed preliminaries included. The book is organized in 4 parts, as follows:

I. We first study the basic differential equations, of low order, notably with a detailed discussion regarding the classical orthogonal polynomials.

II. Then we consider differential equations of arbitrary order, and we develop the standard theory here, using linearization ideas, and tools from linear algebra.

III. As a continuation of this, we further discuss the geometric aspects of the various equations and their solutions, by using tools from differential geometry.

IV. Finally, we restrict the attention to problems coming from classical and celestial mechanics, and we provide an introduction to the advanced theory here.

This book is based on lecture notes from various classes that I taught at Toulouse and Cergy, on differential equations and dynamical systems, and I would like to thank my students. Many thanks as well to my cats, for some help with the computations.

Cergy, September 2025
Teo Banica

Contents

Part I

Differential equations

CHAPTER 1

Basic functions

1a. Basic functions

1b.

1c.

1d.

1e. Exercises

Exercises:

Exercise 1.1.

Exercise 1.2.

Exercise 1.3.

Exercise 1.4.

Exercise 1.5.

Exercise 1.6.

Exercise 1.7.

Exercise 1.8.

Bonus exercise.

CHAPTER 2

Differential equations

2a. Differential equations

Let us start with some basic mechanics. We will be interested in free falls, and the result here, which is something quite familiar, and that we can establish right from the Newton principles, with just a bit of basic calculus, is as follows:

THEOREM 2.1. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity $v_0 = 0$, the equation of the trajectory is*

$$x \simeq R - \frac{gt^2}{2}$$

with the constant being $g = GM/R^2$, called gravity of M, at distance R from it.

PROOF. As before, the equation of motion of our object m is as follows:

$$\ddot{x} = -\frac{Kx}{||x||^3}$$

In one dimension now, things get simpler, and the equation of motion reads:

$$\ddot{x} = -\frac{K}{x^2}$$

Since we assumed $R >> 0$, we must look for a solution of type $x \simeq R + ct^2$, with the lack of the t term coming from $v_0 = 0$. But with $x \simeq R + ct^2$, our equation reads:

$$2c \simeq -\frac{K}{R^2}$$

Now by multiplying by $t^2/2$, and adding R, we obtain as solution:

$$x \simeq R - \frac{Kt^2}{2R^2}$$

Thus, we have indeed $x \simeq R - gt^2/2$, with g being the following number:

$$g = \frac{K}{R^2} = \frac{GM}{R^2}$$

We are therefore led to the conclusion in the statement. □

Along the same lines, as a second result now, which is more advanced, we have:

THEOREM 2.2. *In the context of a free fall from distance $x_0 = R >> 0$, with initial plane velocity vector $v_0 = v$, the equation of the trajectory is*

$$x \simeq R + vt - \frac{gt^2}{2}$$

where $g = GM/R^2$ as usual, and with the quantities R, g in the above being regarded now as vectors, pointing upwards. The approximate trajectory is a parabola.

PROOF. We have several assertions here, the idea being as follows:

(1) Let us first discuss the simpler case where we are still in 1D, as in Theorem 2.1, but with an initial velocity $v_0 = v$ added. In order to find the equation of motion, we can just redo the computations from the proof of Theorem 2.1, with now looking for a general solution of type $x \simeq R + vt + ct^2$, and we get, as stated above:

$$x \simeq R + vt - \frac{gt^2}{2}$$

Alternatively, we can simply argue that, by linearity, what we have to do is to take the solution $x \simeq R - gt^2/2$ found in Theorem 2.1, and add an extra vt term to it.

(2) In the general 2D case now, where the initial velocity $v_0 = v$ is a vector in $\mathbb{R}^2$, the same arguments apply, either by redoing the computations from the proof of Theorem 2.1, or simply by arguing that by linearity we can just take the solution $x \simeq R - gt^2/2$ found there, and add an extra vt term to it. Thus, we have our solution.

(3) Let us study now the solution that we found. In standard (x, y) coordinates, with $v = (p, q)$, and with R, g being now back scalars, our solution looks as follows:

$$x = pt \quad , \quad y \simeq R + qt - \frac{gt^2}{2}$$

From the first equation we get $t = x/p$, and by substituting into the second:

$$y \simeq R + \frac{qx}{p} - \frac{gx^2}{2p^2}$$

We recognize here the approximate equation of a parabola, and we are done. □

Let us discuss now an important topic, namely the conservation of energy, in the gravitational context. The simplest situation is that of a free fall with initial velocity $v_0 = 0$, and our conservation principle here is as follows:

PROPOSITION 2.3. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity $v_0 = 0$, if we define the potential energy to be*

$$V = mgx$$

then the total energy $E = T + V$, with $T = mv^2/2$ as usual, is constant, $E \simeq mgR$.

PROOF. We know that the equation of motion is as follows, with $g = GM/R^2$:

$$x \simeq R - \frac{gt^2}{2}$$

The kinetic energy, from now on to be denoted T, is then given by:

$$T \simeq \frac{mv^2}{2} = \frac{mg^2t^2}{2}$$

Thus with $V = mgx$ as in the statement, and then with $E = T + V$, we have:

$$E = T + V \simeq mgR$$

But this is a constant, and so we have our conservation principle, as desired. □

Along the same lines, as a next result, we have:

THEOREM 2.4. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity vector $v_0 \in \mathbb{R}^2$, if we define the potential energy to be*

$$V = m < g, x >$$

with $g = GM/R^2$ being regarded as usual as a vector pointing upwards, then

$$E = T + V$$

with $T = m||v||^2/2$ as usual, is constant, $E \simeq T_0 + mgR$, with g now back scalar.

PROOF. We can do this in two steps, first by adding an extra parameter to the computation in Proposition 2.3, and then by adding another extra parameter:

(1) Let us first examine the 1D case, where $v_0 = s$ is a vector aligned to x, and so a number. Here the equation of motion is as follows, with $g = GM/R^2$ as usual:

$$x \simeq R + st - \frac{gt^2}{2}$$

The speed being $v \simeq s - gt$, with $V = mgx$ and $E = T + V$ as above, we have:

$$\begin{aligned} E &= T + V \\ &\simeq \frac{m(s-gt)^2}{2} + mg\left(R + st - \frac{gt^2}{2}\right) \\ &= \frac{ms^2}{2} + mgR \\ &= T_0 + mgR \end{aligned}$$

(2) In the general case now, with $v_0 = s$, the equation of motion is as before, with R, g being now vectors pointing upwards, and if we write $s = (a, b)$, then we have:

$$\begin{aligned} T &\simeq \frac{m||s - gt||^2}{2} \\ &= \frac{m((a - gt)^2 + b^2)}{2} \\ &= \frac{m(a^2 + b^2)}{2} - magt + \frac{mg^2t^2}{2} \\ &= T_0 - mg\left(at - \frac{gt^2}{2}\right) \end{aligned}$$

With g vector pointing upwards, the last quantity is $m < g, x - R >$, so if we add $V = m < g, x >$, we obtain $T_0 + mgR$, with g, R being back scalars, as desired. □

With the above done, let us get back to the real thing, 3D gravity. We are interested in the general 2-body problem, where M is fixed at 0, and m moves under the gravitational force of M. The above computations, coming from our "kinetic energy gets converted into height, and vice versa" principle, suggest defining the potential energy as:

$$V \sim ||x||$$

However, this is wrong, because in our formula $V = mgx$ the quantity $g = GM/R^2$ depends on the average height, which is the parameter R, no longer assumed to satisfy $R >> 0$. In view of this, the correct formula for the potential energy should be:

$$V \sim \frac{1}{||x||}$$

In order now to find the constant, it is enough to rewrite $V = mgx$ by getting rid of the parameter $g = GM/R^2$. We obtain in this way, with $K = GM$ as usual:

$$V = mgx = \frac{mGMx}{R^2} \simeq \frac{mGM}{||x||} = \frac{Km}{||x||}$$

Thus, we have our formula for V, and the question now is if $E = T+V$ is constant. And the answer here is unfortunately no, due to some bizarre reasons, with rather $E = T - V$ appearing to be constant, or at least that's what computations tend to suggest.

So, let us change the sign of V, and see what we get. We are led in this way to:

THEOREM 2.5. *In the context of the 2-body problem, with M fixed at 0 and with m moving, if we define the kinetic and potential energy of m to be*

$$T = \frac{m||v||^2}{2} \quad , \quad V = -\frac{Km}{||x||}$$

with $K = GM$ as usual, then the total energy $E = T + V$ is constant.

PROOF. The idea will be that of proving $\dot{E} = 0$. We first have:

$$\begin{aligned} \dot{T} &= \frac{m(< v, \dot{v} > + < \dot{v}, v >)}{2} \\ &= m < v, \dot{v} > \\ &= m < v, a > \end{aligned}$$

Next, let us compute the derivative of the function $f(x) = 1/||x||$. We have:

$$\begin{aligned} \dot{f} &= -\frac{1}{2} \cdot \frac{< x, \dot{x} > + < \dot{x}, x >}{< x, x >^{3/2}} \\ &= -\frac{< x, \dot{x} >}{< x, x >^{3/2}} \\ &= -\frac{< x, v >}{||x||^3} \end{aligned}$$

Thus, getting now to the potential energy V, we have the following formula:

$$\dot{V} = \frac{Km < x, v >}{||x||^3}$$

In order to further process this, remember the equation of motion of m, namely:

$$a = -\frac{Kx}{||x||^3}$$

We will of course jump on this, as to get rid of $||x||^3$, and we finally obtain:

$$\dot{V} = -m < a, v >$$

We are ready now to prove our result. Indeed, we have:

$$\dot{E} = \dot{T} + \dot{V} = m < v, a > -m < a, v >= 0$$

Now since the derivative vanishes, E is constant, as claimed. □

Nice all this, but we still have to understand the relation with Proposition 2.3 and Theorem 2.4, with that sign of V mysteriously switching. And we have here the following result, upgrading Proposition 2.3 and Theorem 2.4, and clarifying the whole thing:

THEOREM 2.6. *In the context of a free fall from distance $x_0 = R >> 0$, with initial velocity $v_0 = 0$, if we define the kinetic and potential energy of m to be*

$$T = \frac{mv^2}{2} \quad , \quad V = -\frac{Km}{x}$$

with $K = GM$ as usual, then the total energy $E = T + V$ is constant. Moreover,

$$V \simeq mgx - 2mgR$$

with $g = GM/R^2$, and so $E' = T + mgx$ is appoximately constant, $E' \simeq mgR$. The same happens for a free fall from $x_0 = R >> 0$, with initial velocity vector $v_0 \in \mathbb{R}^2$.

PROOF. The first assertion is something that we know, coming from Theorem 2.5. In order to clarify now the relation with Proposition 2.3, we first have:

$$V=-\frac{Km}{x}=-\frac{GMm}{x}=-\frac{mgR^2}{x}$$

Now by writing $x=R(1-\varepsilon)$, we obtain the estimate in the statement, namely:

$$\begin{aligned}V&=-\frac{mgR}{1-\varepsilon}\\&\simeq-mgR(1+\varepsilon)\\&=mgR[(1-\varepsilon)-2]\\&=mgx-2mgR\end{aligned}$$

Thus with $V'=mgx$ we have $V\simeq V'-2mgR$, and so $E'=T+V'$ satisfies:

$$\begin{aligned}E'&\simeq E+2mgR\\&=E_0+2mgR\\&=V_0+2mgR\\&=mgR\end{aligned}$$

Finally, the last assertion, which is a bit more general, follows in the same way. □

We will be back to all this later, following Lagrange and Hamilton.

2b.

2c.

2d.

2e. Exercises

Exercises:

EXERCISE 2.7.

EXERCISE 2.8.

EXERCISE 2.9.

EXERCISE 2.10.

EXERCISE 2.11.

EXERCISE 2.12.

EXERCISE 2.13.

EXERCISE 2.14.

Bonus exercise.

CHAPTER 3

Theorems and tricks

Exercises:

Exercise 3.1.

Exercise 3.2.

Exercise 3.3.

Exercise 3.4.

Exercise 3.5.

Exercise 3.6.

Exercise 3.7.

Exercise 3.8.

Bonus exercise.

CHAPTER 4

Orthogonal polynomials

4a. Scalar products

We discuss in this chapter an extension of the basic linear algebra results, obtained by looking at the linear operators $T : H \to H$, with the space H being no longer assumed to be finite dimensional. Our main motivations come from physics, and more specifically quantum mechanics, and in order to get motivated, here is some suggested reading:

(1) Generally speaking, physics is best learned from Feynman [32]. If you already know some, and want to learn quantum mechanics, go with Griffiths [42]. And if you are already a bit familiar with quantum mechanics, a good book is Weinberg [95].

(2) A look at classics like Dirac [22], von Neumann [32] or Weyl [97] can be instructive too. On the opposite, you have as well modern, fancy books on quantum information, such as Bengtsson-Życzkowski [32], Nielsen-Chuang [32] or Watrous [32].

(3) In short, many ways of getting familiar with this big mess which is quantum mechanics, and as long as you stay away from books advertised as "rigorous", "axiomatic", "mathematical", things fine. By the way, you can try as well my book [11].

Getting to work now, physics tells us to look at infinite dimensional complex spaces, such as the space of wave functions $\psi : \mathbb{R}^3 \to \mathbb{C}$ of the electron. In order to do some mathematics on these spaces, we will need scalar products. So, let us start with:

DEFINITION 4.1. *A scalar product on a complex vector space H is a binary operation $H \times H \to \mathbb{C}$, denoted $(x, y) \to < x, y >$, satisfying the following conditions:*

(1) $< x, y >$ *is linear in x, and antilinear in y.*
(2) $\overline{< x, y >} =< y, x >$, *for any x, y.*
(3) $< x, x >> 0$, *for any $x \neq 0$.*

As before in the previous chapters, we use here mathematicians' convention for scalar products, that is, $<,>$ linear at left, as opposed to physicists' convention, $<,>$ linear at right. The reasons for this are quite subtle, coming from the fact that, while basic quantum mechanics looks better with $<,>$ linear at right, advanced quantum mechanics looks better with $<,>$ linear at left. Or at least that's what my cats say.

As a basic example for Definition 4.1, we have the finite dimensional vector space $H = \mathbb{C}^N$, with its usual scalar product, namely:

$$< x, y >= \sum_i x_i \bar{y}_i$$

We will see later in this chapter that in finite dimensions, this is in fact the only example, the point being that algebrically we must have $H \simeq \mathbb{C}^N$, for some $N \in \mathbb{N}$, and then we can always change the basis, as to make it orthogonal with respect to $<,>$, which in practice makes $<,>$ to be given by the above formula. More on this in a moment.

In infinite dimensions now, there are many interesting examples of spaces naturally coming with scalar products, and notably various spaces of L^2 functions, which appear for instance in various problems coming from physics. We will discuss them later.

Summarizing, what we have in Definition 4.1 is a potentially useful generalization of the usual scalar product $<,>$ on the simplest complex vector space, $\mathbb{C}^N$. In order to study now the scalar products, let us formulate the following definition:

DEFINITION 4.2. *The norm of a vector $x \in H$ is the following quantity:*

$$||x|| = \sqrt{< x, x >}$$

We also call this number length of x, or distance from x to the origin.

The terminology comes from what happens in $\mathbb{C}^N$, where the length of the vector, as defined above, coincides with the usual length, given by:

$$||x|| = \sqrt{\sum_i |x_i|^2}$$

In analogy with what happens in finite dimensions, we have two important results regarding the norms. First we have the Cauchy-Schwarz inequality, as follows:

THEOREM 4.3. *We have the Cauchy-Schwarz inequality,*

$$| < x, y > | \leq ||x|| \cdot ||y||$$

and the equality case holds precisely when x, y are proportional.

PROOF. This is something very standard, the idea being as follows:

(1) Consider, and we will understand why in a moment, the following quantity, depending on a real variable $t \in \mathbb{R}$, and on a variable on the unit circle, $w \in \mathbb{T}$:

$$f(t) = ||twx + y||^2$$

By developing f, we see that this is a degree 2 polynomial in t:

$$\begin{aligned} f(t) &= < twx + y, twx + y > \\ &= t^2 < x, x > + tw < x, y > + t\bar{w} < y, x > + < y, y > \\ &= t^2||x||^2 + 2tRe(w < x, y >) + ||y||^2 \end{aligned}$$

(2) Since f is obviously positive, its discriminant must be negative:

$$4Re(w < x, y >)^2 - 4||x||^2 \cdot ||y||^2 \leq 0$$

But this is equivalent to the following condition:

$$|Re(w < x, y >)| \leq ||x|| \cdot ||y||$$

Now the point is that we can arrange for the number $w \in \mathbb{T}$ to be such that the quantity $w < x, y >$ is real. Thus, we obtain the Cauchy-Schwarz inequality:

$$| < x, y > | \leq ||x|| \cdot ||y||$$

(3) Finally, the study of the equality case is straightforward, by using the fact that the discriminant of f vanishes precisely when we have a root. Indeed, this shows that having equality in Cauchy-Schwarz is the same as asking for the following to happen:

$$f(t) = 0$$

But this latter condition is very easy to process, as follows:

$$\begin{aligned} f(t) = 0 &\iff ||twx + y||^2 = 0 \\ &\iff ||twx + y|| = 0 \\ &\iff twx + y = 0 \\ &\iff x \sim y \end{aligned}$$

Thus we are led to the conclusion in the statement, namely that in order to have equality in the Cauchy-Schwarz inequality, the vectors x, y must be proportional. □

As a second main result now, we have the Minkowski inequality:

THEOREM 4.4. *We have the Minkowski inequality*

$$||x + y|| \leq ||x|| + ||y||$$

and the equality case holds precisely when x, y are proportional.

PROOF. This follows indeed from the Cauchy-Schwarz inequality, as follows:

$$\begin{aligned} & ||x + y|| \leq ||x|| + ||y|| \\ \iff & ||x + y||^2 \leq (||x|| + ||y||)^2 \\ \iff & ||x||^2 + ||y||^2 + 2Re < x, y > \leq ||x||^2 + ||y||^2 + 2||x|| \cdot ||y|| \\ \iff & Re < x, y > \leq ||x|| \cdot ||y|| \end{aligned}$$

As for the equality case, this is clear from Cauchy-Schwarz as well. □

In abstract terms, the Minkowski inequality tells us that the following happens:

PROPOSITION 4.5. *The following function is a norm on H,*

$$||x|| = \sqrt{< x, x >}$$

in the usual sense, that of the abstract normed spaces.

PROOF. Recall indeed that a normed space is an abstract vector space X with a function $||.|| : X \to [0, \infty)$, called norm, subject to the following conditions:

– $||x|| > 0$ for $x \neq 0$.

– $||\lambda x|| = |\lambda| \cdot ||x||$.

– $||x + y|| \leq ||x|| + ||y||$.

In our case, the first two axioms are trivially satisfied, and the third axiom, called triangle inequality, is the Minkowski inequality. Thus, the result holds indeed. □

Alternatively, and perhaps more illustrating, we have the following result:

THEOREM 4.6. *The following function is a distance on H,*

$$d(x, y) = ||x - y||$$

in the usual sense, that of the abstract metric spaces.

PROOF. This follows indeed from the Minkowski inequality, which corresponds to the triangle inequality, the other two axioms being trivially satisfied. To be more precise:

(1) Let us first recall that a metric space is an abstract space X with a function $d : X \times X \to [0, \infty)$, called distance, which is subject to the following conditions:

– $d(x, y) > 0$ for $x \neq y$, and $d(x, x) = 0$.

– $d(x, y) = (y, x)$.

– $d(x, y) \leq d(x, z) + d(y, z)$.

(2) Now let us try to check these axioms for $d(x, y) = ||x - y||$. The first axiom is clear, and so is the second axiom, so we are led with checking the third axiom, the triangle inequality one, which in practice means to establish the following inequality:

$$||x - y|| \leq ||x - z|| + ||y - z||$$

(3) But this is clear, because with $x' = x - z$ and $y' = z - y$, our estimate reads:

$$||x' + y'|| \leq ||x'|| + ||y'||$$

And this being the Minkowski inequality, done with the axiom check, as desired. □

The above result is quite important, because it shows that we can normally do geometry and analysis in our present setting, a bit as in the finite dimensional case. In order to do such abstract geometry, we will often need the following key result:

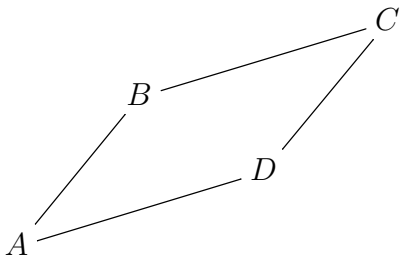
THEOREM 4.7. *The distances on H are subject to the identity*

$$||x+y||^2+||x-y||^2=2(||x||^2+||y||^2)$$

called parallelogram identity.

PROOF. This is something quite fundamental, the idea being as follows:

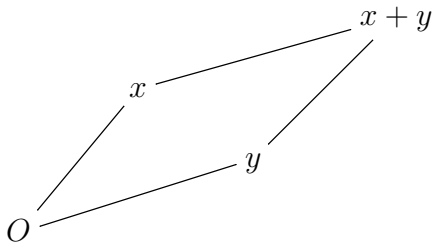
(1) To start with, there is a relation here with a basic result from plane geometry, that you might know or not. Consider indeed a parallelogram in the plane:



The above-mentioned formula from plane geometry is then as follows:

$$AC^2+BD^2=AB^2+BC^2+CD^2+DA^2$$

But this is more or less the formula in the statement. Indeed, if we choose the origin to be A, and relabel x,y the points B,D, our parallelogram becomes:



Now with this done, observe we have the following two formulae:

$$AC^2+BD^2=||x+y||^2+||x-y||^2$$

$$AB^2+BC^2+CD^2+DA^2=2(||x||^2+||y||^2)$$

Thus, the plane geometry formula is the same as the formula in the statement.

(2) In practice now, all this remains a mere remark, because our spaces are complex instead of real, have arbitrary dimension instead of 2, and also because we have not said in the above how the proof of the elementary geometry formula goes. So, better forget about all this, and try to prove the formula in the statement, from scratch.

(3) But here, things are in fact quite straightforward, because we have:

$$\begin{aligned} & ||x+y||^2+||x-y||^2 \\ = \ & <x+y,x+y>+<x-y,x-y> \\ = \ & ||x||^2+||y||^2+<x,y>+<y,x>+||x||^2+||y||^2-<x,y>-<y,x> \\ = \ & 2(||x||^2+||y||^2) \end{aligned}$$

Thus, we have proved our formula, and as a bonus, we have understood as well how the above-mentioned plane geometry formula works. Indeed, our computation above obviously works as well for the real scalar products, and this gives the result. □

As a second result now, which is something fundamental too, everything can be formally recovered in terms of distances, as follows:

THEOREM 4.8. *The scalar products can be recovered from distances, via the formula*

$$4<x,y>=||x+y||^2-||x-y||^2+i||x+iy||^2-i||x-iy||^2$$

called complex polarization identity.

PROOF. This is something that we have already met in finite dimensions. In arbitrary dimensions the proof is similar, as follows:

$$\begin{aligned} & ||x+y||^2-||x-y||^2+i||x+iy||^2-i||x-iy||^2 \\ = \ & ||x||^2+||y||^2-||x||^2-||y||^2+i||x||^2+i||y||^2-i||x||^2-i||y||^2 \\ & +2Re(<x,y>)+2Re(<x,y>)+2iIm(<x,y>)+2iIm(<x,y>) \\ = \ & 4<x,y> \end{aligned}$$

Thus, we are led to the conclusion in the statement. □

Summarizing, all the basic formulae involving scalar products and norms, that we know well from linear algebra, do hold in our abstract vector space setting. As a word of warning here, however, not ever to be forgotten, we have:

WARNING 4.9. *Unlike other things, the basic formula for real scalar products,*

$$<x,y>=||x||\cdot||y||\cdot\cos\alpha$$

does not hold, in our complex vector space setting.

To be more precise here, in what regards the above formula, you certainly know from plane geometry that the formula holds indeed for $\mathbb{R}^2$, and you might know too, from space geometry, that the formula holds as well for $\mathbb{R}^3$. The same goes for any $\mathbb{R}^N$, with similar proof, and going a bit abstract, for any real vector space coming with a scalar product, and this because by Cauchy-Schwarz we have $|<x,y>|\leq||x||\cdot||y||$, so the above formula can stand as a definition for the angle $\alpha\in[0,\pi)$ between our vectors x,y.

In the complex space setting, however, this does not work. Indeed, we still have the Cauchy-Schwarz inequality, telling us that $|<x,y>| \leq ||x|| \cdot ||y||$, but the scalar product being now a complex number, $<x,y> \in \mathbb{C}$, so is its quotient by $||x|| \cdot ||y|| \in \mathbb{R}$, so we cannot come with an angle $\alpha \in [0,\pi)$ whose cosine equals this quotient.

Nevermind. After all, Theorem 4.8 tells us that scalar products, and so the subsequent notion of angle, be that well-defined or not, are not really needed. We will often use this philosophy in what follows, with Theorem 4.8 standing as an answer to Warning 4.9.

