Métodos Matemáticos de la Física

Oscar Reula

Acknowledgements

First of all, I want to thank Gloria Puente for supporting me throughout the time I wrote these notes, which took quite a bit of our shared hours, generally long nights. Secondly, to Bernardo González Kriegel who converted the first versions of these notes to LaTeX, doing a great job with much enthusiasm. Also to Bob Geroch, with whom I discussed several topics of the notes and from whom I was also inspired through his writings and books, not only in content but also in style. Finally, to several generations of students who stoically assimilated a large amount of the material from these courses in a very short time.

Contents

List of Figures

PREFACE

These notes, now turned into a book, originated as an attempt to condense in one place a large set of ideas, concepts, and mathematical tools that I consider basic for the understanding and daily work of a physicist today.

It usually happens that if a problem is formulated from a physical need, such as the description of some natural phenomenon, then it is well formulated, in the sense that a reasonable solution to it exists. This rule has generally been very fruitful and has particularly served as a guide to many mathematicians to make their way in unknown areas. But it has also served, particularly to many physicists, to work without worrying too much about formal aspects, whether analytical, algebraic, or geometric, and thus be able to concentrate on physical and/or computational aspects. Although this allows for the rapid development of some research, in the long run, it leads to stagnation because by proceeding in this way, one avoids facing problems that are very rich in terms of conceptualizing the phenomenon to be described. It is important to verify that the formulated problem has a mathematically and physically correct solution.

An example of this was the development, in the middle of the last century, of the modern theory of partial differential equations. Many of these equations arose because they describe physical phenomena: heat transmission, electromagnetic wave propagation, quantum waves, gravitation, etc. One of the first mathematical responses to the development of these areas was the Cauchy-Kowalevski theorem, which tells us that given a partial differential equation (under quite general circumstances), if an analytic function is given as data on a hypersurface (with certain characteristics), then there is a unique solution in a sufficiently small neighborhood of that hypersurface. It took a long time to realize that this theorem was not really relevant from the point of view of physical applications: there were equations admitted by the theorem such that if the data was not analytic, there was no solution! And in many cases, if they existed, they did not depend continuously on the given data, a small variation of the data produced a completely different solution. It was only in the middle of the last century that substantial progress was made on the problem, finding that there were different types of equations, hyperbolic, elliptic, parabolic, etc., that behaved differently and this reflected the different physical processes they simulated. Due to its relative novelty, this very important set of concepts is not part of the set of tools that many active physicists have, nor are they found in the textbooks usually used in undergraduate courses.

Like the previous one, there are many examples, particularly the theory of or-

dinary differential equations and geometry, without which it is impossible to understand many of the modern theories, such as relativity, elementary particle theories, and many phenomena of solid-state physics. As our understanding of the basic phenomena of nature advances, we realize that the most important tool for their description is geometry. This tool, among other things, allows us to handle a wide range of processes and theories with little in common among each other with a very reduced set of concepts, thus achieving a synthesis. These syntheses are what allow us to acquire new knowledge, since by adopting them we leave space in our minds to learn new concepts, which are in turn ordered more efficiently within our mental construction of the area.

These notes were originally intended for a four-month course. But in reality, they were more suited for an annual course or two semesters. Then, as more topics were incorporated into them, it became increasingly clear that they should be given in two semesters or one annual course. Basically, one course should contain the first chapters that include notions of topology, vector spaces, linear algebra, ending with the theory of ordinary differential equations. The task is considerably simplified if the students have previously had a good course on linear algebra. The correlation with physics subjects should be such that the course is prior to or concurrent with an advanced mechanics course. Emphasizing in it the fact that ultimately one is solving ordinary differential equations with a certain special structure. Using the concepts of linear algebra to find eigenmodes and the stability of equilibrium points. And finally using geometry to describe, albeit briefly, the underlying symplectic structure.

The second course consists of developing the tools to be able to discuss aspects of the theory of partial differential equations. It should be given before or concurrently with an advanced electromagnetism course, where emphasis should be placed on the type of equations that are solved (elliptic, hyperbolic), and the meaning of their initial or boundary conditions, as appropriate. Also using coherently the concept of *distribution*, which is far from being an abstract mathematical concept but is actually a concept that naturally appears in physics.

None of the content of these notes is original material, but some ways of presenting it are: for example, some proofs that are simpler than usual, or the way some content is integrated with the previous ones. Much of the material should be thought of as a first reading or an introduction to the topic. The interested reader should consult the cited books, from which I have drawn upon a lot, being these excellent and difficult to surpass.

Chapter 1

Basic Concepts of Topology

1.1 Introduction

The notion of a set, while telling us that certain objects —the elements that comprise it— have something in common with each other, does not give us any idea of the *closeness* between these elements. On the other hand, if we consider for example the real numbers, this notion is present. We know, for example, that the number 2 is much closer to 1 than 423 is. The concept of a topology in a set that we will define below tries to capture precisely this notion of closeness which, as we will see, admits many gradations.

Definition: A **topological space** consists of a pair $(X, \mathcal{T})$, where X is a set and $\mathcal{T}$ is a collection of subsets of X satisfying the following conditions:

1. The subsets $\emptyset$ and X of X are in $\mathcal{T}$.
2. If O_λ, $\lambda \in I$, is a one-parameter family of subsets of X in $\mathcal{T}$, then $\bigcup_I O_\lambda$ is also in $\mathcal{T}$.
3. If O and O' are in $\mathcal{T}$, then $O \cap O'$ is also in $\mathcal{T}$.

For simplicity, we will often denote the topological space $(X, \mathcal{T})$ simply by X when the topology is understood from context, as is common practice.

The elements of $\mathcal{T}$, subsets of X, are called the **open subsets** of X. The set $\mathcal{T}$ itself is called a **topology** on X. Condition 2) tells us that infinite unions of elements of $\mathcal{T}$ are also in $\mathcal{T}$, while condition 3) tells us that in general only finite intersections remain in $\mathcal{T}$. The following examples illustrate why this asymmetry exists, they also illustrate how giving a topology is essentially giving a notion of closeness between the points of the set in question.

Example: a) Let $\mathcal{T} = \{\emptyset, X\}$, that is, the only open subsets of X are the empty subset and the subset X. It is clear that this collection of subsets is a topology, as it satisfies the three required conditions. This topology is called the **indiscrete topology** on X. We can say that, with this topology, the points of X are arbitrarily close to each other, since if an open set contains one of them it contains all of them.

Example: b) Let $\mathscr{T} = \mathscr{P}(X)$, the collection of all subsets of X. Clearly, this collection also satisfies the conditions mentioned above and therefore it is also a topology on X, the so-called **discrete topology on** X. We can say that in this one all the points are arbitrarily separated from each other since, for example, given any point of X there is an open set that separates it from all the others, which consists of only the point in question.

Example: c) Let X be the set of real numbers, $\mathbb{R}$, and let
$\mathscr{T} = \{O \subseteq \mathbb{R} |$ if $r \in O, \exists \varepsilon > 0$ suchthatif $|r - r'| < \varepsilon,\ r' \in O\}$, that is, the collection of open sets in the usual sense of the real numbers. Let's see that this collection satisfies the conditions to be a topology. Clearly, $\emptyset \in \mathscr{T}$ (since it has no r), as well as $\mathbb{R} \in \mathscr{T}$ (since it contains all r'), and thus condition 1) is satisfied. Let us examine the second condition: let $r \in \bigcup_I O_\lambda$ then $r \in O_\lambda$ for some λ and therefore there will exist $\varepsilon > 0$ such that all r' with $|r - r'| < \varepsilon$ are also in O_λ, and therefore in $\bigcup_I O_\lambda$. Finally, the third condition: let $r \in O \cap O'$ then r is in O and therefore there will exist $\varepsilon > 0$ such that all r' with $|r - r'| < \varepsilon$ will be in O; as r is also in O', there will exist $\varepsilon' > 0$ such that all r' with $|r - r'| < \varepsilon'$ will be in O'. Let $\varepsilon'' = \min\{\varepsilon, \varepsilon'\}$, then all r' with $|r - r'| < \varepsilon''$ will be in O and in O' and therefore in $O \cap O'$, so we conclude that this last set is also in $\mathscr{T}$. $\mathbb{R}$ with this topology is called the **real line**.

Exercise: In the real line, as defined in the previous example, find an infinite intersection of open sets that is not open.

Example: d) Let $X = \mathbb{R} \times \mathbb{R} \equiv \mathbb{R}^2$, that is, the Cartesian product of $\mathbb{R}$with itself —the set of all pairs (x, y), with $x, y \in \mathbb{R}$— and let
$\mathscr{T} = \{O \in \mathbb{R}^2 |$ if$(x, y) \in O, \exists \varepsilon >$ osuchthatif$|x - x'| + |y - y'| < \varepsilon,\ (x', y') \in O\}$.
Following the previous example, it can be seen that this is also a topological space and that this is the topology we usually use in $\mathbb{R}^2$

Definition: A **metric space** is a pair (X, d) consisting of a set X and a map $d : X \times X \to \mathbb{R}$, usually called **distance**, satisfying the following conditions for all $x, x', x'' \in X$:

1. Non-negativity: $d(x, x') \geq 0$, with equality only when $x = x'$.
2. Symmetry: $d(x, x') = d(x', x)$.
3. Triangle inequality: $d(x, x') + d(x', x'') \geq d(x, x'')$.

Exercise: Prove that any metric space has a topology *induced* by its metric in a similar way to $\mathbb{R}$ in the previous example.

Exercise: Prove that, for any set, $d(x, y) = 1$ if $x \neq y$ and $d(x, x) = 0$ defines a distance. What topology does this distance induce on the set?

Clearly, a distance gives us a notion of closeness between points, in the precise form of a numerical value. A topology, by not generally giving us any number, gives us a much vaguer notion of closeness, but still generally interesting.

1.1.1 Terminology

We now give a summary of the usual terminology in this area, which is a direct generalization of the commonly used one.

Definition: We will call the **complement**, O^c, of the subset O of X the subset of all elements of X that are not in O.

Definition: We will say that a subset O of X is **closed** if its complement O^c is open.

Definition: A subset N of X is called a **neighborhood of** $x \in X$ if there is an open set O_x, with $x \in O_x$, contained in N.

Definition: We will call the **interior** of $A \in X$ the subset $Int(A)$ of X formed by the union of all open sets contained in A.

Definition: We will call the **closure** of $A \in X$ the subset $Cl(A)$ of X formed by the intersection of all closed sets containing A.

Definition: We will call the **boundary** of $A \in X$ the subset ∂A of X formed by $Cl(A) - Int(A) \equiv Int(A)^c \cap Cl(A)$.

Exercise: Let (X,d) be a metric vector space, prove that: a) $C_x^1 = \{x' | d(x,x') \leq 1\}$ is closed and is a neighborhood of x. b) $N_x^\varepsilon = \{x' | d(x,x') < \varepsilon\}$, $\varepsilon > 0$ is also a neighborhood of x. c) $Int(N_x^\varepsilon) = N_x^\varepsilon$ d) $Cl(N_x^\varepsilon) = \{x' | d(x,x') \leq \varepsilon\}$ e) $\partial N_x^\varepsilon = \{x' | d(x,x') = \varepsilon\}$.

Exercise: Let $(X,\mathcal{T})$ be a topological space and A a subset of X. Prove that: a) A is open if and only if each $x \in A$ has a neighborhood contained in A. b) A is closed if and only if each x in A^c (that is, not belonging to A) has a neighborhood that does not intersect A.

Exercise: Let $(X,\mathcal{T})$ be a topological space, let $A \in X$ and $x \in X$. Prove that: a) $x \in Int(A)$ if and only if x has a neighborhood contained in A. b) $x \in Cl(A)$ if and only if every neighborhood of x intersects A. c) $x \in \partial A$ if and only if every neighborhood of x contains points in A and points in A^c.

1.2 Derived Concepts

In the previous sections, we have seen that the concept of a topology leads us to a generalization of a series of ideas and derived concepts that we knew how to handle in

$\mathbb{R}^n$, revealing that they did not depend on the usual distance used in these spaces (the so-called Euclidean distance). It is then worth asking if there are still other possible generalizations. In this and the next subsection, we will study two more of them. These in turn open up a vast area of mathematics, which we will not cover in this course but is very important in what concerns modern physics.

The first of these notions is *continuity*.

1.2.1 Continuous Maps

Definition: Let $\varphi : X \to Y$ be a map between two topological spaces (see the box below). We will say that the map φ is **continuous** if given any open set O of Y, $\varphi^{-1}(O)$ is an open set of X.

> Definition: A **map** $\phi : X \to Y$ between a set X and another Y is an assignment to *each* element of X of an element of Y.
> This generalizes the usual concept of a function. Note that the map is defined for every element of X while, in general, its **image**, that is, the set $\phi(X) \equiv \{y \in Y \mid \exists x \in X \text{ such that } \phi(x) = y\}$, is not all of Y. In the case that it is, i.e. $\phi(X) = Y$, we will say that the map is **surjective**. On the other hand, if it is fulfilled that $\phi(x) = \phi(\tilde{x}) \Longrightarrow x = \tilde{x}$, we will say that the map is **injective**. In such a case, there exists the inverse map to ϕ between the set $\phi(X) \subset Y$ and X. If the map is also surjective then its inverse is defined on all of Y and in this case, it is denoted by $\phi^{-1} : Y \to X$. It is also useful to consider the sets $\phi^{-1}(O) = \{x \in X \mid \phi(x) \in O\}$, sometimes called the **pre-image** of the set O under ϕ.

Clearly, the previous definition only uses topological concepts. Does it have anything to do with the usual *epsilon-delta* definition used in $\mathbb{R}^n$? The answer is affirmative, as we will see below in our first theorem. But first, let's see some examples.

Example: a) Let X and Y be any topological spaces and let the topology on X be the discrete one. Then any map between X and Y is continuous. Indeed, for any open set O in Y, $\varphi^{-1}(O)$ is some subset in X, but in the discrete topology, every subset of X is an open set.

Example: b) Let X and Y be any topological spaces and let the topology on Y be the indiscrete one. Then any map between X and Y is also continuous. Indeed, the only open sets in Y are $\emptyset$ and Y, but $\varphi^{-1}(\emptyset) = \emptyset$, while $\varphi^{-1}(Y) = X$, but whatever the topology of X, $\emptyset$ and X are open sets.

From the previous examples, it might seem that our definition of continuity is not very interesting. But that is because we have taken cases with the *extreme* topologies. It is in the intermediate topologies where the definition becomes more useful.

Example: c) Let X and Y be real lines, and let $\varphi : X \to Y$ be a map such that $\varphi(x) = 1$

if $x \geq 0$, $\varphi(x) = -1$ if $x < 0$. This map is not continuous because, for example, while the interval $I = (1/2, 3/2) \subseteq Y$ is open, its pre-image $\varphi^{-1}(I) = \{x \,|\, x \geq 0\}$ is not open.

Theorem 1.1 *The map $\varphi : X \to Y$ is continuous if and only if given any point $x \in X$ and any neighborhood $M \subseteq Y$ of $\varphi(x)$, there exists a neighborhood $N \subseteq X$ of x such that $\varphi(N) \subset M$.*

This theorem provides an equivalent definition of continuity that is much closer to the intuitive concept of continuity.

Proof:

Suppose φ is continuous. Le t x be a point of X, and $M \subseteq Y$ a neighborhood of $\phi(x)$. By definition of neighborhood, there exists an open set $O \subset M$ in Y and containing $\phi(x)$. By continuity, $N = \varphi^{-1}(O)$ is an open set of X, and as it contains x, it is a neighborhood of x. It is then fulfilled that $\varphi(N) \subset O \subset M$. Now to prove the reciprocal, suppose that given any point $x \in X$ and any neighborhood M of $\varphi(x)$, there exists a neighborhood N of x such that $\varphi(N) \subset M$. Then let O be any open set of Y, we must now show that $\varphi^{-1}(O)$ is an open set of X. Let x be any point of $\varphi^{-1}(O)$, then $\varphi(x) \in O$ and therefore O is a neighborhood of $\varphi(x)$, therefore there exists a neighborhood N of x such that $\varphi(N) \subset O$ and therefore $N \subset \varphi^{-1}(O)$. But then $\varphi^{-1}(O)$ contains a neighborhood of each of its points and therefore it is open. ♠

Exercise: Let $\phi : X \to Y$ and $\psi : Y \to Z$ be continuous maps, prove that the composite map $\psi \circ \phi : X \to Z$ is also continuous. (Composition of maps preserves continuity.)

Induced Topology:

Let ϕ be a map between a set X and a topological space $\{Y, \mathcal{T}\}$. This map naturally provides, that is, without the help of any other structure, a topology on X, denoted by $\mathcal{T}_\phi$ and called the **topology induced** by ϕ on X. It is given by $\mathcal{T}_\phi = \{O \subset X \mid O = \phi^{-1}(Q), Q \in \mathcal{T}\}$, that is, O is an open set of X if there exists an open set Q of Y such that $O = \phi^{-1}(Q)$. In other words, the open sets in X are the pre-images of all the open sets in Y. Note that under this topology on X, the map ϕ is automatically continuous. In fact, it is by construction the "minimal" topology on X that renders ϕ continuous.

Exercise: Prove that this construction really defines a topology.
Not all topologies thus induced are of interest and, in general, they depend strongly on the map, as shown by the following example:

Example:
a) Let $X = Y = \mathbb{R}$ with the usual topology and let $\phi : \mathbb{R} \to \mathbb{R}$ be the function $\phi(x) = 17 \ \forall x \in \mathbb{R}$. This function is clearly continuous with respect to the topologies of X and Y, those of the real line. However, the topology induced on X by this map is the indiscrete one!: $\mathcal{T}_\phi = \{\emptyset, X\}$. Interestingly, this "loss" of topological information can be intuitively associated to the information lost by ϕ in mapping the whole $\mathbb{R}$ to a single point, 17. In fact, ϕ would be continuous with respect to *any* topology on X, so its continuity is not very interesting, as it does not tell us anything about the underlying topology.
b) Let X and Y be as in a) and let $\phi(x)$ be an invertible map, then $\mathcal{T}_\phi$ coincides with the topology of the real line.

1.2.2 Compactness

The second generalization that we are interested in corresponds to the concept of *compactness*. For this, we introduce the following definition: let X be a set, A a subset of X, and $C = \{A_\lambda\}$ a collection of subsets of X parameterized by a continuous or discrete variable λ. We say that this collection **covers** A if $A \subset \cup_\lambda A_\lambda$. In such a case, we also say that C is a **cover** of A. When A is a topological space and the sets $\{A_\lambda\}$ are open, we say that C is an **open cover** of A. Finally, a subcollection of C that also covers A is called a **subcover**.

Definition: We say that a subset A of a topological space is **compact** if given any collection $\{A_\lambda\}$ of *open sets* that covers A, there exists a *finite* subcollection of these A_λ that also covers A.

Example: a) Let X be an infinite set of points with the discrete topology. Then a cover of X consists, for example, of all subsets of the form $\{x\}$, with $x \in X$. Since the topology of X is discrete, any cover is an open cover. But no finite number of the sets $\{x\}$ cover all of X, therefore X is not compact in this case. Clearly, if X had only a finite number of elements it would always be compact, regardless of its topology.

Example: b) Let X be any set with the indiscrete topology. Then X is compact. The only open sets of this set are $\emptyset$ and X, so any open cover has X as one of its members and this alone is enough to cover X.

Thus, we see that this property strongly depends on the topology of the set. The relation with the intuitive concept of compactness is clear from the following example and exercise.

Example: c) Let X be the real line and $A = (0, 1)$. This subset is not compact because, for example, the following is an open cover of A that has no finite subcover: $\{A_n = (\frac{1}{n}, \frac{n-1}{n}) : n \in \mathbb{N}\}$

Exercise: Let X be the real line and $A = [0, 1]$. Prove that A is compact.
Proof:

Let $\{A_\lambda\}$ be an open cover of $[0, 1]$. Consider the set S of all $x \in [0, 1]$ such that $[0, x]$ is covered by *finitely* many A_λ. We shall see that $S = [0, 1]$. This set is not empty since 0 is in some open set of the covering, and therefore there exists $x > 0$ such that $[0, x]$ is inside this open set. Furthermore it is bounded above by 1. Therefore, S is a non-empty upper-bounded real set, so by the least-upper-bound property of $\mathbb{R}$ it has a least upper bound. Let a be that number and A_{λ_0} be an element of the cover such that $a \in A_{\lambda_0}$. We assume $a < 1$ otherwise we would be done. Adding this set to the finite subcovering we still get a finite subcover containing a, and therefore also some other element $1 > b > a$. Contradicting the fact that a was a least upper bound.

♠

Now let's see the relation between the two derived concepts of Topology, namely the continuity of maps between topological spaces and compactness. The fact that a map between topological spaces is continuous implies that this map is special, in the sense that *it carries or conveys information about the respective topologies and preserves the topological properties of the sets it associates*. This is seen in the following property, which —as shown by the following example— is very important.

Theorem 1.2 *Let X and Y be two topological spaces and $\phi : X \to Y$ a continuous map between them. Then if A is a compact subset of X, $\phi(A)$ is a compact subset of Y.*

Proof: Let $\{O_\lambda\}$ be a collection of open sets in Y that cover $\phi(A)$. Then the collection $\{\phi^{-1}(O_\lambda)\}$ covers A, but A is compact and therefore there will be a finite subcollection $\{\phi^{-1}(O_{\tilde{\lambda}})\}$ of the former that also covers it. Therefore, the finite subcollection $O_{\tilde{\lambda}}$ will also cover $\phi(A)$. Since this is true for any open cover of $\phi(A)$ we conclude that $\phi(A)$ is compact.

Example: Let A be a compact topological space and let $\phi : A \to \mathbb{R}$ be a continuous map between A and the real line. $\phi(A)$ is then a compact real set and therefore a closed and bounded set, but then this set will have a maximum and a minimum, that is, the map ϕ reaches its maximum and minimum in A.

Finally, another theorem of fundamental importance about compact sets, which shows that they have another property that makes them very interesting. For this, we introduce the following definitions, which also only use topological concepts. A **sequence** in a set X, $\{x_n\} = \{x_1, x_2, ...\}$, with $x_n \in X$, is a map from $\mathbb{N}$ to this set. Given a sequence $\{x_n\}$ in a topological space X, we say that $x \in X$ is a **limit point** of this sequence if given any open set $O \subseteq X$ containing x there exists a number N such that for all $n > N$, $x_n \in O$. We say that $x \in X$ is an **accumulation point** of this sequence if given any open set $O \subseteq X$ containing x, infinitely many elements of the sequence also belong to O.

Exercise: Find an example of a sequence in some topological space with different limit points.

Theorem 1.3 *Let A be compact. Then every sequence in A has an accumulation point.*

Proof: Suppose —contrary to the theorem's assertion— that there exists a sequence $\{x_n\}$ in A without any accumulation points. That is, given any point $x \in A$ there exist a neighborhood O_x containing it and a number N_x such that if $n > N_x$ then $x_n \notin O_x$. Since this is valid for any x in A, the collection of sets $\{O_x | x \in A\}$ is an open cover of A, but A is compact and therefore it has a finite subcover. Let N be the maximum among the N_x of this finite subcover. But then $x_n \notin A$ for all $n > N$ which is absurd. ♠

Exercise: Prove that compact sets in the real line are the closed and bounded ones.

We can now ask the inverse question: if $A \subset X$ is such that every sequence has accumulation points, is it true then that A is compact? An affirmative answer would give us an alternative characterization of compactness, and this is affirmative for the case of the real line. In general, however, the answer is negative: there are topologies in which every sequence in a set has accumulation points in it, but this set is not compact. However, all the topologies we will see are **second countable** [See box] and in these the answer is affirmative.

In the real line, it is true that if $x \in \mathbb{R}$ is an accumulation point of a sequence $\{x_n\}$ then there exists a **subsequence**, $\{\tilde{x}_n\}$ (that is, a restriction of the map defining the sequence to an infinite subset of $\mathbb{N}$) having x as a limit point. This is also not true in the generality of topological spaces, but it is if we consider only those that are **first countable** [See box]. All the spaces we will see in this course are first countable.

***Countability of topological spaces.**

Definition: We say that a topological space $\{X, \mathcal{T}\}$ is first countable if for each $p \in X$ there exists a countable collection of open sets $\{O_n\}$ such that for every neighborhood N of p there exists i such that $O_i \subset N$.

Definition: We say that a topological space $\{X, \mathcal{T}\}$ is second countable if there is a countable collection of open sets such that any open set of X can be expressed as a union of sets from this collection.

Example:
a) X with the indiscrete topology is first countable.
b) X with the discrete topology is first countable. And second countable if and only if its elements are countable.

Exercise: Prove that the real line is first and second countable. Hint: For the first case, use the open sets $O_n = (p - \frac{1}{n}, p + \frac{1}{n})$ and for the second $O_{pqn} = (\frac{p}{q} - \frac{1}{n}, \frac{p}{q} + \frac{1}{n})$

***Separability of topological spaces.**

Definition: A topological space X is Hausdorff if given any pair of points of X, x and y, there exist neighborhoods O_x and O_y such that $O_x \cap O_y = \emptyset$.

Example:
a) X with the indiscrete topology is not Hausdorff.
b) X with the discrete topology is Hausdorff.

Exercise: Find a topology such that the integers are Hausdorff and compact.

Exercise: Prove that if a space is Hausdorff then if a sequence has a limit point, it is unique.

Exercise: Let X be compact, Y Hausdorff, and $\phi : X \to Y$ continuous. Prove that the images of closed sets are closed. Find a counterexample if Y is not Hausdorff.

Bibliography notes: This chapter is essentially a condensed version of Chapters **26, 27, 28, 29,** and **30** of [1], see also [2], [6], and [21]. Topology is one of the most fascinating branches of mathematics, if you delve deeper you will be captivated! Of particular interest in physics is the notion of *Homotopy*, a good place to understand these ideas is Chapter **34** of [1].

CHAPTER 2

LINEAR ALGEBRA

2.1 Vector Spaces

Definition: A **Real Vector Space** consists of three things: — *i*) a set, V, whose elements will be called **vectors**; *ii*) a rule that assigns to each pair of vectors, v, u, a third vector which we will denote by $v+u$ and call their **sum**; and *iii*) a rule that assigns to each vector, v and to each real number a, a vector which we will denote by av and call the **product** of v with a, all subject to the following conditions:

1.a) For any pair of vectors u, $v \in V$ it holds that,

$$u+v=v+u \tag{2.1}$$

1.b) There exists in V a unique element called **zero** and denoted by 0, such that

$$0+v=v \;\forall v \in V. \tag{2.2}$$

1.c) For any vector $u \in V$ there exists a unique vector in V, denoted $-u$, such that,

$$u+(-u)=0 \tag{2.3}$$

2.a) For any pair of real numbers a and a' and any vector v it holds that,

$$a(a'v)=(aa')v.$$

2.b) For any vector v it holds that,

$$1v=v.$$

3.a) For any pair of real numbers a and a' and any vector v it holds that,

$$(a+a')v=av+a'v.$$

3.b) For any real number a and any pair of vectors v and v' it holds that,

$$a(v+v') = av + av'.$$

The first conditions involve only the rule of addition; these are actually the rules that define a group, a structure we will revisit in a later chapter, where addition represents the product between elements of the group. The next conditions involve only the rule of multiplication, while the last two deal with the relation between these two operations. As we will see with examples later, real numbers can be replaced by any field, such as rationals, Q, integers, Z, complex numbers, C, and even finite fields.

Example: The set of all n-tuples of real numbers with the usual operations of componentwise addition and multiplication. This space is denoted by $\mathbb{R}^n$.

Example: Let S be any set and let V be the set of all functions from S to the reals, $v : S \to \mathbb{R}$, with the following operations of addition and multiplication: the sum of the function v with the function v' is the function (element of V) that assigns to the element $s \in S$ the value $v(s) + v'(s)$. The product of $a \in \mathbb{R}$ with the function v is the function that assigns to $s \in S$ the value $av(s)$. This example will appear very often in the following chapters.

Definition: We will say that a set of vectors $\{e_i\}$, $i = 1, \ldots, n$, are **linearly independent** if $\sum_{i=1}^{n} a^i e_i = 0 \; a^i \in \mathbb{R} \Longrightarrow a^i = 0$, $i = 1, \ldots, n$; that is, if any non-trivial linear combination of these vectors gives us a non-zero vector.

Definition: The **span** of a set of vectors $\{u_i\}$ is the set of all possible linear combinations of these elements. They generate a subspace of V, which we will denote by $Span\{u_i\} \subset V$.

If a finite number of linearly independent vectors, n, are sufficient to span V (that is, if any vector in V can be obtained as a linear combination of these n vectors) then we will say that these vectors form a **basis** of V and that the **dimension** of V is n, $\dim V = n$.

Exercise: Show that given a vector $v \in V$ and a basis of V, $\{e_i\}$ $i = 1, \ldots, n$, the linear combination of the basis vectors that equals v is unique. That is, if $v = \sum_i v^i e_i$ and $v = \sum_i \tilde{v}^i e_i$, then $v^i = \tilde{v}^i$ for all $i = 1, \ldots, n$.

Exercise: If we have two bases of a real vector space V of dimension n, $\{e_i\}$ and $\{\tilde{e}_i\}$, we can write the elements of one in terms of the other,

$$\tilde{e}_j = \sum_i P_j{}^i e_i, \quad e_i = \sum_l R_i{}^l e_l,$$

where ${P_j}^i$ and ${R_i}^l$ are the coefficients of two real $n \times n$ matrices. Prove that ${R_i}^l = {P^{-1}_i}^l$.

Exercise: Let $S = \{s_1, s_2, \ldots, s_n\}$ be a finite set. Find a basis of the vector space of all functions from S to $\mathbb{R}$. What is the dimension of this space?

Exercise: Show that the dimension of a vector space V is well defined, i.e. it does not depend on the basis used to define it.

Note that, in the previous example, V has finite dimension[1] because S is finite. If S had an infinite number of elements, V would have an infinite basis and we would say that the dimension of V would be infinite. In the rest of this Chapter, we will assume the vector spaces are finite-dimensional, unless otherwise stated.

Let V be a real vector space of dimension n and a basis of it, $\{\boldsymbol{e}_i\}$ $i = 1, \ldots, n$. Given an n-tuple of real numbers, $\{c^i\}$, we then have determined an element of V, that is, the vector, $v = \sum_{i=1}^{n} c^i \boldsymbol{e}_i$. On the other hand, we have seen in a previous exercise that, given any vector, it determines a unique n-tuple of real numbers, the coefficients of v in that basis. We see that we then have an invertible map between V and the space of n-tuples, $\mathbb{R}^n$. This map is linear, assigning to the sum of two vectors v and $\tilde{v}$ the n-tuple sum of the respective n-tuples. This map depends on the basis, but it still tells us that finite-dimensional real vector spaces do not hold many surprises: they are always copies of $\mathbb{R}^n$.

2.1.1 Covectors and Tensors

Let V be a vector space of dimension n. Associated with this vector space, consider the set of linear maps from V to $\mathbb{R}$, $V^* = \{\omega : V \to \mathbb{R} : \omega \text{is linear}\}$. This is also a vector space, called the **dual space to** V , or **space of covectors**, with addition and multiplication given by:

$$(\omega + \alpha\tau)(v) = \omega(v) + \alpha\tau(v) \quad \forall\, v \in V$$

with $\omega, \tau \in V^*$, $\alpha \in \mathbb{R}$.

What is the dimension of V^*? Note that if $\{\boldsymbol{e}_i\}$ $i = 1, \ldots, n$ is a basis of V, we can define n elements $\theta^i \in V^*$ by the relation

$$\theta^i(\boldsymbol{e}_j) = {\delta^i}_j. \tag{2.4}$$

That is, we define the action of θ^i on the basis vectors $\boldsymbol{e}_j$ as in the equation above, and then extend its action to any vector in V by writing this element in the basis $\{\boldsymbol{e}_i\}$ and using the fact that the θ^i act linearly on vectors.

It can be easily seen that any covector $\rho \in V^*$ can be obtained as a linear combination of the covectors $\{\theta^j\}$, $j = 1, \ldots, n$ and that these are linearly independent, therefore they form a basis and thus the dimension of V^* is also n. We will call this basis the **co-basis** of the basis $\{\boldsymbol{e}_i\}$.

[1] That is, a finite number of linearly independent vectors span V.

Exercise: Prove that V^* is a vector space and that the $\{\theta^i\}$ really form a basis.

Exercise: Prove that if $v = \sum_{i=1}^{n} v^i e_i$ then the coefficients can be obtained by acting on v with the covectors θ^i, that is,

$$v^i = \theta^i(v).$$

Exercise: Let V be the space of functions $S \to \mathbb{R}$, where S is a finite set with n elements. Let a basis of V be given by:

$$e_i(a) := \begin{cases} 1 & \text{if } a \text{ is the i-th element of} S \\ 0 & \text{otherwise} \end{cases}$$

Find the corresponding co-basis of its dual space.

Since V and V^* have the same dimension, they are, as vector spaces, isomorphic. For any basis of V have an associated co-basis of V^*, and we can extend this to associate to any vector in V a covector in V^*. However, this association is dependent on the choice of basis for V: if we choose a different basis, the association would be completely different. Then, this association is not intrinsic to the vector space. Since there is no *natural* map that identifies V with V^*, we have to consider them as intrinsically different, although isomorphic.

What happens if we now take the dual of V^*? Will we get another intrinsically different vector space? The answer is negative, since *there is* a natural identification between V and its double dual V^{**}.

Indeed, to each $v \in V$ we can associate an element X_v of V^{**} (that is, a linear functional from V^* to $\mathbb{R}$) in the following way: $X_v(\omega) := \omega(v) \;\; \forall \;\; \omega \in V^*$. $X_v(\omega)$ is sometimes called, unsurprisingly, the *evaluation map* on v. That is, the element X_v of V^{**} associated with $v \in V$ is the one that, when acting on any covector ω, gives the number $\omega(v)$. Note that X_v acts linearly on the elements of V^* and therefore is an element of V^{**}. Are there elements of V^{**} that do not come from some vector in V? The answer is negative, since the map $X_v : V \to V^{**}$ is injective [$X_v(\omega) = 0 \;\forall\; \omega \Longrightarrow v = 0$] and therefore[2] $\dim X_V = \dim V$. On the other hand $\dim V^{**} = \dim V^*$ since V^{**} is the dual of V^* and thus $\dim V = \dim V^* = \dim V^{**}$, which indicates that the map in question is also surjective and therefore invertible. This allows us to *identify* V and V^{**} and conclude that by dualizing further we will not be able to construct any more interesting vector spaces. In the case where the dimension of the vector space is not finite, this is no longer true and there are cases —frequently used in physics— where $X_V \subset V^{**}$ strictly.

Exercise: Given a basis of V, $\{e_i\}$, and the corresponding co-basis, $\{\theta^j\}$, define the co-co-basis of V^{**}, $\{E_i\}$. Find the relation between the components of a vector of the form X_v in the basis $\{E_i\}$ and those of the vector v in the basis $\{e_i\}$.

[2] Denoting by X_V the image by $X_{(\cdot)}$ of V.

Exercise: Prove that indeed $\dim \boldsymbol{X}_V = \dim V$.

However, nothing prevents us from also considering **multilinear maps** [3] from $\underbrace{V \times V \times \cdots \times V}_{k \text{ times}}$ to $\mathbb{R}$, or more generally,

$$\underbrace{V \times \cdots \times V}_{k \text{ times}} \times \underbrace{V^* \times \cdots \times V^*}_{l \text{ times}} \to \mathbb{R}.$$

The set of these maps, for each given pair (k,l), is also a vector space —with the obvious operations— and its elements are called **tensors of type** $\binom{l}{k}$.

Exercise: What is the dimension of these spaces as a function of their type $\binom{l}{k}$?

Note: In finite dimension, it can be shown that any tensor of type $\binom{l}{k}$ can be written as a linear combination of elements of the Cartesian product of k copies of V^* and l copies of V —where we have identified V with V^{**}—. For example, if $\boldsymbol{t}$ is of type $\binom{0}{2}$, —that is, a map that has as arguments two covectors—, then given a basis $\{\boldsymbol{e}_i\}$ of V, and the corresponding basis of V^{**}, $\{\boldsymbol{E}_i\}$ there will be $n \times n$ real numbers t^{ij}, $i = 1, \ldots, n$ such that

$$\boldsymbol{t}(\boldsymbol{\sigma}, \boldsymbol{\omega}) = \sum_{i,j=1}^{n} t^{ij} \boldsymbol{E}_i(\boldsymbol{\sigma}) \boldsymbol{E}_j(\boldsymbol{\omega}) = \sum_{i,j=1}^{n} t^{ij} \boldsymbol{\sigma}(\boldsymbol{e}_i) \boldsymbol{\omega}(\boldsymbol{e}_j), \quad \forall \boldsymbol{\sigma}, \boldsymbol{\omega} \in V^*. \tag{2.5}$$

But the set of linear combinations of Cartesian products of k copies of V^* and l copies of V is also a vector space, it is called the **outer product** of k copies of V^* and l copies of V and is denoted by

$$\underbrace{V^* \otimes V^* \otimes \cdots \otimes V^*}_{k \text{ times}} \otimes \underbrace{V \otimes \cdots \otimes V}_{l \text{ times}}.$$

Therefore, tensors can also be considered as elements of these outer products.

Example: a) Let $\boldsymbol{t}$ be of type $\binom{0}{2}$, that is, $\boldsymbol{t} \in V^* \otimes V^*$. This is a bilinear map from $V \times V$ to $\mathbb{R}$, $\boldsymbol{t}(\boldsymbol{v}, \boldsymbol{u}) \in \mathbb{R}$. Let $\boldsymbol{t}$ be symmetric $[\boldsymbol{t}(\boldsymbol{v}, \boldsymbol{u}) = \boldsymbol{t}(\boldsymbol{u}, \boldsymbol{v})]$ and non-degenerate $[\boldsymbol{t}(\boldsymbol{v}, \cdot) = 0 \in V^* \Longrightarrow \boldsymbol{v} = 0]$. Since $\boldsymbol{t}$ is non-degenerate, it defines an invertible map between V and its dual. Indeed, given $\boldsymbol{v} \in V$, $\boldsymbol{t}(\boldsymbol{v}, \cdot)$ is an element of V^*. But if $\boldsymbol{v}$ and $\tilde{\boldsymbol{v}}$ determine the same element of V^*, that is, if $\boldsymbol{t}(\boldsymbol{v}, \boldsymbol{u}) = \boldsymbol{t}(\tilde{\boldsymbol{v}}, \boldsymbol{u})$ $\forall\, \boldsymbol{u} \in V$ then $\boldsymbol{v} = \tilde{\boldsymbol{v}}$, which can be seen by taking $\boldsymbol{u} = \boldsymbol{v} - \tilde{\boldsymbol{v}}$ and using that $\boldsymbol{t}$ is non-degenerate. Since the dimensions of V and V^* are equal, the map thus defined is invertible.

Example: b) Let $\boldsymbol{\varepsilon}$ be a tensor of type $\binom{0}{n}$ such that

$$\boldsymbol{\varepsilon}(\ldots, \underbrace{\boldsymbol{v}}_{i}, \ldots, \underbrace{\boldsymbol{u}}_{j}, \ldots) = -\boldsymbol{\varepsilon}(\ldots, \underbrace{\boldsymbol{u}}_{i}, \ldots, \underbrace{\boldsymbol{v}}_{j}, \ldots) \tag{2.6}$$

[3] That is, maps that are separately linear in each of their arguments.

for any box i and j, that is, a totally antisymmetric tensor. Let $\boldsymbol{e}_i$ be a basis of V and $\varepsilon_{123\ldots n} := \boldsymbol{\varepsilon}(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$. Then any other component of $\boldsymbol{\varepsilon}$ in this basis will be either zero or $\varepsilon_{123\ldots n}$ or $-\varepsilon_{123\ldots n}$ depending on whether some $\boldsymbol{e}_i$ is repeated, or is an even permutation of the above, or an odd one. Indeed, for example,

$$\begin{aligned} \varepsilon_{3124\ldots n} &:= \boldsymbol{\varepsilon}(\boldsymbol{e}_3, \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_4, \ldots, \boldsymbol{e}_n) \\ &= -\boldsymbol{\varepsilon}(\boldsymbol{e}_1, \boldsymbol{e}_3, \boldsymbol{e}_2, \boldsymbol{e}_4, \ldots, \boldsymbol{e}_n) \\ &= \boldsymbol{\varepsilon}(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{e}_4, \ldots, \boldsymbol{e}_n) \\ &= \varepsilon_{1234\ldots n}. \end{aligned}$$

Therefore, given a basis, a single number, $\varepsilon_{123\ldots n}$, is enough to determine the tensor $\boldsymbol{\varepsilon}$ and given another tensor $\tilde{\boldsymbol{\varepsilon}}$ not identically zero with the properties mentioned above, there will exist a number α such that $\boldsymbol{\varepsilon} = \alpha\tilde{\boldsymbol{\varepsilon}}$. This last equality does not depend on the basis used and tells us that the dimension of the subspace of antisymmetric tensors of type $\binom{0}{n}$ is 1. Knowing one element is enough to generate the entire space by multiplying it by any real number.

Exercise: Let $\boldsymbol{\varepsilon}$ be a non identically zero, totally antisymmetric tensor of type $\binom{0}{n}$ and $\boldsymbol{u}_i$ a set of $n = \dim V$ vectors of V. Show that these form a basis if and only if

$$\boldsymbol{\varepsilon}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \neq 0. \tag{2.7}$$

Example: c) Let $\boldsymbol{A}$ be a tensor of type $\binom{1}{1}$, mapping

$$\boldsymbol{u} \in V,\ \boldsymbol{v}^* \in V^* \to \boldsymbol{A}(\boldsymbol{u}, \boldsymbol{v}^*) \in \mathbb{R}. \tag{2.8}$$

This implies that $\boldsymbol{A}(\boldsymbol{u}, \cdot)$ is also a vector —identifying V with V^{**}), the one that takes a covector $\boldsymbol{\omega} \in V^*$ and gives the number $\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{\omega})$—. Thus, $\boldsymbol{A}$ determines a **linear map** $V \to V$, that is, a linear operator on V.

Exercise: Continuing the example before, let $\boldsymbol{u}_i$ be a basis of V and let $\boldsymbol{a}_i = \boldsymbol{A}(\boldsymbol{u}_i, \cdot)$. Then

$$\boldsymbol{\varepsilon}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) = \boldsymbol{\varepsilon}(\boldsymbol{A}(\boldsymbol{u}_1, \cdot), \ldots, \boldsymbol{A}(\boldsymbol{u}_n, \cdot))$$

is totally antisymmetric in the $\boldsymbol{u}_i$ and therefore proportional to itself;

$$\boldsymbol{\varepsilon}(\boldsymbol{A}(\boldsymbol{u}_1, \ ), \ldots, \boldsymbol{A}(\boldsymbol{u}_n, \ )) \propto \boldsymbol{\varepsilon}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n).$$

The proportionality constant is called the **determinant** of the operator $\boldsymbol{A}$,

$$\boldsymbol{\varepsilon}(\boldsymbol{A}(\boldsymbol{u}_1, ;;), \ldots, \boldsymbol{A}(\boldsymbol{u}_n, ;;)) =: \det(\boldsymbol{A})\boldsymbol{\varepsilon}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n).$$

Problem 2.1 *Show that this definition does not depend on the $\boldsymbol{\varepsilon}$ used nor the basis and therefore gives truly a function on the space of operators $V \to \mathbb{R}$.*

Exercise: If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two operators on V, then $\boldsymbol{A} \cdot \boldsymbol{B}(\boldsymbol{v}) := \boldsymbol{A}(\boldsymbol{B}(\boldsymbol{v}))$. Show that $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.

2.1.2 Complexification

Another way to obtain vector spaces from a given one, say V, is by extending the field where the multiplication operation is defined, if this is possible. The most common case is the **complexification** of a real vector space; in this case, the product is simply extended to complex numbers, resulting in a vector space of the same dimension (over the complex field). One way to obtain it, for example, is by taking a basis of the real vector space V, and considering all linear combinations with arbitrary complex coefficients. The space thus obtained is denoted by $V^{\mathbb{C}}$. While the components of the vectors in V in the original basis were n-tuples of real numbers, they are now n-tuples of complex numbers. Since the basis is the same, the dimension is also the same. These extensions of vector spaces often appear, and we will see others later.

Multilinear maps must be extended in the same way. That is, for example, the dual of V will consist of all (complex-)linear maps from V to $\mathbb{C}$.

We can also take smaller fields, for example, $\mathbb{Q}^n$ or $\mathbb{Z}^n$.

Example: Consider the vector space $\mathbb{Q}^n$ consisting of all n-tuples of rational numbers. In this space, the field is also the set of rationals. If we extend the field to the reals, we obtain $\mathbb{R}^n$.

Example: Consider a space V and any basis $\{\boldsymbol{e}_i\}$ of V. This choice characterizes a subspace of V, given by all elements of the form,

$$v = \sum_i^n m^i \boldsymbol{e}_i \quad m^i \in Z.$$

Exercise: Now consider the set of all linear maps from this subspace to Z. What form do their elements take?

2.1.3 Quotient Spaces

The last way we will see to obtain vector spaces from other vector spaces is by taking **quotients**. Let V be a vector space and let $W \subset V$ be a subspace of it. We will call the **quotient space** the set of equivalence classes in V, where we will say that two vectors in V are equivalent if their difference is a vector in W. This space is denoted as V/W.

Exercise: Prove that the relation defined above is indeed an equivalence relation. (At the end of the chapter there is a box with a discussion of the relevant definitions and properties of the central concept of equivalence relations.)

Let's see that this is a vector space. The elements of V/W are equivalence classes; two elements of V, v and v', belong to the same equivalence class if $v - v' \in W$. Let ζ and ζ' be two elements of V/W, that is, two equivalence classes of elements of V. We will define the sum and the product of equivalence classes as follows: $\zeta + \alpha\zeta'$ will be the equivalence class corresponding to an element $\tilde{v}$ obtained by taking a

representative vector from ζ, say v, another from ζ', say v', and defining $\tilde{v} := v + \alpha v'$, we have $\tilde{\zeta} = \zeta + \alpha\zeta'$, where $\tilde{\zeta}$ is the equivalence class containing the element $\tilde{v} = v + \alpha v'$. To facilitate notation, the equivalence class containing a given element, say v, is usually represented as $[v]$. In this case, we have,

$$[v] + \alpha[v'] = [v + \alpha v'].$$

Exercise: See that this definition does not depend on the choice of representatives in the equivalence classes taken to perform the operation. That is, consider two other elements in ζ and ζ', say $\hat{v}$ and $\hat{v}'$, and see that with them you obtain an element in the same class as $\tilde{v} = v + \alpha v'$.

Example: Let $V = \mathbb{R}^2$, that is, the space of 2-tuples of real numbers. Let v be any element. This element generates the one-dimensional space W_v consisting of all vectors of the form αv, for $\alpha \in \mathbb{R}$. The quotient space V/W_v is the space composed of lines parallel to v. That is, each line is an element of the quotient space, and there is a notion of addition and scalar multiplication among them.

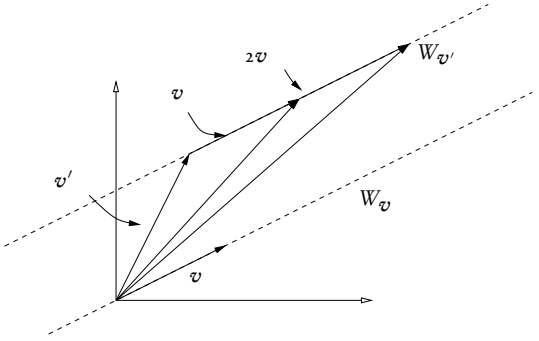

Figure 2.1: Geometric interpretation of the quotient space.

Exercise: Let V be the set of continuous functions from $\mathbb{R}$ to $\mathbb{R}$ and let W be the subset of those that vanish on the interval $[0, 1]$. See that this is a vector subspace. Consider the space V/W. What space can you associate with this?

2.2 Norms

Definition: A **norm** in a vector space V is a map $|x| : V \to \mathbb{R}^+$, satisfying for all $x, y \in V$, $\alpha \in \mathbb{R}$,,

i) Non-negativity: $|x| \geq 0$, with equality only for $x = 0$

ii) Homogeneity: $|\alpha x| = |\alpha|\,|x|$

iii) Triangle inequality: $|\boldsymbol{x}+\boldsymbol{y}| \leq |\boldsymbol{x}|+|\boldsymbol{y}|$

Examples: In $\mathbb{R}^2$:

a) $|(x,y)| := \max\{|x|,|y|\}$;

b) $|(x,y)|_2 := \sqrt{x^2+y^2}$ (Euclidean norm);

c) $|(x,y)|_1 := |x|+|y|$;

d) $|(x,y)|_p := (|x|^p+|y|^p)^{\frac{1}{p}} \quad p \geq 1$;

e) Let $\boldsymbol{t}$ be a positive-definite symmetric tensor of type $\binom{0}{2}$ on a vector space V, that is $\boldsymbol{t}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{t}(\boldsymbol{v},\boldsymbol{u})$, $\boldsymbol{t}(\boldsymbol{u},\boldsymbol{u}) \geq 0$ (with equality only for $\boldsymbol{u}=0$). The function $|\boldsymbol{u}|_{\boldsymbol{t}} = \sqrt{\boldsymbol{t}(\boldsymbol{u},\boldsymbol{u})}$ is a norm on V. Each tensor of this type generates a norm, but there are many norms that do not come from any tensor of this type. Give an example.

Exercise: Prove that $|\boldsymbol{t}(\boldsymbol{u},\boldsymbol{v})|^2 \leq |\boldsymbol{u}|_{\boldsymbol{t}}|\boldsymbol{v}|_{\boldsymbol{t}}$. Hint: Consider the polynomial: $P(\lambda) := \boldsymbol{t}(\boldsymbol{u}+\lambda\boldsymbol{v},\boldsymbol{u}+\lambda\boldsymbol{v})$.

Exercise: Prove that the given examples are indeed norms. Draw the level curves of the first four norms, that is, the sets $S_a = \{(x,y) \in \mathbb{R}^2 \,/\, |(x,y)| = a\}$ and the "balls of radius a", that is $B_a = \{(x,y) \in \mathbb{R}^2 / |(x,y)| \leq a\}$.

Exercise: Prove that the map $d : V \times V \to \mathbb{R}^+$ given by $d(\boldsymbol{x},\boldsymbol{y}) = |\boldsymbol{x}-\boldsymbol{y}|$ for a norm $|\cdot|$ on V defines a distance on V.

What is a norm geometrically? Given a vector $\boldsymbol{x} \neq 0$ of V and any positive number, a, there is a unique number $\alpha > 0$ such that $|\alpha\boldsymbol{x}| = a$. This indicates that the level surfaces of the norm, that is, the hypersurfaces $S_a = \{\boldsymbol{x} \in V / |\boldsymbol{x}| = a\}$, $a > 0$ form a smooth family of layers one inside the other, and each of them divides V into three disjoint sets: the *interior* of S_a —containing the element $\boldsymbol{x} = 0$—, S_a, and the *exterior* of S_a. The *interior* of S_a is a convex set, that is, if $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to the interior of S_a, then $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}$, $\alpha \in [0,1]$ also belongs to it (since $|\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}| \leq \alpha|\boldsymbol{x}|+(1-\alpha)|\boldsymbol{y}| \leq \alpha a+(1-\alpha)a = a$).

A level set completely characterizes a norm in the sense that if we give a subset N of V, such that N: **a)** has the radial property, that is, given $\boldsymbol{x} \neq 0$ there is a unique $\alpha > 0$ such that $\alpha, \boldsymbol{x} \in N$ and $-\alpha, \boldsymbol{x} \in N$ and **b)** is convex, then there is a unique norm such that N is the level surface S_1. This norm is defined as follows: given $\boldsymbol{x}$ we know that there will be a unique $\alpha > 0$ such that $\alpha\boldsymbol{x} \in N$ and then the norm of $\boldsymbol{x}$ will be $|\boldsymbol{x}| := \frac{1}{\alpha}$.

Exercise: Prove that this is a norm. Hint, given any two vectors $\boldsymbol{x}, \boldsymbol{y} \in V$, then $\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and $\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}$ are unitary and therefore are in N. But then we have that $\|\lambda\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}-(1-\lambda)\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}\| \leq 1 \quad \forall\, \lambda \in [0,1]$. Now find a convenient value for λ.

From this perspective, we see that given two norms on a finite-dimensional vector space and a level surface of one of them, there will be level surfaces of the other norm that will have the former inside or outside of it. In the norms a) and b) of the previous example, we see that given a square containing zero, there will be two circles

containing zero, one containing the square and the other contained by it. This leads us to the following theorem.

Theorem 2.1 *Let V be a finite-dimensional vector space. Then all its norms are equivalent to each other, in the sense that given $|\cdot|$ and $|\cdot|'$ there are positive constants M_1 and M_2 such that for all $\boldsymbol{x} \in V$ it holds that $M_1|\boldsymbol{x}| \leq |\boldsymbol{x}|' \leq M_2|\boldsymbol{x}|$.*

Proof: We will show that all are equivalent to the norm $|\boldsymbol{x}|_1 = \sum_{i=1}^{n} |a^i|$, where the a^i are the components of $\boldsymbol{x}$ with respect to a given basis $\{\boldsymbol{e}_i\}$, $\boldsymbol{x} = a^i\, \boldsymbol{e}_i$.

Let $|\cdot|$ be any other norm, then

$$\begin{aligned} |\boldsymbol{x}|| &\leq |\boldsymbol{x}| = |\textstyle\sum_i^n a^i, \boldsymbol{e}_i| \leq \sum_i^n |a^i, \boldsymbol{e}i| \\ &\leq \textstyle\sum_i^n |a^i|, |\boldsymbol{e}i| \leq (max\, j = 1, n|\boldsymbol{e}j|) \sum i = 1^n |a^i| \\ &= (max\, j = 1, n|\boldsymbol{e}_j|); |\boldsymbol{x}|_1. \end{aligned} \tag{2.9}$$

And we have easily obtained the upper bound. Now let's see the lower bound. To do this, we must prove that the norm $|\cdot|$ is, as a function from V to $\mathbb{R}^+$, a continuous function. This easily follows from the already found bound, indeed, let any other vector, $\boldsymbol{y} = b^i, \boldsymbol{e}i$, then

$$\begin{aligned} |, |\boldsymbol{x}| - ||\boldsymbol{y}|, | &\leq |\boldsymbol{x} - \boldsymbol{y}| \\ &= (max\, j = 1, n|\boldsymbol{e}_j|); |\boldsymbol{x} - \boldsymbol{y}|_1 \end{aligned} \tag{2.10}$$

This shows that the norm $|\cdot|$ is a continuous function with respect to the norm $|\cdot|_1$. Let S_1 be the level surface of radius 1 with respect to the metric $|\cdot|_1$. S_1 is a closed and bounded set and therefore compact. Therefore, by continuity, $|\cdot|$ has a maximum value, M_2, and a minimum, M_1, which give the sought inequality. The maximum value we have already found, the minimum is what allows us to bound the norm from below and conclude the theorem.

Notes:

i) In this proof, it is crucial that S_1 is compact. If V is infinite-dimensional, this is not the case, and there are many non-equivalent norms.

ii) For our purposes, any norm is sufficient —since if, for example, $f : V \to \mathbb{R}$ is continuous with respect to one norm, it is also continuous with respect to any other equivalent to it— and for simplicity, from now on, we will use the Euclidean norm.

iii) In this sense, the norms of finite-dimensional vector spaces are equivalent to the one generated by any positive-symmetric element of the outer product of its dual with itself.

iv) Since equivalent norms generate the same topology, we see that in finite-dimensional vector spaces, there is a unique topology associated with all its possible norms. This is usually called the **strong topology**.

Exercise: Prove that the above is indeed an equivalence relation among norms on a vector space.

2.2.1 Induced Norms in $V^\star$

The norms defined in V naturally induce norms in its dual, $V^\star$. This is given by:

$$|\omega| := \max_{|v|=1} |\omega(v)|. \tag{2.11}$$

In other words, the norm of a covector is the maximum value it takes on a unit vector.

Exercise: Show that this is a norm and that $|\omega(v)| \leq |\omega||v|$.

Exercise: Consider $V = \mathbb{R}^2$ with the norm $|(x,y)| := \max\{|x|,|y|\}$. What is the induced norm in $V^\star$?

2.3 Linear Operator Theory

A **linear operator** A on a vector space V is a continuous map[4] from V to V such that $\forall\, x,y \in V,\ \alpha \in \mathbb{R}, A(\alpha x + y) = \alpha A(x) + A(y)$. As we saw earlier, this is equivalent to saying that A is a tensor of type $\binom{1}{1}$.

The set of linear operators $\mathcal{L}$ is an algebra, that is, a vector space with a bilinear product among vectors. Indeed, if $A,\ B \in \mathcal{L}$, $\alpha \in \mathbb{R}$, then $A + \alpha B \in \mathcal{L}$ and also $A \cdot B$ (the operator that sends $x \in V$ to $A(B(x)) \in V$) also belongs to $\mathcal{L}$. Due to this product among vectors, we can also define non-linear functions from $\mathcal{L}$ to $\mathbb{R}$ and maps from $\mathcal{L}$ to $\mathcal{L}$. To study the continuity and differentiability of these maps, we introduce a norm in $\mathcal{L}$, the most convenient being the following norm induced by the one used in V,

$$||A||_{\mathcal{L}} = \max_{||x||_V=1} ||A(x)||_V. \tag{2.12}$$

Similarly as defined for covectors, the norm of a linear operator is the maximum norm attained by applying it to a unit vector.

If V is finite-dimensional (which we will assume from now on), the vector space $\mathcal{L}$ is also finite-dimensional and therefore all its norms are equivalent. In this case, the norm of A is finite because $A : V \to V$ is continuous and $\{x \in V / ||x||_V = 1\}$ is compact. However, in the infinite-dimensional case, neither of these statements is necessarily true, and within $\mathcal{L}$ we only have a subspace of linear operators with finite (bounded) norms.

Exercise: Show that

$$||A(v)|| \leq ||A||_{\mathcal{L}} ||v||.$$

[4] With respect to the topology induced by any of the equivalent norms of V.

Exercise: Using the result of the previous exercise, show that

$$||AB||_{\mathcal{L}} \leq ||A||_{\mathcal{L}} ||B||_{\mathcal{L}}.$$

Exercise: Show that $||\,||_{\mathcal{L}} : \mathcal{L} \to \mathbb{R}^+$ is a norm.

Next, we study various functions in the space of operators.

The determinant of an operator, introduced in the previous section, is a polynomial of degree $n = \dim V$ in A and therefore differentiable. Using the chain rule, we see that $\det(I + \lambda A)$ is differentiable in λ, and indeed a polynomial of degree n in λ. Each of the coefficients of this polynomial is a function of A. Of importance in what follows is the linear coefficient in A, which is obtained using the formula

$$\frac{d}{d\lambda} \det(I + \lambda A)|_{\lambda=0} = \frac{\varepsilon(A(u_1), u_2, \ldots, u_n) + \cdots + \varepsilon(u_1, \ldots, A(u_n))}{\varepsilon(u_1, \ldots, u_n)} \tag{2.13}$$

This function is called the **trace** of A and is denoted $\mathrm{tr}(A)$.

Among the maps from $\mathcal{L}$ to $\mathcal{L}$, consider the exponential map, defined as,

$$e^A = \sum_{i=0}^{\infty} \frac{A^i}{i!} = I + A + \frac{A^2}{2} + \cdots \tag{2.14}$$

Theorem 2.2 *$e^A \in \mathcal{L}$ if $A \in \mathcal{L}$ and e^{tA} is infinitely differentiable with respect to t.*

Proof: Consider the Cauchy sequence $\{e_n^A\}$, where $e_n^A \equiv \sum_{i=0}^{n} \frac{A^i}{i!}$. This sequence is Cauchy because, taking $m > n$, we have

$$\begin{aligned}
||e_m^A - e_n^A||_{\mathcal{L}} &= ||\frac{A^m}{(m)!} + \frac{A^{m-1}}{(m-1)!} + \frac{A^{m-2}}{(m-2)!} + \cdots + \frac{A^{n+1}}{(n+1)!}||_{\mathcal{L}} \\
&\leq ||\frac{A^m}{(m)!}||_{\mathcal{L}} + \frac{A^{m-1}}{(m-1)!}||_{\mathcal{L}} + ||\frac{A^{m-2}}{(m-2)!}||_{\mathcal{L}} + \cdots + ||\frac{A^{n+1}}{(n+1)!}||_{\mathcal{L}} \\
&\leq \frac{||A||_{\mathcal{L}}^m}{(m)!} + \frac{||A||_{\mathcal{L}}^{m-1}}{(m-1)!} + \frac{||A||_{\mathcal{L}}^{m-2}}{(m-2)!} + \cdots + \frac{||A||_{\mathcal{L}}^{n+1}}{(n+1)!} \\
&= |e_m^{||A||_{\mathcal{L}}} - e_n^{||A||_{\mathcal{L}}}| \to 0.
\end{aligned} \tag{2.15}$$

Where $e_n^{||A||_{\mathcal{L}}} \equiv \sum_{i=0}^{n} \frac{||A||_{\mathcal{L}}^i}{i!}$ and the last implication follows from the fact that the

numerical series $e^{\|\boldsymbol{A}\|}\mathscr{L}$ converges. But by completeness[5] of $\mathscr{L}$, every Cauchy sequence converges to some element of $\mathscr{L}$ that we will call $e^{\boldsymbol{A}}$. The differentiability of $e^{t\boldsymbol{A}}$ follows from the fact that if a series $\sum_{i=0}^{\infty} f_i(t)$ is convergent and $\sum_{i=0}^{\infty} \frac{df_i}{dt}$ is uniformly convergent, then $\frac{d}{dt}\sum_{i=0}^{\infty} f_i(t) = \sum_{i=0}^{\infty} \frac{d}{dt} f_i(t)$ ♠

Exercise: Show that
a) $e^{(t+s)\boldsymbol{A}} = e^{t\boldsymbol{A}} \cdot e^{s\boldsymbol{A}}$,
b) If $\boldsymbol{A}$ and $\boldsymbol{B}$ commute, that is if $\boldsymbol{AB} = \boldsymbol{BA}$, then $e^{\boldsymbol{A}+\boldsymbol{B}} = e^{\boldsymbol{A}}\, e^{\boldsymbol{B}}$.
c) $\det(e^{\boldsymbol{A}}) = e^{\mathrm{tr}(\boldsymbol{A})}$.
d) $\frac{d}{dt} e^{t\boldsymbol{A}} = \boldsymbol{A}\, e^{t\boldsymbol{A}}$.
Hint: For point c) use that $e^{\boldsymbol{A}}$ can also be defined as,

$$e^{\boldsymbol{A}} = \lim_{m\to\infty} \left(\boldsymbol{I} + \frac{\boldsymbol{A}}{m}\right)^m .$$

2.3.1 Matrix Representation

To describe certain aspects of linear operators, it is convenient to introduce the following matrix representation.

Let $\{\boldsymbol{u}_i\}$, $i = 1, \ldots, n$, be a basis of V. Applying the operator $\boldsymbol{A}$ to a member of the basis $\boldsymbol{u}_i$, we obtain a vector $\boldsymbol{A}(\boldsymbol{u}_i)$ which in turn can be expanded in the basis, $\boldsymbol{A}(\boldsymbol{u}_i) = \sum_{j=1}^{n} A^j{}_i\, \boldsymbol{u}_j$. The matrix thus constructed, $A^j{}_i$, is a representation of the operator $\boldsymbol{A}$ in that basis. In this language, we see that the matrix $A^j{}_i$ transforms the vector of components $\{v^i\}$ into the vector of components $\{A^j{}_i v^i\}$. Conversely, given a basis, $\{\boldsymbol{u}_i\}$, and a matrix $A^j{}_i$, we can construct a linear operator as follows: let $\{\theta^i\}$ be the associated co-basis, such that $\theta^i(\boldsymbol{u}_j) = \delta^i_j$. Then $\boldsymbol{A} = \sum_{i,j=1}^{n} A^j{}_i\, \boldsymbol{u}_j \otimes \theta^i$ defines a linear operator whose matrix coefficients in the basis are precisely $A^j i$.

If we change the basis, the matrices representing the operators will change. Indeed, if we take another basis $\{\hat{\boldsymbol{u}}_i\}$ and write its components with respect to the previous basis as $\hat{\boldsymbol{u}}_i = P^k{}_i \boldsymbol{u}_k$, and therefore, $\hat{\theta}^j = (P^{-1})^j{}_l \theta^l$, then the relation between the components of the operator $\boldsymbol{A}$ in both bases is given by

$$\hat{A}^j{}_i = A(\hat{\theta}^j, \hat{\boldsymbol{u}}_i) = (P^{-1})^j{}_l\, A^l{}_k\, P^k{}_i \quad \text{or} \quad \hat{A} = P^{-1} A P \tag{2.16}$$

that is, $\hat{A}$ and A are **similar matrices**.

Exercise: From the definition of the trace (equation (2.13)), prove that in terms of a basis we have $\mathrm{tr}A = \sum_{i=1}^{n} A^i{}_i$.

[5] Every finite-dimensional real vector space is complete.

Exercise: See with an example in two dimensions that the definition of determinant conforms with the usual one when we use a basis.

2.3.2 Invariant Subspaces

Definition: Let $A : V \to V$ be an operator and let W be a subspace of V. We will say that W is an **invariant subspace** of A if $AW \subseteq W$.

The invariant subspaces of an operator are important because they allow us to understand its action. Note that given any operator A, there are always at least two invariant spaces, V and $\{0\}$, and in reality, many more. For example, as we will see later, given any number between 1 and n (=dim V), there exists an invariant subspace with that number as its dimension. The ones that truly encode the action of the operator are its **irreducible** invariant subspaces, that is, those that cannot be further decomposed into invariant subspaces such that their direct sum is the whole V [6].

Example: Let V with $\dim(V) = 2$ and $A : V \to V$ given by, $A(\boldsymbol{u}_1) = \lambda_1 \boldsymbol{u}_1$, $A(\boldsymbol{u}_2) = \lambda_2 \boldsymbol{u}_2$, where $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ are linearly independent (and therefore a basis). Note that this completely defines the operator, since given any $\boldsymbol{v} \in V$, we can uniquely write it as $\boldsymbol{v} = v^1 \boldsymbol{u}_1 + v^2 \boldsymbol{u}_2$ and therefore,

$$A(\boldsymbol{v}) = \lambda_1 v^1 \boldsymbol{u}_1 + \lambda_2 v^2 \boldsymbol{u}_2.$$

In this case, the invariant subspaces are $Span\{\boldsymbol{u}_1\}$ and $Span\{\boldsymbol{u}_2\}$, and clearly, we have $V = Span\{\boldsymbol{u}_1\} \oplus Span\{\boldsymbol{u}_2\}$. Since each of these invariant subspaces is one-dimensional, the action of the operator on them is simply a dilation, that is, the multiplication of their elements by a number. Note that in the case where $\lambda_1 = \lambda_2$, the operator is proportional to the identity and therefore we have infinite invariant spaces.

Exercise: Let V be the space from the previous example and let A be given by $A(\boldsymbol{u}_1) = 0$, $A(\boldsymbol{u}_2) = \boldsymbol{u}_1$. Find its irreducible invariant subspaces. Do the same for the operator given by $A(\boldsymbol{u}_1) = \boldsymbol{u}_1$, $A(\boldsymbol{u}_2) = a\boldsymbol{u}_2 + \boldsymbol{u}_1$. What happens when $a = 1$?

We will study in detail the one-dimensional invariant subspaces, note that they are irreducible. To study the invariant spaces, it is convenient to consider the extended action of the linear operator to V^C, that is, the **complexification** of V.

Let's see that an operator always has at least a one-dimensional invariant subspace (and therefore always has a non-trivial irreducible invariant subspace).

Lemma 2.1 *Given $A : V^C \to V^C$, where V^C is finite-dimensional, there always exists $\boldsymbol{u} \in V^C$ and $\lambda \in C$ such that*

$$(A - \lambda I)\boldsymbol{u} = 0 \tag{2.17}$$

[6] A vector space is said to be the direct sum of two of its subspaces, W_1 and W_2, and is denoted by $V = W_1 \oplus W_2$ if each element of V can be uniquely written as the sum of two elements, one from each of these subspaces.

Proof:

A solution to this equation consists of a scalar λ, called an **eigenvalue** of the operator $\boldsymbol{A}$, and a vector $\boldsymbol{u}$, called an **eigenvector** of the operator $\boldsymbol{A}$. The subspace of V^C given by $\{\alpha \boldsymbol{u} \mid \alpha \in C\}$ is the invariant subspace sought.

It is clear that the system has a solution if and only if $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$. But this is a polynomial in λ of order equal to the dimension of V and therefore, by the Fundamental Theorem of Algebra, it has at least one solution or root, (generally complex), λ_1, and therefore there will be, associated with it, at least one $\boldsymbol{u}_1$ solution of (2.17) with $\lambda = \lambda_1$ ♠

The need to consider all these solutions is what leads us to treat the problem for complex vector spaces.

Exercise: In the infinite-dimensional case, the theorem is no longer true. Find an example of an operator on the set of infinite tuples without any eigenvector. Find it by looking for infinite matrices constructed in such a way that they have only one eigenvector and consider the limit to infinite components.

Application: Schur's Triangulation Lemma

Definition: An $n \times n$ matrix $A^j{}_i$ has an upper triangular form if $A^j{}_i = 0 \quad \forall j > i,\ j, i = 1, \ldots, n$. That is, it is a matrix of the form,

$$A = \begin{pmatrix} A^1{}_1 & A^1{}_2 & \cdots & \cdots & A^1{}_n \\ 0 & A^2{}_2 & \cdots & \cdots & A^2{}_n \\ 0 & 0 & \ddots & \ddots & A^3{}_n \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & A^n{}_n \end{pmatrix}. \tag{2.18}$$

As we will see later, in chapter **??**, this is a very convenient form to understand the solutions to systems of ordinary differential equations. And the most attractive thing about it is that any operator has a matrix representation with an upper triangular form! Moreover, if an inner product is present, the basis for this representation can be chosen to be orthonormal.

Lemma 2.2 (Schur) *Let $A : V \to V$ be a linear operator acting on a finite-dimensional complex vector space V, n, and let $(\cdot, \cdot)$ be an inner product on V. Then, there exists an orthonormal basis $\{\boldsymbol{u}_i\}$, $i = 1, \ldots, n$ with respect to which the matrix representation of $\boldsymbol{A}$ is upper triangular.*

Proof: Consider the eigenvalue-eigenvector problem for $\boldsymbol{A}$,

$$(\boldsymbol{A} - \lambda \boldsymbol{I})(\boldsymbol{u}) = 0. \tag{2.19}$$

As we have already seen, this problem always has at least one non-trivial solution, and therefore we have a pair $(\lambda_1, \boldsymbol{u}_1)$ solution of the problem. We take $\boldsymbol{u}_1$ of unit norm as the first element of the basis to be determined. We then have

$$A^j{}_1 := \theta^j(A(\boldsymbol{u}_1)) = \theta^j(\lambda_1 \boldsymbol{u}_1) = \lambda_1 \delta^j{}_1,$$

which gives us the result for the first column of the matrix.

Now consider the space

$$V_{n-1} = Span\{\boldsymbol{u}_1\}^\perp := \{\boldsymbol{u} \in V | (\boldsymbol{u}_1, \boldsymbol{u}) = 0\}$$

and the operator from $V_{n-1} \to V_{n-1}$ given by $A_1 := (\boldsymbol{I} - \boldsymbol{u}_1 \theta^1)\boldsymbol{A}$. Note that as we form an orthonormal basis, we already know the first member of the co-basis, $\theta^1 = (\boldsymbol{u}_1, \cdot)$. The operator $P_1 := \boldsymbol{I} - \boldsymbol{u}_1 \theta^1$ satisfies $P_1(\boldsymbol{u}_1) = 0$, $(P_1(v), \boldsymbol{u}_1) = 0$ and $P_1 \cdot P_1 = P_1$, that is, it is a projection operator to the subspace V_{n-1}.

We then have that $A_1 : V_{n-1} \to V_{n-1}$. Therefore, in this space, we can also pose the eigenvalue-eigenvector equation,

$$(A_1 - \lambda \boldsymbol{I})\boldsymbol{u} = ((\boldsymbol{I} - \boldsymbol{u}_1 \theta^1)\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{0}. \tag{2.20}$$

We thus obtain a new pair $(\lambda_2, \boldsymbol{u}_2)$, with $\boldsymbol{u}_2 \in V_{n-1}$, and therefore perpendicular to $\boldsymbol{u}_1$ and also, $\boldsymbol{A}\boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2 + \boldsymbol{u}_1 \theta^1(A(\boldsymbol{u}_2))$. Therefore

$$A^j{}_2 = \theta^j(A(\boldsymbol{u}_2)) = \theta^j(\lambda_2 \boldsymbol{u}_2 + \boldsymbol{u}_1 \theta^1(A(\boldsymbol{u}_2))) = \lambda_2 \delta^j{}_2 + \delta^j{}_1 A^1{}_2.$$

We thus see that with this choice of basis element, the second column of A satisfies the condition of the Lemma. The next step is to consider the subspace,

$$V_{n-2} = Span\{\boldsymbol{u}_1, \boldsymbol{u}_2\}^\perp = \{\boldsymbol{u} \in V | (\boldsymbol{u}_1, \boldsymbol{u}) = (\boldsymbol{u}_2, \boldsymbol{u}) = 0\},$$

and there the eigenvalue-eigenvector equation for the operator $(\boldsymbol{I} - \boldsymbol{u}_1 \theta^1 - \boldsymbol{u}_2 \theta^2)\boldsymbol{A}$. Proceeding in this way, we generate the entire basis ♠

The previous proof used an inner product that is actually exogenous to the property itself. We did it this way because the proof is conceptually simple and useful in the case that we are in the presence of a given inner product. However, the previous theorem can be proven using the notion of quotient space and thus dispensing with the inner product. This proof can be obtained by doing the following exercises:

Exercise: Let $\boldsymbol{A} : V \to V$ be a linear operator, let $W \subset V$ be invariant under $\boldsymbol{A}$, that is, $\boldsymbol{A}[W] \subset W$. This operator induces an operator in the quotient space V/W as follows: $\hat{\boldsymbol{A}}\zeta = \tilde{\zeta}$ where $\tilde{\zeta}$ is the equivalence class to which $\boldsymbol{A}(\boldsymbol{u})$ belongs when $\boldsymbol{u}$ belongs to ζ. See that this definition is consistent, that is, if we choose another element $v \in \zeta$, we obtain that $\boldsymbol{A}(v) \in \tilde{\zeta}$.