4b. Hilbert spaces

In order to do analysis on our spaces, we need the Cauchy sequences that we construct to converge. This is something which is automatic in finite dimensions, but in arbitrary dimensions, this can fail. It is convenient here to formulate a detailed new definition, as follows, which will be the starting point for our various considerations to follow:

DEFINITION 4.10. *A Hilbert space is a complex vector space H given with a scalar product $<x,y>$, satisfying the following conditions:*

(1) $<x,y>$ *is linear in x, and antilinear in y.*
(2) $\overline{<x,y>} = <y,x>$, *for any x,y.*
(3) $<x,x> > 0$, *for any $x \neq 0$.*
(4) H *is complete with respect to the norm $||x|| = \sqrt{<x,x>}$.*

In other words, what we did here is to take Definition 4.1, and add the condition that H must be complete with respect to the norm $||x|| = \sqrt{<x,x>}$, that we know indeed to be a norm, according to the Minkowski inequality proved above. As a basic example, as before, we have the space $H = \mathbb{C}^N$, with its usual scalar product:

PROPOSITION 4.11. *The space $H = \mathbb{C}^N$, with its usual scalar product, namely*

$$<x,y> = \sum_i x_i \bar{y}_i$$

is a Hilbert space, which is finite dimensional.

PROOF. Here the fact that $<x,y> = \sum_i x_i \bar{y}_i$ is indeed a scalar product on $\mathbb{C}^N$ is something that we know well, and the completness condition is automatic. □

We will see later in this chapter, when talking about orthogonal bases for our spaces, that any finite dimensional Hilbert space H appears as above, $H \simeq \mathbb{C}^N$. Thus, at least we know one thing, done with finite dimensions, no bad surprises here.

More generally now, we have the following construction of Hilbert spaces:

PROPOSITION 4.12. *The sequences of numbers (x_i) which are square-summable,*

$$\sum_i |x_i|^2 < \infty$$

form a Hilbert space $l^2(\mathbb{N})$, with the following scalar product:

$$< x, y >= \sum_i x_i \bar{y}_i$$

In fact, given any index set I, we can construct a Hilbert space $l^2(I)$, in this way.

PROOF. There are several things to be proved, as follows:

(1) Our first claim is that $l^2(\mathbb{N})$ is a vector space. For this purpose, we must prove that $x, y \in l^2(\mathbb{N})$ implies $x + y \in l^2(\mathbb{N})$. But this leads us into proving $||x + y|| \leq ||x|| + ||y||$, where $||x|| = \sqrt{< x, x >}$. Now since we know this inequality to hold on each subspace $\mathbb{C}^N \subset l^2(\mathbb{N})$ obtained by truncating, this inequality holds everywhere, as desired.

(2) Our second claim is that $<, >$ is well-defined on $l^2(\mathbb{N})$. But this follows from the Cauchy-Schwarz inequality, $| < x, y > | \leq ||x|| \cdot ||y||$, which can be established by truncating, a bit like we established the Minkowski inequality in (1) above.

(3) It is also clear that $<, >$ is a scalar product on $l^2(\mathbb{N})$, so it remains to prove that $l^2(\mathbb{N})$ is complete with respect to $||x|| = \sqrt{< x, x >}$. But this is clear, because if we pick a Cauchy sequence $\{x^n\}_{n\in\mathbb{N}} \subset l^2(\mathbb{N})$, then each numeric sequence $\{x_i^n\}_{i\in\mathbb{N}} \subset \mathbb{C}$ is Cauchy, and by setting $x_i = \lim_{n\to\infty} x_i^n$, we have $x^n \to x$ inside $l^2(\mathbb{N})$, as desired.

(4) Finally, the same arguments extend to the case of an arbitrary index set I, leading to a Hilbert space $l^2(I)$, and with the remark here that there is absolutely no problem of taking about quantities of type $||x||^2 = \sum_{i\in I} |x_i|^2 \in [0, \infty]$, even if the index set I is uncountable, because we are summing positive numbers. □

Even more generally, we have the following construction of Hilbert spaces:

THEOREM 4.13. *Given a measured space X, the functions $f : X \to \mathbb{C}$, taken up to equality almost everywhere, which are square-summable,*

$$\int_X |f(x)|^2 dx < \infty$$

form a Hilbert space $L^2(X)$, with the following scalar product:

$$< f, g >= \int_X f(x)\overline{g(x)}dx$$

In the case $X = I$, with the counting measure, we obtain in this way the space $l^2(I)$.

PROOF. This is a straightforward generalization of Proposition 4.12, with the arguments from the proof of Proposition 4.12 carrying over in our case, as follows:

(1) The first part, regarding Cauchy-Schwarz and Minkowski, extends without problems, by using this time approximation by step functions.

(2) Regarding the fact that $<,>$ is indeed a scalar product on $L^2(X)$, there is a subtlety here, because if we want $<f,f>>0$ for $f\neq 0$, we must declare that $f=0$ when $f=0$ almost everywhere, and so that $f=g$ when $f=g$ almost everywhere.

(3) Regarding the fact that $L^2(X)$ is complete with respect to $||f||=\sqrt{<f,f>}$, this is again basic measure theory, by picking a Cauchy sequence $\{f_n\}_{n\in\mathbb{N}}\subset L^2(X)$, and then constructing a pointwise, and hence L^2 limit, $f_n\to f$, almost everywhere.

(4) Finally, the last assertion is clear, because the integration with respect to the counting measure is by definition a sum, and so $L^2(I)=l^2(I)$ in this case. □

As a conclusion to what we did so far, the Hilbert spaces are now axiomatized, and the main examples discussed. In order to do now some geometry on our spaces, in analogy with what we know from finite dimensions, let us start with the following definition:

DEFINITION 4.14. *Let H be a Hilbert space.*

(1) *We call two vectors orthogonal, $x\perp y$, when $<x,y>=0$.*
(2) *Given a subset $S\subset H$, we set $S^\perp=\big\{x\in H\big|x\perp y,\forall y\in S\big\}$.*

Here the first notion is something very familiar and intuitive, with the comment however that in the present complex space setting, orthogonality does not exactly mean that "we have a right angle between our vectors x,y", as explained in Warning 4.9.

In what regards (2), this is something very familiar too, and as an observation here, the subset $S^\perp\subset H$ constructed there is a closed linear space. In finite dimensions a useful, well-known formula here is $E^{\perp\perp}=E$, in case $E\subset H$ is a linear space. As explained below, this generalizes to the infinite dimensional setting as $E^{\perp\perp}=\bar{E}$.

Getting now to what can be done with orthogonality, we have here:

THEOREM 4.15. *Let H be a Hilbert space, and $E\subset H$ be a closed subspace.*

(1) *Given $x\in H$, we can find a unique $y\in E$, minimizing $||x-y||$.*
(2) *With x,y as above, we have $x=y+z$, for a certain $z\in E^\perp$.*
(3) *Thus, we have a direct sum decomposition $H=E\oplus E^\perp$.*
(4) *In terms of $H=E\oplus E^\perp$, the projection $x\to y$ is given by $P(x,y)=x$.*

PROOF. This is something very standard, the idea being as follows:

(1) Given $x \in H$ and two vectors $v, w \in E$, we have the following estimate:

$$\begin{aligned} ||x-v||^2 + ||x-w||^2 &= 2\left(\left|\left|x - \frac{v+w}{2}\right|\right|^2 + \left|\left|\frac{v-w}{2}\right|\right|^2\right) \\ &\geq 2d(x,E)^2 + \frac{||v-w||^2}{2} \end{aligned}$$

But this shows that any sequence in E realizing the inf in the definition of $d(x, E)$ is Cauchy, so it converges to a vector y. Since E is closed we have $y \in E$, so y realizes the inf. Moreover, again from the above inequality, such a y realizing the inf is unique.

(2) In order to prove $x-y \in E^\perp$, let $v \in E$ and choose $w \in \mathbb{T}$ such that $w < x-y, v >$ is a real number. For any $t \in \mathbb{R}$ we have the following equality:

$$||x-y+twv|| = ||x-y||^2 + 2tw < x-y, v > +t^2||v||^2$$

By construction of the vector y we know that this function has a minimum at $t = 0$. But this function is a degree 2 polynomial, so the middle term must vanish:

$$2w < x-y, v >= 0$$

Now since this must hold for any $v \in E$, we must have $x - y \in E^\perp$, as desired.

(3) This is consequence of what we found in (1,2).

(4) This is also a consequence of what we found in (1,2). □

Many things can be said, as a continuation of the above, as for instance with:

PROPOSITION 4.16. *For a closed subspace $E \subset H$, we have:*

$$E^{\perp\perp} = E$$

More generally, for an arbitrary linear subspace $E \subset H$, we have

$$E^{\perp\perp} = \bar{E}$$

and with the closing operation being needed, in infinite dimensions.

PROOF. All this comes indeed as an elementary application of our orthogonal projection technology from Theorem 4.15, and we will leave the details here as an exercise. □

Moving forward now, let us discuss some abstract aspects of the Hilbert spaces. You might know a bit, or not, about the Banach spaces, which are something more general than the Hilbert spaces. In view of this, our goal now will be to see what the general Banach space theory has to say, in the particular case of the Hilbert spaces.

And here, things are very simple, because we have, as a main result:

THEOREM 4.17. *Given a Hilbert space H and a closed subspace $E \subset H$, any linear form $f : E \to \mathbb{C}$ can be extended into a linear form*

$$\tilde{f} : H \to \mathbb{C}$$

having the same norm, and this by using $H = E \oplus E^\perp$, and setting $\tilde{f} = 0$ on $E^\perp$.

PROOF. This is indeed something self-explanatory. Observe that what we have here is the Hahn-Banach theorem, for the Hilbert spaces, coming with a trivial proof. □

Still talking abstract functional analysis, the few other basic Banach space results trivialize in the case of Hilbert spaces, as shown by the following result:

THEOREM 4.18. *Let H be a Hilbert space.*

(1) *Any linear form $f : H \to \mathbb{C}$ must be of type $f(y) =< z, y >$, with $z \in H$.*
(2) *Thus, we have a Banach space isomorphism $H^* \simeq \bar{H}$.*
(3) *In particular, H is reflexive as Banach space, $H^{**} = H$.*

PROOF. This is something that you might already know from Banach space theory, but we have an elementary proof for this, as follows:

(1) Consider a linear form $f : H \to \mathbb{C}$. Choose $v \in H$ such that $f(v) \neq 0$. By linearity we may assume $f(v) = 1$. Then each $z \in H$ decomposes in the following way:

$$z = (z - f(z)v) + f(z)v$$

This shows that we have a direct sum decomposition of H, as follows:

$$H = \ker(f) \oplus \mathbb{C}v$$

Now pick $z \in \ker(f)^\perp$ and consider the kernel of the linear form $f_z(y) =< z, y >$:

$$Ker(f_z) = \{y \in H | < z, y >= 0\} \supset Ker(f)$$

The linear forms f_z and f are then given by the following formulae:

$$f_z(a + \lambda v) = \lambda f_z(v) \quad , \quad f(a + \lambda v) = \lambda$$

It follows that we have $f = \mu f_z$, with $\mu = f_z(v)^{-1}$, and so that we have, as desired:

$$f = f_{\bar{\mu}z}$$

(2) This is just an abstract reformulation of what we found in (1).

(3) This follows from (2), because we have $H^{**} = \bar{H}^* = \bar{\bar{H}} = H$. □

As a conclusion to all this, which is really pleasant, the various general Banach space results are all clear in the Hilbert space setting. However, do not worry, the Hilbert spaces have their own amount of mystery, that we will explore in what follows.

4c. Bases, separability

At a more advanced level now, we can talk about orthonormal bases, and the related notion of dimension of a Hilbert space. However, this is something quite tricky, in the present infinite dimensional setting, that will take us some time to understand.

Let us start with the following result, that you surely know from linear algebras:

THEOREM 4.19. *Any system of linearly independent vectors $\{f_1, \ldots, f_n\}$ can be turned into an orthogonal system $\{e_1, \ldots, e_n\}$ by using the Gram-Schmidt procedure,*

$$\begin{aligned} e_1 &= f_1 \\ e_2 &= f_2 + \alpha_1 f_1 \\ e_3 &= f_3 + \beta_1 f_1 + \beta_2 f_2 \\ e_4 &= f_4 + \gamma_1 f_1 + \gamma_2 f_2 + \gamma_3 f_3 \\ &\vdots \end{aligned}$$

with the needed scalars $\alpha_i, \beta_i, \gamma_i, \ldots$ being uniquely determined.

PROOF. Many things can be said here, depending on how sharp you want to be, with the essentials of what is to be known being as follows:

(1) Let us first study the case $n = 2$. With $e_1 = f_1$ and $e_2 = f_2 + \alpha_1 f_1$ as in the statement, the needed orthogonality condition can be processed as follows:

$$\begin{aligned} e_1 \perp e_2 &\iff < f_1, f_2 + \alpha_1 f_1 >= 0 \\ &\iff \alpha_1 < f_1, f_1 >= - < f_1, f_2 > \\ &\iff \alpha_1 = -\frac{< f_1, f_2 >}{< f_1, f_1 >} \end{aligned}$$

Thus, we get our result, and with the remark that, alternatively, we can set:

$$e_2 = f_2 - Proj_{e_1}(f_2)$$

Indeed, with the above formula of α_1 in hand, the vector $e_2 = f_2 + \alpha_1 f_1$ that we get is precisely this one. Or, we can simply argue that this latter vector e_2 does the job, and with some basic linear algebra telling us that this vector e_2 is indeed unique.

(2) At $n = 3$ now, with e_1, e_2 already constructed, and with $e_3 = f_3 + \beta_1 f_1 + \beta_2 f_2$ as in the statement, the first orthogonality condition can be processed as follows:

$$\begin{aligned} e_1 \perp e_3 &\iff < f_1, f_3 + \beta_1 f_1 + \beta_2 f_2 >= 0 \\ &\iff \beta_1 < f_1, f_1 > + \beta_2 < f_1, f_2 >= - < f_1, f_3 > \end{aligned}$$

As for the second orthogonality condition, this can be now processed as follows:

$$\begin{aligned} e_2 \perp e_3 &\iff < f_2, f_3 + \beta_1 f_1 + \beta_2 f_2 >= 0 \\ &\iff \beta_1 < f_2, f_1 > + \beta_2 < f_2, f_2 >= - < f_2, f_3 > \end{aligned}$$

Thus, we are led to the following system, for the parameters β_1, β_2:

$$\beta_1 < f_1, f_1 > +\beta_2 < f_1, f_2 > = - < f_1, f_3 >$$

$$\beta_1 < f_2, f_1 > +\beta_2 < f_2, f_2 > = - < f_2, f_3 >$$

Now let us compute the determinant of this system. This is given by:

$$\begin{aligned} D &= \begin{vmatrix} < f_1, f_1 > & < f_1, f_2 > \\ < f_2, f_1 > & < f_2, f_2 > \end{vmatrix} \\ &= < f_1, f_1 >< f_2, f_2 > - < f_1, f_2 >< f_2, f_1 > \\ &= ||f_1||^2||f_2||^2 - | < f_1, f_2 > |^2 \end{aligned}$$

But this is exactly the quantity from the Cauchy-Schwarz inequality, so we have $D \geq 0$, with equality when f_1, f_2 are proportional. Now since f_1, f_2 were assumed to be linearly independent, we conclude that we have $D > 0$, so our system has indeed solutions.

(3) Alternatively, we can say at $n = 3$ that with the vectors e_1, e_2 being already constructed, we can construct the vector e_3 as follows, obviously doing the orthogonality job, and with its uniqueness coming from some standard linear algebra:

$$e_3 = f_3 - Proj_{e_1}(f_3) - Proj_{e_2}(f_3)$$

(4) Summarizing, we have two possible proofs for our result. Getting now to the general case, as a first proof, which is perhaps the most straightforward, we can set:

$$e_1 = f_1$$

$$e_2 = f_2 - Proj_{e_1}(f_2)$$

$$e_3 = f_3 - Proj_{e_1}(f_3) - Proj_{e_2}(f_3)$$

$$e_4 = f_4 - Proj_{e_1}(f_4) - Proj_{e_2}(f_4) - Proj_{e_3}(f_4)$$

$$\vdots$$

Indeed, these vectors do indeed the needed orthogonality job, and their uniqueness is clear too, via some basic linear algebra, that we will leave here as an exercise.

(5) Alternatively, by doing some explicit computations, as in (1) and (2), we must prove that a certain determinant is nonzero. To be more precise, at step $k + 1$ of the orthogonalization algorithm, the system to be solved is as follows:

$$x_1 < f_1, f_1 > +x_2 < f_1, f_2 > + \ldots + x_k < f_1, f_k > = - < f_1, f_{k+1} >$$

$$x_1 < f_2, f_1 > +x_2 < f_2, f_2 > + \ldots + x_k < f_2, f_k > = - < f_2, f_{k+1} >$$

$$\vdots$$

$$x_1 < f_k, f_1 > +x_2 < f_k, f_2 > + \ldots + x_k < f_k, f_k > = - < f_k, f_{k+1} >$$

Thus, the determinant to be studied, in order to prove that our system has indeed solutions, is the Gram determinant of $f_1, \ldots, f_k$, given by the following formula:

$$D_k = \begin{vmatrix} < f_1, f_1 > & < f_1, f_2 > & \ldots & < f_1, f_k > \\ < f_2, f_1 > & < f_2, f_2 > & \ldots & < f_2, f_k > \\ & \vdots & \vdots & \\ < f_k, f_1 > & < f_k, f_2 > & \ldots & < f_k, f_k > \end{vmatrix}$$

(6) Now in relation with this latter question, we have already seen in (2) that we have $D_2 > 0$, but with this being something quite complicated, coming from Cauchy-Schwarz. So, not very good news, but fortunately, linear algebra comes to the rescue. Consider the square matrix formed by our vectors $f_1, \ldots, f_k$, arranged horizontally, as follows:

$$F = \begin{pmatrix} (f_1)_1 & \ldots & (f_1)_k \\ & \vdots & \\ (f_k)_1 & \ldots & (f_k)_k \end{pmatrix}$$

We have then the following computation, for any two indices i, j:

$$\begin{aligned} (FF^*)_{ij} &= \sum_l F_{il}(F^*)_{lj} \\ &= \sum_l F_{il}\bar{F}_{jl} \\ &= \sum_l (f_i)_l\overline{(f_j)_l} \\ &= < f_i, f_j > \end{aligned}$$

We conclude that at the matrix level, we have the following formula:

$$FF^* = \begin{pmatrix} < f_1, f_1 > & < f_1, f_2 > & \ldots & < f_1, f_k > \\ < f_2, f_1 > & < f_2, f_2 > & \ldots & < f_2, f_k > \\ & \vdots & \vdots & \\ < f_k, f_1 > & < f_k, f_2 > & \ldots & < f_k, f_k > \end{pmatrix}$$

Thus, at the level of the corresponding determinants we obtain, as desired:

$$D_k = \det(FF^*) = |\det F|^2 > 0$$

(7) Finally, and getting back now to the system, we can work out some explicit formulae for e_i, alternative to those in (4), based on this. To be more precise, we have:

$$e_k = \frac{1}{D_{k-1}} \begin{vmatrix} < f_1, f_1 > & < f_1, f_2 > & \ldots & < f_1, f_k > \\ < f_2, f_1 > & < f_2, f_2 > & \ldots & < f_2, f_k > \\ & \vdots & \vdots & \\ < f_{k-1}, f_1 > & < f_{k-1}, f_2 > & \ldots & < f_{k-1}, f_k > \\ f_1 & f_2 & \ldots & f_k \end{vmatrix}$$

And we will leave some illustrations here as an instructive exercise, and please do better than my students, who usually stop after 2-3 steps. □

Getting back now to our Hilbert space questions, we have the following result:

THEOREM 4.20. *Any Hilbert space H has an orthonormal basis $\{e_i\}_{i\in I}$, which is by definition a set of vectors whose span is dense in H, and which satisfy*

$$< e_i, e_j >= \delta_{ij}$$

with δ being a Kronecker symbol. The cardinality $|I|$ of the index set, which can be finite, countable, or uncountable, depends only on H, and is called dimension of H. We have

$$H \simeq l^2(I)$$

in the obvious way, mapping $\sum \lambda_i e_i \to (\lambda_i)$. The Hilbert spaces with $\dim H = |I|$ being countable, such as $l^2(\mathbb{N})$, are all isomorphic, and are called separable.

PROOF. We have many assertions here, the idea being as follows:

(1) In finite dimensions an orthonormal basis $\{e_i\}_{i\in I}$ can be constructed by starting with any vector space basis $\{f_i\}_{i\in I}$, and using the Gram-Schmidt procedure. As for the other assertions, these are all clear, from basic linear algebra.

(2) In general, the same method works, namely Gram-Schmidt, with a subtlety coming from the fact that the basis $\{e_i\}_{i\in I}$ will not span in general the whole H, but just a dense subspace of it, as it is in fact obvious by looking at the standard basis of $l^2(\mathbb{N})$.

(3) And there is a second subtlety as well, coming from the fact that the recurrence procedure needed for Gram-Schmidt must be replaced by some sort of "transfinite recurrence", using standard tools from logic, and more specifically the Zorn lemma.

(4) Finally, everything at the end, regarding our notion of separability for the Hilbert spaces, is clear from definitions, and from our various results above. □

So long for abstract Hilbert space questions, and orthonormal bases, and many other things can be said here. In practice now, and getting to the essentials, according to Theorem 4.20, there is only one separable Hilbert space, up to isomorphism.

As a first result regarding this unique space that we are interested in, we have:

THEOREM 4.21. *The following happen, in relation with separability:*

(1) *The Hilbert space $H = L^2[-1, 1]$ is separable, with orthonormal basis coming by applying Gram-Schmidt to the basis $\{x^k\}_{k\in\mathbb{N}}$, coming from Weierstrass.*
(2) *In fact, any $H = L^2(\mathbb{R}, \mu)$, with $d\mu(x) = f(x)dx$, is separable, and the same happens in higher dimensions, for $H = L^2(\mathbb{R}^N, \mu)$, with $d\mu(x) = f(x)dx$.*
(3) *More generally, given a separable abstract measured space X, the associated Hilbert space of square-summable functions $H = L^2(X)$ is separable.*

PROOF. Many things can be said here, the idea being as follows:

(1) The fact that $H = L^2[-1, 1]$ is separable is clear indeed from the Weierstrass density theorem, which provides us with the algebraic basis $g_k = x^k$, which can be orthogonalized by using the Gram-Schmidt procedure, as explained in Theorem 4.20.

(2) Regarding now more general spaces, of type $H = L^2(\mathbb{R}, \mu)$, we can use here the same argument, after modifying if needed our measure μ, in order for the functions $g_k = x^k$ to be indeed square-summable. As for higher dimensions, the situation here is similar, because we can use here the multivariable polynomials $g_k(x) = x_1^{k_1} \ldots x_N^{k_N}$.

(3) Finally, the last assertion, regarding the general spaces of type $H = L^2(X)$, which generalizes all this, comes as a consequence of general measure theory, and we will leave working out the details here as an instructive exercise. □

As a conclusion to all this, which is a bit philosophical, we have:

CONCLUSION 4.22. *We are interested in one space, namely the unique separable Hilbert space H, but due to various technical reasons, it is often better to forget that we have $H = l^2(\mathbb{N})$, and say instead that we have $H = L^2(X)$, with X being a separable measured space, or simply say that H is an abstract separable Hilbert space.*

It is also possible to make some physics comments here, with this unique separable Hilbert space being, and no surprise here, the space that we live in.

4d. Orthogonal polynomials

Let us go back now to Theorem 4.20 and its proof, which was something quite subtle. That material leads us into orthogonal polynomials, which are defined as follows:

DEFINITION 4.23. *The orthogonal polynomials with respect to $d\mu(x) = f(x)dx$ are polynomials $P_k \in \mathbb{R}[x]$ of degree $k \in \mathbb{N}$, which are orthogonal inside $H = L^2(\mathbb{R}, \mu)$:*

$$\int_{\mathbb{R}} P_k(x)P_l(x)f(x)dx = 0 \quad , \quad \forall k \neq l$$

Equivalently, these orthogonal polynomials $\{P_k\}_{k\in\mathbb{N}}$, which are each unique modulo scalars, appear from the Weierstrass basis $\{x^k\}_{k\in\mathbb{N}}$, by doing Gram-Schmidt.

Observe that the orthogonal polynomials exist indeed for any real measure $d\mu(x) = f(x)dx$, as explained above. It is possible to be a bit more explicit here, as follows:

THEOREM 4.24. *The orthogonal polynomials with respect to μ are given by*

$$P_k = c_k \begin{vmatrix} M_0 & M_1 & \ldots & M_k \\ M_1 & M_2 & \ldots & M_{k+1} \\ \vdots & \vdots & & \vdots \\ M_{k-1} & M_k & \ldots & M_{2k-1} \\ 1 & x & \ldots & x^k \end{vmatrix}$$

where $M_k = \int_{\mathbb{R}} x^k d\mu(x)$ are the moments of μ, and $c_k \in \mathbb{R}^*$ can be any numbers.