Exercise: The induced operator will have at least one eigenvalue-eigenvector pair in V/W, that is, there will be a pair $(\hat{\lambda}, \zeta)$ such that $\hat{\boldsymbol{A}}\zeta = \hat{\lambda}\zeta$. What does this equation mean in terms of the operator $\boldsymbol{A}$ in V?

Exercise: Specialize the previous case when W is the subspace of V generated by an eigenvector of $\boldsymbol{A}$. $\boldsymbol{A}\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$, $W_1 = Span\{u_1\}$. What does it mean, in terms of the space V and the operator $\boldsymbol{A}$, that $\hat{\boldsymbol{A}} : V/W_1 \rightarrow V/W_1$ has an eigenvalue-eigenvector pair?

Exercise: Iterate the previous procedure, taking at each step an element from each equivalence class of generalized eigenvectors to form a basis where the matrix representation of A is upper triangular.

We continue now with the study of invariant subspaces. If $\det(\boldsymbol{A} - \lambda \boldsymbol{I})$ has $1 \leq m \leq n$ distinct roots, $\{\lambda_i\}$, $i = 1, \ldots, m$, then there will be at least one complex eigenvector $\boldsymbol{u}_i$ associated with each of them. Let's see that these form distinct invariant subspaces.

Lemma 2.3 *Let $\{(\lambda_i, \boldsymbol{u}_i)\}$ $i = 1 \ldots m$ be a set of eigenvalue-eigenvector pairs. If $\lambda_i \neq \lambda_j$ $\forall\, i \neq j$, $i, j = 1 \ldots m$, then these eigenvectors are linearly independent.*

Proof: Suppose by contradiction that they are not, and therefore there exist constants $c^i \in \mathbb{C}$, $i = 1, \ldots, m-1$, such that

$$\boldsymbol{u}_m = \sum_{i=1}^{m-1} c^i \, \boldsymbol{u}_i \tag{2.21}$$

Applying $\boldsymbol{A}$ to both sides, we get

$$\boldsymbol{A}\boldsymbol{u}_m = \lambda_m \, \boldsymbol{u}_m = \sum_{i=1}^{m-1} c^i \, \lambda_i \, \boldsymbol{u}_i \tag{2.22}$$

or,

$$0 = \sum_{i=1}^{m-1} c^i \, (\lambda_m - \lambda_i) \, \boldsymbol{u}_i . \tag{2.23}$$

We conclude that $\{\boldsymbol{u}_i\}$ $i = 1, \ldots, m-1$ are linearly dependent. Due to 2.30 and the hypothesis that the eigenvalues are distinct, at least one of the coefficients must be non-zero, and therefore we can solve for one of the remaining eigenvectors in terms of the other $m-2$. Repeating this procedure $(m-1)$ times, we arrive at the conclusion that $\boldsymbol{u}_1 = 0$, which is a contradiction since, as we have seen, the eigenvector equation always has a non-trivial solution for each distinct eigenvalue ♠

If for each eigenvalue there is more than one eigenvector, then these form a higher-dimensional invariant vector subspace (reducible). Within each of these subspaces, we can take a basis composed of eigenvectors. The previous lemma ensures that all these eigenvectors thus chosen, for all eigenvalues, form a large linearly independent set.

Exercise: Convince yourself that the set of eigenvectors with the same eigenvalue forms a vector subspace.

If a given operator A has all its eigenvalues distinct, then the corresponding eigenvectors are linearly independent and equal in number to the dimension of V, that is, they generate a basis of V^C. In that basis, the matrix representation of A is diagonal, that is, $A^j{}_i = \delta^j{}_i \lambda_i$. Each of its eigenvectors generates an irreducible invariant subspace, and together they generate V^C. In each of them, the operator A acts merely by multiplication by λ_i. Note that the λ_i are generally complex, and therefore such multiplication is actually a rotation and a dilation. Note that, unlike the basis of Schur's triangulation lemma, this is not generally orthogonal with respect to a given inner product.[7]

Example: Let V be a vector space with $\dim(V) = 2$ and let A be given by $A(e_1) = e_2$, $A(e_2) = -e_1$, where (e_1, e_2) are any two linearly independent vectors. If we interpret them as two orthonormal vectors, then A is a rotation by $\pi/2$ in the plane.

Now let's calculate the determinant of $A - \lambda I$,

$$\begin{aligned} \det(A-\lambda I) &= \varepsilon((A-\lambda I)e_1,(A-\lambda I)e_2)/\varepsilon(e_1,e_2) \\ &= \varepsilon(e_2-\lambda e_1,-e_1-\lambda e_2)/\varepsilon(e_1,e_2) \\ &= 1+\lambda^2. \end{aligned} \tag{2.24}$$

and therefore the eigenvalues are $\lambda_1 = \imath$, and $\lambda_2 = -\imath$. The eigenvectors are $u_1 = e_1 + \imath e_2$ and $u_2 = e_1 - \imath e_2 = \bar{u}_1$. We see then that the action of A in these subspaces is multiplication by $\pm\imath$ and that both invariant subspaces are genuinely complex. In this new basis, the space V is generated by all linear combinations of the form $z u_1 + \bar{z} u_2$, and the action of A is simply multiplication by $\imath$ of z.

If the multiplicity of any of the roots of $\det(A-\lambda I) = 0$ is greater than one, there will be fewer eigenvalues than the dimension of the space, and therefore we will not be guaranteed to have enough eigenvectors to form a basis, as we can only guarantee the existence of one independent eigenvector for each eigenvalue.

Example: Let V be the set of 2-tuples of real numbers with a generic element (a, b) and let A be given by $A(a, b) = (\lambda a + \epsilon b, \lambda b)$. Taking a basis, $e_1 = (1, 0)$, $e_2 = (0, 1)$, we see that its matrix representation is:

$$\begin{pmatrix} \lambda & \epsilon \\ 0 & \lambda \end{pmatrix} \tag{2.25}$$

We see that this operator has only one eigenvalue with multiplicity 2. But it has only one independent eigenvector, proportional to $e_1 = (1, 0)$. For future use, note that if

[7] Note, however, that it can be *declared* orthogonal by defining the inner product as $(u, v) = \sum_{i=1}^{n} \theta^i(\bar{u})\theta^i(v)$

we define $\Delta = A - \lambda I$, then $e_1 = \frac{1}{\epsilon}\Delta e_2$, and therefore, in the basis $\{\tilde{e}_1 = e_1, \tilde{e}_2 = \frac{1}{\epsilon}e_2\}$, the operator is represented by the matrix,

$$\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \tag{2.26}$$

We must therefore analyze what happens in these cases. To do this, let us define, given λ_i an eigenvalue of A, the following subspaces:

$$W_{\lambda_i p} = \{u \in V \,|\, (A - \lambda_i I)^p u = 0\} \tag{2.27}$$

Note that these are invariant spaces: $AW_{\lambda_i p} \subset W_{\lambda_i p}$. Moreover, $W_{\lambda_i p} \subset W_{\lambda_i p+1}$, and therefore, for a sufficiently large p ($p \leq n$), we will have that $W_{\lambda_i p} = W_{\lambda_i p+1}$, taking the minimum among the p's where this occurs, we define $W_{\lambda_i} := W_{\lambda_i p}$. Note that if for some λ_i, $p = 1$, then the subspace W_{λ_i} is composed of eigenvectors. These are the maximum invariant spaces associated with the eigenvalue λ_i, indeed we have:

Lemma 2.4 *The only eigenvalue of A in W_{λ_i} is λ_i.*

Proof: Let λ be an eigenvalue of A in W_{λ_i}. Let's see that $\lambda = \lambda_i$. As we have already seen, there will be an eigenvector $\zeta \in W_{\lambda_i}$ with λ as the eigenvalue. Since it is in W_{λ_i}, there will be some $p \geq 1$ such that $(A - \lambda_i I)^p \zeta = 0$, but since it is an eigenvector, we have that $(\lambda - \lambda_i)^p \zeta = 0$, and therefore $\lambda = \lambda_i$ ♠

Now let's see that these subspaces are linearly independent and generate all of V^C. We will prove this theorem again in Chapter **??**.

Theorem 2.3 (See Chapter 5, Theorem 5.3) *Given an operator $A : V \to V$, with eigenvectors $\{\lambda_i\}$, $i = 1...m$, the space V^C admits a direct decomposition into invariant subspaces W_{λ_i}, where in each of them A has only λ_i as an eigenvalue.*

Proof:

The W_{λ_i} are independent. Let $v_1 + \cdots + v_s = 0$, with $v_i \in W_{\lambda_i}$, then we must prove that each $v_i = 0$. Applying $(A - \lambda_2 I)^{p_2} \ldots (A - \lambda_s I)^{p_s}$ to the previous sum, we get, $(A - \lambda_2 I)^{p_2} \ldots (A - \lambda_s I)^{p_s} v_1 = 0$, but since λ_i, $i \neq 1$, is not an eigenvalue of A in W_{λ_1}, the operator $(A - \lambda_2 I)^{p_2} \ldots (A - \lambda_s I)^{p_s}$ is invertible[8] in that subspace, and therefore $v_1 = 0$. Continuing in this way, we see that all the v_i must be zero.

The W_{λ_i} generate all of V^C. Suppose by contradiction that this is not the case, and consider V^C/W, where W is the space generated by all the W_{λ_i}, that is, the

[8] Note that $(A - \lambda_i I)^s|_{W_{\lambda_j}}$ is invertible if its determinant is non-zero. But $\det(A - \lambda_i I)^s = (\det(A - \lambda_i I))^s = (\lambda_j - \lambda_i)^{s \dim(W_{\lambda_j})} \neq 0$

space of all linear combinations of elements in the W_{λ_i}. The operator A acts on V^C/W $[A\{u\} = \{Au\}]$, and therefore it has an eigenvalue-eigenvector pair there. This implies that for some element ζ of V^C in some equivalence class of V^C/W, we have:

$$A\zeta = \lambda\zeta + u_1 + \cdots + u_s \tag{2.28}$$

where the u_i belong to each W_{λ_i}. Now suppose that $\lambda \neq \lambda_i \ \forall\, i = 1..s$, then $A - \lambda I$ is invertible in each W_{λ_i}, and therefore there exist vectors $\zeta_i = (A-\lambda I)^{-1} u_i \in W_{\lambda_i}$. But then $\tilde{\zeta} := \zeta - \zeta_1 - \cdots - \zeta_s$ is an eigenvalue of A! This is impossible since λ is not a root of the characteristic polynomial, nor is $\tilde{\zeta} = 0$, since belonging to ζ to V^C/W is not a linear combination of elements in the W_{λ_i}. We thus have a contradiction. Now suppose that $\lambda = \lambda_j$ for some $j \in \{1..s\}$. We can still define the vectors $\zeta_i = (A-\lambda_j I)^{-1} u_i$ for all $i \neq j$, and $\tilde{\zeta}$, where we only subtract from ζ all the ζ_i with $i \neq j$, therefore we have that

$$(A-\lambda_j I)\tilde{\zeta} = u_i \tag{2.29}$$

But applying $(A-\lambda_j I)^{p_j}$ to this equation, with p_j the minimum value for which $W_{\lambda_j p_j} = W_{\lambda_j p_j+1}$, we get that $\tilde{\zeta} \in W_{\lambda_j}$, and thus another contradiction, therefore it can only be that V^C/W is the trivial space, and the W_{λ_i} generate all of V^C ♠

We see that we only need to study each of these subspaces W_{λ_i} to find all their irreducible parts (from now on we will suppress the subscript i). But in these subspaces, the operator A acts very simply!

Indeed, let $\Delta_\lambda : W_\lambda \to W_\lambda$ be defined by $\Delta_\lambda := A|_{W_\lambda} - \lambda I|_{W_\lambda}$, then Δ has only 0 as an eigenvalue, and therefore it is **nilpotent** , that is, there exists an integer $m \leq n$ such that $\Delta_\lambda^m = 0$.

Lemma 2.5 *Let $\Delta : W \to W$ be such that its only eigenvalue is 0, then Δ is nilpotent.*

Proof: Let $W^p := \Delta^p[W]$, then we have that $W^p \subseteq W^q$ if $p \geq q$. Indeed, $W^p = \Delta^p[W] = \Delta^q[\Delta^{p-q}[W]] \subset \Delta^q[W]]$. Since the dimension of W is finite, it must happen that for some integer p, we will have that $W^p = W^{p+1}$, we see then that Δ^p acts injectively on W^p and therefore cannot have 0 as an eigenvalue. But we have seen that every operator has some eigenvalue, and therefore we have a contradiction unless $W^p = \{0\}$. That is, $\Delta^p = 0$ ♠

Nilpotent operators have the important property of generating a basis of the space in which they act from their repeated application on a smaller set of linearly independent vectorsContinuing now with the study of invariant subspaces. If $\det(A-\lambda I)$ has $1 \leq m \leq n$ distinct roots, $\{\lambda_i\}$, $i = 1,\ldots,m$, then there will be at least one complex eigenvector u_i associated with each of them. Let's see that these form distinct invariant subspaces.

Lemma 2.6 *Let* $\{(\lambda_i, \boldsymbol{u}_i)\}$ $i = 1 \ldots m$ *be a set of eigenvalue-eigenvector pairs. If* $\lambda_i \neq \lambda_j \quad \forall\, i \neq j,\ i, j = 1 \ldots m$, *then these eigenvectors are linearly independent.*

Proof: Suppose by contradiction that they are not, and therefore there exist constants $c^i \in \mathbb{C}$, $i = 1, \ldots, m-1$, such that

$$\boldsymbol{u}_m = \sum_{i=1}^{m-1} c^i \, \boldsymbol{u}_i \tag{2.30}$$

Applying $\boldsymbol{A}$ to both sides, we get

$$\boldsymbol{A}\boldsymbol{u}_m = \lambda_m \, \boldsymbol{u}_m = \sum_{i=1}^{m-1} c^i \, \lambda_i \, \boldsymbol{u}_i \tag{2.31}$$

or,

$$0 = \sum_{i=1}^{m-1} c^i \, (\lambda_m - \lambda_i) \, \boldsymbol{u}_i . \tag{2.32}$$

We conclude that $\{\boldsymbol{u}_i\}$ $i = 1, \ldots, m-1$ are linearly dependent. Due to 2.30 and the hypothesis that the eigenvalues are distinct, at least one of the coefficients must be non-zero, and therefore we can solve for one of the remaining eigenvectors in terms of the others $m-2$. Repeating this procedure $(m-1)$ times, we arrive at the conclusion that $\boldsymbol{u}_1 = 0$, which is a contradiction since, as we have seen, the eigenvector equation always has a non-trivial solution for each distinct eigenvalue ♠

If for each eigenvalue there is more than one eigenvector, then these form a higher-dimensional invariant vector subspace (reducible). Within each of these subspaces, we can take a basis composed of eigenvectors. The previous lemma ensures that all these eigenvectors thus chosen, for all eigenvalues, form a large linearly independent set.

Exercise: Convince yourself that the set of eigenvectors with the same eigenvalue forms a vector subspace.

If a given operator $\boldsymbol{A}$ has all its eigenvalues distinct, then the corresponding eigenvectors are linearly independent and equal in number to the dimension of V, that is, they generate a basis of V^C. In that basis, the matrix representation of $\boldsymbol{A}$ is diagonal, that is, $A^j{}_i = \delta^j{}_i \lambda_i$. Each of its eigenvectors generates an irreducible invariant subspace, and together they generate V^C. In each of them, the operator $\boldsymbol{A}$ acts merely by multiplication by λ_i. Note that the λ_i are generally complex, and therefore such multiplication is actually a rotation plus a dilation. Note that, unlike the basis of Schur's triangulation lemma, this is not generally orthogonal with respect to any given inner product.[9]

[9]Note, however, that it can be *declared* orthogonal by defining the inner product as $(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{n} \theta^i(\tilde{\boldsymbol{u}}) \theta^i(\boldsymbol{v})$

Example: Let V be a vector space with $\dim(V)=2$ and let A be given by $A(e_1)=e_2$, $A(e_2)=-e_1$, where (e_1,e_2) are any two linearly independent vectors. If we interpret them as two orthonormal vectors, then A is a rotation by $\pi/2$ in the plane.

Now let's calculate the determinant of $A-\lambda I$,

$$\begin{aligned}\det(A-\lambda I) &= \varepsilon((A-\lambda I)e_1,(A-\lambda I)e_2)/\varepsilon(e_1,e_2)\\ &= \varepsilon(e_2-\lambda e_1,-e_1-\lambda e_2)/\varepsilon(e_1,e_2)\\ &= 1+\lambda^2. \end{aligned} \tag{2.33}$$

and therefore the eigenvalues are $\lambda_1=\imath$, and $\lambda_2=-\imath$. The eigenvectors are $u_1=e_1+\imath e_2$ and $u_2=e_1-\imath e_2=\bar{u}_1$. We see then that the action of A in these subspaces is multiplication by $\pm\imath$ and that both invariant subspaces are genuinely complex. In this new basis, the space V is generated by all linear combinations of the form $zu_1+\bar{z}u_2$, and the action of A is simply multiplication by $\imath$ of z.

If the multiplicity of any of the roots $\det(A-\lambda I)=0$ is greater than one, there will be fewer eigenvalues than the dimension of the space, and therefore we will not be guaranteed to have enough eigenvectors to form a basis, as we can only guarantee the existence of one for each eigenvalue.

Example: Let V be the set of 2-tuples of real numbers with a generic element (a,b) and let A be given by $A(a,b)=(\lambda a+\epsilon b,\lambda b)$. Taking a basis, $e_1=(1,0)$, $e_2=(0,1)$, we see that its matrix representation is:

$$\begin{pmatrix}\lambda & \epsilon\\ 0 & \lambda\end{pmatrix} \tag{2.34}$$

We see that this operator has only one eigenvalue with multiplicity 2. But it has only one eigenvector, proportional to $e_1=(1,0)$. For future use, note that if we define $\Delta=A-\lambda I$, then $e_1=\frac{1}{\epsilon}\Delta e_2$, and therefore, in the basis $\{\tilde{e}_1=e_1,\tilde{e}_2=\frac{1}{\epsilon}e_2\}$, the operator is represented by the matrix,

$$\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix} \tag{2.35}$$

We must therefore analyze what happens in these cases. To do this, let us define, given λ_i an eigenvalue of A, the following subspaces:

$$W_{\lambda_i p}=\{u\in V\,|\,(A-\lambda_i I)^p u=0\} \tag{2.36}$$

Note that these are invariant spaces: $AW_{\lambda_i p}\subset W_{\lambda_i p}$. Moreover, $W_{\lambda_i p}\subset W_{\lambda_i p+1}$, and therefore, for a sufficiently large p ($p\leq n$), we will have that $W_{\lambda_i p}=W_{\lambda_i p+1}$, taking the minimum among the p's where this occurs, we define $W_{\lambda_i}:=W_{\lambda_i p}$. Note that if for some λ_i, $p=1$, then the subspace W_{λ_i} is composed of eigenvectors. These are the maximum invariant spaces associated with the eigenvalue λ_i, indeed we have:

Lemma 2.7 *The only eigenvalue of A in W_{λ_i} is λ_i.*

Proof: Let λ be an eigenvalue of A in W_{λ_i}. Let's see that $\lambda = \lambda_i$. As we have already seen, there will be an eigenvector $\zeta \in W_{\lambda_i}$ with λ as the eigenvalue. Since it is in W_{λ_i}, there will be some $p \geq 1$ such that $(A - \lambda_i I)^p \zeta = 0$, but since it is an eigenvector, we have that $(\lambda - \lambda_i)^p \zeta = 0$, and therefore $\lambda = \lambda_i$ ♠

Now let's see that these subspaces are linearly independent and generate all of V^C. We will prove this theorem again in Chapter **??**.

Theorem 2.4 (See Chapter ??, Theorem 5.3) *Given an operator $A : V \to V$, with eigenvectors $\{\lambda_i\}$, $i = 1...m$, the space V^C admits a direct decomposition into invariant subspaces W_{λ_i}, where in each of them A has only λ_i as an eigenvalue.*

Proof:

The W_{λ_i} are independent. Let $v_1 + \cdots + v_s = 0$, with $v_i \in W_{\lambda_i}$, then we must prove that each $v_i = 0$. Applying $(A - \lambda_2 I)^{p_2} \ldots (A - \lambda_s I)^{p_s}$ to the previous sum, we get, $(A - \lambda_2 I)^{p_2} \ldots (A - \lambda_s I)^{p_s} v_1 = 0$, but since λ_i, $i \neq 1$, is not an eigenvalue of A in W_{λ_1}, the operator $(A - \lambda_2 I)^{p_2} \ldots (A - \lambda_s I)^{p_s}$ is invertible[10] in that subspace, and therefore $v_1 = 0$. Continuing in this way, we see that all the v_i must be zero.

The W_{λ_i} generate all of V^C. Suppose by contradiction that this is not the case, and consider V^C/W, where W is the space generated by all the W_{λ_i}, that is, the space of all linear combinations of elements in the W_{λ_i}. The operator A acts on V^C/W $[A\{u\} = \{Au\}]$, and therefore it has an eigenvalue-eigenvector pair there. This implies that for some element ζ of V^C in some equivalence class of V^C/W, we have:

$$A\zeta = \lambda\zeta + u_1 + \cdots + u_s \tag{2.37}$$

where the u_i belong to each W_{λ_i}. Now suppose that $\lambda \neq \lambda_i \ \forall\, i = 1..s$, then $A - \lambda I$ is invertible in each W_{λ_i}, and therefore there exist vectors $\zeta_i = (A - \lambda I)^{-1} u_i \in W_{\lambda_i}$. But then $\tilde{\zeta} := \zeta - \zeta_1 - \cdots - \zeta_s$ is an eigenvalue of A! This is impossible since λ is not a root of the characteristic polynomial, nor is $\tilde{\zeta} = 0$, since belonging to ζ to V^C/W is not a linear combination of elements in the W_{λ_i}. We thus have a contradiction. Now suppose that $\lambda = \lambda_j$ for some $j \in \{1..s\}$. We can still define the vectors $\zeta_i = (A - \lambda_j I)^{-1} u_i$ for all $i \neq j$, and $\tilde{\zeta}$, where we only subtract from ζ all the ζ_i with $i \neq j$, therefore we have that

$$(A - \lambda_j I)\tilde{\zeta} = u_i \tag{2.38}$$

[10] Note that $(A - \lambda_i I)^s|_{W_{\lambda_j}}$ is invertible if its determinant is non-zero. But $\det(A - \lambda_i I)^s = (\det(A - \lambda_i I))^s = (\lambda_j - \lambda_i)^{s \dim(W_{\lambda_j})} \neq 0$

But applying $(A-\lambda_j I)^{p_j}$ to this equation, with p_j the minimum value for which $W_{\lambda_j p_j} = W_{\lambda_j p_j+1}$, we get that $\tilde{\zeta} \in W_{\lambda_j}$, and thus another contradiction, therefore it can only be that V^C/W is the trivial space, and the W_{λ_i} generate all of V^C ♠

We see that we only need to study each of these subspaces W_{λ_i} to find all their irreducible parts (from now on we will suppress the subscript i). But in these subspaces, the operator A acts very simply!

Indeed, let $\Delta_\lambda : W_\lambda \to W_\lambda$ be defined by $\Delta_\lambda := A|_{W_\lambda} - \lambda I|_{W_\lambda}$, then Δ has only 0 as an eigenvalue, and therefore it is **nilpotent** , that is, there exists an integer $m \leq n$ such that $\Delta_\lambda^m = 0$.

Lemma 2.8 *Let $\Delta : W \to W$ be such that its only eigenvalue is 0, then Δ is nilpotent.*

Proof: Let $W^p := \Delta^p[W]$, then we have that $W^p \subseteq W^q$ if $p \geq q$. Indeed, $W^p = \Delta^p[W] = \Delta^q[\Delta^{p-q}[W]] \subset \Delta^q[W]]$. Since the dimension of W is finite, it must happen that for some integer p, we will have that $W^p = W^{p+1}$, we see then that Δ^p acts injectively on W^p and therefore cannot have 0 as an eigenvalue. But we have seen that every operator has some eigenvalue, and therefore we have a contradiction unless $W^p = \{0\}$. That is, $\Delta^p = 0$ ♠

Nilpotent operators have the important property of generating a basis of the space in which they act by repeatedly applying them to a smaller set of linearly independent vectors.

Lemma 2.9 *Let $\Delta : W \to W$ be nilpotent, then there exists a basis of W consisting of elements of the form:*

$$\{\{v_1, \Delta v_1, \ldots, \Delta^{p_1} v_1\}, \ldots, \{\{v_d, \Delta v_d, \ldots, \Delta^{p_d} v_d\}\}$$

where p_i is such that $\Delta^{p_i+1} v_i = 0$

Note that if $n = \dim W$ then $n = \sum_{i=1}^{d} p_i$. Each of these sets formed by repeated applications of an operator is called a **cycle** . In this case, the basis is formed by the elements of d cycles. Note that cycles are not necessarily unique entities; indeed, if we have two cycles with the same number of elements, then any linear combination of them will also be a cycle. Note that each cycle contains only one eigenvector.

Example: Consider the matrix,

$$\Delta := \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.39}$$

its powers are,

$$\Delta^2 := \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \Delta^3 := \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \Delta^4 := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.40}$$

In this case, we have a single cycle, corresponding to the eigenvector $\boldsymbol{e}_1 = (1,0,0,0)$, which is the vector whose span is $\Delta^3[I\!R^4]$, the cycle is,

$$\begin{aligned} \boldsymbol{e}_4 &= (0,0,0,1) \\ \boldsymbol{e}_3 &= (0,0,1,0) = \Delta \boldsymbol{e}_4 \\ \boldsymbol{e}_2 &= (0,1,0,0) = \Delta \boldsymbol{e}_3 = \Delta^2 \boldsymbol{e}_4 \\ \boldsymbol{e}_1 &= (1,0,0,0) = \Delta \boldsymbol{e}_2 = \Delta^2 \boldsymbol{e}_3 = \Delta^3 \boldsymbol{e}_4. \end{aligned}$$

Example: Consider the matrix,

$$\Delta := \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.41}$$

its non-trivial powers are,

$$\Delta^2 := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \Delta^3 := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In this case, we have two cycles, corresponding to the two eigenvectors, $\boldsymbol{e}_1$ and $\boldsymbol{e}_3$.

Proof: We will prove it by induction on the dimension of W. If $n = 1$, we take any vector to generate the basis, since in this case $\Delta = 0$. We now assume it is true for any dimension less than n. In particular, since Δ has a zero eigenvalue, $\dim(\ker \Delta) \geq 1$, and therefore we have that $W' (= \Delta(W))$ has a dimension less than n, say n', and by the inductive hypothesis, a basis of the form

$$\{\{\boldsymbol{v}'_1, \Delta \boldsymbol{v}'_1, \ldots, \Delta^{p'_1} \boldsymbol{v}'_1\}, \ldots, \{\{\boldsymbol{v}'_{d'}, \Delta \boldsymbol{v}'_{d'}, \ldots, \Delta^{p'_{d'}} \boldsymbol{v}'_{d'}\}\}.$$

To form a basis of W, we will add to these vectors d' vectors $\boldsymbol{v}_i$ such that $\Delta \boldsymbol{v}_i = \boldsymbol{v}'_i$, $i = 1, \ldots, d'$. This can always be done since $\boldsymbol{v}'_i \in W' = \Delta W$. We thus see that

we have increased the set of vectors to

$$\{\{\boldsymbol{v}_1, \boldsymbol{\Delta}\boldsymbol{v}_1, \ldots, \boldsymbol{\Delta}^{p'_1+1}\boldsymbol{v}_1\}, \ldots, \{\boldsymbol{v}_{d'}, \boldsymbol{\Delta}\boldsymbol{v}_{d'}, \ldots, \boldsymbol{\Delta}^{p'_{d'}+1}\boldsymbol{v}_{d'}\}\},$$

that is, we now have $r = \sum_{i=1}^{d'}(p'_i + 1) = n' + d'$ vectors. To obtain a basis, we must then increase this set with $n - n' - d'$ vectors. Note that this number is non-negative; indeed, $n - n' = \dim(\ker \boldsymbol{\Delta}) \geq \dim(\ker \boldsymbol{\Delta} \cap W') = d'$, and it is precisely the dimension of the subspace of $\ker \boldsymbol{\Delta}$ that is not in W'. We then complete the proposed basis for W by incorporating into the already obtained set $n - n' - d'$ vectors $\{\boldsymbol{z}_i\}, i = 1, ..n - n' - d'$ from the null space of $\boldsymbol{\Delta}$ that are linearly independent among themselves and with the other elements of $\ker \boldsymbol{\Delta}$ in W and that are also not in W'. We have thus obtained a set of $d = d' + n - n' - d' = n - n'$ cycles. Let's see that the set thus obtained is a basis. Since they are n in number, we only need to see that they are linearly independent. We must then prove that if we have constants $\{C_{i,j}\}$, $i = 1..d$, $j = 0..p'_i + 1$ such that

$$0 = \sum_{i=1}^{d} \sum_{j=0}^{p'_i+1} C_{ij} \boldsymbol{\Delta}^{p_i} \boldsymbol{v}_i \tag{2.42}$$

then $C_{ij} = 0$. Applying $\boldsymbol{\Delta}$ to this relation, we get,

$$\begin{aligned} 0 &= \boldsymbol{\Delta} \sum_{i=1}^{d} \sum_{j=0}^{p'_i+1} C_{ij} \boldsymbol{\Delta}^{p_i} \boldsymbol{v}_i \\ &= \sum_{i=1}^{d'} \sum_{j=0}^{p'_i} C_{ij} \boldsymbol{\Delta}^{p'_i} \boldsymbol{v}'_i, \end{aligned} \tag{2.43}$$

where we have used that $\boldsymbol{\Delta}^{p'_i+1}\boldsymbol{v}'_i = 0$. But this is the orthogonality relation of the basis of W', and therefore we conclude that $C_{ij} = 0 \; \forall i \leq d', \;\; j \leq p'_i$. The initial relation is then reduced to

$$\begin{aligned} 0 &= \sum_{i=1}^{d} C_{i\,p'_i+1} \boldsymbol{\Delta}^{p'_i+1} \boldsymbol{v}_i \\ &= \sum_{i=1}^{d'} C_{i\,p'_i} \boldsymbol{\Delta}^{p'_i} \boldsymbol{v}'_i + \sum_{i=d'+1}^{d} C_{i\,1} \boldsymbol{z}_i, \end{aligned} \tag{2.44}$$

but the members of the first summation are part of the basis of W' and therefore linearly independent among themselves, while those of the second are a set of elements outside W' chosen to be linearly independent among themselves and with those of the first summation, and therefore we conclude that all C_{ij} are zero ♠

Alternative Proof: Alternatively, the previous lemma can be proven constructively. Indeed, if $m + 1$ is the power for which $\boldsymbol{\Delta}$ is nullified, we can take the space

$W^m = \Delta^m[W]$ and a basis $\{v_i^m\}$ of it. Note that all elements of W^m are eigenvectors of A, and therefore the elements of the basis. Then we consider the space $W^{m-1} = \Delta^{m-1}[W]$. Note that $W^m = \Delta^m[W] = \Delta^{m-1}[\Delta[W]]$ and $\Delta[W] \subset W$, therefore $W^m \subset W^{m-1}$. Since $W^m = \Delta[W^{m-1}]$, for each vector v_i^m of the basis $\{v_i^m\}$ of W^m there will be a vector v_i^{m-1} such that $\Delta v_i^{m-1} = v_i^m$. Since $W^m \subset W^{m-1}$, the set $\{v_i^m\} \cup \{v_i^{m-1}\}$ is contained in W^{m-1}. Note that $\dim W^m = \dim(\Delta[W^{m-1}]) \leq \dim(\ker \Delta \cap W^{m-1})$, since $W^m \subset W^{m-1}$ and all its elements belong to $\ker \Delta \cap W^{m-1}$.

Adding to the previous set a set $\{z_i\}$ of $\dim(\ker \Delta \cap W^{m-1}) - \dim(\ker \Delta \cap W^m)$ vectors from the null space of Δ in W^{m-1}, such that they are linearly independent among themselves and with the elements of the basis of W^m, we obtain a set of $\dim W^{m-1}$ vectors. Note that the mentioned choice of elements $\{z_i\}$ can be made since it is merely an extension of the basis $\{v_i^m\}$ of W^m to a basis of $\ker \Delta \cap W^{m-1}$. Now let's prove that they are linearly independent and therefore form a basis of W^{m-1}. To do this, we need to prove that if

$$0 = \sum_i C_i^m v_i^m + \sum_i C_i^{m-1} v_i^{m-1} + \sum_j C_j^z z_j \tag{2.45}$$

then each of the coefficients C_i^m, C_i^{m-1}, C_j must be zero. Multiplying the previous expression by Δ, we get,

$$0 = \sum_i C_i^{m-1} \Delta v_i^{m-1} = \sum_i C_i^{m-1} v_i^m, \tag{2.46}$$

but then the linear independence of the basis of W^m ensures that the $\{C_i^{m-1}\}$ are all zero. We then have,

$$0 = \sum_i C_i^m v_i^m + \sum_j C_j^z z_j. \tag{2.47}$$

But these vectors were chosen to be linearly independent among themselves and therefore all the coefficients in this sum must be zero. We thus see that the set $\{v_i^m\} \cup \{v_i^{m-1}\} \cup \{z_i\}$ forms a cyclic basis of W^{m-1}. Continuing with W^{m-2} and so on, we obtain a cyclic basis for all of W ♠

We thus see that the irreducible invariant subspaces of an operator are constituted by cycles within invariant subspaces associated with a given eigenvector. Each cycle contains a unique eigenvalue of the operator. We will denote the subspaces generated by these cycles (and usually also called cycles) by $C_{\lambda_i}^k$, where the lower index refers to the eigenvalue of the cycle and the upper index indexes the different cycles within each W_{λ_i}.

Exercise: Show that the obtained cycles are irreducible invariant subspaces of A.

2.3.3 Jordan Canonical Form

Definition: *Let $A : V \to V$ be a linear operator. We will say that A is of Jordan type with eigenvalue λ if there exists a basis $\{\boldsymbol{u}_i\}$ of V such that*[11]

$$A = \lambda \sum_{i=1}^{n} \boldsymbol{u}_i \theta^i + \sum_{i=2}^{n} \boldsymbol{u}_i \theta^{i-1} \equiv \lambda I + \Delta \tag{2.48}$$

where $\{\theta^i\}$ is the co-basis of the basis $\{\boldsymbol{u}_i\}$.

That is, in this basis, the components of A form a matrix $A^j{}_i$ given by

$$A = \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ & & \lambda & 1 & 0 \\ & 0 & & \lambda & 1 \\ & & & & \lambda \end{pmatrix} \tag{2.49}$$

Note that the matrix Δ is n-nilpotent, that is $\Delta^n = 0$.

Not every operator is of Jordan type —find one that is not— but it is clear that the restriction of any operator to one of its irreducible invariant subspaces (cycles) is. This can be seen by conveniently numbering the elements of the basis generated by the cycle.

Exercise: Find such an ordering of the elements of the basis.

Therefore, we can summarize the results found earlier in the following theorem about the matrix representations of any operator acting in a finite-dimensional space.

Theorem 2.5 *Jordan's Let $A : V \to V$ be a linear operator acting in a complex vector space V. Then, there exists a unique decomposition into a direct sum*[12] *of V into subspaces $C^k_{\lambda_i}$, $V = C^1_{\lambda_1} \oplus \cdots C^{k_i}_{\lambda_1} \oplus \cdots \oplus C^1_{\lambda_d} \oplus \cdots C^{k_d}_{\lambda_d}$, $d \leq n$ such that*

i) The $C^k_{\lambda_i}$ are invariant under the action of A, that is, $A\, C^k_{\lambda_i} \subseteq C^k_{\lambda_i}$

ii) The $C^k_{\lambda_i}$ are irreducible, that is, there are no invariant subspaces of $C^k_{\lambda_i}$ such that their sums are the entire $C^k_{\lambda_i}$.

iii) Due to property i), the operator A induces in each $C^k_{\lambda_i}$ an operator $A_i : C^k_{\lambda_i} \to C^k_{\lambda_i}$, which is of Jordan type with λ_i being one of the roots of the polynomial of degree n_i,

$$\det(A_i - \lambda_i I) = 0. \tag{2.50}$$

[11] In the sense that $A(v) = \lambda \sum_{i=1}^{n} \boldsymbol{u}_i \theta^i(v) + \sum_{i=2}^{n} \boldsymbol{u}_i \theta^{i-1}(v)\ \forall v \in V$

[12] Recall that a vector space V is said to be the direct sum of two vector spaces W and Z, and we denote it as $V = W \oplus Z$ if each vector in V can be obtained in a unique way as the sum of an element in W and another in Z.

This theorem tells us that given $\boldsymbol{A}$, there exists a basis, generally complex, such that the matrix of its components has the form of square diagonal blocks of $n_i \times n_i$, where n_i is the dimension of the subspace $C^k_{\lambda_i}$, each with the form given in 2.49. This form of the matrix is called the **Jordan canonical form**.

Exercise: Show that the roots, λ_i, that appear in the operators $\boldsymbol{A}_i$ are invariant under similarity transformations, that is, $\lambda_i(\boldsymbol{A}) = \lambda_i(\boldsymbol{PAP}^{-1})$.

Example: Let $\boldsymbol{A} : \mathbb{C}^3 \to \mathbb{C}^3$, then $\det(\boldsymbol{A} - \lambda \boldsymbol{I})$ is a polynomial of degree 3, and therefore has three roots. If these are distinct, there will be at least three invariant and irreducible subspaces of $\mathbb{C}^3$, but $\dim \mathbb{C}^3 = 3$ and therefore each of them has $n_i = 1$. The Jordan canonical form is then,
If two of them coincide, we have two possibilities: either we have three subspaces, in which case the Jordan canonical form will be

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \tag{2.51}$$

or, we have two subspaces, one necessarily of dimension 2, the Jordan canonical form will be

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \tag{2.52}$$

If all three roots coincide, then there will be three possibilities,

$$\begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \quad \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \text{ and } \quad \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix} \tag{2.53}$$

Example: We will now illustrate the case of coincident eigenvalues in two dimensions. This case is not generic, in the sense that any perturbation in the system —that is, any minimal change in the equations— separates the roots, making them distinct.

Let $\boldsymbol{A} : \mathbb{C}^2 \to \mathbb{C}^2$. In this case, the characteristic polynomial $\det(\boldsymbol{A} - \lambda \boldsymbol{I})$ has only two roots, which we will assume are coincident, $\lambda_1 = \lambda_2 = \lambda$. We find ourselves with two possibilities: either there are two linearly independent eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, in which case $V = B_1 \oplus B_2$ and $\boldsymbol{A}$ is diagonalizable ($A = \mathrm{diag}(\lambda, \lambda) = \boldsymbol{I}\lambda$), or there is only one eigenvector $\tilde{\boldsymbol{u}}_1$. In this case, let $\tilde{\boldsymbol{u}}_2$ be any vector linearly independent of $\tilde{\boldsymbol{u}}_1$, then $\boldsymbol{A}\tilde{\boldsymbol{u}}_2 = c^1 \tilde{\boldsymbol{u}}_1 + c^2 \tilde{\boldsymbol{u}}_2$ for some scalars c^1 and c^2 in $\mathbb{C}$. Calculating the determinant of $\boldsymbol{A} - \tilde{\lambda}\boldsymbol{I}$ in this basis, we get $(\lambda - \tilde{\lambda})(c^2 - \tilde{\lambda})$, but λ is a double root and therefore $c^2 = \lambda$.

Reordering and rescaling the bases $\boldsymbol{u}_1 = \tilde{\boldsymbol{u}}_1$, $\boldsymbol{u}_2 = c^1 \tilde{\boldsymbol{u}}_2$, we obtain

$$\begin{aligned} A\boldsymbol{u}_1 &= \lambda \boldsymbol{u}_1 \\ A\boldsymbol{u}_2 &= \lambda \boldsymbol{u}_2 + \boldsymbol{u}_1, \end{aligned} \tag{2.54}$$

and therefore

$$A = \lambda(\boldsymbol{u}_1 \otimes \theta^1 + \boldsymbol{u}_2 \otimes \theta^2) + \boldsymbol{u}_1 \otimes \theta^2, \tag{2.55}$$

where $\{\theta^i\}$ is the co-basis of the basis $\{\boldsymbol{u}_i\}$.

Note that $(A - \lambda I)\boldsymbol{u}_2 = \boldsymbol{u}_1$ and $(A - \lambda I)\boldsymbol{u}_1 = 0$, that is, $\Delta^2 = (A - \lambda I)^2 = 0$.

As we will see later in physical applications, the invariant subspaces have a clear physical meaning, they are called *normal modes* in the one-dimensional case and *cycles* in other cases.

2.3.4 Similarity Relation

In physics applications, the following equivalence relation is common: [See box at the end of the chapter.] We will say that the operator $\boldsymbol{A}$ is **similar** to the operator $\boldsymbol{B}$ if there exists an invertible operator $\boldsymbol{P}$ such that

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^{-1}. \tag{2.56}$$

That is, if we *rotate* V with an invertible operator $\boldsymbol{P}$ and then apply $\boldsymbol{A}$, we obtain the same action as if we first apply $\boldsymbol{B}$ and then *rotate* with $\boldsymbol{P}$.

Exercise:
a) Prove that similarity is an equivalence relation.
b) Prove that the functions and maps defined above are invariant within the equivalence classes, that is,

$$\begin{aligned} \det(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}) &= \det(\boldsymbol{A}) \\ \operatorname{tr}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}) &= \operatorname{tr}(\boldsymbol{A}) \\ e^{\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}} &= \boldsymbol{P}e^{\boldsymbol{A}}\boldsymbol{P}^{-1}. \end{aligned} \tag{2.57}$$

2.4 Adjoint Operators

Let A be a linear operator between two vector spaces, $A : V \to W$, that is, $A \in \mathscr{L}(V, W)$. Since V and W have dual spaces, this operator naturally induces a linear operator from W' to V', called its **dual** ,

$$A'(\omega)(v) := \omega(A(v)) \quad \forall \ \omega \in W', \quad v \in V. \tag{2.58}$$

That is, the operator that when applied to an element $\omega \in W'$ gives us the element $A'(\omega)$ of V' which, when acting on $v \in V$, gives the number $\omega(A(v))$. See figure.

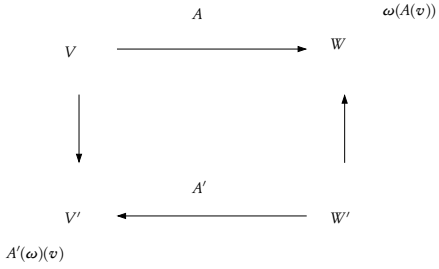

Figure 2.2: Diagram of the dual operator.

Note that this is a linear operator since,

$$\begin{aligned} A'(\alpha\omega + \sigma)(v) &= (\alpha\omega + \sigma)(A(v)) \\ &= \alpha\omega(A(v)) + \sigma(A(v)) \\ &= \alpha A'(\omega)(v) + A'(\sigma)(v). \end{aligned} \tag{2.59}$$

In the matrix representation, this operator is represented merely by the same matrix as the original, but now acting on the left, $A'(w)_i = w_j A^j{}_i$, that is, $A'^j_i = A^j{}_i$.

If there are norms defined in V and W and we define the norm of $A : V \to W$ in the usual way,

$$||A|| := \sup_{||v||_V = 1} \{||A(v)||_W\} \tag{2.60}$$

Then we see that

$$
\begin{aligned}
||A'|| &:= \sup_{||\omega||_{W'}=1} \{||A'(\omega)||_{V'}\} \\
&= \sup_{||\omega||_{W'}=1} \{ \sup_{||v||_V=1} \{|A'(\omega)(v)|\}\} \\
&= \sup_{||\omega||_{W'}=1} \{ \sup_{||v||_V=1} \{|\omega(A(v))|\}\} \\
&\leq \sup_{||\omega||_{W'}=1} \{ \sup_{||v||_V=1} \{||\omega||_{W'}||(A(v))||_W\}\} \\
&= \sup_{||v||_V=1} \{||(A(v))||_W\} \\
&= ||A||. \qquad (2.61)
\end{aligned}
$$

Thus we see that if an operator is bounded, then its dual is also bounded. In fact, the equality of the norms can be proven, but this would require introducing new tools (in the most general case, the Hahn-Banach theorem) that we do not wish to incorporate in this text.

Let's see what the components of the dual of an operator are in terms of the components of the original operator. Let $\{e_i\}$, $\{\theta^i\}$, $i = 1,..n$ be a basis and respectively a co-basis of V, and let $\{\hat{e}_i\}$, $\{\hat{\theta}^i\}$, $i = 1,..m$ be a pair of basis and co-basis of W. We then have that the components of A with respect to these bases are: $A^i{}_j := \hat{\theta}^i(A(e_j))$, that is, $A(v) = \sum_{i=1}^m \sum_{j=1}^n A^i{}_j \hat{e}_i \theta^j(v)$.

Therefore, if v has components $(v^1, v^2, \ldots, v^n)$ in the basis $\{e_i\}$, $A(v)$ has components

$$
(\sum_{i=1}^n A^1{}_i v^i, \sum_{i=1}^n A^2{}_i v^i, \ldots, \sum_{i=1}^n A^m{}_i v^i)
$$

in the basis $\{\hat{e}_i\}$

Now let's see the components of A'. By definition we have,

$$
A'^i{}_j := A'(\hat{\theta}^i)(e_j) = \hat{\theta}^i(A(e_j)) = A^i{}_j.
$$

That is, the same components, but now the matrix acts on the left on the components $(\omega_1, \omega_2, \ldots, \omega_m)$ of an element ω of W' in the co-basis $\{\hat{\theta}^i\}$. The components of $A'(\omega)$ in the co-basis $\{\theta^i\}$ are,

$$
(\sum_{i=1}^m A^i{}_1 \omega_i, \sum_{i=1}^m A^i{}_2 \omega_i, \ldots, \sum_{i=1}^m A^i{}_n \omega_i).
$$

A particularly interesting case of this construction is when $W = V$ and this is a space with an inner product, that is, a Hilbert space. In this case, the inner product gives us a canonical map between V and its dual V':

$$
\phi : V \to V', \quad \phi(v) := \langle v, \cdot \rangle. \qquad (2.62)
$$

This map is injective, and since V and V' have the same dimension, it is also surjective, and therefore invertible. That is, given $\omega \in V'$, there exists $v = \phi^{-1}(\omega) \in V$ such that $\langle v, \cdot \rangle = \omega$. Note then that $\phi^{-1} : V' \to V$ satisfies,

$$\langle \phi^{-1}(\omega), u \rangle = \langle v, u \rangle = \omega(u).$$

If $A : V \to V$, then $A' : V' \to V'$ can also be considered as an operator between V and V which we will call $A^\star$.

With the help of this map, we define $A^\star : V \to V$ given by: (See figure)

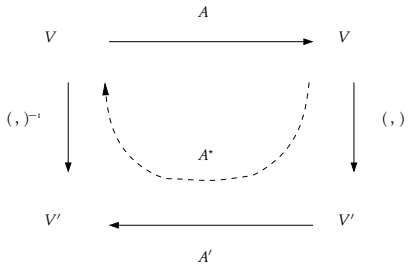

Figure 2.3: Diagram of the star operator.

$$A^\star(v) := \phi^{-1}(A'(\phi(v))). \tag{2.63}$$

In terms of the inner product, this is:

$$\langle A^\star(v), u \rangle = \langle \phi^{-1}(A'(\phi(v))), u \rangle = A'(\phi(v))(u) = \phi(v)(A(u)) = \langle v, A(u) \rangle. \tag{2.64}$$

In its matrix representation, this operator is,

$$A^{\star j}{}_i = t_{il} A^l{}_k (t^{-1})^{kj}, \quad \text{real inner product} \tag{2.65}$$

$$A^{\star j}{}_i = t_{il} \bar{A}^l{}_k (t^{-1})^{kj}, \quad \text{complex inner product} \tag{2.66}$$

where t_{li} is the representation of the inner product and $(t^{-1})^{jk}$ that of its inverse ($t_{il}(t^{-1})^{lk} = \delta^k{}_i$). If we choose a basis such that $t_{ik} = \delta_{ik}$, the Kronecker delta, the matrix representation of the adjoint is simply the transpose matrix, $A^{\star j}{}_i = A^{\dagger j}{}_i = A^i{}_j$,

A particularly interesting subset of operators is those for which $A = A^\star$. These operators are called **Hermitian** or **Self-adjoint** .[13]

Self-adjoint operators have important properties:

Lemma 2.10 *Let $M = Span\{u_1, u_2, \ldots, u_m\}$, where $\{u_i\}$ are a set of eigenvalues of A, a self-adjoint operator. Then M and $M^\perp$ are invariant spaces of A.*

[13] In the case of infinite dimension, these names do not coincide for some authors.

Proof: The first statement is clear and general, the second depends on the Hermiticity of A. Let $v \in M^\perp$ be any vector, let's see that $A(v) \in M^\perp$. Let $u \in M$ be arbitrary, then

$$\langle u, A(v)\rangle = \langle A(u), v\rangle = 0, \tag{2.67}$$

since $A(u) \in M$ if $u \in M$ ♠

This property has the following corollary:

Corolary 2.1 *Let $A : H \to H$ be self-adjoint. Then the eigenvectors of A form an orthonormal basis of H.*

Proof: $A : H \to H$ has at least one eigenvector, let's call it u_1. Now consider its restriction to the space perpendicular to u_1 which we also denote by A since by the previous lemma this is an invariant space, $A : \{u_1\}^\perp \to \{u_1\}^\perp$. This operator also has an eigenvector, say u_2 and $u_1 \perp u_2$. Now consider the restriction of A to $Span\{u_1, u_2\}^\perp$, there we also have $A : Span\{u_1, u_2\}^\perp \to Span\{u_1, u_2\}^\perp$ and therefore an eigenvector of A, u_3 with $u_3 \perp u_1$, $u_3 \perp u_2$. Continuing in this way, we end up with $n = \dim H$ eigenvectors all orthogonal to each other ♠

This theorem has several extensions to the case where the vector space is of infinite dimension. Later in chapter **??** we will see one of them.

Note that in this basis A is diagonal and therefore we have

Corolary 2.2 *Every self-adjoint operator is diagonalizable*

Also note that if u is an eigenvector of A self-adjoint, with eigenvalue λ, then,

$$\bar{\lambda}\langle u, u\rangle = \langle A(u), u\rangle = \langle u, A(u)\rangle = \lambda\langle u, u\rangle \tag{2.68}$$

and therefore $\bar{\lambda} = \lambda$, that is,

Lemma 2.11 *The eigenvalues of a self-adjoint operator are real*

Let's see what the condition of Hermiticity means in terms of the components of the operator in an orthonormal basis. Let A be a self-adjoint operator and let $\{e_i\}$, $i = 1, \ldots, n$ be an orthonormal basis of the space where it acts. We have that $\langle A(e_i), e_j\rangle = \langle e_i, A(e_j)\rangle$ and therefore, noting that $I = \sum_{i=1}^n e_i \theta^i$, we obtain,

$$\begin{aligned} 0 &= \langle \sum_{k=1}^n e_k \theta^k(A(e_i)), e_j\rangle - \langle e_i, \sum_{l=1}^n e_l \theta^l(A(e_j))\rangle \\ &= \sum_{k=1}^n \bar{A}^k{}_i \langle e_k, e_j\rangle - \sum_{l=1}^n A^l{}_j \langle e_i, e_l\rangle \\ &= \sum_{k=1}^n \bar{A}^k{}_i \delta_{kj} - \sum_{l=1}^n A^l{}_j \delta_{li} \\ &= \bar{A}^j{}_i - A^i{}_j \end{aligned} \tag{2.69}$$

from which we conclude that

$$\bar{A}^j{}_i = A^i{}_j \tag{2.70}$$

so the transpose matrix is the complex conjugate of the original. In the case of real matrices, we see that the condition is that in that basis the matrix is equal to its transpose, which is usually denoted by saying that the matrix is symmetric.

An interesting property of self-adjoint operators is that their norm is equal to the supremum of the magnitudes of their eigenvalues. Since the calculation demonstrating this will be used later in the course, we provide a demonstration of this fact below.

Lemma 2.12 *If A is self-adjoint then* $||A|| = \sup\{|\lambda_i|\}$.

Proof: Let $F(\boldsymbol{u}) := \langle \boldsymbol{A}(\boldsymbol{u}), \boldsymbol{A}(\boldsymbol{u})\rangle$ defined on the sphere $||\boldsymbol{u}|| = 1$. Since this set is compact (here we are using the fact that the space is finite-dimensional) it has a maximum which we will denote by $\boldsymbol{u}_o$. Note then that $F(\boldsymbol{u}_o) := ||A||^2$. Since $F(\boldsymbol{u})$ is differentiable on the sphere, it must satisfy

$$\frac{d}{d\lambda} F(\boldsymbol{u}_o + \lambda \boldsymbol{\delta u})|_{\lambda=0} = 0, \tag{2.71}$$

along any curve tangent to the sphere at the point $\boldsymbol{u}_o$, that is, for all $\boldsymbol{\delta u}$ such that $\langle \boldsymbol{u}_o, \boldsymbol{\delta u}\rangle = 0$. See figure.

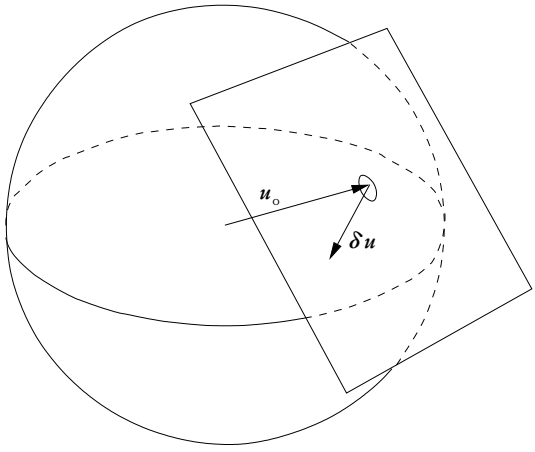

Figure 2.4: Normal and tangent vectors to the sphere.

But

$$\begin{aligned}\frac{d}{d\lambda}F(u_o+\lambda\delta u)|_{\lambda=o} &= \frac{d}{d\lambda}\langle A(u_o+\lambda\delta u),A(u_o+\lambda\delta u)\rangle|_{\lambda=o}\\ &= \langle A(\delta u),A(u_o)\rangle+\langle A(u_o),A(\delta u)\rangle\\ &= 2\Re\langle A^\star A u_o,\delta u\rangle\\ &= o \quad \forall\delta u,\ \ \langle u_o,\delta u\rangle = o \end{aligned} \tag{2.72}$$

Since δu is arbitrary in $\{u_o\}^\perp$ this simply implies

$$A^\star A(u_o) \in \{\delta u\}^\perp = \{u_o\}^{\perp\perp} = \{u_o\}. \tag{2.73}$$

and therefore there will exist $\alpha \in C$ such that

$$A^\star A(u_o) = \alpha u_o. \tag{2.74}$$

Taking the inner product with u_o we get,

$$\begin{aligned}\langle A^\star A(u_o),u_o\rangle &= \bar\alpha\langle u_o,u_o\rangle\\ &= \langle A(u_o),A(u_o)\rangle\\ &= ||A||^2 \end{aligned} \tag{2.75}$$

and therefore we have that $\alpha = \bar\alpha = ||A||^2$.

Now let $v := Au_o - ||A||u_o$, then, now using that A is self-adjoint, we have,

$$\begin{aligned}A(v) &= AA(u_o)-||A||A(u_o)\\ &= A^\star A(u_o)-||A||A(u_o)\\ &= ||A||^2 u_o-||A||A(u_o)\\ &= ||A||(||A||u_o-A(u_o))\\ &= -||A||v, \end{aligned} \tag{2.76}$$

therefore either v is an eigenvector of A, with eigenvalue $\lambda = -||A||$, or $v = o$, in which case u_o is an eigenvector of A with eigenvalue $\lambda = ||A||$ ♠

2.5 Unitary Operators

Another subclass of linear operators that appears very often in physics when there is a privileged inner product is that of **unitary operators** , that is, those such that their action preserves the inner product,

$$\langle U(u),U(v)\rangle = \langle u,v\rangle, \quad \forall u,v \in H. \tag{2.77}$$

The most typical case of a unitary operator is a transformation that sends one orthonormal basis to another. Usual examples are rotations in $\mathbb{R}^n$.
We also observe that $||U|| = \sup_{||v||=1}\{||U(v)||\} = 1$.

Note that

$$\langle U(u), U(v)\rangle = \langle U^* U(u), v\rangle = \langle u, v\rangle, \quad \forall u, v \in H, \tag{2.78}$$

that is,

$$U^* U = I \tag{2.79}$$

and therefore,

$$U^{-1} = U^*. \tag{2.80}$$

Let's see what the eigenvalues of a unitary operator U are. Let v_1 be an eigenvector of U (we know it has at least one), then,

$$\begin{aligned} \langle U(v_1), U(v_1)\rangle &= \lambda_1 \bar{\lambda}_1 \langle v_1, v_1\rangle \\ &= \langle v_1, v_1\rangle \end{aligned} \tag{2.81}$$

and therefore $\lambda_1 = e^{i\theta_1}$ for some angle θ_1.

If the operator U represents a non-trivial rotation in $\mathbb{R}^3$, then, given that we have an odd number of eigenvalues, there will be one that is real, the other two complex conjugates of each other. The eigenvector corresponding to the real eigenvalue defines the axis that remains fixed in that rotation. If we have more than one eigenvalue, then their corresponding eigenvectors are orthogonal, indeed, let v_1 and v_2 be two eigenvectors then

$$\begin{aligned} \langle U(v_1), U(v_2)\rangle &= \bar{\lambda}_1 \lambda_2 \langle v_1, v_2\rangle \\ &= \langle v_1, v_2\rangle \end{aligned} \tag{2.82}$$

and therefore if $\lambda_1 \neq \lambda_2$ we must have $\langle v_1, v_2\rangle = 0$.

Exercise: Show that if A is a self-adjoint operator, then $U := e^{iA}$ is a unitary operator.

Equivalence Relations.

Definition: An **equivalence relation**, $\approx$, between elements of a set X is a relation that satisfies the following conditions:

1. Reflexive: If $x \in X$, then $x \approx x$.
2. Symmetric: If $x, x' \in X$ and $x \approx x'$, then $x' \approx x$.
3. Transitive: If $x, x', x'' \in X$, $x \approx x'$ and $x' \approx x''$, then $x \approx x''$.

Note that the first property ensures that each element of X is related to some element of X, in this case with itself. Equivalence relations often appear in physics, essentially when we describe a physical process using a mathematical entity that has superfluous parts with respect to the process and therefore we would like to ignore them. This is achieved by declaring two mathematical entities that represent the same physical situation as equivalent entities.

Example: Let X be the set of real numbers and let $x \approx y$ if and only if there exists an integer n such that $x = n + y$, this is clearly an equivalence relation. This is used when we are interested in describing something using the straight line but that in reality should be described using a circle of unit circumference.
Given an equivalence relation in a set, we can group the elements of this set into **equivalence classes** of elements, that is, into subsets where all their elements are equivalent to each other and there is no element outside this subset that is equivalent to any of the elements of the subset. (If X is the set, $Y \subset X$ is one of its equivalence classes, and if $y \in Y$, then $y \approx y'$ if and only if $y' \in Y$.)

The fundamental property of equivalence relations is the following.

Theorem 2.6 *An equivalence relation in a set X allows regrouping its elements into equivalence classes such that each element of X is in one and only one of these.*

Proof: Let $x \in X$ and Y be the subset of all elements of X equivalent to x. Let's see that this subset is an equivalence class. Let y and y' be two elements of Y, that is, $y \approx x$ and $y' \approx x$, but by the transitivity property, $y \approx y'$. If $y \in Y$ and $z \notin Y$, then $y \not\approx z$, because otherwise z would be equivalent to x and therefore would be in Y. Finally, note that by reflexivity, x is also in Y. It only remains to see that if y is in Y and also in another equivalence class, say Z, then $Y = Z$. Since $y \in Y$, then y is equivalent to every element of Y, and since $y \in Z$, then y is equivalent to every element of Z, but by transitivity, every element of Y is equivalent to every element of Z, but since these are equivalence classes and therefore each contains all its equivalent elements, both must coincide.

Exercise: What are the equivalence classes of the previous examples?

2.6 Problems

Problem 2.2 *Let the operator $A : V \to V$ where $\dim V = n$, such that $Ax = \lambda x$. Calculate $\det A$ and $tr A$.*

Problem 2.3 *Let $V = \mathbb{R}^3$ and x be any non-zero vector. Find geometrically and analytically the quotient space V/W_x, where W_x is the space generated by x. Take another vector, x', linearly independent of the first and now calculate $V/W_{(x,x')}$.*

Problem 2.4 *The norm on operators is defined as:*

$$||A||_{\mathcal{L}} = max_{||x||_V=1}||A(x)||_V. \tag{2.83}$$

Find the norms of the following operators, given by their matrix representation with respect to a basis and where the norm in the vector space is the Euclidean norm with respect to that basis.

a)

$$\begin{pmatrix} 3 & 5 \\ 4 & 1 \end{pmatrix} \tag{2.84}$$

b)

$$\begin{pmatrix} 3 & 5 & 2 \\ 4 & 1 & 7 \\ 8 & 3 & 2 \end{pmatrix} \tag{2.85}$$

Problem 2.5 *Let V be any vector space and let $||\cdot||$ be a Euclidean norm in that space. The Hilbert-Schmidt norm of an operator is defined as:*

$$||A||^2_{HS} = \sum_{i,j=1}^{n} |A^j{}_i|^2. \tag{2.86}$$

where the basis used has been orthonormal with respect to the Euclidean norm.

a) Show that this is a norm.

b) Show that $||A||_{\mathcal{L}} \leq ||A||_{HS}$.

c) Show that $\sum_{j=1}^{n} |A^j{}_k|^2 \leq ||A||^2_{\mathcal{L}}$ for each k. Therefore $||A||^2_{HS} \leq n||A||^2_{\mathcal{L}}$, and the two norms are equivalent.

Hint: use that $\theta^j(A(u)) = \theta^j(A)(u)$ and then that $|\theta(u)| \leq ||\theta||||u||$.

Problem 2.6 *Calculate the eigenvalues and eigenvectors of the following matrices:*

a)

$$\begin{pmatrix} 3 & 6 \\ 4 & 1 \end{pmatrix} \tag{2.87}$$

b)

$$\begin{pmatrix} 3 & 6 \\ 0 & 1 \end{pmatrix} \tag{2.88}$$

c)

$$\begin{pmatrix} 2 & 4 & 2 \\ 4 & 1 & 0 \\ 3 & 3 & 1 \end{pmatrix} \tag{2.89}$$

Problem 2.7 *Bring the following matrices to upper triangular form: Note: From the transformation of the bases.*

a)

$$\begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix} \tag{2.90}$$

b)

$$\begin{pmatrix} 2 & 4 & 2 \\ 4 & 1 & 0 \\ 3 & 3 & 1 \end{pmatrix} \tag{2.91}$$

c)

$$\begin{pmatrix} 1 & 4 & 3 \\ 4 & 4 & 0 \\ 3 & 3 & 1 \end{pmatrix} \tag{2.92}$$

Problem 2.8 *Show again that* $\det e^{\boldsymbol{A}} = e^{tr\boldsymbol{A}}$. *Hint: express the matrix representation of* $\boldsymbol{A}$ *in a basis where it has the Jordan canonical form. Alternatively, use a basis where* $\boldsymbol{A}$ *is upper triangular and see that the product of upper triangular matrices gives an upper triangular matrix and therefore the exponential of an upper triangular matrix is also upper triangular.*

Bibliography notes: This chapter is based on the following books: [1], [3], [5] and [18]. The following are also of interest, [11] and [12]. Linear algebra is one of the largest and most prolific areas of mathematics, especially when dealing with infinite-dimensional spaces, which is usually called real analysis and operator theory. In my personal experience, most problems end up reducing to an algebraic problem and one feels that progress has been made when that problem can be solved.

CHAPTER 3

GEOMETRY

3.1 Manifolds

There are several reasons that justify the study of the concept of a manifold, or more generally of differential geometry, by physicists. One is that manifolds naturally appear in physics and therefore we cannot avoid them. Only in elementary courses can they be circumvented through vector calculus in $\mathbb{R}^n$. Thus, for example, to study the movement of a particle restricted to move on a sphere, we imagine the latter embedded in $\mathbb{R}^3$ and use the natural coordinates of $\mathbb{R}^3$ to describe its movements.

The second reason is that the concept of a manifold is of great conceptual utility, since, for example, in the case of a particle moving in $\mathbb{R}^3$, it allows us to clearly distinguish between the position of a particle and its velocity vector as mathematical entities of different nature. This fact is masked in $\mathbb{R}^3$ since this special type of manifold has the structure of a vector space.

Since the time of Galileo, we know that the language of physics is mathematics. Like any language, its utility goes beyond its daily use to understand each other and work on our ideas. Language allows for a synthesis of concepts and knowledge that encapsulates an entire area of knowledge into a smaller number of concepts. This allows future generations to understand an immense amount of knowledge that for previous generations were disparate aspects of reality as particular aspects of the same trunk of knowledge. The clearest example of this is the theories of the standard model of particles, which unify under the same phenomenon what we previously understood as distinct properties of matter. In particular, these theories naturally and deterministically incorporate elements of geometry, such as fiber bundles and connections, symmetries, etc.

A manifold is a generalization of Euclidean spaces $\mathbb{R}^n$ in which one preserves the concept of continuity, that is, its topology in the local sense but discards its character as a vector space. A manifold of dimension n is, in imprecise terms, a set of points that locally is like $\mathbb{R}^n$, but not necessarily in its global form.

An example of a two-dimensional manifold is the sphere, S^2. If we look at a sufficiently small neighborhood, U_p, of any point of S^2, we see that it is similar to a neighborhood of the plane, $\mathbb{R}^2$, in the sense that we can define a continuous and invertible map between both neighborhoods. Globally, the plane and the sphere are

topologically distinct, as there is no continuous and invertible map between them. [See figure 3.1.]

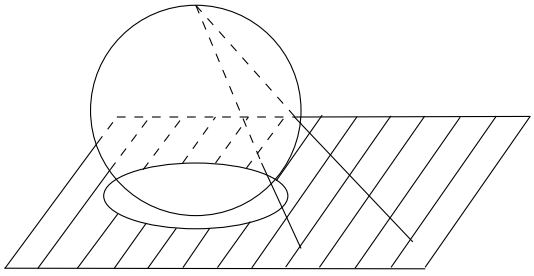

Figure 3.1: An atlas of the sphere.

As we said before, one might object to the need, in the previous example, to introduce the concept of a manifold, since one could consider S^2 as the subset of $\mathbb{R}^3$ such that $x_1^2 + x_2^2 + x_3^2 = 1$. The answer to this objection is that in physics one must follow the rule of economy of concepts and objects and discard everything that is not fundamental to the description of a phenomenon: if we want to describe things happening on the sphere, why do we need a space of more dimensions? This rule of economy forces us to refine concepts and discard everything superfluous. It is in this way that we advance in our maturation as physicists. It is the way we truly penetrate the mysteries of nature.

Note that the first way of working with the sphere is what we normally do when we seek to locate points and paths on the globe. In fact, we use flat maps, formerly called charts, to describe what happens in our cities and countries. When we want to have a collection of maps that cover the entire globe, we acquire an atlas, that is, a set of maps that cover the entire globe and have sectors in common between one and another. Some of these sectors are internal, for example when they describe a city within a country (on the map that covers that country), or on the edges when we go from one sheet to another. Only when we want to see the global structure of the globe, for example if we want to take a plane trip covering a large part of the globe, do we use a small version of the planet as implanted in $\mathbb{R}^3$.

We now give a series of definitions to finally arrive at the definition of an n-dimensional manifold.

Definition: Let M be a set. A **chart of** M is a pair (U, φ) where U is a subset of M and φ an injective map between U and $\mathbb{R}^n$, such that its image, $\varphi[U]$ is open in $\mathbb{R}^n$.

Definition: An **atlas of** M is a collection of charts $\{(U_i, \varphi_i)\}$ satisfying the following conditions: [See figure 3.2.]

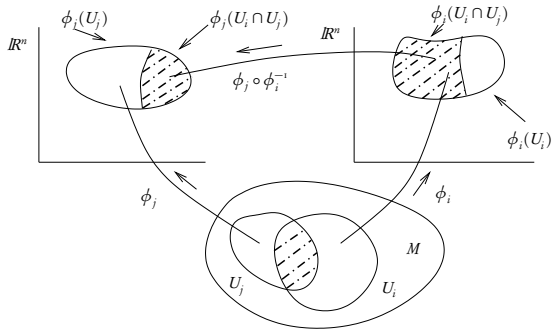

Figure 3.2: Relationship between charts.

1. The U_i cover M, $(M = \bigcup_i U_i)$.

2. If two charts overlap then $\varphi_i(U_i \cap U_j)$ is also an open set in $\mathbb{R}^n$.

3. The map $\varphi_j \circ \varphi_i^{-1} : \varphi_i[U_i \cap U_j] \to \varphi_j[U_i \cap U_j]$ is continuous, injective, and surjective.

Condition 1 gives us a notion of *closeness* in M induced from the analogous notion in $\mathbb{R}^n$. Indeed, we can say that a sequence of points $\{p_k\}$ in M converges to p in U_i if there exists k_o such that $\forall\ k > k_o$, $p_k \in U_i$ and the sequence $\{\varphi_i(p_k)\}$ converges to $\varphi_i(p)\}$. Another way to see this is that if after this construction of a manifold we impose that the maps φ_i are continuous, then we induce a unique topology on M. [See figure 3.3.]

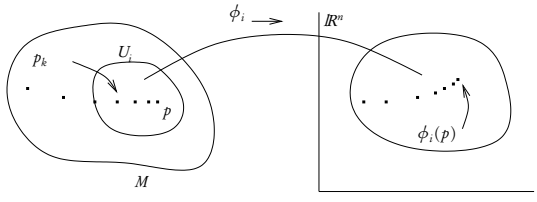

Figure 3.3: Sequences in M.

Condition 2 simply ensures that this notion is consistent. If $p \in U_i \cap U_j$ then the fact that the sequence converges is independent of whether we use the chart (U_i, φ_i) or (U_j, φ_j).

Condition 3 allows us to encode in the maps $\varphi_j \circ \varphi_i^{-1}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ the global topological information necessary to distinguish, for example, whether M is a sphere,

a plane, or a torus. Therein lies, for example, the information that there is no continuous and invertible map between S^2 and $\mathbb{R}^2$. But also, if we require these maps to be differentiable, it is what will allow us to formulate differential calculus on M. Indeed, note that in condition 3 we speak of the continuity of the map $\varphi_j \circ \varphi_i^{-1}$, which is well-defined because it is a map between $\mathbb{R}^n$ and $\mathbb{R}^n$. Similarly, we can speak of the differentiability of these maps.

We will say that **an atlas** $\{(U_i, \varphi_i)\}$ **is** C^p if the maps $\varphi_j \circ \varphi_i^{-1}$ are p-times differentiable and their p-th derivative is continuous.

One might be tempted to define the manifold M as the pair consisting of the set M and an atlas $\{(U_i, \varphi_i)\}$, but this would lead us to consider as different manifolds, for example, the plane with an atlas given by the chart $(\mathbb{R}^2, (x,y) \to (x,y))$ and the plane with an atlas given by the chart $(\mathbb{R}^2, (x,y) \to (x,-y))$.

To remedy this inconvenience, we introduce the concept of equivalence between atlases.

Definition: We will say that **two atlases are equivalent** if their union is also an atlas.

Exercise: Prove that this is indeed an equivalence relation $\approx$, that is, it satisfies:

i) $A \approx A$

ii) $A \approx B \implies B \approx A$

iii) $A \approx B, B \approx C \implies A \approx C$.

With this equivalence relation, we can divide the set of atlases of M into different **equivalent classes**. [Remember that each equivalent class is a set where all its elements are equivalent to each other and such that there is no element equivalent to these that is not in it.]

Definition: We will call **manifold** M **of dimension** n **and differentiability** p the pair consisting of the set M and an equivalent class of atlases, $\{\varphi_i : U_i \to \mathbb{R}^n\}$, in C^p.

It can be shown that to uniquely characterize the manifold M it is sufficient to give the set M and an atlas. If we have two atlases of M, then either they are equivalent and thus represent the same manifold, or they are not and then represent different manifolds.

The definition of a manifold that we have introduced is still too general for usual physical applications, in the sense that the allowed topologies can still be pathological from the point of view of physics. Therefore, in this course, we will impose an extra condition on manifolds. We will assume that they are **separable or Hausdorff**. That is, if p and $q \in M$, distinct, then: either they belong to the domain of the same chart U_i (in which case there exist neighborhoods W_p of $\varphi_i(p)$ and W_q of $\varphi_i(q)$ such that $\varphi_i^{-1}(W_p) \cap \varphi_i^{-1}(W_q) = \emptyset$, that is, the points have disjoint neighborhoods) or there exist U_i and U_j with $p \in U_i$, $q \in U_j$ and $U_i \cap U_j = \emptyset$, which also implies that they have disjoint neighborhoods. This is a property in the topology of M that essentially says we can separate points of M. An example of a non-Hausdorff manifold is the following.

Example: M, as a set, consists of three intervals of the line $I_1 = (-\infty, 0]$, $I_2 = (-\infty, 0]$ and $I_3 = (0, +\infty)$.
An atlas of M is $\{(U_1 = I_1 \cup I_3, \varphi_1 = id), (U_2 = I_2 \cup I_3, \varphi_2 = id)\}$. [See figure 3.4.]

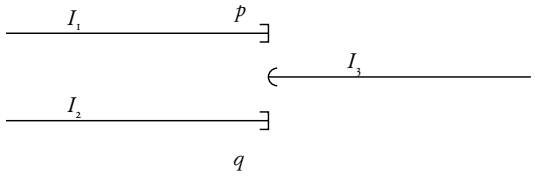

Figure 3.4: Example of a non-Hausdorff manifold.

Exercise: Prove that it is an atlas.
Note that given any neighborhood W_1 of $\varphi_1(0)$ in $\mathbb{R}$ and any neighborhood W_2 of $\varphi_2(0)$ we necessarily have
$\varphi_1^{-1}(W_1) \cap \varphi_2^{-1}(W_2) \neq \emptyset$.

3.2 Differentiable Functions on M

From now on we will assume that M is a C^∞ manifold, that is, all its maps $\varphi_i \circ \varphi_j^{-1}$ are infinitely differentiable. Although mathematically this is a restriction, it is not in physical applications. In these, M is generally the space of possible states of the system and therefore its points cannot be determined with absolute certainty, as every measurement involves some error. This indicates that through measurements we could never know the degree of differentiability of M. For convenience, we will assume it is C^∞.

A function on M is a map $f : M \to \mathbb{R}$, that is, a map that assigns a real number to each point of M. The information encoded in the atlas on M allows us to say how smooth f is.

Definition: We will say that f is **p-times continuously differentiable** at the point $q \in M$, $f \in C_q^p$ if given (U_i, φ_i) with $q \in U_i$, $f \circ \varphi_i^{-1} : \varphi_i(U_i) \subset \mathbb{R}^n \to \mathbb{R}$ is p-times continuously differentiable at $\varphi_i(q)$.
Note that this property is independent of the chart used [as long as we consider only charts from the compatible class of atlases]. We will say that $f \in C^p(M)$ if $f \in C_q^p \ \forall q \in M$. [See figure 3.5.]

In practice, one defines a particular function $f \in C^p(M)$ by introducing functions $f_i : \varphi_i(U_i) \subset \mathbb{R}^n \to \mathbb{R}$ (that is, $f_i(x^j)$ where x^j are the Cartesian coordinates

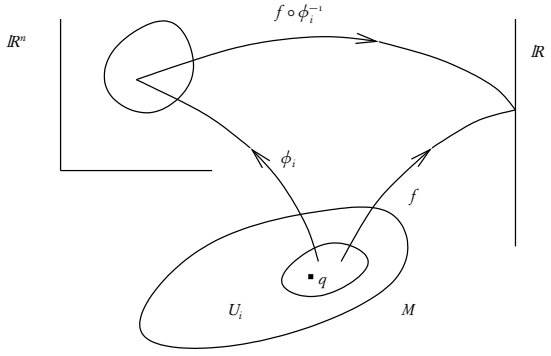

Figure 3.5: Composition of the map of a chart with a function.

in $\varphi_i(U_i) \subset \mathbb{R}^n$) that are C^p in $\varphi_i(U_i)$ and such that $f_i = f_j \circ \varphi_j \circ \varphi_i^{-1}$ in $\varphi_i(U_i \cap U_j)$. This guarantees that the set of f_i determines a unique function $f \in C^p(M)$. The set of $f_i\,(= f \circ \varphi_i^{-1})$ forms a **representation of** f in the atlas $\{(U_i, \varphi_i)\}$. [See figure 3.6.]

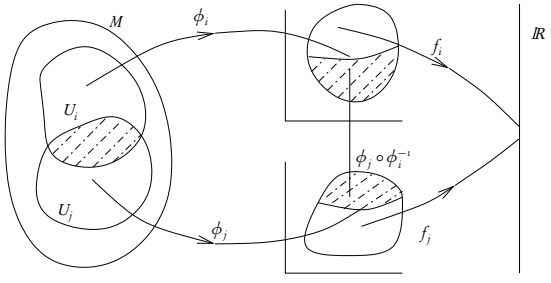

Figure 3.6: The relationship between the f_i.

Exercise: The circle, S^1, can be thought of as the interval $[0,1]$ with its ends identified. What are the functions in $C^2(S^1)$?

Using the previous construction, one can also define maps from M to $\mathbb{R}^m$ that are p-times differentiable. Now we perform the inverse construction, that is, we will define the differentiability of a map from $\mathbb{R}^m$ to M. We will do the case $\mathbb{R} \to M$, in which case the map thus obtained is called a curve. The general case is obvious.

3.3 Curves in M

Definition: **A curve in** M is a map between an interval $I \subset \mathbb{R}$ and M, $\gamma : I \to M$.

Note that the curve is the map and not its graph in M, that is, the set $\gamma[I]$. Thus, it is possible to have two different curves with the same graph. This is not a mathematical whim but a physical necessity: it is not the same for a car to travel the road Córdoba–Carlos Paz at 10 km/h as at 100 km/h, or to travel it in the opposite direction.

Definition: We will say that $\gamma \in C^p_{t_o}$ if given a chart (U_i, φ_i) such that $\gamma(t_o) \in U_i$ the map $\varphi_i \circ \gamma(t) : I_{t_o} \subset I \to \mathbb{R}^n$ is p-times continuously differentiable at t_o. [See figure 3.7.]

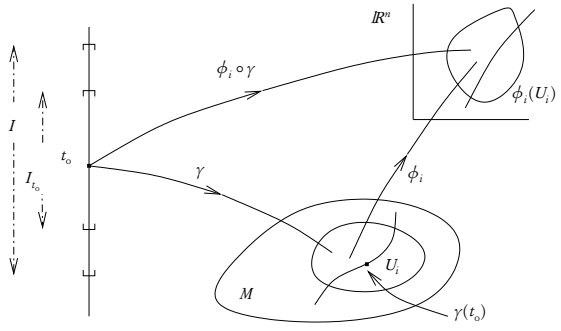

Figure 3.7: Differentiability of curves in M.

Exercise: Prove that the previous definition does not depend on the chart used.

This time we have used the concept of differentiability between maps from $\mathbb{R}$ to $\mathbb{R}^n$.

Definition: A curve $\gamma(t) \in C^p(I)$ if $\gamma(t) \in C^p_t \ \forall\, t \in I$.

Exercise: How would you define the concept of differentiability of maps between two manifolds?

Of particular importance among these are the maps from M to itself $g : M \to M$ that are continuously differentiable and invertible. They are called **Diffeomorphisms**. From now on we will assume that all manifolds, curves, diffeomorphisms, and functions are smooth, that is, they are C^∞.

3.4 Vectors

To define vectors at points of M we will use the concept of directional derivative at points of $\mathbb{R}^n$, that is, we will exploit the fact that in $\mathbb{R}^n$ there is a one-to-one correspondence between vectors $(v^1,\dots,v^n)\big|_{x_0}$ and directional derivatives $v(f)\big|_{x_0} = v^i \frac{\partial}{\partial x^i} f\Big|_{x_0}$.

As we have defined differentiable functions on M we can define derivations, or directional derivatives, at its points and identify with them the tangent vectors.

Definition: A **tangent vector** v **at** $p \in M$ is a map

$$v : C^\infty(M) \to \mathbb{R}$$

satisfying: $\forall f, g \in C^\infty(M)$, $a, b \in \mathbb{R}$

i) Linearity; $v(af + bg)\big|_p = a\, v(f)\big|_p + b\, v(g)\big|_p$.

ii) Leibniz; $v(fg)\big|_p = f(p)v(g)\big|_p + g(p)v(f)\big|_p$.

Note that if $h \in C^\infty(M)$ is the constant function, $h(q) = c \quad \forall q \in M$, then $v(h) = 0$. [$i) \Longrightarrow v(h^2) = v(ch) = c\, v(h)$ while $ii) \Longrightarrow v(h^2) = 2h(p)v(h) = 2c\, v(h)$]. These properties also show that $v(f)$ depends only on the behavior of f at p.

Exercise: Prove this last statement.

Let T_p be the set of all vectors at p. This set has the structure of a vector space and is called the **tangent space at the point** p. Indeed, we can define the sum of two vectors v_1, v_2 as the vector, that is the map, satisfying *i*) and *ii*), $(v_1 + v_2)(f) = v_1(f) + v_2(f)$ and the product of the vector v by the number a as the map $(av)(f) = a\, v(f)$.

As in $\mathbb{R}^n$, the dimension of the vector space T_p, (that is, the maximum number of linearly independent vectors), is n.

Theorem 3.1 *dim* $T_p = dim M$.

Proof: This will consist of finding a basis for T_p. Let $dim\, M = n$ and (U, φ) such that $p \in U$ and $f \in C^\infty(M)$ any. For $i = 1,\dots,n$ we define the vectors $x_i : C^\infty(M) \to \mathbb{R}$ given by,

$$x_i(f) := \frac{\partial}{\partial x^i}(f \circ \varphi^{-1})\Big|_{\varphi(p)} . \tag{3.1}$$

Note that these maps satisfy *i*) and *ii*) and therefore the x_i are really vectors. Note also that the right-hand side of 3.1 is well defined since we have the usual partial derivatives of maps between $\mathbb{R}^n$ and $\mathbb{R}$. These x_i depend on the chart (U, φ) but this

does not matter in the proof since T_p does not depend on any chart. These vectors are linearly independent, that is if $x = \sum_{i=1}^{n} c^i \boldsymbol{x}_i = 0$ then $c^i = 0 \ \forall\, i = 1, \ldots, n$. This is easily seen by considering the functions (strictly defined only in U), $f^j := x^j \circ \varphi$, since $\boldsymbol{x}_i(f^j) = \delta_i^j$ and therefore $0 = \boldsymbol{x}(f^j) = c^j$. It only remains to show that any vector $\boldsymbol{v}$ can be expressed as a linear combination of the $\boldsymbol{x}_i$. For this we will use the following result whose proof we leave as an exercise.

Lemma 3.1 *Let $F : \mathbb{R}^n \to \mathbb{R}$, $F \in C^\infty(\mathbb{R}^n)$ then for each $x_0 \in \mathbb{R}^n$ there exist functions $H_i : \mathbb{R}^n \to \mathbb{R} \in C^\infty(\mathbb{R}^n)$ such that $\forall\, x \in \mathbb{R}^n$ it holds*

$$F(x) = F(x_0) + \sum_{i=1}^{n} (x^i - x_0^i) H_i(x) \ \textit{and} \tag{3.2}$$

also,

$$\left.\frac{\partial F}{\partial x^i}\right|_{x=x_0} = H_i(x_0). \tag{3.3}$$

We now continue the proof of the previous Theorem. Let $F = f \circ \varphi^{-1}$ and $x_0 = \varphi(p)$, then $\forall\, q \in U$ we have

$$f(q) = f(p) + \sum_{i=1}^{n} (x^i \circ \varphi(q) - x^i \circ \varphi(p)) H_i \circ \varphi(q) \tag{3.4}$$

Using *i*) and *ii*) we obtain,

$$\begin{aligned} \boldsymbol{v}(f) \quad &= \boldsymbol{v}(f(p)) + \textstyle\sum_{i=1}^{n} (x^i \circ \varphi(q) - x^i \circ \varphi(p))\big|_{q=p} \ \boldsymbol{v}(H_i \circ \varphi) \\ &\quad + \textstyle\sum_{i=1}^{n} (H_i \circ \varphi)\big|_p \ \boldsymbol{v}(x^i \circ \varphi - x^i \circ \varphi(p)) \\ &= \textstyle\sum_{i=1}^{n} (H_i \circ \varphi)\big|_p \ \boldsymbol{v}(x^i \circ \varphi) \\ &= \textstyle\sum_{i=1}^{n} v^i \ \boldsymbol{x}_i(f) \end{aligned} \tag{3.5}$$

where $v^i \equiv \boldsymbol{v}(x^i \circ \varphi)$, and therefore we have expressed $\boldsymbol{v}$ as a linear combination of the $\boldsymbol{x}_i$, thus concluding the proof ♠

The basis $\{\boldsymbol{x}_i\}$ is called a **coordinate basis** and the $\{v^i\}$, the **components of $\boldsymbol{v}$** in that basis.

Exercise: If $(\tilde{U}, \tilde{\varphi})$ is another chart such that $p \in \tilde{U}$, then it will define another coordinate basis $\{\tilde{\boldsymbol{x}}_i\}$. Show that

$$\boldsymbol{x}_j = \sum_{i=1}^{n} \frac{\partial \tilde{x}^i}{\partial x^j} \tilde{\boldsymbol{x}}_i$$

where $\tilde{x}^i$ is the i-th component of the map $\tilde{\varphi} \circ \varphi^{-1}$. Also show that the relationship between the components is $\tilde{v}^i = \sum_{i=1}^{n} \frac{\partial \tilde{x}^i}{\partial x^j} v^j$.

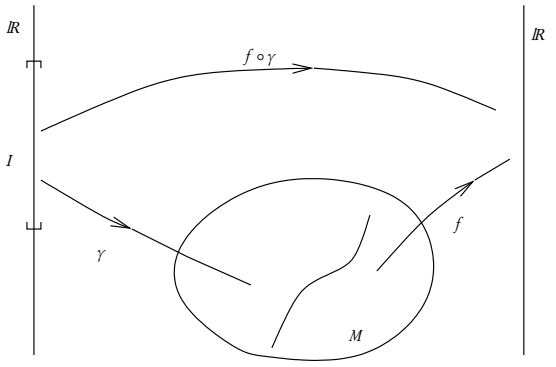

Figure 3.8: Definition of vector.

Example: Let $\gamma : I \to M$ be a curve in M. At each point $\gamma(t_o)$, $t_o \in I$, of M we can define a vector as follows, [See figure 3.8.]

$$\boldsymbol{t}(f) = \frac{d}{dt}\,(f \circ \gamma)|_{t=t_o}\,. \tag{3.6}$$

Its components in a coordinate basis are obtained through the functions

$$x^i(t) = x^i \circ \varphi \circ \gamma(t) \tag{3.7}$$

$$\begin{aligned}
\frac{d}{dt}(f \circ \gamma) &= \frac{d}{dt}(f \circ \varphi^{-1} \circ \varphi \circ \gamma) \\
&= \frac{d}{dt}(f \circ \varphi^{-1}(x^i(t))) \\
&= \sum_{i=1}^{n}(\frac{\partial}{\partial x^i}(f \circ \varphi^{-1}))\frac{dx^i}{dt} \\
&= \sum_{i=1}^{n}\frac{dx^i}{dt}\,\boldsymbol{x}_i(f)
\end{aligned} \tag{3.8}$$

3.5 Vector and Tensor Fields

If to each point q of M we assign a vector $\boldsymbol{v}|_q \in T_q$ we will have a **vector field**. This will be in $C^\infty(M)$ if given any $f \in C^\infty(M)$ the function $\boldsymbol{v}(f)$, which at each point p of M assigns the value $\boldsymbol{v}\,|_p\,(f)$, is also in $C^\infty(M)$. We will denote the set of C^∞ vector fields by TM and it is obviously a vector space of infinite dimension.

3.5.1 The Lie Bracket

Now consider the operation in the set TM of vector fields, $[\cdot,\cdot]: TM \times TM \to TM$. This operation is called the **Lie bracket** and given two vector fields (C^∞) it gives us a third:

$$[x,y](f) := x(y(f)) - y(x(f)). \tag{3.9}$$

Exercise:

1) Show that $[x,y]$ is indeed a vector field.

2) See that the **Jacobi identity** is satisfied:

$$[[x,y],z] + [[z,x],y] + [[y,z,x]] = 0 \tag{3.10}$$

3) Let x^i and x^j be two vector fields coming from a coordinate system, that is $x^i(f) = \frac{\partial f}{\partial x^i}$, etc. Show that $\left[x^i, x^j\right] = 0$.

4) Given the components of x and y in a coordinate basis, what are those of $[x,y]$?

With this operation TM acquires the character of an algebra, called **Lie Algebra.**

3.5.2 Diffeomorphisms and the Theory of Ordinary Differential Equations

Definition: A **one-parameter group of diffeomorphisms** g^t is a map $\mathbb{R} \times M \to M$ such that:

1) For each fixed t it is a diffeomorphism [1] $M \to M$

2) For any pair of real numbers, $t, s \in \mathbb{R}$ we have $g^t \circ g^s = g^{t+s}$ (in particular $g^0 = id$).

We can associate with g^t a vector field in the following way: For a fixed p, $g^t(p)$: $\mathbb{R} \to M$ is a curve that at $t = 0$ passes through p and therefore defines a tangent vector at p, $v|_p$. Repeating the process for every point in M we have a vector field in M. Note that due to the group property satisfied by g^t, the tangent vector to the curve $g^t(p)$ is also tangent to the curve $g^s(g^t(p))$ at $s = 0$.

We can ask the inverse question: Given a smooth vector field v in M, does there exist a one-parameter group of diffeomorphisms that defines it? The answer to this question, which consists of finding all the integrable curves $g^t(p)$ that pass through each $p \in M$, is the theory of ordinary differential equations, –which will be the subject of our study in the following chapters– since it consists of solving the equations $\frac{dx^i}{dt} = v^i(x^j)$ with initial conditions $x^i(0) = \varphi^i(p) \ \forall \ p \in M$. As we will see, the answer is affirmative but only locally, that is, we can only find g^t defined in $I(\subset \mathbb{R}) \times U(\subset M) \to M$.

[1] That is, a smooth map with a smooth inverse.

Example: In $\mathbb{R}^1$ let the vector have the coordinate component x^2, that is $v(x) = x^2 \frac{\partial}{\partial x}$. The ordinary differential equation associated with this vector is $\frac{dx}{dt} = x^2$, whose solution is

$$t - t_0 = \frac{-1}{x} + \frac{1}{x_0} \text{ ó } x(t) = \frac{-1}{t - \frac{1}{x_0}} \tag{3.11}$$

where we have taken $t_0 = 0$. That is, $g^t(x_0) = \frac{-1}{t - \frac{1}{x_0}}$. Note that for any t this map is not defined for all $\mathbb{R}$ and therefore is not a diffeomorphism. Also note that for any interval we take for its definition, the time interval of the solution's existence will be finite, either towards the future or the past.

Example: Let g^t be a linear diffeomorphism in $\mathbb{R}$, that is $g^t(x + \alpha y) = g^t(x) + \alpha\, g^t(y)$. Then it has the form $g^t(x) = f(t)x$. The group property implies $f(t) \cdot f(s) = f(t+s)$ or $f(t) = c\, e^{kt} = e^{kt}$, since $g^0 = id$. Therefore $g^t(x) = e^{kt}\, x$. The associated differential equation is: $x(t) = e^{kt}\, x_0 \implies \dot{x} = k\, e^{kt}\, x_0 = \boxed{k\, x = \dot{x}.}$

Exercise: Plot in a neighborhood of the origin in $\mathbb{R}^2$ the integral curves and therefore g^t of the following linear systems.

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & k \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{3.12}$$

a) $k > 1$ b) $k = 1$ c) $0 < k < 1$, d) $k = 0$, e) $k < 0$

3.5.3 Covector and Tensor Fields

Just as we introduced the notion of a vector field, we can also introduce the notion of a covector field, that is, a smooth map from M to T_p^*. This will act on vector fields giving as a result functions on M. In the following example, we see how the field **differential of** f is defined.

Example: Let $f \in C_p^\infty$. A vector at $p \in M$ is a derivation on functions in C_p^∞, $v(f) \in \mathbb{R}$. But given v_1 and $v_2 \in T_p$, $a \in \mathbb{R}$ and $f \in C_p^\infty$, $(v_1 + a v_2)(f) = v_1(f) + a v_2(f)$ and therefore each given f defines a linear functional $\boldsymbol{df}\,|_p \colon T_p \to \mathbb{R}$, called the differential of f, that is an element of T_p^*,

$$\boldsymbol{df}(v) := v(f), \quad \forall v \in T_p.$$

In this way, the differential of a function, $\boldsymbol{df}$, is a covector that when acting on a vector v gives us the number *the derivative of f at the point p in the direction of v*.

Let f be a smooth function on M, $a \in \mathbb{R}$ and consider the subset S_a of M such that $f(S_a) = a$. It can be seen that if $df \neq 0$ this will be a submanifold of M, that is, a surface embedded in M, of dimension $n - 1$. The condition $df|_p(v) = 0$ on

vectors of T_p with $p \in S_a$ means that these are actually tangent vectors to S_a, that is, elements of $T_p(S_a)$. On the contrary, if $df(v)|_p \neq 0$ then at that point v *pierces* S_a.

Example: The function $f(x,y,z) = x^2 + y^2 + z^2$ in $\mathbb{R}^3$.

$S_a = \{(x,y,z) \in \mathbb{R}^3 | f(x,y,z) = a^2, \ a > 0\}$ is the sphere of radius a, and as we have already seen, a manifold. Let (v^x, v^y, v^z) be a vector at the point $(x,y,z) \in \mathbb{R}^3$, then the condition $\boldsymbol{df}(v) = 2(xv^x + yv^y + zv^z) = 0$ implies that v is tangent to S. Indeed, we see that this is the condition that tells us that v is *perpendicular* to (x,y,z) when we are in the conventional Euclidean structure.

Given a coordinate system (chart) that covers a point $p \in M$, we have seen that we have a canonical basis of T_p associated with it given by the vectors,

$$\boldsymbol{x}_i(f) := \frac{\partial f \circ \phi^{-1}}{\partial x^i}|_{\phi(p)}.$$

What will be the associated cobasis? Note that the coordinate system also gives us a set of n privileged functions, that is, the components of the map ϕ that defines the chart, $\{x^j\}$, $j = 1..n$, $x^i(p) :=$ *value of the i-th coordinate assigned by* ϕ *to the point* p. Note that $x^i \circ \varphi^{-1}$ is then the identity map for the i-th coordinate. If we apply the basis vectors to these functions, we obtain,

$$\boldsymbol{x}_i(x^j) := \frac{\partial x^j \circ \varphi^{-1}}{\partial x^i}|_{\phi(p)} = \delta^j_i,$$

but then, since $\boldsymbol{dx}^j(\boldsymbol{x}_i) = \boldsymbol{x}_i(x^j) = \delta^j_i$, we see that the differentials $\boldsymbol{dx}^j$ are the cobasis of the coordinate basis. In particular, we have that the coordinate components of a vector v are given by:

$$v^j = \boldsymbol{dx}^j(v),$$

and the components of a covector $\boldsymbol{\omega} \in T^*_p$ by,

$$\omega_i = \boldsymbol{\omega}(\boldsymbol{x}_i).$$

Similarly, we define **tensor fields** as the multilinear maps that when acting on vector and covector fields give functions from the manifold to the reals and that at each point of the manifold only depend on the vectors and covectors defined at that point. This last clarification is necessary because otherwise, we would include among the tensors, for example, line integrals over vector fields.

3.5.4 The Metric

Let M be an n-dimensional manifold. We have previously defined on M the notions of curves, vector fields, and covector fields, etc., but not a notion of distance between its points, that is, a function $d : M \times M \to \mathbb{R}$ that takes any two points, p and q of M and gives us a number $d(p,q)$ satisfying,

1. $d(p,q) \geq 0$.

2. $d(p,q) = 0 \leftrightarrow p = q$.

3. $d(p,q) = d(q,p)$.

4. $d(p,q) \leq d(p,r) + d(r,q)$.

This, and in some cases a notion of pseudo-distance [where 1) and 2) are not satisfied], is fundamental if we want to have a mathematical structure that is useful for the description of physical phenomena. For example, Hooke's law, which tells us that the force applied to a spring is proportional to its elongation (a distance), clearly needs this entity. Next, we will introduce a notion of infinitesimal distance, that is, between two infinitesimally separated points, which corresponds to the Euclidean notion of distance and allows us to develop a notion of global distance, that is, between any two points of M.

The idea is then to have a concept of distance (or pseudo-distance) between two *infinitesimally close* points, that is, two points connected by an *infinitesimal displacement*, that is, connected by a vector. The notion we need is then that of the norm of a vector. Since a manifold is locally like $\mathbb{R}^n$, in the sense that the space of tangent vectors at a point p, T_pM is $\mathbb{R}^n$, it is reasonable to consider there the notion of Euclidean distance, that is, the distance between two points x_0 and $x_1 \in \mathbb{R}^n$ is the square root of the sum of the squares of the components (in some coordinate system) of the vector connecting these two points. The problem with this is that such a notion depends on the coordinate system being used and therefore there will be as many distances as coordinate systems covering the point p. This is just an indication that the structure we have so far does not contain a privileged or natural notion of distance. This must be introduced as an additional structure. One way to obtain infinitesimal distances independent of the coordinate system (that is, geometric) is by introducing at each point $p \in M$ a tensor of type $\binom{0}{2}$, symmetric $[\mathbf{g}(\mathbf{u},\mathbf{v}) = \mathbf{g}(\mathbf{v},\mathbf{u}) \ \forall \ \mathbf{u}, \ \mathbf{v} \in T_pM]$ and non-degenerate $[\mathbf{g}(\mathbf{u},\mathbf{v}) = \mathbf{0} \ \forall \mathbf{v} \ \in T_pM \ \Rightarrow \mathbf{u} = \mathbf{0}]$. If we also require that this tensor be positive definite $[\mathbf{g}(\mathbf{u},\mathbf{u}) \geq \mathbf{0} \ (= \ \Leftrightarrow \ \mathbf{u} = \mathbf{0})]$ it can be easily seen that this defines an inner product in T_pM (or pseudo-inner product if $\mathbf{g}(\mathbf{u},\mathbf{u}) = \mathbf{0}$ for some $u \neq 0 \in T_pM$). [2] If we make a smooth choice for this tensor at each point of M we will obtain a smooth tensor field called the **metric** of M. This extra structure, a tensor field with certain properties, is what allows us to build the mathematical foundations to then construct much of physics on it.

Let $\boldsymbol{g}$ be a metric on M, given any point p of M there exists a coordinate system in which its components are

$$g_{ij} = \delta_{ij}$$

and therefore gives rise to the Euclidean inner product, however, in general, this result cannot be extended to a neighborhood of the point and in general, its components will depend there on the coordinates. Note that this is what we wanted to do

[2] Later we will see that an inner product gives rise to a distance, correspondingly a pseudo-inner product gives rise to a pseudo-distance.

initially, but now by defining this norm via a vector we have given it an invariant character.

Restricting ourselves now to positive definite metrics, we define **the norm** of a vector $v \in T_p$ as $|v| = \sqrt{|g(v,v)|}$, that is, as the infinitesimal distance divided by ϵ between p and the point $\gamma(\epsilon)$ where $\gamma(t)$ is a curve such that $\gamma(0) = p$, $\frac{d\gamma(t)}{dt}|_{t=0} = v$. Similarly, we can define the length of a smooth curve $\gamma(t) : [0,1] \to M$ by the formula,

$$L(\lambda) = \int_0^1 \sqrt{\mathbf{g}(\mathbf{v},\mathbf{v})}\, dt, \tag{3.13}$$

where $\mathbf{v}(t) = \frac{d\gamma(t)}{dt}$. We see then that we define the length of a curve by measuring the infinitesimal lengths between nearby points on it and then integrating with respect to t.

Exercise: Prove that the length $L(\gamma)$ is independent of the chosen parameter.

We define the distance between two points $p, q \in M$ as,

$$d_g(p,q) = \inf_{\{\gamma(t)\,:\,\gamma(0)=p,\gamma(1)=q\}} |L(\gamma)| \tag{3.14}$$

That is, as the infimum of the length of all curves connecting p with q.

Exercise: Find an example of a manifold with two points such that the infimum in the previous definition is not a minimum. That is, where there is no curve connecting the two points with the minimum distance between them.

Exercise: a) The Euclidean metric in $\mathbb{R}^2$ is $(dx)^2 + (dy)^2$, where $\{dx, dy\}$ is the cobasis associated with $\{\partial x, \partial y\}$. What is the distance between two points in this case?

Exercise: b) What is the form of the Euclidean metric in $\mathbb{R}^3$ in spherical coordinates? And in cylindrical coordinates?

Exercise: c) The metric of the sphere is $(d\theta)^2 + \sin^2\theta\,(d\varphi)^2$. What is the distance in this case? For which points p, q are there multiple curves γ_i with $L(\gamma_i) = d(p,q)$?

Exercise: d) The metric $(dx)^2 + (dy)^2 + (dz)^2 - (dt)^2$ in $\mathbb{R}^4$ is the Minkowski metric of special relativity. What is the *distance* between the point with coordinates $(0,0,0,0)$ and $(1,0,0,1)$?

A metric gives us a privileged map between the space of tangent vectors at p, T_p, and its dual T_p^* for each p in M, that is, the map that assigns to each vector $v \in T_p$ the

covector $g(v,) \in T_p$. Since this is valid for each p, we thus obtain a map between vector and covector fields.

3.5.5 Diffeomorphisms and the Theory of Ordinary Differential Equations

Definition: A **one-parameter group of diffeomorphisms** g^t is a map $\mathbb{R} \times M \to M$ such that:

1) For each fixed t it is a diffeomorphism [3] $M \to M$

2) For any pair of real numbers, $t, s \in \mathbb{R}$ we have $g^t \circ g^s = g^{t+s}$ (in particular $g^0 = id$).

We can associate with g^t a vector field in the following way: For a fixed p, $g^t(p)$: $\mathbb{R} \to M$ is a curve that at $t = 0$ passes through p and therefore defines a tangent vector at p, $v\,|_p$. Repeating the process for every point in M we have a vector field in M. Note that due to the group property satisfied by g^t, the tangent vector to the curve $g^t(p)$ is also tangent to the curve $g^s(g^t(p))$ at $s = 0$.

We can ask the inverse question: Given a smooth vector field v in M, does there exist a one-parameter group of diffeomorphisms that defines it? The answer to this question, which consists of finding all the integrable curves $g^t(p)$ that pass through each $p \in M$, is the theory of ordinary differential equations, –which will be the subject of our study in the following chapters– since it consists of solving the equations $\frac{dx^i}{dt} = v^i(x^j)$ with initial conditions $x^i(0) = \varphi^i(p) \;\; \forall \; p \in M$. As we will see, the answer is affirmative but only locally, that is, we can only find g^t defined in $I(\subset \mathbb{R}) \times U(\subset M) \to M$.

Example: In $\mathbb{R}^1$ let the vector have the coordinate component x^2, that is $v(x) = x^2 \frac{\partial}{\partial x}$. The ordinary differential equation associated with this vector is $\frac{dx}{dt} = x^2$, whose solution is

$$t - t_0 = \frac{-1}{x} + \frac{1}{x_0} \text{ ó } x(t) = \frac{-1}{t - \frac{1}{x_0}} \tag{3.15}$$

where we have taken $t_0 = 0$. That is, $g^t(x_0) = \frac{-1}{t - \frac{1}{x_0}}$. Note that for any t this map is not defined for all $\mathbb{R}$ and therefore is not a diffeomorphism. Also note that for any interval we take for its definition, the time interval of the solution's existence will be finite, either towards the future or the past.

Example: Let g^t be a linear diffeomorphism in $\mathbb{R}$, that is $g^t(x + \alpha y) = g^t(x) + \alpha\, g^t(y)$. Then it has the form $g^t(x) = f(t)x$. The group property implies $f(t) \cdot$

[3] That is, a smooth map with a smooth inverse.

$f(s) = f(t+s)$ or $f(t) = c\,e^{kt} = e^{kt}$, since $g^0 = id$. Therefore $g^t(x) = e^{kt}\,x$. The associated differential equation is: $x(t) = e^{kt}\,x_0 \implies \dot{x} = k\,e^{kt}\,x_0 = \boxed{k\,x = \dot{x}.}$

Exercise: Plot in a neighborhood of the origin in $\mathbb{R}^2$ the integral curves and therefore g^t of the following linear systems.

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & k \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{3.16}$$

a) $k > 1$ b) $k = 1$ c) $0 < k < 1$, d) $k = 0$, e) $k < 0$

3.5.6 Covector and Tensor Fields

Just as we introduced the notion of a vector field, we can also introduce the notion of a covector field, that is, a smooth map from M to T_p^*. This will act on vector fields giving as a result functions on M. In the following example, we see how the field **differential of** f is defined.

Example: Let $f \in C_p^\infty$. A vector at $p \in M$ is a derivation on functions in C_p^∞, $v(f) \in \mathbb{R}$. But given v_1 and $v_2 \in T_p$, $a \in \mathbb{R}$ and $f \in C_p^\infty$, $(v_1 + a v_2)(f) = v_1(f) + a v_2(f)$ and therefore each given f defines a linear functional $\boldsymbol{df}\,|_p$: $T_p \to \mathbb{R}$, called the differential of f, that is an element of T_p^*,

$$\boldsymbol{df}(v) := v(f), \quad \forall v \in T_p.$$

In this way, the differential of a function, $\boldsymbol{df}$, is a covector that when acting on a vector v gives us the number *the derivative of f at the point p in the direction of v*.

Let f be a smooth function on M, $a \in \mathbb{R}$ and consider the subset S_a of M such that $f(S_a) = a$. It can be seen that if $df \neq 0$ this will be a submanifold of M, that is, a surface embedded in M, of dimension $n-1$. The condition $df|_p(v) = 0$ on vectors of T_p with $p \in S_a$ means that these are actually tangent vectors to S_a, that is, elements of $T_p(S_a)$. On the contrary, if $df(v)|_p \neq 0$ then at that point v *pierces* S_a.

Example: The function $f(x,y,z) = x^2 + y^2 + z^2$ in $\mathbb{R}^3$.

$S_a = \{(x,y,z) \in \mathbb{R}^3 | f(x,y,z) = a^2,\ a > 0\}$ is the sphere of radius a, and as we have already seen, a manifold. Let (v^x, v^y, v^z) be a vector at the point $(x,y,z) \in \mathbb{R}^3$, then the condition $\boldsymbol{df}(v) = 2(xv^x + yv^y + zv^z) = 0$ implies that v is tangent to S. Indeed, we see that this is the condition that tells us that v is *perpendicular* to (x,y,z) when we are in the conventional Euclidean structure.

Given a coordinate system (chart) that covers a point $p \in M$, we have seen that we have a canonical basis of T_p associated with it given by the vectors,

$$\boldsymbol{x}_i(f) := \frac{\partial f \circ \phi^{-1}}{\partial x^i}\Big|_{\phi(p)}.$$

What will be the associated cobasis? Note that the coordinate system also gives us a set of n privileged functions, that is, the components of the map ϕ that defines the chart, $\{x^j\}$, $j = 1..n$, $x^i(p) :=$ *value of the i-th coordinate assigned by ϕ to the point p*. Note that $x^i \circ \varphi^{-1}$ is then the identity map for the i-th coordinate. If we apply the basis vectors to these functions, we obtain,

$$\boldsymbol{x}_i(x^j) := \frac{\partial x^j \circ \varphi^{-1}}{\partial x^i}|_{\phi(p)} = \delta_i^j,$$

but then, since $\boldsymbol{dx}^j(\boldsymbol{x}_i) = \boldsymbol{x}_i(x^j) = \delta_i^j$, we see that the differentials $\boldsymbol{dx}^j$ are the cobasis of the coordinate basis. In particular, we have that the coordinate components of a vector $\boldsymbol{v}$ are given by:

$$v^j = \boldsymbol{dx}^j(v),$$

and the components of a covector $\boldsymbol{\omega} \in T_p^*$ by,

$$\omega_i = \boldsymbol{\omega}(\boldsymbol{x}_i).$$

Similarly, we define **tensor fields** as the multilinear maps that when acting on vector and covector fields give functions from the manifold to the reals and that at each point of the manifold only depend on the vectors and covectors defined at that point. This last clarification is necessary because otherwise, we would include among the tensors, for example, line integrals over vector fields.

3.5.7 The Metric

Let M be an n-dimensional manifold. We have previously defined on M the notions of curves, vector fields, and covector fields, etc., but not a notion of distance between its points, that is, a function $d : M \times M \to \mathbb{R}$ that takes any two points, p and q of M and gives us a number $d(p,q)$ satisfying,

1. $d(p,q) \geq 0$.
2. $d(p,q) = 0 \leftrightarrow p = q$.
3. $d(p,q) = d(q,p)$.
4. $d(p,q) \leq d(p,r) + d(r,q)$.

This, and in some cases a notion of pseudo-distance [where 1) and 2) are not satisfied], is fundamental if we want to have a mathematical structure that is useful for the description of physical phenomena. For example, Hooke's law, which tells us that the force applied to a spring is proportional to its elongation (a distance), clearly needs this entity. Next, we will introduce a notion of infinitesimal distance, that is, between two infinitesimally separated points, which corresponds to the Euclidean notion of distance and allows us to develop a notion of global distance, that is, between any two points of M.

The idea is then to have a concept of distance (or pseudo-distance) between two *infinitesimally close* points, that is, two points connected by an *infinitesimal displacement*, that is, connected by a vector. The notion we need is then that of the norm of a vector. Since a manifold is locally like $\mathbb{R}^n$, in the sense that the space of tangent vectors at a point p, T_pM is $\mathbb{R}^n$, it is reasonable to consider there the notion of Euclidean distance, that is, the distance between two points x_0 and $x_1 \in \mathbb{R}^n$ is the square root of the sum of the squares of the components (in some coordinate system) of the vector connecting these two points. The problem with this is that such a notion depends on the coordinate system being used and therefore there will be as many distances as coordinate systems covering the point p. This is just an indication that the structure we have so far does not contain a privileged or natural notion of distance. This must be introduced as an additional structure. One way to obtain infinitesimal distances independent of the coordinate system (that is, geometric) is by introducing at each point $p \in M$ a tensor of type $\binom{0}{2}$, symmetric $[\mathbf{g}(\mathbf{u},\mathbf{v}) = \mathbf{g}(\mathbf{v},\mathbf{u}) \ \forall \ \mathbf{u}, \ \mathbf{v} \in T_pM]$ and non-degenerate $[\mathbf{g}(\mathbf{u},\mathbf{v}) = \mathbf{0} \ \forall \mathbf{v} \ \in T_pM \ \Rightarrow \mathbf{u} = \mathbf{0}]$. If we also require that this tensor be positive definite $[\mathbf{g}(\mathbf{u},\mathbf{u}) \geq \mathbf{0} \ (= \ \Leftrightarrow \ \mathbf{u} = \mathbf{0})]$ it can be easily seen that this defines an inner product in T_pM (or pseudo-inner product if $\mathbf{g}(\mathbf{u},\mathbf{u}) = \mathbf{0}$ for some $u \neq 0 \in T_pM$). [4] If we make a smooth choice for this tensor at each point of M we will obtain a smooth tensor field called the **metric** of M. This extra structure, a tensor field with certain properties, is what allows us to build the mathematical foundations to then construct much of physics on it.

Let $\boldsymbol{g}$ be a metric on M, given any point p of M there exists a coordinate system in which its components are

$$g_{ij} = \delta_{ij}$$

and therefore gives rise to the Euclidean inner product, however, in general, this result cannot be extended to a neighborhood of the point and in general, its components will depend there on the coordinates. Note that this is what we wanted to do initially, but now by defining this norm via a vector we have given it an invariant character.

Restricting ourselves now to positive definite metrics, we define **the norm** of a vector $v \in T_p$ as $|v| = \sqrt{|g(v,v)|}$, that is, as the infinitesimal distance divided by ϵ between p and the point $\gamma(\epsilon)$ where $\gamma(t)$ is a curve such that $\gamma(0) = p$, $\frac{d\gamma(t)}{dt}|_{t=0} = v$. Similarly, we can define the length of a smooth curve $\gamma(t) : [0,1] \to M$ by the formula,

$$L(\lambda) = \int_0^1 \sqrt{\mathbf{g}(\mathbf{v},\mathbf{v})}\, dt, \tag{3.17}$$

where $\mathbf{v}(t) = \frac{d\gamma(t)}{dt}$. We see then that we define the length of a curve by measuring the infinitesimal lengths between nearby points on it and then integrating with respect to t.

[4] Later we will see that an inner product gives rise to a distance, correspondingly a pseudo-inner product gives rise to a pseudo-distance.

Exercise: Prove that the length $L(\gamma)$ is independent of the chosen parameter.

We define the distance between two points $p, q \in M$ as,

$$d_g(p,q) = \inf_{\{\gamma(t)\,:\,\gamma(0)=p,\gamma(1)=q\}} |L(\gamma)| \tag{3.18}$$

That is, as the infimum of the length of all curves connecting p with q.

Exercise: Find an example of a manifold with two points such that the infimum in the previous definition is not a minimum. That is, where there is no curve connecting the two points with the minimum distance between them.

Exercise: a) The Euclidean metric in $\mathbb{R}^2$ is $(dx)^2+(dy)^2$, where $\{dx, dy\}$ is the cobasis associated with $\{\partial x, \partial y\}$. What is the distance between two points in this case?

Exercise: b) What is the form of the Euclidean metric in $\mathbb{R}^3$ in spherical coordinates? And in cylindrical coordinates?

Exercise: c) The metric of the sphere is $(d\theta)^2 + \sin^2\theta\,(d\varphi)^2$. What is the distance in this case? For which points p, q are there multiple curves γ_i with $L(\gamma_i) = d(p,q)$?

Exercise: d) The metric $(dx)^2+(dy)^2+(dz)^2-(dt)^2$ in $\mathbb{R}^4$ is the Minkowski metric of special relativity. What is the *distance* between the point with coordinates $(0,0,0,0)$ and $(1,0,0,1)$?

A metric gives us a privileged map between the space of tangent vectors at p, T_p, and its dual T_p^* for each p in M, that is, the map that assigns to each vector $v \in T_p$ the covector $g(v, \,) \in T_p$. Since this is valid for each p, we thus obtain a map between vector and covector fields.

Since g is non-degenerate, this map is invertible, that is, there exists a symmetric tensor of type $\binom{2}{0}$, g^{-1}, such that

$$g(g^{-1}(\theta, \,), \,) = \theta \tag{3.19}$$

for any co-vector field θ. This indicates that when we have a manifold with a metric, it becomes irrelevant to distinguish between vectors and co-vectors or, for example, between tensors of type $\binom{0}{2}$, $\binom{2}{0}$, or $\binom{1}{1}$.

3.5.8 Abstract Index Notation

When working with tensorial objects, the notation used so far is not the most convenient because it is difficult to remember the type of each tensor, in which slot it "eats" other objects, etc. One solution is to introduce a coordinate system and work

with the components of the tensors, where having indices makes it easy to know what objects they are or to introduce general bases. In this way, for example, we represent the vector $\boldsymbol{l} = l^i \dfrac{\partial}{\partial x^i}$ by its components $\{l^i\}$. A convenience of this notation is that "eating" becomes "contracting," since, for example, we represent the vector $\boldsymbol{l}$ "eating" a function f by the contraction of the coordinate components of the vector and the differential of f:

$$\boldsymbol{l}(f) = \sum_{i=1}^{n} l^i \frac{\partial f}{\partial x^i}.$$

But a serious drawback of this representation is that it initially depends on the coordinate system and therefore all the expressions we construct with it have the potential danger of depending on such a system.

We will remedy this by introducing **abstract indices** (which will be Latin letters) that indicate where the coordinate indices would go but nothing more, that is, they do not depend on the coordinate system and do not even take numerical values, that is, l^a does not mean the n-tuple $(l^1, l^2, \ldots, l^n)$ as if they were indices. In this way, l^a will denote the vector l, θ_a the co-vector θ, and g_{ab} the metric g. A contraction such as $g(v, \ )$ will be denoted $g_{ab}v^a$ and we will denote this co-vector by v_b, that is, the action of g_{ab} is to lower the index of v^a and give the co-vector $v_b \equiv v^a g_{ab}$. Similarly, we will denote $\boldsymbol{g}^{-1}$ (the inverse of $\boldsymbol{g}$) as g^{ab}, that is, $\boldsymbol{g}$ with the indices raised.

The symmetry of $\boldsymbol{g}$ is then equivalent to $g_{ab} = g_{ba}$.

Exercise: How would you denote an antisymmetric tensor of type $\binom{0}{2}$?

Using repeated indices for contraction, we see that $l(f)$ can be denoted by $l^a \nabla_a f$ where $\nabla_a f$ denotes the differential co-vector of f, while the vector $\nabla^a f := g^{ab}\nabla_b f$ is called the **gradient** of f and we see that it depends not only on f but also on $\boldsymbol{g}$.

3.6 Covariant Derivative

We have seen that in M there is the notion of the derivative of a scalar field f, which is the differential co-vector of f that we denote $\nabla_a f$. Is there the notion of the derivative of a tensor field? For example, is there an extension of the operator ∇_a to vectors such that if l^a is a differentiable vector then $\nabla_a l^b$ is a tensor of type $\binom{1}{1}$? To fix ideas, let us define this extension of the differential ∇_a, called the covariant derivative, by requiring it to satisfy the following properties: [5]

i) Linearity: If $A^{a_1\cdots a_k}_{b_1\ldots b_l}, B^{a_1\cdots a_k}_{b_1\ldots b_l}$ are tensors of type $\binom{k}{l}$ and $\alpha \in \mathbb{R}$ then

$$\nabla_c \left(\alpha A^{a_1\cdots a_k}_{b_1\ldots b_l} + B^{a_1\cdots a_k}_{b_1\ldots b_l}\right) = \alpha \nabla_c A^{a_1\cdots a_k}_{b_1\ldots b_l} + \nabla_c B^{a_1\cdots a_k}_{b_1\ldots b_l} \tag{3.20}$$

[5] Note that they are an extension of those required to define derivations.

ii) Leibnitz:

$$\nabla_e \left(A^{a_1 \cdots a_k}_{b_1 \ldots b_l} B^{c_1 \ldots c_m}_{d_1 \ldots d_n} \right) = A^{a_1 \cdots a_k}_{b_1 \ldots b_l} \left(\nabla_e B^{c_1 \ldots c_m}_{d_1 \ldots d_n} \right) + \left(\nabla_e A^{a_1 \cdots a_k}_{b_1 \ldots b_l} \right) B^{c_1 \ldots c_m}_{d_1 \ldots d_n} \tag{3.21}$$

iii) Commutativity with contractions:

$$\nabla_e \left(\delta^c{}_d A^{a_1 \ldots, d, \ldots, a_k}_{b_1 \ldots, c, \ldots, b_k} \right) = \delta^c{}_d \nabla_e A^{a_1 \ldots, d, \ldots, a_k}_{b_1 \ldots, c, \ldots, b_k}, \tag{3.22}$$

where $\delta^c{}_d$ is the identity tensor. That is, if we first contract some indices of a tensor and then take its derivative, we obtain the same tensor as if we first take the derivative and then contract.

iv) Consistency with the differential: If l^a is a vector field and f a scalar field, then

$$l^a \nabla_a f = l(f) \tag{3.23}$$

v) Zero torsion: If f is a scalar field then

$$\nabla_a \nabla_b f = \nabla_b \nabla_a f \tag{3.24}$$

Example: Let $\{x^i\}$ be a global coordinate system in $\mathbb{R}^n$ and let ∇_c be the operator that, when acting on $A^{a_1 \cdots a_k}_{b_1 \ldots b_l}$, generates the tensor field that in these coordinates has components

$$\partial_j A^{i_1 \ldots i_k}_{j_1 \ldots j_l} \tag{3.25}$$

By definition, it is a tensor and clearly satisfies all the conditions of the definition, since it satisfies them in this coordinate system, so it is a covariant derivative. If we take another coordinate system, we will obtain another covariant derivative, generally different from the previous one. For example, let us act ∇_c on the vector l^a, then

$$(\nabla_c l^a)^i_j = \partial_j l^i. \tag{3.26}$$

In another coordinate system $\{\tilde{x}^i\}$ this tensor has components

$$(\nabla_c l^a)^l_k = \sum_{i,j=1}^{n} \frac{\partial \tilde{x}^l}{\partial x^i} \frac{\partial x^j}{\partial \tilde{x}^k} \frac{\partial l^i}{\partial x^j} \tag{3.27}$$

which are not, in general, the components of the covariant derivative $\tilde{\nabla}_c$ that these new coordinates define, indeed

$$\begin{aligned} \left(\tilde{\nabla}_c l^a \right)^l_k &\equiv \frac{\partial \tilde{l}^l}{\partial \tilde{x}^k} = \sum_{j=1}^n \left(\frac{\partial x^j}{\partial \tilde{x}^k} \right) \frac{\partial}{\partial x^j} \sum_{i=1}^n \left(\frac{\partial \tilde{x}^l}{\partial x^i} l^i \right) = \\ &= \sum_{i,j=1}^n \left(\frac{\partial \tilde{x}^l}{\partial x^i} \right) \left(\frac{\partial x^j}{\partial \tilde{x}^k} \right) \frac{\partial l^i}{\partial x^j} + \sum_{i,j=1}^n \frac{\partial x^j}{\partial \tilde{x}^k} \left(\frac{\partial^2 \tilde{x}^l}{\partial x^j \partial x^i} \right) l^i \\ &= (\nabla_c l^a)^i_j + \sum_{i,j=1}^n \frac{\partial x^j}{\partial \tilde{x}^k} \left(\frac{\partial^2 \tilde{x}^l}{\partial x^j \partial x^i} \right) l^i, \end{aligned} \tag{3.28}$$

which clearly shows that they are two different tensors and that their difference is a **tensor** that depends linearly and **not differentially** on l^a. Is this true in general? That is, given two connections, ∇_c and $\bar{\nabla}_c$, is their difference a tensor (and not a differential operator)? We will see that this is true.

Theorem 3.2 *The difference between two connections is a tensor.*

Proof: Note that by properties $iii)$ and $iv)$ of the definition, if we know how ∇_c acts on co-vectors, we know how it acts on vectors and thus by $i)$ and $ii)$ on any tensor. Indeed, if we know $\nabla_c w_a$ for any w_a, then $\nabla_c l^a$ is the tensor of type $\binom{1}{1}$ such that when contracted with an arbitrary w_a gives us the co-vector

$$(\nabla_c l^a) w_a = \nabla_c (w_a l^a) - l^a (\nabla_c w_a), \tag{3.29}$$

which we know since by $iv)$ we also know how ∇_c acts on scalars. Therefore, it is sufficient to see that

$$\left(\bar{\nabla}_c - \nabla_c\right) w_a = C^b{}_{ca} w_b \tag{3.30}$$

for some tensor $C^b{}_{ca}$. First, let us prove that given any $p \in M$, $\left(\bar{\nabla}_c - \nabla_c\right) w_a|_p$ depends only on $w_a|_p$ and not on its derivative. Let w'_a be any other co-vector such that at p they coincide, that is, $(w_a - w'_a)|_p = 0$. Then given a smooth co-basis $\{\mu^i_a\}$ in a neighborhood of p, we will have that $w_a - w'_a = \sum_i f_i \mu^i_a$ with f_i smooth functions that vanish at p. At p we then have

$$\begin{array}{rcl} \bar{\nabla}_c (w_a - w'_a) - \nabla_c (w_a - w'_a) & = & \sum_i \bar{\nabla}_c \left(f_i \mu^i_a\right) - \sum_i \nabla_c \left(f_i \mu^i_a\right) \\ & = & \sum_i \mu^i_a \left(\bar{\nabla}_c f_i - \nabla_c f_i\right) = 0 \end{array} \tag{3.31}$$

since by *iv)* $\bar{\nabla}_c$ and ∇_c must act in the same way on scalars –and in particular on the f_i–. This shows that

$$\left(\bar{\nabla}_c - \nabla_c\right) w'_a = \left(\bar{\nabla}_c - \nabla_c\right) w_a \tag{3.32}$$

and therefore that $\left(\bar{\nabla}_c - \nabla_c\right) w_a$ depends only on $w_a|_p$ and obviously in a linear way. But then $\left(\bar{\nabla}_c - \nabla_c\right)$ must be a tensor of type $\binom{1}{2}$ that is waiting to "eat" a co-vector to give us the tensor of type $\binom{0}{2}$, $\left(\bar{\nabla}_c - \nabla_c\right) w_a$. That is, $\left(\bar{\nabla}_c - \nabla_c\right) w_a = C^b{}_{ca} w_b$, which proves the theorem.

Note that condition $v)$ tells us that $\nabla_a \nabla_b f = \nabla_b \nabla_a f$, taking $w_a = \nabla_a f$ we get

$$\begin{array}{rcl} \bar{\nabla}_a \bar{\nabla}_b f & = & \bar{\nabla}_a \nabla_b f \\ & = & \bar{\nabla}_a w_b \\ & = & \nabla_a w_b + C^c_{ab} \nabla_c f \\ & = & \nabla_a \nabla_b f + C^c_{ab} \nabla_c f. \end{array} \tag{3.33}$$

Since $\nabla_c f|_p$ can be any co-vector, we see that the condition of no torsion implies that $C^c{}_{ab}$ is symmetric in the lower indices, $C^c{}_{ab} = C^c{}_{ba}$.

Exercise: How does $\left(\bar{\nabla}_c - \nabla_c\right)$ act on vectors?

Exercise: Express the Lie bracket in terms of any connection and then explicitly prove that it does not depend on the connection used.

Exercise: Let $A_{b,\ldots,z}$ be a totally antisymmetric tensor. Show that $\nabla_{[a}A_{b,\ldots,z]}$, that is, the total antisymmetrization of $\nabla_a A_{b,\ldots,z}$, does not depend on the covariant derivative used.

The difference between any connection ∇_c and one coming from a coordinate system $\{x^i\}$ is a tensor called the **Christoffel symbol** of ∇_c with respect to the coordinates $\{x^i\}$, Γ^b_{ca},

$$\nabla_c w_a = \partial_c w_a + \Gamma^b_{ca} w_b. \tag{3.34}$$

The knowledge of this tensor is very useful in practice, as it allows us to express ∇_c in terms of the corresponding coordinate connection, ∂_c.

As we have seen, in a manifold M there are infinite ways to *take the derivative of a tensor*. Is there any natural or privileged one? The answer is no, unless we add more structure to M. Intuitively, the reason for this is that in M we do not know how to compare $l^a|_p$ with $l^a|_q$ if p and q are two different points. [6]

Is the presence of a metric in M sufficient to make this comparison? The answer is yes!

Theorem 3.3 *Let g_{ab} be a (smooth) metric on M, then there exists a unique covariant derivative ∇_c such that $\nabla_c g_{ab} = 0$.*

Proof: Let $\bar{\nabla}_c$ be any connection and let ∇_c be such that $\nabla_c g_{ab} = 0$, it is sufficient to show that this condition uniquely determines the difference tensor, $C^d{}_{ca}$. But,

$$0 = \nabla_a g_{bc} = \bar{\nabla}_a g_{bc} - C^d{}_{ab} g_{dc} - C^d{}_{ac} g_{bd} \tag{3.35}$$

that is

$$C_{cab} + C_{bac} = \bar{\nabla}_a g_{bc} \tag{3.36}$$

but also

$$\begin{array}{lll} a \leftrightarrow b & C_{cba} + C_{abc} & = \bar{\nabla}_b g_{ac} \\ c \leftrightarrow b & C_{bca} + C_{acb} & = \bar{\nabla}_c g_{ab}. \end{array} \tag{3.37}$$

Adding the last two, subtracting the first and using the symmetry of the last two indices we get

$$2C_{abc} = \bar{\nabla}_b g_{ac} + \bar{\nabla}_c g_{ab} - \bar{\nabla}_a g_{bc} \tag{3.38}$$

[6] Note that one way to compare infinitesimally close vectors, given a vector field m^a, is with the Lie bracket of m^a with l^a, $[m,l]^a$. This is not appropriate since $[m,l]^a|_p$ depends on the derivative of m^a at p.

or

$$C^a{}_{bc} = \frac{1}{2} g^{ad} \{ \bar{\nabla}_b g_{dc} + \bar{\nabla}_c g_{db} - \bar{\nabla}_d g_{bc} \} \tag{3.39}$$

Note that the existence of a metric is not equivalent to the existence of a connection. There are connections ∇_c for which there is no metric g_{ab} such that $\nabla_a g_{bc} = 0$, that is, there are tensors $C^a{}_{bc}$ for which there is no g_{ab} that satisfies (3.39).

Exercise: If $\bar{\nabla}_c$ is a derivative corresponding to a coordinate system $\bar{\nabla}_c = \partial_c$ the corresponding Christoffel symbol is

$$\Gamma^a{}_{bc} = \frac{1}{2} g^{ad} \{ \partial_b g_{dc} + \partial_c g_{db} - \partial_d g_{bc} \}. \tag{3.40}$$

Show that its components in that coordinate system are given by,

$$\Gamma^i{}_{jk} = \frac{1}{2} g^{il} \{ \partial_j g_{lk} + \partial_k g_{lj} - \partial_l g_{jk} \}, \tag{3.41}$$

where g_{ij} are the components of the metric in the same coordinate system.

Exercise: The Euclidean metric of $\mathbb{R}^3$ in spherical coordinates is

$$ds^2 = (dr)^2 + r^2((d\theta)^2 + \sin^2\theta (d\phi)^2).$$

Calculate the Laplacian $\Delta = g^{ab} \nabla_a \nabla_b$ in these coordinates.

***The Riemann Tensor**
Given a covariant derivative on a manifold, is it possible to define tensor fields that depend only on it and therefore give us information about it?
The answer is yes! The following tensor is called the **Riemann or Curvature Tensor** and depends only on the connection:

$$2\nabla_{[a} \nabla_{b]} l^c := (\nabla_a \nabla_b - \nabla_b \nabla_a) l^c := R^c{}_{dab} l^d \quad \forall l^c \in TM.$$

Exercise: Show that the definition above makes sense, that is, that the left-hand side, evaluated at any point p in M, depends only on $l^c|_p$ and therefore we can write the right-hand side for some tensor $R^c{}_{dab}$.

Exercise: Let $\bar{\nabla}_a$ be another covariant derivative. Calculate the difference between the respective Riemann tensors in terms of the tensor that appears as the difference of the two connections.

Bibliography notes: I recommend the book by Wald [6], especially for its modern notation, see also [21]. The language of intuition is geometry, it is the tool that allows us to visualize problems, touch them, turn them around to our liking and then reduce them to algebra. Understanding geometry is the most efficient way to understand physics since it is only fully understood when translated into a geometric language. Do not abuse it, it is a very vast area and it is easy to get lost, learn the basics well and then only what is relevant to you.

Chapter 4

Ordinary Differential Equations

4.1 Introduction

In this chapter we will begin the study of systems of ordinary differential equations –ODE from now on–. These systems constantly appear in physics and together with the systems of partial differential equations, –PDE from now on– that we will study in the second part of this course, form the mathematical skeleton of the exact sciences. This is because most physical phenomena can be described in terms of these equations. If the free parameter –or independent variable– has the interpretation of being a time, then we are dealing with **evolution equations**. These allow us, for each initial state of the system, to know its temporal evolution, that is, they give us the ability to make predictions from initial conditions, which is the characteristic aspect of any physical theory. [1]

There are other physical phenomena described by ODE that do not have an evolutionary character. In general, in these cases, we are interested in equilibrium states of the system. For example, if we want to see what shape the clothesline at home describes. These cases will not be included in the theory of ODE that we will develop below, but their simplest cases will be treated as special (one-dimensional) cases of elliptic PDE.

Although quantum physics, or more precisely quantum field theory, shows us that a description of physical phenomena at that microscopic scale is not possible through the systems of differential equations mentioned above, we usually manage to find partial aspects of these phenomena, or certain approximations where a description in terms of these systems is possible. This is not due to a whim or stubbornness, but to the fact that our knowledge of ODE, and to a lesser extent of PDE, is quite deep, which allows us to handle them with relative ease and comfort. On the other hand, and it is important to emphasize this, our knowledge of the type of equations that appear in non-linear quantum field theories is, in comparison, practically nil.

[1] In some cases, the independent variable is not time but some other parameter, but the equation can also be considered as describing evolution. For example, if we want to see how the density of a medium varies when the temperature increases.

4.2 The Case of a Single First-Order Ordinary Equation

4.2.1 First-Order Autonomous Equation

This has the form:

$$\dot{x} = \frac{dx}{dt} = f(x) \tag{4.1}$$

that is, an equation for a map $x(t) : I \subset \mathbb{R} \to U \subset \mathbb{R}$.
We will assume $f : U \to \mathbb{R}$ continuous. A solution of this equation will be a differentiable map $\varphi(t)$ defined for all I and with image in U. We will say that $\varphi(t)$ satisfies the **initial condition** x_0 at $t = t_0 \in I$ if

$$\varphi(t_0) = x_0. \tag{4.2}$$

As we will see below, under certain hypotheses, the initial condition that a solution satisfies determines it uniquely.

The equation 4.1 can be interpreted geometrically in the following way. At each point of the plane (x, t) we mark a "little line", that is, a field of directions such that the angle it forms with the axis $x =$cte has a tangent equal to $f(x)$.

Example: $\dot{x} = x^{\frac{1}{2}}$. [See figure (4.1).]

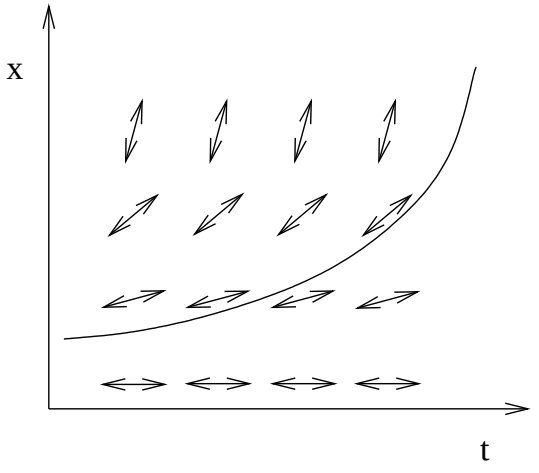

Figure 4.1: Geometric interpretation of the equation $f(x) = x^{1/2}$.

Since $\varphi(t)$ has a derivative $f(\varphi(t))$, at the point $(\varphi(t), t)$, $\varphi(t)$ will be tangent to the little line at that point. If we stand at some point (x_0, t_0) through which it passes, we can find the solution by following the little lines. Therefore, we see that given a point through which the solution passes, by following a little line, we will determine a unique solution. This, in particular, tells us that in this graph, most solutions do

not intersect. As we will see later, the uniqueness of the solution occurs if the little line at each point has a non-zero tangent. On the other hand, it is clear from the case in the previous figure, 4.1, that if we start with the point $(x_o = o, t_o)$, we would have two options: either draw the already drawn line (t^2, t) or simply the line (o, t).

Exercise: Extend this graphical intuition to the case $\dot{x} = f(x,t)$.

Example: In fact, the equation $\dot{x} = x^{1/2}$ has two solutions that pass through (o, o),

$$\varphi(t) = o \text{ y } \varphi = \frac{t^2}{4} \tag{4.3}$$

Indeed, suppose $x(t) \neq o$ then $\frac{dx}{x^{1/2}} = dt$ or $2(x^{1/2} - x_o^{1/2}) = t - t_o$, taking $x_o = t_o = o$

$$x = \left(\frac{t}{2}\right)^2 \tag{4.4}$$

The other is trivially a solution.

The uniqueness of the solutions is obtained, even in the case where $f(x)$ vanishes, if instead of asking that $f(x)$ be continuous, we ask a little more, and that is that, where it vanishes, say at x_o, it is fulfilled that

$$\lim_{\epsilon \to o} \int_{x_o + \varepsilon}^{x} \frac{dx}{f(x)} = \infty, \tag{4.5}$$

or alternatively that for $\epsilon > o$ sufficiently small, there exists $k > o$ such that,

$$|f(x)| < k \,|\, x - x_o \,|\ \text{ si } |\, x - x_o \,| < \epsilon. \tag{4.6}$$

That is, $f(x)$ goes to zero when $x \to x_o$ at most as quickly as a linear function. This condition is called the Lipschitz condition. From now on, we will assume that $f(x)$ is differentiable and therefore Lipschitz.

Exercise: See that this condition is weaker than asking that f be differentiable at x_o.
Solution: The function $f(x) = x \sin(\frac{1}{x})$ is Lipschitz at x_{+o} but not differentiable.

Thus, we arrive at our first theorem on ordinary equations.

Theorem 4.1 *Let $f(x) : U \to \mathbb{R}$ be continuous and Lipschitz at the points where it vanishes. Then*

i) For each $t_o \in \mathbb{R}$, $x_o \in U$ there exists an interval $I_{t_o} \in \mathbb{R}$ and a solution $\varphi(t)$ of the equation 4.1 defined $\forall\ t \in I_{t_o}$ and the initial condition $\varphi(t_o) = x_o$;

ii) Any two solutions φ_1, φ_2 of 4.1 satisfying the same initial condition coincide in some neighborhood of $t = t_o$

iii) the solution φ of 4.1 is such that

$$t - t_0 = \int_{x_0}^{\varphi(t)} \frac{dx}{f(x)} \quad si\ f(x) \neq 0 \tag{4.7}$$

Note: The solution whose existence is guaranteed by this theorem is valid in an open interval I_{t_0}, which the theorem does not determine. Therefore, uniqueness occurs only in the maximum interval that they have in common.

Proof:

If $f(x_0) = 0$ then $\varphi(t) = x_0$ is a solution.

If $f(x_0) \neq 0$ then by continuity there exists a neighborhood of x_0 where $f(x) \neq 0$ and therefore $1/f(x)$ is a continuous function, this implies that the integral in 4.29 is a differentiable function of its integration limits, which we will call $\psi(x) - \psi(x_0)$. Therefore, we have,

$$t - t_0 = \psi(x) - \psi(x_0) \tag{4.8}$$

That is, given x we obtain t and also this relationship is automatically fulfilled when $x = x_0$, $t = t_0$. We now want to solve for x from this relationship, that is, find $\phi(t)$. But

$$\left.\frac{d\psi}{dx}\right|_{x=\zeta} = \frac{1}{f(\zeta)} \neq 0\,, \tag{4.9}$$

Therefore, the inverse function theorem assures us that there exists an I_{t_0} and a *unique* $\varphi(t)$ differentiable, defined for all $t \in I_{t_0}$ and such that $\varphi(t_0) = x_0$ and $t - t_0 = \psi(\varphi(t)) - \psi(x_0)$. Taking its derivative, we obtain,

$$\frac{d\varphi}{dt} = \left.\frac{d\psi^{-1}}{dt}\right|_{x=\varphi(t)} = \left.\left[\frac{1}{f(x)}\right]^{-1}\right|_{x=\varphi(t)} = f(\varphi(t)) \tag{4.10}$$

which shows that it is a solution, necessarily unique of 4.1 with the appropriate initial condition.

It only remains to see that in the case where $f(x_0) = 0$ the solution is unique. For this, we will use the following Lemma.

Lemma 4.1 *Let f_1 and f_2 be real functions in $U \subset \mathbb{R}$ such that $f_1(x) < f_2(x)\ \forall\, x \in U$ and let φ_1 and φ_2 be the respective solutions to the differential equations that they generate, defined in an interval $I \in \mathbb{R}$ and such that $\varphi_1(t_1) = \varphi_2(t_1)$ for some $t_1 \in I$. Then*

$$\varphi_1(t) \leq \varphi_2(t) \quad \forall\ t > t_1 \in I \tag{4.11}$$

Proof: Trivial, the one who runs faster in each sector of the path wins! Let the set $S = \{t > t_1 \in I | \phi_1(t) > \phi_2(t)\}$. Let $T \geq t_1$, $T \in I$ be the greatest lower bound of S. At this moment $\varphi_1(T) = \varphi_2(T)$, but

$$\left.\frac{d\varphi_1}{dt}\right|_{t=T} < \left.\frac{d\varphi_2}{dt}\right|_{t=T} \tag{4.12}$$

Therefore, there exists $\varepsilon > 0$ such that for all $t \in [T, \varepsilon)$, $\varphi_1(t) < \varphi_2(t)$ which is a contradiction, unless it is the last value in I.

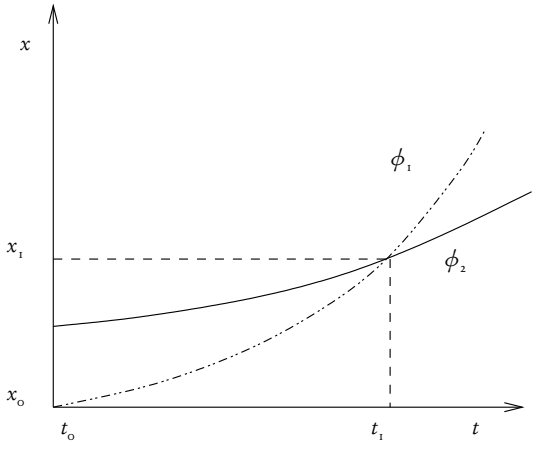

Figure 4.2: Proof of Lemma 4.1.

With the help of this Lemma, we will prove the uniqueness of the solution. Let $\varphi_1(t)$ be a solution with $\varphi_1(t_0) = x_0$ but different from the solution $\varphi(t) = x_0 \; \forall \; t$. Suppose then that there exists $t_1 > t_0$ such that $\varphi_1(t_1) = x_1 > x_0$, the other cases are treated similarly. [See figure 4.3.]

If we choose x_1 close enough to x_0, it is fulfilled that

$$f(x) < k \,|\, x - x_0 \,| \quad \forall \; x_0 \leq x \leq x_1 \tag{4.13}$$

and therefore that $\varphi_2(t) < \varphi_1(t)$, $\forall \; t_0 \leq t \leq t_1$, where $\varphi_2(t)$ is the solution of the equation $\dot{x} = k(x - x_0)$ with the initial condition $\varphi_2(t_1) = x_1$. But $\varphi_2(t) = x_0 + (x_1 - x_0)e^{k(t-t_1)}$ which tends to x_0 only when $t \to -\infty$. Therefore, $\varphi_1(t)$ cannot take the value x_0 for any finite $t < t_1$ and therefore we have a contradiction ♠

The solution whose existence and uniqueness we have just demonstrated can not only be thought of as a map between I_{t_0} and U but also as a map g_x^t between $U_{x_0} \times I_{t_0}$ and U (U_{x_0} neighborhood of x_0). The map g_x^t takes (x, t) to the point $\varphi(t)$ where φ is the solution of 4.1 with the initial condition $\varphi(0) = x$.

These maps, or local diffeomorphisms, are called **local phase flows**, (since in general they cannot be defined in all $I_{t_0} \times U$) and satisfy the following important group property; if $t, s \in I_0$ then $g_x^{s+t} = g^s \, o \, g_x^t$.

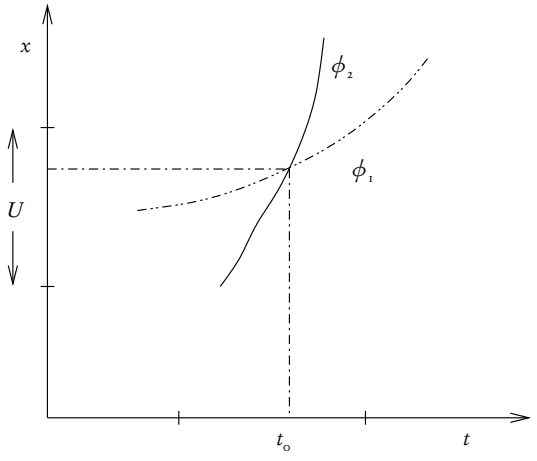

Figure 4.3: Proof of the uniqueness of the solution.

By the way the initial conditions appear in the previous Theorem, it is easy to see, using again the inverse function theorem, that if $f(x_o) \neq 0$ then g_x^t is also differentiable with respect to x.

Corolary 4.1 *if* $f(x_o) \neq 0$ *then* $g_x^t : U_{x_o} \times I_{t_o} \to U$ *is differentiable in both* t *and* x.

Note: The restriction $f(x_o) \neq 0$ is unnecessary as we will see later.

From the use of the inverse function theorem in Theorem 4.1, it also follows that if we consider the set of solutions distinguishing each one by its initial condition, [$\varphi(x_o, t_o, t)$ is the solution that satisfies $\varphi(t_o) = x_o$], then the function $\varphi(x_o, t_o, t) : U \times I \times I \to U$ is differentiable with respect to all arguments. That is, if we slightly change the initial conditions, then the solution will only change slightly. Later we will give a proof of all the properties, including Theorem 4.1 for general systems, that is, equation 4.30.

Example: Let $x(t)$ be the number of bacteria in a culture vessel at time t. If the number is large enough, we can think of it as a smooth, differentiable function. If this number is not so large that its density in the vessel is low enough so that they do not have to compete for air and food, then their reproduction is such that $x(t)$ satisfies the equation,

$$\dot{x} = a\,x \qquad a = \text{const.} \simeq 2.3 \frac{1}{\text{day}}, \tag{4.14}$$

that is, the growth rate is proportional to the number of bacteria present. The constant a is the difference between the reproduction rate and the extinction rate. Applying Theorem 4.1, we know that the equation allows us, given the number of bacteria

x_0 at time t_0, to know the number of bacteria at any time. Indeed, the solution of 4.14 is

$$\frac{dx}{x} = a\, dt \tag{4.15}$$

Integrating both sides

$$\ln \frac{x}{x_0} = a(t - t_0) \quad \text{or;} \quad x(t) = x_0\, e^{a(t-t_0)} \tag{4.16}$$

That is, the number of bacteria grows exponentially, unless of course $x_0 = 0$, and depends continuously on the initial number. It should be noted that although the dependence is continuous, the difference between solutions with different initial conditions grows exponentially.