PROOF. Let us first see what happens at small values of $k \in \mathbb{N}$. At $k = 0$ our formula is as follows, stating that the first polynomial P_0 must be a constant, as it should:

$$P_0 = c_0|M_0| = c_0$$

At $k = 1$ now, again by using $M_0 = 1$, the formula is as follows:

$$P_1 = c_1 \begin{vmatrix} M_0 & M_1 \\ 1 & x \end{vmatrix} = c_1(x - M_1)$$

But this is again the good formula, because the degree is 1, and we have:

$$\begin{aligned} < 1, P_1 > &= c_1 < 1, x - M_1 > \\ &= c_1(< 1, x > - < 1, M_1 >) \\ &= c_1(M_1 - M_1) \\ &= 0 \end{aligned}$$

At $k = 2$ now, things get more complicated, with the formula being as follows:

$$P_2 = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ 1 & x & x^2 \end{vmatrix}$$

However, no need for big computations here, in order to check the orthogonality, because by using the fact that x^k integrates up to M_k, we obtain:

$$< 1, P_2 >= \int_{\mathbb{R}} P_2(x) d\mu(x) = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_0 & M_1 & M_2 \end{vmatrix} = 0$$

Similarly, again by using the fact that x^k integrates up to M_k, we have as well:

$$< x, P_2 >= \int_{\mathbb{R}} x P_2(x) d\mu(x) = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_1 & M_2 & M_3 \end{vmatrix} = 0$$

Thus, result proved at $k = 0, 1, 2$, and the proof in general is similar. □

In practice now, all this leads us to a lot of interesting combinatorics, and countless things can be said. For the simplest measured space $X \subset \mathbb{R}$, which is the interval $[-1,1]$, with its uniform measure, the orthogonal basis problem can be solved as follows:

THEOREM 4.25. *The orthonormal polynomials for $L^2[-1,1]$, subject to*

$$\int_{-1}^{1} P_k(x)P_l(x)\,dx = \delta_{kl}$$

and called Legendre polynomials, satisfy the following differential equation,

$$(1-x^2)P_k''(x) - 2xP_k'(x) + k(k+1)P_k(x) = 0$$

which is the Legendre equation from physics. Moreover, we have the formula

$$(k+1)P_{k+1}(x) = (2k+1)xP_k(x) - kP_{k-1}(x)$$

called Bonnet recurrence formula, as well as the formula

$$P_k(x) = \frac{1}{2^k k!} \cdot \frac{d^k}{dx^k}\left(1-x^2\right)^k$$

called Rodrigues formula for the Legendre polynomials.

PROOF. As a first observation, we are not lost somewhere in abstract math, because of the occurrence of the Legendre equation. As for the proof, this goes as follows:

(1) The first assertion is clear, because the Gram-Schmidt procedure applied to the Weierstrass basis $\{x^k\}$ can only lead to a certain family of polynomials $\{P_k\}$, with each P_k being of degree k, and also unique, if we assume that it has positive leading coefficient, with this $\pm$ choice being needed, as usual, at each step of Gram-Schmidt.

(2) In order to have now an idea about these beasts, here are the first few of them, which can be obtained say via a straightforward application of Gram-Schmidt:

$$\begin{aligned}
P_0 &= 1\\
P_1 &= x\\
P_2 &= (3x^2-1)/2\\
P_3 &= (5x^3-3x)/2\\
P_4 &= (35x^4-30x^2+3)/8\\
P_5 &= (63x^5-70x^3+15x)/8
\end{aligned}$$

(3) Now thinking about what Gram-Schmidt does, this is certainly something by recurrence. And examining the recurrence leads to the Legendre equation, as stated. As for the Bonnet recurrence formula, the story here is similar.

(4) Regarding the Rodrigues formula, by uniqueness no need to try to understand where this formula comes from, and we have two choices here, either by verifying that $\{P_k\}$ is orthonormal, or by verifying the Legendre equation. And both methods work. □

The above result is just the tip of the iceberg, and as a continuation, we have:

THEOREM 4.26. *The orthogonal polynomials for $L^2[-1,1]$, with measure*

$$d\mu(x) = (1-x)^a(1+x)^b dx$$

called Jacobi polynomials, satisfy as well a degree 2 equation, namely

$$(1-x^2)P_k''(x) + (b-a-(a+b+2)x)P_k'(x) + k(k+a+b+1)P_k(x) = 0$$

as well as an order 2 recurrence relation, and are given by the following formula:

$$P_k(x) = \frac{(-1)^k}{2^k k!}(1-x)^{-a}(1+x)^{-b}\frac{d^k}{dx^k}\left[(1-x)^a(1+x)^b(1-x^2)^k\right]$$

At $a=b=0$ we recover the Legendre polynomials, and at $a=b=\pm\frac{1}{2}$ we recover the Chebycheff polynomials of the first and second kind, from trigonometry.

PROOF. There are many things going on here, the idea being as follows:

(1) To start with, in what regards the precise statement, the order 2 recurrence relation mentioned there is something quite complicated, as follows:

$$\begin{aligned} & 2k(k+a+b)(2k+a+b-2)P_k(x) \\ = \ & (2k+a+b-1)\left[(2k+a+b)(2k+a+b-2)x+a^2-b^2\right]P_{k-1}(x) \\ - \ & 2(k+a-1)(k+b-1)(2k+a+b)P_{k-2}(x) \end{aligned}$$

(2) Regarding now the proof, the statement itself appears as a generalization of Theorem 4.25, which corresponds to the particular case $a=b=0$, and the proof is quite similar. We will leave learning more about all this as an interesting exercise.

(3) For completness, let us record as well a few numerics, as follows:

$$\begin{aligned} P_0 &= 1 \\ P_1 &= (a+1)+(a+b+2)\,\frac{x-1}{2} \\ P_2 &= \frac{(a+1)(a+2)}{2}+(a+2)(a+b+3)\,\frac{x-1}{2} \\ &+ \frac{(a+b+3)(a+b+4)}{2}\left(\frac{x-1}{2}\right)^2 \end{aligned}$$

(4) Regarding now the main particular cases of the Jacobi polynomials, these are the Gegenbauer polynomials, appearing at $a=b$. However, there is not that much of a simplification when passing from general parameters a,b to equal parameters, $a=b$, so in practice, the main particular cases are those indicated in the statement, namely:

– The Legendre polynomials, that we know well from Theorem 4.25, appearing at the simplest values of the parameters, namely $a=b=0$.

– The Chebycheff polynomials of the first kind T_k, which are given by the formula $T_k(\cos t) = \cos(kt)$ from trigonometry, appearing at $a = b = -\frac{1}{2}$.

– The Chebycheff polynomials of the second kind U_k, which are given by the formula $U_k(\cos t)\sin t = \sin((k+1)t)$, appearing at $a = b = \frac{1}{2}$.

(5) So, this was for the story of the Jacobi polynomials, and their main particular cases, and in practice, we will leave some further learning here as an exercise, coming as a continuation of the further learning of Theorem 4.25, and its details. □

Getting now to other spaces $X \subset \mathbb{R}$, of particular interest here is the following result, which complements well Theorem 4.25, for the needs of basic quantum mechanics:

THEOREM 4.27. *The orthogonal polynomials for $L^2[0,\infty)$, with scalar product*

$$< f, g >= \int_0^\infty f(x)g(x)e^{-x}\,dx$$

are the Laguerre polynomials $\{P_k\}$, satisfying the following differential equation,

$$xP_k''(x) + (1-x)P_k'(x) + kP_k(x) = 0$$

as well as the following order 2 recurrence relation,

$$(k+1)P_{k+1}(x) = (2k+1-x)P_k(x) - kP_{k-1}(x)$$

and which are given by the following formula,

$$P_k(x) = \frac{e^x}{k!} \cdot \frac{d^k}{dx^k}\left(e^{-x}x^k\right)$$

called Rodrigues formula for the Laguerre polynomials.

PROOF. The story here is very similar to that of the Legendre and Jacobi polynomials, and many further things can be said here, with exercise for you to learn a bit about all this. Let us record as well a few numeric values, for the Laguerre polynomials:

$$\begin{aligned}
P_0 &= 1\\
P_1 &= 1-x\\
P_2 &= (x^2-4x+2)/2\\
P_3 &= (-x^3+9x^2-18x+6)/6\\
P_4 &= (x^4-16x^3+72x^2-96x+24)/24
\end{aligned}$$

Finally, for the story to be complete, no discussion about the Laguerre polynomials would be complete without a word about their use, in quantum mechanics. And here, as usual, we will leave some exploration of this as an instructive exercise. □

Finally, regarding the space $X = \mathbb{R}$ itself, we have here the following result:

THEOREM 4.28. *The orthogonal polynomials for $L^2(\mathbb{R})$, with scalar product*

$$< f, g >= \int_0^\infty f(x)g(x)e^{-x^2}\,dx$$

are the Hermite polynomials $\{P_k\}$, satisfying the following differential equation,

$$P_k''(x) - 2xP_k'(x) + P_k(x) = 0$$

as well as the following order 2 recurrence relation,

$$P_{k+1}(x) = 2xP_k(x) - 2kP_{k-1}(x)$$

and which are given by the following formula,

$$P_k(x) = (-1)^k e^{x^2} \cdot \frac{d^k}{dx^k}\left(e^{-x^2}\right)$$

called Rodrigues formula for the Hermite polynomials.

PROOF. As before, the story here is quite similar to that of the Legendre and other orthogonal polynomials, and exercise for you to learn a bit about all this. Let us record as well a few numeric values, for the Hermite polynomials:

$$\begin{aligned}
P_0 &= 1\\
P_1 &= 2x\\
P_2 &= 4x^2 - 2\\
P_3 &= 8x^3 - 12x\\
P_4 &= 16x^4 - 48x^2 + 12\\
P_5 &= 32x^5 - 160x^3 + 120x\\
P_6 &= 64x^6 - 480x^4 + 720x^2 - 120
\end{aligned}$$

With of course, exercise for you to deduce all these formulae. □

And with this, good news, end of the story with the orthogonal polynomials, at least at the very introductory level, and this due to the following fact, which is something quite technical, and that we will not attempt to prove, or even explain in detail here:

FACT 4.29. *From an abstract point of view, coming from degree 2 equations, and Rodrigues formulae for the solutions, there are only three types of "classical" orthogonal polynomials, namely the Jacobi, Laguerre and Hermite ones, discussed above.*

Finally, as already mentioned, the above results are very useful in the context of basic quantum mechanics, and more specifically, for solving the hydrogen atom, following Heisenberg and Schrödinger. Again, exercise for you to learn a bit about this.

4e. Exercises

Exercises:

Exercise 4.30.

Exercise 4.31.

Exercise 4.32.

Exercise 4.33.

Exercise 4.34.

Exercise 4.35.

Exercise 4.36.

Exercise 4.37.

Bonus exercise.

Part II

General theory

CHAPTER 5

Linear equations

5a. Linear equations

In order to discuss order and chaos, in the context of classical mechanics, let us start with some abstract mathematics. Here is a good, concrete question, which appears in mathematics, physics, and related disciplines, that we would like to solve:

QUESTION 5.1. *How to solve differential equations?*

Obviously, this question is quite broad, and as a first concrete example, let us examine the case of a falling object. If we denote by $x = x(t) : \mathbb{R} \to \mathbb{R}^3$ the position of our falling object, then its speed $v = v(t) : \mathbb{R} \to \mathbb{R}^3$ and acceleration $a = a(t) : \mathbb{R} \to \mathbb{R}^3$ are given by the following formulae, with the dots standing for derivatives with respect to time t:

$$v = \dot{x} \quad , \quad a = \dot{v} = \ddot{x}$$

Regarding now the equation of motion, this is as follows, coming from Newton, with m being the mass of our object, and with F being the gravitational force:

$$m \cdot a(t) = F(x(t))$$

Thus, in terms of derivatives as above, in order to have as only unknown the position vector $x = x(t) : \mathbb{R} \to \mathbb{R}^3$, the equation of motion is as follows:

$$m \cdot \ddot{x}(t) = F(x(t))$$

Which looks nice, but since what we have here is a degree 2 equation, instead of degree 1, which would be better, was it really a good idea to get rid of speed $v : \mathbb{R} \to \mathbb{R}^3$ and acceleration $a : \mathbb{R} \to \mathbb{R}^3$, and reformulate everything in terms of position $x : \mathbb{R} \to \mathbb{R}^3$.

Nevermind. So going all over again, with the aim this time of reaching to a degree 1 equation, let us replace our 3-dimensional unknown $x : \mathbb{R} \to \mathbb{R}^3$ with the 6-dimensional unknown $(x, v) : \mathbb{R} \to \mathbb{R}^6$. And with this done, surprise, we have our degree 1 system:

$$\begin{cases} \dot{x}(t) = v(t) \\ \dot{v}(t) = \frac{1}{m} F(x(t)) \end{cases}$$

Which was a nice trick, wasn't it. So, before going further, let us record the following conclusion, that we will come back to in a moment, after done with gravity:

CONCLUSION 5.2. *We can convert differential equations of higher order into differential equations of first order, by suitably enlarging the size of our unknown vectors.*

Now back to gravity and free falls, and to the degree 1 system found above, we will assume in what follows that our object is subject to a free fall under a uniform gravitational field. In practice, this means that the force F is given by the following formula, with $m > 0$ being as usual the mass of our object, and with $g > 0$ being a certain constant:

$$F(x) = -mg \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

With this data, the system that we found takes the following form:

$$\begin{cases} \dot{x}(t) = v(t) \\ \\ \dot{v}(t) = -g \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \end{cases}$$

But this latter system is very easy to solve. Indeed, the second equation gives:

$$v(t) = v(0) - g \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t$$

Now by integrating once again, we can recover as well the formula of x, as follows:

$$x(t) = x(0) + v(0)t - \frac{g}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t^2$$

Which is very nice, good work that we did here, so let us record our findings, along with a bit more, in the form of a complete statement, as follows:

THEOREM 5.3. *For a free fall in a uniform gravitational field, with gravitational acceleration constant $g > 0$, the equation of motion is*

$$x(t) = x(0) + v(0)t - \frac{g}{2} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} t^2$$

and the trajectory is a parabola, unless in the case where the free fall is straight downwards, where the trajectory is a line.

PROOF. This is a conclusion to what we found above, namely equation of motion, and its obvious implications, and the level of the corresponding trajectory. □

Now back to theory, let us go back to Conclusion 5.2, which was our main theoretical finding so far, and further comment on that. Of course in the case of extremely simple equations, like the above uniform gravity ones, there is no really need to use this trick, because you can directly integrate twice, and so on. However, in general, this remains a very useful trick, worth some discussion, and we will discuss this now.

Let us start with some generalities in one variable. We have here:

DEFINITION 5.4. *A general ordinary differential equation (ODE) is an equation as follows, with a function $x = x(t) : \mathbb{R} \to \mathbb{R}$ as unknown,*

$$F(t, x, \dot{x}, \ldots, x^{(k)}) = 0$$

depending on a given function $F : U \to \mathbb{R}$, with $U \subset \mathbb{R}^{k+2}$ being an open set.

As a first observation, under suitable assumptions on our function $F : U \to \mathbb{R}$, and more specifically non-vanishing of its partial derivatives, in all directions, we can use the implicit function theorem, in order to reformulate our equation as follows, for a certain function $f : V \to \mathbb{R}$, with $V \subset \mathbb{R}^{k+1}$ being a certain open set:

$$x^{(k)} = f(t, x, \dot{x}, \ldots, x^{(k-1)})$$

In practice, we will make this change, which often comes by default, when investigating questions coming from physics, and these will be the ODE that we will be interested in.

Now moving to several variables, more generally, let us formulate:

DEFINITION 5.5. *A standard system of ODE is a system as follows,*

$$x_1^{(k)} = f_1(t, x, \dot{x}, \ldots, x^{(k-1)})$$

$$\vdots$$

$$x_N^{(k)} = f_N(t, x, \dot{x}, \ldots, x^{(k-1)})$$

with the unknown being a vector function $x = x(t) : \mathbb{R} \to \mathbb{R}^N$.

Here the adjective "standard" refers to the implicit function theorem manipulation made in the above, which can be of course made in the context of several variables too.

Now with these abstract definitions in hand, we can go back to Conclusion 5.2, and formulate a more precise version of that observation, as follows:

THEOREM 5.6. *We can convert any standard system of ODE into a standard order 1 system of ODE, by suitably enlarging the size of the unknown vector.*

PROOF. This is indeed clear from definitions, because with $y = (x, \dot{x}, \ldots, x^{(k-1)})$, in the context of Definition 5.1, the system there takes the following form, as desired:

$$\begin{aligned}
\dot{y}_1 &= y_2\\
\dot{y}_2 &= y_3\\
&\vdots\\
\dot{y}_{k-1} &= y_k\\
\dot{y}_k &= f(t, y)
\end{aligned}$$

Thus, we are led to the conclusion in the statement. There are of course many explicit applications of this method, and further comments that can be made. □

We will be back to this in chapter 6 below, when investigating more in detail the ODE, and the related notion of dynamical systems, by using methods from linear algebra.

Getting now to the point where we wanted to get, in order to get truly started with all this, with some mathematics going on, let us have a look at the systems of ODE which are linear. That is, we would like to solve equations as follows, with f_i being linear:

$$\begin{aligned}
x_1^{(k)} &= f_1(t, x, \dot{x}, \ldots, x^{(k-1)})\\
&\vdots\\
x_N^{(k)} &= f_N(t, x, \dot{x}, \ldots, x^{(k-1)})
\end{aligned}$$

By doing the manipulation in Theorem 5.6, and assuming that we are in the "autonomous" case, where there is no time t in our linear function which produces the system, we are led to a vector equation as follows, with $A \in M_N(\mathbb{R})$ being a certain matrix:

$$x' = Ax$$

But here, we are in familiar territory, namely very standard calculus, because in the 1D case, the solution simply appears by exponentiating, as follows:

$$x = e^{tA}x_0$$

Which is something very nice, and with this understood, we can now go back to our original Question 5.1, from the beginning of this chapter. As already mentioned, that question was something very broad, and as something more concrete now, we have:

QUESTION 5.7. *The solution of a system of linear differential equations,*

$$x' = Ax \quad , \quad x(0) = x_0$$

with $A \in M_N(\mathbb{R})$, is normally given by $x = e^{tA}x_0$, and this because we should have:

$$(e^{tA}x_0)' = Ae^{tA}x_0$$

But, what exactly is e^{tA}, and then, importantly, how to explicitly compute e^{tA}?

To be more precise, again as with Question 5.1, this question appears indeed in a myriad contexts, all across physics and science, and with all this needing no further presentation. Observe also that, due to Theorem 5.6, this question allows us to deal with differential equations of higher order too, by enlarging the size of our vectors.

5b. Matrix exponential

So, let us attempt to solve Question 5.7. As a first task, and forgetting now about time t and differential equations, we would like to talk about exponentials of matrices. But here, the answer can only be given by the following formula:

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

Which leads us into analysis over $M_N(\mathbb{R})$, or over $M_N(\mathbb{C})$, if we want to deal directly with the complex case. So, getting started with our study, let us begin with:

THEOREM 5.8. *The following quantity, with sup over the norm 1 vectors,*

$$||A|| = \sup_{||x||=1} ||Ax||$$

where $||x|| = \sqrt{\sum |x_i|^2}$ as usual, is a norm on the space of matrices $M_N(\mathbb{C})$.

PROOF. This is indeed clear from definitions. In fact, we already saw such things in Part I, in an indirect form, when talking about density results inside $M_N(\mathbb{C})$. Note also that $M_N(\mathbb{C})$ being finite dimensional, all possible norms on it are equivalent. □

Now with the above result in hand, we can do analysis over $M_N(\mathbb{C})$, and in particular we can investigate our exponentiation problem, with the following conclusions:

THEOREM 5.9. *We can talk about the exponentials of matrices $A \in M_N(\mathbb{C})$, given by*

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

and these exponentials have the following basic properties:

(1) $||e^A|| \leq e^{||A||}$.
(2) *If $D = diag(\lambda_i)$ then $e^D = diag(e^{\lambda_i})$.*
(3) *If P is invertible, $e^{PDP^{-1}} = Pe^DP^{-1}$.*
(4) *If $A = PDP^{-1}$ with $D = diag(\lambda_i)$, then $e^A = Pdiag(e^{\lambda_i})P^{-1}$.*

PROOF. The fact that our exponential series converges indeed follows from (1), so we are left with proving (1-4), and this can be done as follows:

(1) We have indeed the following computation, using the various properties of the norm, and notably the formula $||AB|| \leq ||A|| \cdot ||B||$, which is clear from definitions:

$$\begin{aligned}
||e^A|| &= \left|\left|\sum_{k=0}^{\infty} \frac{A^k}{k!}\right|\right| \\
&\leq \sum_{k=0}^{\infty} \left|\left|\frac{A^k}{k!}\right|\right| \\
&= \sum_{k=0}^{\infty} \frac{||A^k||}{k!} \\
&\leq \sum_{k=0}^{\infty} \frac{||A||^k}{k!} \\
&= e^{||A||}
\end{aligned}$$

(2) This is clear from definitions, with the computation being as follows:

$$\begin{aligned}
\exp\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix} &= \sum_{k=0}^{\infty} \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}^k \Big/ k! \\
&= \sum_{k=0}^{\infty} \begin{pmatrix}\lambda_1^k & & \\ & \ddots & \\ & & \lambda_N^k\end{pmatrix} \Big/ k! \\
&= \begin{pmatrix}\sum_{k=0}^{\infty} \lambda_1^k/k! & & \\ & \ddots & \\ & & \sum_{k=0}^{\infty} \lambda_N^k/k!\end{pmatrix} \\
&= \begin{pmatrix}e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_N}\end{pmatrix}
\end{aligned}$$

(3) Again, this is clear from definitions, the computation being as follows:

$$\begin{aligned}
e^{PDP^{-1}} &= \sum_{k=0}^{\infty}\frac{(PDP^{-1})^k}{k!}\\
&= \sum_{k=0}^{\infty}\frac{PDP^{-1}\cdot PDP^{-1}\ldots PDP^{-1}}{k!}\\
&= \sum_{k=0}^{\infty}\frac{PD^kP^{-1}}{k!}\\
&= P\left(\sum_{k=0}^{\infty}\frac{D^k}{k!}\right)P^{-1}\\
&= Pe^DP^{-1}
\end{aligned}$$

(4) This follows indeed by combining (2) and (3). □

As a consequence of our theory, we can now state, in relation with Question 5.7:

THEOREM 5.10. *Given a matrix $A\in M_N(\mathbb{C})$, the vector function*

$$x=e^{tA}x_0$$

satisfies the system of linear differential equations $x'=Ax$, $x(0)=x_0$.

PROOF. In what regards the first formula, this comes from:

$$\begin{aligned}
x' &= (e^{tA}x_0)'\\
&= \left(\sum_{k=0}^{\infty}\frac{(tA)^kx_0}{k!}\right)'\\
&= \sum_{k=0}^{\infty}\frac{kt^{k-1}A^kx_0}{k!}\\
&= A\sum_{k=1}^{\infty}\frac{t^{k-1}A^{k-1}x_0}{(k-1)!}\\
&= A\sum_{l=0}^{\infty}\frac{t^lA^lx_0}{l!}\\
&= Ae^{tA}x_0\\
&= Ax
\end{aligned}$$

As for the second formula, this is clear from $e^{0_N}=1_N$, that is, from the fact that the exponential of the null $N\times N$ matrix is the identity $N\times N$ matrix. □

As a key result now, which shows that things are certainly more complicated with matrices than with real numbers, when talking exponentials, we have:

THEOREM 5.11. *We have the following formula, when A, B commute:*

$$e^{A+B} = e^A e^B$$

When the matrices A, B do not commute, this formula might fail.