The phase flow in this case is $g_x^t = x\, e^{a\,t}$ and here it is global. With this growth, it is easy to realize that in a few days the number of bacteria and therefore their density will be so large that they will start to compete with each other for food, which will decrease the growth rate. It is experimentally observed that this fact can be taken into account by including in the 4.14 a term proportional to the square of the number of bacteria.

$$\dot{x} = a\, x - b\, x^2 \qquad b = \frac{a}{375}. \tag{4.17}$$

If x is small, the first term dominates and x begins to grow, as we have seen, exponentially until the second term becomes important and the growth rate decreases, tending to zero when $x = x_s$ such that $a\, x_s - b\, x_s^2 = 0$ or $a - b\, x_s = 0$ or $x_s = a/b \simeq 375$ bacteria. Since $\varphi(t) = x_s$ is a solution, due to uniqueness, the solution that tends to x_s that we described above cannot reach the value x_s in finite time. If the number of bacteria present is greater than 375, the growth rate will be negative and the solution will again tend asymptotically to x_s.

The general solution of 4.17 can be obtained similarly to that of 4.14 and is,

$$x(t) = \frac{a/b}{1 + (\frac{a}{b x_0} - 1) e^{-a(t-t_0)}} \tag{4.18}$$

which confirms the previous analysis.

This example allows us to introduce two key concepts in the area. The first is that of a **stationary or equilibrium solution**. These are the constant solutions, that is, they are such that $\dot{x}_s = f(x_s) = 0$, that is, the roots of the function $f(x)$. For equation 4.14 $x_s = 0$, for equation 4.17 $x_s = 0$ and a/b.

Note that stationary solutions do not always exist, for example, the equation with $f(x) = 1$ does not have them, but when they do exist, the problem of finding them reduces to solving at most a transcendental equation. Due to the uniqueness of the solutions, this allows us to divide the plane (x, t) into strips such that if a solution has initial conditions in this strip, it must remain in it.

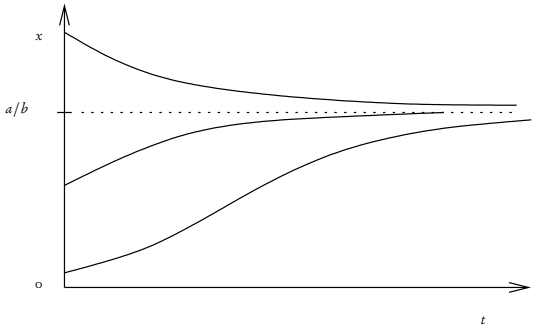

Figure 4.4: Different solutions of the bacterial growth equation: $\varphi_1(t_o) = 0$, $\varphi_2(t_o) = x_s = \frac{a}{b}$, $\varphi_3(t) = x_o < x_s$ and $\varphi_4(t) = x_o > x_s$

The second concept is the **stability** of these solutions. The solution $x_s = 0$ of 4.14 and 4.17 is not stable in the sense that if we choose as an initial condition a point in any neighborhood of this solution, the corresponding solution will initially move away exponentially from the previous one. On the contrary, the solution of 4.17 $x_s = a/b$ is stable in the sense that if we take initial values close to it, the corresponding solutions will asymptotically approach (in the future) the previous one.

Unstable solutions have no physical interest since the slightest perturbation of the system will generate a solution that quickly moves away from the unstable solution. This is not the case in the previous example since the number of bacteria is discrete and the equation only represents an approximation. If the culture vessel is sterilized and hermetically sealed, then there is no possibility of bacterial growth. Later we will analyze the problem of stability in detail.

4.2.2 Extending the Local Solution

If we forget about the physical model we described with 4.17 and take $z_o < 0$, $t_o = 0$, for example, there will be a maximum time, $t_d = \frac{1}{a} \ln[\frac{a/b - z_o}{-z_o}]$, finite for which the solution diverges, that is,

$$\lim_{t \to t_d} z(t) = \infty. \tag{4.19}$$

This shows that the previous theorem can only ensure the existence of an interval I for which the solution exists. The most we can conclude in general, that is, without giving more information about f, is the following Corollary of Theorem 4.1.

Corolary 4.2 *Let $U \subset \mathbb{R}$ be the domain of definition of f, U_c a compact interval [that is, closed and bounded] of U, $z_o \in U_c$. Then the solution $(\varphi(z_o, t_o, t), I)$ of equation 4.1 can be extended either indefinitely or up to the boundary of U_c. This extension is unique in the sense that any pair of solutions, $(\varphi_1(z_o, t_o, t), I_1)$, $\varphi_2(z_o, t_o, t), I_2)$ coincide in the intersection of their definition intervals, $I = I_1 \cap I_2$.*

Proof: First, we prove the global version of the uniqueness of the solution. Let T be the smallest of the upper bounds of the set $\{\tau \,|\, \varphi_1(z_o,t) = \varphi_2(z_o,t) \, forall \, t_o \leq t \leq \tau\}$. Suppose in contradiction that T is an interior point of $I_1 \cap I_2$. By continuity we have, $\varphi_1(z_o,T) = \varphi_2(z_o,T)$, but by the result of Theorem 4.1, using initial conditions $\varphi_1(z_o,T)$, at T we see that both solutions must coincide in a neighborhood of T, which contradicts that T is the maximum. Thus, T must be an endpoint of one of the definition intervals, so the solutions coincide in $I_1 \cap I_2 \cap \{t | t \geq t_o\}$. The case $t \leq t_o$ is treated similarly, and therefore the solutions coincide in all $I_1 \cap I_2$.

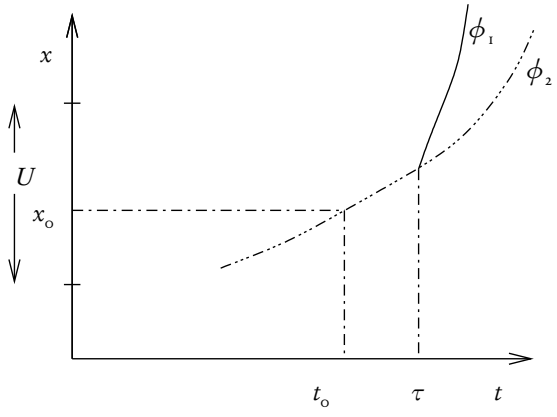

Figure 4.5: Global uniqueness.

Second, we construct the extension. The idea is that if two solutions coincide in $I_1 \cap I_2$, then we can combine them and form one in $I_1 \cup I_2$,

$$\varphi(t) = \begin{cases} \varphi_1(t) & forall t \in I_1 \\ \varphi_2(t) & forall t \in I_2. \end{cases} \tag{4.20}$$

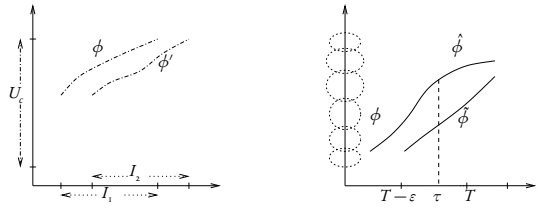

Figure 4.6: Extending the solution.

Let T be the smallest upper bound of $\{\tau | \varphi(t)$ exists and is contained in $U_c \, forall \, t_o \leq t \leq \tau\}$. If $T = \infty$ there is nothing to prove, so suppose $T < \infty$. We will prove that there exists $\varphi(t)$ defined for all $t_o \leq t \leq T$ and such that $\varphi(T)$ is one of the endpoints

of U_c. Corollary 4.1 tells us that given any $\hat{z} \in U$, there exist $\epsilon_{\hat{z}} > 0$ and a neighborhood of $\hat{z}$, $V_{\hat{z}} \subset U$ such that for any $z \in V_{\hat{z}}$ there exists a solution with initial conditions $\varphi(t_o) = z$ defined for all t in $|t - t_o| < \epsilon_{\hat{z}}$. Since U_c is compact, we can choose a finite number of points $\hat{z}_i$ such that $U_c \subset \bigcup V_{\hat{z}_i}$. Let $\epsilon > 0$ be the minimum of the $\epsilon_{\hat{z}_i}$. Since T is the smallest upper bound, there exists τ between $T - \epsilon$ and T such that $\varphi(t) \in U_c \, forall \, t_o \leq t \leq \tau$, but since $\varphi(\tau) \in U_c$ is in one of the $V_{\hat{z}_i}$, there exists a solution $\hat{\varphi}(t)$ defined for all t in $|t - \tau| < \epsilon$ with initial condition $\hat{\varphi}(\tau) = \varphi(\tau)$. Using $\hat{\varphi}(t)$ we can now extend φ, whose extension we will call $\tilde{\varphi}$, to the interval $t_o \leq t < \tau + \epsilon$. But $\tilde{\varphi}(\theta) \in U_c \, forall \, t_o < \theta < T$ since $\tilde{\varphi}(\theta) = \varphi(\theta) \in U_c$, and otherwise T would not be a smallest upper bound. By continuity $\tilde{\varphi}(T) \in U_c$ and as by definition of T for any open interval, I_T, containing T, $\tilde{\varphi}[I_T] \notin U_c$ we conclude that $\tilde{\varphi}(T)$ is an endpoint of U_c ♠

4.2.3 The Non-autonomous Case

This is the case where instead of 4.1 we have

$$\frac{d\dot{z}}{dt} = f(z,t). \tag{4.21}$$

The geometric interpretation is the same, only now the *little lines* have an angle that also depends on the variable t, therefore we expect that if we start at a point (t_o, z_o) and trace a curve tangent to the little lines, we will also obtain a unique solution. This is indeed the case, but its discussion will be included in the general theory of first-order autonomous systems that we will state later.

Assuming certain conditions for $f(z,t)$ – for example, that it is Lipschitz with respect to both variables – it can be seen that local flows also exist, but these now also depend on t_o and not just on the difference between the initial and final time. We will denote these flows as $g^t_{t_o}(z)$ and the semigroup property they satisfy is $g^t_{t_1} \circ g^{t_1}_{t_o}(z) = g^t_{t_o}(z)$.

There is a special class of non-autonomous equations that can be reduced to the already studied case. This is the class where $f(z,t)$ can be written as the quotient of two functions, one of z and another of t,

$$f(z,t) = \frac{g(z)}{h(t)}. \tag{4.22}$$

In this case, the geometric interpretation is that the parameter t is not the most suitable to describe this system. This is remedied by choosing another parameter τ given by, $\frac{d\tau}{dt} = \frac{1}{h(t)}$, or

$$\tau - \tau_o = \int_{t_o}^{t} \frac{dt}{h(t)}. \tag{4.23}$$

Once $t(\tau)$ is obtained [See Theorem 4.1], we must solve,

$$\frac{dz}{d\tau} = g(z(\tau)), \tag{4.24}$$

[Using Theorem 4.1 again], obtaining $\varphi(\tau)$. The solution with respect to the original parameter is obtained by defining $\varphi(t) = \varphi(\tau(t))$.

4.3 Reduction to First-Order Systems

Definition: The **general ordinary differential equation of order** m is an equation of the form:

$$F(x, x^{(1)}, \ldots, x^{(m)}, t) = 0 \tag{4.25}$$

That is, an implicit equation of x, a map between $\tilde{I} \subset \mathbb{R}$ and $\tilde{U} \subset \mathbb{R}^{m \times n}$, (i.e., x is a vector of n components), and its derivatives (where $x^{(i)} \equiv \frac{d^i x}{dt^i}$] denotes the i-th derivative with respect to the independent parameter, t) up to order n [2].

Definition: We will say that the m–times differentiable map $x(t) : I \to \mathbb{R}^n$ is **a solution** of the previous equation if F evaluated in this map is well-defined and identically zero over the entire interval I.

In this course, we will assume that 4.25 can be solved for x^m, that is, that 4.25 is equivalent to the following equation:

$$x^{(n)} = f(x, x^{(1)}, \ldots, x^{(n-1)}, t) \tag{4.26}$$

in open sets $I \subset \tilde{I} \subset \mathbb{R}$ and $U \subset \tilde{U} \subset \mathbb{R}^m$

Example: The ODE $F(x, x^{(1)}, t) = ((x^{(1)})^2 - 1) = 0$ implies one of the following ODEs in our sense:

$$\begin{aligned} & x^{(1)} = 1 \\ & or \\ & x^{(1)} = -1 \end{aligned} \tag{4.27}$$

If we know the values of x and its derivatives up to the $(m-1)$-th order at a point t_0, then equation 4.26 allows us to know the m-th derivative at that point. If we assume that f is continuous, then this allows us to know x and its derivatives up to the $(m-1)$-th order at a point $t_0 + \varepsilon$ sufficiently close (the error incurred will be of the order of $\varepsilon \times$ *derivatives of* f), but then we can use equation 4.26 again and thus approximately know the n-th derivative of x at $t_0 + \varepsilon$. Proceeding in this way, we can achieve an *approximate solution* in a given interval. At least intuitively, it would seem that by making the interval ε smaller and smaller, we should obtain a solution. This is indeed the case! And we will prove it later in the course. The important thing for now is the fact that to obtain a solution, we must give at a point t_0 the value of all the derivatives of the unknown variable of order less than the one that determines the system, that is, the one that appears on the left side of equation 4.26.

[2] In reality, to correctly define 4.25, the open sets $\tilde{U}^i$ for the i-th derivative with respect to x where F is defined must be given.

Introducing $\tilde{u}^1 \equiv x$, $\tilde{u}^2 \equiv x^{(1)}$, ... , $\tilde{u}^n \equiv x^{(m-1)}$, we can write 4.26 in the form:

$$\begin{aligned} \dot{\tilde{u}}^1 &= \tilde{u}^2 \\ \dot{\tilde{u}}^2 &= \tilde{u}^3 \\ &\vdots \\ \dot{\tilde{u}}^m &= f(\tilde{u}^1, \tilde{u}^2, \ldots, \tilde{u}^m, t) \end{aligned} \tag{4.28}$$

where we have denoted with a dot over the function its derivative with respect to the parameter t. Therefore, 4.26 is equivalent to a first-order equation for the vector of vectors $\tilde{\boldsymbol{u}} \subset U \subset \mathbb{R}^{n\times m}$,

$$\dot{\tilde{\boldsymbol{u}}} = \tilde{V}(\tilde{\boldsymbol{u}}, t) \tag{4.29}$$

where $\tilde{V}$ is also a vector in some subset of $\mathbb{R}^{n\times m}$. If $\tilde{V}(\tilde{u}, t)$ depends explicitly on t, then we can add to $\tilde{\boldsymbol{u}}$ another component, the $(n \times m + 1)$-th with $\tilde{u}^{n\times m+1} = t$, and therefore we have that $\dot{\tilde{u}}^{n\times m+1)} = 1$. Equation 4.29, plus this last equation, take the form,

$$\dot{\boldsymbol{u}} = V(\boldsymbol{u}) \tag{4.30}$$

where $\boldsymbol{u} = (\tilde{\boldsymbol{u}}, \tilde{u}^{n\times m+1})$ and

$$V^i(\boldsymbol{u}) = \begin{cases} \tilde{V}^i(\tilde{\boldsymbol{u}}, \tilde{u}^{n\times m+1}) & \text{if } 1 \le i \le m \times n \\ 1 & \text{if } i = n \times m + 1) \,, \end{cases} \tag{4.31}$$

a map between $U \subset \mathbb{R}^{n\times m+1}$ and $\mathbb{R}^{n\times m+1}$.
Thus, we arrive at the following theorem.

Theorem 4.2 (Reduction) *If the function f of the system 4.26 of m ordinary differential equations of order n is differentiable, then this system is equivalent to a system of $m \times n + 1$ first-order ordinary differential equations that do not explicitly depend on the independent variable. That is, for each solution of the system 4.26, we can construct a solution of the system 4.30 and vice versa.*

Proof: Clearly, given a sufficiently differentiable solution of the system 4.26, we can write with it a vector $\boldsymbol{u}$ and it will satisfy (because we constructed it this way) the corresponding system 4.30. Therefore, every solution of the original system is a solution of the associated first-order system.

To prove the inverse, that is, that every solution $\boldsymbol{u}(t)$ of the system 4.30 gives rise to a solution of the system 4.26, it is only necessary to take repeated derivatives of the first components, $x(t) = u^1(t)$, and use the equations in the corresponding order until obtaining the original system satisfied by the nth derivative of $x(t)$ ♠

This theorem is important because it allows us to encompass the entire theory of ODEs by studying equation 4.30, which, as we will see, has a clear geometric meaning. However, it should be noted that many times in physics there appear very special subclasses of equations with particular properties of great physical importance that the general theory does not contemplate.

Example: Mathematical Pendulum

$$\ddot{x} = k\,x \quad u^1 = x\,,\, u^2 = \dot{x}$$

$$\frac{d}{dt}\begin{pmatrix} u^1 \\ u^2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ k & 0 \end{pmatrix}\begin{pmatrix} u^1 \\ u^2 \end{pmatrix} \tag{4.32}$$

Note that if $u = \gamma(t) : I \to U$ is a solution of 4.30 then $\forall s \in \mathbb{R}$ such that $t+s \in I$, $\gamma(t+s)$ is also a solution. Indeed,

$$\left.\frac{d}{dt}\gamma(t+s)\right|_{t=t_o} = \left.\frac{d}{dt}\gamma(t)\right|_{t=t_o+s} = V(\gamma(t_o+s)) = \left. V(\gamma(t+s))\right|_{t=t_o}. \tag{4.33}$$

Due to this, from now on we will take t_o, the point where we give the initial condition, equal to zero.

Exercise: If $u = \sin(t) : \mathbb{R} \to [0,1]$, is a solution of 4.30 prove that $u = \cos(t)$ is also a solution.

4.4 ODE Systems

Do ODE systems have a geometric interpretation? As we have seen, they have the generic form:

$$\frac{dx^i}{dt} = v^i(x^j) \qquad i = 1,\ldots,n. \tag{4.34}$$

Let $\gamma(t)$ be a curve between an interval $I \in \mathbb{R}$ and M a manifold of n dimensions. Let $p = \gamma(t_0)$ be a point of M and $(U, \varphi = (x^1,\ldots,x^n))$ a chart with $p \in U$. Then $\varphi \circ \gamma(t) = (x^1(t),\ldots,x^n(t))$ is a map between I and an open set of $\mathbb{R}^n$, and $\dfrac{dx^i}{dt}$ are the components of the tangent vector to γ in the coordinate basis $\{\boldsymbol{x_i}\}$ at the point $\gamma(t)$ of M. Therefore, $v^i(x^j)$ are the components of a vector field $\boldsymbol{v}$ in the coordinate basis $\{\boldsymbol{x_i}\}$ at the point $p \in M$ corresponding to $\varphi(p) = (x^1,\ldots,x^n)$. Thus, we see that ODE systems can be interpreted as a vector field $\boldsymbol{v}$ on a manifold M, and their solutions as curves $\gamma(t)$ such that their tangent at every point is $\boldsymbol{v}$.

Example: The Physical Pendulum

$$\begin{cases} \ddot{\theta} &= -\sin\theta \quad \text{ó} \\ \dot{\theta} &= z \\ \dot{z} &= -\sin\theta \end{cases} \tag{4.35}$$

Here the vector field is $z\dfrac{\partial}{\partial\theta} - \sin\theta\dfrac{\partial}{\partial z}$,

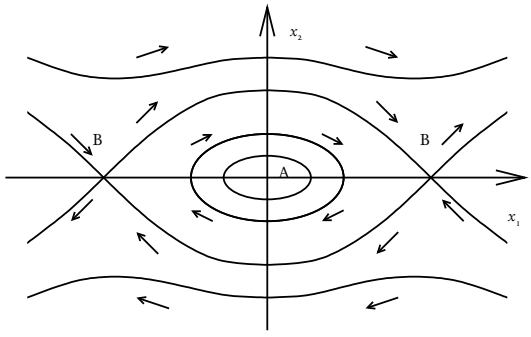

Figure 4.7: The Physical Pendulum.

note that the diagram in the plane repeats every 2π along the θ axis. In reality, the equation makes physical sense on a cylinder, with θ being the angular variable, so the manifold is a cylinder. The **integral curves**, i.e., the images of the maps $\gamma(t)$ that are solutions, are constructed by following the vectors so that they are always tangent to the curves.

From the example, it is clear that in general, if we give a point of M and follow the vectors from it, we can find a solution $\gamma(t)$ that passes through this point. If we have two solutions $\gamma_1(t)$, $\gamma_2(t)$, they cannot cross each other – without being tangent – at points where the vector field is non-zero [since otherwise, their tangents would give two different vectors at the point]. If we assume that the vector field is differentiable, then it can be seen that different curves can never touch, and therefore the solutions are unique.

Of course, as seen for example in figure 4.8, the same curve can intersect itself (note that this can only happen if the system is autonomous, as otherwise, the inclusion of the temporal variable prevents any curve from intersecting itself). In such a case, it can be proven that this solution is **periodic**, meaning that there exists $T > 0$ such that $\gamma(t+T) = \gamma(t)$. Note in particular that this implies that γ is defined for all t in $\mathbb{R}$.

Problem 4.1 *If M is a compact manifold (example: a sphere or a torus), it can be proven using Corollary 4.2 that all its phase curves are defined for all t. However, there are some that are not closed, give an example.*

4.4.1 First Integrals

Given a vector field v on M, if there exists $f : M \to \mathbb{R}$ non-constant such that $v(f) = 0$ on M, then we will say that f is a **first integral or constant of motion** of the ODE system generated by v.

It can be shown that if n is the dimension of M, then locally, around points where

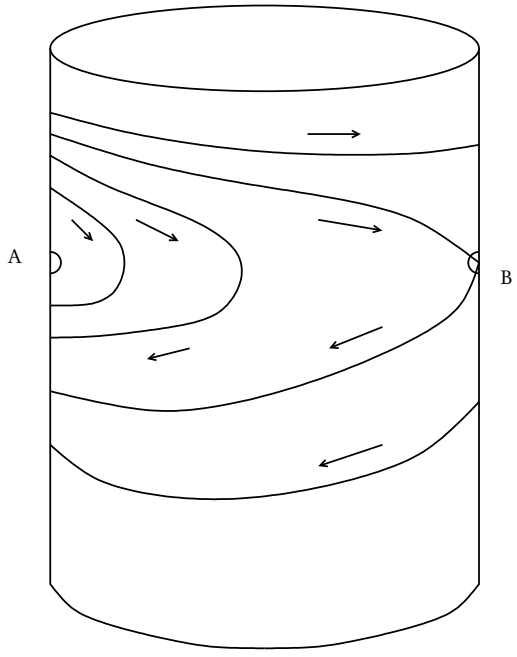

Figure 4.8: The Physical Pendulum Manifold.

$v \neq 0$, there exist $n-1$ functionally independent first integrals. However, it is easy to find vector fields that do not have any first integrals.

Examples–Problems

Problem 4.2 *The system*

$$\begin{cases} \dot{x}_1 = x_1 \\ \dot{x}_2 = x_2 \end{cases}$$

has none, why? Hint: Draw the integral curves.

Problem 4.3 *Consider the system*

$$\begin{cases} \dot{x}_1 = k_1 \\ \dot{x}_2 = k_2 \end{cases}$$

on the torus (obtained by identifying the points (x_1, x_2) and $(x_1 + n, x_2 + m)$ of $\mathbb{R}$). For which values of k_1, k_2 does a first integral exist and for which does it not? Hint: same as 1).

Problem 4.4 *The Hamilton equations of classical mechanics form an ODE system,*

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \tag{4.36}$$

$$\dot{p}_i = -\frac{\partial H}{\partial q_i} \tag{4.37}$$

$i = 1, \ldots, n$, where $H : \mathbb{R}^{2n} \to \mathbb{R}$ is the Hamiltonian function or energy. Show that H is a first integral of this system.

The previous examples show that there are ODE systems that globally are not derivable from a variational principle.

Just as knowing the energy of a Hamiltonian system is an invaluable aid in understanding the quantitative form of its motion, knowing a first integral of a general ODE system also provides great help since we know that the phase curves are restricted to their level surfaces ($f =$ constant), which allows us to effectively reduce the order of the equation by one: we only have to consider the problem on the manifold given by the points $f = constant$.

4.4.2 Fundamental Theorem of ODE Systems

Theorem 4.3 *Let v be a continuously differentiable (C^1) vector field in a neighborhood of $p \in M$. Then,*

i) There exists a neighborhood U_p of p and an open set $I \subset \mathbb{R}$, such that given any $q \in U_p$ ($t_o \in I$), there exists a differentiable curve $\gamma_q(t) : I \to M$ satisfying

$$\begin{aligned} \left.\frac{d\gamma_q}{dt}\right|_t &= v|_{\gamma_q(t)} \\ \gamma_q(t_o) &= q. \end{aligned} \tag{4.38}$$

ii) If γ_q^1 and γ_q^2 satisfy condition i), then $\gamma_q^1 = \gamma_q^2$ in $I^1 \cap I^2$.

iii) $\gamma_q(t)$ is also C^1 when considered as a function from $I \times U_p \to M$. That is, it is continuously differentiable with respect to the initial condition.

iv) Let U be a compact (bounded) region of M, $v \in C^1$, and $p \in U$, then $\gamma_p(t)$ can be extended (forward and/or backward) to the boundary of U. In particular, if M is compact and $v \in C^1(M)$, then $\gamma_p(t)$ exists globally (for all $t \in \mathbb{R}$).

We will prove this theorem later, after introducing the necessary mathematical tools.

4.4.3 Parameter Dependence, Variation Equation

A very important consequence of the fundamental theorem in applications is the following:

Corolary 4.3 *Let v_λ be a differentiable vector field in $U \subset M$ that depends differentiably on a parameter $\lambda \in A \subset \mathbb{R}$. Then given $p \in U$, $t_o \in I$, and $\lambda_o \in A$, there exist neighborhoods U_p, I_{t_o}, and A_{λ_o} such that for any triplet $(q, t, \lambda) \in U_p \times I_{t_o} \times A_{\lambda_o}$, there exists a unique integral curve of $v_\lambda, \gamma_\lambda(t) : I_{t_o} \times A_{\lambda_o} \to U_p$ with $\gamma_\lambda(o) = q$. This depends differentiably on q, t, and λ.*

Proof: Consider the vector field $(v_\lambda, o) : U \times A \to TM \times \mathbb{R}$. By hypothesis, this field is differentiable and therefore has integral curves $(\gamma_\lambda\ ,\ \lambda)$ that depend differentiably on the initial conditions and therefore on λ ♠

Warning: The interval $I_{t_o} \times A_{\lambda_o}$ where the solution is differentiable is not necessarily the interval of definition of the solution. For example, it is not true that even when the solution is defined for all t in $\mathbb{R}$, the limits $\lim_{T\to\infty} \gamma_\lambda(t)$ and $\lim_{\lambda\to\lambda_o} \gamma_\lambda(t)$ commute.

The practical importance of this corollary is that it allows us to obtain approximate solutions to already known ones. Indeed, suppose that for $\lambda = o$ we know some integral curve of $v|_{\lambda=o}$, $\gamma_o(t)$ with $\gamma_o(o) = p$. Then, applying the Corollary and a Taylor series expansion, we have

$$\gamma_\lambda(t) = \gamma_o(t) + \lambda\,\gamma_1(t) + O(\lambda^2) \tag{4.39}$$

that is, $\gamma_o(t) + \lambda\,\gamma_1(t)$ is a good approximation to $\gamma_\lambda(t)$ for λ sufficiently small.

It remains to find the equation that $\gamma_1(t)$ satisfies. This is obtained by differentiating with respect to λ the differential equation at $\lambda = 0$,

$$\frac{d}{d\lambda}\left(\frac{d\gamma_\lambda(t)}{dt} = v_\lambda(\gamma_\lambda(t))\right)\Bigg|_{\lambda=0}, \tag{4.40}$$

expanding and using coordinates

$$\frac{d}{dt}\gamma_1^j = \left.\frac{\partial v^j}{\partial x^i}\right|_{(\gamma_0,\lambda=0)} \cdot \gamma_1^i + \left.\frac{dv^j}{d\lambda}\right|_{(\gamma_0,\lambda=0)}, \tag{4.41}$$

$$= A^j{}_i(t)\,\gamma_1^i + b^j(t).$$

that is, the equation for $\gamma_1(t)$ –usually called the variation equation– is a linear inhomogeneous non-autonomous equation. Since we took the initial condition $\gamma_\lambda(0) = p \ \ \forall\ \lambda$, the initial condition we will consider for 4.4.3 is $\gamma_1(0) = 0$.

Problem 4.5 *What would be the initial condition for γ_1 if the condition for γ_λ were $\gamma_\lambda(0) = p(1-\lambda) + \lambda q$? Consider only the one-dimensional case but think about the general case.*

Problem 4.6 *If we decided to consider an approximation up to the second order $\gamma_\lambda = \gamma_0 + \lambda\gamma_1 + \lambda^2\gamma_2 + O(\lambda^3)$, what equations would we obtain for γ_2?*

Example: **a)** A body falls vertically in a low-viscosity medium, which depends on its position and velocity.

$$\ddot{x} = -g + \varepsilon\, F(x,\dot{x}) \qquad \varepsilon \ll 1 \tag{4.42}$$

Calculate the effect of this resistance on the motion.

The solution, $x(t)$, will depend smoothly on ε, so we can expand it in a Taylor series with respect to ε.

$$x(t) = x_0(t) + \varepsilon\, x_1(t) + O(\varepsilon^2). \tag{4.43}$$

The solution for $\varepsilon = 0$ is clearly

$$x_0(t) = x_i + v_i t - g\frac{t^2}{2}, \tag{4.44}$$

where x_i and v_i are the initial position and velocity, respectively, while the variation equation is

$$\ddot{x}_1 = F(x_0,\dot{x}_0), \qquad x_1(0) = 0,\ \ \dot{x}_1(0) = 0. \tag{4.45}$$

Integrating, we obtain

$$x_1(t) = \int_0^t \int_0^\tau F(x_0(\tilde{\tau}),\dot{x}_0(\tilde{\tau}))\, d\tilde{\tau}\, d\tau. \tag{4.46}$$

b) The second example is self-oscillations.
Consider the system,

$$\begin{aligned} \dot{x}_1 &= x_2 + \varepsilon f_1(x_1, x_2) \\ \dot{x}_2 &= -x_1 + \varepsilon f_2(x_1, x_2), \quad \text{for } \varepsilon \ll 1. \end{aligned} \tag{4.47}$$

Without loss of generality, we can assume $f_i(0,0) = 0$. When $\varepsilon = 0$, we obtain the equations of the pendulum –in the small amplitude approximation–, that is, a conservative system. Its energy –first integral– is given by $E(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ and its phase curves are then circles of radius $\sqrt{2E}$.

When $\varepsilon > 0$, the phase curves are not necessarily closed, but due to the previous Corollary, they can at most be spirals with a distance of the order of ε between turns. These can tend towards some stationary point within the circle $\sqrt{2E_i}$, where $E_i =$ initial energy, that is, where $x_2 + \varepsilon f_1(x_1, x_2) = -x_1 + \varepsilon f_2(x_1, x_2) = 0$, or they can asymptotically approach some closed orbit or finally diverge. That these are the only options is the result of a classic theorem by **PoincarÃ©–Bendixson**, which shows that in general these are the only possibilities in two dimensions.

To study which case it is among these three, we consider the variation of energy between two consecutive turns.

Taking a time derivative of the energÃ expression defined above, and using the evolution equations we obtain:

$$\dot{E}(x_1, x_2) = \varepsilon (x_1 f_1 + x_2 f_2). \tag{4.48}$$

With this expression we can estimate the change on the energy to first order after a revolution around an unperturbed orbit:

$$\begin{aligned} \Delta E &= \varepsilon \int_{t_i}^{t_f} (x_1 f_1 + x_2 f_2)\, dt + O(\varepsilon^2) \\ &= \varepsilon \int_{t_i}^{t_f} (-\tfrac{dx_2}{dt} f_1 + \tfrac{dx_1}{dt} f_2)\, dt + O(\varepsilon^2) \\ &= \varepsilon \oint (-f_1\, dx_2 + f_2\, dx_1) + O(\varepsilon^2) \\ &= \varepsilon F(E) + O(\varepsilon^2) \end{aligned} \tag{4.49}$$

where the integral is along a circle of radius $\sqrt{2E}$ in the direction of motion. That is, we have approximated the curve with $\varepsilon > 0$ by the curve with $\varepsilon = 0$.

If $F(E)$ is positive, the energy increases and the system undergoes growing oscillations.

If $F(E)$ is negative, the oscillations decrease in amplitude and the system eventually tends to an equilibrium point.

If $F(E)$ changes sign, say at $E = E_0$, it can be proven that near this circle there is a closed phase curve Γ_ε.

If $F'(E)|_{E=E_0}$ is negative, Γ is a stable limit cycle –any nearby phase curve asymptotically approaches it–. If $F'(E)|_{E=E_0}$ is positive, Γ is unstable.

Problem 4.7 *Show that if there are two stable limit cycles, then there must also be at least one that is unstable.*

From the two previous examples, the practical importance of the Corollary is clear. The differentiable dependence on the initial conditions also leads to the variation equation, therefore this equation, which is linear and inhomogeneous, is of utmost importance and in the next chapter, we will study it in more detail.

4.5 Problems

Problem 4.8 (Kiseliov) *Find all the solutions of*

$$\frac{dx}{dt} = \frac{3}{2}x^{2/3}. \tag{4.50}$$

Hint: there are infinitely many, and they are obtained from segments of some particular solutions. Graph some of these solutions.

Problem 4.9 (Kiseliov) *Apply the existence and uniqueness theorem to determine the regions where the following equations have a unique solution:*

a) $\dot{x} = x + 3x^{1/3}$.

b) $\dot{x} = 1 - \cot x$.

c) $\dot{x} = \sqrt{1 - x^2}$.

Problem 4.10 (Kiseliov) *Solve the following equations, in all cases give the general solutions as a function of an initial value* x_o *for the initial time* t_o.

a) $t\dot{x} + x = \cos(t)$. *(Use the first integral of the homogeneous part.)*

b) $\dot{x} + 2x = e^t$. *(Use variation of constants.)*

c) $(1 - t^2)\dot{x} + tx = 2t$. *(First solve the homogeneous part and then add a particular inhomogeneous solution.)*

d) $x\dot{x} = t - 2t^3$.

Problem 4.11 (Kiseliov) *Solve the following equations by making a change of variable in the independent variable (t).*

a) $\dot{x}x = -t$. *(*$t = \cos(s)$*.)*

b) $\dot{x}t - x = 0$.

c) $\dot{x} + e^{\dot{x}} = t$.

Problem 4.12 (Kiseliov) *Graph the isoclines (lines of equal slope in the* (t, x) *plane) and then trace the solutions of the following equations:*

a) $\dot{x} = 2t - x$.

b) $\dot{x} = \sin(x + t)$.

c) $\dot{x} = x - t^2 + 2t - 2$.

d) $\dot{x} = \frac{x-t}{x+t}$.

Problem 4.13 *Solve the equation:*

$\dot{x} = A(t)x + B(t)x^n$ *Hint: Use the change of variable* $y = x^{1-n}$.

Problem 4.14 *Solve the equation*

$$\frac{dx}{dt} = i\lambda x + Ae^{i\omega t} \quad \lambda,\ \omega \in \mathbb{R} \tag{4.51}$$

and see how its real part behaves. Examine the cases:
a) $A = 0$, *b)* $(A \neq 0,\ \lambda \neq \omega)$ *and c)* $(A \neq 0,\ \lambda = \omega)$.

Problem 4.15 *The equation of the physical pendulum.*
a) Graph the vector field of the equation

$$\frac{d^2\theta}{dt^2} = -\sin(\theta), \quad -\pi \leq \theta \leq \pi. \tag{4.52}$$

b) Graph some integral curves around $(\theta = 0, z = \frac{d\theta}{dt} = 0)$.
c) Graph the integral curves that pass through the point $\theta = \pm\pi, z = 0$). *Infer that the time required to reach this point along these curves is infinite.*
d) Graph the vector field along a line $z = z_0$. *Infer from this that the solutions can never exceed the speed acquired when passing through the point* $\theta = 0$. *Hint: Conclude this first for the region* $0 \leq \theta \leq \pi, \quad 0 \leq z \leq z_0$ *and then use the symmetry of the solution.*
e) Let $E(z,\theta) := \frac{z^2}{2} - \cos(\theta)$. *See that* $\frac{dE}{dt} = 0$. *Use this quantity to analyze the behavior of the solutions with initial data* $(z_0, \theta_0 = 0)$. *In particular, determine which cross the* $z = 0$ *axis and which do not.*

Problem 4.16 *Consider the equation:*

$$\frac{dz}{dt} = \begin{cases} f(z) & |z| < 1 \\ iz & |z| \geq 1 \end{cases} \tag{4.53}$$

Where $f(z)$ *is continuous with* iz *at* $|z| = 1$. *Infer that no matter the form of* $f(z)$, *the solutions never escape the region* $\{|z(t)| \leq \max\{|z(0)|, 1\}\}$.

Problem 4.17 *Consider a body affected by a central force, that is,*

$$m\frac{d^2\vec{x}}{dt^2} = f(r)\vec{x} \quad r := |\vec{x}|. \tag{4.54}$$

a) Find an equivalent first-order system.
b) Verify that given any (constant) vector $\vec{c}$, *then* $F(\vec{x}, \vec{p}) := \vec{c}\cdot(\vec{x}\wedge\vec{p})$, *where* $\vec{p} := \frac{d\vec{x}}{dt}$, *is a first integral.*

Problem 4.18 *Consider the equation:*

$$\frac{d\vec{x}}{dt} = \vec{x}\wedge\vec{\omega}(\vec{x}). \tag{4.55}$$

a) See that $R := \vec{x}\cdot\vec{x}$ *is a first integral.*
b) Conclude that the solutions of this system exist for all time.

Problem 4.19 (Strichartz) *Show that*

$$J_k(t)=\sum_{j=0}^{\infty}\frac{(-1)^j(t/2)^{k+2j}}{j!(k+j)!}, \tag{4.56}$$

has an infinite radius of convergence and satisfies the Bessel equation, $x''(t)+(1/t)x'(t)+(1-k^2/t^2)x(t)=0$.

Problem 4.20 *Solve the system:*

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} -x_2+\varepsilon(x_1^2+x_2^2) \\ x_1 \end{pmatrix} \tag{4.57}$$

For the initial data $(x_1(0)=1, x_2(0)=0)$. *Now investigate the solution near* $\varepsilon=0$ *by obtaining the coefficients in its Taylor series with respect to the parameter* ε. *Find which equations these coefficients satisfy.*

Problem 4.21 (Arnold) *Consider the equation of the physical pendulum:*

$$\frac{d^2x}{dt^2}=-\sin(x).$$

See how the frequency varies as a function of the amplitude for small solutions. Hint: assume a solution in a Taylor series of the amplitude and solve until finding the first nontrivial contribution to the linearized solution. Find the period by locating the return points (zero velocity).

Bibliography notes: For this and the next three chapters, I recommend the following books: [3], [7], and [8]. Especially the first one. Ordinary differential equations are the basis of classical mechanics, although in the latter there is also a particular geometric structure that is fundamental. Almost everything we have seen in this chapter is the *local* theory, the *global* theory, that is, the behavior of the solutions for large values of the independent variable (time in most cases) has only been understood in recent years, giving rise to new concepts such as chaos, attractors, fractal dimensions, etc. Unfortunately, these aspects have not yet been sufficiently synthesized, and therefore I cannot recommend any book on the subject. There are too many interesting ideas, but each of them requires a great deal of information.

CHAPTER 5

LINEAR SYSTEMS

5.1 Homogeneous Linear System

Theorem 5.1 *Let $A(t)$ be continuous on $I \subset \mathbb{R}$, closed. Then there exists a unique solution $\boldsymbol{x}(t) : I \to V$ of the equation,*

$$\frac{d\boldsymbol{x}(t)}{dt} = A(t)\boldsymbol{x}(t), \tag{5.1}$$

with initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0 \in V$, $t_0 \in I$.

Proof: Let $\boldsymbol{x}, \boldsymbol{y} \in V$, then

$$||A(t)\boldsymbol{x} - A(t)\boldsymbol{y}||_V \leq ||A(t)||_{\mathscr{L}} ||\boldsymbol{x} - \boldsymbol{y}||_V \tag{5.2}$$

and therefore $A(t)\boldsymbol{x}$ is Lipschitz in V. The fundamental theorem guarantees local existence and uniqueness. The theorem will be proven if we show that $\boldsymbol{x}(t)$ cannot become infinite in finite time. For this, we use,

$$\begin{aligned} \frac{d}{dt}||\boldsymbol{x}||_V &= \lim_{t_1 \to t} \frac{||\boldsymbol{x}(t_1)||_V - ||\boldsymbol{x}(t)||_V}{t_1 - t} \leq \lim_{t_1 \to t} \frac{||\boldsymbol{x}(t_1) - \boldsymbol{x}(t)||}{t_1 - t} \\ &= \left|\left| \frac{d\boldsymbol{x}}{dt} \right|\right|_V = ||A\boldsymbol{x}||_V \leq ||A||_{\mathscr{L}} ||\boldsymbol{x}||_V. \end{aligned} \tag{5.3}$$

Integrating this inequality between t_0 and $t \in I$ we obtain,

$$||\boldsymbol{x}(t)||_V \leq ||\boldsymbol{x}(t_0)||_V e^{\int_{t_0}^{t} ||A(\tilde{t})||_{\mathscr{L}} d\tilde{t}}, \tag{5.4}$$

which shows that $\boldsymbol{x}(t)$ cannot escape to infinity in finite time and therefore completes the proof.

Consider the set of solutions of 5.1 defined on a single closed interval $I \subset \mathbb{R}$, $Sol(A, I)$. This set has the structure of a vector space. Indeed, if $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ are two solutions of 5.1, then $\boldsymbol{x}(t) + \alpha \boldsymbol{y}(t)$, $\alpha \in \mathbb{R}$, is also a solution. What is its dimension? The following theorem answers this question and shows that any solution of 5.1 can be expressed as a linear combination of a finite number $(n) = \dim V$ of solutions.

Theorem 5.2 $\dim Sol(A,I) = \dim V$.

Proof: Let $\{\boldsymbol{u}_i^\circ\}$, $i = 1,\ldots,n$ be a basis of V, $t_\circ \in I$ and $\{\boldsymbol{u}_i(t)\}$, $i = 1,\ldots,n$ $t \in I$ the set of solutions of 5.1 with initial condition $\boldsymbol{u}_i(t_\circ) = \boldsymbol{u}_i^\circ$. We will show that $\{\boldsymbol{u}_i(t)\}$ form a basis of $Sol(A,I)$, and therefore $\dim Sol(A,I) = \dim V$. First, let's see that the solutions $\{\boldsymbol{u}_i(t)\}$ are linearly independent.

Suppose there are constants c^i such that $\boldsymbol{x}(t) = \sum_{i=0}^n c^i \boldsymbol{u}_i(t)$ is zero for some $\tilde{t}$ in I. Since $\boldsymbol{x}(t)$ is a solution of 5.1 and these are unique when an initial condition is specified, taking in this case $\boldsymbol{x}(\tilde{t}) = 0$, we see that $\boldsymbol{x}(t) = 0 \;\forall\; t \in I$. In particular, we have that $\boldsymbol{x}(t_\circ) = \sum_{i=0}^n c^i \boldsymbol{u}_i^\circ = 0$ and the independence of the set $\{\boldsymbol{u}_i^\circ(t)\}$ implies that $c^i = 0 \;\forall\; i = 1,\ldots,n$, proving linear independence. Note that we have not only proven that $\{\boldsymbol{u}_i(t)\}$ are linearly independent as elements of $Sol(A,I)$ – for which we would only have had to prove that if $\sum_{i=0}^n c^i \boldsymbol{u}_i(t) = 0 \;\forall\; t \in I$ then $c^i = 0 \;\forall\; i = 1,\ldots,n$ which is trivial since $t_\circ \in I$ – but also that for each $t \in I$ the $\{\boldsymbol{u}_i(t)\}$ are linearly independent as elements of V, this result will be used later.

To complete the theorem, let's now see that any solution of 5.1 $\boldsymbol{x}(t) : I \to V$, that is, any element of $Sol(A,I)$ can be written as a linear combination of $\{\boldsymbol{u}_i(t)\}$. Let $\boldsymbol{x}(t) \in Sol(A,I)$, since $\{\boldsymbol{u}_i^\circ\}$ form a basis of V there will be constants c^i, $i = 1,\ldots,n$ such that $\boldsymbol{x}(t_\circ) = c^i \boldsymbol{u}_i^\circ$. Consider then the solution of 5.1 given by $\varphi(t) = c^i \boldsymbol{u}_i(t)$. Since $\varphi(t_\circ) = \boldsymbol{x}(t_\circ)$ and the solutions are unique, we have,

$$\boldsymbol{x}(t) = \varphi(t) = c^i \boldsymbol{u}_i(t) \;\forall\; t \in I \tag{5.5}$$

The previous theorem tells us that if we know n linearly independent solutions of A, $\{\boldsymbol{u}_i(t)\}$, we know all its solutions, since these will be linear combinations of $\{\boldsymbol{u}_i(t)\}$. The dependence on the initial data is also linear, that is, if $\boldsymbol{x}(t)$, $\boldsymbol{y}(t) \in Sol(A,I)$ and $\boldsymbol{x}(t_\circ) = \boldsymbol{x}_\circ$ and $\boldsymbol{y}(t_\circ) = \boldsymbol{y}_\circ$ then $\boldsymbol{x}(t) + \alpha \boldsymbol{y}(t)$, $\alpha \in \mathbb{R}$, is the solution with initial data $\boldsymbol{x}_\circ + \alpha \boldsymbol{y}_\circ$ and therefore the map $g_{t_\circ}^t$ – which in this case we will call $\boldsymbol{X}_{t_\circ}^t$ – that takes initial data given at $t = t_\circ$ to solutions with those data, at time t is a linear operator from V to V. Indeed, if $\{\theta_\circ^i\}$ is the co-basis of the basis $\{\boldsymbol{u}_i\}$, i.e. $\theta_\circ^j(\boldsymbol{u}_i^\circ) = \delta_i^j$. Then $\boldsymbol{X}_{t_\circ}^t = \sum_{i=1}^n \boldsymbol{u}_i(t)\theta_\circ^i$. Due to the linear independence of $\{\boldsymbol{u}_i(t)\}$ as elements of V, it can be seen that the map $\boldsymbol{X}_{t_\circ}^t : V \to V$ is injective and therefore invertible, for each $t \in I$. Its inverse, which we will denote $(\boldsymbol{X}_{t_\circ}^t)^{-1}$, takes the solution at t to its initial data at $t = t_\circ$, that is, $(\boldsymbol{X}_{t_\circ}^t)^{-1} = \boldsymbol{X}_t^{t_\circ}$.

Exercise: Prove that $(\boldsymbol{X}_{t_\circ}^t)^{-1} = \boldsymbol{X}_t^{t_\circ}$.

Exercise: Prove that $\boldsymbol{X}_{t_\circ}^t$ does not depend on the basis used.

The map $\boldsymbol{X}_{t_\circ}^t$ is actually the one-parameter family of diffeomorphisms from V

to V generated by the vector field Ax. In this case, these are linear. Indeed, $X^t_{t_o}x_o$ is the curve that passes through the point $x_o \in V$ at $t = t_o$ and whose tangent vector is $AX^t_{t_o}x_o$.

We have seen then that if we know a set of n linearly independent solutions $\{u_i(t)\}$ we can construct any solution using the operator $X^t_{t_o}$ applied to the initial data of our choice.

How do we know in practice that a set of n solutions $\{u_i(t)\}$ is linearly independent? As we have seen, it is sufficient to see that these are linearly independent, as elements of V, at any time t. That is, the scalar

$$w(t) = \varepsilon(u_1(t), u_2(t), \ldots, u_n(t)) \tag{5.6}$$

is different from zero. This function is called the **Wronskian** of the system and satisfies a particularly simple equation whose solution, called the **Liouville formula**, is

$$w(t) = w(t_o)e^{\int_{t_o}^{t} tr(A(\tilde{t}))d\tilde{t}} \tag{5.7}$$

Proof:

$$\begin{aligned} \dot{w}(t) &= \varepsilon(\dot{u}_1, u_2, \ldots, u_n) + \varepsilon(u_1, \dot{u}_2, \ldots, u_n) + \cdots + \varepsilon(u_1, u_2, \ldots, \dot{u}_n) \\ &= \varepsilon(Au_1, u_2, \ldots, u_n) + \varepsilon(u_1, Au_2, \ldots, u_n) + \cdots + \varepsilon(u_1, u_2, \ldots, Au_n) \\ &= tr(A)w(t), \end{aligned} \tag{5.8}$$

whose solution is 5.7.

The Liouville formula is an independent demonstration of the result that $\{u_i(t)\}$ form a basis of V for each $t \in I$. Note that if $tr(A) \equiv 0$ the Wronskian is constant and can be useful to determine a solution in terms of others already known.

5.2 Inhomogeneous Linear System – Variation of Constants

Here we will deal with the system,

$$\frac{dx}{dt} = A(t)x + b(t), \tag{5.9}$$

where $b(t) : I \to V$ is integrable. We will see that if we know the operator $X^t_{t_o}$ corresponding to the homogeneous system

$$\frac{dx}{dt} = A(t)x, \tag{5.10}$$

Then we know all the solutions of system 5.9. The method we will use is called **variation of constants** and it is also useful for obtaining approximate solutions to

non-linear systems. The method consists of assuming that the solution of 5.9 will be of the form,

$$\varphi(t) = X^t_{t_o}\, c(t) \tag{5.11}$$

for some map $c(t) : I \to V$, differentiable. Note that if $c(t) = c \in V$, that is, a constant map, then $\varphi(t)$ satisfies 5.10. Substituting $\varphi(t)$ in 5.9 we get,

$$\begin{aligned} \tfrac{d}{dt}\dot{\varphi}(t) &= \left(\tfrac{d}{dt}X^t_{t_o}\right) c(t) + X^t_{t_o}\tfrac{d}{dt}c(t) \\ &= A(t)X^t_{t_o}\, c(t) + X^t_{t_o}\tfrac{d}{dt}c(t) \quad = \quad A(t)X^t_{t_o}\, c(t) + b(t). \end{aligned} \tag{5.12}$$

From which we see that for $\varphi(t)$ to be a solution, $c(t)$ must satisfy the equation $X^t_{t_o}\frac{d}{dt}c(t) = b(t)$.

But since $X^t_{t_o}$ is invertible and its inverse is $X^{t_o}_t$ we get

$$\frac{d}{dt}c(t) = X^{t_o}_t\, b(t). \tag{5.13}$$

Integrating we get

$$c(t) = \int_{t_o}^{t} X^{t_o}_{\tilde{t}}\, b(\tilde{t})\, d\tilde{t} + c(t_o) \tag{5.14}$$

or

$$\varphi(t) = X^t_{t_o}\left[\int_{t_o}^{t} X^{t_o}_{\tilde{t}}\, b(\tilde{t})\, d\tilde{t}\right] + X^t_{t_o}\,\varphi(t_o), \tag{5.15}$$

where $\varphi(t_o)$ is any initial condition.

Due to the existence theorem of solutions of the homogeneous system we know that $X^{t_o}_t$ exists and is differentiable $\forall\ t \in I$ and therefore $\varphi(t)$ also exists and is differentiable $\forall\ t \in I$, since $b(\tilde{t})$ is integrable. Since the initial data for $\varphi(t)$ can be given arbitrarily, and the solutions of 5.9 are unique, we conclude that 5.15 is the general solution of system 5.9.

5.3 Homogeneous Linear Systems: Constant Coefficients

The equation we will deal with here is

$$\frac{dx}{dt} = Ax, \tag{5.16}$$

where $x : I \subset \mathbb{R} \to V$ is a curve in V and A is a linear operator of V, $A : V \to V$.

We already saw in an exercise that $x(t) = e^{At}\, x_o$, $x_o \in V$ is a solution of 5.16 with initial condition $x(0) = x_o$. Note that since e^{At} is defined for all $t \in \mathbb{R}$, the solutions we have found are also defined in all $\mathbb{R}$.

Let $\tilde{x}(t)$ be any solution of 5.16 defined in an interval $\tilde{I} \subset \mathbb{R}$ and let $t_1 \in \tilde{I}$. Then the curve $x(t) = e^{A(t_1 - t)}\,\tilde{x}(t_1)$ is also a solution of 5.16 and $x(t_1) = \tilde{x}(t_1)$, but by the

uniqueness of the solutions of 5.16, $\boldsymbol{x}(t)$ and $\tilde{\boldsymbol{x}}(t)$ must coincide in all $\tilde{I}$ and $\boldsymbol{x}(t)$ is the maximum extension of $\tilde{\boldsymbol{x}}(t)$. Thus we see that through the exponential $e^{\boldsymbol{A}t}$ we obtain all the solutions of 5.16. In fact, we have shown that the one-parameter family of linear diffeomorphisms $g^t = e^{\boldsymbol{A}t}$ is globally defined $[\forall t \in \mathbb{R}, \forall \boldsymbol{x}_o \in V]$ and is the generator of the vector field $v(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$. [With the definition given in 5.1 in this case $\boldsymbol{X}^t_{t_o} = g^{t-t_o} = e^{\boldsymbol{A}(t-t_o)}$.]

What is the general form of this family? Or in other words, what is the functional form of the exponential map? For this it is convenient to take a fixed basis $\{\boldsymbol{u}_i\}$ (independent of t) in V and express the equation in terms of the n-tuple of components $\{x^i\}$ of $\boldsymbol{x}$. In this way we obtain,

$$\frac{dx^i}{dt} = \sum_{j=1}^{n} A^i{}_j\, x^j, \tag{5.17}$$

that is, an equation for the n-tuple $\{x^i\}$. The advantage of this change is that we can take a convenient basis, in particular the orthogonal basis of Lemma 2.2 (of Schur) in which the matrix $A^i{}_j$ is upper triangular. The price paid for the simplification is that, since the Schur basis is generally complex, both the n-tuple $\{x^i\}$ and the components of $\boldsymbol{A}$, $A^i{}_j$, are complex and therefore we have moved to a problem in C^n, or abstractly in the complexification of V.

For that basis we have that the nth component satisfies a particularly simple equation!,

$$\frac{dx^n}{dt} = A^n{}_n\, x^n \tag{5.18}$$

and therefore, $x^n(t) = e^{\lambda_n t} x^n_o$ (remember that the diagonal of an upper triangular matrix is composed of its eigenvalues), where x^n_o is the nth component of the initial condition at $t = o$. The equation for the (n-1)th component is,

$$\begin{aligned} \frac{dx^{n-1}}{dt} &= A^{n-1}{}_{n-1}\, x^{n-1} + A^{n-1}{}_n\, x^n & (5.19)\\ &= \lambda_{n-1}\, x^{n-1} + A^{n-1}{}_n\, e^{\lambda_n t} x^n_o. & (5.20) \end{aligned}$$

Using the formula (5.15) for the one-dimensional case, (and $\boldsymbol{A}$ constant) we get,

$$\begin{aligned} x^{n-1}(t) &= e^{\lambda_{n-1} t} x^{n-1}_o + e^{\lambda_{n-1} t} \int_o^t e^{-\lambda_{n-1} s} (A^{n-1}{}_n\, e^{\lambda_n s} x^n_o)\, ds & (5.21)\\ &= e^{\lambda_{n-1} t} x^{n-1}_o + e^{\lambda_{n-1} t} \int_o^t e^{(\lambda_n - \lambda_{n-1}) s}\, ds\; A^{n-1}{}_n\, x^n_o, & (5.22) \end{aligned}$$

and we see that if $\lambda_n - \lambda_{n-1} \neq o$ then the solution is a sum of exponentials, $x^{n-1}(t) = e^{\lambda_{n-1} t} x^{n-1}_o + \frac{A^{n-1}{}_n\, x^n_o}{\lambda_n - \lambda_{n-1}} (e^{\lambda_n t} - e^{\lambda_{n-1} t})$. If the eigenvalues coincide ($\lambda_n - \lambda_{n-1} = o$) then

a linear term in t appears, $x^{n-1}(t) = e^{\lambda_{n-1}t}(x_o^{n-1} + tA^{n-1}{}_n\, x_o^n)$. Knowing $x^{n-1}(t)$ we can now calculate, using again (5.15), $x^{n-2}(t)$ and so on for all the components of $\boldsymbol{x}(t)$ in this basis. The final result will be that the components are all sums of exponentials by polynomials. Indeed note that the integral of a polynomial of degree m by an exponential is also a polynomial of the same degree by the same exponential plus an integration constant (unless the exponential is zero, in which case the result is a polynomial of degree $m+1$).

Exercise: Prove this last statement by induction on the degree of the polynomial and the use of integration by parts, that is, use that $e^{at}P_m(t) = \frac{d}{dt}(\frac{e^{at}}{a}P_m(t)) - \frac{e^{at}}{a}\frac{d}{dt}P_m(t)$ and that $\frac{d}{dt}P_m(t)$ is a polynomial of order $m-1$.

Rearranging terms we see that the general solution will be of the form:

$$\boldsymbol{x}(t) = \sum_{i=1}^{m} e^{\lambda_i t}\boldsymbol{v}_i(t) \tag{5.23}$$

where the sum is over the different eigenvalues and where the vectors $\boldsymbol{v}_i(t)$ are polynomials in t.

An important property of the algebraic structure of this solution is manifested with the following lemma.

Lemma 5.1 *The vector spaces*

$$E_\lambda := \{f(t) : \mathbb{R} \mapsto V^C | f(t) = e^{\lambda t}P_m(t), \quad \textit{with } P_m(t) \textit{ a polynomial}\}$$

are linearly independent.

Proof: We must then prove that if we have a finite sum of elements of the E_{λ_i}, $\{\boldsymbol{u}_i\}$ and this sum is zero, then each of these vectors must be identically zero. That is

$$\sum_{i=1}^{m}\boldsymbol{u}_i = 0, \quad \boldsymbol{u}_i \in E_{\lambda_i} \quad \lambda_i \neq \lambda_j \quad \Rightarrow \boldsymbol{u}_i = 0. \tag{5.24}$$

We will prove it by induction on the number of elements in the sum. The case $m = 1$ is trivial. We will then assume that the statement is true for $m-1$ and we will prove it for m. We will also assume by contradiction that there is a sum of m non-zero elements that is zero, that is:

$$\sum_{i=1}^{m}\boldsymbol{u}_i = 0, \quad \boldsymbol{u}_i \in E_{\lambda_i}, \quad \lambda_i \neq \lambda_j, \quad \boldsymbol{u}_i \neq 0. \tag{5.25}$$

dividing by $e^{\lambda_1 t}$ we get

$$0 = S(t) := e^{-\lambda_1 t}\sum_{i=1}^{m} \boldsymbol{u}_i = P_{m_1}(t) + \sum_{i=2}^{m} e^{(\lambda_i-\lambda_1)t} P_{m_i}(t). \tag{5.26}$$

Taking $m_1 + 1$ derivatives of $S(t)$, where m_1 is the order of the polynomial of the first term we get:

$$0 = \frac{d^{m_1+1}}{dt^{m_1+1}} S(t) = \sum_{i=2}^{m} e^{(\lambda_i-\lambda_1)t} \tilde{P}_{m_i}(t). \tag{5.27}$$

where it is easy to see that the polynomials $\tilde{P}_{m_i}(t)$ are of the same degree as the originals. We are thus under the inductive hypothesis and therefore since these polynomials are non-zero we have a contradiction ♠

Let us now see that in fact each of the terms in this sum is a solution of equation 5.16. Given any $\lambda \in C$ consider the vector space of functions from $\mathbb{R} \to V^C$ of the form $e^{\lambda t}\boldsymbol{w}(t)$, where $\boldsymbol{w}(t)$ is a polynomial [1] in t. This space is invariant under the action of $\frac{d}{dt}$, since taking its derivative we get a similar expression, that is, the product of the same exponential by another polynomial. On the other hand it is invariant under the action of A, since as A does not depend on t its action on any $e^{\lambda t}\boldsymbol{w}(t)$ gives us another similar expression. Therefore, the action of $\frac{d}{dt} - A$ will also keep us in the same space. Thus we see that if we have a sum of terms with this structure and with different values of λ, as is the case in the previous equation 5.23, then if the application of $\frac{d}{dt} - A$ gives us zero this means that the application of $\frac{d}{dt} - A$ to each term of the sum must give zero. That is, each term is a solution of equation 5.16! Thus we conclude that each term in the sum 5.16 is of the form $e^{\boldsymbol{A}t}\boldsymbol{v}_0$ for some $\boldsymbol{v}_0 \in V^C$. Let us now consider the subset W_λ of V^C, such that if $\boldsymbol{v}_0 \in W_\lambda$ then $e^{\boldsymbol{A}t}\boldsymbol{v}_0 = e^{\lambda t}\boldsymbol{v}(t)$. It is clear that the only non-trivial subspaces will be those with $\lambda = \lambda_i$, where $\{\lambda_i\}$, $i = 1..d$ are the eigenvalues of A. Therefore we must consider only these subsets. Since this relationship is linear, W_λ is a subspace of V^C. Since $e^{\boldsymbol{A}t}A\boldsymbol{v}_0 = Ae^{\boldsymbol{A}t}\boldsymbol{v}_0 = e^{\lambda t}A\boldsymbol{v}(t) = e^{\lambda t}\boldsymbol{v}'(t)$ and $\boldsymbol{v}'(t)$ is also polynomial in t, we see that W_λ is an invariant subspace of A. While using a basis of W_λ in which the restriction to that space of A is upper triangular, we see that λ is the only eigenvalue that such a restriction can have. Finally, since every solution of 5.16 has the form 5.23, the existence theorem of solutions assures us that any initial data, that is, any element of V^C, can be expressed as a linear combination of elements in W_{λ_i}. Indeed, let $\boldsymbol{v}_0$ be any element of V^C, from the existence theorem of solutions, we then have a unique solution of 5.16 $\boldsymbol{x}(t)$ with $\boldsymbol{x}(0) = \boldsymbol{v}_0$. But then

$$\begin{aligned} \boldsymbol{x}(t) &= \sum_{i=1}^{m} e^{\lambda_i t}\boldsymbol{v}_i(t) \\ &= \sum_{i=1}^{m} e^{At}\boldsymbol{v}_{0i}, \end{aligned} \tag{5.28}$$

[1] That is, a linear combination of vectors in V^C with polynomial coefficients in t.

for some set of vectors $\{v_{oi}\} \in W_{\lambda_i}$. Evaluating this expression at $t = 0$, we obtain $v_o = \sum_{i=1}^{m} v_{oi}$ and we see that the W_{λ_i} generate V^C. On the other hand, the uniqueness of the solutions implies that no element of a given W_{λ_i} can be written as a linear combination of elements of the other W_λ [otherwise the same initial data would give rise to two different solutions (since their functional dependence would be different)]. Indeed, let $0 = v_{o1} + \cdots + v_{os}$ be any linear combination of elements of W_{λ_i}, $i = 1..s$, we see that each of them must be zero. By the uniqueness of the solutions to the ordinary equations, we then have $0 = \sum_{i=1}^{m} e^{At} v_{oi} = \sum_{i=1}^{m} e^{\lambda_i t} v_i(t)$ as seen earlier, each element in the last sum must be zero and therefore, evaluating these at $t = 0$, we obtain $v_{oi} = 0 \ \forall\, i = 1..s$. Summarizing the above, we have,

Theorem 5.3 *Given an operator $A : V^C \to V^C$, there exists a set of invariant subspaces of A, $\{W_{\lambda_i}\}$, where λ_i are the eigenvalues of A such that:*

a) $V^C = W_{\lambda_1} \oplus W_{\lambda_2} \oplus \cdots W_{\lambda_d}$

b) The only eigenvalue of the restriction of A in $\{W_{\lambda_i}\}$ is λ_i.

We conclude the study of the solutions of equation 5.16, that is, of the function e^{At} with a detailed description of their form obtained using the matrix representation of A in the basis where it acquires the Jordan canonical form. That is, for every invariant subspace W_{λ_i} the restriction of A in this subspace has the form,

$$A = \lambda_i + \Delta_i, \tag{5.29}$$

where the numbers λ_i are the eigenvalues corresponding to A –generally complex, and therefore we are now considering A as an operator from C^n to C^n– and Δ_i is a matrix whose only non-zero components are ones and can only be in the immediate upper diagonal. The matrix Δ_i has the important property of being nilpotent, that is, there exists an integer m_i less than or equal to the multiplicity of λ_i, such that $\Delta_i^{m_i} = 0$.

Since $\lambda_i I$ and Δ_i commute, we have that

$$e^{t\lambda_i I + t\Delta_i} = e^{t\lambda_i I} e^{t\Delta_i}, \tag{5.30}$$

but $e^{t\lambda_i I} = e^{t\lambda_i} I$ and $e^{t\Delta_i} = \sum_{j=0}^{m_i - 1} \dfrac{(t\Delta_i)^j}{j!}$, that is, a finite sum.

Since the exponential of A is a sum of powers of A, it follows that the invariant spaces of this are the same as those of A and therefore, $e^{tA}|_{B_i} = e^{tA_i}$. From this, we conclude that the matrix e^{tA} is of the form

$$e^{tA} = \begin{pmatrix} k_0 & & & \\ & k_1 & & \\ & & \ddots & \\ & & & k_p \end{pmatrix} \tag{5.31}$$

with each k_i a square sub-matrix, the $k_0 = diag(e^{t\lambda_1}, \ldots, e^{t\lambda_n})$ and if $i \neq 0$

$$k_i = e^{t\lambda_i} \begin{bmatrix} 1 & t & t^2/2 & \cdots & \cdots & t^{n-1}/(n-1)! \\ & 1 & t & \cdots & \cdots & \vdots \\ & & 1 & \cdots & \cdots & \vdots \\ & & & \ddots & \cdots & t^2/2 \\ & 0 & & & \ddots & t \\ & & & & & 1 \end{bmatrix}, \tag{5.32}$$

where the number of rows or columns is the same as those of the J_i corresponding to A in the Jordan composition, this is less than or equal to the multiplicity of λ_i as a root of the characteristic polynomial.

The basis in which A has the Jordan canonical form is generally complex, it is actually a basis in C^n. If we wish to use a real basis, which is possible since we start with A as an operator from $\mathbb{R}^n$ to $\mathbb{R}^n$, and we make the corresponding transformation, the matrix e^{At} will undergo the corresponding similarity transformation. Since this transformation is independent of time, although the components of e^{At} will not have the previous form, they will be sums of exponential terms by polynomials in t. Therefore, each component of the general solution of 5.16 will be of the form,

$$x^i(t) = \sum_{p=1}^{q} \left\{ (e^{t\lambda_p} + e^{t\bar{\lambda}_p}) P_p^i(t) + i(e^{t\lambda_p} - e^{t\bar{\lambda}_p}) Q_p^i(t) \right\} \tag{5.33}$$

where q is the number of distinct eigenvalues –counting complex pairs as one– and P_p^i, Q_p^i are polynomials in t whose degree is less than or equal to the multiplicity with which the eigenvalue λ_p appears as a root of the characteristic polynomial. This information is useful in two ways. Practically, because although the method of constructing the solution using a basis in which A has the Jordan canonical form is straightforward, for large dimension systems it becomes cumbersome. In some cases, it is convenient to calculate the eigenvalues and their multiplicity and then assume a solution of the form 5.33 and calculate the polynomials P_p^i, Q_p^i.

This method is also useful in the sense that it allows us to know the global behavior of the solutions. For example, we see that if the real part of the eigenvalues is not positive and those whose real part is zero do not appear repeated, then all solutions are bounded [There exists $C > 0$ such that $||\boldsymbol{x}(t)|| < C$.] If in addition, all have negative real part, then all solutions tend asymptotically to the trivial solution [$\lim_{t\to\infty} \boldsymbol{x}(t) = 0$].

We will now analyze in detail the case where all eigenvalues are distinct. Note that although this is the generic case –in the sense that if a system has coincident eigenvalues, then there are arbitrarily small modifications to it that make the eigenvalues distinct– systems with coincident eigenvalues do appear in physics. This case also contemplates the situation of a general system where the initial data belongs to one of the one-dimensional subspaces B_i.

Since the operator A is real, its eigenvalues will be real or complex conjugates [if $det(A-\lambda I)=0$ then $det(\overline{A-\lambda I})=det(A-\bar{\lambda}I)=0$]. If λ_i is real, then its eigenvector can be chosen to be real. Indeed, if $(A-\lambda I)\boldsymbol{u}_i=0$, then also $(A-\lambda I)\bar{\boldsymbol{u}}_i=0$, but the roots are simple and therefore each λ_i has only one eigenvector –modulo a complex scalar–, that is $\bar{\boldsymbol{u}}_i=\alpha\boldsymbol{u}_i$ $\alpha\in C$. Choosing $\boldsymbol{v}_i=\boldsymbol{u}_i+\bar{\boldsymbol{u}}_i=(1+\alpha)\boldsymbol{u}_i$ we obtain a real eigenvector.

In this case, the component x^i_0 of $\boldsymbol{x}_0$ in the eigenbasis in the direction $\boldsymbol{v}_i$ evolves as

$$x^i(t)=x^i_0 e^{\lambda_i t}, \tag{5.34}$$

that is, it grows or decays exponentially with time according to the sign of λ_i.

If the eigenvalue λ_i is complex, then we can choose its eigenvector $\boldsymbol{u}_i$ such that it is the complex conjugate of the chosen one corresponding to $\bar{\lambda}_i$. This pair of eigenvectors generates a 2-dimensional complex subspace. If $\boldsymbol{x}_0$ belongs to this subspace and is real, then it will have the form $\boldsymbol{x}_0=a(\boldsymbol{u}_i+\bar{\boldsymbol{u}}_i)-ib(\boldsymbol{u}_i-\bar{\boldsymbol{u}}_i)$ with a and b real, that is $\boldsymbol{x}_1=\boldsymbol{u}_i+\bar{\boldsymbol{u}}_i$ and $\boldsymbol{x}_2=i(\boldsymbol{u}_i-\bar{\boldsymbol{u}}_i)$ form a real basis.

How do these vectors change if we apply the operator e^{At} to them? That is, what are the solutions of the equation $\dot{\boldsymbol{x}}=A\boldsymbol{x}$ with initial conditions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$? Calling these $\boldsymbol{x}_1(t)$ and $\boldsymbol{x}_2(t)$ respectively and using that $e^{At}\boldsymbol{u}_i=e^{\lambda_i t}\boldsymbol{u}_i$ we obtain,

$$\begin{aligned}\boldsymbol{x}_1(t) &= e^{\alpha_i t}(\boldsymbol{x}_1\cos w_i t-\boldsymbol{x}_2\sin w_i t)\\ \boldsymbol{x}_2(t) &= e^{\alpha_i t}(\boldsymbol{x}_2\cos w_i t+\boldsymbol{x}_1\sin w_i t),\end{aligned} \tag{5.35}$$

with $\lambda_i=\alpha_i+i\,w_i$.

We see that the action of the operator e^{At} in this case is to dilate the vectors by a factor of $e^{\alpha_i t}$ and to rotate them by an angle $w_i t$.

5.4 Problems

Problem 5.1 *Given a set of functions $\{f_i(t)\}$, $i=1,..,n$, vectors are defined as $\boldsymbol{u}_i(t):=(f_i(t),f_i^{(1)}(t),\dots f^{(n-1)})$ and the Wronskian of the system as $W(\{f_i\})(t):=\varepsilon(\boldsymbol{u}_1(t),\boldsymbol{u}_2(t),\dots,\boldsymbol{u}_n(t))$. If the Wronskian of a set of functions does not vanish, then the functions are linearly independent, that is, no non-trivial linear combination of them (with constant coefficients) vanishes. The converse is not true. Calculate the Wronskian of the following sets:*

a) $\{4,t\}$
b) $\{t,3t,t^2\}$
c) $\{e^t,te^t,t^2e^t\}$
d) $\{\sin(t),\cos(t),\cos(2t)\}$
e) $\{1,\sin(t),\sin(2t)\}$

Problem 5.2 *Decide if the following set of functions is linearly dependent or not. Then calculate the Wronskian.*

$$f_1(t)=\begin{cases}0, & 0\le x\le 1/2\\ (x-1/2)^2, & 1/2\le x\le 1\end{cases} \tag{5.36}$$

$$f_2(t) = \begin{cases} (x - 1/2)^2, & 0 \le x \le 1/2 \\ 0, & 1/2 \le x \le 1 \end{cases} \tag{5.37}$$

Problem 5.3 *Use the theory of ordinary differential equations to prove the following identities:*

a)

$$e^{sA} e^{tA} = e^{(s+t)A}, \tag{5.38}$$

b)

$$e^A e^B = e^{A+B}, \quad \textit{if and only if } [A, B] := AB - BA = 0. \tag{5.39}$$

Problem 5.4 *Graph the vector fields corresponding to the following systems and some typical sets of solutions.*

a)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.40}$$

b)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.41}$$

c)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.42}$$

d)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.43}$$

e)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.44}$$

f) (compare the solutions with those of point b)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.45}$$

g)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.46}$$

h)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1+i & 0 \\ 0 & 1-i \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.47}$$

i)

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1+i & 0 \\ 1 & 1-i \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{5.48}$$

Problem 5.5 *Reduce these equations to first-order systems and find the general solution of the equations:*

a) $\frac{d^3x}{dt^3} - 2\frac{d^2x}{dt^2} - 3\frac{dx}{dt} = 0$

b) $\frac{d^3x}{dt^3} + 2\frac{d^2x}{dt^2} + \frac{dx}{dt} = 0$

c) $\frac{d^3x}{dt^3} + 4\frac{d^2x}{dt^2} + 13\frac{dx}{dt} = 0$

d) $\frac{d^4x}{dt^4} + 4\frac{d^3x}{dt^3} + 8\frac{d^2x}{dt^2} + 8\frac{dx}{dt} + 4 = 0$

Problem 5.6 (Newton's Law) *Consider the equation* $\frac{d^2x}{dt^2} + f(x) = 0$.

a) Prove that $\frac{1}{2}\dot{x}^2 + \int_{x_0}^{x} f(s)ds$ *is a first integral.*

b) Find the first integral of $\frac{d^2x}{dt^2} - x + x^2/2 = 0$.

c) Graph the corresponding vector field and some of its solutions. Find its stationary solutions (or equilibrium points) and study their neighborhoods by linearizing the equation at these points.