PROOF. We have two assertions here, the idea being as follows:

(1) As a first observation, when two matrices A, B commute we can compute the powers $(A+B)^k$ as for the usual numbers, and we get a binomial formula, namely:

$$\begin{aligned}(A+B)^k &= (A+B)(A+B)\ldots(A+B)\\ &= A^k + kA^{k-1}B + \ldots + kAB^{k-1} + B^k\\ &= \sum_{r=0}^{k}\binom{k}{r}A^rB^{k-r}\end{aligned}$$

Now by using this binomial formula for A, B we obtain, as for the usual numbers:

$$\begin{aligned}e^{A+B} &= \sum_{k=0}^{\infty}\frac{(A+B)^k}{k!}\\ &= \sum_{k=0}^{\infty}\sum_{r=0}^{k}\binom{k}{r}\frac{A^rB^{k-r}}{k!}\\ &= \sum_{k=0}^{\infty}\sum_{r=0}^{k}\frac{A^rB^{k-r}}{r!(k-r)!}\\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{A^rB^s}{r!s!}\\ &= \sum_{r=0}^{\infty}\frac{A^r}{r!}\sum_{s=0}^{\infty}\frac{B^s}{s!}\\ &= e^Ae^B\end{aligned}$$

(2) In order to find now a counterexample to $e^{A+B} = e^Ae^B$, we need some matrices which do not commute, $AB \neq BA$, and the simplest such matrices are as follows:

$$J = \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix} \quad , \quad J^* = \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}$$

Indeed, the products of two matrices are given by the following formulae:

$$JJ^* = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} \quad , \quad J^*J = \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$$

Now observe that, since these two products are both diagonal, we can compute right away their exponentials, and we are led to the following conclusion:

$$e^{JJ^*}=\begin{pmatrix}e&0\\0&0\end{pmatrix}\neq\begin{pmatrix}0&0\\0&e\end{pmatrix}=e^{J^*J}$$

Thus, we have a counterexample to $e^{AB}=e^{BA}$, but bad luck, this being not the counterexample we were looking for, still some work to do. So, let us exponentiate our matrices. Regarding J, by using the formula $J^2=0$, we obtain:

$$\begin{aligned}e^J&=\sum_{k=0}^{\infty}\begin{pmatrix}0&1\\0&0\end{pmatrix}^k\Big/k!\\&=\begin{pmatrix}1&0\\0&1\end{pmatrix}+\begin{pmatrix}0&1\\0&0\end{pmatrix}\\&=\begin{pmatrix}1&1\\0&1\end{pmatrix}\end{aligned}$$

Similarly, regarding J^*, by using the formula $(J^*)^2=0$, we obtain:

$$\begin{aligned}e^{J^*}&=\sum_{k=0}^{\infty}\begin{pmatrix}0&0\\1&0\end{pmatrix}^k\Big/k!\\&=\begin{pmatrix}1&0\\0&1\end{pmatrix}+\begin{pmatrix}0&0\\1&0\end{pmatrix}\\&=\begin{pmatrix}1&0\\1&1\end{pmatrix}\end{aligned}$$

Now by making the products, we obtain the following formulae:

$$e^Je^{J^*}=\begin{pmatrix}1&1\\0&1\end{pmatrix}\begin{pmatrix}1&0\\1&1\end{pmatrix}=\begin{pmatrix}2&1\\1&1\end{pmatrix}$$

$$e^{J^*}e^J=\begin{pmatrix}1&0\\1&1\end{pmatrix}\begin{pmatrix}1&1\\0&1\end{pmatrix}=\begin{pmatrix}1&1\\1&2\end{pmatrix}$$

But these two formulae give, at least in theory, our counterexample to the multiplication formula $e^{A+B}=e^Ae^B$, due to the following logical implication:

$$e^Je^{J^*}\neq e^{J^*}e^J\implies e^{J+J^*}\neq e^Je^{J^*}\text{ or }e^{J^*+J}\neq e^{J^*}e^J$$

This being said, let us do a clean work, and find out the explicit counterexample. For this purpose, we must compute e^{J+J^*}. The matrix to be exponentiated is:

$$J+J^*=\begin{pmatrix}0&1\\1&0\end{pmatrix}$$

Now this matrix being a symmetry, $(J+J^*)^2=1$, we are led to the following formula, with R,S being certain sums, still in need to be computed:

$$\begin{aligned}
e^{J+J^*} &= \sum_{k=0}^{\infty}\begin{pmatrix}0&1\\1&0\end{pmatrix}^k\Big/k!\\
&= \sum_{l=0}^{\infty}\begin{pmatrix}1&0\\0&1\end{pmatrix}\Big/(2l)!+\sum_{l=0}^{\infty}\begin{pmatrix}0&1\\1&0\end{pmatrix}\Big/(2l+1)!\\
&= \begin{pmatrix}R&S\\S&R\end{pmatrix}
\end{aligned}$$

It remains to compute R,S. But these are given by the following formulae:

$$R=\sum_{l=0}^{\infty}\frac{1}{(2l)!}=\frac{e+e^{-1}}{2}=\cosh 1$$

$$S=\sum_{l=0}^{\infty}\frac{1}{(2l+1)!}=\frac{e-e^{-1}}{2}=\sinh 1$$

Thus, as a conclusion, the matrix e^{J+J^*} is something quite complicated, as follows:

$$e^{J+J^*}=\begin{pmatrix}\cosh 1&\sinh 1\\\sinh 1&\cosh 1\end{pmatrix}$$

Which looks quite exciting, isn't this good mathematics, and more on such things in a moment. But in any case, this matrix being flagrantly different from $e^Je^{J^*}$, and from $e^{J^*}e^J$ too, we have now our counterexample to $e^{A+B}=e^Ae^B$, as desired. □

Moving forward, in order to compute the exponential, with our knowledge so far, the main workhorse remains the formula from Theorem 5.9 (4), for the diagonalizable matrices. So, let us see how that formula works, in practice. We can actually use here as input the symmetry $J+J^*$ from the previous proof, which diagonalizes as follows:

$$\begin{pmatrix}0&1\\1&0\end{pmatrix}=\frac{1}{2}\begin{pmatrix}1&1\\1&-1\end{pmatrix}\begin{pmatrix}1&0\\0&-1\end{pmatrix}\begin{pmatrix}1&1\\1&-1\end{pmatrix}$$

Now by using Theorem 5.9 (4) we obtain, as established in the previous proof:

$$\begin{aligned}\exp\begin{pmatrix}0&1\\1&0\end{pmatrix} &= \frac{1}{2}\begin{pmatrix}1&1\\1&-1\end{pmatrix}\begin{pmatrix}e&0\\0&e^{-1}\end{pmatrix}\begin{pmatrix}1&1\\1&-1\end{pmatrix}\\ &= \frac{1}{2}\begin{pmatrix}1&1\\1&-1\end{pmatrix}\begin{pmatrix}e&e\\e^{-1}&-e^{-1}\end{pmatrix}\\ &= \frac{1}{2}\begin{pmatrix}e+e^{-1}&e-e^{-1}\\e-e^{-1}&e+e^{-1}\end{pmatrix}\\ &= \begin{pmatrix}\cosh 1&\sinh 1\\\sinh 1&\cosh 1\end{pmatrix}\end{aligned}$$

Beyond the diagonalizable case, the only computations that we have so far are those for the matrices J, J^*, from the above proof. But these computations, crucially based on the fact that J, J^* are nilpotent, suggest formulating a general result, as follows:

THEOREM 5.12. *Assuming that $A \in M_N(\mathbb{C})$ is nilpotent, $A^s = 0$, we have:*

$$e^A = \sum_{k=0}^{s-1} \frac{A^k}{k!}$$

More generally, assuming $A^s = 0$, we have the following formula,

$$e^{\lambda+A} = e^\lambda \sum_{k=0}^{s-1} \frac{A^k}{k!}$$

valid for any extra parameter $\lambda \in \mathbb{C}$.

PROOF. The first formula is clear from definitions, and the second one follows from it, by using the fact that the matrices λI and A commute, as follows:

$$\begin{aligned}e^{\lambda I+A} &= e^{\lambda I}e^A\\ &= (e^\lambda I)\sum_{k=0}^{s-1}\frac{A^k}{k!}\\ &= e^\lambda\sum_{k=0}^{s-1}\frac{A^k}{k!}\end{aligned}$$

Thus, we are led to the conclusions in the statement. □

Before going further with our study, which normally means going head-first into the non-diagonalizable case, let us have a listen to cat, who's meowing something, as usual since I started this mathematics chapter, about the diagonalizable matrices being dense. Good point, cat, and double meal for you tonight, because thinking well, by using that density result we can indeed say something very nice about exponentials, as follows:

THEOREM 5.13. *We have the following formula,*

$$\det(e^A) = e^{Tr(A)}$$

valid for any matrix $A \in M_N(\mathbb{C})$.

PROOF. This is something quite tricky, because according to the definition of the exponential, the computation that we have to do looks of extreme difficulty:

$$\det\left(\sum_{k=0}^{\infty}\frac{A^k}{k!}\right) =?$$

But we won't be discouraged by this. For the diagonal matrices, we have:

$$\begin{aligned}\det\left[\exp\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}\right] &= \det\begin{pmatrix}e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_N}\end{pmatrix}\\ &= e^{\lambda_1+\ldots+\lambda_N}\\ &= \exp\left[Tr\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}\right]\end{aligned}$$

Next, by using this, for the diagonalizable matrices, $A = PDP^{-1}$, we have:

$$\begin{aligned}\det(e^A) &= \det(e^{PDP^{-1}})\\ &= \det(Pe^DP^{-1})\\ &= \det(e^D)\\ &= e^{Tr(D)}\\ &= e^{Tr(PDP^{-1})}\\ &= e^{Tr(A)}\end{aligned}$$

And finally, since the diagonalizable matrices are dense, as it is well-known, and more on this in a moment, we get by continuity our result in general. As simple as that. □

So long for the matrix exponential, using beautiful mathematics and tricks. But, everything has to come to an end, and time now to get into some dirty work.

5c. The Jordan form

In order to reach to the Jordan form, we must first do some abstract algebra, for the eigenspaces. Let us begin with some general discussion about these eigenspaces. The basic diagonalization theory, formulated in terms of matrices, is as follows:

PROPOSITION 5.14. *A vector $v \in \mathbb{C}^N$ is called eigenvector of $A \in M_N(\mathbb{C})$, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av = \lambda v$$

In the case where $\mathbb{C}^N$ has a basis $v_1, \ldots, v_N$ formed by eigenvectors of A, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$, in this new basis A becomes diagonal, as follows:

$$A \sim \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

Equivalently, if we denote by $D = diag(\lambda_1, \ldots, \lambda_N)$ the above diagonal matrix, and by $P = [v_1 \ldots v_N]$ the square matrix formed by the eigenvectors of A, we have:

$$A = PDP^{-1}$$

In this case we say that the matrix A is diagonalizable.

PROOF. This is something which is clear, the idea being as follows:

(1) The first assertion is clear, because the matrix which multiplies each basis element v_i by a number λ_i is precisely the diagonal matrix $D = diag(\lambda_1, \ldots, \lambda_N)$.

(2) The second assertion follows from the first one, by changing the basis. We can prove this by a direct computation as well, because we have $Pe_i = v_i$, and so:

$$PDP^{-1}v_i = PDe_i = P\lambda_i e_i = \lambda_i Pe_i = \lambda_i v_i$$

Thus, the matrices A and PDP^{-1} coincide, as stated. □

In order to study now the diagonalization problem, the idea is that the eigenvectors can be grouped into linear spaces, called eigenspaces, as follows:

THEOREM 5.15. *Let $A \in M_N(\mathbb{C})$, and for any eigenvalue $\lambda \in \mathbb{C}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

$$E_\lambda = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda v \right\}$$

These eigenspaces E_λ are then in a direct sum position, in the sense that given vectors $v_1 \in E_{\lambda_1}, \ldots, v_k \in E_{\lambda_k}$ corresponding to different eigenvalues $\lambda_1, \ldots, \lambda_k$, we have:

$$\sum_i c_i v_i = 0 \implies c_i = 0$$

In particular, we have $\sum_\lambda \dim(E_\lambda) \leq N$, with the sum being over all the eigenvalues, and our matrix is diagonalizable precisely when we have equality.

PROOF. We prove the first assertion by recurrence on $k \in \mathbb{N}$. Assume by contradiction that we have a formula as follows, with the scalars $c_1, \ldots, c_k$ being not all zero:

$$c_1 v_1 + \ldots + c_k v_k = 0$$

By dividing by one of these scalars, we can assume that our formula is:

$$v_k = c_1 v_1 + \ldots + c_{k-1} v_{k-1}$$

Now let us apply A to this vector. On the left we obtain:

$$Av_k = \lambda_k v_k = \lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1}$$

On the right we obtain something different, as follows:

$$\begin{aligned} A(c_1 v_1 + \ldots + c_{k-1} v_{k-1}) &= c_1 A v_1 + \ldots + c_{k-1} A v_{k-1} \\ &= c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1} \end{aligned}$$

We conclude from this that the following equality must hold:

$$\lambda_k c_1 v_1 + \ldots + \lambda_k c_{k-1} v_{k-1} = c_1 \lambda_1 v_1 + \ldots + c_{k-1} \lambda_{k-1} v_{k-1}$$

On the other hand, we know by recurrence that the vectors $v_1, \ldots, v_{k-1}$ must be linearly independent. Thus, the coefficients must be equal, at right and at left:

$$\lambda_k c_1 = c_1 \lambda_1$$

$$\vdots$$

$$\lambda_k c_{k-1} = c_{k-1} \lambda_{k-1}$$

Now since at least one of the numbers c_i must be nonzero, from $\lambda_k c_i = c_i \lambda_i$ we obtain $\lambda_k = \lambda_i$, which is a contradiction. Thus our proof by recurrence of the first assertion is complete. As for the second assertion, this follows from the first one. □

In order to reach to more advanced results about diagonalization, we can use the characteristic polynomial, which appears via the following fundamental result:

THEOREM 5.16. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial:*

$$P(x) = \det(A - x1_N)$$

The eigenvalues of A are then the roots of P. Also, we have the inequality

$$\dim(E_\lambda) \leq m_\lambda$$

where m_λ is the multiplicity of λ, as root of P.

PROOF. The first assertion follows from the following computation, using the fact that a linear map is bijective when the determinant of the associated matrix is nonzero:

$$\begin{aligned} \exists v, Av = \lambda v &\iff \exists v, (A - \lambda 1_N)v = 0 \\ &\iff \det(A - \lambda 1_N) = 0 \end{aligned}$$

Regarding now the second assertion, given an eigenvalue λ of our matrix A, consider the dimension $d_\lambda = \dim(E_\lambda)$ of the corresponding eigenspace. By changing the basis of $\mathbb{C}^N$, as for the eigenspace E_λ to be spanned by the first d_λ basis elements, our matrix becomes as follows, with B being a certain smaller matrix:

$$A \sim \begin{pmatrix} \lambda 1_{d_\lambda} & 0 \\ 0 & B \end{pmatrix}$$

We conclude that the characteristic polynomial of A is of the following form:

$$P_A = P_{\lambda 1_{d_\lambda}} P_B = (\lambda - x)^{d_\lambda} P_B$$

Thus the multiplicity m_λ of our eigenvalue λ, as a root of P, satisfies $m_\lambda \geq d_\lambda$, and this leads to the conclusion in the statement. □

We can now put together Theorem 5.15 and Theorem 5.16, and by using as well the well-known fact that any complex polynomial of degree N has exactly N complex roots, when counted with multiplicities, we obtain the folowing result:

THEOREM 5.17. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial*

$$P(X) = \det(A - X1_N)$$

then factorize this polynomial, by computing the complex roots, with multiplicities,

$$P(X) = (-1)^N (X - \lambda_1)^{n_1} \ldots (X - \lambda_k)^{n_k}$$

and finally compute the corresponding eigenspaces, for each eigenvalue found:

$$E_i = \left\{ v \in \mathbb{C}^N \middle| Av = \lambda_i v \right\}$$

The dimensions of these eigenspaces satisfy then the following inequalities,

$$\dim(E_i) \leq n_i$$

and A is diagonalizable precisely when we have equality for any i.

PROOF. This follows by combining the above results. Indeed, by summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Theorem 5.16, we obtain an inequality as follows:

$$\sum_\lambda \dim(E_\lambda) \leq \sum_\lambda m_\lambda \leq N$$

On the other hand, we know from Theorem 5.15 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

This was for the main result of linear algebra. There are countless applications of this, and we will see illustrations in a moment, and generally speaking, advanced linear algebra, including the Jordan theory to come, consists in building on Theorem 5.17.

Let us record as well a useful algorithmic version of the above result:

THEOREM 5.18. *The square matrices $A \in M_N(\mathbb{C})$ can be diagonalized as follows:*

(1) *Compute the characteristic polynomial.*
(2) *Factorize the characteristic polynomial.*
(3) *Compute the eigenvectors, for each eigenvalue found.*
(4) *If there are no N eigenvectors, A is not diagonalizable.*
(5) *Otherwise, A is diagonalizable, $A = PDP^{-1}$.*

PROOF. This is an informal reformulation of Theorem 5.17, with (4) referring to the total number of linearly independent eigenvectors found in (3), and with $A = PDP^{-1}$ in (5) being the usual diagonalization formula, with P, D being as before. □

As an illustration for all this, which is a must-know computation, we have:

PROPOSITION 5.19. *The rotation of angle $t \in \mathbb{R}$ in the plane diagonalizes as:*

$$\begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

Over the reals this is impossible, unless $t = 0, \pi$, where the rotation is diagonal.

PROOF. Observe first that, as indicated, unlike we are in the case $t = 0, \pi$, where our rotation is $\pm 1_2$, our rotation is a "true" rotation, having no eigenvectors in the plane. Fortunately the complex numbers come to the rescue, via the following computation:

$$\begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix} = \begin{pmatrix} \cos t - i \sin t \\ i \cos t + \sin t \end{pmatrix} = e^{-it} \begin{pmatrix} 1 \\ i \end{pmatrix}$$

We have as well a second complex eigenvector, coming from:

$$\begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 1 \\ -i \end{pmatrix} = \begin{pmatrix} \cos t + i \sin t \\ -i \cos t + \sin t \end{pmatrix} = e^{it} \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

Thus, we are led to the conclusion in the statement. □

At the level of basic examples of diagonalizable matrices, we first have the following result, which provides us with the "generic" examples:

THEOREM 5.20. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent,*

(1) *The eigenvalues are different, $\lambda_i \neq \lambda_j$,*
(2) *The characteristic polynomial P has simple roots,*
(3) *The characteristic polynomial satisfies $(P, P') = 1$,*
(4) *The resultant of P, P' is nonzero, $R(P, P') \neq 0$,*
(5) *The discriminant of P is nonzero, $\Delta(P) \neq 0$,*

and in this case, the matrix is diagonalizable.

PROOF. The last assertion holds indeed, due to Theorem 5.17. As for the equivalences in the statement, these are all standard, by using the basic theory of the resultant R and of the discriminant Δ, for which we refer to any advanced linear algebra book. □

As already mentioned, one can prove that the matrices having distinct eigenvalues are "generic", and so the above result basically captures the whole situation. We have in fact the following collection of density results, which are quite advanced:

THEOREM 5.21. *The following happen, inside $M_N(\mathbb{C})$:*

(1) *The invertible matrices are dense.*
(2) *The matrices having distinct eigenvalues are dense.*
(3) *The diagonalizable matrices are dense.*

PROOF. These are quite advanced results, which can be proved as follows:

(1) This is clear, intuitively speaking, because the invertible matrices are given by the condition $\det A \neq 0$. Thus, the set formed by these matrices appears as the complement of the hypersurface $\det A = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(2) Here we can use a similar argument, this time by saying that the set formed by the matrices having distinct eigenvalues appears as the complement of the hypersurface given by $\Delta(P_A) = 0$, and so must be dense inside $M_N(\mathbb{C})$, as claimed.

(3) This follows from (2), via the fact that the matrices having distinct eigenvalues are diagonalizable, that we know from Theorem 5.20. There are some other proofs as well, for instance by putting the matrix in Jordan form, and more on this in a moment. □

Moving forward, in order to reach to the Jordan form, we must do some more abstract algebra, for the eigenspaces. But this is something which is quite routine, by further building on the material above, and more specifically, by working out what happens, when we have strict inequalities in the various inequalities that we established.

We are led in this way to the Jordan form, which applies to any matrix:

THEOREM 5.22. *Any matrix $A \in M_N(\mathbb{C})$ can be written, up to a base change, as*

$$A = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix}$$

with each J_i being a Jordan block, meaning a matrix as follows,

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}$$

with our usual convention that blank spaces stand for 0 entries.

PROOF. This follows indeed from the above discussion. In fact, we have already met Jordan blocks in the above, on various occasions, and we are quite familiar with them. □

In practice, there are many illustrations for this theorem. We will be back to this, on a regular basis, in the remainder of this chapter, and in the next chapter too.

In analogy with our diagonalization algorithm given above, we can talk about algorithms, regarding the Jordan form, the precise statement here being as follows:

ALGORITHM 5.23. *In order to find the Jordan blocks of our matrix,*

$$J_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}$$

we must, for each eigenvalue λ_i, do a number of computations.

To be more precise, all this is based of course on Theorem 5.22, which gives the result as formulated above, and in practice, nothing better than working out some particular cases, for some matrices of your choice, chosen of course not to be diagonalizable.

For instance, as a must-do computation here, which is very illustrating, for all the above, you can try to see what happens in the particular case of the 2×2 matrices.

We will actually do a number of such exercises right next, when talking applications, in relation with the differential equations questions that we started with.

5d. Basic applications

There are many concrete illustrations for the Jordan decomposition theorem, and we can, for instance, explicitly compute the Jordan form of any 2×2 matrix.

As a first application now, we can go back to exponentials, and compute e^A for any matrix, decomposed in Jordan form. In fact, we have already seen such computations, in the proof of Theorem 5.11, and the computations in general are quite similar.

To be more precise, let us write the matrix to be exponentiated in Jordan form, as in Theorem 5.22, as follows, with P denoting the passage matrix used there:

$$A = P \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix} P^{-1}$$

According to Theorem 5.9, the exponential is then given by the following formula:

$$e^A = P \begin{pmatrix} e^{J_1} & & \\ & \ddots & \\ & & e^{J_k} \end{pmatrix} P^{-1}$$

Thus, it is enough to know how to exponentiate Jordan blocks. So, consider a Jordan block, as follows, with our usual convention that blank spaces stand for 0 entries:

$$J = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

In order to exponentiate this matrix, the best is to use Theorem 5.12. Indeed, what we have here is a multiple of the identity, summed with a nilpotent matrix:

$$J = \lambda + \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

Thus, we have the following formula for the exponential of our Jordan block:

$$e^J = e^\lambda \exp \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

So, we are led to the question of exponentiating the matrix on the right, namely:

$$N = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix}$$

Now in order to exponentiate this latter matrix, we can use the fact that this matrix is nilpotent. Indeed, the square of this matrix is given by the following formula:

$$N^2 = \begin{pmatrix} 0 & 0 & 1 & & & & \\ & 0 & 0 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 0 & 0 & 1 & \\ & & & & 0 & 0 & \\ & & & & & 0 & \end{pmatrix}$$

Then, the third power of this matrix is given by the following formula:

$$N^3 = \begin{pmatrix} 0 & 0 & 0 & 1 & & & & \\ & 0 & 0 & 0 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ & & & 0 & 0 & 0 & 1 & \\ & & & & 0 & 0 & 0 & \\ & & & & & 0 & 0 & \\ & & & & & & 0 & \end{pmatrix}$$

And so on up to the $(s-1)$-th power, with s being the size of our matrix, which is given by the following formula, with our usual convention for blank spaces:

$$N^s = \begin{pmatrix} 0 & & \dots & \dots & 0 & 1 \\ & 0 & & & & 0 \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & 0 & \\ & & & & & 0 \end{pmatrix}$$

Now by using the exponentiating formula in Theorem 5.12, for this nilpotent matrix N, we obtain the following formula, for its exponential:

$$e^N = \begin{pmatrix} 1 & 1 & \frac{1}{2} & \frac{1}{6} & & \dots & & \frac{1}{(s-1)!} \\ & 1 & 1 & \frac{1}{2} & \frac{1}{6} & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \ddots & \ddots & \\ & & & & 1 & 1 & \frac{1}{2} & \frac{1}{6} \\ & & & & & 1 & 1 & \frac{1}{2} \\ & & & & & & 1 & 1 \\ & & & & & & & 1 \end{pmatrix}$$

Summarizing, done with our computation, and we can now formulate:

THEOREM 5.24. *For a matrix written in Jordan form, as follows,*

$$A = P\begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix} P^{-1}$$

the corresponding exponential is given by the following formula,

$$e^A = P\begin{pmatrix} e^{J_1} & & \\ & \ddots & \\ & & e^{J_k} \end{pmatrix} P^{-1}$$

with the exponential of each Jordan block being computed by the formula

$$\exp\begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix} = e^\lambda \begin{pmatrix} 1 & 1 & \frac{1}{2} & \frac{1}{6} & & \cdots & & \frac{1}{(s-1)!} \\ & 1 & 1 & \frac{1}{2} & \frac{1}{6} & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \vdots \\ & & & \ddots & \ddots & \ddots & \ddots & \\ & & & & 1 & 1 & \frac{1}{2} & \frac{1}{6} \\ & & & & & 1 & 1 & \frac{1}{2} \\ & & & & & & 1 & 1 \\ & & & & & & & 1 \end{pmatrix}$$

with s being the size of our Jordan block.

And good news, this is all we need to know, this being obviously something very powerful, closing any further mathematical discussion about exponentiation.

But with this is hand, we can now go back to the linear differential equations, and say more about them. We will be back to all this in the next chapter, on a more systematic basis, when discussing dynamical systems, at a quite general level.

As another application of our theory, we can recover the density of the diagonalizable matrices, that we can get via the Jordan form, by perturbing the diagonal.

As another application, we can apply other complex functions to our matrices, under suitable assumptions. All this is quite technical, called "functional calculus", and as a basic result here, coming from the Cauchy formula, we can apply any holomorphic function to any matrix. More on such things, which can be quite technical, later, when needed.