Problem 5.7 *Study the following system*

$$\begin{aligned} \dot{x}_1 &= x_1 - x_1x_2 - x_2^3 + x_3(x_1^2 + x_2^2 - 1 - x_1 + x_1x_2 + x_2^3) \\ \dot{x}_2 &= x_1 - x_3(x_1 - x_2 + x_1x_2) \\ \dot{x}_3 &= (x_3 - 1)(x_3 + 2x_3x_2^2 + x_3^3) \end{aligned} \tag{5.49}$$

a) Find the equilibrium points.

b) Show that the planes $x_3 = 0$ *and* $x_3 = 1$ *are invariant sets, (that is, solutions that start in them never leave them).*

c) Consider the invariant set $x_3 = 1$ *and see if it has periodic solutions.*

CHAPTER 6

STABILITY

Stationary or equilibrium solutions are very important in physics because many systems (especially those with dissipation) behave in such a way that the solution approaches these solutions in their evolution.

Another particularity that makes them important is the fact that they are simply given by the points where the vector field vanishes, and therefore at most one must solve an algebraic equation to find them, in contrast to the general case where we must solve a differential equation. In some cases, it is even possible to infer that the vector field must vanish, such as if the manifold is a two-dimensional sphere, since any continuous vector field defined on it is such that there are at least two points where it vanishes, and therefore in this case there will always be stationary solutions.

Exercise: Convince yourself that this is so.

From the above, it follows that it is important to study the behavior of solutions that originate from initial data close to a stationary solution. For this, we define the following concepts of stability.

Definition: Let v be a vector field in M and let $p \in M$ such that $v(p) = 0$. We will say that the stationary solution $\gamma(t) \equiv p$ is **stable** if given a neighborhood U_p of p there exists another neighborhood V_p of p such that any solution $\sigma(t)$ with $\sigma(0) \in V_p$ satisfies $\sigma(t) \in U_p \quad \forall \ t \geq 0$. [See figure 6.1.]

Definition: We will say that the previous solution is **asymptotically stable** if it is stable and also if $\sigma(0) \in V_p$ then $\lim_{t \to +\infty} \sigma(t) = p$.

If a stationary solution is not stable, then it has little physical interest since the slightest perturbation will *move it far* from it.

Examples:
a) The bacterial growth equation $\dot{x} = ax - bx^2$ has stationary solutions $x(t) \equiv 0$ and

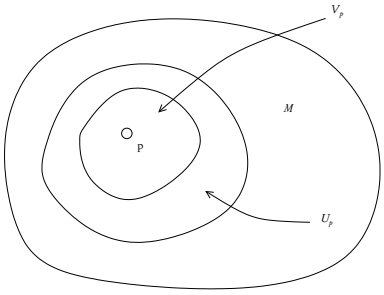

Figure 6.1: Stability.

$x(t) \equiv \frac{a}{b}$. Since the general solution is,

$$x(t) = \frac{x(0)e^{at}}{1 + \frac{bx(0)}{a}(e^{at} - 1)}, x(0) \geq 0 \tag{6.1}$$

it is clear that $x(t) = 0$ is not a stable solution (we will call these unstable) and that $x(t) \equiv \frac{a}{b}$ is asymptotically stable. If we contaminate a culture container with a single bacterium, this is enough for it to reproduce [1] until it reaches (asymptotically) the concentration $\frac{a}{b}$. If we now remove or add some bacteria to change their concentration, the bacteria will reproduce or annihilate until they reach again *the stable concentration* $\frac{a}{b}$.

b) The equation $\dot{x} = Ax$ has among its stationary solutions the one given by $x(t) \equiv 0$. What are the others? These will be stable when the eigenvalues of A, λ_i, satisfy $\Re(\lambda_i) \leq 0$, $\lambda_i \neq \lambda_j$ or $\Re(\lambda_i) < 0$ if $\lambda_i = \lambda_j$, where $\Re$ indicates the real part. This is clear since the general solution is $x(t) = e^{At}x(0)$ and the mentioned condition implies that $||e^{At}||_{\mathscr{L}} < C, C > 0 \, \forall t \geq 0$. In this case, we can take as $U_{p=0} = \{x \in \mathbb{R}^n | \, |x| < \epsilon\}$ and as $V_{p=0} = \{x \in \mathbb{R}^n | \, |x| < \frac{\epsilon}{C}\}$. If it is also fulfilled that $\Re(\lambda_i) < 0 \, \forall i = 1, ..., n$ then $x(t) \equiv 0$ is asymptotically stable.

The following theorem provides a very practical tool to know when a stationary solution is stable or not.

Theorem 6.1 (Stability) *Let $\gamma(t) \equiv p$ be a stationary solution of $v \in TM$, that is $v(p) = 0$, and let $A : T_pM \to T_pM$ be defined by,*

$$Ax \equiv \frac{d}{ds} v(\sigma_x(s))|_{s=0}, \tag{6.2}$$

[1] Note that in the real process we do not have a continuous variable and therefore in it the perturbations are at least one bacterium, which makes it possible to have sterile containers and therefore the mathematical instability does not manifest in reality.

where $\sigma_x(s)$ is a curve in M satisfying: $\sigma_x(0) = p$ and $\frac{d}{ds}\sigma_x|_{s=0} = \boldsymbol{x} \in T_pM$, that is, it is any smooth curve that passes through p when $s = 0$ and at that point has as tangent to $\boldsymbol{x} \in T_pM$. If $\Re(\lambda_i) < 0$, where λ_i are the eigenvalues of A, then $\gamma(t) \equiv p$ is asymptotically stable. If any λ_i has a positive real part then $\gamma(t) = p$ is unstable.

Exercises:
a) Show that A is really a linear operator.

b) Show that $A\boldsymbol{x} = [\boldsymbol{v}, \tilde{\boldsymbol{x}}]|_p$, where $\tilde{\boldsymbol{x}}$ is any vector field such that $\tilde{\boldsymbol{x}}|_p = \boldsymbol{x}$. Hint: do not forget that $\boldsymbol{v}|_p = 0$.

Examples:
a) Consider the bacterial growth equation at $x = 0$ and $x = \frac{a}{b}$. For the first case, we take $\sigma_{\delta x}(s) = \delta x s$ and obtain,

$$A\delta x = \frac{d}{ds}(a\delta x s - b\,(\delta x)^2 s^2)|_{s=0} = a\delta x, \tag{6.3}$$

which shows that $x = 0$ is an unstable solution since $a > 0$. For the second case, we take $\sigma_{\delta x} = \frac{a}{b} + \delta x s$ then

$$A\delta x = \frac{d}{ds}(\frac{a^2}{b} + a\delta x s - b(\frac{a}{b} + \delta x s)^2)|_{s=0} = a\delta x - 2a\delta x = -a\delta x, \tag{6.4}$$

which shows that it is asymptotically stable.

b) Physical pendulum with friction: $\ddot{\theta} = -\sin\theta - k\dot{\theta}$, $k > 0$. The vector field in this case is given by,

$$\begin{array}{rcl} \dot{\theta} & = & z \\ \dot{z} & = & -\sin\theta - kz, \end{array} \tag{6.5}$$

that is, the vector with components $(z, -\sin\theta - kz)$ in the phase space which in this case is a cylinder, $z \in [-\infty, +\infty]$, $\theta \in [0, 2\pi]$.

This vector vanishes only at $p_1 = (\theta = 0, z = 0)$ and $p_2 = (\theta = \pi, z = 0)$. For the first stationary solution, using $\sigma(s)$ such that,

$$\phi \circ \sigma(s) = \begin{pmatrix} \delta\theta\, s \\ \delta z\, s \end{pmatrix}, \tag{6.6}$$

we obtain,

$$A\begin{pmatrix} \delta\theta \\ \delta z \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & -k \end{pmatrix}\begin{pmatrix} \delta\theta \\ \delta z \end{pmatrix}. \tag{6.7}$$

In this case, the eigenvalue equation is $\lambda^2 + k\lambda + 1 = 0$ from which we obtain, $\lambda_\pm = \frac{-k \pm \sqrt{k^2 - 4}}{2}$, which implies $\Re(\lambda_\pm) < 0$ and thus stability.

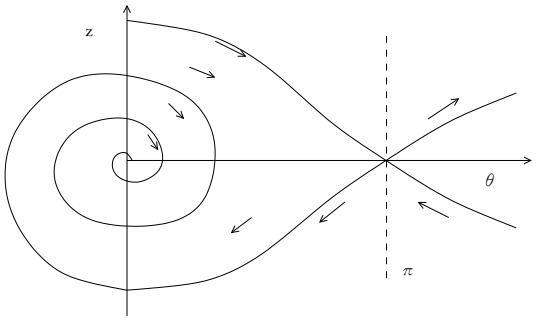

Figure 6.2: Physical pendulum with friction.

For the second stationary solution, we take $\sigma(s)$ such that,

$$\phi \circ \sigma(s) = \begin{pmatrix} \delta\theta s + \pi \\ \delta z \end{pmatrix} \tag{6.8}$$

and obtain

$$A \begin{pmatrix} \delta\theta \\ \delta z \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -k \end{pmatrix} \begin{pmatrix} \delta\theta \\ \delta z \end{pmatrix}. \tag{6.9}$$

with eigenvalues $\lambda_{\pm} = \dfrac{-k \pm \sqrt{k^2+4}}{2}$, which implies $\Re(\lambda_+) > 0$ and thus instability.

From the examples, it is seen that this theorem has broad application. Since stability is a local notion, to facilitate the demonstration we will take $M = I\!R^n$ and a Cartesian coordinate system with origin at the stable solution to be considered. We will use several previous results that we prove below.

Theorem 6.2 (Lyapunov) *Let $A : I\!R^n \to I\!R^n$ be a linear operator whose eigenvalues have a negative real part. Then there exists a tensor of type $(2,0)$, $\rho(\cdot,\cdot)$, symmetric and positive definite [$\rho(w,v) = \rho(v,w)$ and $\rho(w,w) \geq 0$ $(= \text{ if } f \ w = 0)$] such that,*

$$Ax(\rho(x,x)) < 0 \ \ \forall \ x \neq 0,$$

that is, the derivative of $\rho(x,x)$ in the direction Ax is negative.

The geometric interpretation of this condition is that $\rho(x,x)$ defines a norm whose level surfaces are such that on each of these the vector Ax at $p = x$ points inward. See figure.

Proof: If all the eigenvalues of A are distinct, then there exists a Jordan basis $\{u_i\}$, $i = 1,...,n$ and the corresponding co-basis $\{\theta^i\}$, $i = 1,...,n$ (with $\theta^i(u_j) = \delta^i_j$) such that $A = \sum_{i=1}^n \lambda_i u_i \theta^i$. In this case, let $\rho = \sum_{i=1}^n \theta^i \bar{\theta}^i$. If $z = \sum_{i=1}^n z^i u_i$ and

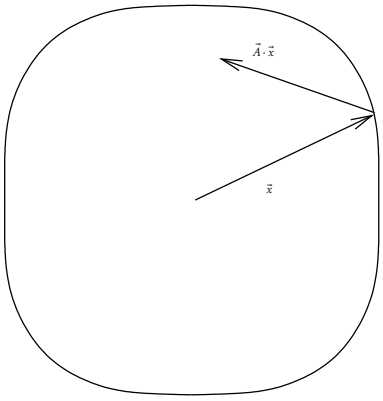

Figure 6.3: The norm $\rho(x,x)$.

$y=\sum_{i=1}^{n} y^i \boldsymbol{u}_i \in C^n$, then $\rho(z,y)=\sum_{i=1}^{n'} z^i y^i + \sum_{n'+1}^{(n-n')/2}(z^i\bar{y}^i+\bar{z}^i y^i)$ where we have separated the sum into that of the real eigenvectors and that of the complex conjugates. It is easy to see that the $\rho(,)$ obtained by restricting z and y to $\mathbb{R}^n$ is the usual Cartesian norm in the corresponding real basis. Now let's calculate $Ax(\rho(x,x))$.

$$\begin{aligned} Ax(\rho(x,x)) &= \lim_{\varepsilon\to 0}\frac{\rho(x+\varepsilon Ax, x+\varepsilon Ax)-\rho(x,x)}{\varepsilon} \\ &= \rho(Ax,x)+\rho(x,Ax) \\ &= 2\sum_{i=1}^{n}(\Re\lambda_i)x^i x^i < 0. \end{aligned} \tag{6.10}$$

Thus, we have proved the theorem for the diagonalizable case. The case where A is not is more complicated, and for this, we will use the following lemmas.

Lemma 6.1 *Given $\epsilon > 0$ there exists a basis $\{\boldsymbol{u}_i\}$ such that in that basis*

$$A = diag.(\lambda_1,...,\lambda_n)+\epsilon\Delta,$$

with Δ a matrix with non-zero components (and equal to one) at most on the upper diagonal, that is,

$$A=\begin{pmatrix} \lambda_1 & \varepsilon & 0 & . & . & . & 0 \\ 0 & \lambda_2 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & . & . & . & . & . & \varepsilon \\ 0 & . & . & . & . & 0 & \lambda_r \end{pmatrix}. \tag{6.11}$$

Proof: It is a simple rescaling of the Jordan basis of A. For example, in C^3 if $\{\tilde{\boldsymbol{u}}_i\}$ is

such that in it,

$$A = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}. \tag{6.12}$$

defining $\boldsymbol{u}_1 = \frac{\tilde{\boldsymbol{u}}_1}{\epsilon^2}$, $\boldsymbol{u}_2 = \frac{\tilde{\boldsymbol{u}}_2}{\epsilon}$, and $\boldsymbol{u}_3 = \tilde{\boldsymbol{u}}_3$, we see that in this new basis,

$$A = \begin{pmatrix} \lambda & \varepsilon & 0 \\ 0 & \lambda & \varepsilon \\ 0 & 0 & \lambda \end{pmatrix}. \tag{6.13}$$

$[A\boldsymbol{u}_1 = A\frac{\tilde{\boldsymbol{u}}_1}{\epsilon^2} = \frac{\lambda\tilde{\boldsymbol{u}}_1}{\epsilon^2} = \lambda\boldsymbol{u}_1, A\boldsymbol{u}_2 = A\frac{\tilde{\boldsymbol{u}}_2}{\epsilon} = \frac{\tilde{\boldsymbol{u}}_1 + \lambda\tilde{\boldsymbol{u}}_2}{\epsilon} = \epsilon\boldsymbol{u}_1 + \lambda\boldsymbol{u}_2$, etc.].

Lemma 6.2 *The set of positive symmetric tensors is open in the set of symmetric forms, that is, if ρ_0 is symmetric and positive and ρ_1 is symmetric, then there exists $\epsilon > 0$ such that $\rho_0 + \epsilon\,\rho_1$ is also positive.*

Proof: Let B_1 be the unit sphere with respect to some norm in $\mathbb{R}^n$ and consider $\rho_0(\boldsymbol{x},\boldsymbol{x}) : B_1 \to \mathbb{R}^+$. Since B_1 is compact, ρ_0 reaches its minimum there, α_{min}. Similarly, ρ_1 reaches its maximum there, which we will call β_{max}. Taking $\varepsilon < \frac{\alpha_{min}}{\beta_{max}}$ satisfies the requirement.

Example: In $\mathbb{R}^2$ let $\rho_0((x_1,y_1),(x_2,y_2)) = x_1x_2 + y_1y_2$ and $\rho_1((x_1,y_1),(x_2,y_2)) = x_1y_2 + x_2y_1$, then $\rho_0 + \varepsilon\,\rho_1$ is positive if $|\varepsilon| < 2$.

To complete the proof of Lyapunov's theorem we will take

$$\rho = \sum_i \theta^i \otimes \bar{\theta}^i,$$

where $\{\theta^i\}$ is the co-basis found in lemma 6.1 with $\varepsilon > 0$ to be determined. Then,

$$\begin{aligned} -A\boldsymbol{x}(\,\rho(\boldsymbol{x},\boldsymbol{x})) &= -\rho(A\boldsymbol{x},\boldsymbol{x}) - \rho(\boldsymbol{x},A\boldsymbol{x}) \\ &= -2\textstyle\sum_{i=1}^{n} \Re(\lambda_i)x^i x^i - 2\varepsilon \sum_{i=1}^{n-1} f_i x^i x^{i+1}, \end{aligned} \tag{6.14}$$

with $f_i = 1$ or 0 depending on whether there is ε or not in that place of the immediate upper diagonal. The first term is a positive definite form ρ_0 evaluated in $\boldsymbol{x}$ in both entries. The second is a symmetric form $\varepsilon\,\rho_1$ evaluated in $\boldsymbol{x}$ in both entries. The second lemma tells us that taking ε small $\rho_0 + \varepsilon\,\rho_1$ is also positive definite. This concludes the proof of the theorem.

Note now that not only the ρ we have found is symmetric and positive definite and therefore a norm in $\mathbb{R}^n$, but also $-Ax(\rho(x,x))$ is positive definite and defines a norm [since $-Ax(\rho(x,x)) = -\rho(Ax,x) - \rho(x,Ax)$]. But we have already seen that in $\mathbb{R}^n$ all norms are equivalent and therefore there will exist $\gamma > 0$ such that

$$-\frac{1}{2\gamma}\rho(x,x) \leq Ax(\rho(x,x)) \leq -2\gamma\rho(x,x). \tag{6.15}$$

This is the result we will use for the proof of the stability theorem given below.

Proof of the Stability Theorem: We have seen that if $\Re(\lambda_i) < 0$ then there exists a constant $\gamma > 0$ and a symmetric positive definite form $\rho(\ ,\ )$ such that

$$Ax(\rho(x,x)) \leq -2\gamma\rho(x,x). \tag{6.16}$$

Applying the vector field that defines the equation, $v(x)$ to $\rho(x,x)$ we get,

$$v(x)(\rho(x,x)) = Ax(\rho(x,x)) + O(\rho(x,x)^{3/2}), \tag{6.17}$$

where we have assumed v is differentiable and therefore $v(x) = Ax + O(|x|^2)$.

Example: $v(x) = (ax + bx^2)\frac{\partial}{\partial x}$ and $\rho(x,x) = \alpha x^2$ then,

$$\begin{aligned} v(x)(\rho(x,x)) &= ax\tfrac{\partial}{\partial x}\rho(x,x) + 2b\alpha x^3 \\ &= ax\tfrac{\partial}{\partial x}\rho(x,x) + \tfrac{2b}{\alpha^{1/2}}\rho(x,x)^{3/2} \end{aligned} \tag{6.18}$$

If x is sufficiently small, [$|x| < C$ for some $C > 0$] it holds that $O(\rho(x,x)^{3/2}) < \gamma(\rho(x,x))$ and therefore we have that

$$v(x)(\rho(x,x)) \leq -2\gamma\rho(x,x) + \gamma\rho(x,x) \leq -\gamma\rho(x,x). \tag{6.19}$$

Let $\varphi(t)$ be a solution with initial data sufficiently close to $x = 0$, that is $\dot{\varphi}(t) = v(\varphi(t))$, $|\varphi(0)| < C$. Defining [2]

$$r(t) = \ln \rho(\varphi(t), \varphi(t)) \tag{6.20}$$

and differentiating with respect to t we get,

$$\begin{aligned} \dot{r}(t) &= 2\frac{\rho(\dot{\varphi}(t),\varphi(t))}{\rho(\varphi(t),\varphi(t))} = \frac{v(\varphi(t))(\rho(\varphi(t),\varphi(t)))}{\rho(\varphi(t),\varphi(t))} \\ &\leq -\gamma. \end{aligned} \tag{6.21}$$

Therefore $r(t) \leq r(0) - \gamma t$, and integrating we conclude that

$$\rho(\varphi(t),\varphi(t)) \leq \rho(\varphi(0),\varphi(0))e^{-\gamma t}, \tag{6.22}$$

[2] By uniqueness of the solution $\varphi(t) \neq 0$ and therefore $\rho(\varphi(t),\varphi(t)) > 0$ and the logarithm is well defined.

which implies that $\varphi(t) \underset{t\to\infty}{\to} 0$ and concludes the proof of the theorem. [3]

It is instructive to observe that the proof is based on the construction of a norm, $\rho(x,x)$, especially adapted to the problem in the sense that it assures us that on its level surfaces the vector $v(x)$ (if x is sufficiently small) points inward.

6.1 Problems

Problem 6.1 (The Volterra-Lotka Equation) *Consider the system:*

$$\begin{aligned} \dot{x}_1 &= (a - b x_2) x_1 \\ \dot{x}_2 &= -(c - f x_1) x_2, \end{aligned} \tag{6.23}$$

$a, b, c, f \geq 0$.

a) Perform a coordinate and time transformation to bring it to the form,

$$\begin{aligned} \dot{x}_1 &= (1 - x_2) x_1 \\ \dot{x}_2 &= -(e - x_1) x_2 \end{aligned} \tag{6.24}$$

b) Plot the vector field and see that there are no non-trivial first integrals in the quadrants where at least one of the coordinates is negative.

c) Find the equilibrium solutions and determine which are stable and which are not.

d) Examine the positive quadrant using the transformation: $x_1 = e^{q_1}$, $x_2 = e^{q_2}$ and see that the quantity $f(q_1, q_2) := e q_1 + q_2 - (e^{q_1} + e^{q_2})$ is an integral of motion. Use this information to infer that in this quadrant the trajectories remain in bounded regions.

e) Examine the linearized equations around the equilibrium solution in the positive quadrant and determine the frequency of oscillations of the variations near equilibrium that the system would have.

Note, this equation describes the population of two competing species. What we have just seen is that the species do not grow indefinitely nor disappear. This last point tells us that the approximation is not very good... Note that x_2 represents a predator that depends exclusively on the prey x_1 for its subsistence, since if the prey becomes extinct, the predator follows.

Problem 6.2 *If in the previous system the species have other alternative means of subsistence then the resulting equations are:*

$$\begin{aligned} \dot{x}_1 &= (1 - x_1 - a x_2) x_1 \\ \dot{x}_2 &= (1 - x_2 + b x_1) x_2 \end{aligned} \tag{6.25}$$

where another parameterization has been used.

[3] In reality, we must also prove that if $|\varphi(0)| < C$ then $\varphi(t)$ exists for all t. But the last inequality proved tells us that $\varphi(t)$ cannot leave the compact region $\{x |\ \rho(x,x) \leq \rho(\varphi(0), \varphi(0))\}$ and therefore the extension theorem assures us that $\varphi(t)$ exists for all $t \in [0, +\infty)$.

a) Find the equilibrium solutions for the cases i) $0 < a < 1$*, and ii)* $1 < a$*.*

b) It is observed that $b \approx 3a$ *(a indicates how aggressive the predator is). What is the value of a if its instinct leads it to maximize the population of its species while maintaining a stable equilibrium?*

c) Look at the system near its equilibrium solution and determine the frequency of the oscillations in the number of species that you would expect in that environment.

Problem 6.3 (Limit Cycle) *Consider the system:*

$$\begin{aligned} \dot{x}_1 &= x_2 + x_1(1-(x_1^2+x_2^2)) \\ \dot{x}_2 &= -x_1 + x_2(1-(x_1^2+x_2^2)). \end{aligned} \tag{6.26}$$

a) Transform this system into a pair of decoupled equations,

$$\begin{aligned} \dot{r} &= f(r) \\ \dot{\theta} &= -1. \end{aligned} \tag{6.27}$$

b) Study the equilibrium points of the first equation and their stability. What is the solution of the original system corresponding to this equilibrium point?

c) Plot solutions near these points, in the (r,θ) *plane and in the* (x_1,x_2) *plane.*

d) Use the method described at the end of chapter **??** *to corroborate the stability found in point b).*

Problem 6.4 (Verhulst) *Find the equilibrium solutions (and analyze their stability) of the system:*

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= x - 2x^3 \end{aligned} \tag{6.28}$$

Plot some solutions.

Problem 6.5 (Verhulst) *Find the equilibrium solutions (and analyze their stability) of the system:*

$$\begin{aligned} \dot{x} &= x^2 - y^3 \\ \dot{y} &= 2x(x^2 - y) \end{aligned} \tag{6.29}$$

Plot some solutions.

Problem 6.6 (Verhulst) *Find the equilibrium solutions (and analyze their stability) of the system:*

$$\begin{aligned} \dot{x} &= -x \\ \dot{y} &= 1-(x^2+y^2) \end{aligned} \tag{6.30}$$

Plot some solutions.

Chapter 7

Proof of the Fundamental Theorem

We will only prove points i) and ii), for which we will use the method of successive approximations by Picard, which is important in applications. The proof of point iii) is of a technical nature and does not provide any additional significant insight. Point iv) follows from points i) and ii), and its proof is identical to that of Corollary 1.2 of the analogous Theorem for the case of an ODE.

To prove these points, we will need to develop some ideas and results from the mathematical theory of infinite-dimensional vector spaces. We will restrict ourselves to the minimum required by the theorem since these topics will be developed more extensively in the second part of this course.

Definition: We will say that a vector space, V, is of **infinite dimension** if it has an infinite number of linearly independent vectors.
Examples:
a) Let V be the set of all sequences $\{x_i\}$, $i = 1, \ldots, \infty$ of real numbers. This is a vector space if we define the **sum** and the **product** of sequences by the following formula,

$$\{x_i\} + \alpha\{y_i\} = \{x_i + \alpha\, y_i\}. \tag{7.1}$$

These vectors can also be written as $\{x_i\} = (x_1, x_2, x_3, \ldots)$, which shows that it is the extension of $\mathbb{R}^n$ with $n \to \infty$. Clearly, the vectors $\boldsymbol{u}_1 = (1, 0, 0, \ldots)$; $\boldsymbol{u}_2 = (0, 1, 0, \ldots)$; $\boldsymbol{u}_3 = (0, 0, 1, \ldots)$; are linearly independent and infinite in number.

b) Let V be the set of continuous functions on the interval $[0, 1]$. This is a vector space since if f and g are continuous on $[0, 1]$, then $h = f + \alpha g$, $\alpha \in \mathbb{R}$, is also continuous on $[0, 1]$.

The set of vectors $u_n = x^n$, $n \in \boldsymbol{N}$, is linearly independent and infinite ($\sum_{n=0}^{M} c^n\, u_n = \sum_{n=0}^{M} c^n\, x^n = 0 \Longrightarrow c^n = 0 \;\forall\; n \leq M$) since a polynomial of degree M has at most M roots.

Infinite-dimensional spaces can also be assigned norms, but in this case, these norms are not equivalent, and therefore one must be careful not to confuse the resulting structures. To avoid confusion, we will assign different names to spaces with different norms.

Examples: a) The normed vector space l^2 is the space of infinite sequences with the norm $||\{x_i\}||_2 = \sqrt{\sum_{i=1}^{\infty} x_i^2} < \infty$.
b) The normed vector space l^∞ is the space of infinite sequences with the norm $||\{x_i\}||_\infty = sup_i\{|x_i|\}$. That is, the space of all bounded sequences.
c) The normed vector space $C[a,b]$ is the space of continuous functions with the norm $||f||_c = sup_{x\in[a,b]}\{|f(x)|\}$. That is, the space of bounded continuous functions on $[a,b]$.

Exercises:
1) Show that the norms defined in the previous example are indeed norms.
2) Show that there are sequences in l^∞ that do not belong to l^2. Hint: Find one of them.
3) Show that $||\{x_i\}||_n := \sqrt[n]{\sum_{i=1}^{\infty} |x_i|^n} \underset{n\to\infty}{\longrightarrow} sup_i\{|x_i|\}$.

Unlike the finite-dimensional case, an infinite-dimensional normed space is not necessarily complete. To illustrate this, consider the normed vector space l_0^∞, which is a subspace of l^∞ consisting of all bounded sequences with only a finite number of non-zero terms. Each sequence $\{x_i\}_n = (1, 1/2, 1/3, \ldots, 1/n, 0, 0, \ldots)$ is in l_0^∞, the sequence of sequences $\{\{x_i\}_n\}$ is Cauchy

$$\left(||\{x_i\}_m - \{x_i\}_n||_\infty = \frac{1}{n+1} \underset{n\to\infty}{\longrightarrow} 0\right)$$

and converges to the sequence $\{x_i\}_\infty = (1, 1/2, 1/3, \ldots) \in l^\infty$ *which does not belong to* l_0^∞. Therefore, l_0^∞ is not complete.

Definition: We will say that a normed vector space $(V, ||\cdot||)$ is a **Banach space** if it is complete.
Examples: $\mathbb{R}^n$ with any of its norms, l^2 and l^∞ are Banach spaces.

An important result, crucial in the proof of the Fundamental Theorem, is the following theorem.

Theorem 7.1 *The vector space of bounded continuous functions, $C[a,b]$, is complete.*

Proof: Let $\{f_n(x)\}$ be a Cauchy sequence in $C[a,b]$, that is, each $f_n(x)$ is a continuous and bounded function on $[a,b]$ and it holds that given $\varepsilon > 0$ there exists $N > 0$ such that $\forall\ m,n > N \quad ||f_n - f_m||_c = sup_{x\in[a,b]}|f_n(x) - f_m(x)| < \varepsilon$. But then for each $x \in [a,b]$ the sequence of real numbers $\{f_n(x)\}$ is Cauchy. But the real numbers are complete and therefore for each $x \in [a,b]$ $\{f_n(x)\}$ converges to a number that we will call $f(x)$. But then given $\varepsilon > 0$ for all N such that $m,n \geq N$ implies $||f_n - f_m||_c < \varepsilon$, we have that

$$\begin{aligned} sup_{x\in[a,b]}|f(x)-f_N(x)| &= sup_{x\in[a,b]}\lim_{n\to\infty}|f_n(x)-f_N(x)| \\ &\leq sup_{n\geq N}sup_{x\in[a,b]}|f_n(x)-f_N(x)| \\ &= sup_{n\geq N}||f_n-f_N||_c<\varepsilon. \end{aligned} \tag{7.2}$$

Therefore, if we could prove that $f\in C[a,b]$, then we would have that $||f-f_n||_c\to 0$, $n\to\infty$ and therefore that $\{f_n\}\to f$ in $C[a,b]$ and the theorem would be proven.

Let $x\in[a,b]$ be any point, we will prove that f is continuous at x and thus on all $[a,b]$. Let $\varepsilon>0$, we want to find a δ such that $|x-y|<\delta$ implies $|f(x)-f(y)|<\varepsilon$. Take N such that $||f-f_N||_c<\varepsilon/3$ and δ such that $|x-y|<\delta$ implies $|f_N(x)-f_N(y)|<\varepsilon/3$ [This is possible since the $f_N(x)$ are continuous on $[a,b]$]. Then $|x-y|<\delta$ implies

$$\begin{aligned} |f(x)-f(y)| &\leq |f(x)-f_N(x)|+|f_N(x)-f_N(y)|+|f_N(y)-f(y)| \\ &< \tfrac{1}{3}\varepsilon+\tfrac{1}{3}\varepsilon+\tfrac{1}{3}\varepsilon=\varepsilon. \end{aligned} \tag{7.3}$$

and therefore the continuity of $f(x)$. Since $[a,b]$ is a compact set (closed and bounded), f is bounded on $[a,b]$ and therefore belongs to $C[a,b]$ ♠

Definition: Let $T:V\to V$ be a map from a Banach space V to itself. We will say that T is a **contraction** if there exists $\lambda<1$ such that:

$$||T(\boldsymbol{x})-T(\boldsymbol{y})||_V\leq\lambda||\boldsymbol{x}-\boldsymbol{y}||_V. \tag{7.4}$$

Examples:

a) The linear map in l^2; $A\{x_i\}=\left\{\frac{x_i}{i+1}\right\}$.

b) The map from $\mathbb{R}^2$ to $\mathbb{R}^2$ that sends the point (x,y) to the point $(x_0+x/2,y/2)$.

c) Any Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ with a modulus of continuity (k) less than one.

The important property of these maps is the following theorem

Theorem 7.2 *Let $T:V\to V$ be a contraction. Then there exists a unique $v\in V$ such that $T(v)=v$, and the sequence $T^n(\boldsymbol{u})$ converges to v for any $\boldsymbol{u}$.*

Proof: Let $||T(\boldsymbol{u})-\boldsymbol{u}||=d$, then $||T^{n+1}(\boldsymbol{u})-T^n(\boldsymbol{u})||\leq\lambda^n d$ and if $m>n$

$$\begin{aligned} &||\,T^m(\boldsymbol{u})-T^n(\boldsymbol{u})|| \\ &\quad= ||T^m(\boldsymbol{u})-T^{m-1}(\boldsymbol{u})+T^{m-1}(\boldsymbol{u})-T^{m-2}(\boldsymbol{u})+\cdots T^n(\boldsymbol{u})|| \\ &\quad\leq ||T^m(\boldsymbol{u})-T^{m-1}(\boldsymbol{u})||+||T^{m-1}(\boldsymbol{u})-T^{m-2}(\boldsymbol{u})||+\cdots \\ &\quad\leq d\,\textstyle\sum_n^m\lambda^m. \end{aligned}$$

Since $\sum_{n=0}^{\infty}\lambda^n$ converges, $\{T^n(\boldsymbol{u})\}$ is a Cauchy sequence. But V is complete and therefore there exists $v\in V$ such that $\lim_{n\to\infty}T^n\,\boldsymbol{u}=v$.

Since every contraction is continuous, we have that

$$T(v) = T(\lim_{n\to\infty} T^n(u)) = \lim_{n\to\infty} T^{n+1}(u) = v.$$

It only remains to prove that v is unique. Suppose by contradiction that there exists $w \in V$ different from v and such that $Tw = w$. But then $||w - v||_V = ||T(w) - T(v)||_V \leq \lambda ||w - v||_V$ which is a contradiction since $\lambda \neq 1$ ♠

Exercise: Let $T : B_{R,x_o} \to B_{R,x_o}$, $B_{R,x_o} \in V$ be a closed ball of radius R around x_o, a contraction [1]. Prove for this case the same statements as in the previous theorem.

Proof of points *i*) and *ii*) of the fundamental theorem. Since we only want to see local existence and uniqueness, that is, only in a neighborhood of a point p of M, it is sufficient to consider the system in $\mathbb{R}^n$. This will allow us to use the Euclidean norm present there. To see this, take a chart (U, φ) with $p \in U$ and for simplicity $\varphi(p) = 0 \in \mathbb{R}^n$. Using this map, we can translate the vector field v in M to a vector field $\tilde{v}$ defined in a neighborhood of zero in $\mathbb{R}^n$. There we will treat the problem of finding its integral curves $g(t,x)$ that pass through the point $x \in \varphi(U)$ at time $t = 0$. Then, through the map φ^{-1}, we will obtain in M the one-parameter families of diffeomorphisms $g^t(q)$, $q \in U$, which will be tangent at every point to the vector field v.

With this in mind, it only remains to see that for $R > 0$ and $\varepsilon > 0$ sufficiently small there exist integral curves, $g(t,x) : [0,\varepsilon] \times B_R = \{x \in \mathbb{R}^n \mid ||x||_V < R\} \to \mathbb{R}^n$, of the vector $\tilde{v}$, that is, maps satisfying

$$\frac{dg(t,x)}{dt} = \tilde{v}(g(t,x)) \quad , \quad g(0,x) = x, \tag{7.5}$$

with $g(t,x)$ continuous with respect to the second argument, that is, with respect to the initial condition. By the assumption that v is Lipschitz [2], we have that there exists $k > 0$ such that $\forall\, x, y \in B_R$

$$||\tilde{v}(x) - \tilde{v}(y)||_V < k\, ||x - y||_V \,. \tag{7.6}$$

Now consider the Banach space $C([0,\varepsilon] \times B_R)$, which consists of all continuous maps (in t and x) from $[0,\varepsilon] \times B_R$ to $\mathbb{R}^n$ with the norm

$$||h||_C = \sup_{\substack{x \in B_R \\ t \in [0,\varepsilon].}} ||h(t,x)||_V \,, \tag{7.7}$$

Let the map of the ball of radius R in $C([0,\varepsilon] \times B_R)$ into itself be given by,

$$T(h) = \int_0^t \tilde{v}(x + h(\tau, x))\, d\tau \,. \tag{7.8}$$

[1] Also adjust the definition of a contraction for this case.

[2] To prove the local existence and uniqueness of solutions, it is only necessary to assume that v is Lipschitz.

For this map to be well-defined, we will assume R is sufficiently small so that $\tilde{v}$ is defined and satisfies the Lipschitz condition in B_{2R} and ε is less than R/C, where $C = \max_{x \in B_{2R}} |\tilde{v}(x)|$, so that if $||h||_C < R$ then $||x + h(\tau, x)||_V < 2R \ \ \forall \tau \in [0, \varepsilon]$ and therefore $||T(h)||_C < R$ in all $C([0, \varepsilon] \times B_R)$. [See figure 7.1.]

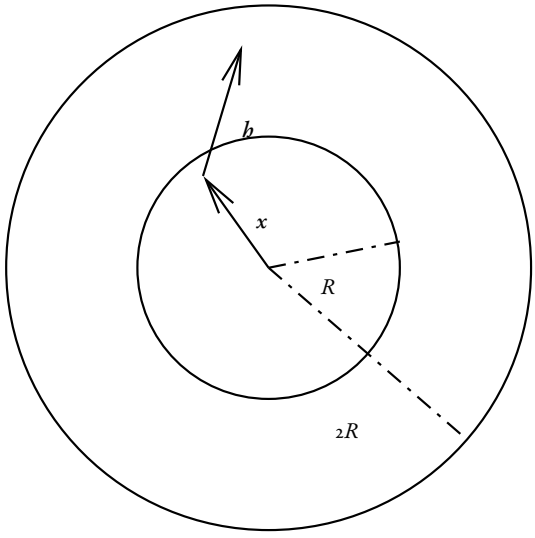

Figure 7.1: Neighborhoods used in the proof of the Fundamental Theorem.

We will only prove points i) and ii), for which we will use the method of successive approximations by Picard, which is important in applications. The proof of point iii) is of a technical nature and does not provide any additional significant insight. Point iv) follows from points i) and ii), and its proof is identical to that of Corollary 1.2 of the analogous Theorem for the case of an ODE.

To prove these points, we will need to develop some ideas and results from the mathematical theory of infinite-dimensional vector spaces. We will restrict ourselves to the minimum required by the theorem since these topics will be developed more extensively in the second part of this course.

Definition: We will say that a vector space, V, is of **infinite dimension** if it has an infinite number of linearly independent vectors.
Examples:
a) Let V be the set of all sequences $\{x_i\}$, $i = 1, \ldots, \infty$ of real numbers. This is a vector space if we define the **sum** and the **product** of sequences by the following formula,

$$\{x_i\} + \alpha\{y_i\} = \{x_i + \alpha\, y_i\}. \tag{7.9}$$

These vectors can also be written as $\{x_i\} = (x_1, x_2, x_3, \ldots)$, which shows that it is the extension of $\mathbb{R}^n$ with $n \to \infty$. Clearly, the vectors $\boldsymbol{u}_1 = (1, 0, 0, \ldots)$; $\boldsymbol{u}_2 = (0, 1, 0, \ldots)$; $\boldsymbol{u}_3 = (0, 0, 1, \ldots)$; are linearly independent and infinite in number.

b) Let V be the set of continuous functions on the interval $[0, 1]$. This is a vector space since if f and g are continuous on $[0, 1]$, then $h = f + \alpha g$, $\alpha \in \mathbb{R}$, is also continuous on $[0, 1]$.

The set of vectors $u_n = x^n$, $n \in \boldsymbol{N}$, is linearly independent and infinite ($\sum_{n=0}^{M} c^n u_n = \sum_{n=0}^{M} c^n x^n = 0 \Longrightarrow c^n = 0 \ \forall\, n \leq M$) since a polynomial of degree M has at most M roots.

Infinite-dimensional spaces can also be assigned norms, but in this case, these norms are not equivalent, and therefore one must be careful not to confuse the resulting structures. To avoid confusion, we will assign different names to spaces with different norms.

Examples: a) The normed vector space l^2 is the space of infinite sequences with the norm $||\{x_i\}||_2 = \sqrt{\sum_{i=1}^{\infty} x_i^2} < \infty$.

b) The normed vector space l^∞ is the space of infinite sequences with the norm $||\{x_i\}||_\infty = sup_i\{|x_i|\}$. That is, the space of all bounded sequences.

c) The normed vector space $C[a, b]$ is the space of continuous functions with the norm $||f||_c = sup_{x\in[a,b]}\{|f(x)|\}$. That is, the space of bounded continuous functions on $[a, b]$.

Exercises:

1) Show that the norms defined in the previous example are indeed norms.

2) Show that there are sequences in l^∞ that do not belong to l^2. Hint: Find one of them.

3) Show that $||\{x_i\}||_n := \sqrt[n]{\sum_{i=1}^{\infty} |x_i|^n} \underset{n\to\infty}{\longrightarrow} sup_i\{|x_i|\}$.

Unlike the finite-dimensional case, an infinite-dimensional normed space is not necessarily complete. To illustrate this, consider the normed vector space l_0^∞, which is a subspace of l^∞ consisting of all bounded sequences with only a finite number of non-zero terms. Each sequence $\{x_i\}_n = (1, 1/2, 1/3, \ldots, 1/n, 0, 0, \ldots)$ is in l_0^∞, the sequence of sequences $\{\{x_i\}_n\}$ is Cauchy

$$\left(||\{x_i\}_m - \{x_i\}_n||_\infty = \frac{1}{n+1} \underset{n\to\infty}{\longrightarrow} 0 \right)$$

and converges to the sequence $\{x_i\}_\infty = (1, 1/2, 1/3, \ldots) \in l^\infty$ *which does not belong to* l_0^∞. Therefore, l_0^∞ is not complete.

Definition: We will say that a normed vector space $(V, ||\cdot||)$ is a **Banach space** if it is complete.

Examples: $\mathbb{R}^n$ with any of its norms, l^2 and l^∞ are Banach spaces.

An important result, crucial in the proof of the Fundamental Theorem, is the following theorem.

Theorem 7.3 *The vector space of bounded continuous functions, $C[a,b]$, is complete.*

Proof: Let $\{f_n(x)\}$ be a Cauchy sequence in $C[a,b]$, that is, each $f_n(x)$ is a continuous and bounded function on $[a,b]$ and it holds that given $\varepsilon > 0$ there exists $N > 0$ such that $\forall\ m,n > N \quad ||f_n - f_m||_c = sup_{x\in[a,b]}|f_n(x)-f_m(x)| < \varepsilon$. But then for each $x \in [a,b]$ the sequence of real numbers $\{f_n(x)\}$ is Cauchy. But the real numbers are complete and therefore for each $x \in [a,b]$ $\{f_n(x)\}$ converges to a number that we will call $f(x)$. But then given $\varepsilon > 0$ for all N such that $m,n \geq N$ implies $||f_n - f_m||_c < \varepsilon$, we have that

$$\begin{aligned} sup_{x\in[a,b]}|f(x)-f_N(x)| &= sup_{x\in[a,b]} \lim_{n\to\infty} |f_n(x)-f_N(x)| \\ &\leq sup_{n\geq N} sup_{x\in[a,b]}|f_n(x)-f_N(x)| \\ &= sup_{n\geq N}||f_n - f_N||_c < \varepsilon. \end{aligned} \tag{7.10}$$

Therefore, if we could prove that $f \in C[a,b]$, then we would have that $||f - f_n||_c \to 0$, $n \to \infty$ and therefore that $\{f_n\} \to f$ in $C[a,b]$ and the theorem would be proven.

Let $x \in [a,b]$ be any point, we will prove that f is continuous at x and thus on all $[a,b]$. Let $\varepsilon > 0$, we want to find a δ such that $|x-y| < \delta$ implies $|f(x)-f(y)| < \varepsilon$. Take N such that $||f - f_N||_c < \varepsilon/3$ and δ such that $|x-y| < \delta$ implies $|f_N(x)-f_N(y)| < \varepsilon/3$ [This is possible since the $f_N(x)$ are continuous on $[a,b]$]. Then $|x-y| < \delta$ implies

$$\begin{aligned} |f(x)-f(y)| &\leq |f(x)-f_N(x)|+|f_N(x)-f_N(y)|+|f_N(y)-f(y)| \\ &< \tfrac{1}{3}\varepsilon + \tfrac{1}{3}\varepsilon + \tfrac{1}{3}\varepsilon = \varepsilon. \end{aligned} \tag{7.11}$$

and therefore the continuity of $f(x)$. Since $[a,b]$ is a compact set (closed and bounded), f is bounded on $[a,b]$ and therefore belongs to $C[a,b]$ ♠

Definition: Let $T : V \to V$ be a map from a Banach space V to itself. We will say that T is a **contraction** if there exists $\lambda < 1$ such that:

$$||T(\boldsymbol{x})-T(\boldsymbol{y})||_V \leq \lambda ||\boldsymbol{x}-\boldsymbol{y}||_V. \tag{7.12}$$

Examples:

a) The linear map in l^2; $A\{x_i\} = \left\{\frac{x_i}{i+1}\right\}$.

b) The map from $\mathbb{R}^2$ to $\mathbb{R}^2$ that sends the point (x,y) to the point $(x_0+x/2, y/2)$.

c) Any Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ with a modulus of continuity (k) less than one.

The important property of these maps is the following theorem

Theorem 7.4 *Let $T : V \to V$ be a contraction. Then there exists a unique $v \in V$ such that $T(v) = v$, and the sequence $T^n(u)$ converges to v for any u.*

Proof: Let $||T(u) - u|| = d$, then $||T^{n+1}(u) - T^n(u)|| \leq \lambda^n d$ and if $m > n$

$$\begin{aligned} &|| T^m(u) - T^n(u)|| \\ &\quad = \quad ||T^m(u) - T^{m-1}(u) + T^{m-1}(u) - T^{m-2}(u) + \cdots T^n(u)|| \\ &\quad \leq \quad ||T^m(u) - T^{m-1}(u)|| + ||T^{m-1}(u) - T^{m-2}(u)|| + \cdots \\ &\quad \leq \quad d \sum_n^m \lambda^m. \end{aligned}$$

Since $\sum_{n=0}^{\infty} \lambda^n$ converges, $\{T^n(u)\}$ is a Cauchy sequence. But V is complete and therefore there exists $v \in V$ such that $\lim_{n\to\infty} T^n u = v$.

Since every contraction is continuous, we have that

$$T(v) = T(\lim_{n\to\infty} T^n(u)) = \lim_{n\to\infty} T^{n+1}(u) = v.$$

It only remains to prove that v is unique. Suppose by contradiction that there exists $w \in V$ different from v and such that $Tw = w$. But then $||w - v||_V = ||T(w) - T(v)||_V \leq \lambda ||w - v||_V$ which is a contradiction since $\lambda \neq 1$ ♠

Exercise: Let $T : B_{R,x_0} \to B_{R,x_0}$, $B_{R,x_0} \in V$ be a closed ball of radius R around x_0, a contraction [3]. Prove for this case the same statements as in the previous theorem.

Proof of points *i*) and *ii*) of the fundamental theorem. Since we only want to see local existence and uniqueness, that is, only in a neighborhood of a point p of M, it is sufficient to consider the system in $\mathbb{R}^n$. This will allow us to use the Euclidean norm present there. To see this, take a chart (U, φ) with $p \in U$ and for simplicity $\varphi(p) = 0 \in \mathbb{R}^n$. Using this map, we can translate the vector field v in M to a vector field $\tilde{v}$ defined in a neighborhood of zero in $\mathbb{R}^n$. There we will treat the problem of finding its integral curves $g(t, x)$ that pass through the point $x \in \varphi(U)$ at time $t = 0$. Then, through the map φ^{-1}, we will obtain in M the one-parameter families of diffeomorphisms $g^t(q)$, $q \in U$, which will be tangent at every point to the vector field v.

With this in mind, it only remains to see that for $R > 0$ and $\varepsilon > 0$ sufficiently small there exist integral curves, $g(t, x) : [0, \varepsilon] \times B_R = \{x \in \mathbb{R}^n \mid ||x||_V < R\} \to \mathbb{R}^n$, of the vector $\tilde{v}$, that is, maps satisfying

$$\frac{d\, g(t,x)}{d\, t} = \tilde{v}(g(t,x)) \quad , \quad g(0,x) = x, \tag{7.13}$$

with $g(t, x)$ continuous with respect to the second argument, that is, with respect to the initial condition. By the assumption that v is Lipschitz [4], we have that there

[3] Also adjust the definition of a contraction for this case.

[4] To prove the local existence and uniqueness of solutions, it is only necessary to assume that v is Lipschitz.

exists $k > 0$ such that $\forall\ x, y \in B_R$

$$||\tilde{v}(x) - \tilde{v}(y)||_V < k\,||x - y||_V\ . \tag{7.14}$$

Now consider the Banach space $C([0,\varepsilon] \times B_R)$, which consists of all continuous maps (in t and x) from $[0,\varepsilon] \times B_R$ to $\mathbb{R}^n$ with the norm

$$||h||_C \quad = \quad \underset{\substack{x \in B_R \\ t \in [0,\varepsilon].}}{sup} \quad ||h(t,x)||_V\ , \tag{7.15}$$

Let the map of the ball of radius R in $C([0,\varepsilon] \times B_R)$ into itself be given by,

$$T(h) = \int_0^t \tilde{v}(x + h(\tau, x))\, d\tau\ . \tag{7.16}$$

For this map to be well-defined, we will assume R is sufficiently small so that $\tilde{v}$ is defined and satisfies the Lipschitz condition in B_{2R} and ε is less than R/C, where $C = \max_{x \in B_{2R}} |\tilde{v}(x)|$, so that if $||h||_C < R$ then $||x + h(\tau, x)||_V < 2R\ \ \forall \tau \in [0,\varepsilon]$ and therefore $||T(h)||_C < R$ in all $C([0,\varepsilon] \times B_R)$. [See figure 7.1.]

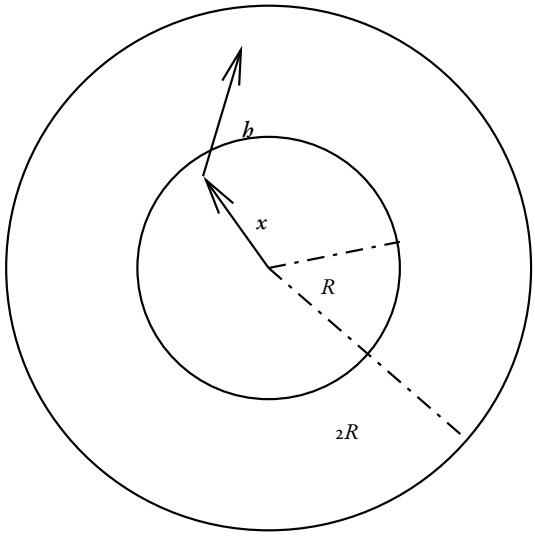

Figure 7.2: Environments used in the proof of the Fundamental Theorem.

Lemma 7.1 *If ε is sufficiently small then T is a contraction.*

Proof:

$$\begin{aligned} ||T(h_1)-T(h_2)||_V &= \int_0^t ||\tilde{v}(x+h_1(\tau,x))-\tilde{v}(x+h_2(\tau,x))||_V \, d\tau \\ &\leq \int_0^t k\,||h_1(\tau,x)-h_2(\tau,x)||_V \, d\tau \\ &\leq k\,\varepsilon\,||h_1-h_2||_C \end{aligned}$$

and therefore

$$||T(h_1)-T(h_2)||_C \leq k\,\varepsilon\,||h_1-h_2||_C \qquad \forall\, h_1, h_2 \in C([0,\varepsilon]\times B_R).$$

Taking $\varepsilon < 1/k$ we complete the proof of the Lemma.

This Lemma and Theorem 7.4 ensure that there exists a unique map $h(t,x)$ –the fixed point of T– satisfying

$$h(t,x) = T(h(t,x)) = \int_0^t \tilde{v}(x+h(\tau,x))\,d\tau. \tag{7.17}$$

Let $g(t,x) \equiv x + h(t,x)$, this function is continuous in both arguments –since $h(t,x) \in C([0,\varepsilon]\times B_R)$– and by construction continuously differentiable in t –since it satisfies (7.17)–. Differentiating (7.17) with respect to t we see that $g(t,x)$ satisfies the equation (7.13) and its initial condition, which completes the proof of points i) and ii) of the Fundamental Theorem. ♠

7.1 Problems

Problem 7.1 *See that l^2 the unit ball is not compact. Hint: find an infinite sequence in the unit ball that has no accumulation point.*

Problem 7.2 *Prove that the condition*

$$||A(\boldsymbol{x})-A(\boldsymbol{y})|| < ||\boldsymbol{x}-\boldsymbol{y}|| \tag{7.18}$$

is not sufficient to guarantee the existence of a fixed point. Hint: construct a counterexample. There are some very simple ones using functions from the line to itself.

Problem 7.3 *Let $f : [a,b] \to [a,b]$ be a Lipschitz function with a constant less than one throughout the interval $[a,b]$. Prove using the fixed point theorem for contractions that the equation $f(x) = x$ always has a solution. Plot in a diagram $(y = f(x), x)$ the sequence given by $x_i = f(x_{i-1})$ for the case of f with positive slope and less than one at every point. What happens when the slope becomes greater than one in some interval?*

Problem 7.4 *Let $g : [a,b] \to \mathbb{R}$ be a continuously differentiable function such that $g(a) < 0$, $g(b) > 0$ and $0 < c_1 < g'(x)$. Use the fixed point theorem for contractions to prove that there is a unique root $g(x) = 0$ in the interval. Hint: define the function $f(x) = x - \lambda g(x)$ for a conveniently chosen constant λ and find a fixed point: $f(x) = x$. Note that the approximating sequence is that of the Newton method for finding roots.*

CHAPTER 8

BASIC ELEMENTS OF FUNCTIONAL ANALYSIS

8.1 Introduction

This area of mathematics studies functional spaces, that is, spaces whose elements are certain functions. Usually, these are of infinite dimension. Thus, its main results form two classes. Some are general results, valid for more general vector spaces than those whose elements are functions. These have a geometric or topological character that we will try to rescue at all times. Others are particular results that relate different functional spaces to each other. These are closer to the results of usual analysis. We will present here results of both types since from their conjunction we will obtain some aspects of the theory of PDEs. For reasons of brevity, we will only consider the points that will be useful to us or those for which a minimum of extra effort is required to obtain them and their cultural importance justifies it.

8.2 Completing a Normed Space

In the previous chapter, we introduced Banach spaces, that is, vector spaces with a norm defined on them and that were complete with respect to it. There we also proved that the set of continuous functions on an interval $[a, b] \in \mathbb{R}$ with the norm

$$||f||_1 = \sup_{x \in [a,b]} \{|f(x)|\} \tag{8.1}$$

was complete and therefore Banach.

One may ask if, given a normed vector space W, it is possible to "fatten it up," that is, add vectors to it, and thus make it complete. Note that this is what is done with the rationals $\mathbb{Q}$ (which is a vector space if we only allow the multiplication of its elements by rational numbers!). The fattened space is in this case that of the real numbers. The answer is affirmative and is given by the following theorem.

Theorem 8.1 *Let W be a normed vector space. Then there exists a Banach space V and a continuous linear map $\varphi : W \to V$ such that $||\varphi(w)||_V = ||w||_W$ and $\varphi[W]$ is a dense subspace of V, that is, the closure of $\varphi[W]$ is V.*

Proof: The details of this can be found in, for example, [13], page 56. Here we will only give the ideas. If W is complete then we take $V = W$ and $\varphi = id$. Therefore we will assume that W is not complete. Then there will be Cauchy sequences $\{w_n\}$ in W that do not converge to any point of W. The idea is to take these sequences as new points with which to fatten up W. As many different sequences could tend to the same point, in order not to fatten up W too much, all of them should be considered as a single element. This is achieved by taking equivalence classes of sequences as elements of V. We will say that two sequences, $\{w_n\}$, $\{w'_n\}$, are equivalent if their difference tends to the zero element,

$$\lim_{n\to\infty} ||w_n - w'_n||_W = 0. \tag{8.2}$$

As the set of Cauchy sequences of elements of W forms a vector space, the set of equivalence classes of sequences is also a vector space, this will be V. This space inherits a norm naturally from W, given by,

$$||\{w_n\}||_V = \lim_{n\to\infty} ||w_n||_W, \tag{8.3}$$

which is clearly independent of the particular sequence that one chooses, to calculate it, within the equivalence class. It can be easily proved that with this norm V is complete and therefore Banach.

Exercise: Prove that the Cauchy sequence of Cauchy sequences, $\{\{w_n\}_N\}$ converges in this norm to the sequence $\{\bar{w}_n\} := \{\{w_n\}_n\}$.

What is the map φ? This takes an element $w \in W$ and gives us an element in V, that is, an equivalence class of sequences. This is the equivalence class that converges to w and a representative is, for example,

$$\{w_n\} = (w, w, w, \ldots). \tag{8.4}$$

The previous theorem tells us that we can always complete a normed vector space and this in an essentially unique way. In this way then we can talk about completing a normed space W in one V. As W is dense in V and the map φ is continuous all the properties that are continuous in W automatically hold in V. The previous theorem also tells us that the elements of the completed space can have a very different character from the elements of the original space. Due to this, in practice one must be very careful when attributing properties to the elements of a given Banach space. As an example of the above we will see the Lebesgue integral.

8.3 *Lebesgue Integral

The Lebesgue integral is an extension of the Riemann integral to a more general class of functions than that where the Riemann integral is defined. We have two ways to define it, one as the norm of a completed Banach space, whose elements are then the

functions integrable in the sense of Lebesgue. The other way is using a limit process similar to that used to define the Riemann integral. We will see both.

Let W be the space of continuous functions on the interval $[a, b]$ and let

$$||f||_w = \int_a^b |f(x)|\, dx \tag{8.5}$$

be its norm, where the integral is in the sense of Riemann. This definition makes sense since the elements of W are continuous functions and therefore the integral is well defined. Note also that this is a norm since as f is continuous if the integral of its modulus is zero, its modulus, and thus f, is also zero. Let $L^1([a, b])$ be the completed space of W. This is the space of Lebesgue integrable functions. What functions are there?[1] Let $\{f_n\}$ be the sequence given by the graph of fig. 8.1.

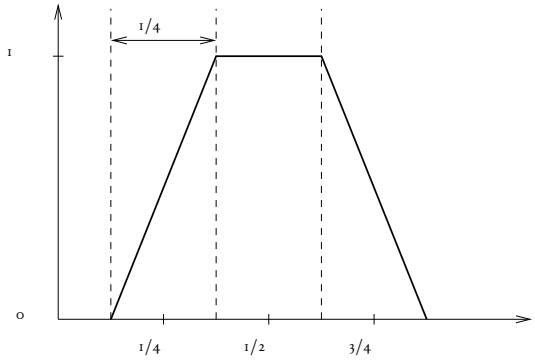

Figure 8.1: A Cauchy sequence.

This sequence is Cauchy in $L^1([0, 1])$ and therefore converges to an element of $L^1([0, 1])$, the function,

$$f(x) = \begin{cases} 0 & 0 \le x \le 1/4,\ 3/4 \le x \le 1 \\ 1 & 1/4 < x < 3/4. \end{cases} \tag{8.6}$$

This is not surprising since although this function is not continuous it is integrable even in the sense of Riemann. Let now the sequence $\{f_n\}$ be given by the graph of fig. 8.2,

this is also Cauchy and tends to the function,

$$f(x) = \begin{cases} 0 & 0 \le x < 1/2 < x \le 1 \\ 1 & x = 1/2, \end{cases} \tag{8.7}$$

[1] In reality the functions are not properly elements of L^1 but rather these are the image by the map φ (defined in the previous section) of these.

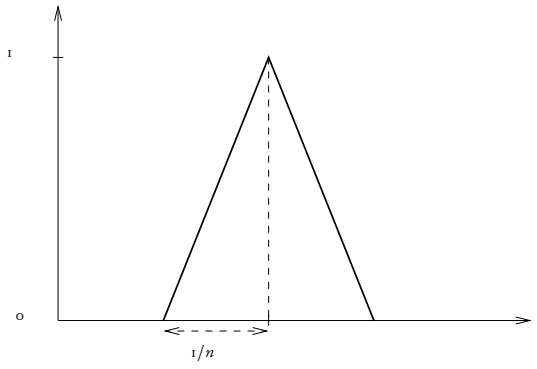

Figure 8.2: Another Cauchy sequence.

which is very strange. Whatever the meaning of the Lebesgue integral the integral of this function is expected to be zero, and in fact it is since

$$\lim_{n\to\infty} \int_0^1 |f_n(x)|\, dx = 0. \tag{8.8}$$

But then the norm of f is zero and therefore it would seem we have a contradiction, unless f is zero. The resolution of this problem consists in noting that when we complete the space W we take as its elements certain equivalence classes. The function described above is in the equivalence class corresponding to the zero element. As we will see later, the elements of L^1 are functions, but only defined at **almost all points**.

The second method consists in defining the Lebesgue integral in a manner similar to how the Riemann integral is defined. For this, we must define what is called a **measure** on certain subsets of $\mathbb{R}$, that is, a function from these subsets to the positive reals, which generalizes the concept of length (or measure) of an open interval (a, b), $\mu((a, b)) = b - a$.

Once this concept of measure is introduced, the Lebesgue integral of a positive function $f : \mathbb{R} \to \mathbb{R}^+$ is defined as,

$$\int f(x)\, dx = \lim_{n\to\infty} \sum_n (f) = \lim_{n\to\infty} \sum_{m=0}^{\infty} \frac{m}{n} \mu\left(f^{-1}\left[[\frac{m}{n}, \frac{m+1}{n}) \right] \right),$$

where we have used the same symbol to denote the Lebesgue integral as that normally used for the Riemann integral, which is natural since as we saw in the first definition if a function is integrable in the sense of Riemann it is also integrable in the sense of Lebesgue and the value of the integrals coincides.

The interpretation of this definition is as follows: the image of f is divided into regular intervals of length $\frac{1}{n}$, the image by f^{-1} of these intervals is considered, they

are *measured* with μ and these measures are summed conveniently. See figure. Finally, the limit is taken as n goes to infinity, that is, the intervals go to zero. Note that $\sum_{2n}(f) \geq \sum_n(f)$ and therefore $\lim_{n\to\infty} \sum_n(f) = \sup_n \{\sum_n(f)\}$ exists (it could be infinite).

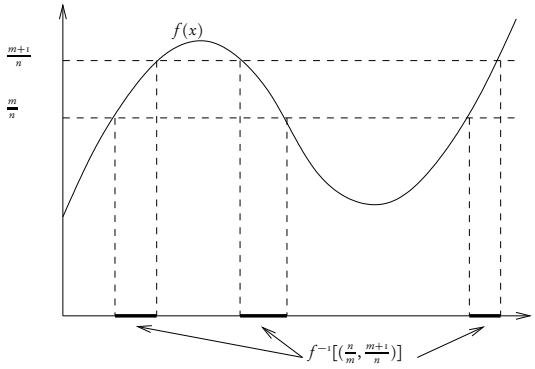

Figure 8.3: Lebesgue Integral.

For which functions is this operation defined? The condition that f be positive is not really a restriction since whatever it is, it can always be written as $f = f_+ - f_-$ with f_+ and f_- both positive and the integral is a linear operation [which is not obvious from the definition given above]. The condition that is restrictive is that $f^{-1}(A)$, with A open, be a measurable subset, since as we will see if we ask the measure function to satisfy certain natural properties then not every subset of $\mathbb{R}$ can be in the domain of this function. To study this restriction we must be clear about what properties we want the measure to satisfy, find the collection of subsets of $\mathbb{R}$ that are measurable and finally define the measure. These properties will define the notion of measure, or more specifically of measurable space, since the properties we assign depend both on the measure and its domain of definition.

Definition: A **measurable space** consists of a triple (X, M, μ) where X is a set (the domain of the functions to be integrated), M a collection of subsets of X called the **measurable subsets** of X and μ is a function (called the **measure**) from M to $\mathbb{R}^* = \mathbb{R}^+ \cup \{\infty\}$ satisfying the following conditions:

i) $\emptyset \in M$ and $\mu(\emptyset) = 0$.

ii) If $A \in M$ then A^c (the complement of A in X) is also in M.

iii) If $A_i \in M, i = 1, 2, \ldots,$ then $\bigcup_i A_i \in M$.

iv) If $A_i \in M, i = 1, 2, \ldots$ and $A_i \cap A_j = \emptyset$ if $i \neq j$ then $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$.

Intuitively, measurable sets are sets that admit a notion of *length* or *area* and their measure is the value of this *length* or *area*.

Exercise: Show that:

1. $X \in M$
2. If $A_i \in M$, $i = 1, 2, \ldots$ then $\bigcap_i A_i \in M$. This is the reason why *ii)* is required, since we would expect that $\mu\left(\bigcap_i A_i\right) \leq \sum_i \mu(A_i)$.
3. If $A \subset B$ then $\mu(A) \leq \mu(B)$.

Examples:

1. Let X be any set and let M be the collection of all subsets of X. Let μ be such that $\mu(\emptyset) = 0$ and $\mu(A) = \infty$ for all non-empty $A \in M$.
2. Let $M = \{\emptyset, X\}$ and let μ be such that $\mu(\emptyset) = 0$ and $\mu(X) = 7$.
3. Let $X = \mathbb{Z}^+ = \{\text{positive integers}\}$ and let $\{x_i\}, x_i \in \mathbb{R}^+$, be a sequence. Let M be the collection of all subsets of $\mathbb{Z}^+$ and $\mu(A) = \sum_{i \in A}(x_i)$.

   The following is not a measurable space (why?) but it will be useful to construct the one we desire later.
4. Let $X = \mathbb{R}$ and $M = \mathscr{J}$ be the set of all (countable) unions of disjoint open intervals, that is, an element I of $\mathscr{J}$ is a subset of $\mathbb{R}$ of the form,

$$I = \bigcup_i (a_i, b_i), \tag{8.9}$$

   with $a_1 \leq b_1 < a_2 \leq b_2 < a_3 \cdots$. Let $\mu := m : \mathscr{J} \to \mathbb{R}^*$ be defined as,

$$m(I) = \sum_i (b_i - a_i). \tag{8.10}$$

Exercise: Show that the third example is a measurable space.

As can be seen from these examples, the notion of a measurable space is broad. Different examples of measurable spaces appear in various branches of physics. From now on, we will restrict ourselves to a measurable space, which is the Lebesgue space, which we will construct from the fourth example.

Let $X = \mathbb{R}$, $\bar{M}$ be any subset of X, and $\bar{\mu} : \bar{M} \to \mathbb{R}^*$ given by,

$$\bar{\mu}(A) = \inf_{\substack{I \in \mathscr{J} \\ A \subseteq I}} (m(I)) \tag{8.11}$$

That is, given a subset of $\mathbb{R}$, A, we consider all elements of $\mathscr{J}$ such that A is contained in them, calculate their measure, and take the infimum over all elements of $\mathscr{J}$.
Examples:

1. Let $A = (0, 1)$, then a candidate is $I_1 = (-1, 1) \cup (3, 5)$, $m(I_1) = 2 + 2 = 4$, but we also have $I_2 = (0, 1)$ with $m(I_2) = 1$ and clearly $\bar{\mu}(A) = 1$.
2. Let $B = 1 \cup 2 \cup 3 \cup \ldots$ then $\bar{\mu}(B) = 0$ since we can cover B with $I = (1 - \epsilon, 1 + \epsilon) \cup (2 - \epsilon, 2 + \epsilon) \cup \ldots$.

As we see, this triplet $(\mathbb{R}, \bar{M}, \bar{\mu})$ seems to have the desired conditions to be the measure we seek, but in reality, it is not even a measurable space!

Counterexample: Let (S^1, M, μ), where S^1 is the circle, be a measurable space with μ a finite measure invariant under translations, that is, $\mu(A) = \mu(A_r)$ where $A_r = \{a + r | a \in A\}$ -note that our $\bar{\mu}$ is, but this result is much more general-. Then M cannot be the collection of all subsets of S^1.

The idea is to find a subset A of S^1 such that if we assume it is measurable, we obtain a contradiction. To do this, we think of S^1 as $\mathbb{R}$ with its ends identified (0=1). We now introduce an equivalence relation: We will say that two points of S^1 are equivalent $a \approx b$ if $a - b$ is a rational number.

Exercise: Show that this is an equivalence relation.

We construct A by taking exactly one element from each equivalence class -note that there are infinitely many ways to choose a set A- Suppose A is measurable with $\mu(A) = \alpha \in \mathbb{R}^+$ and let $A_r = \{a | (a + r) \ (mod\ 1) \in A\}$, with r rational, that is, a translation by $-r$ of A. Then it is easy to see that if $r \neq r'$ then $A_r \cap A_{r'} = \emptyset$, that is, the A_r are disjoint [If $x \in A_r$ and $x \in A_{r'}$ then $x = a + r = b + r'$, with $a, b \in A$, but then $a - b = r - r' \in \mathbb{Q}$ which is a contradiction.] and that $S^1 = \bigcup_{r \in \mathbb{Q}} A_r$ [Let $x \in S^1$, then x belongs to one of the equivalence classes into which we have separated S^1, but then there exists $a \in A$ and $r \in \mathbb{Q}$ such that $a + r = x$, that is, $x \in A_r$.]. Since by hypothesis $\mu(A_r) = \mu(A) = \alpha$ then,

$$1 = \mu(S^1) = \mu\left(\bigcup_{r \in \mathbb{Q}} A_r\right) = \sum_{r \in \mathbb{Q}} \mu(A_r) = \sum^{\infty} \alpha, \tag{8.12}$$

which leads to a contradiction since if $\alpha = 0$ then the sum is zero and if $\alpha \neq 0$ the sum is infinite.

This counterexample then tells us that we must restrict M to be a subset of $\bar{M}$ if we want to have a measure. There are many ways to characterize this restriction, one of which is given by the following theorem (which we will not prove).

Theorem 8.2 *Let M be the subspace of $\bar{M}$ such that if $A \in M$ then*

$$\bar{\mu}(E) = \bar{\mu}(A \cap E) + \bar{\mu}((X - A) \cap E) \; \forall E \in \bar{M}. \tag{8.13}$$

Let μ be the restriction of $\bar{\mu}$ to M, then (X, M, μ) is the Lebesgue measurable space.

Intuitively, we see that measurable sets are those that *when used to separate other sets into two parts give a division that is* **additive** *with respect to $\bar{\mu}$.*

Since in general we only have $\bar{\mu}(E) \leq \bar{\mu}(A \cap E) + \bar{\mu}((X - A) \cap E)$ we then see that non-measurable sets are those *whose points are distributed in X in such a way that when one tries to cover $A \cap E$ and $(X - A) \cap E$ with open sets, they overlap so much that in reality one can obtain a smaller infimum by covering E directly with open sets.*

There are a large number of theorems that give us information about which sets are measurable. Suffice it to say that all open sets of $\mathbb{R}$ (and many more) are in M and that if $f : \mathbb{R} \to \mathbb{R}$ is a continuous function then $f^{-1}[M] \in M$.

We now return to the Lebesgue integral. Naturally, we will say that f is **measurable** if $f^{-1}[(a, b)] \in M$ for every open interval $(a, b) \in \mathbb{R}$.

Theorem 8.3 *Measurable functions have the following properties:*

a) If f and g are measurable and $\lambda \in \mathbb{R}$ then $f + \lambda g$ is measurable.

b) Also measurable are fg, $max\{f, g\}$ and $min\{f, g\}$.

c) Let $\{f_n(x)\}$ be a sequence of measurable functions that converge pointwise to $f(x)$, that is, $\lim_{n \to \infty} f_n(x) = f(x)$, then $f(x)$ is also measurable.

Note that $|f| = max\{f, -f\}$ and $f_{\pm} =^{max}_{min} \{f, 0\}$.

The first part of this theorem tells us that the set of measurable functions is a vector space. This remains true if we restrict ourselves to the space of integrable functions in the sense of Lebesgue, that is, f measurable and $\int |f| \, dx < \infty$. We will denote this space as $\mathscr{L}_1$ or $\mathscr{L}_1(\mathbb{R})$. Similarly, we will define $\mathscr{L}_1[a, b]$ as the space of integrable functions $f : [a, b] \to \mathbb{R}$ where in this case the integral is defined as before but extending the function f to all $\mathbb{R}$ with zero value outside the interval $[a, b]$.

As we saw before, to make $\mathscr{L}_1$ a normed space we must take as its elements the equivalent classes of functions where we will say that $f, g \in \mathscr{L}_1$ are **equivalent** if $\int |f - g| dx = 0$. This is equivalent to saying that the subset of $\mathbb{R}$ where f is different from g is of measure zero, or in other words that f is equal to g at **almost every point**.

We will denote the space of equivalent classes of integrable (Lebesgue) functions with L_1 and its elements with tildes. Note that an element $\tilde{f}$ of L_1 is an equivalent class of functions and therefore in general *the value of $\tilde{f}$ at x, $\tilde{f}(x)$*, makes no sense since we can have functions in the equivalent class $\tilde{f}$, f_1 and f_2 with $f_1(x) \neq f_2(x)$ for some $x \in \mathbb{R}$. Therefore, we must be cautious in this regard.

Is this space L_1 the same as the one we obtained previously? The answer is yes and it follows trivially from the following theorems.

Theorem 8.4 (Riesz-Fischer) *: L_1 is complete.*

Theorem 8.5 *$C^1[a,b]$ is dense in $L^1[a,b]$, that is, $L^1[a,b]$ is the completed space of $C^1[a,b]$.*

8.4 Hilbert Spaces

Hilbert spaces are Banach spaces[2] whose norm comes from an inner product, that is, from a map $\langle\cdot,\cdot\rangle$ from $V\times V\to C$ satisfying:

i) $\langle x, y+cz\rangle = \langle x,y\rangle + c\langle x,z\rangle$,
$\langle x+cy, z\rangle = \langle x,z\rangle + \bar{c}\langle y,z\rangle$
for any x,y,z in V and c in C.

ii) $\overline{\langle x,y\rangle} = \langle y,x\rangle$

iii) $\langle x,x\rangle \geq 0$ (0 sii $x=0$).

The first part of condition *i)* indicates that the inner product map is linear with respect to its second argument. The second part indicates that it is linear in the first argument, except for the fact that the scalar is taken as the complex conjugate. This map is said to be **anti-linear** with respect to the first argument. Condition *ii)* tells us that the map is as symmetric as it can be, given that it is anti-linear in the first argument. This condition guarantees that $\langle x,x\rangle$ is a real number and, along with *iii*), that it is non-negative.

Exercise: Prove that given a tensor of type $(2,0)$, $\boldsymbol{t}(\cdot,\cdot)$, symmetric, real, and positive definite, then $\langle \boldsymbol{x},\boldsymbol{y}\rangle := \boldsymbol{t}(\bar{\boldsymbol{x}},\boldsymbol{y})$ is an inner product.

The norm induced by this inner product is simply the function, $||\cdot||_V : \boldsymbol{V}\to\mathbb{R}^+$ given by,

$$||x||_V = \sqrt{\langle x,x\rangle}. \tag{8.14}$$

That this is indeed a norm follows from the following lemmas:

Lemma 8.1 (Schwarz Inequality) *: $|\langle x,y\rangle| \leq ||x||\,||y||$.*

Proof: For any $x,y\in V$, $\lambda\in\mathbb{R}$ we have by *iii)*,

$$\begin{aligned} 0 &\leq \langle y+\lambda\langle x,y\rangle x, y+\lambda\langle x,y\rangle x\rangle \\ &= ||y||^2+\lambda^2|\langle x,y\rangle|^2\,||x||^2 + \\ &\quad +\lambda\overline{\langle x,y\rangle}\langle x,y\rangle + \lambda\langle x,y\rangle\langle y,x\rangle \\ &= ||y||^2 + 2\lambda|\langle x,y\rangle|^2 + \lambda^2|\langle x,y\rangle|^2\,||x||^2. \end{aligned} \tag{8.15}$$

[2]From now on and essentially for the same reason we gave in the case of the Jordan canonical form theorem, we will consider complex vector spaces.

Since this relation must hold for all $\lambda \in \mathbb{R}$, the discriminant of the polynomial in λ on the right must be non-positive, that is,

$$4|\langle x,y\rangle|^4 - 4||x||^2||y||^2|\langle x,y\rangle|^2 \leq 0. \tag{8.16}$$

which gives us the desired inequality.

Lemma 8.2 (Triangle Inequality) *:* $||x+y|| \leq ||x|| + ||y||$.

Proof:

$$\begin{aligned} ||x+y||^2 &= ||x||^2 + ||y||^2 + 2\,\mathrm{Re}\,\langle x,y\rangle \\ &\leq ||x||^2 + ||y||^2 + 2\,|\langle x,y\rangle| \\ &\leq ||x||^2 + ||y||^2 + 2\,||x||\,||y|| \\ &\leq (||x|| + ||y||)^2, \end{aligned}$$

which gives us the desired inequality.
Examples:

1. C^n, the vector space of n-tuples of complex numbers.
   Let $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$, then

$$\langle x,y\rangle = \sum_{j=1}^{n} \bar{x}_j y_j.$$

2. l^2, the vector space of sequences of complex numbers $\{x_i\}$ such that

$$||\{x_j\}||_2 := \sqrt{\sum_{j=1}^{\infty} |x_j|^2} < \infty,$$

   with the inner product,

$$\langle \{x_j\}, \{y_j\} \rangle = \sum_{j=1}^{\infty} \bar{x}_j\, y_j.$$

   Note that the Schwarz inequality guarantees that the inner product is well-defined for any pair of vectors in l^2.

   To ensure that l^2 is a Hilbert space, we must prove that it is complete, that is, that every Cauchy sequence (with respect to the l^2 norm) converges to an element of l^2.

   Lemma 8.3 *l^2 is a Hilbert space.*

Proof: Let $\{\{x_i\}_N\}$ be a sequence of sequences. That this is Cauchy means that given $\varepsilon > 0$ there exists $\bar{N}$ such that

$$||\{x_i\}_N - \{x_i\}_M||^2 = \sum_{i=1}^{\infty} |x_i^N - x_i^M|^2 < \varepsilon^2 \quad \forall \quad N, M > \bar{N},$$

but this implies that for each i

$$|x_i^N - x_i^M| < \varepsilon \tag{8.17}$$

that is, the sequence of complex numbers (in N, fixed i) $\{x_i^N\}$ is Cauchy. But the complex plane is complete and therefore for each i, $\{x_i^N\}$ converges to a complex number which we will denote $\bar{x}_i$

Let $\{\bar{x}_i\}$ be the sequence (in i) of these numbers, we will now prove that $\{\bar{x}_i\} \in l^2$ and that $\{\{x_i\}_N\} \longrightarrow \{\bar{x}_i\}$ as $N \to \infty$.

Taking the limit $N \to \infty$ we see that if $M > \bar{N}$ then

$$\sum_{j=1}^{k} |x_j^M - \bar{x}_j|^2 < \varepsilon^2$$

But this is an increasing sequence (in k) of real numbers that is bounded (by ε^2) and therefore convergent. Now taking the limit $k \to \infty$ we see that $\{\{x_i\}_N\} - \{\bar{x}_i\} \in l^2$ and that if $\{\bar{x}_i\} \in l^2$ then $\{\{x_i\}_N\} \to \{\bar{x}_i\}$ in norm. But

$$\begin{aligned} ||\{\bar{x}_i\}||^2 &= ||\{\{x_i\}_N\} + (\{\bar{x}_i\} - \{\{x_i\}_N\})||^2 \\ &\leq ||\{\{x_i\}_N\}||^2 + ||\{\bar{x}_i\} - \{\{x_i\}_N\}||^2, \end{aligned}$$

and therefore $\{\bar{x}_i\} \in l^2$ ♠

This example and the next one are classic examples to keep in mind, essentially every Hilbert space we will deal with is some variant of these.

3. L^2 (or H^0), the space of measurable functions with square integrable in $\mathbb{R}$ and identified with each other if their difference is in a set of measure zero ($f \sim g$ if $\int |f - g|^2 \, dx = 0$). The inner product is $\langle f, g \rangle = \int \bar{f} \, g \, dx$ and its norm is obviously $||f||_{H^0} = \sqrt{\int |f|^2 \, dx}$.

4. (Sobolev Spaces) Let the norm be

$$\begin{aligned} ||f||_{H^m}^2 &= \int_\Omega \{|f|^2 + \sum_{i=1}^{n} |\partial_i f|^2 + \sum_{i,j=1}^{n} |\partial_i \partial_j f|^2 + \cdots \\ &+ \underbrace{\sum_{i,j,\ldots,k=1}^{n}}_{m} |\partial_i \partial_j \ldots \partial_k f|^2\}, \end{aligned}$$

where the partial derivatives are with respect to a Cartesian coordinate system in $\mathbb{R}^n$. We define the Sobolev space of order m as, $\boldsymbol{H}^m(\Omega) = \{$ Completion of the space of functions differentiable m times in $\Omega \subset \mathbb{R}^n$ with respect to the norm $||\ ||_{H^m}\}$

It is obvious what the corresponding inner product is, also note that by definition H^m is complete. H^0 coincides with the one defined in the previous example, as continuous functions are dense in L^2.

As we see from these examples, Hilbert spaces are the immediate generalization of $\mathbb{R}^n$ (or C^n) to infinite dimensions where we have preserved the notion not only of the magnitude of a vector but also of the angle between two vectors. This makes Hilbert spaces have more interesting properties than Banach spaces in general. The most interesting one refers to the subspaces of H. Let M be a closed subspace of H. This subspace inherits the inner product defined in H (simply by restricting the map $\langle\cdot,\cdot\rangle$ to act only on elements of M) and being closed it is complete, therefore it is also a Hilbert space.

Examples:

a) Let M be the subspace generated by the vector $(1,0,0)$ in C^3, that is, all vectors of the form $(c,0,0)$ with $c \in C$.

b) Let M be the subspace of H^0 consisting of all functions that vanish in the interval $(0, 1)$ except on a subset of null measure. M is closed, since if a Cauchy sequence of functions vanishes in $(0, 1)$ then the limit function (which exists since H is complete) also vanishes in that interval and therefore is in M.

c) Let $M = C[a,b]$ be the subspace of continuous functions of $H^0([a,b])$. This is not closed and therefore not a Hilbert space. [Show, by finding a Cauchy sequence of continuous functions that does not have a continuous function as its limit, that this subspace is not closed.]

d) Let K be any subset of H and consider the intersection of all closed subspaces of H that contain K. This intersection gives us the smallest subspace that contains K and is called the subspace generated by K. In example a) $K = \{(1,0,0)\}$ generates the complex plane containing this vector. In example c) $K = C[a,b]$ generates all of $H^0([a,b])$.

The concept defined above does not use the inner product and therefore is also valid for Banach spaces in general (a closed subspace of a Banach space is also a Banach space).

The inner product allows us to introduce the concept of orthogonality and thus define the orthogonal complement of M, that is, the set,

$$M^\perp = \{x \in H \,|\, \langle x, y\rangle = 0 \ \forall\ y \in M\}, \tag{8.18}$$

which has the following property,

Theorem 8.6 *Let M be a closed subspace of a Hilbert space H. Then $M^\perp$ is also a Hilbert space. Moreover, M and $M^\perp$ are complementary, that is, every vector in H can be uniquely written as the sum of a vector in M and another in $M^\perp$.*

$$H = M \oplus M^\perp \tag{8.19}$$

Proof:

It is immediate that $M^\perp$ is a vector subspace and that it is closed. [If $\{x_i\}$ is a sequence in $M^\perp$ converging to x in H then $|\langle x, y\rangle| = |\langle x - x_i, y\rangle| \leq ||x - x_i||\,||y|| \to 0 \ \forall\, y \in M$ and therefore $x \in M^\perp$.] We only need to prove the complementarity. To do this, given $x \in H$, we will look for the element z in M closest to x, see figure.

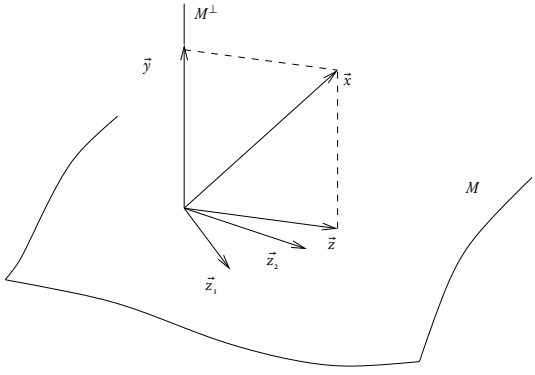

Figure 8.4: The perpendicular space.

Let $d = inf_{w \in M} ||x - w||_H$ and choose a sequence $\{z_n\}$ in M such that $||x - z_n||_H \to d$. This is possible by the definition of infimum.

Then

$$\begin{aligned}
||z_n - z_m||_H^2 &= ||(z_n - x) - (z_m - x)||_H^2 \\
&= 2||z_n - x||_H^2 + 2||z_m - x||_H^2 \\
&\quad -||(z_n - x) + (z_m - x)||_H^2 \\
&= 2||z_n - x||_H^2 + 2||z_m - x||_H^2 - 4||\tfrac{z_n + z_m}{2} - x||^2 \\
&\leq 2||z_n - x||_H^2 + 2||z_m - x||_H^2 - 4d^2 \\
&\underset{\substack{m\to\infty \\ n\to\infty}}{\longrightarrow} 2d^2 + 2d^2 - 4d^2 = 0
\end{aligned}$$

The second equality comes from the so-called parallelogram law, see figure.

This tells us that $\{z_n\}$ is Cauchy and therefore converges to a unique z in M.

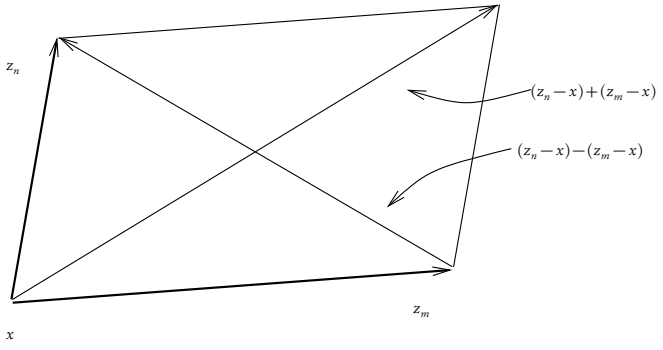

Figure 8.5: Parallelogram law.

Let $y = x - z$, it only remains to see that y is in $M^\perp$, that is, $\langle y, v\rangle = 0 \ \forall \ v \in M$. Let $v \in M$ and $t \in I\!R$, then

$$\begin{aligned} d^2 &:= ||x-z||^2 \leq ||x-\langle z + t\, v\rangle||^2 = ||y - t\, v||^2 \\ &= d^2 - 2\, t\, Re\langle y, v\rangle + t^2\, ||v||^2 \end{aligned} \tag{8.20}$$

and therefore $-2t\, Re\langle y, v\rangle + t^2\, ||v||^2 \geq 0 \ \forall \ t$ which implies that $Re\langle y, v\rangle = 0$. Taking it we get that $Im\langle y, v\rangle = 0$ and the proof is complete ♠

Problem 8.1 *In the proof, only the parallelogram law was used,*

$$||x+y||^2 + ||x-y||^2 = 2(||x||^2 + ||y||^2) \tag{8.21}$$

Show that a norm satisfies it if and only if it comes from an inner product. Hint: Use the so-called polarization identity to define an inner product from the norm,

$$\langle x, y\rangle = \frac{1}{4}\{[||x+y||^2 - ||x-y||^2] - \Im[||x+\Im y||^2 - ||x-\Im y||^2] \tag{8.22}$$

This theorem has important corollaries that we will see below.

Corolary 8.1 $(M^\perp)^\perp = M$

Proof:

We will prove this by showing that $M \subset (M^\perp)^\perp$ and that $(M^\perp)^\perp \subset M$. The first inclusion is obvious since $(M^\perp)^\perp$ is the set of vectors orthogonal to $M^\perp$ which in turn is the set of vectors orthogonal to M. The second inclusion follows from decomposing any vector $x \in (M^\perp)^\perp$ into its part in M and part in $M^\perp$, $x = x_1 + x_2$, $x_1 \in M$, $x_2 \in M^\perp$, the first inclusion tells us that $x_1 \in (M^\perp)^\perp$ but $(M^\perp)^\perp$ and $M^\perp$ are also complementary and therefore $x_2 = 0$ ♠

This tells us that by taking complements we will only obtain one extra Hilbert space.

To formulate the next corollary, it is necessary to introduce a new concept, the dual space of a Hilbert space. We could define the dual of a Hilbert space in the same way as we did for finite-dimensional vector spaces, that is, as the set of linear maps from H to $\mathbf{C}$. The vector space thus obtained does not inherit any interesting property from H because it is too large. To achieve a smaller space with attractive properties, we will restrict the linear maps to those that are continuous. That is, the **dual space** of H, H' will be the set of continuous linear maps from H to $\mathbf{C}$. Which maps are in H'? Note that if $y \in H$ then the map $\varphi_y : H \to \mathbf{C}$ defined by $\varphi_y(x) = \langle y, x\rangle$ is linear and since $|\varphi_y(x)| \leq ||y||_H \, ||x||_H$ it is also continuous. This map gives us an injective (not canonical, since it depends on the inner product) correspondence between the elements of H and those of its dual. Thus we see that H is naturally contained in H'.

Exercise: Reflect on what the natural norm in H' is. Prove that with this norm we have $||\phi_y||_{H'} = ||y||_H$.

Problem 8.2 *Let $\phi : H \to C$ be a linear map. Show that the continuity of the map at the origin ensures the continuity of the map at every point.*

If the dimension of H is finite, then we know that H' will have the same dimension and we will have a (not canonical) one-to-one correspondence between vectors and covectors. In principle, this does not have to be the case in the infinite-dimensional case, and indeed if we use an analogous definition and define the dual of a Banach space, in general, there will be no relationship between it and the original space. In the case of Hilbert spaces, everything is simpler, as shown by the following corollary.

Corolary 8.2 (Riesz Representation Theorem) *: Let $\varphi \in H'$, then there exists a unique $y \in H$ such that $\varphi(x) = \langle y, x\rangle \forall\ x \in H$, that is, $H \approx H'$ in the sense that there exists a natural invertible map between H and H'.*

Proof: Let $M = \{x \in H \text{ such that } \varphi(x) = 0\}$. Since φ is linear, M is a subspace of H, and since φ is continuous, it is closed. By the previous theorem, $M^\perp$ is a Hilbert space and complements M. If $M^\perp$ is zero, then $M = H$ and φ is the zero map and $y = 0$ will be its representative in H. Suppose then that $M^\perp \neq \{0\}$ and choose $w \in M^\perp$ such that $\varphi(w) = 1$[3]. For any $v \in M^\perp$, $v - \varphi(v)w \in M^\perp$, but $\varphi(v - \varphi(v)w) = 0$ and therefore also $v - \varphi(v)w \in M$. Thus we conclude that $v - \varphi(v)w = 0\ \forall\ v \in M^\perp$, that is, $v = \varphi(v)w\ \forall\ v \in M^\perp$, meaning that $M^\perp$ is one-dimensional. Now let's see that $\varphi(x) = \langle \frac{w}{||w||^2}, x\rangle\ \forall x \in H$. Indeed, by the previous theorem, $x = \alpha w + y$ with

[3]The reader should convince themselves that if φ is not the zero map, then there always exists $w \in H$ such that $\varphi(w) = 1$.

$\alpha \in C$ and $y \in M$, and therefore, $\varphi(x) = \varphi(\alpha w + y) = \alpha\varphi(w) + \varphi(y) = \alpha$, but on the other hand, $\langle \frac{w}{||w||^2}, x\rangle = \alpha$ ♠

Exercise: Conclude the proof by proving uniqueness.

To state the third corollary, we need the concept of an orthonormal basis. An **orthonormal basis** of H is a subset of vectors of H that have norm one, are mutually orthogonal, and generate H (that is, given $x \in H$ and $\varepsilon > 0$ there exists a **finite** linear combination of elements of this basis y such that $||x - y||_H < \varepsilon$.

Example: In l^2 let $e_1 = (1, 0, 0, ...)$, $e_2 = (0, 1, 0, ...)$, etc.

Exercise: Show that if $a \in l^2$ then $||a - \sum_{n=1}^{i} \langle e_n, a\rangle e_n||_{l^2_{i\to\infty}} \longrightarrow 0$.

Corolary 8.3 *Every Hilbert space has an orthonormal basis.*

This result, like the previous one, is very powerful because it tells us that we can always approximate different elements of H using a given basis. We will not prove this Corollary because it requires tools that will not be covered in this course. But we will prove another simpler result, for which we introduce the following definition.

Definition: We will say that a Banach space (and in particular a Hilbert space) is **separable** if it has a dense subset with a countable number of elements.
Examples:

a) The subset S of sequences l^2 that have a finite number of *rational* elements is dense in l^2 [since as we will see later, any element of l^2 can be expressed as the limit of a sequence S] and countable [since the rationals are countable].

b) The subset $S(\mathbb{R})$ consisting of functions $f_{r,s,t}(x) = re^{-s|x-t|}$ (note that these are smooth), where r, s, t are rational numbers and $s > 0$, is dense and countable in $L^2(\mathbb{R})$.

Most Hilbert spaces that appear in practice are separable and therefore have the following property:

Theorem 8.7 *A Hilbert space H is separable if and only if it has a countable orthonormal basis S. If S has a finite number of elements, N, then H is isomorphic to C^N (that is, there exists a map $\varphi : H \to C^N$, in this case linear, with $||\varphi(x)||_{C^N} = ||x||_H \forall x \in H$ that is continuous and invertible and its inverse is also continuous). If S has an infinite number of elements, then H is isomorphic to l^2.*

This theorem tells us that among separable Hilbert spaces, essentially there is no more structure than that already present in C^N or l^2.

Proof: To obtain the orthonormal (countable) basis, we first discard some elements from a dense and countable subset of H, S, until we obtain a subcollection of vectors

(by construction countable) that are linearly independent and still span S. Then we apply the Gram-Schmidt process to this subcollection to obtain the orthonormal basis. To prove the rest of the theorem, we need the following lemmas.

Lemma 8.4 (Pythagoras) *: Let $\{x_n\}_{n=1}^N$ be an orthonormal set in H, not necessarily a basis. Then,*

$$||x||_H^2 = \sum_{n=1}^{N} |\langle x_n, x\rangle|^2 + ||x - \sum_{n=1}^{N} \langle x_n, x\rangle x_n||^2 \quad \forall \ x \in H. \tag{8.23}$$

Proof: Let $v = \sum_{n=1}^N \langle x_n, x\rangle x_n$ and $w = \langle x - \sum_{n=1}^N \langle x_n, x\rangle x_n\rangle$. It is easy to see that $v \in V$, the Hilbert subspace generated by the $\{x_n\}_{n=1}^N$ and $w \in V^\perp$, but then

$$\begin{aligned} ||x||_H^2 &= \langle x, x\rangle \\ &= \langle v + w, v + w\rangle \\ &= \langle v, v\rangle + \langle w, w\rangle \\ &= \sum_{n=1}^{N} |\langle x_n, x\rangle|^2 + ||w||_H^2. \end{aligned}$$

♠

Lemma 8.5 (Bessel's Inequality) *:*
Let $\{x_n\}_{n=1}^N$ be an orthonormal set in H, not necessarily a basis. Then,

$$||x||_H^2 \geq \sum_{n=1}^{N} |\langle x, x_n\rangle|^2 \quad \forall\, x \in H \tag{8.24}$$

Proof: Obvious conclusion from the previous lemma ♠

Lemma 8.6 *Let $S = \{x_n\}$ be a countable orthonormal basis[4] of H. Then $\forall\, y \in H$,*

$$y = \sum_n \langle x_n, y\rangle x_n \tag{8.25}$$

and

$$||y||^2 = \sum_n |\langle x_n, y\rangle|^2, \tag{8.26}$$

where the first equality means that the sum converges with respect to the norm of H to $y \in H$. The last equality is called **Parseval's relation** *and the coefficients $\langle x_n, y\rangle$ are the* **Fourier coefficients** *of y. Conversely, if $\{c_n\} \in l^2$ then $\sum_n c_n x_n \in H$.*

[4] A similar result is valid even if the basis is not countable, that is, even if H is not separable.

Proof: From Bessel's inequality, it follows that given any finite subset $\{x_n\}_{n=1}^N$,

$$a_N = \sum_{n=1}^{N} |\langle x_n, y\rangle|^2 \leq ||y||^2 \tag{8.27}$$

which implies that the sequence $\{a_N\}$, which is monotonically increasing, is bounded and therefore converges to a finite limit. This in turn implies that $\{a_N\}$ is Cauchy. Let $y_N = \sum_{n=1}^{N} \langle x_n, y\rangle x_n$, then for $N > M$

$$||y_N - y_M||_H^2 = || \sum_{j=M+1}^{N} \langle x_j, y\rangle x_j ||_H^2 = \sum_{j=M+1}^{N} |\langle x_j, y\rangle|^2 = a_N - a_M$$

The convergence of $\{a_N\}$ thus guarantees that $\{y_N\}$ is a Cauchy sequence and therefore converges to an element y' of H. Let us then prove that $y' = y$. But for any element of S, x_n

$$\begin{aligned} \langle y - y', x_n\rangle &= \lim_{N\to\infty} \langle y - \sum_{j=1}^{N} \langle x_j, y\rangle x_j, x_n\rangle \\ &= \langle y, x_n\rangle - \langle y, x_n\rangle = 0 \end{aligned} \tag{8.28}$$

and therefore $y - y'$ is perpendicular to the space generated by S, which is all of H and therefore must be the zero element. It only remains to see that if $\{c_n\} \in l^2$ then $\sum_{n=1}^{\infty} c_n x_n \in H$. An identical calculation to the previous one shows that $\{y_N = \sum_{n=1}^{N} c_n x_n\}$ is a Cauchy sequence and therefore that $\lim_{N\to\infty} y_N \in H$ ♠

We now continue the proof of Theorem 8.7 The last argument of Lemma 8.6 shows us that given a countable basis $\{x_n\}_{n=1}^{\infty}$ of H, the countable set $Span_{\mathbb{Q}}\{S\} = \{x | x = \sum_{n=1}^{\infty} c_n x_n$ with $\{c_n\}$ a finite sequence of rationals $\}$ is dense in H and therefore H is separable. This concludes the *iff* part of the proof, it only remains to find an isomorphism between H and C^N or l^2. Let $\{x_n\}$ be a countable basis of H and let $\varphi : H \to C^N$ or l^2 be the map

$$\varphi(y) = \{\langle x_n, y\rangle\}. \tag{8.29}$$

If the basis is finite, with dimension N, this is clearly a map into C^N. If the basis is infinite, the image of H under φ is included in the space of infinite complex sequences, but using Lemma 8.6 we see that,

$$||\varphi(y)||_{l^2}^2 = \sum_{n=1}^{\infty} |\langle x_n, y\rangle|^2 = ||y||_H^2 \tag{8.30}$$

and therefore that this image is contained in l^2. The continuity and invertibility of the map are left as an exercise for the reader ♠

The previous theorem, among other things, serves to give a characterization that differentiates finite-dimensional Hilbert spaces from infinite-dimensional ones. (In fact, this characterization is valid for Banach spaces in general).

Theorem 8.8 *Let H be a Hilbert space and let*

$$B_1(H) = \{x \in H / \ ||x||_H \leq 1\}$$

the unit ball in H. Then $B_1(H)$ is compact iff H is finite-dimensional. [It can be seen that in this case a subset B of H is compact iff every sequence $\{x_n\}$ of elements of H in B has a subsequence that converges to an element of B.]

Proof: We will only see the case where the space is separable. If H is finite-dimensional, it is isomorphic to C^N, but $B_1(C^N)$ is a closed and bounded subset, and therefore compact. If H is infinite-dimensional (and separable), then it is isomorophic to l^2, but the sequence in $B_1(l^2)$, $\{x_n\}$ with $\{x_n = (0, ..., 0, \underset{n}{1}, 0, ...)\}$ clearly has no convergent subsequence ♠

Exercise: See that no subsequence of the previous sequence is Cauchy.

Example: Let $H = L^2([0, 2\pi])$ and $S = \{\frac{1}{\sqrt{2\pi}} e^{inx}, n = 0, \pm 1, \pm 2, \ldots\}$. The development of this example will be our occupation until the end of this chapter.

Exercise: Show that the elements of $L^2(0, 2\pi)$ of the form $f_n(x) = \frac{1}{\sqrt{2\pi}} e^{inx}$, $n = 0, \pm 1, \pm 2, ...$, form an orthonormal set.

8.5 Fourier Series

Theorem 8.9 $S = \left\{ \frac{e^{inx}}{\sqrt{2\pi}}, ; n = 0, \pm 1, \pm 2 \ldots \right\}$ *is a complete set of orthonormal vectors (i.e., an orthonormal basis) of $L^2[0, 2\pi]$.*

Note that by Lemma 8.6 this is equivalent to ensuring that if $f \in L^2[0, 2\pi]$ then $f_N(x)$ converges (in the norm of $L^2[0, 2\pi]$) to f when $N \to \infty$.

Proof: The orthogonality is trivial and part of a previous exercise, so we will only focus on completeness. The space of Lipschitz functions on $[0, 2\pi]$, $Lip[0, 2\pi]$ is dense [essentially by definition! since it contains all smooth functions] in $L^2[0, 2\pi]$, but then its subspace consisting of periodic functions $Lip_p[0, 2\pi]$ is also dense [since the elements of $L^2[0, 2\pi]$ are equivalence classes of functions and in each class, there is always one that is periodic]. Now suppose that

$$S_N(g)(x) := \frac{1}{\sqrt{2\pi}} \sum_{n=-N}^{N} g_n e^{inx} \to g(x) \tag{8.31}$$

pointwise and uniformly if $g(x) \in Lip_p([0, 2\pi])$. Then by density, given any function $f(x) \in L^2([0, 2\pi])$ and $\varepsilon > 0$ there will exist $f_\varepsilon \in Lip_p([0, 2\pi])$ such that $|f - f_\varepsilon|_{L^2} < \varepsilon/3$ and therefore we will have,

$$\begin{aligned} |f - S_N(f)|L^2 &= |f - S_N(f) - f\varepsilon + S_N(f_\varepsilon) + f_\varepsilon - S_N(f_\varepsilon)|L^2 \\ &\leq |f - f\varepsilon|L^2 + |f\varepsilon - S_N(f_\varepsilon)|L^2 + |S_N(f - f\varepsilon)|L^2 \end{aligned}$$

but

$$\begin{aligned} |S_N(f - f\varepsilon)|L^2 &= \sum n = -N^N |(f - f_\varepsilon)n|^2 \\ &\leq |f - f\varepsilon|^2 \\ &\leq \varepsilon/3 \end{aligned} \tag{8.32}$$

and therefore choosing N such that $|f_\varepsilon(x) - S_N(f_\varepsilon)(x)| < \frac{\varepsilon}{3}$, which follows from our previous assumption (still not proven), we have that $|f_\varepsilon - S_N(f_\varepsilon)|L^2 < \frac{\varepsilon}{3}$ and therefore we conclude that for such N, $|f - S_N(f)|L^2 < \varepsilon$ ♠

Thus we see that it only remains to prove our assumption of uniform convergence of $S_N(f)$ to f for periodic Lipschitz functions. The proof of this statement is very instructive and shows us how the Fourier series works. Preparatory to this result, we prove a series of lemmas:

Lemma 8.7 *If f is integrable ($f \in L^1$) then,*

$$\sup_n |f_n| \leq \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} |f(x)| dx. \tag{8.33}$$

Proof:

$$\begin{aligned} |f_n| &= |\langle \frac{e^{inx}}{\sqrt{2\pi}}, f \rangle| \\ &\leq \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} |e^{-inx} f(x)| dx \\ &\leq \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} |f(x)| dx \end{aligned} \tag{8.34}$$

♠

Lemma 8.8 *If f has an integrable derivative, then $f_n \to 0$ as $n \to \infty$.*

Proof: Integrating by parts we have,

$$\begin{aligned} f_n &= \langle \frac{e^{inx}}{\sqrt{2\pi}}, f \rangle \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-inx} f(x) dx \\ &= \frac{1}{in\sqrt{2\pi}} \int_0^{2\pi} e^{-inx} f'(x) dx - \frac{1}{in\sqrt{2\pi}} e^{-inx} f(x)|_0^{2\pi} \\ &= \frac{1}{in\sqrt{2\pi}} \int_0^{2\pi} e^{-inx} f'(x) dx - \frac{1}{in\sqrt{2\pi}} [f(2\pi) - f(0)] \end{aligned}$$

and therefore

$$|f_n| \leq \frac{1}{n\sqrt{2\pi}} [|f'|_{L^1} + |f(2\pi) - f(0)|], \tag{8.35}$$

with which the lemma is proven ♠

This lemma and its generalization to a greater number of derivatives indicate that the more differentiable a function is, the faster its Fourier coefficients decay (asymptotically).

Exercise: Prove a similar lemma that gives a better bound for f_n if the function is periodic and m times differentiable.

Lemma 8.9 (Riemann-Lebesgue) *If $f \in L^1([0, 2\pi])$, i.e., an integrable function. Then,*

$$\lim_{n \to \infty} f_n = 0 \tag{8.36}$$

Proof: If $f \in L^1([0, 2\pi])$, then it is approximable by smooth functions, in particular, given any $\varepsilon > 0$ there exists $f_\varepsilon : [0, 2\pi]$, smooth, such that $|f - f_\varepsilon|_{L^1} < \frac{\varepsilon}{2\sqrt{2\pi}}$. But since,

$$\begin{aligned} |f_n - f_{\varepsilon n}| &= |\langle \frac{e^{inx}}{\sqrt{2\pi}}, f - f_\varepsilon \rangle| \\ &\leq \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} |f(x) - f_\varepsilon(x)| dx \\ &\leq \frac{1}{\sqrt{2\pi}} |f - f_\varepsilon|_{L^1} \\ &\leq \varepsilon/2 \end{aligned} \tag{8.37}$$

we have that

$$|f_n| \leq |f_n - f_{\varepsilon n}| + |f_{\varepsilon n}| \leq \varepsilon/2 + |f_{\varepsilon n}|. \tag{8.38}$$

Applying the previous lemma and noting that f_ε is differentiable, we see that $f_{\varepsilon n} \to 0$ and therefore, given $\varepsilon > 0$ I can choose N such that for all $n > N$ $|f_{\varepsilon n}| < \varepsilon/2$

with which $|f_n| < \varepsilon$ for all $n > N$ and therefore we conclude that $f_n \to 0$ as $n \to \infty$
♠

We are now in a position to prove the theorem on pointwise convergence of Lipschitz functions.

Theorem 8.10 *Let $f : [0, 2\pi] \to C$ be periodic and Lipschitz. Then*

$$S_N(f)(x) := \frac{1}{\sqrt{2\pi}} \sum_{n=-N}^{N} f_n e^{inx} \tag{8.39}$$

converges pointwise and uniformly to $f(x)$.

Proof: We will only prove pointwise convergence, uniform convergence follows easily and its proof adds nothing new. We begin with the following calculation,

$$\begin{aligned} S_N(f)(x) &:= \frac{1}{\sqrt{2\pi}} \sum_{n=-N}^{N} f_n e^{inx} \\ &= \frac{1}{2\pi} \sum_{n=-N}^{N} \int_0^{2\pi} e^{inx'} f(x') e^{inx} dx' \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(x') \Big(\sum_{n=-N}^{N} e^{in(x-x')} \Big) dx' \\ &= \int_0^{2\pi} f(x') D_N(x-x') dx' \end{aligned} \tag{8.40}$$

where we have defined the **Dirichlet kernel**

$$D_N(x-x') := \frac{1}{2\pi} \sum_{n=-N}^{N} e^{in(x-x')}. \tag{8.41}$$

We now study some properties of the Dirichlet kernel. First, note that,

$$\begin{aligned} \int_0^{2\pi} D_N(x-x') dx' &= \frac{1}{2\pi} \sum_{n=-N}^{N} \int_0^{2\pi} e^{in(x-x')} dx' \\ &= \frac{1}{2\pi} \Big[\sum_{n=-N, n\neq 0}^{N} \frac{-i}{n} [e^{in(x-2\pi)} - e^{inx}] + 2\pi \Big] \\ &= 1. \end{aligned} \tag{8.42}$$

On the other hand, we have that

$$\begin{aligned} 2\pi D_N(x) &= \sum_{n=-N}^{N} e^{inx} \\ &= \sum_{n=-N}^{N} (e^{ix})^n \end{aligned} \tag{8.43}$$

and calling $q := e^{ix}$ to simplify the notation, we have,

$$\begin{aligned}
& 2\pi D_N(x) \\
= & \sum_{n=0}^{N} q^n + \sum_{n=0}^{N} q^{-n} - 1 \\
= & \frac{q^{N+1}-1}{q-1} + \frac{q^{-(N+1)}-1}{q^{-1}-1} - 1 \\
= & \frac{(q^{N+1}-1)(q^{-1}-1)+(q^{-(N+1)}-1)(q-1)-(q-1)(q^{-1}-1)}{(q-1)(q^{-1}-1)} \\
= & \left[(q^{N+1/2}-q^{-1/2})(q^{-1/2}-q^{1/2})+(q^{-(N+1/2)}-q^{1/2})(q^{1/2}-q^{-1/2})\right. \\
- & \left.(q^{1/2}-q^{-1/2})(q^{-1/2}-q^{1/2})\right]/(q^{1/2}-q^{-1/2})(q^{-1/2}-q^{1/2}) \\
= & \frac{q^{N+1/2}-q^{-N-1/2}}{q^{1/2}-q^{-1/2}} \\
= & \frac{e^{ix(N+1/2)}-e^{-ix(N+1/2)}}{e^{ix/2}-e^{-ix/2}} \\
= & \frac{\sin((N+1/2)x)}{\sin(x/2)}.
\end{aligned} \tag{8.44}$$

Therefore

$$D_N(x) = \frac{1}{2\pi}\frac{\sin((N+1/2)x)}{\sin(x/2)}. \tag{8.45}$$

We have therefore,

$$\begin{aligned}
S_N(f)(x) &= \int_0^{2\pi} f(x')D_N(x-x')dx' \\
&= \int_0^{2\pi} f(x-x')D_N(x')dx' \\
&= \frac{1}{2\pi}\int_0^{2\pi}[f(x-x')-f(x)]\frac{\sin((N+1/2)x')}{\sin(x'/2)}dx' + f(x)
\end{aligned}$$

where in the second equality we have used that the integral of a periodic function over its period is independent of the point where the integration interval begins, and in the last we have subtracted and added $f(x)$ and used that the integral of the Dirichlet kernel is one (8.42).

If $f(x)$ is Lipschitz, $|f(x-x')-f(x)| \leq k|x'|$ and therefore

$$g_x(x') := \frac{|f(x-x')-f(x)|}{\sin(x'/2)} \tag{8.46}$$

is continuous in $(0, 2\pi)$ and bounded at the ends, therefore integrable. Applying now the Riemann-Lebesgue Lemma we conclude then that the integral tends to zero with N and therefore that $\lim_{N\to\infty} S_N(f)(x) = f(x)$ at every point where $f(x)$ is Lipschitz. ♠

To prove that S is complete, we only need to see that if $\langle e^{inx}, g\rangle = 0 \ \forall\ n$ then $g = 0$. Let $f \in C_p^1[0, 2\pi]$ and c_n be its Fourier coefficients with respect to S, then

$$\langle f, g\rangle = \lim_{M\to\infty} \langle \sum_{-M}^{M} c_n \frac{e^{inx}}{\sqrt{2\pi}}, g\rangle = 0 \tag{8.47}$$

We then conclude that g is orthogonal to all $f \in C_p^1[0, 2\pi]$, but as we saw this space is dense in $L^2[0, 2\pi]$ and therefore by continuity we must have that $\langle f, g\rangle = 0 \ \forall\ f \in H$, that is $g = 0$ ♠

Example [**Application of the Fourier series**] Let S be a metal ring of circumference 2π and let $T_0(\theta)$ be a temperature distribution that we assume to be square integrable ($\in L^2(S)$). The temporal evolution of $T(\theta, t)$, (neglecting losses to the surrounding medium by conduction or radiation) is given by,

$$\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial \theta^2} \tag{8.48}$$

with k a positive constant. We assume $T(\theta, 0) = T_0(\theta)$. Assuming also that $T(\theta, t)$ admits a Fourier series decomposition,

$$T(\theta, t) = \sum_{n=-\infty}^{\infty} T_n(t)\frac{e^{in\theta}}{\sqrt{2\pi}} \tag{8.49}$$

and that the derivatives can be passed inside the infinite summation we obtain,

$$\begin{aligned} 0 &= \frac{\partial T}{\partial t} - k\frac{\partial^2 T}{\partial \theta^2} \\ &= \sum_{n=-\infty}^{\infty} \left(\frac{dT_n(t)}{dt} + kn^2 T_n(t)\right)\frac{e^{in\theta}}{\sqrt{2\pi}} \end{aligned} \tag{8.50}$$

from the linear independence of the basis elements, or simply by taking the inner product with $\frac{e^{in\theta}}{\sqrt{2\pi}}$ we see that it must be satisfied,

$$\frac{dT_n(t)}{dt} = -kn^2 T_n(t) \quad \forall\ n \tag{8.51}$$

that is,

$$T_n(t) = T_n(0)e^{-kn^2 t}. \tag{8.52}$$

If the initial temperature was

$$T_0(\theta) = \sum_{n=-\infty}^{\infty} T_n^0 \frac{e^{in\theta}}{\sqrt{2\pi}} \tag{8.53}$$

we then see that $T_n(t) = T_n^0 e^{-kn^2 t}$ and

$$T(\theta, t) = \sum_{n=-\infty}^{\infty} T_n^0 e^{-kn^2 t} \frac{e^{in\theta}}{\sqrt{2\pi}} \tag{8.54}$$

gives us the solution to the problem.

Note that: a) The temperature distribution in the bar only depends on the initial distribution. b) The fact that we were able to reduce a partial differential problem to an ordinary differential problem is due to the fact that we chose to represent the functions in a basis of eigenvectors of the derivative operator. Indeed, what we have fundamentally used is that $\frac{\partial}{\partial\theta} e^{in\theta} = ine^{in\theta}$. c) No matter how bad (non-differentiable) the initial distribution $T_0(\theta)$ is, as long as it is in $L^2(S)$, the solution smooths out for positive times. Indeed, if for example the initial distribution is only Lipschitz, then for all $t > 0$ the solution is infinitely differentiable, both in t and in θ. To see this, for example, take,

$$\frac{\partial^p T(\theta, t)}{\partial t^p} = \sum_{n=-\infty}^{\infty} T_n^0 (-kn^2)^p e^{-kn^2 t} \frac{e^{in\theta}}{\sqrt{2\pi}}, \tag{8.55}$$

but since for $t > 0$, $n^{2p} e^{-kn^2 t} \to 0$ quickly as $n \to \infty$ the series converges absolutely. On the contrary, for a generic initial data, even infinitely differentiable, the solution does not exist for negative times, since in that case the series coefficients grow rapidly with n.

Exercise: Prove that an orthogonal set $\{x_n\}$ is complete if and only if $(x_n, g) = 0 \ \forall n \Rightarrow g = 0$.

Problem 8.3 *: Use Gram-Schmidt to obtain an orthonormal basis from the monomials*

$$1, x, x^2, \ldots, x^n, \ldots, \tag{8.56}$$

with respect to the Hilbert spaces obtained from the following inner products:

1. $\langle f, g\rangle = \int_{-1}^{1} \bar{f} g \, dx$ *(In this case you will obtain the Legendre polynomials.)*
2. $\langle f, g\rangle = \int_{-\infty}^{\infty} \bar{f} g e^{-x^2} \, dx$ *(In this case you will obtain the Hermite polynomials.)*
3. $\langle f, g\rangle = \int_{0}^{\infty} \bar{f} g e^{-x} \, dx$ *(In this case you will obtain the Laguerre polynomials.)*

These polynomial bases are generically called systems of Chebyshev polynomials.

The fact that the Legendre polynomials are a basis follows from the Weierstrass Approximation Theorem and the fact that continuous functions are dense in L^2.

***Gram – Schmidt Method.**
Given a countable set of linearly independent elements of H, $\{\boldsymbol{x}_i\}$ we recursively generate the following sets:

$$\boldsymbol{y}_i = \boldsymbol{x}_i - \sum_{l=1}^{i-1} \langle \boldsymbol{u}_l, \boldsymbol{x}_i \rangle \boldsymbol{u}_l, \quad \boldsymbol{u}_i = \frac{\boldsymbol{y}_i}{||\boldsymbol{y}_i||}.$$

Note that the second operation is well defined since $\boldsymbol{y}_i \neq 0$. This assertion follows from the fact that the right-hand side in the definition of $\boldsymbol{y}_i$ is a linear combination of the $\boldsymbol{x}_j$, $j = 1, ..., i$ and we have assumed that these are linearly independent.
Now let's see that $(\boldsymbol{u}_i, \boldsymbol{u}_j) = \delta_{ij}$. To do this, we will prove by induction that given i, $(\boldsymbol{u}_i, \boldsymbol{u}_j) = 0 \; \forall \; j < i$ positive. This is true for $i = 1$ [since there is no j]. Suppose then that it is true for $i - 1$ and let's see that it is also true for i. But, given $j < i$ we have,

$$\langle \boldsymbol{u}_j, \boldsymbol{y}_i \rangle = [\langle \boldsymbol{u}_j, \boldsymbol{x}_i \rangle - \sum_{l=1}^{j-1} \langle \boldsymbol{u}_l, \boldsymbol{x}_i \rangle \langle \boldsymbol{u}_j, \boldsymbol{u}_l \rangle] = [\langle \boldsymbol{u}_j, \boldsymbol{x}_i \rangle - \langle \boldsymbol{u}_j, \boldsymbol{x}_i \rangle] = 0.$$

If the starting set is a complete set, that is, a set that is not a subset of a larger set of linearly independent vectors, then the resulting set is an orthogonal basis. In particular, if $(\boldsymbol{x}, \boldsymbol{u}_i) = 0 \; \forall i$, then $\boldsymbol{x} = 0$. Otherwise, we take $\boldsymbol{u} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$ and we would have another element of the basis, which would be a contradiction with respect to the completeness of the $\{\boldsymbol{u}_i\}$.

8.6 Problems

Problem 8.4 *Let $\phi : H \to C$ be a linear map. Show that ϕ is continuous if and only if it is bounded.*

Problem 8.5 *Show that the map $I : C[a, b] \to \Re$ given by*

$$I(f) := \int_a^b f(x) dx, \tag{8.57}$$

is a linear and continuous map.

Problem 8.6 *Let V be a finite-dimensional space and let $\{\boldsymbol{u}_i\}$, $i = 1..n$ be a basis and $\{\theta^i\}$, $i = 1..n$ the corresponding dual basis. Let $\boldsymbol{x} = \sum_{i=1}^n x^i \boldsymbol{u}_i$ be any vector in V and $\boldsymbol{\omega} = \sum_{i=1}^n \omega_i \theta^i$ be any linear functional, i.e., an element of V'. Consider the norm in V*

$$||\boldsymbol{x}||_p := (\sum_{i=1}^n |x^i|^p)^{\frac{1}{p}}. \tag{8.58}$$

See that this is a norm and prove that the norm induced in V' by this is given by,

$$||\boldsymbol{\omega}||_q := (\sum_{i=1}^n |\omega_i|^q)^{\frac{1}{q}}, \tag{8.59}$$

where

$$\frac{1}{p}+\frac{1}{q}=1 \qquad (p,q\geq 1); \tag{8.60}$$

Hint: Express $\boldsymbol{\omega}(\boldsymbol{x})$ in components with respect to the given basis/dual basis and then use (prove) the inequality:

$$|\sum_{i=1}^{n} x^i\,\omega_i| \leq (\sum_{i=1}^{n}|x^i|^p)^{\frac{1}{p}}(\sum_{i=1}^{n}|\omega_i|^q)^{\frac{1}{q}}. \tag{8.61}$$

Problem 8.7 *Let c_0 be the space of sequences $\{x\}=(x_1,x_2,\dots)$ converging to zero with the norm*

$$||\{x\}||_{c_0} := \sup_i\{|x_i|\}. \tag{8.62}$$

Prove that the dual of the space c_0 is the space l_1 of absolutely summable sequences $\{\omega\}=(\omega_1,\omega_2,\dots)$ with the norm

$$||\{\omega\}||_{l_1} := \sum_{i=1}^{\infty}|\omega_i|. \tag{8.63}$$

Hints: Note that given an element of l_1, $\{\omega\}=(\omega_1,\omega_2,\dots)$, we have a linear functional given by,

$$\boldsymbol{\omega}(\{x\}) := \sum_{i=1}^{\infty} x_i\,\omega_i. \tag{8.64}$$

Prove that this satisfies

$$||\boldsymbol{\omega}|| \leq ||\{\omega\}||_{l_1}. \tag{8.65}$$

Then find an element of norm equal to or less than one in c_0 and with its help see that

$$||\boldsymbol{\omega}|| \geq ||\{\omega\}||_{l_1}. \tag{8.66}$$

from which it is concluded that the norms are the same. It only remains to see that for each element of the dual of c_0, $\boldsymbol{\omega}$, there exists an element of l_1, $\{\omega\}=(\omega_1,\omega_2,\dots)$ such that equation 8.64 holds. To do this, construct a basis of c_0 and the respective basis of its dual. Note: at some point, you will have to use that the considered linear functionals are continuous.

8.7 Fourier Series Problems

Problem 8.8 *Let f be an integrable function of period T. Show that*

$$\int_0^T f(x)dx = \int_a^{T+a} f(x)dx, \quad \forall a\in\mathbb{R} \tag{8.67}$$

Problem 8.9 *a.- Find the Fourier series of the function $f(x):=x$ in the interval $[-\pi,\pi]$.*
b.- Use Parseval's relation to prove that

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\pi^2/6 \tag{8.68}$$

Problem 8.10 *a.- Find the Fourier series of the function* $f(x) := e^{sx}$ *in the interval* $[-\pi, \pi]$.

b.- Use Parseval's relation to prove that

$$\pi cot\, h(\pi s)/s = \sum_{n=-\infty}^{\infty} \frac{1}{s^2 + n^2} \tag{8.69}$$

Problem 8.11 *Let* $S_n : L^2 \to L^2$ *be the map that sends* $f \in L^2$ *to the partial Fourier series,*

$$S_n(f) := \sum_{m=-n}^{n} c_m e^{imx}, \qquad c_m := \frac{1}{2\pi}\langle e^{imx}, f(x)\rangle. \tag{8.70}$$

Show that the S_n *are orthogonal projections and that* $S_n S_m = S_m S_n = S_m$ *if* $m \leq n$.

Problem 8.12 *In this problem, we attempt to prove that the Fourier series of a continuous function is CesÃ ro summable at every point. Let* $f(\theta)$ *be a periodic function in* L^2, $f(\theta) \in L^2[0, 2\pi]$, $c_m := \frac{1}{2\pi}(e^{im\theta}, f(\theta))$ *and* $S_n(f) := \sum_{m=-n}^{n} c_m e^{imx}$

a.- Prove that

$$S_n(f)(\theta) = \frac{1}{2\pi}\int_0^{2\pi} f(\theta + x)\frac{\sin((n + 1/2)x)}{\sin(x/2)}dx \tag{8.71}$$

b.- Let $SS_n(f)(\theta) = \frac{1}{n+1}\sum_0^n S_m(f)(\theta)$ *(CesÃ ro sum), prove that:*

$$SS_n(f)(\theta) = \frac{1}{2\pi(n+1)}\int_0^{2\pi} f(\theta + x)\frac{\sin^2((n+1)x/2)}{\sin^2(x/2)}dx \tag{8.72}$$

c.- Let $K_n(x) = \frac{\sin^2((n+1)x/2)}{2\pi(n+1)\sin^2(x/2)}$ *prove that for all* $\delta > 0$, $K_n(x) \to 0$ *uniformly on* $[\delta, 2\pi - \delta]$.

d.- Prove that $SS_n(f)(\theta_0) \to f(\theta_0)$ *if* f *is bounded and continuous at* θ_0.

e.- Prove that if f *is continuous and periodic, then* $SSL_n(f)(\theta) \to f(\theta)$ *uniformly in* θ. *Hint: remember that if* f *is continuous on* $[0.2\pi]$ *then* f *is uniformly continuous.*

f.- Show that $||f - S_n(f)|| \leq ||f - SSL_n(f)||$ *and conclude that* $S_n(f) \to f$ *in* L^2 *if* f *is continuous.*

Problem 8.13 *Suppose that* $f \in C_p^1([0, 2\pi])$ *and let* $c_n = \langle e^{in\theta}, f\rangle$ *and* $b_n = \langle e^{in\theta}, f'\rangle$.

a.- See that $\sum_{-\infty}^{\infty} |b_n|^2 < \infty$ *and conclude that* $\sum_{-\infty}^{\infty} n^2|c_n|^2 < \infty$.

b.- Prove that $\sum_{-\infty}^{\infty} |c_n| < \infty$.

c.- Prove that $\sum_{m=-n}^{n} c_m e^{mi\theta}$ *is uniformly convergent for* $n \to \infty$.

d.- Use item f.- of the previous problem to conclude that $\sum_{m=-n}^{n} c_m e^{mi\theta} \to 2\pi f(\theta)$ *uniformly.*

Problem 8.14 *Consider the Fourier series expansion for the following function:*

$$f(x) := \begin{cases} -1 & 0 < x < \pi \\ +1 & \pi < x 2\pi \end{cases} \tag{8.73}$$

$f(\theta + 2\pi) = f(\theta)$. *The sum of the first n terms produces a function that has an absolute maximum near* 0 *of height* $1 + \delta_n$. *Show that* $\lim_{n\to\infty} \delta_n \approx 0.18$. *This is known as the Gibbs phenomenon.*

Bibliography notes: These notes are based on the following books: [9], [10], [13], [1] and [14]. This is one of the most beautiful and useful areas of mathematics, fundamental for almost everything, particularly quantum mechanics. Do not fail to delve a little deeper into it, especially I recommend the books [9] and [1] for pleasant reading.

CHAPTER 9

DISTRIBUTIONS

9.1 Introduction

To make the space of square-integrable functions $L^2(\mathbb{R})$ a normed space, it was necessary to generalize the concept of a function (as a map from $\mathbb{R}$ to $\mathbb{R}$) in the sense that the elements of L^2 are only functions defined almost everywhere, that is, equivalent classes of square-integrable functions under the equivalence relation $f \approx g$ if $\int |f-g|^2\, dx = 0$.

Although this generalization is useful since, among other things, it allows us to group functions into Hilbert spaces and thus use the powerful geometric structure they have, it is convenient to consider an even greater generalization which, as we will see later, will provide us with an important tool in formal calculus in mathematical physics.

The generalized functions that we will define below, called distributions, have many interesting properties, among them that the differentiation operation is closed in this space, that is, the derivative of a distribution is another distribution. This is particularly surprising considering that among the distributions there are functions that are not even continuous!

What is the idea behind this generalization? The Riesz representation theorem showed us that the dual of $L^2(\mathbb{R})$ is that same space. Now, if instead of the dual of $L^2(\mathbb{R})$ we consider the dual of a subspace of $L^2(\mathbb{R})$, we will obtain a space larger than $L^2(\mathbb{R})$ and that contains it in a natural way. This linear space, which contains the usual functions, is a space of generalized functions.

It is clear that there are many spaces of generalized functions since we can not only consider different subspaces of $L^2(\mathbb{R})$, but we can also consider different notions of continuity weaker than the continuity coming from the norm of $L^2(\mathbb{R})$ to define the dual spaces.

Which one to study? The answer is: The one that is most convenient for the treatment of the problem for which they are to be used. Here we will deal with those obtained from a fairly small subspace, which results in a sufficiently broad generalization to cover most of the problems in Physics. It is necessary to emphasize that the concept of distribution that we will introduce is not a physical necessity, in the sense that physical theories can be stated using simply infinitely differentiable

functions[1] , but it is a very useful tool that allows a more "condensed" formulation of some of these laws. The subspace of $L^2(\mathbb{R})$ that we will use is that of infinitely differentiable functions with compact support $C_o^\infty(\mathbb{R})$. [Recall that the support of a classical function $f(x)$ is given by the subset of $\mathbb{R}$,

$$Cl\{f^{-1}[\mathbb{R}-\{0\}]\},$$

where Cl means taking the closure. Since the support of a function is automatically closed, being compact (as a subset of $\mathbb{R}$ or $\mathbb{R}^n$) merely means that it is bounded.]

Exercise: Show that $C_o^\infty(\mathbb{R})$ is indeed a vector space.

Exercise: Show that it is also an algebra with respect to the usual product. What is the support of $f \cdot g$?

What notion of continuity will we introduce in the functionals of $C_o^\infty(\mathbb{R})$ to define the dual? Unfortunately, there is no *natural* norm in this space, in particular, there is none in which it is complete [2].

There is, however, a convenient topology in this space. The corresponding notion of continuity of this topology is obtained from the following convergence criterion:

Definition: We will say that a sequence $\{\varphi_n\}$, $\varphi_n \in C_o^\infty(\mathbb{R})$ **converges** to $\varphi \in C_o^\infty(\mathbb{R})$ if:

1. There exists a compact $K \subset \mathbb{R}$ such that $support(\varphi_n) \subset K \quad \forall n$.

2. The sequences $\{\varphi_n^{(p)}\}$ of their derivatives of order p converge uniformly in K to $\varphi^{(p)}$ for all $p = 0, 1, 2, \ldots,$ that is, given p and $\varepsilon > 0$ there exists N such that for all $n > N$ it holds that

$$\sup_{x \in K} |\varphi_n^{(p)}(x) - \varphi^{(p)}(x)| < \varepsilon. \tag{9.1}$$

Note that the first condition restricts us to consider as convergent sequences those that can only converge to a function with compact support. This is fundamental

[1] Here we refer to fundamental theories. There are approximations, such as fluid theory, where distributions appear naturally.

[2] Even in the case that, for example, we used as a norm

$$\|f\| = \sum_{n=0}^{\infty} \frac{1}{n!} \sup_{x \in \mathbb{R}} \{|f^{(n)}(x)|\}$$

we would not obtain a complete space since in this space there are Cauchy sequences that tend to infinitely differentiable functions but whose support is not compact.

for the completeness of the space and for the uniform convergence in the second condition to make sense using the supremum. With this convergence criterion, we associate the following notion of continuity on the functionals of $C_o^\infty(\mathbb{R})$ in $\mathbb{R}$.

Definition: We will say that the functional $F : C_o^\infty(\mathbb{R}) \to \mathbb{R}$ (or C) is **continuous** at $\varphi \in C_o^\infty(\mathbb{R})$ if given any convergent sequence $\{\varphi_n\}$ in $C_o^\infty(\mathbb{R})$ to φ, it holds that

$$F(\varphi_n) \longrightarrow F(\varphi). \tag{9.2}$$

This notion comes from the aforementioned topology.

Exercise: Let B be a Banach space with the notion of convergence given by its norm. Show that in that case the notion of continuity defined above coincides with the usual (ε, δ).

With this notion of continuity, the space $C_o^\infty(\mathbb{R})$ is called the **space of test functions** of $\mathbb{R}$ and denoted by $\mathcal{D}(\mathbb{R})$.

Definition: The dual space to the space of test functions, $\mathcal{D}'$, that is, the space of continuous linear functionals $T : C_o^\infty(\mathbb{R}) \to \mathbb{R}$ is called the space of **distributions**.

Examples:
a) Let f be continuous and let the linear functional

$$T_f(\varphi) = \int_{\mathbb{R}} f\,\varphi\,dx. \tag{9.3}$$

Since

$$|T_f(\varphi)| \leq \left(\int_K |f|\,dx\right) sup_{x\in K}|\varphi| \tag{9.4}$$

where K is any compact containing the support of φ, we see that T_f is continuous and therefore a distribution. Thus, we see that continuous functions give rise to distributions, that is, they are naturally included in the space of distributions.

Exercise: Show that if $f \neq g$ then $T_f \neq T_g$.

b) Let f be integrable (in the sense of Lebesgue), that is, an element of $\mathcal{L}^1(\mathbb{R})$ and let

$$T_f(\varphi) = \int_{\mathbb{R}} f\,\varphi\,dx \quad \forall\ \varphi \in C_o^\infty(\mathbb{R}). \tag{9.5}$$

But $|T_f(\varphi)| \leq \sup_{x\in K}|\varphi| \int_{\mathbb{R}}|f|\,dx$ and therefore if $\{\varphi_n\} \to 0$ then $T_f(\varphi_n) \to 0$ which ensures us (by linearity) that T_f is continuous and thus a distribution. Note

that if $f = g$ almost everywhere ($f \sim g$) then $T_f = T_g$ so we conclude that it is actually the elements of L^1 that define these distributions. The distributions obtained in this way are called **regular**.

c) Let $T_a : C_o^\infty(\mathbb{R}) \to \mathbb{R}$, $a \in \mathbb{R}$, be given by $T_a(\varphi) = \varphi(a)$, this map is clearly linear,

$$T_a(\varphi + \alpha\psi) = \varphi(a) + \alpha\,\psi(a) = T_a(\varphi) + \alpha\, T_a(\psi),$$

and continuous

$$|T_a(\varphi)| = |\varphi(a)| \leq \sup_{x\in\mathbb{R}} |\varphi(x)|$$

and therefore a distribution. This is called the Dirac delta at the point a. Is there any continuous function, f, such that $T_a = T_f$? Suppose so, and that $f(r) \neq 0$ with $r \neq a$. Choosing φ non-zero only in a sufficiently small neighborhood of r that does not contain a and with the same sign as $f(r)$, we obtain $\varphi(a) = 0$ and $T_f(\varphi) \neq 0$ which implies that $f(r) = 0\ \forall r \neq a$ but by continuity we then conclude that $f \equiv 0$ and therefore that $T_f(\varphi) = 0\ \forall\ \varphi \in C_o^\infty(\mathbb{R})$. We see then that this distribution does not come from any continuous function and it can be seen that it actually does not come from any element of $L^1(\mathbb{R})$, thus it is an **irregular distribution**. These are the **extra** elements that the defined generalization gives us. Usually, in formal manipulations, it is pretended that this distribution comes from a function, called the Dirac delta and denoted by $\delta(x-a)$. With it, things like

$$\int_{\mathbb{R}} \delta(x-a)\,\varphi(x)\,dx = \varphi(a). \tag{9.6}$$

are written. As we have seen, there is actually no function for which this expression makes sense, so it is only **formal** and should be considered with caution, that is, always as a linear and continuous map of the space of test functions.

What structure does $\mathcal{D}'(\mathbb{R})$ have? Being a dual space, it is a vector space with the sum of its elements and the product of these by real numbers defined in the obvious way, that is, if $T, \tilde{T} \in \mathcal{D}$, $\alpha \in \mathbb{R}$, then $\left(T + \alpha\tilde{T}\right)(\varphi) = T(\varphi) + \alpha\left(\tilde{T}(\varphi)\right)$. These operations generalize the operations defined on integrable functions since $T_f + \alpha\, T_g = T_{f+\alpha g}$.

Is the product of distributions defined? The answer is that in general it is not – just as it is not defined between integrable functions. It is defined if one of them comes from a test function, that is

$$T_\varphi\, \tilde{T}(\psi) := \tilde{T}(\varphi\psi)\ \ \forall \psi \in \mathcal{D}, \tag{9.7}$$

where we have used that the elements of $\mathcal{D}$ form an algebra. Note that this generalizes the operation defined on integrable functions:

$$T_\varphi\, T_f(\psi) = \int_{\mathbb{R}} f\,\varphi\,\psi\ = T_f(\varphi\psi) = T_{\varphi f}(\psi) \tag{9.8}$$

since if $f \in L^1$ then $f\,\varphi \in L^1$ if $\varphi \in \mathcal{D}$.

9.2 The derivative of a distribution

If f is continuously differentiable then its derivative also gives rise to a distribution $T_{f'}$. What is the relationship between T_f and $T_{f'}$? Note that

$$T_{f'}(\varphi) = \int_{\mathbb{R}} f'\,\varphi\,dx = -\int_{\mathbb{R}} f\,\varphi'\,dx = -T_f(\varphi') \quad \forall\ \varphi \in \mathscr{D} \tag{9.9}$$

where by integrating by parts we used that the φ in $\mathscr{D}$ have compact support. This suggests the possibility of extending the notion of derivative to any distribution by means of the formula

$$T'(\varphi) := -T(\varphi'), \quad \forall\ \varphi \in \mathscr{D}, \tag{9.10}$$

where the right-hand side is well defined since if $\varphi \in \mathscr{D}$ then $\varphi' \in \mathscr{D}$. Note that this operation satisfies the conditions one would expect since they are the same as those satisfied by the usual derivative, considering that we are now in this broader space where the product of functions is not generally defined. Indeed, this operation is linear.

$$\begin{aligned} \left(T + \alpha\tilde{T}\right)'(\varphi) &= -\left(T + \alpha\tilde{T}\right)(\varphi') = -\left(T(\varphi') + \alpha\tilde{T}(\varphi')\right) \\ &= T'(\varphi) + \alpha\tilde{T}'(\varphi) \end{aligned} \tag{9.11}$$

and satisfies the Leibniz rule **as much as possible** since, as we saw, the product of distributions is not generally defined. When it is, that is when one of them is an element of $\mathscr{D}$, we have that

$$\begin{aligned} \left(T_\varphi\,\tilde{T}\right)'(\psi) &= -\left(T_\varphi\,\tilde{T}\right)(\psi') = -\tilde{T}(\varphi\,\psi') \\ &= -\tilde{T}((\varphi\,\psi)' - \varphi'\psi) \\ &= -\tilde{T}((\varphi\,\psi)') + \tilde{T}(\varphi'\psi) \\ &= \tilde{T}'(\varphi\,\psi) + T_{\varphi'}\tilde{T}(\psi) \\ &= T_\varphi\,\tilde{T}'(\psi) + T_{\varphi'}\,\tilde{T}(\psi) \\ &= \left(T_\varphi\,\tilde{T}' + T_{\varphi'}\,\tilde{T}\right)(\psi). \end{aligned} \tag{9.12}$$

But these are the conditions that define a derivation. We see therefore that this is an extension of the derivative of a function in $\mathscr{D}$. Note that by generalizing the notion of function to that of distribution, we have broadened it so much that now objects like the derivative of discontinuous functions (even at all their points!) are included among these and are even infinitely differentiable themselves [$(T)^{(n)}(\psi) \equiv (-1)^n\,T(\psi^{(n)})$].

Example: T_a', the derivative of the Dirac function at a, is the distribution such that $T_a'(\varphi) = -\varphi'(a)\ \forall\ \varphi \in \mathscr{D}$.

We start with the space of infinitely differentiable functions with compact support and also end with an infinitely differentiable space [with a different notion of

derivative], it is worth asking if there is a notion of support of a distribution. Obviously, we cannot use the same notion as for continuous functions and we will have to proceed in an indirect way.

Let O be a bounded open set in $\mathbb{R}$ and $\mathcal{D}(O)$ the space of test functions with support in the closure of O, $\bar{O}$. We will say that a distribution T **vanishes in** O if $T[\mathcal{D}(O)] = 0$, that is, if it vanishes for every test function with support in O. We will call the **support of** T the complement of the union of all open sets where T vanishes. Since it is the complement of an open set, this set is closed.

Example: The support of T_o is $\{0\}$. Let $O_n = (1/n, n) \cup (-n, -1/n)$, then $\mathbb{R} - \bigcup_n O_n = \{o\}$.

Exercise: Let f be continuous. Prove that $support\{f\} = support\{T_f\}$

Exercise: How would you extend the notions of even functions ($f(x) = f(-x)$) and odd functions ($f(x) = -f(-x)$)? What properties does this extension preserve?

Exercise: Let

$$g(x) = \begin{cases} x, & x \geq 0 \\ 0, & x \leq 0 \end{cases}$$

$x \in \mathbb{R}$. Clearly $g(x)$ is continuous but not differentiable (in the classical sense). Find the first 3 derivatives of g in the sense of distributions.

Exercise: The principal part, in the sense of Cauchy, of a function,

$$\mathcal{P}(1/x)(f) = \lim_{\varepsilon \to 0} \int_{|x| \geq \varepsilon} \frac{1}{x} f(x)\, dx$$

is a distribution. How should the formula be interpreted?

$$\lim_{\varepsilon \to 0} \frac{1}{x - x_0 + i\varepsilon} = \mathcal{P}\left(\frac{1}{x - x_0}\right) - i\,\pi\,\delta(x - x_0)$$

We have seen that a distribution is differentiable, that is, given $T \in \mathcal{D}'$ there exists $S \in \mathcal{D}'$ such that $T' = S$. It is worth asking the opposite, that is, if given $S \in \mathcal{D}'$ there exists T such that the above formula holds, that is

$$-T(\varphi') = S(\varphi) \quad \forall\, \varphi \in \mathcal{D}. \tag{9.13}$$

This is a generalization to distributions of the simplest of the ordinary differential equations already studied and the answer to the posed problem is affirmative. Note

that given T the equation (9.13) defines S, that is, its derivative, but if we give S then (9.13) does not completely define T since this formula only tells us how T acts on test functions (φ') whose integral (φ) is also a test function. This is to be expected since in the case of functions the primitive of a function is only determined up to a constant. This indeterminacy is remedied by giving **generalized initial values** which is achieved by requiring that $T(\theta)$ has a given value, T_θ, for some $\theta \in \mathcal{D}$ that is not the primitive of another function in $\mathcal{D}$, that is, $\varphi(x) = \int_{-\infty}^{x} \theta(\tilde{x})\, d\tilde{x}$ does not have compact support (or $\int_{\mathbb{R}} \theta(\tilde{x})\, d\tilde{x} \neq 0$).

Theorem 9.1 *Given $S \in \mathcal{D}'$, $\theta \in \mathcal{D}$ such that $\int_{\mathbb{R}} \theta\, dx \neq 0$ and $T_\theta \in \mathbb{R}$, there exists a unique T satisfying*

$$\begin{array}{rcll} -T(\varphi') & = & S(\varphi) & \forall\, \varphi \in \mathcal{D} \\ T(\theta) & = & T_\theta & \end{array} \tag{9.14}$$

Proof: We only need to know the action of T on an arbitrary $\psi \in \mathcal{D}$. Without loss of generality, take θ such that $\int_{\mathbb{R}} \theta = 1$. We see that given $\psi \in \mathcal{D}$ there exists a unique $\lambda_\psi \in \mathbb{R}$ and a unique $\varphi_\psi \in \mathcal{D}$ such that, $\psi - \lambda_\psi \theta = \varphi'_\psi$, that is, it is a test function with a primitive. Indeed, let

$$\varphi_\psi(x) = \int_{-\infty}^{x} (\psi - \lambda_\psi \theta)\, d\tilde{x}$$

then the condition for φ_ψ to have compact support (and therefore be a test function) is that

$$0 = \int_{-\infty}^{\infty} (\psi - \lambda_\psi \theta)\, d\tilde{x} = \int_{-\infty}^{\infty} \psi\, dx - \lambda_\psi = 0 \tag{9.15}$$

that is

$$\lambda_\psi = \int_{-\infty}^{\infty} \psi\, dx. \tag{9.16}$$

Then let

$$T(\psi) = \lambda_\psi\, T_\theta - S(\varphi_\psi), \tag{9.17}$$

this distribution satisfies the equations in the theorem's statement. Note that $\lambda_\psi\, T_\theta$ is a distribution,

$$\lambda_\psi\, T_\theta = T_\theta \int_{\mathbb{R}} \psi\, dx,$$

and that,

$$T(\theta) = \lambda_\theta\, T_\theta - S(\varphi_\theta) = T_\theta - S(0) = T_\theta.$$

Let's see the uniqueness, let $\tilde{T}$ be another distribution satisfying 9.14, then the difference, $\delta T = T - \tilde{T}$ will satisfy,

$$\begin{array}{rcll} -\delta T(\varphi') & = & 0 & \forall\, \varphi \in \mathscr{D} \\ \delta T(\theta) & = & 0. & \end{array} \tag{9.18}$$

As we have seen that any test function ψ can be written as, $\psi = \lambda_\psi \theta + \varphi'_\psi$ we have,

$$\delta T(\psi) = \lambda_\psi \delta T(\theta) + \delta T(\varphi'_\psi) = 0.$$

This concludes the proof of uniqueness. From this we conclude that, as in the case of functions, any two solutions of (9.13) differ by a constant ♠

9.3 Note on the completeness of $\mathscr{D}$ and its dual $\mathscr{D}'$

Using the notion of convergence introduced in $\mathscr{D}$, we can define an analogous concept to that of a Cauchy sequence:

Definition: We will say that a sequence of test functions $\{\varphi_n\}$, $\varphi_n \in \mathscr{D}$ is **convergent** if:
1) There exists a compact $K \in I\!R$ such that $sop(\varphi_n) \subset K \;\; \forall\, n$.
2) Given p and $\varepsilon > 0$, there exists N such that for all $n, m > N$ it holds that

$$sup_{x \in K} \left| f_n^{(p)}(x) - f_m^{(p)}(x) \right| < \varepsilon$$

With this notion of convergence, the space $\mathscr{D}$ is complete, that is, every convergent sequence converges to an element of $\mathscr{D}$. To discuss the completeness of $\mathscr{D}'$, we must introduce similar notions in this space. The appropriate notion of convergence is the following.

Definition: We will say that the sequence $\{T_n\}$, $T_n \in \mathscr{D}'$ **converges** to $T \in \mathscr{D}'$ if $T_n(\varphi) \to T(\varphi)$ for all $\varphi \in \mathscr{D}$ [3].
Examples:
a) Let T_n be the distribution associated with the function $e^{-|x-n|^2}$. Then T_n converges to the zero distribution. This shows how weak this type of convergence is.
b) Let T_n be the distribution associated with some function f_n satisfying

1. $f_n(t) \geq 0$ if $|t| < 1/n$ and zero if $|t| \geq 1/n$.
2. $\int_a^b f_n(t)\, dt = 1$,

[3] Again, we are introducing a topology indirectly, this time in $\mathscr{D}'$.

Then $T_n \to T_0$, the Dirac function with support at zero.

Similarly, we can define the notion of convergence of distributions.

Definition: $\{T_n\}$, $T_n \in \mathscr{D}'$ is **convergent** if for each $\varphi \in \mathscr{D}$ and $\varepsilon > 0$ there exists N such that if $n, m > N$ then

$$|T_n(\varphi) - T_m(\varphi)| < \varepsilon$$

With this notion of convergence, the space $\mathscr{D}'$ is complete.