Passed the holomorphic functions, things become more complicated. In the normal case, we can apply continuous functions, and even measurable ones, to our matrices. Indeed, this follows from the well-known spectral theorem for the normal matrices.

5e. Exercises

Exercises:

Exercise 5.25.

Exercise 5.26.

Exercise 5.27.

Exercise 5.28.

Exercise 5.29.

Exercise 5.30.

Exercise 5.31.

Exercise 5.32.

Bonus exercise.

CHAPTER 6

Ordinary equations

6a. Differential equations

Let us go back to the general ordinary differential equations (ODE), briefly discussed in the beginning of chapter 5. We recall from there that a standard system of ODE is a system as follows, with the unknown being a vector function $x = x(t) : \mathbb{R} \to \mathbb{R}^N$:

$$x_1^{(k)} = f_1(t, x, \dot{x}, \ldots, x^{(k-1)})$$

$$\vdots$$

$$x_N^{(k)} = f_N(t, x, \dot{x}, \ldots, x^{(k-1)})$$

The point now is that, up to suitably enlarging the size of the unknown vector, we can convert this standard system of ODE into a standard order 1 system of ODE. Indeed, with $y = (x, \dot{x}, \ldots, x^{(k-1)})$, the system takes the following form, as desired:

$$\dot{y}_1 = y_2$$

$$\dot{y}_2 = y_3$$

$$\vdots$$

$$\dot{y}_{k-1} = y_k$$

$$\dot{y}_k = f(t, y)$$

Moreover, in the autonomous case, that where the function f does not depend on time t, we can further set $z = (t, y)$, and we are led in this way to a system as follows:

$$\dot{z}_1 = 1$$

$$\dot{z}_2 = z_3$$

$$\vdots$$

$$\dot{z}_k = z_{k+1}$$

$$\dot{z}_{k+1} = f(z)$$

Our first goal in this chapter will be that of finding existence and uniqueness results for the solutions of such systems of ODE. But let us begin with some examples, in 1D. More specifically, we will be interested in the following type of equations:

DEFINITION 6.1. *An autonomous order 1 ODE is an equation of type*

$$\dot{x} = f(x) \quad , \quad x(0) = x_0$$

with $f \in C(\mathbb{R})$ being a certain function.

In order to solve now our equation, assume that we are in the case $f(x_0) \neq 0$. Then, around $t = 0$, we can divide our equation by $f(x(s))$, and then integrate:

$$\int_0^t \frac{\dot{x}(s)}{f(x(s))}\, ds = t$$

In view of this observation, consider the following function:

$$F(x) = \int_{x_0}^{x} \frac{1}{f(y)}\, dy$$

We have then the following computation, taking into account our equation:

$$\begin{aligned} F(x(t)) &= \int_{x_0}^{x(t)} \frac{1}{f(y)}\, dy \\ &= \int_0^t \frac{\dot{x}(s)}{f(x(s))}\, ds \\ &= t \end{aligned}$$

Obviously, the converse holds too, so our original equation is equivalent to:

$$F(x(t)) = t$$

Now recall that we assumed $f(x_0) \neq 0$. But this means that $F(x)$ is monotone around x_0, and so invertible, so we have a unique solution to our equation, given by:

$$\varphi(t) = F^{-1}(t)$$

Note also that we have, as we should, as required by Definition 6.1:

$$\varphi(0) = F^{-1}(0) = x_0$$

With this discussion made, which was something local, let us turn now to global questions. We have here the following question, that we would like to solve:

QUESTION 6.2. *In the context of the above autonomous order 1 ODE, and discussion, what is the interval where the solution is defined? And, when is this interval $\mathbb{R}$ itself?*

In order to discuss this latter question, in view of $f(x_0) \neq 0$, assume that we are in the case $f(x_0) > 0$, with the other case, $f(x_0) < 0$, being similar. We have then $f > 0$ on a certain interval (x_1, x_2) around x_0. Now consider the following two limits:

$$T_+ = \lim_{x \nearrow x_2} F(x) \in (0, \infty]$$

$$T_+ = \lim_{x \searrow x_1} F(x) \in [-\infty, 0)$$

We have then $f \in C^1(T_-, T_+)$, and the following equalities hold:

$$\lim_{t \nearrow T_+} \varphi(t) = x_2$$

$$\lim_{t \searrow T_-} \varphi(t) = x_1$$

Now according to the first equation, φ exists for any $t > 0$ precisely when:

$$T_+ = \int_{x_0}^{x_2} \frac{1}{f(y)}\, dy = \infty$$

Also, according to the second equation, φ exists for any $t < 0$ precisely when:

$$T_- = \int_{x_1}^{x_0} \frac{1}{f(y)}\, dy = -\infty$$

Summarizing, we are led to the following answer to Question 6.2:

ANSWER 6.3. *In the context of the above autonomous order 1 ODE, and discussion involving the interval (x_1, x_2) around x_0, the solution φ is as follows:*

(1) *φ exists for any $t > 0$ when $1/f$ is not integrable around x_2.*
(2) *φ exists for any $t < 0$ when $1/f$ is not integrable around x_1.*

All this was quite theoretical, so let us work out now some examples. For $f(x) = x$, and with $x_0 > 0$, we have $(x_1, x_2) = (0, \infty)$, and the function F is given by:

$$F(x) = \log\left(\frac{x}{x_0}\right)$$

Also, we have $T_\pm = \pm\infty$, and the solution is as follows, defined on the whole $\mathbb{R}$:

$$\varphi(t) = x_0 e^t$$

As a second example now, let us take $f(x) = x^2$, and $x_0 > 0$. In this case we have $(x_1, x_2) = (0, \infty)$, and the function F is given by:

$$F(x) = \frac{1}{x_0} - \frac{1}{x}$$

Also, in this case we have $T_+ = 1/x_0$ and $T_- = -\infty$, and the solution of our equation is as follows, defined on the interval $(-\infty, 1/x_0)$:

$$\varphi(t) = \frac{x_0}{1 - x_0 t}$$

We will see some other examples for all this, in what follows.

As a continuation of the above discussion, dealing with the case $f(x_0) \neq 0$, it remains now to discuss the case $f(x_0) = 0$. Here we have the trivial solution $\varphi(t) = x_0$, and we

can have as well non-trivial solutions. Assume for instance that we have:

$$\left|\int_{x_0}^{x_0+\varepsilon}\frac{1}{f(y)}\,dy\right|<\infty$$

Then, we have the following non-trivial solution, to our equation:

$$\varphi(t)=F^{-1}(t)\quad,\quad F(x)=\int_{x_0}^{x}\frac{1}{f(y)}\,dy$$

Again, in order to understand this, nothing better than an explicit example. Let us take $f(x)=\sqrt{|x|}$. In the case $x_0>0$, studied before, we have $(x_1,x_2)=(0,\infty)$, then $F(x)=2(\sqrt{x}-\sqrt{x_0})$, and the solution is as follows, with $t\in(-2\sqrt{x_0},\infty)$:

$$\varphi(t)=\left(\sqrt{x_0}+\frac{t}{2}\right)^2$$

In the case $x_0=0$, however, we have several solutions, that can be obtained by gluing the trivial solution, and the generic solution. Indeed, we can take:

$$\varphi(t)=\begin{cases}-\frac{(t-t_0)^2}{4} & \text{for } t\leq t_0\\ 0 & \text{for } t_0\leq t\leq t_1\\ \frac{(t-t_1)^2}{4} & \text{for } t_1\leq t\end{cases}$$

Based on the above study, and on our various examples, let us formulate:

Conclusion 6.4. *In the context of the above autonomous order 1 ODE:*

(1) *Even when the function f is C^∞, we can only have local solutions.*
(2) *Also, in general, we do not have the uniqueness of the solution.*

Before getting into a heavier theoretical study of the existence and uniqueness of solutions, let us discuss as well a few tricks for the ODE, sometimes leading to explicit solutions. A useful method is that of using a change of variables, as follows:

$$(t,x)\to(s,y)$$

To be more precise, we are looking for suitable functions σ,η, as follows:

$$s=\sigma(t,x)\quad,\quad y=\eta(t,x)$$

In order to have a change of variables, our transformation must be of course invertible. However, this assumption is not enough, at the level of solutions, because by rotating the graph of a function, we do not necessarily obtain the graph of a function.

In view of this, a reasonable assumption is that our transformations must preserve the fibers, with "fiber" meaning here corresponding to constant time. That is, we are looking for changes of variables, suitably adapted to our ODE, of the following special type:

$$s=\sigma(t)\quad,\quad y=\eta(t,x)$$

Now assume that we have such a transformation, which is invertible, as any change of variables should be, with inverse given by formulae as follows:

$$t = \tau(s) \quad , \quad x = \xi(s, y)$$

Then $\varphi(t)$ is a solution of $\dot{x} = f(t, x)$ precisely when $\psi(s) = \eta(\tau(s), \varphi(\tau(s)))$ is a solution of the following equation, where $\tau = \tau(s)$ and $\xi = \xi(s, y)$:

$$\dot{y} = \dot{\tau}\left(\frac{d\eta}{dt}(\tau, \xi) + \frac{d\eta}{dx}(\tau, \xi) f(t, \xi)\right)$$

Which is quite nice, becase we can get some concrete results in this way, that is, explicit solutions for explicit ODE, by doing some reverse engineering, based on this.

Finally, for ending this preliminary section on general ODE theory, let us discuss some well-known equations. First we have the Bernoulli equations, which are as follows:

$$\dot{x} = f(t)x + g(t)x^n$$

Assuming $n \neq 1$, we can set $y = x^{1-n}$, and our equation takes the following form:

$$\dot{y} = (1 - n)f(t)y + (1 - n)g(t)$$

But this is a linear equation, that we can solve by using the linear algebra methods from chapter 5. We will be back to this later, with further details.

As a second class of well-known equations, again coming from a variety of questions from physics, we have the Riccati equations, which are as follows:

$$\dot{x} = f(t)x + g(t)x^2 + h(t)$$

Now assuming that we have found a particular solution $x_p(t)$, we can set:

$$y = \frac{1}{x - x_p(t)}$$

With this change of variables, our equation takes the following form:

$$\dot{y} = -(f(t) + 2x_p(t)g(t))y - g(t)$$

But this is again a linear equation, that we can solve by using the linear algebra methods from chapter 5. We will be back to this later, with further details.

6b. Functional analysis

With the above discussed, which remains something a bit ad-hoc, let us try now to develop some general theory. We would like to solve the following problem:

PROBLEM 6.5. *Do we have the local existence and uniqueness of the solutions of*

$$\dot{x} = f(t, x) \quad , \quad x(t_0) = x_0$$

under suitable assumptions on the function $f \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open?

In order to solve this latter question, we have a strategy which is quite straightforward. Indeed, we can integrate our equation, which takes the following form:

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

Based on this observation, consider the following function:

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

In terms of this function, our original equation reads:

$$K(x) = x$$

So, all in all, we are into a fixed point problem. But, as you certainly know from basic calculus, such questions can be simply solved by iterating. Thus, we are led to:

QUESTIONS 6.6. *In relation with the above strategy, for solving Problem 6.5:*

(1) *Can we develop a theory of infinite dimensional complete normed spaces?*
(2) *Do we have fixed point theorems, inside such complete normed spaces?*
(3) *Can we apply these fixed point theorems, as to solve our ODE problem?*

We will see in what follows that the answers to these latter questions are yes, yes, yes. However, this is something quite technical, which will take some time. In order to solve the first question, in relation with the normed spaces, let us start with:

DEFINITION 6.7. *A normed space is a complex vector space V, which can be finite or infinite dimensional, together with a map*

$$||.|| : V \to \mathbb{R}_+$$

called norm, subject to the following conditions:

(1) $||x|| = 0$ *implies* $x = 0$.
(2) $||\lambda x|| = |\lambda| \cdot ||x||$, *for any* $x \in V$, *and* $\lambda \in \mathbb{C}$.
(3) $||x + y|| \leq ||x|| + ||y||$, *for any* $x, y \in V$.

As a basic example here, which is finite dimensional, we have the space $V = \mathbb{C}^N$, with the norm on it being the usual length of the vectors, namely:

$$||x|| = \sqrt{\sum_i |x_i|^2}$$

Indeed, for this space (1) is clear, (2) is clear too, and (3) is something well-known, which is equivalent to the triangle inequality in $\mathbb{C}^N$, and which can be deduced from the Cauchy-Schwarz inequality. More on this, with some generalizations, in a moment.

Getting back now to the general case, we have the following result:

PROPOSITION 6.8. *Any normed vector space V is a metric space, with*

$$d(x,y) = ||x-y||$$

as distance. If this metric space is complete, we say that V is a Banach space.

PROOF. This follows from the definition of the metric spaces, as follows:

(1) The first distance axiom, $d(x,y) \geq 0$, and $d(x,y) = 0$ precisely when $x = y$, follows from the fact that the norm takes values in $\mathbb{R}_+$, and from $||x|| = 0 \implies x = 0$.

(2) The second distance axiom, which is the symmetry one, $d(x,y) = d(y,x)$, follows from our condition $||\lambda x|| = |\lambda| \cdot ||x||$, with $\lambda = -1$.

(3) As for the third distance axiom, which is the triangle inequality $d(x,y) \leq d(x,z) + d(y,z)$, this follows from our third norm axiom, namely $||x+y|| \leq ||x|| + ||y||$. □

Very nice all this, and it is possible to develop some general theory here, but before everything, however, we need more examples, besides $\mathbb{C}^N$ with its usual norm.

However, these further examples are actually quite tricky to construct, needing some inequality know-how. Let us start with a very basic result, as follows:

THEOREM 6.9 (Jensen). *Given a convex function $f : \mathbb{R} \to \mathbb{R}$, we have the following inequality, for any $x_1, \ldots, x_N \in \mathbb{R}$, and any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1,*

$$f(\lambda_1 x_1 + \ldots + \lambda_N x_N) \leq \lambda_1 f(x_1) + \ldots + \lambda_N x_N$$

with equality when $x_1 = \ldots = x_N$. In particular, by taking the weights λ_i to be all equal, we obtain the following inequality, valid for any $x_1, \ldots, x_N \in \mathbb{R}$,

$$f\left(\frac{x_1 + \ldots + x_N}{N}\right) \leq \frac{f(x_1) + \ldots + f(x_N)}{N}$$

and once again with equality when $x_1 = \ldots = x_N$. We have a similar statement holds for the concave functions, with all the inequalities being reversed.

PROOF. This is indeed something quite routine, the idea being as follows:

(1) First, we can talk about convex functions in a usual, intuitive way, with this meaning by definition that the following inequality must be satisfied:

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}$$

(2) But this means, via a simple argument, by approximating numbers $t \in [0,1]$ by sums of powers 2^{-k}, that for any $t \in [0,1]$ we must have:

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

$$f\left(\frac{x_1+\ldots+x_N}{N}\right)\leq\frac{f(x_1)+\ldots+f(x_N)}{N}$$

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1,\ldots,\lambda_N>0$ summing up to 1:

$$f(\lambda_1x_1+\ldots+\lambda_Nx_N)\leq\lambda_1f(x_1)+\ldots+\lambda_Nx_N$$

(3) Summarizing, all our Jensen inequalities, at $N=2$ and at $N\in\mathbb{N}$ arbitrary, are equivalent. The point now is that, if we look at what the first Jensen inequality, that we took as definition for the convexity, means, this is simply equivalent to:

$$f''(x)\geq 0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that f is concave precisely when $-f$ is convex, and get the results from what we have. □

As a basic application of the Jensen inequality, we have:

PROPOSITION 6.10. *For $p\in(1,\infty)$ we have the following inequality,*

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\leq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

and for $p\in(0,1)$ we have the following reverse inequality,

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\geq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

with in both cases equality precisely when $|x_1|=\ldots=|x_N|$.

PROOF. This follows indeed from Theorem 6.9, because we have:

$$(x^p)''=p(p-1)x^{p-2}$$

Thus x^p is convex for $p>1$ and concave for $p<1$, which gives the results. □

Observe that at $p=2$ we obtain as particular case of the above inequality the Cauchy-Schwarz inequality, or rather something equivalent to it, namely:

$$\left(\frac{x_1+\ldots+x_N}{N}\right)^2\leq\frac{x_1^2+\ldots+x_N^2}{N}$$

As another basic application of the Jensen inequality, we have:

THEOREM 6.11 (Young). *We have the following inequality,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

valid for any $a, b \geq 0$, and any exponents $p, q > 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$.

PROOF. We use the logarithm function, which is concave on $(0, \infty)$, due to:

$$(\log x)'' = \left(-\frac{1}{x}\right)' = -\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned} \log\left(\frac{a^p}{p} + \frac{b^q}{q}\right) &\geq \frac{\log(a^p)}{p} + \frac{\log(b^q)}{q} \\ &= \log(a) + \log(b) \\ &= \log(ab) \end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

Observe that for the simplest exponents, namely $p = q = 2$, the Young inequality gives something which is trivial, but is very useful and basic, namely:

$$ab \leq \frac{a^2 + b^2}{2}$$

In general, the Young inequality is something non-trivial, and the idea with it is that "when stuck with a problem, and with $ab \leq \frac{a^2+b^2}{2}$ not working, try Young".

Moving forward now, as a consequence of the Young inequality, we have:

THEOREM 6.12 (Hölder). *Assuming that $p, q \geq 1$ are conjugate, in the sense that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

we have the following inequality, valid for any two vectors $x, y \in \mathbb{C}^N$,

$$\sum_i |x_i y_i| \leq \left(\sum_i |x_i|^p\right)^{1/p} \left(\sum_i |y_i|^q\right)^{1/q}$$

with the convention that an ∞ exponent produces a $\max |x_i|$ quantity.

PROOF. This is something very standard, the idea being as follows:

(1) Assume first that we are dealing with finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that x, y are normalized, in the following way:

$$\sum_i |x_i|^p = \sum_i |y_i|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\sum_i |x_i y_i| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any i:

$$|x_i y_i| \leq \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$$

By summing now over $i = 1, \ldots, N$, we obtain from this, as desired:

$$\begin{aligned} \sum_i |x_i y_i| &\leq \sum_i \frac{|x_i|^p}{p} + \sum_i \frac{|y_i|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) In the case $p = 1$ and $q = \infty$, or vice versa, the inequality holds too, trivially, with the convention that an ∞ exponent produces a max quantity, according to:

$$\lim_{p \to \infty} \left(\sum_i |x_i|^p \right)^{1/p} = \max |x_i|$$

Thus, we are led to the conclusion in the statement. □

As a consequence now of the Hölder inequality, we have:

THEOREM 6.13 (Minkowski). *Assuming $p \in [1, \infty]$, we have the inequality*

$$\left(\sum_i |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_i |x_i|^p \right)^{1/p} + \left(\sum_i |y_i|^p \right)^{1/p}$$

for any two vectors $x, y \in \mathbb{C}^N$, with our usual conventions at $p = \infty$.

PROOF. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q \in [1, \infty]$, given by $1/p + 1/q = 1$:

$$\begin{aligned}
\sum_i |x_i + y_i|^p &= \sum_i |x_i + y_i| \cdot |x_i + y_i|^{p-1} \\
&\leq \sum_i |x_i| \cdot |x_i + y_i|^{p-1} + \sum_i |y_i| \cdot |x_i + y_i|^{p-1} \\
&\leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |x_i + y_i|^{(p-1)q} \right)^{1/q} \\
&+ \left(\sum_i |y_i|^p \right)^{1/p} \left(\sum_i |x_i + y_i|^{(p-1)q} \right)^{1/q} \\
&= \left[\left(\sum_i |x_i|^p \right)^{1/p} + \left(\sum_i |y_i|^p \right)^{1/p} \right] \left(\sum_i |x_i + y_i|^p \right)^{1-1/p}
\end{aligned}$$

Here we have used the following fact, at the end:

$$\frac{1}{p} + \frac{1}{q} = 1 \implies \frac{1}{q} = \frac{p-1}{p} \implies (p-1)q = p$$

Now by dividing both sides by the last quantity at the end, we obtain:

$$\left(\sum_i |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_i |x_i|^p \right)^{1/p} + \left(\sum_i |y_i|^p \right)^{1/p}$$

Thus, we are led to the conclusion in the statement. □

Good news, done with inequalities, and as a consequence of the above results, and more specifically of the Minkowski inequality obtained above, we can formulate:

THEOREM 6.14. *Given an exponent $p \in [1, \infty]$, the formula*

$$||x||_p = \left(\sum_i |x_i|^p \right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on $\mathbb{C}^N$, making it a Banach space.

PROOF. Here the normed space assertion follows from the Minkowski inequality, established above, and the Banach space assertion is trivial, because our space being finite dimensional, by standard linear algebra all the Cauchy sequences converge. □

Very nice all this, but you might wonder at this point, what is the relation of all this with functions. In answer, Theorem 6.14 can be reformulated as follows:

THEOREM 6.15. *Given an exponent $p \in [1,\infty]$, the formula*

$$||f||_p = \left(\int |f(x)|^p\right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on the space of functions

$$f : \{1, \ldots, N\} \to \mathbb{C}$$

making it a Banach space.

PROOF. This is a just fancy reformulation of Theorem 6.14, by using the fact that the space formed by the functions $f : \{1, \ldots, N\} \to \mathbb{C}$ is canonically isomorphic to $\mathbb{C}^N$. □

In order to further extend the above result, let us start with:

THEOREM 6.16. *Given two functions $f, g : X \to \mathbb{C}$ and an exponent $p \geq 1$, we have*

$$\left(\int_X |f+g|^p\right)^{1/p} \leq \left(\int_X |f|^p\right)^{1/p} + \left(\int_X |g|^p\right)^{1/p}$$

called Minkowski inequality. Also, assuming that $p, q \geq 1$ satisfy $1/p + 1/q = 1$, we have

$$\int_X |fg| \leq \left(\int_X |f|^p\right)^{1/p} \left(\int_X |g|^q\right)^{1/q}$$

called Hölder inequality. These inequalities hold as well for ∞ values of the exponents.

PROOF. This is very standard, exactly as in the case of sequences, as follows:

(1) Let us first prove Hölder, in the case of finite exponents, $p, q \in (1,\infty)$. By linearity we can assume that f, g are normalized, in the following way:

$$\int_X |f|^p = \int_X |g|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\int_X |fg| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any $x \in X$:

$$|f(x)g(x)| \leq \frac{|f(x)|^p}{p} + \frac{|g(x)|^q}{q}$$

By integrating now over $x \in X$, we obtain from this, as desired:

$$\begin{aligned} \int_X |fg| &\leq \int_X \frac{|f(x)|^p}{p} + \int_X \frac{|g(x)|^q}{q} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

(2) Let us prove now Minkowski, again in the finite exponent case, $p \in (1,\infty)$. We have the following estimate, using the Hölder inequality, and the conjugate exponent:

$$\begin{aligned}
\int_X |f+g|^p &= \int_X |f+g| \cdot |f+g|^{p-1} \\
&\leq \int_X |f| \cdot |f+g|^{p-1} + \int_X |g| \cdot |f+g|^{p-1} \\
&\leq \left(\int_X |f|^p\right)^{1/p} \left(\int_X |f+g|^{(p-1)q}\right)^{1/q} \\
&+ \left(\int_X |g|^p\right)^{1/p} \left(\int_X |f+g|^{(p-1)q}\right)^{1/q} \\
&= \left[\left(\int |f|^p\right)^{1/p} + \left(\int_X |g|^p\right)^{1/p}\right] \left(\int_X |f+g|^p\right)^{1-1/p}
\end{aligned}$$

Thus, we are led to the Minkowski inequality in the statement.

(3) Finally, in the infinite exponent cases we have similar results, which are trivial this time, with the convention that an ∞ exponent produces an essential supremum, according to the following formula, which follows from the measure theory that we know:

$$\lim_{p\to\infty} \left(\int_X |f|^p\right)^{1/p} = \text{ess sup}|f|$$

Thus, we are led to the conclusion in the statement. □

We can now extend Theorem 6.16, into something very general, as follows:

THEOREM 6.17. *Given a measured space X, and $p \in [1,\infty]$, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^p(X) = \left\{f : X \to \mathbb{C} \middle| \int_I |f(x)|^p dx < \infty\right\}$$

is a vector space, and the following quantity

$$||f||_p = \left(\int_X |f(x)|^p\right)^{1/p}$$

is a norm on it, making it a Banach space.

PROOF. This follows indeed from Theorem 6.16, with due attention to the null sets, and this because of the first normed space axiom, namely:

$$||x|| = 0 \implies x = 0$$

To be more precise, in order for this axiom to hold, we must identify the functions up to equality almost everywhere, as indicated in the statement. □

6c. Existence, uniqueness

Getting now towards our ODE business, existence and uniqueness results, as explained before, we would like to use some fixed point technology. So, let us formulate:

DEFINITION 6.18. *Let V be a Banach space, and $K : C \subset V \to C$ be a linear map, with C being closed. We say that K is a contraction if*

$$||K(x) - K(y)|| \leq \theta ||x - y||$$

for some $\theta \in [0, 1)$. Also, we call fixed point of K any $x \in C$ such that $K(x) = x$.

Observe that the fixed point of a contraction, if it exists, is unique, due to our assumption $\theta < 1$. Now with these notions in hand, we have the following result:

THEOREM 6.19. *Any contraction $K : C \subset V \to C$ has a unique fixed point $\bar{x} \in C$, which can be obtained by starting with any point $x \in C$, and iterating K:*

$$\bar{x} = \lim_{n \to \infty} K^n(x)$$

In addition, we have the following estimate,

$$||K^n(x) - \bar{x}|| \leq \frac{\theta^n}{1 - \theta} ||K(x) - x||$$

valid for any $x \in C$, regarding the convergence $K^{(n)}(x) \to \bar{x}$.

PROOF. As explained in the above, the uniqueness of the fixed point is clear, coming from our assumption $\theta < 1$. Regarding now the existence part, and the precise estimate in the statement too, pick $x = x_0 \in C$, and set $x_n = K^n(x_0)$. We have then:

$$\begin{aligned} ||x_{n+1} - x_n|| &\leq \theta ||x_n - x_{n-1}|| \\ &\leq \theta^2 ||x_{n-1} - x_{n-2}|| \\ &\vdots \\ &\leq \theta^n ||x_1 - x_0|| \end{aligned}$$

Now by using the triangle inequality, we obtain from this, for $n > m$:

$$\begin{aligned} ||x_n - x_m|| &\leq \sum_{j=m+1}^{n} ||x_j - x_{j-1}|| \\ &\leq \theta^m \sum_{j=0}^{n-m-1} \theta^j ||x_1 - x_0|| \\ &\leq \frac{\theta^m}{1 - \theta} ||x_1 - x_0|| \end{aligned}$$

Thus the sequence $\{x_n\}$ is Cauchy, and since we are in a Banach space, this sequence converges. Moreover, since $C \subset V$ was chosen closed, the limit belongs to C:

$$x_n \to \bar{x} \in C$$

Now since our map K was assumed to be a contraction, it is continuous, and by continuity we obtain, as desired, that we have indeed a fixed point, due to:

$$||K(\bar{x}) - \bar{x}|| = \lim_{n\to\infty} ||x_{n+1} - x_n|| = 0$$

Finally, in what regards the estimate at the end, in the statement, let us go back to the main estimate obtained before, which was as follows, for any $n > m$:

$$||x_n - x_m|| \leq \frac{\theta^m}{1-\theta}||x_1 - x_0||$$

But this gives, with $m \to \infty$, the estimate in the statement, as desired. □

Now by getting back to our ODE questions, recall from before that the map which was needing fixed points was as follows:

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

Thus, we are led into the question on whether such a map K is a contraction or not. In order to discuss this, let us introduce the following technical definition:

DEFINITION 6.20. *A map $f \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, is called locally Lipschitz with respect to x, uniformly with respect to t, if for any $V \subset U$ compact we have*

$$\frac{|f(t,x) - f(t,y)|}{||x - y||} \leq L$$

for any $(t, x) \neq (t, y) \in V$, for a certain number $L \in (0, \infty)$.

Observe that in the case $L \leq 1$, our map is a contraction, at any t. Now with this notion in hand, we can formulate, following Cauchy-Lipschitz and Picard-Lindelöf:

THEOREM 6.21. *An equation as follows, with $f \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, has a unique local solution,*

$$\dot{x} = f(t, x) \quad , \quad x(t_0) = x_0$$

provided that f is locally Lipschitz with respect to x, uniformly with respect to t.

PROOF. Consider, as already indicated above, the following map:

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

We assume for simplifying $t_0 = 0$. In order to verify that K is a contraction, for $t > 0$ small, consider the following Banach space, with $T > 0$ to be determined later:

$$V = C(I, \mathbb{R}^N) \quad , \quad I = [0, T]$$

Let also $\delta > 0$, and consider the following closed ball, inside this space V:

$$C = \bar{B}_\delta(x_0)$$

We would like to apply Theorem 6.19, and in order to do so, we need to check two things, namely that we have indeed $K : C \to C$, and that K is a contraction.

(1) Let us first check that we have $K : C \to C$. For this purpose, let us set:

$$W = [0, T] \times C \subset U$$

We have then the following estimate, coming from definitions:

$$\begin{aligned} |K(x)(t) - x_0| &\leq \int_0^t |f(s, x(s))| ds \\ &\leq t \max_{(t,v) \in W} |f(t, x)| \end{aligned}$$

In view of this, consider the number appearing on the right, namely:

$$M = \max_{(t,v) \in W} |f(t, x)|$$

With this notation, we conclude from our estimate above that we have:

$$TM \leq \delta \implies |K(x)(t) - x_0| \leq \delta, \ \forall t \in [0, T]$$

On the other hand, inside the Banach space $C([0, T], \mathbb{R}^N)$, we have:

$$||K(x) - x_0|| = \sup_{t \in [0,T]} |K(x)(t) - x_0|$$

Thus, under the above assumption $TM \leq \delta$, the following happens:

$$||K(x) - x_0|| \leq \delta$$

But this shows that we have $K(x) \in \bar{B}_\delta(x_0) = C$, and so that we have, as desired:

$$K : C \to C$$

(2) With this done, let us turn now to the second check, that of the fact that our linear map K is indeed a contraction. For this purpose, we use the Lipschitz property of f from the statement, or rather from Definition 6.20, namely:

$$\frac{|f(t, x) - f(t, y)|}{||x - y||} \leq L$$

By using this, and integrating, we obtain the following estimate:

$$\begin{aligned}\int_0^t |f(s,x(s)) - f(s,y(s))|ds &\leq L\int_0^t |x(s) - y(s)|ds\\ &\leq Lt \sup_{0\leq s\leq t} |x(s) - y(s)|\end{aligned}$$

Thus, in terms of our linear map K, we have the following estimate:

$$||K(x) - K(y)|| \leq LT||x - y||$$

But this shows that, with $T \leq 1/L$, we have indeed a contraction, as desired.

(3) Summarizing, we have shown that we have $K : C \to C$, and that this map is a contraction. Thus Theorem 6.19 applies, and gives the result. □

Before getting into further theory, let us discuss a simple application of the above. Consider the following linear equation, that we certainly know how to solve:

$$\dot{x} = x \quad , \quad x(0) = 1$$

Observe that $f(t,x) = x$ is indeed Lipschitz as in Definition 6.20, with $L = 1$. Regarding now the linear map K, this is given by the following formula:

$$\begin{aligned}K(x)(t) &= x_0 + \int_{t_0}^t f(s,x(s))ds\\ &= 1 + \int_0^t x(s)ds\end{aligned}$$

By choosing now $y = 1$ as starting point, the iteration goes as follows:

$$K(y) = 1 + \int_0^t 1ds = 1 + t$$

$$K^2(y) = 1 + \int_0^t (1 + s)ds = 1 + t + \frac{t^2}{2}$$

$$K^3(y) = 1 + \int_0^t \left(1 + s + \frac{s^2}{2}\right) ds = 1 + t + \frac{t^2}{2} + \frac{t^3}{6}$$

$$\vdots$$

Thus we obtain in the limit, as we should, the following solution:

$$K^\infty(y) = \sum_{n=0}^\infty \frac{t^n}{n!} = e^t$$

There are of course many other illustrations, and more on this later.

6d. Gronwall estimates

Getting now to technical comments, in relation with Theorem 6.21, many things can be said here, and here are two of them, which are of particular importance:

(1) In the context of Theorem 6.21, it is possible to prove that if $f \in C^k(U, \mathbb{R}^N)$ with $k \geq 1$, then the solution is C^{k+1}. This is indeed elementary, by recurrence on k.

(2) Also in the context of Theorem 6.21, assume that $[t_0, T] \times \mathbb{R}^N \subset U$ is such that:

$$\int_{t_0}^{T} L(t)dt < \infty \quad , \quad L(t) = \sup_{x \neq y \in \mathbb{R}^N} \frac{|f(t,x) - f(t,y)|}{||x - y||}$$

Then, by suitably changing the Banach space norm, and suitably modifying the contraction principle too, it is possible to prove that the solution is defined on $[t_0, T]$.

We refer to the ODE literature for more on the above, which is something quite standard. As a main question now that we would like to solve, we have:

QUESTION 6.22. *How does the solution depend on the initial data, and also, on the equation itself?*

This is something quite general. To be more precise, assume that we have two functions $f, g \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, which are both locally Lipschitz with respect to x, uniformly with respect to t. In this case Theorem 6.21 applies to the following two equations, and provides us with local solutions x, y to them, which are unique:

$$\dot{x} = f(t, x) \quad , \quad x(t_0) = x_0$$

$$\dot{y} = g(t, y) \quad , \quad y(t_0) = y_0$$

The problem that we would like to solve is that of finding an estimate for the quantity $||x(t) - y(t)||$. And, we will prove in what follows that we have indeed such an estimate, which looks as follows, with $M, L > 0$ being certain constants, depending on f, g:

$$||x(t) - y(t)|| \leq ||x_0 - y_0||e^{L|t-t_0|} + \frac{M}{L}\left(e^{L|t-t_0|} - 1\right)$$

Obviously, such things are of key importance, in relation with our order vs chaos problematics. However, such questions non-trivial, and our tools so far, which are quite abstract, do not provide a direct answer to them. So, we have to work some more.

In order to solve our question, let us begin with a key technical statement, of classical analysis type, not obviously related to equations, due to Gronwall, as follows:

PROPOSITION 6.23. *Assume that a function ψ satisfies the estimate*

$$\psi(t) \leq \alpha(t) + \int_0^t \beta(s)\psi(s)ds$$

for any $t \in [0,T]$, with $\alpha(t) \in \mathbb{R}$, and $\beta(t) > 0$. We have then

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s) \exp\left(\int_s^t \beta(r)dr\right) ds$$

for any $t \in [0,T]$. Moreover, assuming that α is increasing, we have

$$\psi(t) \leq \alpha(t) \exp\left(\int_0^t \beta(s)ds\right)$$

for any $t \in [0,T]$.

PROOF. This is something quite tough, and for the story, it happened to me more than once, when teaching this to our graduate math students in Cergy, for one student to leave the class during or after the proof, in protest, never to be seen again. Well, in the hope that these protesting kids got some friends, spouses and jobs, not quite sure about that, and here is the proof of the result, that I personally find quite cute:

(1) Let us first prove the first assertion, which is the main one. For this purpose, we use a trick. Consider the following function:

$$\phi(t) = \exp\left(-\int_0^t \beta(s)ds\right)$$

We have then the following computation, using the Leibnitz rule for derivatives, and also using at the end our assumption on ψ from the statement:

$$\begin{aligned}
& \frac{d}{dt}\left[\phi(t)\int_0^t \beta(s)\psi(s)ds\right] \\
= & \left[\frac{d}{dt}\phi(t)\right]\int_0^t \beta(s)\psi(s)ds + \phi(t)\left[\frac{d}{dt}\int_0^t \beta(s)\psi(s)ds\right] \\
= & -\beta(t)\phi(t)\int_0^t \beta(s)\psi(s)ds + \phi(t)\beta(t)\psi(t) \\
= & \beta(t)\psi(t)\left(\psi(t) - \int_0^t \beta(s)\psi(s)ds\right) \\
\leq & \alpha(t)\beta(t)\phi(t)
\end{aligned}$$

Now by integrating with respect to t, we obtain from this:

$$\phi(t)\int_0^t \beta(s)\psi(s)ds \leq \int_0^t \alpha(s)\beta(s)\phi(s)ds$$

We conclude that we have the following estimate:

$$\int_0^t \beta(s)\psi(s)ds \leq \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds$$

By adding now $\alpha(t)$ to both sides, we obtain the following estimate:

$$\alpha(t) + \int_0^t \beta(s)\psi(s)ds \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds$$

But in this situation, we can use once again our assumption on ψ from the statement, and we obtain the following estimate:

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds$$

Now let us look at the fraction on the right. This is given by:

$$\begin{aligned}
\frac{\phi(s)}{\phi(t)} &= \frac{\exp\left(-\int_0^s \beta(r)dr\right)}{\exp\left(-\int_0^t \beta(r)dr\right)} \\
&= \exp\left(\int_0^t \beta(r)dr - \int_0^s \beta(r)dr\right) \\
&= \exp\left(\int_s^t \beta(r)dr\right)
\end{aligned}$$

We conclude that the estimate that we found above reads:

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds$$

But this is precisely what we wanted to prove, the first estimate in the statement.

(2) With this done, let us turn now to the second assertion in the statement. So, assume that the function α there is increasing. We have then:

$$\begin{aligned}
\psi(t) &\leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds \\
&\leq \alpha(t) + \int_0^t \alpha(t)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds \\
&= \alpha(t)\left[1 + \int_0^t \beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds\right] \\
&= \alpha(t)\left[1 + \int_0^t \beta(s)\exp\left(\int_0^t \beta(r)dr - \int_0^s \beta(r)dr\right)ds\right] \\
&= \alpha(t)\left[1 + \exp\left(\int_0^t \beta(r)dr\right)\int_0^t \beta(s)\exp\left(-\int_0^s \beta(r)dr\right)ds\right]
\end{aligned}$$

Now recall that we can consider, as in (1), the following function:

$$\phi(t) = \exp\left(-\int_0^t \beta(s)ds\right)$$

The derivative of this function satisfies then the following formula:

$$\phi'(t) = -\beta(t)\phi(t)$$

Thus, we have the following formula, for this derivative:

$$\phi'(s) = -\beta(s)\exp\left(-\int_0^s \beta(r)dr\right)$$

We conclude that the estimate found before reformulates as:

$$\begin{aligned}
\psi(t) &\leq \alpha(t)\left[1+\exp\left(\int_0^t \beta(r)dr\right)\int_0^t \beta(s)\exp\left(-\int_0^s \beta(r)dr\right)ds\right]\\
&= \alpha(t)\left[1+\exp\left(\int_0^t \beta(r)dr\right)(-\phi')\Big|_0^t\right]\\
&= \alpha(t)\left[1+\exp\left(\int_0^t \beta(r)dr\right)(1-\phi(t))\right]
\end{aligned}$$

In order to finish, consider the following number, depending on t:

$$K = \int_0^t \beta(r)dr$$

In terms of this number, the estimate that we found above reads:

$$\begin{aligned}
\psi(t) &\leq \alpha(t)(1+e^K(1-e^{-K}))\\
&= \alpha(t)(1+e^K-1)\\
&= \alpha(t)e^K
\end{aligned}$$

Thus, as a conclusion, we have reached to the following estimate:

$$\psi(t) \leq \alpha(t)\exp\left(\int_0^t \beta(s)ds\right)$$

But this is exactly what we wanted to prove, namely second estimate in the statement, and so good news, eventually, done with the proof of the present statement. □

Very good all this, welcome to analysis, and still with me, I hope.

As a continuation now of the above, we won't leave such beautiful things like this, we would definitely love to spend more time with them, we have:

PROPOSITION 6.24. *Assume that a function ψ satisfies the estimate*

$$\psi(t) \leq \alpha(t) + \int_0^t (\beta\psi(s) + \gamma)ds$$

for any $t \in [0,T]$, with $\alpha \in \mathbb{R}$, $\beta \geq 0$ and $\gamma \in \mathbb{R}$. We have then

$$\psi(t) \leq \alpha \exp(\beta t) + \frac{\gamma}{\beta}(\exp(\beta t) - 1)$$

for any $t \in [0,T]$.

PROOF. In order to prove this result, consider the following function:

$$\widetilde{\psi}(t) = \psi(t) + \frac{\gamma}{\beta}$$

In terms of this function $\widetilde{\psi}$, our assumption on ψ in the statement reads:

$$\widetilde{\psi} - \frac{\gamma}{\beta} \leq \alpha + \beta \int_0^t \widetilde{\psi}(s)ds$$

Thus, our modified function $\widetilde{\psi}$ satisfies the following estimate:

$$\widetilde{\psi} \leq \left(\alpha + \frac{\gamma}{\beta}\right) + \beta \int_0^t \widetilde{\psi}(s)ds$$

Thus, we can apply the second assertion in Proposition 6.23, with the following values for the functions $\alpha(t)$ and $\beta(t)$ there, both chosen to be constant functions:

$$\alpha(t) = \alpha + \frac{\gamma}{\beta} \quad , \quad \beta(t) = \beta$$

We obtain in this way the following estimate, for our modified function $\widetilde{\psi}$:

$$\widetilde{\psi} \leq \left(\alpha + \frac{\gamma}{\beta}\right) \exp(\beta t)$$

But this gives, in terms of the original function ψ, the following estimate:

$$\begin{aligned} \psi(t) &\leq \left(\alpha + \frac{\gamma}{\beta}\right) \exp(\beta t) - \frac{\gamma}{\beta} \\ &= \alpha \exp(\beta t) + \frac{\gamma}{\beta}(\exp(\beta t) - 1) \end{aligned}$$

Thus, we have reached to the conclusion in the statement. □

Now back to the ODE, the above results apply, and we can answer Question 6.22. To be more precise, in the general context of Theorem 6.21, we have the following result:

THEOREM 6.25. *Assume that $f, g \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, are locally Lipschitz with respect to x, uniformly with respect to t. If x, y are solutions of*

$$\dot{x} = f(t, x) \quad , \quad x(t_0) = x_0$$

$$\dot{y} = g(t, y) \quad , \quad y(t_0) = y_0$$

then we have the following estimate, for any t in the interval of definition of x, y,

$$||x(t) - y(t)|| \leq ||x_0 - y_0||e^{L|t-t_0|} + \frac{M}{L}\left(e^{L|t-t_0|} - 1\right)$$

with the constant M on the right being given by the following formula,

$$M = \sup_{(t,x)\in U} |f(t, x) - g(t, x)|$$

and with $L > 0$ being a common Lipschitz constant for both f, g.

PROOF. We know from Theorem 6.21 that the above equations have indeed solutions. We can assume for simplifying that we have $t_0 = 0$. Now observe that we have:

$$\begin{aligned}
& ||x(t) - y(t)|| \\
\leq\ & ||x_0 - y_0|| + \int_0^t |f(s, x(s)) - g(s, y(s))|ds \\
\leq\ & ||x_0 - y_0|| + \int_0^t \Big(|f(s, x(s)) - f(s, y(s))| + |f(s, y(s)) - g(s, y(s))|\Big)ds \\
\leq\ & ||x_0 - y_0|| + \int_0^t \Big(L||x(s) - y(s)|| + M\Big)ds
\end{aligned}$$

In view of this estimate, consider the following function:

$$\psi(t) = ||x(t) - y(t)||$$

In terms of this function, the estimate that we found above reads:

$$\psi(t) \leq ||x_0 - y_0|| + \int_0^t (L\psi(s) + M)ds$$

But this shows that the Gronwall estimate from Proposition 6.24 applies, with the following choices for the constants $\alpha \in \mathbb{R}$, $\beta \geq 0$ and $\gamma \in \mathbb{R}$ appearing there:

$$\alpha = ||x_0 - y_0|| \quad , \quad \beta = L \quad , \quad \gamma = M$$

So, let us apply Proposition 6.24, with these values of α, β, γ. We obtain:

$$\begin{aligned}
\psi(t) &\leq \alpha \exp(\beta t) + \frac{\gamma}{\beta}(\exp(\beta t) - 1) \\
&= ||x_0 - y_0||e^{L|t-t_0|} + \frac{M}{L}\left(e^{L|t-t_0|} - 1\right)
\end{aligned}$$

But this is exactly the estimate in the statement, as desired. □

6e. Exercises

Exercises:

Exercise 6.26.

Exercise 6.27.

Exercise 6.28.

Exercise 6.29.

Exercise 6.30.

Exercise 6.31.

Exercise 6.32.

Exercise 6.33.

Bonus exercise.

CHAPTER 7

Dynamical systems

7a. Dynamical systems

Generally speaking, a dynamical system is an action of a semigroup $(G, \cdot)$ on a space M. That is, we must have a map as follows, satisfying $T_g T_h = T_{gh}$, for any $g, h \in G$:

$$T : G \times M \to M \quad , \quad (g, x) \to T_g(x)$$

All this is quite general. As a first remark, the dynamical systems fall into two classes, namely discrete, where $G = \mathbb{N}, \mathbb{Z}, \ldots$, and continuous, where $G = \mathbb{R}^+, \mathbb{R}, \ldots$

The discrete systems are quite easy to construct. Indeed, as a basic example here, with $G = \mathbb{N}$, you can take any function $f : I \to I$, and then set $T_n = f^n$, for any $n \in \mathbb{N}$. Observe that when the function f is invertible, we can extend this into a system with $G = \mathbb{Z}$, again by setting $T_n = f^n$, but this time for any $n \in \mathbb{Z}$.

In what follows we will be mainly interested in continuous dynamical systems, with $G = \mathbb{R}$, coming from the ODE. To be more precise, consider, as in the previous chapter, an autonomous system as follows, with $f \in C^k(M, \mathbb{R}^N)$, and $M \subset \mathbb{R}^N$ open:

$$\dot{x} = f(x) \quad , \quad x(0) = x_0$$

Assuming that the system has solutions, coming for instance via the general existence results from the previous chapter, we call these solutions integral curves of the system.

The point now is that, for any initial data $x = x_0$, we can talk about the maximal integral curve of our system, passing through x, which is by definition the solution on the maximal possible interval, that we will denote, as usual, as follows:

$$I_x = (T_-(x), T_+(x))$$

With this notion in hand, consider now the following space $W \subset \mathbb{R} \times M$:

$$W = \bigcup_{x \in M} I_x \times \{x\}$$

The solutions φ of our system are then encoded into a map, as follows:

$$\Phi : W \to M \quad , \quad (t, x) \to \varphi(t, x)$$

We will call this map Φ, encoding the solutions, the flow of the system.

The flow has a number of basic properties, which can be summarized as follows:

PROPOSITION 7.1. *The flow of a system $\dot{x} = f(x)$, $x(0) = x_0$, written as*

$$\Phi : W \to M \quad , \quad (t, x) \to \varphi(t, x)$$

with $W = \bigcup_{x \in M} I_x \times \{x\}$ as above, has the following properties:

(1) *Its domain W is open.*
(2) *If $f \in C^k(M, \mathbb{R}^N)$ then $\Phi \in C^k(W, M)$.*
(3) *We have $\Phi(0, x) = x$, for any x.*
(4) *Flow property: $\Phi(s + t, x) = \Phi(s, \Phi(t, x))$.*

PROOF. These are all obvious properties, coming from our general existence and uniqueness results for the solutions, that we assumed, as explained above, to apply. □

In relation now with the abstract notion of dynamical system, as axiomatized before, assuming that we have $I_x = \mathbb{R}$ for any x, we can consider the following map:

$$\Phi_t(x) = \Phi(t, x)$$

Then, according to the flow property, (4) above, we have, for any s, t:

$$\begin{aligned} \Phi_{s+t}(x) &= \Phi(s + t, x) \\ &= \Phi(s, \Phi(t, x)) \\ &= \Phi_s(\Phi(t, x)) \\ &= \Phi_s \Phi_t(x) \end{aligned}$$

Thus, we have indeed an abstract dynamical system, with $G = \mathbb{R}$.

With this discussed, let us formulate now a key definition, as follows:

DEFINITION 7.2. *The orbit of a point x is the following set:*

$$\gamma(x) = \Phi(I_x, x) \subset M$$

We say that x is a fixed point when $\gamma(x) = x$, and that x is regular, otherwise.

As a first remark about the orbits, as constructed above, observe that these are by definition disjoint. Thus, we have an equivalence relation, given by:

$$x \simeq y \iff \gamma(x) = \gamma(y)$$

In what follows we will need as well a refinement of this. With $I_x = (T_-(x), T_+(x))$, as usual, let us define the backwards orbit of x as being the following set:

$$\gamma_-(x) = \Phi((T_-(x), 0), x)$$

Similarly, we can talk about the forward orbit of x, as being the following set:

$$\gamma_+(x) = \Phi((0, T_+(x)), x)$$

We have then the following equality, coming from definitions:

$$\gamma(x) = \gamma_-(x) \cup \{x\} \cup \gamma_+(x)$$

As yet another related notion, that we will need, let us introduce:

DEFINITION 7.3. *We say that x is a periodic point when $\Phi(T, x) = x$, for some $T > 0$. In this case we call the number*

$$T(x) = \inf \left\{ T > 0 \middle| \Phi(T, x) = x \right\}$$

the period of our point x.

Observe that, due to the flow property in Proposition 7.1 (4), under the above circumstances, the orbit is indeed periodic, of period T, as shown by:

$$\Phi(t + T(x), x) = \Phi(t, x)$$

As another remark, in terms of the backwards orbit $\gamma_-(x)$ and forward orbit $\gamma_+(x)$, constructed above, the fact that the point is periodic is equivalent to:

$$\gamma_-(x) \cap \gamma_+(x) \neq \emptyset$$

Indeed, this follows again from the flow property. Finally, observe too that when x is periodic, any point in its orbit $\gamma(x)$ is periodic too, and of the same period.