9.4 Weak Convergence and Compactness

In a normed space, H, we have the notion of convergence with respect to the norm, which we will call strong convergence

$$\{x_n\} \xrightarrow{f} x \quad \text{if} \quad \lim_{n\to\infty} ||x_n - x||_H = 0.$$

In these spaces, there is another notion of convergence, called weak convergence, which uses the existence of the dual space of H, H'.

Definition: We will say that $\{x_n\}$ **converges weakly** to x,

$$\{x_n\} \xrightarrow{d} x, \quad \text{if} \quad \sigma(x_n) \to \sigma(x) \quad \forall \ \sigma \in H'.$$

If H is a Hilbert space (which we will assume from now on), the Riesz Representation Theorem tells us that $\{x_n\} \xrightarrow{d} x$ if and only if $(x_n, y) \to (x, y)\ \forall\ y \in H$.

Clearly, this notion of convergence is the weakest such that the elements of H' are continuous functionals [In the sense that f is continuous at x if **given any sequence** $\{x_n\}$ **converging to** x then $\lim_{n\to\infty} f(x_n) = f(x)$.], where we say that a notion of convergence is weaker than another if every sequence that converges with respect to the second also converges with respect to the first, and there are sequences that converge with respect to the first but not with respect to the second. Let's see, as an example, that norm convergence, or strong convergence, is in fact stronger than the so-called weak convergence. Suppose then that $\{x_n\} \xrightarrow{f} x$, that is, $\lim_{n\to\infty} ||x - x_n||_H = 0$, then since the elements of H' are bounded linear functionals ($\Longleftrightarrow$ continuous), it holds that

$$|\sigma(x) - \sigma(x_n)| \leq ||\sigma||_{H'} ||x - x_n|| \quad \forall \quad \sigma \in H', \tag{9.19}$$

and therefore,

$$\lim_{n\to\infty} |\sigma(x) - \sigma(x_n)| = 0, \tag{9.20}$$

that is, $\{x_n\} \xrightarrow{d} x$. The following is an example of a sequence that converges weakly and not strongly.

Exercise: Show that the sequence $\{x_n = (0, ..., 0, \underset{n}{1}, 0, ...)\}$ converges weakly in l^2 but not strongly.

The previous example was also used to show that the unit ball in l^2 was not compact with respect to strong convergence. Will it be weakly compact? That is, given a sequence $\{x_n\} \in B_1(l^2)$, will there exist a subsequence that converges weakly? The answer is affirmative and it is one of the most useful tools in functional analysis.

Theorem 9.2 *B_1 is weakly compact.*

Proof: We will only prove the case where H is separable. Let $\{x_n\}$ with $||x_n||_H \leq 1$ and $S = \{e_m\}$ a countable orthonormal basis of H. We will construct, using induction, a subsequence $\{x_n^\infty\}$ such that,

$$(x_n^\infty, e_m) \overset{n\to\infty}{\to} \alpha_m \quad \forall\, e_m \in S. \tag{9.21}$$

Let $m = 1$, then $|(x_n, e_1)| \leq ||x_n||_H ||e_1||_H \leq 1$. We see then that $\{(x_n, e_1)\}$ is a bounded sequence in $\boldsymbol{C}$. But the unit ball in $\boldsymbol{C}$ is compact and therefore there will be some subsequence (x_n^1, e_1) converging to some α_1 in $\boldsymbol{C}$. Now suppose we have a subsequence $\{x_n^{m-1}\}$ such that

$$(x_n^{m-1}, e_p) \to \alpha_p \quad \forall\, 1 \leq p \leq m-1. \tag{9.22}$$

In the same way as we did for the case $m = 1$, considering in this case $\{(x_n^{m-1}, e_m)\}$, we obtain a subsequence $\{x_n^m\}$ of $\{x_n^{m-1}\}$ that satisfies,

$$(x_n^m, e_m) \overset{n\to\infty}{\to} \alpha_m, \tag{9.23}$$

which completes the induction. We thus have a map $\sigma : \{e_m\} \to \boldsymbol{C}$ given by $\sigma(e_m) = \alpha_m$, since $\{e_m\}$ is a basis, we can extend this map linearly to all of H. Since $\{x_n^\infty\}$ is bounded,

$$|\sigma(y)| = \lim |(x_n^\infty, y)| \leq ||y||_H \tag{9.24}$$

and σ is also bounded, therefore continuous. Using the Riesz Representation Theorem, we know then that there exists $x \in H$ such that

$$(x, y) = \sigma(y) = \lim (x_n^\infty, y) \quad \forall\, y \in H. \tag{9.25}$$

We thus conclude that $\{x_n^\infty\} \xrightarrow{d} x$ ♠

In the next chapter, we will use this property to prove an important result in Sobolev spaces.

Bibliography notes: I recommend reading: [10], [1] [9]. Although it was a physicist, Dirac, who introduced the concept of distribution, many physicists dismiss them as something *mathematical* and use them as a useful abbreviation for calculations. Usually, the person manipulating them knows what they are doing and does not make mistakes, but it is quite easy to make them if the rules are not carefully followed and the nature of what they are is lost. This, for example, leads to errors such as assigning meaning to the product of two arbitrary distributions. It is not difficult to understand the basic concept of distribution nor that one should not deviate from the operational rules, follow them always and you will not go wrong.

CHAPTER 10

THE FOURIER TRANSFORM

10.1 Introduction

Consider the following problem on the circle S^1 whose circumference is 2π: Given ρ continuous on S^1, find f on S^1 such that

$$\frac{\partial^2 f}{\partial \theta^2} = \rho. \tag{10.1}$$

One way to solve this problem is by using the orthogonal Fourier basis, $\left\{\frac{1}{\sqrt{2\pi}} e^{in\theta}\right\}$ of $L^2(S^1)$. Let $F : L^2(S^1) \to l^2$ be the map between these spaces generated by this basis, that is,

$$F(f) = \left\{\left\langle \frac{1}{\sqrt{2\pi}} e^{in\theta}, f\right\rangle\right\} \equiv \{c_n\}. \tag{10.2}$$

Then

$$\begin{aligned} F\left(\frac{\partial^2 f}{\partial \theta^2} - \rho\right) &= \left\{\left\langle \frac{1}{\sqrt{2\pi}} e^{in\theta}, \frac{d^2 f}{d\theta^2} - \rho\right\rangle\right\} \\ &= \left\{-n^2 c_n - a_n\right\} \end{aligned} \tag{10.3}$$

where $\{a_n\} = F(\rho)$.

Since $F(0) = \{0\}$ we see that (10.1) implies an algebraic equation in l^2,

$$-n^2 c_n = a_n. \tag{10.4}$$

If ρ is L^2 orthogonal to $f = cte.$, that is to the only solutions of (10.1) with $\rho = 0$, which implies $a_0 = 0$, the sequence with c_0 arbitrary and $c_n = -\frac{a_n}{n^2}$, $n \neq 0$ satisfies (10.4). The inverse map $F^{-1} : l^2 \to L^2(S^1)$ defines $f(\theta) = \sum_{n=-\infty}^{\infty} \frac{c_n}{\sqrt{2\pi}} e^{in\theta}$, which at least formally[1] is a solution of (10.1).

[1] The map defines an element of L^2 since $\rho \in L^2$ and therefore $\sum_{-\infty}^{\infty} |a_n|^2 < \infty$ which implies that $\{c_0$ arbitrary, $c_n = -\frac{a_n}{n^2}$, $n \neq 0\}$ is also an element of l^2. It remains to be seen that $f = F^{-1}\{c_n\}$ is twice differentiable.

Note that in this application we have not directly used the fact that the functions $\{\frac{1}{\sqrt{2\pi}}e^{in\theta}\}$ form an orthogonal basis but only that $\frac{d}{d\theta}e^{in\theta} = in\,e^{in\theta}$ and certain properties of the map F generated by this basis. In particular, this property of the basis is induced in the map in the sense that if $f \in L^2$ is differentiable and $F(f) = \{c_n\}$ then $F(\frac{df}{d\theta}) = \{+in\,c_n\}$.

These observations are very useful since there are cases, such as $L^2(\mathbb{R})$, in which interesting orthogonal bases are not known, but similar maps to F with interesting properties are known. One of them is the Fourier transform that we will study now. The problem in $L^2(\mathbb{R})$ is that we would like to have a basis $\{\varphi_n\}$ whose functions satisfy $\frac{d}{dx}\varphi_n = i\,c_n\varphi_n$, but the solutions to this equation are $\varphi_n = a_n\,e^{ic_n x}$, which are not square-integrable functions for any c_n or $a_n \in \mathbb{R}$ (except of course if we take $a_n \equiv 0$). However, although these functions do not form an orthogonal basis, they do generate a map $\mathscr{F}$, this time from $L^2(\mathbb{R})$ to another copy (considered as distinct) of $L^2(\mathbb{R})$, with properties similar to those of the map F considered earlier.

Theorem 10.1 (Fourier's Theorem) *The Fourier transform*

$$\hat{f}(\lambda) := \mathscr{F}(f)(\lambda) := \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} e^{-ix\cdot\lambda}\,f(x)\,d^n x \tag{10.5}$$

where $x\cdot\lambda = \sum_{i=1} x^i\,\lambda_i$.

Satisfies:

a) $\mathscr{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ *is a linear, continuous, and invertible map from* $L^2(\mathbb{R}^n)$ *onto itself that preserves the norm (i.e., unitary),* $||\hat{f}||_{L^2} = ||\mathscr{F}(f)||_{L^2} = ||f||_{L^2}$ *(Plancherel's identity).*

b) Its inverse is given by

$$\check{g}(x) = \mathscr{F}^{-1}(\hat{g}) = \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} e^{ix\cdot\lambda}\,\hat{g}(\lambda)\,d^n\lambda. \tag{10.6}$$

c) If the derivative of $f \in L^2(\mathbb{R}^n)$ *in the distributional sense is also in* $L^2(\mathbb{R}^n)$*, then*

$$\mathscr{F}\left(\frac{\partial f}{\partial x_j}\right)(\lambda) = i\,\lambda_j\,\mathscr{F}(f)(\lambda). \tag{10.7}$$

This theorem tells us that this transform has all the properties that the one generated by the Fourier basis had and therefore will be as useful as that one in similar applications, but now in $\mathbb{R}^n$.

The proof of this theorem uses several of the most common and powerful techniques of functional analysis and therefore a careful and non-automatic reading of it is recommended.

Proof: The map $\mathcal{F}$ is obviously linear and clearly defined for any function in $C_o^\infty(\mathbb{R}^n)$, that is, infinitely differentiable and compactly supported. First, let's see that $\mathcal{F}$ preserves the $L^2(\mathbb{R}^n)$ norm for these functions. Let $f \in C_o^\infty(\mathbb{R}^n)$ be any function and take an n-cube C_ε of volume $(\frac{2}{\varepsilon})^n$ with $\varepsilon > 0$ small enough so that $sop(f) \subset C_\varepsilon$. Let $K_\varepsilon = \{k \in \mathbb{R}^n \mid k_i = \pi p \varepsilon \text{ for some integer } p\}$, [for example in $\mathbb{R}^3$ the vector $(\pi\varepsilon 5, \pi\varepsilon 17, 0)$ is an element of K_ε]. Then the set of functions $\left\{ \left(\frac{\varepsilon}{2}\right)^{n/2} e^{ik\cdot x} \mid k \in K_\varepsilon \right\}$ forms an orthogonal basis of $L^2(C_\varepsilon)$.

But $f \in L^2(C_\varepsilon)$ and is continuously differentiable, therefore

$$\sum_{k\in K_\varepsilon} \langle \left(\frac{\varepsilon}{2}\right)^{n/2} e^{ik\cdot\tilde{x}}, f(\tilde{x})\rangle \left(\frac{\varepsilon}{2}\right)^{n/2} e^{ik\cdot x} = \sum_{k\in K_\varepsilon} \frac{\hat{f}(k)}{(2\pi)^{n/2}} e^{ik\cdot x} (2\pi\varepsilon/2)^n, \tag{10.8}$$

is the Fourier series representation of f which converges uniformly to f in C_ε and therefore in $\mathbb{R}^n$. We then have that

$$\begin{aligned} ||f||_{L^2}^2 &= \int_{\mathbb{R}^n} |f(x)|^2 \, d^n x \\ &= \int_{C_\varepsilon} |f(x)|^2 \, d^n x \\ &= \sum_{k\in K_\varepsilon} \left| \langle (\tfrac{\varepsilon}{2})^{n/2} e^{ik\cdot x}, f(x)\rangle \right|^2 \\ &= \sum_{k\in K_\varepsilon} |\hat{f}(k)|^2 \, (\pi\varepsilon)^n. \end{aligned} \tag{10.9}$$

Since $\mathbb{R}^n$ is the union of n-cubes with sides $\pi\varepsilon$ (i.e., of volume $(\pi\varepsilon)^n$ around the points of K_ε, the right-hand side of (10.9) is simply the Riemann series of the function $|\hat{f}(k)|^2$. If this function were continuous then the Riemann series would converge to $||\hat{f}||_{L^2}^2$ and we would have proven that the Fourier transform of a function in $C_o^\infty(\mathbb{R}^n)$ preserves the $L^2(\mathbb{R}^n)$ norm. But

$$\begin{aligned} sup_k |\tfrac{\partial \hat{f}(k)}{\partial k_j}| &= sup_k |\tfrac{1}{(2\pi)^{n/2}} \int_{R^n} -i\, x^j \, e^{-ik\cdot x} \, f(x) \, d^n x| \\ &\le \tfrac{1}{(2\pi)^{n/2}} \int_{R^n} |x^j f(x)| \, d^n x < \infty, \end{aligned} \tag{10.10}$$

since $f \in C_o^\infty(\mathbb{R}^n)$. We then see that the derivatives of $\hat{f}(k)$ are bounded and therefore that $\hat{f}$ is continuous. This result not only completes the proof that $\mathcal{F}$ pre-

serves the norm but also, since $\hat{f}(k)$ is continuous, that the Riemann series on the right-hand side of the equation (10.8) converges to the integral and therefore that $\mathscr{F}^{-1}(\mathscr{F}(f)) = f(x)\ \forall\ f \in C_o^\infty(\mathbb{R}^n)$.

If $f \in C_o^\infty(\mathbb{R}^n)$ then the identity in point c) is shown trivially. It only remains to extend these results to arbitrary f in $L^2(\mathbb{R}^n)$, but $C_o^\infty(\mathbb{R}^n)$ is a dense subspace of $L^2(\mathbb{R}^n)$, that is, every element $f \in L^2(\mathbb{R}^n)$ can be obtained as the limit (with respect to the $L^2(\mathbb{R}^n)$ norm) of a sequence $\{f_n\}$ of functions in $C_o^\infty(\mathbb{R}^n)$, this allows us to extend the action of $\mathscr{F}$ to every element of $L^2(\mathbb{R}^n)$. Indeed, let $f \in L^2(\mathbb{R}^n)$ be arbitrary and let $\{f_n\} \to f$ with $f_n \in C_o^\infty(\mathbb{R}^n)$ then the sequence $\hat{f}_n(k) = \mathscr{F}(f_n)$ satisfies,

$$||\hat{f}_n(k) - \hat{f}_m(k)||_{L^2(R^n)} = ||f_n(x) - f_m(x)||_{L^2(\mathbb{R}^n)}. \tag{10.11}$$

Since $\{f_n\}$ is convergent it is Cauchy and therefore the equality of norms implies that $\{\hat{f}_n\}$ is also Cauchy in $L^2(\mathbb{R}^n)$, but $L^2(\mathbb{R}^n)$ is complete and therefore there exists a unique $\hat{f} \in L^2(\mathbb{R}^n)$ such that $\{\hat{f}_n\} \to \hat{f}$. We will extend the map $\mathscr{F}$ to $L^2(\mathbb{R}^n)$ by defining $\mathscr{F}(f) := \hat{f}$. This extension is clearly linear and bounded, therefore it is continuous. Using the same reasoning we extend $\mathscr{F}^{-1}$ to all $L^2(\mathbb{R}^n)$ and see that $\mathscr{F}^{-1}(\mathscr{F}(f(x))) = f(x)\ \forall\ f(x) \in L^2(\mathbb{R}^n)$ which proves that $\mathscr{F}$ is invertible, that its inverse is $\mathscr{F}^{-1}$ and is continuous. The same argument shows that the formula in c) is also valid for any f such that its gradient is also in $L^2(\mathbb{R}^n)$. This completes the proof of the theorem ♠

Note that the strategy has been to first take a space ($C_o^\infty(\mathbb{R}^n)$) where all properties obviously hold, then take a space that contains the former and where it is dense, see that the map is there continuous (with respect to the notion of continuity of the larger space) and finally take the extension that this continuity provides us.

Are there other possible extensions? The answer is yes and now we will see one of them that will allow us to use the Fourier transform in distributions. We will do it in the form of a problem divided into several exercises.

Schwartz Notation: I_+^n will denote the set of n-tuples of non-negative integers $\alpha =< \alpha_1, \ldots, \alpha_n >$ and $|\alpha| = \sum_{i=1}^n \alpha_i$. We define, $x^\alpha := (x^1)^{\alpha_1}(x^2)^{\alpha_2}\cdots(x^n)^{\alpha_n}$ Similarly, we denote by $D^\alpha \equiv \dfrac{\partial^\alpha}{\partial x^\alpha} \equiv \dfrac{\partial^{|\alpha|}}{\partial (x^1)^{\alpha_1}\cdots\partial (x^n)^{\alpha_n}}$, that is, D^α is the differential operator that takes α_1 partial derivatives with respect to x_1, α_2 partial derivatives with respect to x_2, etc. The degree of D^α is $|\alpha|$.

Definition: We will call the **Schwartz space** $\mathscr{S}(\mathbb{R}^n)$ the vector space of infinitely differentiable functions that decay asymptotically faster than the inverse of any polynomial, that is,

$$||\varphi||_{p,q} \equiv \sum_{\substack{|\tilde{\alpha}| \le p \\ |\tilde{\beta}| \le q}} sup_{x\in\mathbb{R}^n} |x^\alpha D^\beta \varphi(x)| < \infty \tag{10.12}$$

$\forall\ \alpha, \beta \in I_+^n$, with the following notion of convergence: we will say that $\{\varphi_n\}$, $\varphi_n \in \mathcal{S}$ converges to $\varphi \in \mathcal{S}$ if given $\varepsilon > 0$ for each pair of non-negative integers, p, q, there exists $N_{p,q}$ such that if $n > N_{p,q}$ then $\|\varphi - \varphi_n\|_{p,q} < \varepsilon$.

Exercises:
1) How would you define the notion of a convergent sequence in this space?
2) Show that if $\{\varphi_n\} \to \varphi$ then $\{\varphi_n\}$ converges in the above sense.
3) Show that $\mathcal{D}$ is strictly contained in $\mathcal{S}$. Hint: Find $\varphi \in \mathcal{S}$ with $sop(\varphi) = \mathbb{R}^n$.

The properties of this space that interest us are the following:

Lemma 10.1 *With the corresponding notion of convergence, the space $\mathcal{S}(\mathbb{R}^n)$ is complete.*

Lemma 10.2 *The space $C_0^\infty(\mathbb{R}^n)$ is dense in $\mathcal{S}$, that is, any element $\varphi \in \mathcal{S}$ can be obtained as the limit of a convergent sequence with each of its members in $C_0^\infty(\mathbb{R}^n)$.*

10.2 Exercises and Definitions

Exercise: Prove Lemma 10.2.

Lemma 10.3 *$\mathcal{F} : \mathcal{S} \to \mathcal{S}$ is continuous and invertible with a continuous inverse.*

Exercise: Prove Lemma 10.3. Hint: Prove that given integers p and q there exist $\tilde{p}$ and $\tilde{q}$ and $c > 0$ such that for all $\varphi \in \mathcal{S}$ it holds

$$\|\hat{\varphi}\|_{p,q} \leq c\, \|\varphi\|_{\tilde{p},\tilde{q}}. \tag{10.13}$$

Exercise: Find $\mathcal{F}(e^{-\alpha x^2/2})$.

Definition: The dual space to $\mathcal{S}$, $\mathcal{S}'$ is called the space of **tempered distributions**.

Exercise: Show that $\mathcal{S}'$ is strictly contained in $\mathcal{D}'$. Hint: Find f such that $T_f \in \mathcal{D}'$ and not in $\mathcal{S}'$.

How to extend $\mathcal{F}$ to $\mathcal{S}'$? Using the Plancherel identity and the polarization identity[2] we have that $\langle \varphi, \psi \rangle = \langle \hat{\varphi}, \hat{\psi} \rangle$. But then

$$T_{\hat{\sigma}}(\hat{\psi}) = \int \hat{\sigma}\, \hat{\psi}\, dk = \int \sigma\, \psi\, dx = T_\sigma(\psi) := \hat{T}_\sigma(\hat{\psi}). \tag{10.14}$$

[2] That is, $\langle x, y \rangle = \frac{1}{4}\{\|x+y\|^2 - \|x-y\|^2 + i\|x-iy\|^2 - i\|x+iy\|^2\}$

This leads us to define in general

$$\hat{T}(\hat{\varphi}) := T(\varphi). \tag{10.15}$$

Example: The Fourier transform of the Dirac delta. From the previous definition we have that,

$$\hat{\delta}_a(\hat{\phi}) = \phi(a),$$

but

$$\phi(a) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} e^{ika} \hat{\phi}(k) dk,$$

and therefore the Fourier transform of the Dirac delta is the regular distribution given by,

$$\hat{\delta}_a = T_{e^{ika}}.$$

Exercise: Calculate $\mathcal{F}(\delta'_a)$.

What other properties does the Fourier transform have? Note that from the identity in point c) of the Fourier Theorem it follows that if $x^\alpha D^\beta f \in L^2(\mathbb{R}^n)$ then $k^\beta D^\alpha \hat{f} \in L^2(\mathbb{R}^n)$. This essentially tells us that differentiability in x is equivalent to decay in k and vice versa. In particular this tells us that the Fourier transform of a compactly supported function is infinitely differentiable.[3] Moreover, it can be proven that it is analytic.

Another important property of the Fourier transform is that it maps the product of functions to the convolution of functions and vice versa.

Definition: Let $f, g \in \mathcal{S}(\mathbb{R}^n)$. The **convolution** of f with g denoted $f * g$ is the function, also in $\mathcal{S}(\mathbb{R}^n)$, given by

$$(f * g)(y) = \int_{\mathbb{R}^n} f(y-x)\, g(x)\, d^n x. \tag{10.16}$$

Theorem 10.2 *a) For each $f \in \mathcal{S}(\mathbb{R}^n)$, the linear map $g \to f * g$ from $\mathcal{S}(\mathbb{R}^n)$ to $\mathcal{S}(\mathbb{R}^n)$ is continuous.*
b) $\widehat{f\, g} = \frac{1}{(2\pi)^{n/2}} \hat{f} * \hat{g}$ *and* $\widehat{f * g} = (2\pi)^{n/2} \hat{f}\, \hat{g}$.
c) $f * g = g * f$ *and* $f * (g * h) = (f * g) * h$.

[3] In fact, the previous argument tells us that if $sop(f)$ is compact then $D^\alpha \hat{f} \in L^2(\mathbb{R}^n) \ \ \forall\, \alpha \in I_+^n$. As we will see later, this implies that it is infinitely differentiable.

Proof: Once point b) is proven, the others follow trivially. Make sure of it! We will then prove b). As we have seen $\langle\varphi,\psi\rangle = \langle\hat{\varphi},\hat{\psi}\rangle \quad \forall\, \varphi,\psi \in \mathscr{S}(\mathbb{R}^n)$. Applying this identity to $e^{ik\cdot x}\bar{f}(x)$ and $g(x)$ we obtain,

$$\langle e^{ik\cdot x}\bar{f}, g\rangle = \widehat{\langle e^{ik\cdot x}\bar{f}, \hat{g}\rangle},$$

but

$$\langle e^{ik\cdot x}\bar{f}, g\rangle = \int_{\mathbb{R}^n} e^{-ik\cdot x} f(x)\, g(x)\, d^n x = (2\pi)^{n/2}\, \widehat{f g}(k) \tag{10.17}$$

and

$$\begin{aligned}
\widehat{\langle e^{ik\cdot x}\bar{f}, \hat{g}\rangle} &= \int_{\mathbb{R}^n} \overline{\left((2\pi)^{-n/2}\int_{\mathbb{R}^n} e^{-i\lambda\cdot x + ik\cdot x}\,\bar{f}(x)\, d^n x\right)}\, \hat{g}(\lambda)\, d^n\lambda \\
&= \int_{\mathbb{R}^n} \hat{f}(k-\lambda)\,\hat{g}(\lambda)\, d^n\lambda \\
&= \hat{f} * \hat{g}.
\end{aligned} \tag{10.18}$$

The other formula is obtained by applying $\mathscr{F}^{-1}$ to the previous one ♠

Another interesting property of the convolution is obtained by noting that since $\mathscr{S}(\mathbb{R}^n)$ is closed with respect to the operation of convolution ($f, g \in \mathscr{S}(\mathbb{R}^n) \Longrightarrow f * g \in \mathscr{S}(\mathbb{R}^n)$) we can define the convolution of a tempered distribution with a function in $\mathscr{S}(\mathbb{R}^n)$ as,

$$(T * f)(\varphi) := T(\tilde{f} * \varphi) \quad \text{where } \tilde{f}(x) := f(-x). \tag{10.19}$$

But,

$$T_{[x]}(\tilde{f} * \varphi) = T_{[x]}\Big(\int f(y-x)\phi(y)\, dy\Big) = \int T_{[x]}(f(y-x))\phi(y)\, dy, \tag{10.20}$$

where we have included a sub-expression in T to indicate that the distribution acts on the test function as a function of x. Since $f \in \mathscr{S}(\mathbb{R}^n)$, $T_{[x]}(f(y-x)) := g(y)$ gives us a finite value for each $y \in \mathbb{R}^n$ considered as a parameter, and therefore a function $g(y)$, this function is in fact continuous. [4] Therefore the integral in the previous expression is well-defined and we conclude that,

$$(T * f)(\varphi) = T_g(\phi). \tag{10.21}$$

That is, this convolution is a regular distribution!

Exercise: Apply (10.19) to $T = T_g$ for $g \in \mathscr{S}(\mathbb{R}^n)$ and see that this definition makes sense.

[4] That is, for any $\{y_n\} \to y$ we have, $g(y_n) = T_{[x]}(f(y_n - x)) \to T_{[x]}(f(y-x)) = g(y)$, since $(f(y_n - x) \to f(y-x)$ in $\mathscr{S}(\mathbb{R}^n)$ and distributions are continuous as maps from $\mathscr{S}(\mathbb{R}^n)$ to $\mathbb{R}$.

On the other hand, note that if f is an integrable function and $g \in \mathscr{S}(\mathbb{R}^n)$, $f * g$ is infinitely differentiable! This follows from the fact that, as

$$\begin{aligned}\frac{d(f*g)}{dx}(x) &= \frac{d}{dx}\int_{-\infty}^{\infty} f(x-y)g(y)\,dy \\ &= \int_{-\infty}^{\infty} \frac{df}{dx}(x-y)g(y)\,dy \\ &= -\int_{-\infty}^{\infty} \frac{df}{dy}(x-y)g(y)\,dy \\ &= \int_{-\infty}^{\infty} f(x-y)\frac{dg}{dy}(y)\,dy, \end{aligned} \tag{10.22}$$

all derivatives are well defined and bounded, and therefore such convolution is infinitely differentiable even if f was only integrable.
This interesting property can even be extended to distributions as seen in the following lemma.

Lemma 10.4 *If $f \in \mathscr{S}(\mathbb{R}^n)$, and $T \in \mathscr{S}'(\mathbb{R}^n)$, $T * f$ is equivalent to an infinitely differentiable function.*

In fact, we have,

$$\begin{aligned}(T*f)'(\phi) &= (T*f)(-\phi') \\ &= T(\tilde{f} * (-\phi')) \\ &= T(\tilde{f}' * \phi) \\ &= (T*f')(\phi). \end{aligned} \tag{10.23}$$

Exercise: Combine this result with the previous one, equation (10.21) and find the corresponding regular distribution.

10.3 *Basic properties of Sobolev Spaces

As an application of the Fourier transform we will now see the basic properties of Sobolev spaces. These will be used later when we study the theory of partial differential equations. The first step will be a generalization of Sobolev spaces. As we saw, Sobolev spaces are Hilbert spaces with a norm derived from the following inner product.

$$\langle f, g\rangle_{H^m} = \sum_{k=0}^{m}\sum_{i,j,p=1}^{n} \langle \underbrace{\partial_i\partial_j\ldots\partial_p}_{k \text{ times}} f, \partial_i\partial_j\ldots\partial_p g\rangle_{L^2}, \tag{10.24}$$

where f and g are functions defined in any open set Ω of $\mathbb{R}^n$. If $\Omega = \mathbb{R}^n$ then from the properties of the Fourier transform we get

$$\begin{aligned} \langle f, g\rangle_{H^m} &= \sum_{k=0}^{m}\sum_{i,j,p=1}^{n}\langle \lambda_i\lambda_j\dots\lambda_p\hat{f}(\lambda), \lambda_i\lambda_j\dots\lambda_p, \hat{g}(\lambda)\rangle_{L^2} \\ &= \langle \hat{f}(\lambda)\sqrt{\sum_{k=0}^{m}|\lambda|^{2k}}, \hat{g}(\lambda)\sqrt{\sum_{k=0}^{m}|\lambda|^{2k}}\rangle_{L^2}, \end{aligned} \tag{10.25}$$

that is, the functions $f(x) \in H^m(\mathbb{R}^n)$ are those whose Fourier transform $\hat{f}$ decays sufficiently fast so that $\hat{f}(\lambda)\sqrt{\sum_{k=0}^{m}|\lambda|^{2k}}$ is square integrable.

This suggests generalizing the spaces by allowing non-integer and even negative indices through the following inner product,

$$\langle f, g\rangle_{H^s} = \langle \hat{f}(\lambda)(1+|\lambda|^2)^{s/2}, \hat{g}(\lambda)(1+|\lambda|^2)^{s/2}\rangle_{L^2}. \tag{10.26}$$

Exercise: Show that this is an inner product.

Note that for $s = m$ instead of the polynomial $\sum_{k+0}^{m}|\lambda|^{2k}$, we now have $(1+|\lambda|^2)^m$. Since there are positive constants c_1 and c_2 such that

$$c_1(1+|\lambda|^2)^m \le \sum_{k=0}^{m}|\lambda|^{2k} \le c_2(1+|\lambda|^2)^m \tag{10.27}$$

this change in the norm is trivial in the sense that these norms are equivalent. In fact, as we will see in a special case, even the 1 in the polynomial can be ignored and still obtain an equivalent norm. The first property we will see is the following lemma:

Lemma 10.5 *If $s' > s$ then $H^{s'} \subset H^s$.*

Proof: Trivial ♠

In particular, if $s > 0$ then $H^s \subset H^0 = L^2$, if $s < 0$ then $L^2 \subset H^s$. What are the elements of H^s for negative s?

Given $g \in H^{-s}$ I can define the following map from H^s to $\mathbb{C}$,

$$\Psi_g(f) := \langle \hat{g}, \hat{f}\rangle_{H^0} = \langle \frac{\hat{g}}{(1+|\lambda|^2)^{s/2}}, \hat{f}(1+|\lambda|^2)^{s/2}\rangle_{H^0}.$$

This map is linear, well defined $\forall\, f, g \in C_0^\infty$ and can be extended to all H^s (and H^{-s}) by continuity, since,

$$|\Psi_g(f)| \le ||g||_{H^{-s}}||f||_{H^s}.$$

Therefore we have a map between H^{-s} and the dual of H^s, which preserves the norm, moreover, it can be proven that this map is an isomorphism (that is, it is also

invertible) between these spaces.[5] Therefore we can identify H^{-s} with the dual of H^s.

Exercise: Using that $C^\infty(\mathbb{R}^n)$ is dense in $H^s(\mathbb{R}^n)$ show that the Dirac delta is in $H^s(\mathbb{R}^n)$ for all $s < -n/2$.

Perhaps the most important property of $H^s(\mathbb{R}^n)$ is the following,

Lemma 10.6 (of Sobolev) *If $m < s - n/2$ then $H^s(\mathbb{R}^n) \subset C^m(\mathbb{R}^n)$.*

Proof: It is enough to prove it for $m = 0$, the rest follows by induction. It is also enough to prove the inequality. $||f||_{C^0} < C\,||f||_{H^s}$, $s > \frac{n}{2}$ for some constant $C > 0$ independent of f and for all $f \in C^\infty(\mathbb{R}^n)$, the rest follows by the continuity of the norm. But

$$\begin{aligned} ||f||_{c^0} = sup_{x\in R^n}|f(x)| &= sup_{x\in R^n} \tfrac{1}{(2\pi)^{n/2}} \left| \int_{R^n} e^{i\lambda\cdot x}\, \hat{f}(\lambda)\, d^n x \right| \\ &\leq \tfrac{1}{(2\pi)^{n/2}} \int_{R^n} |\hat{f}(\lambda)|\, d^n\lambda \\ &\leq \tfrac{1}{(2\pi)^{n/2}} \int_{R^n} \frac{|\hat{f}(\lambda)|(1+|\lambda|^2)^{s/2}}{(1+|\lambda|^2)^{s/2}}\, d^n\lambda \qquad (10.28) \\ &\leq \tfrac{1}{(2\pi)^{n/2}} ||\hat{f}(\lambda)(1+|\lambda|^2)^{s/2}||_{L^2}\, ||\tfrac{1}{(1+||\lambda|^2)^{s/2}}||_{L^2} \\ &\leq C\,||f||_{H^s}. \end{aligned}$$

where we have used that $s > n/2$ to prove that $||\frac{1}{(1+|\lambda|^2)^{s/2}}||_{L^2} < \infty$ ♠

This lemma tells us that if $f \in H^m(\mathbb{R}^n)$ for sufficiently large m then f can be identified with an ordinary, continuous, and even differentiable function.

Let f be a continuous function in $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \;\; x_n \geq 0\}$ and let τf be the restriction of that function to the hyperplane $x_n = 0$, that is, $(\tau f)(x_1, x_2, \ldots, x_{n-1}) = f(x_1, x_2, \ldots, x_n = 0)$. Clearly, τ is a linear map and if f is continuous in $\mathbb{R}^n_+$ then τf is continuous in $\mathbb{R}^{n-1} = \{x \in \mathbb{R}^n \,/\, x_n = 0\}$. Can this map be extended to more general functions? The answer is the following lemma which also shows why it is necessary to extend Sobolev spaces to non-integer indices.

Lemma 10.7 (Trace) *Let $m > 0$ be an integer, then:*
i) $\tau : H^m(\mathbb{R}^n_+) \to H^{m-1/2}(\mathbb{R}^{n-1})$ is continuous.
ii) it is surjective.

[5] By the Riesz Representation Theorem, each $\Psi : H^s \to \boldsymbol{C}$ can also be written as $\Psi(f) = (\tilde{g}, f)_{H^s}$, $\tilde{g} \in H^s$. If we take $g = \mathscr{F}^{-1}((1+|\lambda||^2)^s \mathscr{F}(\tilde{g})) \in H^{-s}$ then $\Psi(f) = \Psi_g(f)$.

Proof: By the same argument as in the previous lemma to prove i) it is enough to prove that there exists $c > 0$ such that

$$||\tau f||_{H^{1/2}(R^{n-1/2})} < C\,||f||_{H^1(R^n_+)} \quad \forall\, f \in C^\infty_0(I\!R^n_+) \tag{10.29}$$

Let $\hat{f}(\lambda', x_n)$ be the Fourier transform of $f(x)$ with respect to the coordinates $(x_1, x_2, \ldots, x_{n-1})$ of $I\!R^n_+$, then

$$|\hat{f}(\lambda', 0)|^2 = -2\,Re \int_0^\infty \frac{\partial \hat{f}}{\partial x_n}(\lambda', t)\overline{\hat{f}(\lambda', t)}\, dt. \tag{10.30}$$

Multiplying both sides of this equality by $(1 + |\lambda'|^2)^{1/2}$ and integrating with respect to λ' we obtain,

$$\begin{aligned} ||\tau f||_{H^{1/2}(R^{n-1})} &= \int_{R^{n-1}} |(1 + |\lambda'|^2)^{1/2} \hat{f}(\lambda', 0)|^2\, d^{n-1}\lambda' \\ &= 2\left| \int_{R^{n-1}} \int_0^\infty (1 + |\lambda'|^2)^{1/2} \frac{\partial \hat{f}(\lambda', t)}{\partial x_n} \overline{\hat{f}(\lambda', t)}\, dt\, d^{n-1}\lambda' \right| \\ &\leq 2\left\{ \int_{R^{n-1}} \int_0^\infty \left| \frac{\partial \hat{f}}{\partial x_n}(\lambda', t) \right|^2 dt\, d^{n-1}\lambda' \right\}^{1/2} \times \\ &\quad \times \left\{ \int_{R^{n-1}} \int_0^\infty (1 + |\lambda'|^2)\, |f(\lambda', t)|^2\, dt\, d^{n-1}\lambda' \right\}^{1/2}. \end{aligned} \tag{10.31}$$

Using that $|2\,a\,b| \leq a^2 + b^2$ and Plancherel's identity we obtain

$$\begin{aligned} ||\tau f||^2_{H^{1/2}(R^{n-1})} &\leq \int_{R^n_+} \left\{ |f(x)|^2 + \textstyle\sum_{k=0}^{n-1} |\partial_k f|^2 + |\partial_n f|^2 \right\} d^n x \\ &= ||f||_{H^1(R^n_+)}. \end{aligned} \tag{10.32}$$

To prove surjectivity we need to see that given $g \in H^{1/2}(I\!R^{n-1})$ there exists at least one (in fact infinitely many) $f \in H^1(I\!R^n_+)$ such that $\tau f = g$. Again it is enough to define an anti-trace K on $C^\infty(I\!R^n)$ and prove that there exists $C > 0$ such that

$$||K\,g||_{H^1(R^n_+)} < C\,||g||_{H^{1/2}(R^{n-1})} \quad \forall\, g \in C^\infty_0. \tag{10.33}$$

Let $K(g) = \mathscr{F}^{-1}(e^{-(1+|\lambda'|^2)^{1/2} x^n} \mathscr{F}(g))$ and noting that the argument of $\mathscr{F}^{-1}$ is in $\mathscr{S}(I\!R^{n-1})$ we leave the proof of the above inequality as an exercise.

In applications, we will have to consider functions defined only in open sets of $I\!R^n$, Ω and their boundaries $\partial\Omega$. We will assume that Ω is such that its boundary, $\partial\Omega = \bar{\Omega} - \Omega$, is a smooth manifold.

Let $H^m(\Omega)$ be the Sobolev space obtained by taking the integral in the inner product simply over Ω and completing the space $C^\infty_0(\Omega)$ with respect to its norm and

let $H_o^m(\Omega)$ be the one obtained with the same norm but completing the space $C^\infty(\Omega)$. If $m = 0$ or if $\Omega = \mathbb{R}^n$ then these spaces coincide. But if $m \geq 1$ and $\partial\Omega$ is non-empty then they are different (obviously $H_o^m \subset H^m$) since as derivatives are controlled in the norm, the sequences of compactly supported functions cannot converge to a non-zero function on the boundary. How will we extend the results obtained in $\mathbb{R}^n$ to Ω? The key lemma for this is the following.

Lemma 10.8 *Let γ be the restriction of functions in $\mathbb{R}^n$ to functions in Ω, then*

$$\gamma : H^m(\mathbb{R}^n) \to H^m(\Omega)$$

is continuous and surjective.

Proof: The continuity is clear, it only remains to see the surjectivity. We will prove surjectivity for the case $\Omega = \{x \in \mathbb{R}^n,\ x_n > 0\}$, $\partial\Omega = \{x \in \mathbb{R}^n \ / \ x_n = 0\}$. Let $f \in H^m(\Omega)$ then by the previous Lemma $\tau(\partial_n^k f) \in L^2$ for $k = 0, 1, \ldots, m-1$. We extend f for x_n negative as,

$$f(x) \equiv \left[\sum_{k=1}^{m-1} \frac{(-x_n)^k}{k!}\, \tau(\partial_n^k f)\right] \left(1 - e^{1/x_n^2}\right),\ x_n < 0 \tag{10.34}$$

The summation makes the derivatives of f continuous at $x_n = 0$ and the exponential makes the H^m norm bounded. But this norm will only depend on the norms of $\tau(\partial_n^k f)$ which in turn are bounded by $||f||_{H^m(\Omega)}$ and therefore there will exist $c > 0$ such that $||f(x)||_{H^m(R^n)} < C\ ||f||_{H^m(\Omega)}\ \ \forall\ f \in C_o^\infty$. The general case is technically more cumbersome, the basic idea is to use charts that map regions containing part of the boundary of Ω into $\mathbb{R}^{n+}$ mapping boundary to boundary. In each one of these charts we know how to prove the corresponding inequality. The global inequality is proved by assuming, for the sake of contradiction, that it does not hold ♠

This lemma allows us to immediately generalize the previous lemmas.

Definition: We will say that $f \in H^s(\Omega)$ if it admits an extension $\tilde{f}$ to $\mathbb{R}^n$ such that $\tilde{f} \in H^s(\mathbb{R}^n)$ [Note that this agrees with what was proved for positive integer s].

It is immediate then that if $s' > s$ then $H^{s'}(\Omega) \subset H^s(\Omega)$, that if $m < s - n/2$ then $H^s(\Omega) \subset C^m(\Omega)$ and that if $m > 0$ integer $\tau : H^m(\Omega) \to H^{m-1/2}(\partial\Omega)$ is continuous and surjective. In the last statement we use $H^{m-1/2}(\partial\Omega)$ where in general $\partial\Omega \neq$ open in $\mathbb{R}^{n-1}$ and therefore does not fit the previous definition. In the case where $\partial\Omega$ is compact we will say that $f \in H^{m-1/2}(\partial\Omega)$ if, given any sufficiently small open cover $\{U_i\}$ of $\partial\Omega$ such that there exists φ_i, such that (U_i, φ_i) is a chart of $\partial\Omega$ then $f \in H^{m-1/2}(U_i)\ \forall\ i$.

10.4 Weak Compactness and Compact Embeddings

We conclude this chapter with two lemmas. The second of them, a consequence of the first, is of great importance in the theory of partial differential equations.

Lemma 10.9 *Let Γ_d be a cube in $\mathbb{R}^n$ with sides of length $d > 0$. If $u \in H^1(\Gamma_d)$ then,*

$$||u||^2_{H^0(\Gamma_d)} \leq d^{-n} \left| \int_{\Gamma_d} u d^n x \right|^2 + \frac{nd^2}{2} \sum_{j=1}^{n} ||\partial_j u||^2_{H^0(\Gamma_d)}. \tag{10.35}$$

Proof: It is sufficient to consider $u \in C^1(\Gamma_d)$. For any x and y in Γ_d we have,

$$u(y) - u(x) = \sum_{j=1}^{n} \int_{x^j}^{y^j} \partial_j u(y^1, ..., y^{j-1}, s, x^{j+1}, ..., x^n) ds. \tag{10.36}$$

Taking its square and using the Schwarz inequality we obtain,

$$\begin{aligned} |u(x)|^2 + |u(y)|^2 \quad &- \quad 2\Re(u(x)u(y)) \\ &\leq \quad nd \sum_{j=1}^{n} \int_{-d/2}^{d/2} |\partial_j u(y^1, \ldots, y^{j-1}, s, x^{j+1}, \ldots, x^n)|^2 ds. \end{aligned}$$

Integrating with respect to all x^j and y^j we obtain,

$$2\, d^n\, ||u||^2_{H^0(\Gamma_d)} \leq 2 \left| \int_{\Gamma_d} u\ d^n x \right|^2 + n\, d^{n+2} \sum_{j=1}^{n} ||\partial_j\, u||^2_{H^0(\Gamma_d)}, \tag{10.37}$$

from which the desired inequality trivially follows ♠

Lemma 10.10 *Let $\Omega \subset \mathbb{R}^n$ be compact with smooth boundary. If a sequence $\{u_p\}$ in $H^1_o(\Omega)$ is bounded, then there exists a subsequence that converges (strongly) in $H^0(\Omega)$. That is, the natural map [6] $H^1_o(\Omega) \to H^0(\Omega)$ is compact.*

Proof: Let $k = sup\{||u_p||_{H^1_o(\Omega)}\}$. Since Ω is bounded we can enclose it in a cube Γ_D and extend each u_n as zero in $\Gamma_D - \Omega$. Let $\varepsilon > 0$ and M be large enough such that $\frac{2nk^2D^2}{M^2} < \varepsilon$. Decomposing Γ_D into M^n cubes Γ^j_d with $d = D/M$ and using that since $\{u_n\}$ is also bounded in $H^0(\Omega)$, the weak compactness of balls in Hilbert spaces ensures that there exists a subsequence $\{\tilde{u}_p\}$ and $u \in H^0(\Omega)$ such that it converges weakly to u. We then see that there exists an integer N such that if p and $q > N$,

$$\left| \int_{\Gamma^j_d} (\tilde{u}_p - \tilde{u}_q)\, d^n x \right|^2 < \frac{\varepsilon}{2} \left(\frac{D}{M} \right)^{2n} \frac{1}{D^n}. \tag{10.38}$$

[6] By natural map between two Banach spaces we mean the following: recall that the elements of each of these spaces are equivalence classes of Cauchy sequences, which we can assume consist of smooth elements of some dense space, for example $C^\infty(\Omega)$. The natural map is defined as the one that sends element by element a Cauchy sequence from one space to the other. Since the norm of the output space bounds the norm of the input space, the sequence obtained in the output space is also Cauchy and therefore determines an equivalence class there, hence an element of the space.

If we apply lemma (10.9) to each Γ_d^j and sum over j we obtain that $\forall\ p$ and $q > N$,

$$\begin{aligned} ||\tilde{u}_p - \tilde{u}_q||^2_{H^\circ(\Omega)} &\leq \left(\frac{D}{M}\right)^{-n}\left(\sum_{j=1}^{M^n} \frac{\varepsilon}{2}\left(\frac{D}{M}\right)^{2n} \frac{1}{D^n}\right) + \frac{n}{2}\left(\frac{D}{M}\right)^2 (2k^2) \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &\leq \varepsilon. \end{aligned} \tag{10.39}$$

But then $\{\tilde{u}_p\}$ is a Cauchy sequence in $H^\circ(\Omega)$ and therefore converges strongly to u in $H^\circ(\Omega)$ ♠

Bibliography notes: I recommend the books: [15] and [9]. Sobolev spaces allowed us to understand much of the theory of nonlinear equations and previously the theory of linear equations with non-smooth coefficients. They are not complex ideas, but very useful and essential for research in partial differential equations.

CHAPTER 11

THEORY OF PARTIAL DIFFERENTIAL EQUATIONS

11.1 Introduction

Definition: A **partial differential equation of order** m in M is an equation of the form

$$F\left(p, u, \nabla_a u, \nabla_a \nabla_b u, \ldots, \overbrace{\nabla_a \cdots \nabla_c}^{m \text{ times}} u\right) = 0, \tag{11.1}$$

where ∇_a is some connection in M. More generally, u can be a tuple of tensor fields and F can have range in some other tuple of tensor fields.

Examples:

a) The Laplace equation in $\mathbb{R}^3$ with respect to a metric g_{ab},

$$\Delta u \equiv g^{ab} \nabla_a \nabla_b u = 0 \tag{11.2}$$

If g_{ab} is the Euclidean metric, then in Cartesian coordinates

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}. \tag{11.3}$$

b) The Poisson equation,

$$\Delta u - \rho = 0, \tag{11.4}$$

where ρ is a given function.

c) The wave equation in $\mathbb{R}^{n+1}$. This has the form of the Laplace equation but for a metric of the form $-(dx_0)^2 + \sum_{i=1}^{n} (dx^i)^2$. For example in $\mathbb{R}^2$, $g_{ab} = -(dt)^2_{ab} + (dx)^2_{ab}$,

$$g^{ab} \nabla_a \nabla_b u = -\frac{\partial^2 u}{\partial t^2} + \frac{\partial^2 u}{\partial x^2}. \tag{11.5}$$

d) Maxwell's equations,

$$\begin{aligned} \nabla_a F^{ab} &= j^b \\ \nabla_{[a} F_{bc]} &= 0 \end{aligned} \tag{11.6}$$

where M is $\mathbb{R}^4$, the metric (used both to raise the indices of F_{ab} and to define ∇_c) is the Minkowski metric, $g_{ab} = -(dx^0)^2 + dx^1)^2 + (dx^2)^2 + (dx^3)^2$, F_{ab} is an antisymmetric tensor field in M, and y^b is a vector field (the four-current) in M.

e) The elasticity equations in $\mathbb{R}^3$(Euclidean)$\times\mathbb{R}$

$$\rho \frac{\partial^2 u^a}{\partial t^2} = \mu \Delta u^a + (\lambda + \mu) \nabla^a (\nabla_c u^c), \tag{11.7}$$

where u^a is the displacement vector (in $\mathbb{R}^3$), ρ the density of the elastic medium and λ and μ the LamÃ© constants of the medium.

f) The heat conduction equation in $\mathbb{R}^3$(Euclidean)$\times\mathbb{R}$

$$\frac{\partial u}{\partial t} = k \Delta u, \qquad k > 0 \tag{11.8}$$

g) The SchrÃ¶dinger equation in $\mathbb{R}^3$(Euclidean)$\times\mathbb{R}$,

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \Delta \psi + V \psi, \tag{11.9}$$

where ψ is a complex function and V a potential.

h) The Navier-Stokes equation for a viscous and incompressible fluid (e.g., water) in $\mathbb{R}^3$(Euclidean)$\times\mathbb{R}$

$$\frac{\partial u^a}{\partial t} + u^b \nabla_b u^a + \frac{1}{\rho} \nabla^a p - \gamma \Delta u^a = 0 \tag{11.10}$$

$$\nabla_a u^a = 0, \tag{11.11}$$

where u^a is the velocity vector of the fluid, p its pressure, ρ its density (constant) and γ the kinematic viscosity.

From the cited examples, of which only the last one is not linear, we see the tremendous physical importance of having a general theory of these equations, and thus this is one of the most active branches of mathematics. Due to its complexity, in the large number of different cases it presents, the general theory is far from complete, however, there are cases or classes of equations where this has been achieved. One of these is the case of a single first-order equation, where as we will see the problem reduces to that of ordinary differential equations.

Another of these cases is that of linear equations with constant coefficients (that is, for which there exists a coordinate system in which all the coefficients are constant - for example, the Laplacian for a Euclidean metric). This is mainly due to the use of transforms such as the Fourier transform. However, I note that there are recent works showing new results, even in the case of the Laplacian! If we allow the coefficients to vary, the problem becomes more complicated, however, certain subclasses of these have been completely studied. If we add non-linearity in a not too drastic form – what is known as quasi-linear equations – the knowledge is drastically reduced, although some subclasses have been overcome and some particular equations completely studied. The case where the non-linearity is drastic has not yet been successfully tackled - of any kind. Fortunately, the physical problems that we have been able to model or describe by partial differential equations so far have equations at most of the quasi-linear type. In this course, we will see in detail only the theory of some of the simplest equations [essentially the equations in examples a), b), c) and f)], always trying to use methods that can be applied to similar but more complex equations. This is not only due to the simplicity of these equations, which allows their complete and detailed knowledge but also because on the one hand, they represent the *"canonical examples"* of different classes of equations. The solutions of the equations in each of these classes behave very similarly while they do so in a radically different way from the solutions to equations in the other classes. On the other hand, these are the most used equations in physics and appear in a multitude of different problems, even as particular cases of the equations in examples d), e), f) and g)!

11.2 The First Order Equation

This is an equation of the form

$$F(p, u, \nabla_a u) = 0. \tag{11.12}$$

where u is a function in some manifold M. This equation can be tackled very successfully and results in ordinary equations, whose theory we already know. For simplicity, we will consider here only the quasi-linear case and in $\mathbb{R}^2$, that is, an equation of the form,

$$a(x,y,u)\,u_x + b(x,y,u)\,u_y = c(x,y,u), \tag{11.13}$$

where $u_x = \frac{\partial u}{\partial x}$ and $u_y = \frac{\partial u}{\partial y}$.

For geometric reasons, it is useful to represent the solutions of this equation in $\mathbb{R}^3$ or more precisely in a region Ω of $\mathbb{R}^3$ where a, b, and c are defined, that is to associate a solution $u(x,y)$ of (11.13) with the hypersurface of $\mathbb{R}^3$ given by $\tau = z - u(x,y) = cte$. These hypersurfaces are called integral surfaces of the equation (11.13). The gradient of τ in these coordinates is $(\nabla_a \tau) = (-u_x, -u_y, 1)$, so we see that the equation (11.13) is simply the condition that τ is constant along the vector field $(l^a) = (a(x,y,z),\ b(x,y,z),\ c(x,y,z))$, that is $l^a \nabla_a \tau = 0$, which is equivalent to saying that l^a is **tangent** to the integral surfaces. Note that if $l^a \nabla_a \tau = 0$ then

$(f\,l^a)\nabla_a\,\tau = 0$ so what determines the equation (11.13) is not l^a but its direction. This field of directions is called **characteristic directions**, and their integral curves **characteristic curves**. [1] The theory of ODEs then tells us that through each point of Ω passes a unique characteristic curve. The knowledge of these curves is fundamental since if we form a surface S by taking the union of certain characteristic curves then clearly S will be an integral surface of (11.13), but on the other hand, given an integral surface S and any $p \in S$ the characteristic curve that passes through p will be tangent to S at every point and therefore will be a submanifold of S so that S will be formed by the union of characteristic curves.

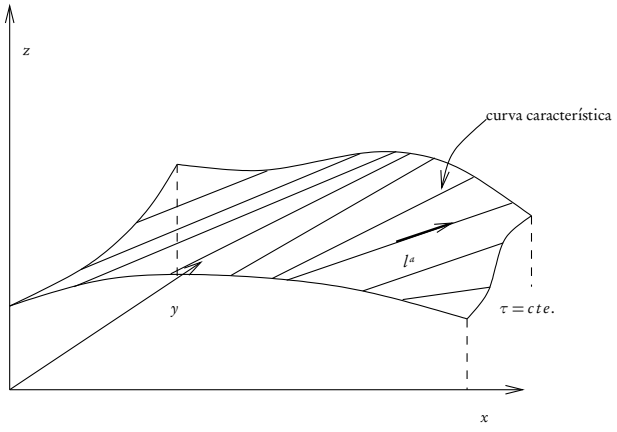

Figure 11.1: Characteristic curves.

In particular, note that if two integral surfaces S, S', that is two solutions of (11.13), u and u', have a point p in common then they must have an entire characteristic curve in common, since through p only one of such curves passes and it cannot leave either of the two surfaces. On the other hand, if S and S' intersect in a curve γ this must be integral. To see this, take a point p of γ and consider T_pS and T_pS'; since the surfaces intersect in a curve these two subspaces of $T_p\mathbb{R}^3$ intersect in a line, as both must contain the direction given by l^a this will be the line. But this is true for every point of γ and therefore the tangent vector to γ is proportional to l^a and γ is characteristic.

[1] Recall that an integral curve is the "image" of a curve $\gamma(t)$ (in this case $= (x(t), y(t), z(t))$ solution of an ODE, in this case

$$\frac{d\gamma(t)}{dt} = \frac{d}{dt}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a(x,y,z) \\ b(x,y,z) \\ c(x,y,z) \end{pmatrix}. \tag{11.14}$$

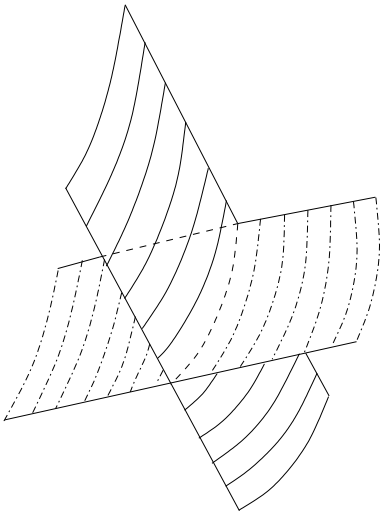

Figure 11.2: Intersection of solutions.

11.2.1 The Cauchy Problem

The **Cauchy problem** of an equation is the problem of finding certain data such that giving these there exists a unique solution of this equation. That is, finding a certain set whose elements are called data and a map between this space and the set of solutions of the equation. For example, in the case of ODEs, the problem consists of given a smooth vector field v^a in M finding some set such that each element of this corresponds to an integral curve of v^a. Clearly, we could take as this set the points of M where $v^a \neq 0$ since through each of these points passes a unique integral curve. Clearly, we could also take as a set of data – at least locally – a hypersurface s of M such that at none of its points v^a is tangent to it. In such a case, we also have the very desirable property that each point of S determines (locally) a unique solution, that is, we do not count each solution more than once.

What will these data be in the case of the equation (11.13)?
Let $\gamma(s) = (x_0(s), y_0(s), z_0(s))$, $s \in [0, 1]$, be a curve in $\mathbb{R}^3$. We will look for a solution such that its integral surface contains γ, that is, a $u(x, y)$ such that it satisfies,[2]

$$z_0(s) = u(x_0(s), y_0(s)) \qquad \forall\, s \in [0, 1]. \tag{11.15}$$

We will first consider the case where $\gamma(s)$ is not a characteristic curve. Taking each point $\gamma(s)$ as the initial point for the ordinary differential equation that determines l^a and solving this we obtain for each s the characteristic curve that passes through $\gamma(s)$. (See figure.)

[2]Only the image of the curve matters and not its parameterization, so we will take one in which the range of S is the interval $[0, 1]$.

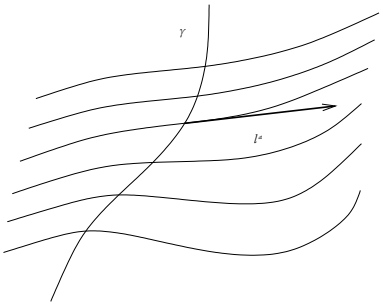

Figure 11.3: Constructing the solution from the curve γ.

We thus obtain a map $\gamma(s,t): I_s \times I_t \to \mathbb{R}^3$ given by,

$$\gamma(s,t) = (x(s,t), y(s,t), z(s,t)) \tag{11.16}$$

with $x(s,0) = x_0(s)$, $y(s,0) = y_0(s)$ and $z(s,0) = z_0(s)$ and where at fixed s

$$\frac{d\gamma(s,t)}{dt} = l^a(\gamma(s,t)). \tag{11.17}$$

If we could invert the functions $x(s,t)$ and $y(s,t)$ and thus obtain s and t as functions of x and y then

$$u(x,y) \equiv z(s(x,y), t(x,y)) \tag{11.18}$$

would be the sought solution, since such u satisfies by construction (11.13) and (11.15).

It is not always possible to make such an inversion. The reason is that in general, there will be values of s and t such that the tangent plane to $\gamma(s,t)$ at that point contains the z axis. Let's see if there are conditions that ensure that such an inversion is possible at least locally, that is, in a neighborhood of some point $(x(s_0,0), y(s_0,0))$ on $\gamma(s)$.

The implicit function theorem tells us that this will be possible if the differential of the transformation at that point $(s_0,0)$ is invertible, that is, if its determinant (the Jacobian of the transformation) is not zero. At this point we have,

$$J = \begin{vmatrix} \left.\frac{\partial x}{\partial s}\right|_{(s_0,0)} & \left.\frac{\partial y}{\partial s}\right|_{(s_0,0)} \\ \left.\frac{\partial x}{\partial t}\right|_{(s_0,0)} & \left.\frac{\partial y}{\partial t}\right|_{(s_0,0)} \end{vmatrix} = \frac{\partial x}{\partial s}(s_0)\, b_0 - \frac{\partial y}{\partial s}(s_0) a_0 \neq 0, \tag{11.19}$$

where $a_0 = a(x(s_0), y(s_0), z(s_0))$ and $b_0 = b(x(s_0), y(s_0), z(s_0))$. This is then the condition for the local existence of solutions and tells us that $\gamma(s)$ must be chosen such that its projection in the (x,y) plane has a tangent vector that is not proportional to the projection in that plane of the vector (a,b,c).

Example: In some applications, the coordinate y is time. In such a case, it is natural to specify u at an instant of time, say $y = 0$, that is, to give its **initial value**. Thus, the **initial value problem** simply consists of choosing $\gamma(s) = (s, 0, h(s))$. The equation (11.15) then results in $h(s) = u(s, 0)$ or $h(x) = u(x, 0)$, that is, $h(s)$ will be the initial value that the solution will have at $y = 0$. In this case, there will be a local solution as long as b, the coefficient of $\dfrac{\partial u}{\partial y}$, does not vanish at $(x, 0, h(x))$. If γ were a characteristic curve, then there would be infinite (local) solutions since given a point $\gamma(s)$ of γ and a non-characteristic curve $\gamma^\star(r)$ passing through this point, we could construct, using the previous procedure, but now with $\gamma^\star(r)$, a solution (a surface) $\gamma^\star(r, t)$ that would necessarily contain γ.

Exercise: Solve the equations using the described method:

a)

$$\frac{\partial u}{\partial y} + c\frac{\partial u}{\partial x} = 0$$
$$u(x, 0) = h(x).$$

b)

$$\frac{\partial u}{\partial y} + u\frac{\partial u}{\partial x} = 0$$
$$u(x, 0) = -x.$$

c) For how long ($y = t$) can the solutions of b) be extended?

11.3 Classification of Partial Differential Equations

To facilitate the classification, we will proceed similarly to how we did when we studied ordinary differential equations, that is, we will reduce the systems of equations to first-order systems. To do this, we will take as independent variables all the derivatives of lower order than the highest order in which each of the variables appears.

Example: Let $\phi : \mathbb{R}^2 \to \mathbb{R}$ satisfy

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = \rho, \tag{11.20}$$

that is, the Laplacian in a two-dimensional flat manifold. Define $u^1 := \phi$, $u^2 := \frac{\partial \phi}{\partial x}$ and $u^3 := \frac{\partial \phi}{\partial y}$, then this equation is equivalent to the following system

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \partial_x \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \partial_y \begin{pmatrix} u^1 \\ u^2 \\ u^3 \end{pmatrix} = \begin{pmatrix} \rho \\ u^2 \\ u^3 \\ 0 \end{pmatrix} \tag{11.21}$$

The reason why we added the fourth equation will be seen later, but we anticipate that it allows us to deduce, without resorting to the equations of the second and third rows, that the components u^2 and u^3 are the components of the differential of some function. If we know a solution of 11.21, (u^1, u^2, u^3) we can prove that u^1 also satisfies the Laplacian. Indeed, taking a derivative with respect to x of the second row and a derivative with respect to y of the third, summing them and then using the first row we obtain the desired result. Of all the first-order systems (and basically for the same reason we gave when we considered systems of ordinary equations) we will only consider those of the form,

$$M^a_{A'B} \nabla_a u^B = I_{A'}, \tag{11.22}$$

with $M^a_{A'B}$ and $I_{A'}$ smooth functions of the point in the manifold (fields) and of u^A.

The indices we are using are abstract indices and can denote not only a set of scalars, but also a large vector made up of various vector fields. We will see this in examples. Coordinate systems and bases can also be taken and then the problem can be posed in components, as we did in the previous example, in which case we can think of u^A as a large vector array of scalar fields. We have used primed and unprimed indices to make it clear that the (co-vectorial) space of the primed indices does not have the same dimension as the vector space of the unprimed indices.

The type of system we have just defined is called quasi-linear, since the derivatives appear linearly. This is not a loss of generality from the point of view of physics:

¡*All known physical systems are of this form!*

Historically, the classification we will give below arises from the attempt to frame all equations in the Cauchy Problem, that is, to take a hypersurface Σ of M, give u as data there and obtain through the equations its derivative in the transverse direction and from these data try to construct solutions in a neighborhood of Σ in M (local solutions). This attempt was not generally successful, since in general this is not the correct way to give data, but the classification of these equations thus obtained was successful, since the properties of the solutions to the equations in each of these classes in which we will classify them are very similar and in each of these classes the data are also prescribed in a similar way.

To fix ideas, let M be of dimension n and let $p \in M$. We want to find the solutions in a neighborhood of p giving u^A as data on some hypersurface Σ of M that contains p. For simplicity, we will do the necessary calculations in a coordinate system adapted to the problem in the sense that we will choose the coordinate x^n in such a way that the submanifold $x^n = 0$ is the surface Σ and p is the origin of coordinates. The data will then be $\Phi^A(x^1, ..., x^{n-1})$, which will correspond to the solution u^A restricted to Σ, that is $\Phi^A = u^A|_\Sigma$.

Since we know what u^A will be on Σ, we know what all its derivatives (of any order) will be with respect to the coordinates x^i, $i = 1, ..., n-1$. The idea is now to use the equation to find $\partial_n u^A|_\Sigma$ and thus successively find all the derivatives of u at p (or any other point of Σ). If we could achieve this, we could, at least formally and in a

neighborhood of p, construct u^A as its Taylor series around p. Therefore, if we could solve for the normal derivatives to Σ, if in addition the Φ^A were analytical data, and if the coefficients of the equation were also analytical, then we could prove (Cauchy-Kowalevski Theorem) that the solution u^A formally constructed above actually exists and is analytical in a neighborhood of p in M.

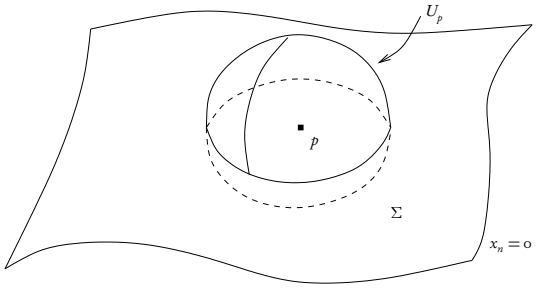

Figure 11.4: Constructing the local solution.

What requirements must the equation satisfy for this to be possible? Using the mentioned coordinate system it can be seen that

$$M(\Phi^C, q)^n_{A'B} \partial_n u^B|_\Sigma = \text{Terms in } (\Phi^A, \partial_i \Phi^C, q)_{A'}, \tag{11.23}$$

It is clear then that we can solve for $\partial_n u^A$ only if $M(\Phi^C, q)^n_{A'B}$ is invertible. In general (as in the example we gave) the target space of the map $M^n_{A'B}$ does not have the same dimension as the source space and therefore the map is not generally invertible, so we will only ask that **the rank of this map has the maximum dimension** (that is, the dimension of the source space –that of the vectors u^A–). In particular, this implies that we are assuming that the dimension of the target space is greater than or equal to that of the source space. We are asking for the minimum possibility of having unique solutions, indeed, the maximality of the rank of $M(\Phi^C, q)^n_{A'B}$ is equivalent to its kernel being of zero dimension and therefore, if a solution exists it will be unique.

If this does not happen, that is, if the kernel is not trivial, we will not be able to solve for the normal derivatives of the solution in terms of the initial data and therefore there will not be a unique solution. Note that it may still happen that having to satisfy more equations than the unknowns present may make the equations inconsistent and there may be no solution. In some cases where this happens, the resolution of this problem consists in realizing that the data to be freely given will be a number less than the dimension of the source space, only some components of u^A can be given as initial data.

What is geometrically the condition of maximality of the rank of $M^a_{A'B}$ in terms of the chosen coordinates?

If the surface $x^n = 0$ is a smooth surface then we can assume that $\nabla_a x^n$ exists and is non-zero, in this case the previous condition is simply the condition that the matrix

$M^a_{A'B}\nabla_a x^n$ has maximum rank, but this condition is independent of the coordinate system and only depends on Σ, since here x^n is simply a function in M that is constant on Σ and whose gradient does not vanish. If we take another function with the same characteristics, say $\tilde{x}^n$, then we will have that their gradients are proportional, that is, there will exist α such that $\nabla_a \tilde{x}^n = \alpha \nabla_a x^n$. The matrices will then be proportional and the conditions on the kernels identical.

The surfaces where the previous condition is violated at all points are called **characteristic surfaces**. We will classify the equations according to the number of characteristic surfaces that intersect at a given point. Note that the classification only depends on the tensor $M^a_{A'B}$ and not on the lower-order terms. $M^a_{A'B}\nabla_a u^B$ is called the **principal part** of the equation. Also note that since $M^a_{A'B}$ is a tensor field not necessarily constant, the condition defining the characteristic surfaces can be very different from point to point, therefore the classification we will introduce is generally valid only for the point in question. Fortunately, in applications, equations where their type changes from region to region rarely appear. Also note that for non-linear systems the condition also depends on the solution one is considering! Since our classification is based only on the principal part of the equation, it must contain all the information about the equation, which is why in the previous example we added the last row in the system of equations; without it and considering the principal system, we would not know that the last two components of u^A had to be the components of the differential of a function.

We will say that an equation is **elliptic at** $p \in M$ if p is not intersected by any characteristic surface. That is, if Rank $M^a_{A'B}k_a$ is not maximal then $k_a = 0$ [since we are at a point, the condition that k_a is a gradient is irrelevant].

Exercise: The canonical example of an elliptic equation is the Laplacian in M, $\Delta u := g^{ab}\nabla_a\nabla_b\phi = \rho$ where g^{ab} is a Riemannian metric. Show that the given example is an elliptic system.

We will say that an equation is **parabolic at** $p \in M$ if p is intersected by a unique characteristic surface. That is, there exists a unique $k_a \neq 0$ –up to a multiplicative factor– such that Rank $M^a_{A'B}k_a$ is not maximal.

The canonical example of a parabolic (or diffusion) equation is the heat equation in $\mathbb{R}\times\mathbb{R}^{n-1}$,

$$\partial_t u = \Delta u. \tag{11.24}$$

Exercise: Consider the previous equation in $\mathbb{R}\times\mathbb{R}$, $\partial_t u = \frac{\partial^2 u}{\partial x^2}$. Reduce this system to first order and find the characteristic surfaces of this equation.

We will say that an equation is **hyperbolic at** $p \in M$ if it is intersected by more than one characteristic surface.

The canonical example in this case is the wave equation in M

$$\Delta u = g^{ab}\nabla_a\nabla_b u, \tag{11.25}$$

where g^{ab} is a metric such that given any point $p \in M$ there exists a coordinate system in which g_{ab} takes the form,

$$g_{\mu\nu}|_p = \{-(dt)^2 + \sum_i (dx^i)^2\}|_p. \tag{11.26}$$

Exercise: Consider in $\mathbb{R}^2$ the metric $ds^2 = -(dt)^2 + (dx)^2$ (at every point) and the equation $g^{ab}\nabla_a\nabla_b u = \rho$.
a) Reduce the equation to first order.
b) Find the characteristics of the equation.
c) Do the same in $\mathbb{R} \times S^1$ (a spacetime cylinder) with the following two metrics, $ds^2 = -(dt)^2 + (d\theta)^2$ and $ds^2 = -(dt)^2 + t^2(d\theta)^2$.

Bibliography notes: Recommended reading for this and the following chapters: [15], [16], [17] and [19]. What is presented in these chapters is the minimum and essential to have an understanding of this area. Much more is known and at the same time, as always, there is much more that we do not know, weak solutions, global existence, shock waves, boundary conditions, stability, etc., etc. This is probably the most active area, with the most people working and with the largest number of applications in all of mathematics. Most of these applications have traditionally been in the area of engineering and particularly in fluids, which has meant that only certain specific types of equations, quite difficult by the way, were treated, and not the most used in other areas of physics. This has evolved in recent years and now there is a considerable shift of attention towards many of the problems of modern physics.

CHAPTER 12

ELLIPTIC EQUATIONS

12.1 The Laplace Equation

We will take the Laplace equation in $\mathbb{R}^n$ with Euclidean metric $ds^2 = \sum_{i=1}^{n} (dx^i)^2$, or an open region Ω of $\mathbb{R}^n$ with the same metric, as a model of an elliptic equation. The results we obtain that do not depend on special functions can be generalized to elliptic equations of the form (12.45) if it is assumed that $c \leq 0$.

As we mentioned before, the Cauchy program, that is, formulating the problem of obtaining solutions by giving u^A on a hypersurface Σ, does not generally work, it only works for hyperbolic equations. If we consider non-analytic data, and physical data are generally non-analytic, there are generally not even local solutions. What is the appropriate way then to give data for the Laplace equation? We will obtain the answer to this by considering a physical phenomenon described by this equation.

Consider a drum and apply a (small) force on its membrane (or patch) perpendicular to it.

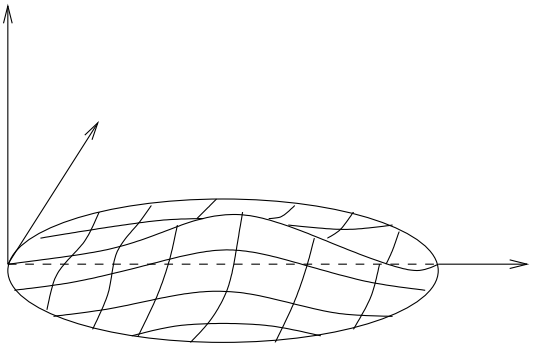

Figure 12.1: The drum membrane.

The membrane will move from its flat (rest) position and acquire a new equilibrium position generating an elastic force that exactly cancels the applied one. What

will this new shape of the membrane be? If we denote by $u(x,y)$ the displacement of it in the direction of the vertical axis (z) and f the force per unit area applied (in the vertical direction), it can be seen that u must satisfy the equation,

$$\Delta u = f, \text{in } \Omega = \{(x,y) | \sqrt{x^2+y^2} < \text{drum radius}\}. \tag{12.1}$$

Exercise: Convince yourself that (12.1) is the equilibrium equation for the case where the applied force is small enough to produce a small deformation of the membrane. Hint: First do the one-dimensional case.