As a summary to this preliminary discussion about orbits, let us formulate:

CONCLUSION 7.4. *The points of our dynamical system can be of 3 types:*

(1) *Fixed points, $\gamma(x) = \{x\}$.*
(2) *Regular periodic points, $0 < T < \infty$.*
(3) *Non-periodic points.*

Hang on, still not done with the definitions, several more still to come. We first have the following notions, which are something quite useful as well:

DEFINITION 7.5. *We say that a point x is:*

(1) *$+$ complete, if $T_+(x) = \infty$.*
(2) *$-$ complete, if $T_-(x) = -\infty$.*
(3) *Complete, if it is both $+$ and $-$ complete.*

Observe that any periodic point is complete, since we can indefinitely travel forward, or backwards, on its orbit. The converse of this does of course not hold, in general.

As another remark, in relation with the above notions, when our system is complete, in the sense that any point is complete, then, as explained in the discussion following Proposition 7.1, we have an abstract dynamical system, with $G = \mathbb{R}$.

Here is now a key definition, which will be of importance, in what follows:

DEFINITION 7.6. *We say that a subset $U \subset M$ is:*

(1) *+ invariant, if $x \in U \implies \gamma_+(x) \subset U$.*
(2) *− invariant, if $x \in U \implies \gamma_-(x) \subset U$.*
(3) *Invariant, if it is both + and − invariant.*

Obviously, this is something which will allow us to do some geometry.

As a preliminary remark here, the above various types of invariant sets are stable under taking intersections, taking unions, and under taking closures too. Indeed, these properties are clear from definitions. We will use them many times, in what follows.

Moving ahead, let us attempt now to study the orbits, in the non-periodic case. In order to do so, we will need one more definition, as follows:

DEFINITION 7.7. *We let $w_\pm(x)$ be the set of points $y \in M$ satisfying:*

$$\exists t_n \to \pm\infty \quad , \quad \Phi(t_n, x) \to y$$

That is, $w_\pm(x)$ are the limit points of the forward/backwards orbit of x.

Observe that, for a periodic point, both the above sets $w_\pm(x)$ coincide with the orbit. Also, when our point is not complete, these sets $w_\pm(x)$ are both empty.

And with this, good news, end of definitions, and time now for some theorems. As a first result, in relation with the sets $w_\pm(x)$ introduced above, we have:

PROPOSITION 7.8. *The sets $w_\pm(x)$ are closed, and invariant.*

PROOF. We have two things to be proved, the idea being as follows:

(1) Closedness. Consider indeed a point in the closure of one of our sets:

$$y \in \overline{w_\pm(x)}$$

Thus, for any $n \in \mathbb{N}$, we can find a point $y_n \in w_\pm(x)$ such that:

$$|y - y_n| < \frac{1}{2n}$$

According now to our definition of $w_\pm(x)$, we can find $t_n \to \pm\infty$ such that:

$$|\Phi(t_n, x) - y_n| < \frac{1}{2n}$$

We therefore obtain, by adding, the following inequality:

$$|\Phi(t_n, x) - y_n| < \frac{1}{n}$$

But this shows, with $n \to \infty$, that we have $y \in w_\pm(x)$, as desired.

(2) Invariance. Assume indeed that we have $\Phi(t_n, x) \to y$. We obtain:

$$\begin{aligned}\Phi(t_n + t, x) &= \Phi(t, \Phi(t_n, x)) \\ &\to \Phi(t, y) \\ &\in \gamma(y)\end{aligned}$$

Thus the orbit of y is included in $w_\pm(x)$, as desired. □

As a question now, we would like to understand if for a complete point, the sets $w_\pm(x)$ are empty or not. For this purpose, let us see what happens for the following equation:

$$\dot{x} = -x$$

Here the solutions are very easy to find, given by the following formula:

$$x = e^{-t}x_0$$

Regarding now the sets $w_\pm(x)$, observe first that we have $w_+(x) = \{0\}$, for any $x \in \mathbb{R}$, and this because $e^{-t}x \to 0$ with $t \to \infty$. Similarly, we have $w_-(x) = \emptyset$, for any $x \neq 0$, and this because $e^{-t}x \to \pm\infty$ with $t \to -\infty$, for $x \neq 0$. Thus, as a conclusion:

Conclusion 7.9. *We can have complete points with $w_\pm(x) = \emptyset$.*

Getting now to more general theory, we have the following result, coming as a complement to what we already know about the sets $w_\pm(x)$, from Proposition 7.8:

Theorem 7.10. *If $\gamma_\pm(x) \subset C$, compact, then the set $w_\pm(x)$ is:*

(1) *Non-empty.*
(2) *Compact.*
(3) *Connected.*

Proof. The first two assertions, regarding the non-emptiness and the compactness, are both clear. Regarding now the connectedness assertion, assume that, inside our compact set C, we can separate $w_\pm(x)$ by two parts U_1, U_2, lying at distance $\delta > 0$. Now let us pick an increasing or decreasing sequence $t_n \to \pm\infty$, such that:

$$\Phi(t_{2n+1}, x) \in U_1 \quad , \quad \Phi(t_{2n}, x) \in U_2$$

Since $\Phi((t_{2n}, t_{2n+1}), x)$ is connected, we can find $t_n' \in (t_{2n}, t_{2n+1})$ such that:

$$\Phi(t_n', x) \in C - (U_1 \cup U_2)$$

But, by choosing a suitable subsequence of $\{t_n'\}$, this would give us a limit point, which cannot be in U_1, nor in U_2, and so we have here a contradiction, as desired. □

7b. Stability issues

Getting now to the whole point with the dynamical systems, let us discuss stability issues. We will be interested in the stable points, in the following sense:

DEFINITION 7.11. *Let x_0 be a fixed point.*

(1) *We say that x_0 is stable if, for any neighborhood $U(x_0)$, there is a smaller neighborhood $V(x_0) \subset U(x_0)$ such that any solution departing from a point of $V(x_0)$ remains inside $U(x_0)$, at any time $t \geq 0$.*
(2) *Also, we say that x_0 is asymptotically stable if it is stable, in the above sense, and in addition, there is a neighborhood $U(x_0)$ such that $\lim_{t\to\infty} |\Phi(t,x) - x_0| = 0$ holds, for any point $x \in U(x_0)$.*

These notions are both quite intuitive, and of obvious interest, when thinking for instance mechanics. As an illustration, consider the following equation:

$$\dot{x} = ax$$

The solution of this equation is then trivial to find, given by:

$$x = e^{at}x_0$$

We have a fixed point, $x_0 = 0$, and according to our conventions above, this fixed point is stable when $a \leq 0$, and asymptotically stable when $a < 0$.

In order to study the stable points, we will need the following key notion:

DEFINITION 7.12. *Let x_0 be a fixed point, and $U(x_0)$ be an open neighborhood of it. A Lyapunov function for x_0, on $U(x_0)$, is a continuous function*

$$L : U(x_0) \to \mathbb{R}$$

satisfying $L(x_0) = 0$, and $L(x) > 0$ for $x \neq x_0$, and which is such that

$$t_0 < t_1 \implies L(\Phi(t_0)) \geq L(\Phi(t_1))$$

for any solution Φ, provided that $\Phi(t_j) \in U(x_0) - \{x_0\}$. That is, L must decrease on the integral curves. We say that L is strict, when the above inequality is strict.

As a first observation, assuming that a Lyapunov function as above exists, the set $U(x_0) - \{x_0\}$ contains no periodic orbits. This is indeed clear from definitions.

The interest in the Lyapunov functions comes from the following key result:

THEOREM 7.13. *Assuming that x_0 is fixed, and that a Lyapunov function as above*

$$L : U(x_0) \to \mathbb{R}$$

exists, then x_0 must be stable.

PROOF. This is something quite technical, the idea being as follows:

(1) For any $\delta > 0$, let us denote by S_δ the connected component containing x_0 of the following set, which contains indeed x_0, according to our conventions above:

$$\left\{x \in U(x_0)\middle| L(x) \leq \delta\right\}$$

Our first claim is then that the following happens:

$$\forall \delta > 0, \exists \varepsilon > 0, S_\varepsilon \subset B_\delta(x_0)$$

(2) We will prove this claim by contradiction. So, assume by contradiction that, for a certain fixed $\delta > 0$, for any $n \in \mathbb{N}$ we can find $x_n \in S_{1/n}$, such that:

$$|x_n - x_0| \geq \delta$$

Now recall that $S_{1/n}$ was chosen connected. Thus, we can assume that we have:

$$|x_n - x_0| = \delta$$

Now since the spheres are compact, we can assume that our sequence is convergent:

$$x_n \to y$$

By using now the continuity of the Lyapunov function L, we have:

$$L(x_n) \to L(y)$$

On the other hand, from our assumption $x_n \in S_{1/n}$, we know that we have:

$$L(x_n) \leq \frac{1}{n}$$

Thus $L(y) = 0$, and so $y = x_0$, which contradicts $|y - x_0| = \delta > 0$, as desired.

(3) Our next claim now is that, conversely, the following happens:

$$\forall \delta > 0, \exists \varepsilon > 0, B_\varepsilon(x_0) \subset S_\delta$$

Again, we will prove this by contradiction. So, assume by contradiction that we can find a sequence of points x_n satisfying the following two conditions:

$$|x_n - x_0| \leq \frac{1}{n} \quad , \quad L(x_n) > \delta$$

We therefore obtain, by taking the $n \to \infty$ limit, in this situation:

$$\delta \leq \lim_{n \to \infty} L(x_n) = L(x_0) = 0$$

But this is a contradiction, as desired, and so our present second claim is proved.

(4) With the above in hand, namely the claims in (1) and (3), we can now finish the proof. Assume indeed that x_0 is a fixed point, and that $U(x_0)$ is a neighborhood for it, with a Lyapunov function. We fix $\delta > 0$, such that the following happens:

$$B_\delta(x_0) \subset U(x_0)$$

By using now our claims in (1) and (3), we can find $\varepsilon, \varepsilon' > 0$ such that:

$$S_\varepsilon \subset B_\delta(x_0) \quad , \quad B_{\varepsilon'}(x_0) \subset S_\varepsilon$$

In order to prove the theorem, that is, in order to prove that x_0 is indeed stable, in the sense of Definition 7.11, consider the following neighborhood of x_0:

$$V(x_0) = B_{\varepsilon'}(x_0)$$

We will show in what follows that any solution departing from a point of $V(x_0)$ stays in $U(x_0)$, and so that x_0 is indeed stable, in the sense of Definition 7.11.

(5) In order to do so, it is enough to show that our solution stays inside S_ε, and this because, according to our various choices above, we have inclusions as follows:

$$S_\varepsilon \subset B_\delta(x_0) \subset V(x_0)$$

Now since we know as well that we have $B_{\varepsilon'}(x_0) \subset S_\varepsilon$, it is enough to prove that S_ε is $+$ invariant. Thus, as a conclusion to all this, we must show that S_ε is $+$ invariant.

(6) We will prove this, as usual, by contradiction. So, assume by contradiction that the solution $\Phi(t)$ exists our set S_δ at a certain time $t_0 > 0$. We then set:

$$x = \Phi(t_0)$$

With this done, we can then find a suitable ball $B_r(x) \subset U(x_0)$, around our point x, such that the following happens, for $\varepsilon > 0$ small:

$$\Phi(t_0 + \varepsilon) \in B_r(x) - S_\delta$$

But this shows that, in relation with the Lyapunov function L, we have:

$$\Phi(t_0 + \varepsilon) \notin S_\delta = \left\{ x \in U(x_0) \middle| L(x) \leq \delta \right\}$$

In order words, we have reached here to the following conclusion:

$$L(\Phi(t_0 + \varepsilon)) > \delta$$

But this contradicts our assumption on the Lyapunov function L, from Definition 7.12, that this must decrease on the integral curves. Thus, our theorem is now proved. □

And with this, good news, end of our general theoretical discussion regarding the dynamical systems, in general. In what follows we will go towards more concrete questions, and we will also see, of course, some illustrations for the above general results.

7c. Integral equations

In mathematics, it all comes down to linearization. You surely know about this general principle, for instance from basic calculus, where the functions, be them of one or several variables, can be thought of as being locally linear, with the help of the derivative.

Further illustrations of this general linearization principle include the fact that the smooth manifolds are locally linear too, with the help of the tangent space. And even complicated beasts like continuous groups of transformations are locally linear too, again with the help of the same ideas, namely tangent vectors and spaces.

We discuss in this section, and in the remainder of this chapter, a very fruitful linearization idea, in the context of the dynamical systems, as follows:

IDEA 7.14 (Linearization). *In order to deal with an arbitrary, non-linear system*

$$\dot{x} = f(x) \quad , \quad x_0 = 0$$

we can write the function f as follows, with $A = f'(x) \in M_N(\mathbb{R})$ being its derivative,

$$f(x) = Ax + o(||x||)$$

and then use, by perturbing, the results regarding the linear system $\dot{x} = Ax$.

Which sounds very good, normally this type of idea will lead us into classical and rock-solid mathematics, as classical and rock-solid mathematics can get.

Before getting head-first into this, however, let us go back to the general theory of the linear equations $\dot{x} = Ax$, as developed in detail in chapters 5-6. Obviously, in view of our above idea, what we need to do is to scan the material there, in search for things that can be perturbed into results about non-linear systems $\dot{x} = f(x)$, as above.

But, and here comes the point, what we did in chapters 5-6 is not exactly satisfying, from this perspective, and so, good news, we will have to work some more.

So, working some more on the linear systems $\dot{x} = Ax$, with the above type of ideas in mind. To start with, the solution of the system is, as we know well:

$$x(t) = e^{tA}x_0$$

But this leads us, as explained in chapters 5-6, into diagonalizing A, and more generally, when this is not possible, into putting A into Jordan form.

Now based on this, what we know about the Jordan form, let us formulate:

NOTATIONS 7.15. *Given $A \in M_N(\mathbb{R})$, we write its characteristic polynomial as*

$$P(z) = \prod_i (z - \alpha_i)^{a_i}$$

so that we have a direct sum decomposition of the ambient space, as follows:

$$\mathbb{C}^N = \bigoplus_i \ker\left[(A - \alpha_i)^{a_i}\right]$$

We also consider the corresponding geometric multiplicities, given by

$$g_i = \dim \ker(A - a_i)$$

and satisfying $g_i \leq a_i$, with equalities when A is diagonalizable.

We refer to chapters 5-6 for more on all this, theory and applications. Now back to our equation $\dot{x} = Ax$, let us formulate the following key definition:

DEFINITION 7.16. *We say that a linear system $\dot{x} = Ax$ is hyperbolic when*

$$Re(\alpha) \neq 0$$

for any eigenvalue α. In this case, we consider the linear spaces

$$E^{\pm} = \bigoplus_{\pm Re(\alpha_i) < 0} \ker\left[(A - \alpha_i)^{a_i}\right]$$

which are therefore in direct sum position, $\mathbb{C}^N = E^+ \oplus E^-$.

So, studying these hyperbolic linear systems, and then extending our results to the hyperbolic non-linear systems, according to Idea 7.14, will be our job, in what follows.

In what regards the study in the linear case, this is something quickly done, by using the general theory developed in chapters 5-6, the result here being as follows:

THEOREM 7.17. *For a hyperbolic linear system $\dot{x} = Ax$, the following happen:*

(1) *The spaces $E^{\pm}$ are both invariant by the flow.*
(2) *Any integral curve departing from $E^{\pm}$ converges to 0, with $t \to \pm\infty$.*
(3) *In fact, we have the following explicit estimate for the decay,*

$$|e^{tA} x_{\pm}| \leq C e^{\pm t\alpha} |x_{\pm}|$$

for any $\pm t > 0$ and any $x_{\pm} \in E^{\pm}$, with $\alpha > 0$ subject to

$$\alpha < \min\left\{|Re(\alpha_i)| : \pm Re(\alpha_i) < 0\right\}$$

and with $C > 0$ depending on α.

PROOF. This is something quite straightforward, the idea being as follows:

(1) This is something which is obvious.

(2) This is something that we already know, as a consequence of our general results from chapters 5-6, and which follows also from (3), that we will prove next.

(3) We will just discuss here the proof of the "+" result, with the proof of the "−" result being similar, or just by replacing $A \to -A$. We put our matrix A in Jordan form, as explained in chapters 5-6, and we consider the following quantity:

$$m = \min\left\{|Re(\alpha_i)| : Re(\alpha_i) < 0\right\}$$

Now let $\alpha < m$ as in the statement, and let us set:

$$\varepsilon = m - \alpha$$

Then, for any eigenvalue satisfying $Re(\alpha_i) < 0$, the entry of maximal absolute value, say M_i, of the corresponding component $e^{tJ_{a_i}}$ of the matrix e^{tA}, appearing by exponentiating t times the corresponding Jordan block J_{a_i}, can be estimated as follows:

$$\begin{aligned} M_i &= \frac{|t^n e^{a_i t}|}{F} \\ &\leq \frac{|t^n e^{-\varepsilon t}|e^{-\alpha t}}{F} \\ &\leq Ce^{-\alpha t} \end{aligned}$$

To be more precise, here F is a certain factorial, namely $F = (s-1)!$, with s being the size of the Jordan block, and $C > 0$ at the end is a certain constant, depending on this nunber F, and on α. Thus, we are led to the conclusion in the statement. □

Good news, with this in hand, we can go back now to the non-linear systems. Indeed, inspired by Idea 7.14, let us formulate the following definition:

DEFINITION 7.18. *We say that a non-linear system*

$$\dot{x} = f(x) \quad , \quad x_0 = 0$$

is hyperbolic when the associated linear system

$$\dot{x} = Ax \quad , \quad x_0 = 0$$

with $A = f'(x) \in M_N(\mathbb{R})$ being the derivative of f, is hyperbolic.

With this done, our goal now will be to extend what we have in Theorem 7.17, to the case of the non-linear hyperbolic systems. But this can be done indeed, with a lot of routine approximation work, and with some differential geometry helping too, the point being that, in the non-linear case, the spaces $E^{\pm}$ considered there become manifolds.

Getting started now, we first have the following result:

THEOREM 7.19. *Consider a non-linear system, written in the following form, with $A = f'(x) \in M_N(\mathbb{R})$ being the derivative of f, and $g(x) = o(|x|)$ being the remainder:*

$$\dot{x} = f(x) \quad , \quad f(x) = Ax + g(x)$$

This system is then equivalent to the following equation,

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-r)A}g(x_r)dr$$

called Volterra integral equation.

PROOF. This is indeed something elementary, which follows from a direct computation, by computing the time derivative of the function in the statement. Let us set:

$$x(t) = e^{tA}C(t)$$

The derivative of this function is then given by the following formula:

$$\dot{x}(t) = Ae^{tA}C(t) + e^{tA}\dot{C}(t)$$

Now observe that this latter formula reads:

$$Ax(t) + g(x(t)) = Ax(t) + e^{tA}\dot{C}(t)$$

We conclude from this that we have the following equality:

$$e^{tA}\dot{C}(t) = g(x(t))$$

Equivalently, we have the following formula:

$$\dot{C}(t) = e^{-tA}g(x(t))$$

Now by integrating, this gives the following formula:

$$C(t) = x(0) + \int_0^t e^{-rA}g(x(r))dr$$

Thus, the solution is given by the following formula:

$$\begin{aligned} x(t) &= e^{tA}\left(x_0 + \int_0^t e^{-rA}g(x(r))dr\right) \\ &= e^{tA}x(0) + \int_0^t e^{(t-r)A}g(x(r))dr \end{aligned}$$

We are therefore led to the conclusion in the statement. □

In order to study now the integral equations as above, we will need some functional analysis tools, namely the following result, called uniform contraction principle:

THEOREM 7.20. *Given a Banach space X, and a closed subset $C \subset X$, assume that we have maps as follows, depending on parameters $\lambda \in \Lambda$, in some Banach space,*

$$K_\lambda : C \to C$$

which are continuous with respect to them, and which are uniform contractions,

$$||K_\lambda(x) - K_\lambda(y)|| \leq \theta ||x - y||$$

with $\theta \in [0, 1)$. Then, these maps have unique fixed points $\bar{x}(\lambda)$, which are continuous with respect to λ. Moreover, if $\lambda_n \to \lambda$, then with $x_{n+1} = K_{\lambda_n}(x_n)$ we have $x_n \to \bar{x}(\lambda)$.

PROOF. The existence and uniqueness of each of the fixed points $\bar{x}(\lambda)$ follows from the usual contraction principle, that we know well from before. Thus, we are left with proving the last assertions, regarding the continuity properties of the following map:

$$\lambda \to \bar{x}(\lambda)$$

(1) Let us first prove that this map is indeed continuous. We have:

$$\begin{aligned} ||\bar{x}(\lambda) - \bar{x}(\eta)|| &= ||K_\lambda(\bar{x}(\lambda)) - K_\eta(\bar{x}(\eta))|| \\ &\leq \theta ||\bar{x}(\lambda) - \bar{x}(\eta)|| + ||K_\lambda(\bar{x}(\eta)) - K_\eta(\bar{x}(\eta))|| \end{aligned}$$

We deduce from this that we have the following estimate:

$$\begin{aligned} ||\bar{x}(\lambda) - \bar{x}(\eta)|| &\leq \frac{1}{1-\theta} ||K_\lambda(\bar{x}(\eta)) - K_\eta(\bar{x}(\eta))|| \\ &= \frac{1}{1-\theta} ||(K_\lambda - K_\eta)\bar{x}(\eta)|| \end{aligned}$$

Now since $\lambda \to \eta$ implies $K_\lambda \to K_\eta$, the map $\lambda \to \bar{x}(\lambda)$ is indeed continuous.

(2) Let us prove now the last assertion of the theorem. For this purpose, pick a point $x_0 \in C$, and construct a sequence as in the statement, namely:

$$x_{n+1} = K_{\lambda_n}(x_n)$$

We want to show that the following happens, as claimed in the statement:

$$\lambda_n \to \lambda \implies x_n \to \bar{x}(\lambda)$$

For this purpose, consider the following two quantities:

$$\Delta_n = ||x_n - \bar{x}(\lambda)|| \quad , \quad \varepsilon_n = ||\bar{x}(\lambda_n) - \bar{x}(\lambda)||$$

We have then the following estimate, by using the triangle inequality, then our contraction assumption on the maps K_λ, and then the triangle inequality again:

$$\begin{aligned}
\Delta_{n+1} &= ||x_{n+1} - \bar{x}(\lambda)|| \\
&\leq ||x_{n+1} - \bar{x}(\lambda_n)|| + ||\bar{x}(\lambda_n) - \bar{x}(\lambda)|| \\
&= ||K_{\lambda_n}(x_n) - K_{\lambda_n}(\bar{x}(\lambda_n))|| + \varepsilon_n \\
&= ||K_{\lambda_n}(x_n - \bar{x}(\lambda_n))|| + \varepsilon_n \\
&\leq \theta||x_n - \bar{x}(\lambda_n)|| + \varepsilon_n \\
&\leq \theta(||x_n - \bar{x}(\lambda)|| + ||\bar{x}(\lambda) - \bar{x}(\lambda_n)||) + \varepsilon_n \\
&= \theta(\Delta_n + \varepsilon_n) + \varepsilon_n \\
&= \theta\Delta_n + (1+\theta)\varepsilon_n
\end{aligned}$$

We conclude from this, by iterating, that we have the following estimate:

$$\Delta_n \leq \theta^n\Delta_0 + (1+\theta)\sum_{j=1}^{n}\theta^{n-j}\varepsilon_{j-1}$$

Now since we have $\varepsilon_n \to 0$, this estimate gives then $\Delta_n \to 0$, as desired. □

In practice, we will need as well the following version of the above result:

THEOREM 7.21. *Given a Banach space X, and a closed subset $C \subset X$, assume that we have maps as follows, depending on parameters $\lambda \in \Lambda$, in some Banach space,*

$$K_\lambda : C \to C$$

which are continuous with respect to them, and are uniform contractions, satisfying

$$||K_{\lambda_n} \ldots K_{\lambda_1}(x) - K_{\lambda_n} \ldots K_{\lambda_1}(y)|| \leq \theta_n||x - y||$$

with $\sum_n \theta_n < \infty$. Then, these maps have unique fixed points $\bar{x}(\lambda)$, which are continuous with respect to λ. Moreover, if $\lambda_n \to \lambda$, then with $x_{n+1} = K_{\lambda_n}(x_n)$ we have $x_n \to \bar{x}(\lambda)$.

PROOF. This is something more technical, the idea being as follows:

(1) Consider the following maps, depending on parameters $\lambda' = (\lambda_1, \ldots, \lambda_n) \in \Lambda^n$:

$$K_{\lambda'} = K_{\lambda_n} \ldots K_{\lambda_1}$$

In order to prove the result, the idea will be to show that these maps are continuous with respect to their parameters $\lambda' \in \Lambda^n$, and then apply the previous theorem.

(2) So, let us prove the above-mentioned continuity property. We do this by recurrence on $n \in \mathbb{N}$, with the case $n = 1$ being clear from definitions. So, assume that we have continuity at $n-1$, and let us try to prove that we have continuity at n.

(3) For this purpose, consider two parameters at $n-1$, denoted as follows:

$$\lambda' = (\lambda_1, \ldots, \lambda_{n-1}) \quad , \quad \eta' = (\eta_1, \ldots, \eta_{n-1})$$

We have then the following estimate, obtained by using the triangle inequality, and then our assumptions from the statement, on our maps K_λ:

$$\begin{aligned} & ||K_{\lambda_n}K_{\lambda'}(x) - K_{\eta_n}K_{\eta'}(x)|| \\ \leq\ & ||K_{\lambda_n}K_{\lambda'}(x) - K_{\lambda_n}K_{\eta'}(x)|| + ||K_{\lambda_n}K_{\eta'}(x) - K_{\eta_n}K_{\eta'}(x)|| \\ \leq\ & \theta_1||K_{\lambda'}(x) - K_{\eta'}(x)|| + ||(K_{\lambda_n} - K_{\eta_n})K_{\eta'}(x)|| \end{aligned}$$

Now when assuming $(\lambda_1, \ldots, \lambda_n) \to (\eta_1, \ldots, \eta_n)$, both terms at the end go to 0, and so the quantity itself, that we estimated above, goes to 0 too, as desired.

(4) Summarizing, we have proved the continuity claim in (1). But with this in hand, the previous theorem applies, with the remark that our assumption $\sum_n \theta_n < \infty$ forces indeed $\theta_n < 1$, for $n \in \mathbb{N}$ big enough, and this gives the result. □

Very nice all this, so we have now functional analysis tools for dealing with the Volterra integral equations from Theorem 7.19. It is actually convenient to go beyond the framework of Theorem 7.19, with more general results. Let us formulate indeed:

DEFINITION 7.22. *A Volterra integral operator is an operator of type*

$$K_\lambda(x)(t) = k(t, \lambda) + \int_0^t K(s, x(s), \lambda)ds$$

depending on functions as follows,

$$k \in C(I \times \Lambda, U) \quad , \quad K \in C(I \times U \times \Lambda, \mathbb{R}^N)$$

with $I = [-T, T]$ being an interval, $U \subset \mathbb{R}^N$ an open set, and $\Lambda \subset \mathbb{R}^N$ a compact set.

Observe that the previous Volterra integral equations, from Theorem 7.19, make appear indeed such operators. There are some other interesting examples as well.

The point now is that we can apply our fixed point theorems, and we obtain:

THEOREM 7.23. *Assume that there exists $L > 0$ such that*

$$||K(t, x, \lambda) - K(t, y, \lambda)|| \leq L||x - y||$$

holds, for any t, any λ, and any $x, y \in U$. Then the equation

$$K_\lambda(x) = x$$

has a unique solution, which is continuous with respect to λ.

PROOF. This follows indeed by applying our general fixed point theorems, with the existence of $L > 0$ as above guaranteeing that the contraction conditions are satisfied. □

7d. Linearization

Getting now to more advanced topics, we would like to know more about the solutions. For this purpose, we will need some classical analysis results. Let us start with:

THEOREM 7.24. *Assume that $f_n \to f$ pointwise, with $f_n, f : U \subset \mathbb{R}^N \to \mathbb{R}^N$ being integrable, and assume in addition that we have*

$$|f_n(x)| \leq g(x)$$

for some integrable function $g : U \subset \mathbb{R}^N \to \mathbb{R}^N$. Then, the following happens:

$$\lim_{n\to\infty} \int f_n(x) = \int f(x)$$

Moreover, this latter conclusion can fail, without our assumption using g.

PROOF. This is a very standard analysis theorem, called dominated convergence theorem, and for a complete proof of this, see for instance Rudin [79]. □

Next, we have the following result, which is something more specialized:

THEOREM 7.25. *Assume that $f_n \to f$ and $f_n' \to g$ pointwise, and that we have*

$$|f_n'| \leq \gamma$$

for some integrable function γ. Then f is differentiable, with derivative given by:

$$f'(x) = g(x)$$

Moreover, this latter conclusion can fail, without our assumption using γ.

PROOF. It is enough to deal with the case of the one-variable functions, $f : \mathbb{R} \to \mathbb{R}$. But here, we can use the following formula:

$$f_n(x) = f_n(x_0) + \int_{x_0}^{x} f_n'(t)dt$$

Indeed, with $n \to \infty$ we obtain from this, by using Theorem 7.24 on the right:

$$f(x) = f(x_0) + \int_{x_0}^{x} g(t)dt$$

On the other hand, we know that we have the following formula:

$$f(x) = f(x_0) + \int_{x_0}^{x} f(t)dt$$

We conclude that the following equality must hold, for any x:

$$\int_{x_0}^{x} f(t)dt = \int_{x_0}^{x} g(t)dt$$

But now, by differentiating we obtain $f = g$, as desired. As for the counterexample at the end, we will leave this as an instructive exercise. □

Hang on, we are not done yet with classical analysis. Here is in fact the result that we will need, in connection with the questions that we are interested in:

THEOREM 7.26. *Assume that $f(x,\lambda)$ is integrable with respact to x, for any λ, and C^1 with respect to λ, for any x. Assume in addition that*

$$\left|\frac{df}{d\lambda}(x,\lambda)\right| \leq g(x)$$

for some integrable function g. Then the function given by

$$F(\lambda) = \int f(x,\lambda)dx$$

is C^1, and its derivative is given by the following formula:

$$\frac{dF}{d\lambda}(\lambda) = \int \frac{df}{d\lambda}(x,\lambda)dx$$

Moreover, this latter conclusion can fail, without our assumption using g.

PROOF. As before with Theorem 7.25, it is enough to do this in 1 dimension. In order to simplify the notations, let us denote by f' the derivative of f with respect to λ:

$$f' = \frac{df}{d\lambda}$$

We have then the following formula, coming from standard calculus:

$$f(x,\lambda+\varepsilon) - f(x,\lambda) = \varepsilon \int_0^1 f'(x,\lambda+\varepsilon t)dt$$

Thus, in terms of the function F from the statement, we have:

$$\frac{F(\lambda+\varepsilon) - F(\lambda)}{\varepsilon} = \int \int_0^1 f'(x,\lambda+\varepsilon t)dtdx$$

According to our assumption $|f'| \leq g$, the following estimate holds:

$$|f'(x,\lambda+\varepsilon t)| \leq g(x)$$

Thus we can apply Theorem 7.24, and we obtain in this way:

$$\begin{aligned}\lim_{\varepsilon\to 0}\int_0^1 f'(x,\lambda+\varepsilon t)dt &= \int_0^1 \lim_{\varepsilon\to 0} f'(x,\lambda+\varepsilon t)dt\\ &= f'(x,\lambda)\end{aligned}$$

On the other hand, we have as well the following estimate:

$$\left|\int_0^1 f'(x,\lambda+\varepsilon t)dt\right| \leq g(x)$$

Thus, we can apply Theorem 7.24 again, and we obtain in this way:

$$\begin{aligned}\lim_{\varepsilon\to 0}\frac{F(\lambda+\varepsilon)-F(\lambda)}{\varepsilon} &= \lim_{\varepsilon\to 0}\int\int_0^1 f'(x,\lambda+\varepsilon t)dtdx\\ &= \int\lim_{\varepsilon\to 0}\int_0^1 f'(x,\lambda+\varepsilon t)dtdx\\ &= \int f'(x,\lambda)dx\end{aligned}$$

But this is exactly the formula that we wanted to establish, so done. □

Good news, we can now go back to our integral equations. Let us recall indeed from Definition 7.22 that a Volterra integral operator was an operator of the following type, depending on functions $k\in C(I\times\Lambda,U)$ and $K\in C(I\times U\times\Lambda,\mathbb{R}^N)$:

$$K_\lambda(x)(t)=k(t,\lambda)+\int_0^t K(s,x(s),\lambda)ds$$

As explained in Theorem 7.23, we are interested in the corresponding fixed points:

$$K_\lambda(x)=x$$

And the point now is that we have the following result regarding these fixed points, coming as a complement to what we already know from Theorem 7.23:

THEOREM 7.27. *In the context of a Volterra integral operator*

$$K_\lambda(x)(t)=k(t,\lambda)+\int_0^t K(s,x(s),\lambda)ds$$

if k,K are assumed both C^r with respect to λ,x, then the solution of

$$K_\lambda(x)=x$$

is C^r too, with respect to its variable.

PROOF. This comes indeed by applying Theorem 7.26, as follows:

(1) As a first observation, by suitably modifying the function $K(t,x,\lambda)$, we can assume that we have $k(t,\lambda)=0$. That is, we can assume that our operator is as follows:

$$K_\lambda(x)(t)=\int_0^t K(s,x(s),\lambda)ds$$

(2) The idea will be that of proceeding by recurrence on r. Let us first prove that the solution is C^0. For this purpose, we use the triangle inequality, which gives:

$$|\bar{x}(t,\lambda)-\bar{x}(s,\eta)|\leq|\bar{x}(t,\lambda)-\bar{x}(s,\lambda)|+|\bar{x}(s,\lambda)-\bar{x}(s,\eta)|$$

Now observe that with $(t,\lambda)\to(s,\eta)$, we have the following estimate:

$$|\bar{x}(t,\lambda)-\bar{x}(s,\lambda)|\leq\left|\int_0^t K(r,\bar{x}(r,\lambda),\lambda)dr\right|\to 0$$

As for the other term appearing above, again with $(t,\lambda)\to(s,\eta)$, from the fixed point theorem that we used in order to construct the solution, we know that we have:

$$|\bar{x}(s,\lambda)-\bar{x}(s,\eta)|\to 0$$

As a conclusion, with $(t,\lambda)\to(s,\eta)$, we have, as desired:

$$|\bar{x}(t,\lambda)-\bar{x}(s,\eta)|\to 0$$

(2) In order to discuss now the case $r\geq 1$, the idea will be that of constructing integral equations for the partial derivatives of the solutions, that we can solve afterwards by using the calculus rules coming from Theorem 7.24, and its various versions above.

(3) To be more precise, let us first discuss the usual differentiability, $r=1$. By following the above idea, let us consider the following function:

$$\bar{y}(t,\lambda)=\frac{d}{d\lambda}\,\bar{x}(t,\lambda)$$

Consider as well the following modification of our original integral operator:

$$\bar{K}_\lambda(x,y)(t)=\int_0^t\frac{d}{d\lambda}\,K_\lambda(s,x(s),\lambda)y(s)ds$$

Our claim is then that $\bar{y}(t,\lambda)$ is the solution of the following equation:

$$\bar{K}_\lambda(\bar{x}(\lambda),y)=y$$

(4) Indeed, by our result from (2) above we know that this latter equation has a certain continuous solution, say $\tilde{y}$. In order to prove now that we have $\bar{y}=\tilde{y}$, let us set:

$$(x_0(t),y_0(t))=(0,0)\quad,\quad(x_{n+1},y_{n+1})=(K_\lambda(x_n),\bar{K}_\lambda(x_n,y_n))$$

By using the fixed point theorem, in its second, technical version, we obtain:

$$(x_n,y_n)\to(\bar{x},\tilde{y})$$

Now since (x_n,y_n) is uniformly bounded with respect to λ, we can apply Theorem 7.26, and we obtain $\bar{y}=\tilde{y}$, as desired, proving our claim in (3).

(5) So, this was for the idea of the proof at $r=1$, and the proof in general, at $r\in\mathbb{N}$, is similar, by recurrence. We will leave the details here as an instructive exercise. □

Along the same lines, we have as well the following useful result:

THEOREM 7.28. *In the context of a Volterra integral operator*

$$K_\lambda(x)(t) = k(t,\lambda) + \int_0^t K(s, x(s), \lambda)ds$$

the solution of the equation $K_\lambda(x) = x$ exists and is unique on

$$C([-T_0, T_0] \times \Lambda, U)$$

and satisfies the following explicit estimate,

$$|\bar{x}(t,\lambda) - k(t,\lambda)| \leq e^{LT_0} \sup_{\lambda\in\Lambda} \int_{-T_0}^{T_0} |K(s, k(s,\lambda), \lambda)|ds$$

with the number $T_0 > 0$ depending on K and k.

PROOF. This is something very standard, exactly as in the case without parameters, which was worked out in the above, and we will leave clarifying the details, including working out the formula of $T_0 > 0$, as function of K and k, as an instructive exercise. □

As a conclusion to this, we have a quite good understanding of the Volterra integral equations, that can be applied for instance to questions related to the linear equations.

Getting now to the non-linear case, the idea here will be that of using the linearization strategy explained earlier in this chapter, in the hyperbolic case.

So, consider such a non-linear equation $\dot{x} = f(x)$, and denote as usual by $\Phi(t,x)$ its flow, describing the solution in time t, with initial data $x(0) = x$. We have:

DEFINITION 7.29. *We associate to the equation $\dot{x} = f(x)$ the following sets,*

$$W^\pm(x_0) = \left\{ x \,\middle|\, \lim_{t\to\pm\infty} \Phi(t,x) = x_0 \right\}$$

gathering the initial data x such that the solution converges to x_0, with $t \to \pm\infty$.

Observe that both the above sets $W^\pm(x_0)$ are stable under the flow. In order now to compute these sets, we use our linearization idea. So, let us introduce as well:

DEFINITION 7.30. *We associate to the equation $\dot{x} = f(x)$ the sets*

$$M^{\pm,\alpha} = \left\{ x \,\middle|\, \gamma_\pm(x) \subset U(x_0),\ \sup_{\pm t\geq 0} e^{\pm\alpha t}|\Phi(t,x) - x_0| < \infty \right\}$$

and then we consider the intersection of these sets, over eigenvalues,

$$M^\pm(x_0) = \bigcup_{\alpha>0} M^{\pm,\alpha}$$

which in the linear case, $\dot{x} = Ax$, are the spaces $E^\pm$ that we knew from before.

To be more precise here, in the linear case, $\dot{x} = Ax$, the spaces $M^{\pm,\alpha}$ constructed above correspond to the spaces $E^{\pm,\alpha}$ spanned by the eigenvectors of A corresponding to the eingenvalues satisfying $Re(\lambda) \geq \alpha$ and $Re(\lambda) \leq -\alpha$, and so by intersecting, we obtain indeed the spaces $E^{\pm}$ that we knew from before, as claimed in the above.

Observe also that, in general, the spaces constructed above are invariant by the flow.

We can now formulate our main linearization result, as follows:

THEOREM 7.31. *For a hyperbolic point x_0, the following happen:*

(1) $M^{\pm}(x_0)$ *is a* C^1 *manifold.*
(2) $M^{\pm}(x_0)$ *is tangent to* $E^{\pm}$ *at* 0.
(3) $M^{\pm}(x_0) = W^{\pm}(x_0)$.

PROOF. The idea here will be that of using the standard direct sum decomposition $\mathbb{R}^N = E^+ \oplus E^-$, in order to decompose everything, and then using the theory of integral equations developed in the above, in order to prove the various assertions.

(1) Let us begin with some notations. We let $P^{\pm}$ be the orthogonal projection onto the linear space $E^{\pm}$, and we consider the following quantities:

$$x_{\pm} = P^{\pm}x(0) \quad , \quad g_{\pm}(x) = P^{\pm}g(x) \quad , \quad x_{\pm}(t) = P^{\pm}x(t)$$

(2) Our first claim is that, assuming that $x(t)$ is bounded with $t > 0$, any solution solves the following equation, where $P(t) = P^+$ for $t > 0$, and $P(t) = -P^-$ for $t \leq 0$:

$$x(t) = K(x)(t) \quad , \quad K(x)(t) = e^{tA}x_+ + \int_0^{\infty} e^{(t-r)A}P(t-r)g(x(r))dr$$

But this follows indeed from a routine computation, based on Theorem 7.19.

(2) Our second claim is that, assuming that $f \in C^k$, and that $\alpha > 0$ is such that $A + \alpha 1_N$ is hyperbolic, we can find a neighborhood $U(x_0) = x_0 + U$ and a function $h^{t,\alpha} \in C^k(E^{t,\alpha} \cap U, E^{-,\alpha})$ such that both $h^{t,\alpha}$ and its derivative vanish at 0, and that:

$$M^{+,\alpha}(x_0) \cap U(x_0) = \left\{x_0 + a + h^{t,\alpha}(a)\middle| a \in E^{t,\alpha} \cap U\right\}$$

Moreover, we also claim that in this situation, the following happen:

$$\alpha_1 \leq \alpha_2 \implies M^{+,\alpha_2}(x_0) \leq M^{+,\alpha_1}(x_0)$$

$$E^{+,\alpha_2} = E^{+,\alpha_1} \implies M^{+,\alpha_2}(x_0) = M^{+,\alpha_1}(x_0)$$

But all this can be proved, by carefully applying our contraction principles above.

(3) Now the point is that the first claim in (2) proves that $M^{\pm}(x_0)$ is indeed tangent to $E^{\pm}$ at 0, and so, with a bit more work, we are led to the proof of the theorem. We will leave the details here, including learning about manifolds, as an instructive exercise. □

7e. Exercises

Exercises:

Exercise 7.32.

Exercise 7.33.

Exercise 7.34.

Exercise 7.35.

Exercise 7.36.

Exercise 7.37.

Exercise 7.38.

Exercise 7.39.

Bonus exercise.

CHAPTER 8

Advanced aspects

Exercises:

Exercise 8.1.

Exercise 8.2.

Exercise 8.3.

Exercise 8.4.

Exercise 8.5.

Exercise 8.6.

Exercise 8.7.

Exercise 8.8.

Bonus exercise.

Part III

Geometric aspects

CHAPTER 9

9a.

9b.

9c.

9d.

9e. Exercises

Exercises:

Exercise 9.1.

Exercise 9.2.

Exercise 9.3.

Exercise 9.4.

Exercise 9.5.

Exercise 9.6.

Exercise 9.7.

Exercise 9.8.

Bonus exercise.

CHAPTER 10

10a.

10b.

10c.

10d.

10e. Exercises

Exercises:

Exercise 10.1.

Exercise 10.2.

Exercise 10.3.

Exercise 10.4.

Exercise 10.5.

Exercise 10.6.

Exercise 10.7.

Exercise 10.8.

Bonus exercise.

CHAPTER 11

11a.

11b.

11c.

11d.

11e. Exercises

Exercises:

Exercise 11.1.

Exercise 11.2.

Exercise 11.3.

Exercise 11.4.

Exercise 11.5.

Exercise 11.6.

Exercise 11.7.

Exercise 11.8.

Bonus exercise.

CHAPTER 12

12a.

12b.

12c.

12d.

12e. Exercises

Exercises:

Exercise 12.1.

Exercise 12.2.

Exercise 12.3.

Exercise 12.4.

Exercise 12.5.

Exercise 12.6.

Exercise 12.7.

Exercise 12.8.

Bonus exercise.

Part IV

Advanced mechanics

CHAPTER 13

13a.

13b.

13c.

13d.

13e. Exercises

Exercises:

Exercise 13.1.

Exercise 13.2.

Exercise 13.3.

Exercise 13.4.

Exercise 13.5.

Exercise 13.6.

Exercise 13.7.

Exercise 13.8.

Bonus exercise.

CHAPTER 14

14a.

14b.

14c.

14d.

14e. Exercises

Exercises:

Exercise 14.1.

Exercise 14.2.

Exercise 14.3.

Exercise 14.4.

Exercise 14.5.

Exercise 14.6.

Exercise 14.7.

Exercise 14.8.

Bonus exercise.

CHAPTER 15

15a.

15b.

15c.

15d.

15e. Exercises

Exercises:

Exercise 15.1.

Exercise 15.2.

Exercise 15.3.

Exercise 15.4.

Exercise 15.5.

Exercise 15.6.

Exercise 15.7.

Exercise 15.8.

Bonus exercise.

CHAPTER 16

16a.

16b.

16c.

16d.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter.

Bibliography

[1] V.I. Arnold, Ordinary differential equations, Springer (1973).

[2] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] V.I. Arnold, Catastrophe theory, Springer (1974).

[5] V.I. Arnold and B.A. Khesin, Topological methods in hydrodynamics, Springer (1998).

[6] M.F. Atiyah, K-theory, CRC Press (1964).

[7] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[8] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).

[9] T. Banica, Calculus and applications (2024).

[10] T. Banica, Linear algebra and group theory (2024).

[11] T. Banica, Introduction to modern physics (2025).

[12] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[13] N. Berline, E. Getzler and M. Vergne, Heat kernels and Dirac operators, Springer (2004).

[14] B. Blackadar, K-theory for operator algebras, Cambridge Univ. Press (1986).

[15] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[16] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[17] A.R. Choudhuri, Astrophysics for physicists, Cambridge Univ. Press (2012).

[18] A. Connes, Noncommutative geometry, Academic Press (1994).

[19] A. Connes and M. Marcolli, Noncommutative geometry, quantum fields and motives, AMS (2008).

[20] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[21] P.A. Davidson, Introduction to magnetohydrodynamics, Cambridge Univ. Press (2001).

[22] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[23] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).

[24] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).

[25] S. Dodelson, Modern cosmology, Academic Press (2003).

[26] S.K. Donaldson, Riemann surfaces, Oxford Univ. Press (2004).

[27] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[28] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[29] L.C. Evans, Partial differential equations, AMS (1998).

[30] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[31] E. Fermi, Thermodynamics, Dover (1937).

[32] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics, Caltech (1963).

[33] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[34] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[35] A.P. French, Special relativity, Taylor and Francis (1968).

[36] W. Fulton, Algebraic topology, Springer (1995).

[37] W. Fulton and J. Harris, Representation theory, Springer (1991).

[38] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[39] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[40] M.B. Green, J.H. Schwarz and E. Witten, Superstring theory, Cambridge Univ. Press (2012).

[41] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[42] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[43] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[44] P. Griffiths and J. Harris, Principles of algebraic geometry, Wiley (1994).

[45] A. Grothendieck and J. Dieudonné, Éléments de géométrie algébrique, IHES (1967).

[46] A. Grothendieck et al., Séminaire de géométrie algébrique, IHES (1972).

[47] G.H. Hardy and E.M. Wright, An introduction to the theory of numbers, Oxford Univ. Press (1938).

[48] J. Harris, Algebraic geometry, Springer (1992).

[49] R. Hartshorne, Algebraic geometry, Springer (1977).

[50] A. Hatcher, Algebraic topology, Cambridge Univ. Press (2002).

[51] H. Hofer and E. Zehnder, Symplectic invariants and Hamiltonian dynamics, Birkhäuser (1994).

[52] L. Hörmander, The analysis of linear partial differential operators, Springer (1983).

[53] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[54] K. Huang, Introduction to statistical physics, CRC Press (2001).

[55] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).

[56] J.E. Humphreys, Linear algebraic groups, Springer (1975).

[57] K. Ireland and M. Rosen, A classical introduction to modern number theory, Springer (1982).

[58] N. Jacobson, Basic algebra, Dover (1974).

[59] V.F.R. Jones, Index for subfactors, *Invent. Math.* **72** (1983), 1–25.

[60] V.F.R. Jones, A polynomial invariant for knots via von Neumann algebras, *Bull. Amer. Math. Soc.* **12** (1985), 103–111.

[61] V.F.R. Jones, Hecke algebra representations of braid groups and link polynomials, *Ann. of Math.* **126** (1987), 335–388.

[62] V.F.R. Jones, On knot invariants related to some statistical mechanical models, *Pacific J. Math.* **137** (1989), 311–334.

[63] V.F.R. Jones, Planar algebras I (1999).

[64] M. Karoubi, K-theory: an introduction, Springer (1978).

[65] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[66] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[67] L.D. Landau and E.M. Lifshitz, Course of theoretical physics, Pergamon Press (1960).

[68] S. Lang, Algebra, Addison-Wesley (1993).

[69] S. Lang, Abelian varieties, Dover (1959).

[70] P. Lax, Linear algebra and its applications, Wiley (2007).

[71] P. Lax, Functional analysis, Wiley (2002).

[72] J.M. Lee, Introduction to topological manifolds, Springer (2011).

[73] J.M. Lee, Introduction to smooth manifolds, Springer (2012).

[74] J.M. Lee, Introduction to Riemannian manifolds, Springer (2019).

[75] D. McDuff and D. Salamon, Introduction to symplectic topology, Oxford Univ. Press (2017).

[76] P. Petersen, Linear algebra, Springer (2012).

[77] P. Petersen, Riemannian geometry, Springer (2006).

[78] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[79] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[80] W. Rudin, Fourier analysis on groups, Dover (1974).

[81] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[82] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[83] W. Schlag, A course in complex analysis and Riemann surfaces, AMS (2014).

[84] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[85] J.P. Serre, A course in arithmetic, Springer (1973).

[86] J.P. Serre, Linear representations of finite groups, Springer (1977).

[87] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[88] J.H. Silverman, The arithmetic of elliptic curves, Springer (1986).

[89] J.H. Silverman and J.T. Tate, Rational points on elliptic curves, Springer (2015).

[90] B. Singh, Basic commutative algebra, World Scientific (2011).

[91] C.H. Taubes, Differential geometry, Oxford Univ. Press (2011).

[92] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[93] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[94] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[95] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[96] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[97] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[98] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[99] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

[100] B. Zwiebach, A first course in string theory, Cambridge Univ. Press (2004).

Index