Since the edge of the drum holds the membrane there, we will have the following **boundary condition**,

$$u|_{\partial\Omega} = 0, \;\; \partial\Omega = \{(x,y) \mid \sqrt{x^2+y^2} = \text{drum radius}\}. \tag{12.2}$$

Thus we have the following problem: Given f in Ω, find u satisfying (12.1) and (12.2). From our physical experience, we know that this problem should be solvable! Moreover, the problem should be solvable where the drum is not circular but has an arbitrary shape, allowing $\partial\Omega$ to be an arbitrary but smooth edge and also allowing $\partial\Omega$ not to be in the plane $z = 0$, or -equivalently- to be in such a plane but allowing u to take any –but smooth– value ϕ on $\partial\Omega$.

We thus arrive at the following **Dirichlet Problem:** Given Ω with smooth $\partial\Omega$, $f : \Omega \to \mathbb{R}$ smooth and $\phi_0 : \partial\Omega \to \mathbb{R}$, also smooth, find $u : \Omega \to \mathbb{R}$ satisfying,

1. $\Delta u = f \quad \text{in } \Omega$,
2. $u|_{\partial\Omega} = \phi_0$.

Later we will reaffirm our intuition by seeing that this problem can always be solved.

Now suppose we allow the edges of the membrane to slide vertically but place vertical springs on the edge with a Hooke constant that depends on the position, $k : \partial\Omega \to \mathbb{R}$, and arranged in such a way that the equilibrium position before applying f is $u = 0$. When we apply f, the membrane will move until again in the interior we have,

$$\Delta u = f \quad \text{in } \Omega \tag{12.3}$$

and on the edge

$$(ku + n^a \nabla_a u)|_{\partial\Omega} = 0, \tag{12.4}$$

where n^a is the normal to $\partial\Omega$. This **Mixed Problem** can also be solved.

A particular case of this is when instead of using springs to apply the edge force, we simply use a given force $\phi_1 : \partial\Omega \to \mathbb{R}$, but taking care that its total contribution, $\int_{\partial\Omega} \phi_1 dS$, exactly cancels the total contribution of f –otherwise we would have an accelerating membrane–. In such a case we have the **Neumann Problem:**

$$\Delta u = f \quad \text{in } \Omega \tag{12.5}$$

$$n^a \nabla u|_{\partial\Omega} = \phi_1, \text{ on the edge } \partial\Omega, \tag{12.6}$$

$$\text{with} \int_{\partial\Omega} \phi_1 \, dS = \int_{\partial\Omega} n^a \nabla_a u \, dS = \int_\Omega \Delta u \, dV = \int_\Omega f \, dV. \tag{12.7}$$

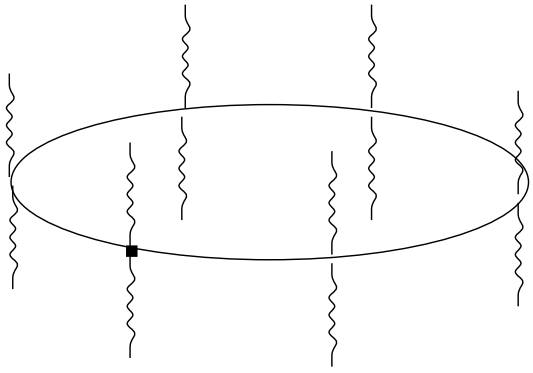

Figure 12.2: Mixed problem.

12.1.1 Existence

Next, we will see that the Dirichlet problem always has a unique solution (assuming that f, ϕ_o, and $\partial\Omega$ are sufficiently smooth). The other problems are solved similarly.

Suppose there exists $u \in H^2(\Omega)$ satisfying, $\Delta u = f$ in Ω and $u|_{\partial\Omega} = 0$ –which implies $u \in H^1_o(\Omega)$–. Then using the divergence theorem we obtain the **first Green's identity**, $\forall v \in C^\infty(\Omega)$.

$$\int_\Omega v\bar{f}\, d^n x = \int_\Omega v\Delta\bar{u}\, d^n x = -\int_\Omega e^{ab}\nabla_a v\nabla_b\bar{u}\, d^n x + \int_{\partial\Omega} v(n^a\nabla_a\bar{u})\, d^{n-1}x, \quad (12.8)$$

If we assume that $v \in H^1_o(\Omega)$ the identity is still valid and in this case reduces to,

$$\int_\Omega e^{ab}\nabla_a v\nabla_b\bar{u}\, d^n x = -\int_\Omega v\bar{f}\, d^n x, \quad \forall\ v \in H^1_o(\Omega). \quad (12.9)$$

But note that this identity is still valid if we simply assume that $u \in H^1_o(\Omega)$ –not necessarily in $H^2(\Omega)$– and that $f \in H^{-1}(\Omega)$.

Thus we have the **Weak Dirichlet Problem** (with $\phi_o = 0$): Find $u \in H^1_o(\Omega)$ such that given $f \in H^{-1}(\Omega)$, (12.9) is satisfied.

If the right-hand side of (12.9) were an inner product, this would give rise (by completion) to a Hilbert space H and then (12.9) would take the form

$$\langle u, v\rangle_H = \Phi_f(v) := -\int_\Omega \bar{f}v\, d^n x, \forall v \in H^1_o(\Omega). \quad (12.10)$$

If $H = H^1_o(\Omega)$ and if $\Phi_f : H \to C$ were continuous, then the Riesz representation theorem would tell us that there exists a unique u in H satisfying (12.9) and therefore (12.10). As we will see below (PoincarÃ©-Hardy lemma), in this case, the right-hand

side of (12.9) is an inner product equivalent to that of $H^1_o(\Omega)$ and therefore $H = H^1_o(\Omega)$. But then Φ_f is clearly continuous since $H^{-1}(\Omega)$ is the dual of $H^1(\Omega) \supset H^1_o(\Omega)$ and $u \in H^1_o(\Omega)$. We have thus proved:

Theorem 12.1 (existence and uniqueness) *Given $f \in H^{-1}(\Omega)$ there exists a unique $u \in H^1_o(\Omega)$ satisfying the weak Dirichlet problem.*

Corolary 12.1 *The map $(\Delta, \tau) : H^1(\Omega) \to H^{-1}(\Omega) \times H^{1/2}(\partial\Omega)$ given by,*

$$\begin{aligned} \Delta u &= f \in H^{-1}(\Omega), \\ \tau u &= \phi_o \in H^{1/2}(\partial\Omega), \end{aligned} \tag{12.11}$$

is an isomorphism [1].

Proof: It is clear that the map is continuous –since it is linear and bounded–. It is also injective since if $\phi_o = 0$ then $u \in H^1_o(\Omega)$ and if $f = 0$ the previous theorem (uniqueness) tells us that $u = 0$. Let's see that it is surjective. Let $\phi_o \in H^{1/2}(\Omega)$, then there exists $w \in H^1(\Omega)$ such that its restriction to $\partial\Omega$, $\tau w = \phi_o$ and let $f \in H^{-1}(\Omega)$. Since also $\Delta w \in H^{-1}(\Omega)$ the previous theorem guarantees that there exists $\bar{u} \in H^1_o(\Omega)$ such that

$$\Delta \bar{u} = f - \Delta w \text{ in } \Omega. \tag{12.12}$$

But then $u = \bar{u} + w$ satisfies $\Delta u = f$ in Ω and $\tau u = \phi_o$ on $\partial\Omega$.

It only remains to prove then that the previous inner product is indeed an inner product and the norm thus obtained is equivalent to that of H^1_o. This follows from the following result,

Lemma 12.1 (of PoincarÃ©-Hardy) *There exists $C > 0$ such that for all $u \in H^1_o(\Omega)$,*

$$\int_\Omega |u|^2 \, d^n x \leq C \int_\Omega e^{ab} \nabla_a \bar{u} \nabla_b u \, d^n x. \tag{12.13}$$

This tells us that the inner product in $H^1_o(\Omega)$ is equivalent to the inner product used previously.

Proof: Since $C^\infty_o(\Omega)$ is dense in $H^1_o(\Omega)$ it is sufficient to prove the inequality for these functions. Let Γ_d be an n-cube of side d containing Ω. Extending the functions in $C^\infty_o(\Omega)$ as zero in $\Gamma_d - \Omega$ we obtain,

$$\begin{aligned} |u|^2(x) &= |\textstyle\int_{-d/2}^{x^1} \partial_1 u(\xi^1, x^2, ..., x^n) \, d\xi^1|^2 \\ &\leq (x^1 + \tfrac{d}{2}) \textstyle\int_{-d/2}^{x^1} |\partial_1 u|^2 \, d\xi^1 \\ &\leq d \textstyle\int_{-d/2}^{d/2} |\partial_1 u|^2 \, d\xi^1. \end{aligned} \tag{12.14}$$

[1] Always understanding these equations in their *weak* or distributional form

Therefore,

$$\int_{-d/2}^{d/2} |u(x)|^2 dx^1 \leq d^2 \int_{-d/2}^{d/2} |\partial_1 u|^2 \, d\xi^1, \tag{12.15}$$

and

$$\int_{\Gamma_d} |u|^2(x) \, d^n x \leq d^2 \int_{\Gamma_d} |\partial_1 u|^2 \, d^n x \leq C \int_{\Gamma_d} \nabla_a \bar{u} \nabla^a u \, d^n x, \tag{12.16}$$

with $C = d^2$. This proves the lemma and concludes the proof of existence and uniqueness ♠♠

Both the existence and uniqueness theorem and the regularity theorem (which we give below) can be generalized to elliptic equations with non-constant coefficients as long as they are smooth, it is satisfied that $c \leq 0$ and that $a^{ab} l_a l_b > k g^{ab} l_a l_b$, where k is a constant and g^{ab} is a positive definite metric such that the volume of Ω with respect to this metric is finite. The proof we gave is valid only if $Vol_g(\Omega) < \infty$, since we used the PoincarÃ©-Hardy lemma. If $\Omega = I\!R^n$ then the proof is still valid if we substitute the space H_0^1 by the space H_1^1, that is, the space of functions that decay to infinity, with norm

$$||f||^2_{H_1^1(R^n)} = \int_{R^n} \frac{|f|^2}{r^2} + e^{ab} \nabla_a \bar{f} \nabla_b f.$$

In this case, it can be proven that the solution to the problem will not only be smooth but also decay asymptotically like $1/r$.

12.1.2 *Regularity of Solutions

The solution we have found is only in the weak sense. If we now assume that f and ϕ_0 have some regularity, we can also conclude that,

Theorem 12.2 (Regularity) *Let $u \in H^1(\Omega)$ be a weak solution of the equation $\Delta u = f$ in Ω with boundary condition $u|_{\partial\Omega} = \phi_0$ and let $f \in H^k(\Omega)$, $\phi_0 \in H^{k+\frac{1}{2}}(\partial\Omega)$, $k \geq -1$ then $u \in H^{k+2}(\Omega)$. In particular, if $f \in C^\infty(\Omega)$ and $\phi_0 \in C^\infty(\partial\Omega)$, then $u \in C^\infty(\Omega)$.*

Before proceeding with the proof, we will define the finite difference operator and see its properties. Let $\{x^i\}$ be a coordinate system in an open subset of Ω.

Definition:

$$\Delta_i^h u(x^1, ..., x^i, ..., x^n) := \frac{u(x^1, ..., x^i + h, ...x^n) - u(x^1, ..., x^i, ..., x^n)}{h}.$$

Lemma 12.2 *If $u \in H^1(\Omega)$ and $\Omega' \subset\subset \Omega$ (strictly contained, $\partial\Omega \cap \Omega' = \emptyset$) then,*

$$||\Delta_i^h u||_{H^0(\partial\Omega)} \leq ||\partial_i u||_{H^0(\Omega)}. \tag{12.17}$$

Proof: It is sufficient to consider the case $u \in C^1(\Omega)$. If $h < dist(\partial\Omega, \Omega')$, then

$$\Delta_i^h u(x) = \frac{1}{h}\int_0^h \partial_i u(x^1, ..., x^i + \xi, ..., x^n) d\xi \quad \forall x \in \Omega', \tag{12.18}$$

therefore, using the Schwarz inequality, we see that,

$$|\Delta_i^h u(x)|^2 \leq \frac{1}{h}\int_0^h |\partial_i u(x^1, ..., x^i + \xi, ..., x^n)|^2 \, d\xi. \tag{12.19}$$

Integrating over Ω' we obtain,

$$\int_{\Omega'} |\Delta_i^h u(x)|^2 d^n x \leq \frac{1}{h}\int_0^h \int_{\Omega'} |\partial_i u|^2 d^n x \, d\xi \leq \int_\Omega |\partial_i u|^2 \, d^n x. \tag{12.20}$$

♠

Lemma 12.3 *Let $u \in H_0(\Omega)$ (and therefore $\Delta_i^h u \in H_0(\Omega')$) and suppose there exists $k < 0$ such that $||\Delta_i^h u||_{H_0(\Omega')} \leq k \;\; \forall \; h > 0$ and $\Omega' \subset\subset \Omega$, with $h < dist(\partial\Omega, \Omega')$. Then the weak derivative $\partial_i u$ exists and satisfies, $||\partial_i u||_{H_0(\Omega)} \leq k$.*

Proof: By the weak compactness of closed and bounded sets in $H_0(\Omega')$, there will exist a sequence $\{h_m\} \to 0$ and $v_i \in H_0(\Omega)$ with $||v_i||_{H_0(\Omega')} \leq k$ such that $\Delta_i^{h_m} u \xrightarrow{d} v_i$, that is

$$\int_\Omega \phi \Delta_i^{h_m} u \, d^n x \to \int_\Omega \phi v_i \, d^n x, \quad \forall \, \phi \in C_0^1(\Omega). \tag{12.21}$$

On the other hand, if $h_m < dist(sop.\phi, \partial\Omega)$, then

$$\int_\Omega \phi \Delta_i^{h_m} u \, d^n x = -\int_\Omega u(\Delta_i^{-h_m}\phi) \, d^n x \to -\int_\Omega u \partial_i \phi \, d^n x. \tag{12.22}$$

We thus conclude that,

$$\int_\Omega \phi v_i \, d^n x = -\int_\Omega u \partial_i \phi \, d^n x, \tag{12.23}$$

meaning that v_i is the weak (distributional) derivative of u in the direction x^i ♠

Proof: (Regularity Theorem): We will be content to prove that u is regular in any Ω' strictly contained in Ω, the extension of the proof to the whole Ω does not add any new concept. We will prove the statement for $k = 0$, the rest follows by induction on k. Since u is a weak solution, we have that

$$\int_\Omega \nabla^a \bar{u} \nabla_a v \, d^n x = \int_\Omega \bar{f} v \, d^n x, \forall v \in C_0^1(\Omega). \tag{12.24}$$

Replacing v by $-\Delta_i^{-h} v$ with $|2h| < dist(sop.v, \partial\Omega)$, we have,

$$\int_\Omega (\Delta_i^h \nabla^a \bar{u}) \nabla_a v \, d^n x = -\int_\Omega \bar{f} \Delta_i^h v \, d^n x, \forall v \in C_o^1(\Omega) \tag{12.25}$$

and therefore (Using Lemma (12.2)),

$$|| \int_\Omega (\Delta_i^h \nabla^a \bar{u}) \nabla_a v \, d^n x || \leq ||f||_{H_o} ||\partial_i v||_{H_o}. \tag{12.26}$$

Now taking $v = \Delta_i^h u$ we obtain,

$$|| \int_\Omega (\Delta_i^h \nabla^a \bar{u}) \Delta_i^h \nabla_a u \, d^n x || \leq ||f||_{H_o} ||\partial_i \Delta_i^h u||_{H_o}, \tag{12.27}$$

that is,

$$||\Delta_i^h \nabla^a u||_{H_o}^2 \leq ||f||_{H_o} ||\Delta_i^h \nabla_a u||_{H_o}, \tag{12.28}$$

which finally implies,

$$||\Delta_i^h \nabla^a u||_{H_o} \leq ||f||_{H_o} < k. \tag{12.29}$$

Lemma (12.3) then tells us that $\partial_i u \in H^1(\partial\Omega)$ and completes the first part of the proof.

Now, to complete the proof, let's see that if $f \in C^\infty(\Omega)$ then $u \in C^\infty(\Omega)$. But this is obvious since if $u \in H^p(\Omega)$, $\Omega \subset \mathbb{R}^n$, then by Sobolev's Lemma $u \in C^{p-\frac{n}{2}-\varepsilon}(\Omega) \quad \forall\, \varepsilon > 0$ and in particular if $u \in H^p(\Omega) \quad \forall\, p$ then $u \in C^\infty(\Omega)$ ♠

12.2 Spectral Theorem

Consider a rod made of a material with thermal conductivity $q = 1$ and length $L = 2\pi$ and suppose we want to describe the evolution of its temperature distribution $T(t,x)$. To do this, we will assume an initial distribution $T_o(x)$ and that the ends of the rod are connected to an infinite heat reservoir at zero degrees. We must then solve the mathematical problem,

$$\begin{aligned} \frac{d}{dt} T &= -\frac{d^2}{dx^2} T, \quad t \geq 0 \\ T(t_o, x) &= T_o(x), \\ T(t, 0) &= T(t, 2\pi) = 0 \quad t \geq 0. \end{aligned} \tag{12.30}$$

To solve it, we will use a Fourier series expansion, that is, we will propose a solution of the form,

$$T(t,x) = \sum_{n=1}^{\infty} C_n(t) \sin(\frac{nx}{2}), \tag{12.31}$$

which clearly satisfies the boundary conditions. Applying the equation and using the orthogonality of the functions we obtain,

$$\frac{d}{dt}C_n(t) = -\frac{n^2}{4}C_n(t), \tag{12.32}$$

which has the solution,

$$C_n(t) = C_n(0)e^{-\frac{n^2}{4}t}. \tag{12.33}$$

Therefore, if we give the initial condition $T_0 \in L^2(0, 2\pi)$, it will determine a sequence $\{C_n^0 = (T_0, \sin(\frac{nx}{2}))_{L^2}\}$ in l^2 that we will use as the initial condition and thus obtain $T(t,x)$. Since

$$\sum_{n=1}^{\infty}|C_n(t)|^2 = \sum_{n=1}^{\infty}|C_n^0|^2 e^{-\frac{n^2}{2}t} \leq \sum_{n=1}^{\infty}|C_n^0|^2 < \infty \tag{12.34}$$

the formal solution –since we do not know if it is differentiable– will be in $L^2(0, 2\pi)$ for all $t \geq 0$. For any $t > 0$ and $q \in \mathbb{N}$ we have that $n^{2q}e^{-\frac{n^2}{2}t}$ tends exponentially to zero as n tends to infinity and therefore $T(t,x) \in H^q(0, 2\pi)$ which implies $T(t,x) \in C^\infty((0,+\infty)\times[0,2\pi])$ [2].
We have thus completely solved this problem.

What have we used to construct these solutions? We have used that any function that vanishes at the ends can be expanded in terms of its Fourier series, that is, in terms of the functions $\sin(\frac{nx}{2})$ and also that,

$$\frac{d^2}{dx^2}\sin(\frac{nx}{2}) = -\frac{n^2}{4}\sin(\frac{nx}{2}) \tag{12.35}$$

In analogy with the theory of operators between finite-dimensional vector spaces, we will call the above equation the eigenvector-eigenvalue equation of the linear operator $\frac{d^2}{dx^2}$. Given a linear operator $L : D \subset L^2 \to L^2$ the problem of finding its eigenvalues and eigenvectors with given boundary conditions is called the Sturm-Liouville problem. As we will see below, there is a large number of operators for which this problem has a solution. Is it a coincidence that the set of eigenvalues of the operator in the example is a basis of L^2? [3] The following theorem and its corollary tell us that it is not, if the operator in question satisfies certain conditions.

Theorem 12.3 (Spectral for the Laplacian) *Let $\Omega \subset \mathbb{R}^n$ be bounded. Then the Laplacian has a countable and discrete set of eigenvalues, $\Sigma = \{\lambda_i\}$, whose eigenfunctions (eigenvectors) expand $H_0^1(\Omega)$.*

[2] It can also be shown that $T(t,x)$ is analytic in both variables in $(0,+\infty)\times[0,2\pi]$)

[3] To obtain pointwise convergence (and not just in L^2) in the case of non-zero boundary conditions, one must add the eigenvectors (with zero eigenvalue) $f_0(x) = 1$ and $f_1(x) = x$.

Proof: [Spectral Theorem] Consider the functional $\mathcal{J} : H^1_o(\Omega) \to \mathbb{R}^+$ given by,

$$\mathcal{J}(u) = \frac{\int_\Omega \nabla_a \bar{u} \nabla^a u \; d^n x}{\int_\Omega \bar{u} u \; d^n x}. \tag{12.36}$$

Since $\mathcal{J}(u)$ is non-negative, there will exist $\lambda_o(\Omega) \in \mathbb{R}^+$ such that

$$\lambda_o = \inf_{u \in H^1_o(\Omega)} \mathcal{J}(u) \; . \tag{12.37}$$

Exercise: Relate this λ_o with the constant in the PoincarÃ©-Hardy Lemma.

As we will see below, this λ_o is the smallest of the eigenvalues of the Laplacian, and its eigenfunction u_o is the one that minimizes $\mathcal{J}$. Indeed, suppose there exists u_o in $H^1_o(\Omega)$ that minimizes $\mathcal{J}$. Since $\mathcal{J}$ is a differentiable functional, we must have

$$\frac{d}{ds} \mathcal{J}(u_o + sv)|_{s=o} = o \quad \forall \, v \in H^1_o(\Omega). \tag{12.38}$$

But

$$\begin{aligned} \frac{d}{ds} \mathcal{J}(u_o + sv)|_{s=o} &= \frac{1}{\int_\Omega \bar{u} u \; d^n x} [\int_\Omega (\nabla^a \bar{v} \nabla_a u_o + \nabla^a \bar{u}_o \nabla_a v) d^n x \\ &- \lambda_o [\int_\Omega (\bar{v} u_o + \bar{u}_o v) d^n x]] \end{aligned} \tag{12.39}$$

and therefore (12.38) is equivalent to u_o satisfying the weak version of the eigenvalue-eigenvector equation. Now let's see that u_o exists. Since λ_o is the infimum of $\mathcal{J}$, there exists a sequence $\{u_p\} \in H^1_o(\Omega)$ such that $||u_p||_{H^o(\Omega)} = 1$ and $\mathcal{J}(u_p) \to \lambda_o$. But such a sequence is bounded in $H^1_o(\Omega)$ and therefore by lemma (10.10) there exists a subsequence $\{\tilde{u}_p\}$ that converges strongly in $H^o(\Omega)$ to a function u_o with $||u_o||_{H^o(\Omega)} = 1$. Since $Q(u) := \int_\Omega \nabla^a \bar{u} \nabla_a u \; d^n x$ is a norm derived from an inner product, the parallelogram law holds,

$$Q(\frac{\tilde{u}_p - \tilde{u}_q}{2}) + Q(\frac{\tilde{u}_p + \tilde{u}_q}{2}) = \frac{1}{2}(Q(\tilde{u}_p) + Q(\tilde{u}_q)), \tag{12.40}$$

which implies

$$Q(\frac{\tilde{u}_p - \tilde{u}_q}{2}) \leq \frac{1}{2}(Q(\tilde{u}_p) + Q(\tilde{u}_q)) - \lambda_o ||\frac{\tilde{u}_p + \tilde{u}_q}{2}||^2_{H^o(\Omega)} \to \lambda_o - \lambda_o = o. \tag{12.41}$$

Where we have used that since $\{\tilde{u}_q\} \to u_o$ in $H^o(\Omega)$, $\{\frac{\tilde{u}_p + \tilde{u}_q}{2}\}$ also does, and therefore $\{||\frac{\tilde{u}_p + \tilde{u}_q}{2}||_{H^o(\Omega)}\} \to 1$. But since the norm $Q(u)$ is equivalent to that of $H^1_o(\Omega)$, we see

that $\{\tilde{u}_p\}$ is also Cauchy in $H^1_o(\Omega)$ and therefore $u_o \in H^1_o(\Omega)$. Using the regularity theorem with $f = \lambda_o u_o$, we see that $u_o \in C^\infty(\Omega) \cap H^1_o(\Omega)$ and therefore u_o is an eigenvalue in the classical sense ($\Delta u_o + \lambda_o u_o = o$). Now let's prove the existence of the other eigenvalues-eigenvectors. Let $H(1) = \{u \in H^1_o(\Omega) | \langle u, u_o \rangle_{H^o(\Omega)} = o\}$. This is a vector subspace and is closed, therefore it is a Hilbert space [4]. This space is invariant with respect to the action of the Laplacian. Indeed, if $v \in H(1)$, then,

$$\begin{aligned}
\langle u_o, \Delta v \rangle &= \int_\Omega \bar{u}_o \Delta v \, d^n x \\
&= \int_\Omega \Delta \bar{u}_o v \, d^n x + \int_{\partial\Omega} [\bar{u}_o n^a \nabla_a v + v n^a \nabla_a \bar{u}_o] \, d^{n-1} x \\
&= \int_\Omega \Delta \bar{u}_o v \, d^n x \\
&= \lambda_o \int_\Omega \bar{u}_o v \, d^n x \\
&= \lambda_o \langle u_o, v \rangle, \\
&= o
\end{aligned}$$

which tells us that $\Delta v \in H(1)$. Repeating the previous proof, we conclude that therefore there will exist λ_1, such that

$$\lambda_1 = \inf_{u \in H(1)} \mathscr{J}(u). \tag{12.42}$$

and that λ_1, will be an eigenvalue, that is, there will exist an eigenfunction u_1, with eigenvalue λ_1.

Defining $H(2) = \{u \in H^1_o(\Omega) | \langle u, u_o \rangle_{H^o(\Omega)} = \langle u, u_1 \rangle_{H^o(\Omega)} = o\}$, etc. we can continue indefinitely and obtain Σ and, correspondingly, an orthonormal set [5], in $H^o(\Omega)$, of eigenvectors. To complete the proof, let's see that this set expands $H^1_o(\Omega)$. Since u_i satisfies $\Delta u_i + \lambda_i u_i = o$ and $\langle u_i, u_j \rangle_{H^o(\Omega)} = o$ if $i \neq j$ we have that

$$o = \langle u_j, \Delta u_i + \lambda_i u_i \rangle_{H^o(\Omega)} = -\int_\Omega \nabla^a \bar{u}_j \nabla_a u_i d^n x \tag{12.43}$$

and therefore also that

$$\langle u_j, u_i \rangle_{H^1_o(\Omega)} = o. \tag{12.44}$$

We then see that this set is orthogonal in $H^1_o(\Omega)$ and that by construction its perpendicular subspace is $\{o\}$, which implies that this set is a basis of $H^1_o(\Omega)$ ♠

[4] Note that $H(1) \neq$ the perpendicular space to u_o in the $H^1_o(\Omega)$ norm.

[5] After normalizing them appropriately.

Exercise: Find the eigenvalues and eigenvectors of the Laplacian in $H^1_o(\Omega)$ when: a) $\Omega \subset \mathbb{R}^2$ is a square with side L. b) $\Omega \subset \mathbb{R}^3$ is a sphere with radius R. Construct in both cases using these eigenfunctions the Green's function of the problem in question.

For which equations can the previous theorem be generalized? For the proof, specific properties of the Laplacian were used only to assert that $\mathcal{J}$ was bounded below –to conclude that the infimum existed– and that $Q(u)$ was a norm derived from an inner product –to conclude that the parallelogram law held–. If

$$L(u) = a^{ab}\nabla_a \nabla_b u + b^a \nabla_b u + cu, \tag{12.45}$$

with a^{ab}, b^a and c smooth (real) fields in Ω, and such that there exists $k > 0$ such that $a^{ab} l_a l_b \geq k\ g^{ab} l_a l_b$, for every vector field l_a in Ω (ellipticity) then there exist positive constants c_1 and c_2 such that

$$\langle u, -L(u)\rangle_{H^0(\Omega)} = -\int_\Omega \bar{u} L(u) d^n x \leq c_1 \int_\Omega g^{ab}\nabla_a \bar{u} \nabla_b u d^n x - c_2 \int_\Omega |u|^2 d^n x. \tag{12.46}$$

Exercise: Prove this.

Therefore in this case we also have that

$$\mathcal{J}(u) := \frac{\langle u, -L(u)\rangle_{H^0(\Omega)}}{\langle u, u\rangle_{H^0(\Omega)}}, \tag{12.47}$$

is bounded below. The condition that $Q(u) := \langle u, -L(u)\rangle_{H^0(\Omega)}$ satisfies the *parallelogram law*, even if it is not positive definite, is much more restrictive, and is equivalent to requiring that L satisfies

$$\langle v, L(u)\rangle_{H^0(\Omega)} = \langle L(v), u\rangle_{H^0(\Omega)} \quad \forall\ u, v \in H^1_o(\Omega). \tag{12.48}$$

Operators that satisfy this relation are called **self-adjoint** or **Hermitian**. Note that this condition also ensures that the operator L leaves invariant the respective subspaces $H(i)$ that need to be considered in the previous proof.

Exercise: Show that

$$L(u) = \nabla_a(a^{ab}\nabla_b u + b^a u) - b^a \nabla_a u + cu, \tag{12.49}$$

with a^{ab}, b^a and c real tensor fields is self-adjoint.

Exercise: Show that if L is self-adjoint then the eigenvectors are real.

Exercise: Find the eigenvalues and eigenvectors in $H^1_o(\Omega)$ of

$$L(u) = \frac{d^2}{dx^2} u + cx^2 u. \tag{12.50}$$

We thus arrive at the following generalization:

Theorem 12.4 (Spectral) *Let Ω be bounded and L an elliptic self-adjoint operator with coefficients a^{ab}, b^a and c in $C^\infty(\Omega)$ [A condition that can be considerably weakened]. Then the eigenvalue problem $L(u_i) = \lambda_i u_i$, $u_i \in H_o^1(\Omega)$, has a countable and discrete set of real eigenvalues, whose eigenfunctions $u_i \in C^\infty(\Omega)$ expand $H_o^1(\Omega)$.*

Exercise:
a) Prove the following **corollary:**

If L is elliptic and self-adjoint such that its eigenvalues are different from zero, then the Dirichlet problem

$$L(u) = f, \tag{12.51}$$

$f \in H^o(\Omega)$, $u \in H^1(\Omega)$, has a unique solution.
b) If some $\lambda_i = 0$ then the previous problem has a solution iff $\langle u_i, f \rangle_{H^o(\Omega)} = 0$ for every eigenfunction with zero eigenvalue.

Chapter 13

Symmetric–Hyperbolic Equations

13.1 Introduction

In this chapter we will study systems of hyperbolic equations under the following restriction:

Definition: We will say that a system is **symmetric–hyperbolic** if:

a.) The target space of the linear map $M^a_{A'B}k_a$ is of the same dimension as the domain space for all $k_a \neq 0$. Therefore, from now on we will use unprimed indices.

b.) The map $M^a_{AB}k_a$ is symmetric for all $k_a \neq 0$.

c.) At each point of the manifold there exists a co-vector t_a such that the map, $H_{AB} := M^a_{AB}t_a$ is positive definite. (That is, $H_{AB}u^A u^B \geq 0$ $(= 0$ iff $u^A = 0)$.) [1]

Note that this last condition implies that H_{AB} is a metric in the space of independent variables. This and its inverse, which we will denote by H^{AB}, will be used to raise and lower indices.

This is also not a significant restriction from the point of view of physics, since all physical systems we know of are symmetric–hyperbolic.

Also, but only for simplicity in the exposition as this will avoid some technical complications, we will only consider linear systems.

We will begin this chapter with a simple example that illustrates the basic characteristics of this class of equations.

13.2 An Example

Consider an infinite string in the x, y plane and let $y = u(x,t)$ be the position of the string at time t in that plane. By adjusting the dimensions (of length or time) it can

[1] Since the set of symmetric and positive maps is open, if x_a is any other co-vector, then $H_{AB}(\varepsilon) := M^a_{AB}(t_a + \varepsilon x_a)$ will also be positive for ε sufficiently close to zero. Therefore, level surfaces $\tau = const.$ can be locally defined such that $H_{AB} := M^a_{AB}\nabla_a \tau$ is positive on the entire local surface.

be seen that $u(x,t)$ satisfies the equation,

$$-\frac{\partial^2 u}{\partial t^2}+\frac{\partial^2 u}{\partial x^2}=f(x,t), \tag{13.1}$$

where $f(x,t)$ is the force density exerted on the string and which we assume does not depend on the position of the string with respect to the y coordinate.[2] We thus have the mathematical problem of finding the solutions to the equation,

$$\Box u = g^{ab}\nabla_a\nabla_b u = f, \tag{13.2}$$

in $M=\mathbb{R}^2$ with pseudo-Euclidean metric $g_{ab}=-(dt)^2+(dx)^2$. To handle this equation it is convenient to introduce a coordinate system appropriate to this metric, that is, one that has its characteristic lines as axes,

$$\begin{array}{rcl}\xi &=& x+t=const.,\\ \eta &=& x-t=const.\end{array} \tag{13.3}$$

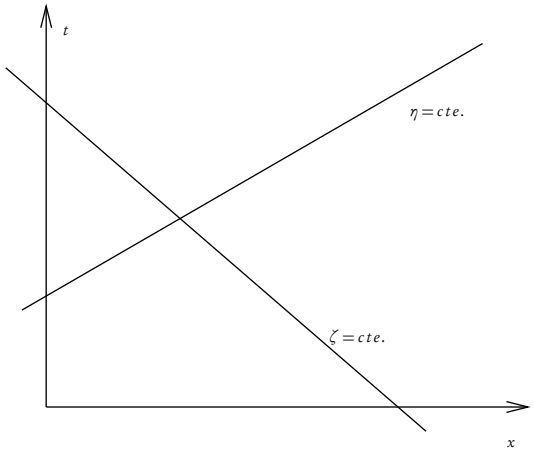

Figure 13.1: Null coordinate system.

we then have that

$$\begin{array}{rcl}x &=& \frac{\xi+\eta}{2}\\ t &=& \frac{\xi-\eta}{2}\end{array} \tag{13.4}$$

and

$$\begin{array}{rcl}g_{ab}=-(dt)^2+(dx)^2 &=& \frac{1}{4}[\{-(d\xi)^2-(d\eta)^2+d\xi\otimes d\eta+d\eta\otimes d\xi\}\\ && +\{(d\xi)^2+(d\eta)^2+d\xi\otimes d\eta+d\eta\otimes d\xi\}]\\ &=& \frac{1}{2}[d\xi\otimes d\eta+d\eta\otimes d\xi]\end{array} \tag{13.5}$$

[2] Otherwise we would have to consider $f(x,t,u)$ which complicates the problem.

Noting that $g^{ab} = 2[\partial\xi \otimes \partial\eta + \partial\eta \otimes \partial\xi]$ and that the Christoffel symbols vanish because the metric has constant components we have,[3]

$$\Box u = 4\frac{\partial^2 u}{\partial\eta\partial\xi} = f(\eta,\xi). \tag{13.6}$$

This equation can be integrated immediately, obtaining

$$\begin{aligned} \frac{\partial u}{\partial\eta}(\eta,\xi) &= \tfrac{1}{4}\int_{\xi_0}^{\xi} f(\eta,\tilde{\xi})\,d\tilde{\xi} + C(\eta) \\ u(\eta,\xi) &= \tfrac{1}{4}\int_{\eta_0}^{\eta}\int_{\xi_0}^{\xi} f(\tilde{\eta},\tilde{\xi})d\tilde{\xi}\,d\tilde{\eta} + u_I(\xi) + u_{II}(\eta), \end{aligned} \tag{13.7}$$

where $u_I(\xi)$ and $u_{II}(\eta)$ are arbitrary functions. Let us first consider the case $f \equiv 0$, that is, the homogeneous equation. Its solutions are then the sum of any function of ξ and any function of η. Returning to the x, t coordinates we obtain

$$u(x,t) = u_I(x+t) + u_{II}(x-t). \tag{13.8}$$

For example,

$$u_I(x+t) = \begin{cases} e^{\frac{1}{(x+t)^2-1}} & x+t \in [-1,1] \\ 0 & x+t \in (-\infty,-1]\cup[1,+\infty) \end{cases} \tag{13.9}$$

is a solution that represents a **wave** ($\equiv$ solution of the homogeneous equation) moving to the left without changing shape and with speed 1.

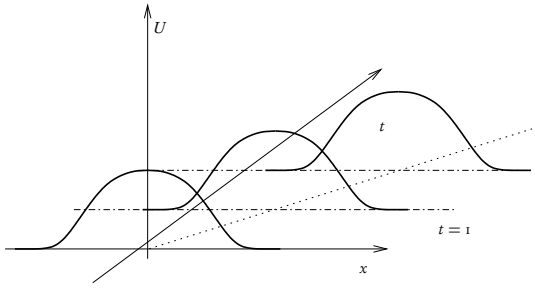

Figure 13.2: Wave propagation.

Similarly u_{II} represents a wave moving to the right. See figure. Let us now see that the Cauchy problem in this case has a solution. This is extremely important in physics: it tells us that if we take a non-characteristic surface (for example $t = 0$) and

[3] Note also that $g(\partial_\eta, \partial_\eta) = g(\partial\xi, \partial\xi) = 0$, that is, these coordinate vectors have null norm.

give there as data u and its time derivative we will obtain a unique solution for future times. This is what allows us, if we know the present, to predict the future, that is, if we prepare an experiment, to predict the result. This fact is what distinguishes physics from other natural sciences. Suppose then that at $t = 0$ ($\xi = \eta = x$) we give $u(x,0) = u_0(x)$ and its derivative $\frac{\partial u}{\partial t}(x,0) = u_1(x)$. We then have that

$$u_0(x) = u(x,0) = u_I(x) + u_{II}(x), \tag{13.10}$$

$$u_1(x) = \frac{\partial u}{\partial t}(x,0) = u_I'(x) - u_{II}'(x). \tag{13.11}$$

Differentiating (13.10) with respect to x and solving the linear system thus obtained we have,

$$\begin{aligned} u_I'(x) &= \frac{u_0'(x) + u_1(x)}{2} \\ u_{II}' &= \frac{u_0'(x) - u_1(x)}{2}, \end{aligned} \tag{13.12}$$

and integrating,

$$\begin{aligned} u_I(x) &= \frac{u_0(x)}{2} + \frac{1}{2}\int_0^x u_1(\tilde{x})d\tilde{x} + C_I, \\ u_{II}(x) &= \frac{u_0(x)}{2} - \frac{1}{2}\int_0^x u_1(\tilde{x})d\tilde{x} + C_{II}. \end{aligned} \tag{13.13}$$

For (13.10) to be satisfied we must have $C_I = -C_{II}$ and therefore

$$u(x,t) = \frac{2}{(}u_0(x+t) + u_0(x-t)) + \frac{2}{\int}_{x-t}^{x+t} u_1(\tilde{x})d\tilde{x}. \tag{13.14}$$

We then see that if we give as data $u_0(x) \in C^2(\mathbb{R})$ and $u_1(x) \in C^1(\mathbb{R})$ we obtain a solution $u(x,t) \in C^2(\mathbb{R}\times\mathbb{R})$. By construction this solution is unique.

Exercise: Explicitly show that (13.14) satisfies (13.1) with $f \equiv 0$.

Exercise: Use the general solution found earlier to see that the homogeneous solutions of the wave equation in two dimensions satisfy $u(t,x) = u(0,x+t) + u(0,x-t) - u(-t,x)$.

Equation (13.14) tells us that $u(x,t)$ is contributed *only* by the average of the values of u_0 at $x-t$ and $x+t$ and the integral of u_1, between these two values. [See figure 13.3.]

What happens if we have a source $f(x,t)$? Since we already have the general solution (for arbitrary Cauchy data) of the homogeneous equation, we only need

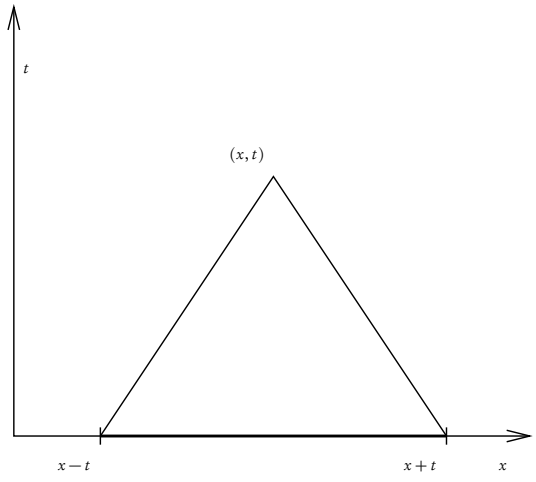

Figure 13.3: General homogeneous solution in 1+1 dimensions.

the solution of the inhomogeneous equation with zero data. This is achieved by integrating $f(\xi,\eta)$ first with respect to $\tilde{\xi}$ between $\tilde{\eta}$ and ξ and then $\tilde{\eta}$ between ξ and η.

$$v(\eta,\xi)=\frac{1}{4}\int_{\xi}^{\eta}\left(\int_{\tilde{\eta}}^{\xi} f(\tilde{\xi},\tilde{\eta})d\tilde{\xi}\right)d\tilde{\eta}. \tag{13.15}$$

Exercise: Show that

$$v(x,t)=\int_{0}^{t}\left(\int_{x-(t-\tilde{t})}^{x+(t-\tilde{t})} f(\tilde{x},\tilde{t})d\tilde{x}\right)d\tilde{t}, \tag{13.16}$$

that

$$v(x,t)|_{t=0}=\frac{\partial v}{\partial t}(x,t)|_{t=0}=0, \tag{13.17}$$

and that $v(x,t)$ satisfies (13.1).

We then see that the solution sought is,

$$u(x,t)=\frac{2}{(}u_{0}(x+t)+u_{0}(x-t))+\frac{2}{\int}_{x-t}^{x+t}u_{1}(\tilde{x})d\tilde{x}+\int_{0}^{t}\left(\int_{x-(t-\tilde{t})}^{x+(t-\tilde{t})} f(\tilde{x},\tilde{t})d\tilde{x}\right)d\tilde{t}, \tag{13.18}$$

which by construction is unique and that $u(x_0,t_0)$ depends on the initial values and on f in the conical region with vertex (x_0,t_0) given by,

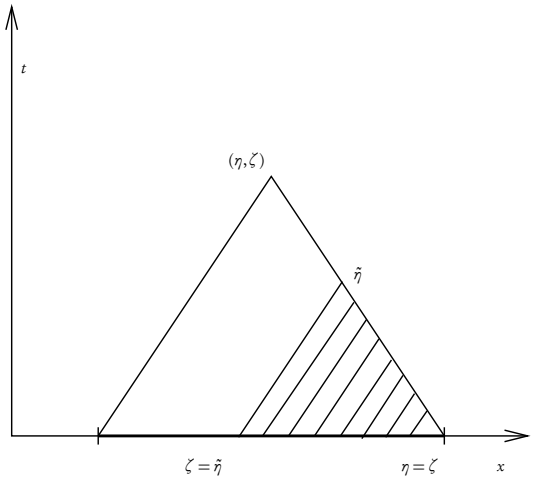

Figure 13.4: General inhomogeneous solution.

$$\begin{cases} t \leq t_o \\ |x - x_o| \leq t_o - t. \end{cases} \tag{13.19}$$

This region is called the **domain of dependence** of the point (x_o, t_o), only what happens in this region can affect the value of u at that point. Similarly, the **domain of influence** of a point (x_o, t_o) is defined as the set of points where the value of u can be changed if the value of u, its derivative, or f is changed at (x_o, t_o). In this case, this is given by: $\{(x,t)|t \geq t_o, |x - x_o| \leq t - t_o\}$.

The behavior of the solutions of hyperbolic equations is generally the same as that of this simple example: Given generic Cauchy data there will be a unique solution (both to the future and to the past). In the case of linear equations this solution can be extended indefinitely in both temporal directions, in the nonlinear case the solutions are only valid in a finite temporal interval and it is an interesting physical problem to see if the physical nonlinear equations can or cannot be extended indefinitely and what it means physically the appearance of singularities in the solutions. [4] A quantity of great physical and mathematical importance related to the wave equation is the energy of the solutions. In two dimensions this is given by,

$$E(u, t_o) = \frac{2}{\int_{t=t_o}} [(\frac{\partial u}{\partial t})^2 + (\frac{\partial u}{\partial x})^2] dx. \tag{13.20}$$

[4] A singularity is a point where u ceases to be sufficiently differentiable for the equation to make sense or worse where u ceases to make sense even as a distribution.

Observe that the energy is positive and its rate of change is given by,

$$\begin{aligned}\frac{dE}{dt}(u,t_o) &= \int_{t=t_o}(\frac{\partial u}{\partial t}\frac{\partial^2 u}{\partial t^2}+\frac{\partial^2 u}{\partial t\partial x}\frac{\partial u}{\partial x})dx \\ &= \int_{t=t_o}[\frac{\partial u}{\partial t}(\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 u}{\partial x^2})+\frac{\partial}{\partial x}(\frac{\partial u}{\partial t}\frac{\partial u}{\partial x})]dx \\ &= \int_{t=t_o}[\frac{\partial u}{\partial t}f]dx,\end{aligned} \tag{13.21}$$

where in the last equality we have used (13.1) and assumed that $\lim_{x\to\infty}\frac{\partial u}{\partial t}\frac{\partial u}{\partial x}=0$. If $f=0$ then the energy is conserved and this gives us an alternative proof of the uniqueness of the solutions.

Theorem 13.1 (Uniqueness) *At most, there exists a unique solution $u(x,t)$ to the wave equation among the functions in $u(x,t)\in H^1(\mathbb{R})$, $\frac{\partial u}{\partial t}(x,t)\in H^0(\mathbb{R})$ (where $\mathbb{R}$ is the surface $t=const$) for a given Cauchy data.*

Proof: Suppose there are two solutions u_1 and u_2 with the same Cauchy data at, say $t=0$. Then $\delta u=u_1-u_2$ satisfies the homogeneous equation and has zero Cauchy data. Therefore $E(\delta u,t=0)=0$, but the energy of δu is conserved and thus $E(\delta u,t-t_o)=0;;;\forall t_o$. This implies that $|\frac{\partial\delta u}{\partial t}(x,t)|t=t_o|H^0(R)=0$ and therefore $\frac{\partial\delta u}{\partial t}(x,t)|t=t_o=0$ at almost every point. Similarly, we have that $|\frac{\partial\delta u}{\partial x}(x,t)|t=t_o|H^0(R)=0$ and therefore $\frac{\partial\delta u}{\partial x}(x,t)|t=t_o=0$ or $\delta u=const$. But constants are not square integrable in $\mathbb{R}$ and therefore $\delta u(x,t)=u_1(x,t)-u_2(x,t)=0$ as an element of $H^1(\mathbb{R})$ ♠

In this section, we will consider a general linear symmetric-hyperbolic system. That is, a system of the form:

$$M^a_{AB}\nabla_a u^B=I_A, \tag{13.22}$$

with M^a_{AB} and I_A generally dependent on position, with M^a_{AB} symmetric in the uppercase indices and such that there exists a function τ with gradient t_a such that H_{AB} is positive definite and therefore invertible.

Let Σ_t be the family of surfaces given by the level surfaces $\tau=t$. Let Γ be any region of Σ_o and let Ω be a region such that $\Omega\cap\Sigma_o=\Gamma$ and also $\Gamma(t)=\Omega\cap\Sigma_t$.

Let u^A be a solution of 13.22 and let,

$$E(u^A,t)=\int_{\Gamma(t)}n_aM^a_{AB}u^Au^B, \tag{13.23}$$

that is, the integral of $H_{AB}u^Au^B$ over a region of the hypersurface $\tau=t=const.$, where we have defined H_{AB} using $n_a=t_a/|t_a|$, that is, we have normalized t_a.

Let us see the difference of this quantity between two surfaces as shown in the figure.

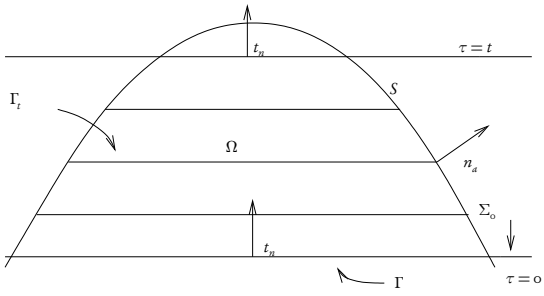

Figure 13.5: Energy inequality

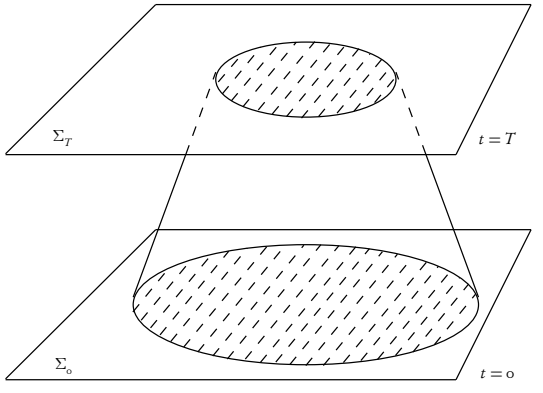

Figure 13.6: Energy inequality, perspective view.

To do this, we will use the divergence theorem, which tells us that,

$$E(u^A, t) - E(u^A, 0) + E(u^A, s) = \int_{\Omega(0,t)} \nabla_a (M^a_{AB} u^A u^B), \tag{13.24}$$

where the negative sign of the second term on the left is due to the fact that in the definition of $E(u^A, 0)$ we take the inward normal to the region $\Omega(0, t)$. The last term represents the integral of the energy over a surface S which we will assume is given by the equation $\sigma = s$ where σ is a smooth function and $s \in I\!R$. This term represents the energy that escapes from the region through the surface S.

Using equation 13.22 on the right-hand side we get,

$$\begin{aligned} E(u^A,t)-E(u^A,0)+E(u^A,s) &= \int_{\Omega(0,t)}[(\nabla_a M^a_{AB})u^A u^B + 2I_A u^A] \qquad (13.25)\\ &\leq \int_0^t\int_{\Sigma_{\tilde{t}}}\{|(\nabla_a M^a_{AB})u^A u^B| + 2|I_A u^A|\}d\tilde{t}\\ &\leq \int_0^t\int_{\Sigma_{\tilde{t}}}\{|CH_{AB}u^A u^B|\\ &+ 2\sqrt{|I_A I_B H^{AB}|}\sqrt{|H_{AB}u^A u^B|}\}d\tilde{t}\\ &\leq \int_0^t[(C+1)E(u^A,\tilde{t})+D(\tilde{t})]d\tilde{t}, \end{aligned}$$

where in the first member of the second inequality we have used the Schwartz inequality (see exercise) and have defined,

$$C^2 := \sup_{\Omega}(\nabla_a M^a_{AB})(\nabla_b M^b_{CD})H^{AC}H^{BD}. \qquad (13.26)$$

In the third inequality we have used that $2ab \leq a^2 + b^2$ and have defined,

$$D(t) := \int_{\Sigma_t} H^{AB}I_A I_B. \qquad (13.27)$$

Exercise: Let H_{AB} be symmetric and positive definite, with inverse H^{AB}. Prove:
a.) $S^{AB}S^{CD}H_{AC}H_{BD} \geq 0, (= 0; iff; S^{AB} = 0)$.
b.) $|S_{AB}u^A u^B| \leq \sqrt{H^{AC}H^{BD}S_{AB}S_{CD}}; H_{AB}u^A u^B$.

We will now make an important assumption that we will discuss in detail later:

we will assume from now on that $E(u^A,s) \geq 0; \forall; u^A$.

With this assumption, we can ignore this term in the previous inequality and obtain,

$$E(u^A,t)-E(u^A,0) \leq \int_0^t[(C+1)E(u^A,\tilde{t})+D(\tilde{t})]d\tilde{t}. \qquad (13.28)$$

Differentiating this integral inequality, we will obtain a maximum bound for the energy. Indeed, differentiating this expression and noting that the sign of the inequality is maintained, we obtain the following differential inequality,

$$\frac{d}{dt}E(u^A,t) \leq (1+C)E(u^A,t)+D(t), \qquad (13.29)$$

The differential equality has as a solution (using the method of variation of constants, section 5.2),

$$Y(t) = e^{(1+C)t)};[Y(0) + \int_0^t e^{-(1+C)\tilde{t}} D(\tilde{t}) d\tilde{t}]. \tag{13.30}$$

Using now lemma **??** we see that $E(u^A, t) \leq Y(t); \forall; t \geq 0$ if $E(u^A, 0) = Y(0)$, and therefore we have that,

$$E(u^A, t) \leq e^{(1+C)t)};[E(u^A, 0) + \int_0^t e^{-(1+C)\tilde{t}} D(\tilde{t}) d\tilde{t}]. \tag{13.31}$$

This inequality is extremely important, not only does it allow us to infer the uniqueness of solutions (as we will see below) but it also plays a fundamental role in proving the existence of solutions and achieving convergent and reliable numerical algorithms.

13.3 Uniqueness of solutions

Using the inequality obtained earlier, we will prove the following theorem:

Theorem 13.2 *Let there be a symmetric-hyperbolic equation on a manifold M. Let Σ_0 be a surface given by the equation $\tau = 0$ such that $M^a_{AB} \nabla_a \tau$ is positive definite. Let Γ be any region of Σ_0 and let Ω be a region such that $\Omega \cap \Sigma_0 = \Gamma$ and such that $E(u^A, s) \geq 0; \forall u^A$. [See previous figure.] Then if u^C and $\tilde{u}^C$ are two solutions that coincide in Γ, they coincide in all of Ω.*

Proof: Let $\delta^A := u^A - \tilde{u}^A$. Then δ^A satisfies,

$$M^a_{AB} \nabla_a \delta^A = 0. \tag{13.32}$$

Therefore, we have an energy inequality for δ^C with $D(t) \equiv 0$ and also with $E(\delta^C, 0) = 0$ since the two solutions coincide in Γ. But then the inequality tells us that $E(\delta^C, t) = 0$ for all t and therefore $\delta^C = 0$ in all of Ω thus proving the theorem ♠

13.4 Domain of dependence

The previous uniqueness theorem was based on the assumption that

$$E(u^C, s) = \int_S n_a M^a_{AB} u^A u^B \geq 0 \tag{13.33}$$

and therefore it is important to determine what are the possible regions where this happens. In particular, given a region Γ, the largest region Ω where we have uniqueness of solutions with identical initial data in Γ is called the **domain of dependence of** Γ, it is the region that depends completely on the initial data given in Γ, that is,

giving initial data in Γ we can completely control the value of the solution at any point in its domain of dependence.

First, let's see that this domain of dependence is not empty. To do this, take Γ compact in Σ_o and consider $H_{AB} = t_a M^a_{AB}$. Now let $\sigma = \tau - \delta\xi$ with ξ a smooth function in a neighborhood of Γ positive inside this set and negative in $\Sigma_o - \Gamma$ (that is, it vanishes on its boundary) and δ a real number that we will assume small. [See figure 13.7.]

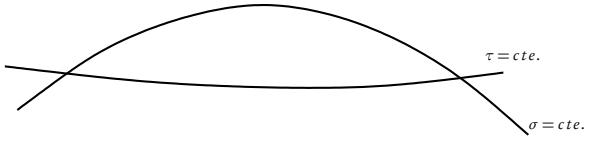

Figure 13.7: Bubble-shaped region

At each point $p \in \Gamma$ H_{AB} is a positive definite metric and therefore, as the set of positive definite metrics is open in the space of all symmetric tensors, given any other covector w_a there will exist $\varepsilon > 0$ small enough such that $(t_a + \varepsilon w_a)M^a_{AB} = H_{AB} + \varepsilon w_a M^a AB$ is also positive definite. Since Γ is compact given a w_a in it there will be a minimum and positive ε such that the tensor defined above will be positive [5] By the same continuity argument, there will be a compact region around Γ and a $\varepsilon > 0$, a little smaller than the previous one such that there $\nabla_a \sigma M^a_{AB}$ will be positive definite. We have thus managed to have a region Ω between the level surfaces $\tau = 0$ and $\sigma = 0$, that is, a small *bubble* where the integral of the outgoing energy through $S = p \in M | \sigma(p) = 0$ is positive for any solution u^A.

How large can we make this bubble, that is, how much more can we tilt the surface S and still maintain positivity? This question has a lot to do with the following: how much can we tilt t_a at each point and still have positivity of $t_a M^a_{AB}$ at that point?

First, note that if $t_a M^a_{AB}$ is positive then $\alpha t_a M^a_{AB}, \alpha > 0$, is also positive, so the set of covectors for which we have positivity forms a cone. This also ensures that not all covectors are in this cone, since if t_a is in it, $(-t_a)$ is not.

Second, note that if t_a and $\tilde{t}_a$ are in the cone, [that is, each of them gives a positive definite metric] then all covectors of the form, $\alpha t_a + (1-\alpha)\tilde{t}a, ; \alpha \in [0, 1]$ are also in it, since $(\alpha t_a M^a AB + (1-\alpha)\tilde{t}a M^a AB)u^A u^B$ is positive if the coefficients of α and $(1-\alpha)$ are. That is, the set of covectors that give positive definite metrics form a convex cone in T^*_p, at each point $p \in M$. This cone is called the **characteristic cone**. [See figure 13.8.]

What happens to the covectors on the boundary of this cone? There the condition of positive definiteness must fail, that is, given a covector t_a on this boundary

[5] To see this consider the map between $(B_1 \times \Gamma) \times (B_1 \times \Gamma) \times \Gamma \to \mathbb{R}$ given by, $w_a(p) M^a_{AB}(p) u^A(p) u^B(p)$ where $B_1(p) = u^A(p) | H_{AB}(p) u^A u^B = 1$. This is a continuous map and its domain is compact therefore it has a maximum, $m < \infty$. We can then take $0 < \varepsilon < 1/m$.

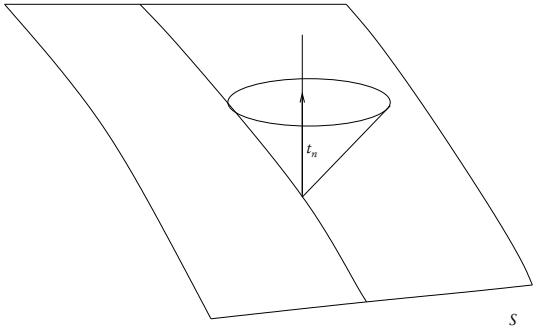

Figure 13.8: Characteristic cone

there will be some u^A such that $t_a M^a_{AB} u^A u^B = 0$. This implies that there the rank of $t_a M^a_{AB}$ ceases to be maximum.

13.4.1 Construction of a characteristic surface

Now we will construct the boundary surface of the maximum domain of dependence. From a region Γ given by $q \in \Sigma_0; |; \sigma_0(p) = 0$ we will construct a surface S such that its normal at each of its points belongs to the boundary of the characteristic cone. This surface will be given by a function $\sigma = 0$ such that $\sigma|\Sigma_0 = \sigma_0$. To find the equation that this function must satisfy, it is convenient to introduce an appropriate coordinate system, the same one we used in our classification of partial differential equations. One of these coordinates will be $t = \tau$ and we will call the others x^i, also for convenience we will define $\sigma_t = \frac{\partial \sigma}{\partial t}$ and $\sigma_i = \frac{\partial \sigma}{\partial x^i}$. With these coordinates we obtain that,

$$\begin{aligned} \nabla_a \sigma M^a AB &= \sigma_t M^t_{AB} + \sum_i \sigma_i M^i_{AB} \\ &= \sigma_t H_{AB} + \sum_i \sigma_i M^i_{AB}. \end{aligned} \tag{13.34}$$

Multiplying by the inverse of H_{AB}, H^{AB}, we obtain

$$H^{CA} \nabla_a \sigma M^a_{AB} = \sigma_t \delta^C{}_B \sum_i \sigma_i H^{CA} M^i_{AB}, \tag{13.35}$$

which is an operator, that is, a linear map from a vector space to itself, and therefore we can take its determinant, obtaining,

$$det \left(\sigma_t \delta^C{}_B \sum_i \sigma_i H^{CA} M^i_{AB} \right) = 0, \tag{13.36}$$

since the determinant of $H^{CA}\nabla_a \sigma M^a_{AB}$ vanishes because we have assumed that the rank of $\nabla_a \sigma M^a_{AB}$ ceased to be maximum. For each fixed value of the spatial derivatives σ_i this equation will generally have n real solutions (roots), σ_t, (the eigenvalues of the operator $\sum_i \sigma_i H^{CA} M^i_{AB}$), of all of them we will take the one that gives us the boundary of the smallest cone containing t_a, that is, the largest root. We will thus have for this root an equation of the form:

$$\sigma_t + H(\sigma_i, x^i, t) = 0. \tag{13.37}$$

The function H has a very important property, note that if (σ_t, σ_i) is a solution so is $(\alpha\sigma_t, \alpha\sigma_i)$ and therefore H must be **homogeneous of first degree**, that is $H(\alpha\sigma_i, x^i, t) = \alpha H(\sigma_i, x^i, t)$. These equations, with H homogeneous of first degree are called **eikonal equations** and are particular cases of the Hamilton–Jacobi equation studied in mechanics. This type of equations can be solved by transforming them into an equivalent problem in ordinary derivatives derived from a Hamiltonian. Indeed, consider the system of ordinary Hamiltonian equations:

$$\begin{aligned} \frac{dx^i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial x^i}, \end{aligned} \tag{13.38}$$

where $H(p_i, x^i, t) := H(\sigma_i, x^i, t)|\sigma_i = p_i$. Integrating this system with initial conditions:

$$\begin{aligned} x^i(0) &= x^i_0 \\ p_i(0) &= \sigma_0 i, \end{aligned} \tag{13.39}$$

and then restricting the integral curves obtained in the phase space (x^i, p_j) to the configuration space x^i, we obtain a series of curves in our manifold emanating from the surface Γ. We will call these curves, **characteristic curves** since they have the important property that along them the function σ we are looking for is constant! To prove this first note that with the chosen initial conditions $p_i(t) = \sigma_i(x^j(t), t); \forall; t$. But taking the derivative with respect to x^i of equation 13.37 we obtain:

$$\begin{aligned} \frac{\partial \sigma_i}{\partial t} &= \frac{\partial^2 \sigma}{\partial t \partial x^i} \\ &= \frac{\partial^2 \sigma}{\partial x^i \partial t} \\ &= -\sum_j \frac{\partial H}{\partial \sigma_j}(\sigma_k, x^k, t)\frac{\partial \sigma_i}{\partial x^j} - \frac{\partial H}{\partial x^i}(\sigma_k, x^k, t), \end{aligned} \tag{13.40}$$

and therefore

$$\begin{aligned}
\frac{d(p_i-\sigma_i)}{dt} &= -\frac{\partial H}{\partial x^i}(p_k,x^k,t)\sum_j \frac{\partial \sigma_i}{\partial x^j}\frac{dx^j}{dt} - \frac{\partial \sigma_i}{\partial t} \\
&= -\frac{\partial H}{\partial x^i}(p_k,x^k,t)\sum_j \frac{\partial \sigma_i}{\partial x^j}\frac{\partial H}{\partial p_j}(p_k,x^k,t) \\
&+ \sum_j \frac{\partial H}{\partial \sigma_j}(\sigma_k,x^k,t)\frac{\partial \sigma_i}{\partial x^j} + \frac{\partial H}{\partial x^i}(\sigma_k,x^k,t) \\
&= -\frac{\partial H}{\partial x^i}(p_k,x^k,t) + \frac{\partial H}{\partial x^i}(\sigma_k,x^k,t) \\
&- \sum_j \frac{\partial \sigma_i}{\partial x^j}\left(\frac{\partial H}{\partial p_j}(p_k,x^k,t)\frac{\partial H}{\partial \sigma_j}(\sigma_k,x^k,t)\right), \qquad (13.41)
\end{aligned}$$

where in the second equality we have used equations 13.38 and 13.40. This last equation, with the chosen initial conditions has a trivial solution, but the uniqueness theorem of solutions of ordinary equations guarantees that this is the only one and therefore the sought equality. Therefore from now on we should not distinguish in the argument of H whether it is evaluated in σ_i or in p_i. But then note that the derivative along a characteristic curve of σ is,

$$\begin{aligned}
\frac{d\sigma}{dt} &= \sum_i \sigma_i \frac{dx^i}{dt} + \sigma_t \\
&= \sum_i \frac{\partial H}{\partial \sigma_i}\sigma_i - H(\sigma_k,x^k,t) \\
&= 0, \qquad (13.42)
\end{aligned}$$

where in the last equality we have used that since H is homogeneous of first degree in σ_i the equality $H(\sigma_k,x^k,t) = \sum_i \frac{\partial H}{\partial \sigma_i}\sigma_i$ holds.

We have thus demonstrated that σ will be constant along the integral lines of equation 13.38 with initial conditions given by 13.39. Therefore we know S, this will be the surface ruled by the characteristic curves emanating from $\partial\Gamma$ [6]. [See figure 13.9.]

13.4.2 Domain of dependence, examples

Next we give a couple of examples of the construction of characteristic curves and determination of the domains of dependence.

[6] In general the characteristic curves intersect each other and therefore even giving a region Γ with smooth boundary after a certain time the ruled surface will develop singularities and σ will be multivalued. However these singularities are perfectly well known and it is known how to discard regions until obtaining domains of dependence with continuous boundaries.

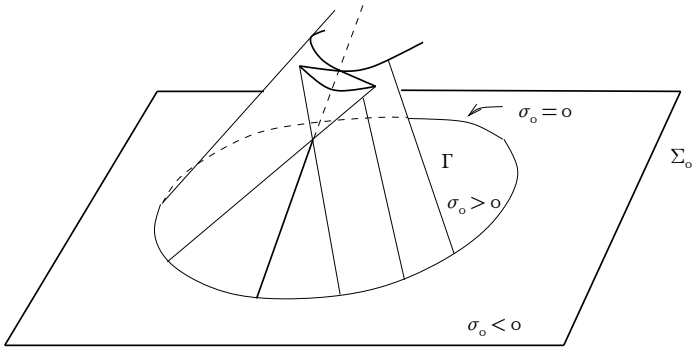

Figure 13.9: Construction of S and a singularity in S

Example: **Fluids in one dimension**

Consider a fluid with average density ρ_o, moving with average velocity v_o and with an equation of state for the pressure as a function of density $p = p(\rho)$. We are interested in small fluctuations of these quantities around the equilibrium state (ρ_o, v_o), that is, in the theory of sound in this fluid. In this case $u^A = (\rho, v)$ will be these fluctuations and the fluid equations are

$$\begin{aligned} \frac{\partial \rho}{\partial t} - v_o \frac{\partial \rho}{\partial x} - \rho_o \frac{\partial v}{\partial x} &= 0 \\ \frac{\partial v}{\partial t} - v_o \frac{\partial v}{\partial x} - \frac{1}{\rho_o} \frac{\partial p}{\partial x} &= 0. \end{aligned} \tag{13.43}$$

If $c^2 := \frac{dp}{d\rho}|\rho_o > 0$, ($c$ is the speed of sound in the medium) then the system is symmetric–hyperbolic with $M^a AB$ obtained by rewriting the equations as:

$$\begin{pmatrix} c^2 & 0 \\ 0 & \rho_o^2 \end{pmatrix} \begin{pmatrix} \rho \\ v \end{pmatrix}_t + \begin{pmatrix} -c^2 v_o & -c^2 \rho_o \\ -c^2 \rho_o & -\rho_o^2 v_o \end{pmatrix} \begin{pmatrix} \rho \\ v \end{pmatrix}_x = 0, \tag{13.44}$$

where we have used that $\frac{\partial p}{\partial x} = \frac{\partial p}{\partial \rho} \frac{\partial \rho}{\partial x}$.

The determinant we must study is then:

$$det \begin{pmatrix} \sigma_t c^2 - \sigma_x c^2 v_o & -\sigma_x c^2 \rho_o \\ -\sigma_x c^2 \rho_o & \sigma_t \rho_o^2 - \sigma_x \rho v \end{pmatrix}, \tag{13.45}$$

which has as roots,

$$\sigma_t = (v_o \pm c)\sigma_x. \tag{13.46}$$

Suppose $c > v_o$ (normal fluid, subsonic), and that $\Gamma = [0, 1]$ with $\sigma_o = x(x-1)$. At $x = 0$ σ_{ox} is negative and then the largest root is $v_o - c$ and the solution is $\sigma_- =$

$t(c-v_o)-x$. At $x=1$ σ_{ox} is positive and then the largest root is v_o+c and the solution is $\sigma_- = t(v_o+c)-(x-1)$. [See figure.]

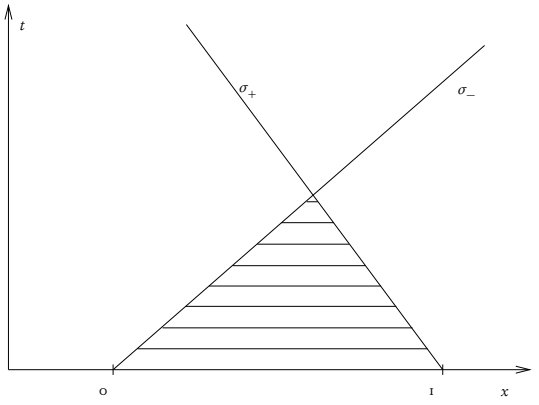

Figure 13.10: Domain of dependence of a fluid

Example: **Wave equation in $2+1$ dimensions.**

The equation is:

$$-\frac{\partial^2\phi}{\partial t\partial t}+\frac{\partial^2\phi}{\partial x\partial x}+\frac{\partial^2\phi}{\partial y\partial y}=\rho, \tag{13.47}$$

and using $u^A=(\phi,\frac{\partial\phi}{\partial t},\frac{\partial\phi}{\partial x},\frac{\partial\phi}{\partial y})$ the system can be written as,

$$\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\begin{pmatrix}u^1\\u^2\\u^3\\u^4\end{pmatrix}_t-\begin{pmatrix}0&0&0&0\\0&0&1&0\\0&1&0&0\\0&0&0&0\end{pmatrix}\begin{pmatrix}u^1\\u^2\\u^3\\u^4\end{pmatrix}_x-\begin{pmatrix}0&0&0&0\\0&0&0&1\\0&0&0&0\\0&1&0&0\end{pmatrix}\begin{pmatrix}u^1\\u^2\\u^3\\u^4\end{pmatrix}_y=\begin{pmatrix}-u^2\\-\rho\\0\\0\end{pmatrix}$$

Exercise: Prove that if the initial data is such that $u^2=\frac{\partial u^1}{\partial x}$ and $u^3=\frac{\partial u^1}{\partial y}$, then u^1 satisfies 13.47.

The equations for σ are:

$$det\begin{pmatrix}\sigma_t&0&0&0\\0&\sigma_t&\sigma_x&\sigma_y\\0&\sigma_x&\sigma_t&0\\0&\sigma_y&0&\sigma_t\end{pmatrix}=(\sigma_t)^2[(\sigma_t)^2-(\sigma_x)^2-(\sigma_y)^2], \tag{13.48}$$

that is

$$\sigma_t = \sqrt{(\sigma_x)^2 + (\sigma_y)^2} := -H(\sigma_x, \sigma_y). \tag{13.49}$$

The Hamilton equations for this system are:

$$\begin{aligned}
\frac{dx}{dt} &= \frac{\partial H}{\partial \sigma_x} = \frac{-\sigma_x}{\sqrt{(\sigma_x)^2 + (\sigma_y)^2}} \\
\frac{dy}{dt} &= \frac{\partial H}{\partial \sigma_y} = \frac{-\sigma_y}{\sqrt{(\sigma_x)^2 + (\sigma_y)^2}} \\
\frac{d\sigma_x}{dt} &= -\frac{\partial H}{\partial x} = 0 \\
\frac{d\sigma_y}{dt} &= -\frac{\partial H}{\partial y} = 0,
\end{aligned} \tag{13.50}$$

and their solutions are:

$$\begin{aligned}
x(t) &= \frac{-\sigma_{ox}}{\sqrt{(\sigma_{ox})^2 + (\sigma_{oy})^2}} t + x_o \\
y(t) &= \frac{-\sigma_{oy}}{\sqrt{(\sigma_o x)^2 + (\sigma_{oy})^2}} t + y_o.
\end{aligned} \tag{13.51}$$

Chapter 14

Parabolic Equations

14.1 Introduction

Here we will deal with the archetype of parabolic equation, the heat equation,

$$\begin{aligned} \frac{\partial u}{\partial t} - \Delta u &= f \text{ in } \Omega, \\ u|_{t=0} &= u^0, \\ u|_L &= 0, \end{aligned} \tag{14.1}$$

where Ω is a cylindrical region of the form $[0, T] \times S$ and $L = [0, T] \times \partial S$. See figure.

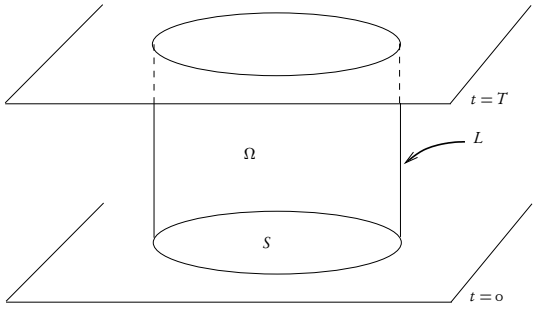

Figure 14.1: Boundary conditions for the heat equation.

We can also consider the problem where $n^a \nabla_a u|_L = 0$, or mixed problems, here we will only deal with the first one, as the treatment of the others does not involve major changes. Non-homogeneous boundary conditions are treated similarly to the case of elliptic equations.

Using the Spectral Theorem, 12.2, we decompose u, f and u^0 into eigenfunctions

of the Laplacian in S in $H^1_0(S)$ and obtain,

$$\begin{aligned} u(t,x^j) &= \sum_{n=0}^{\infty} u_n(t)v_n(x^j), \\ f(t,x^j) &= \sum_{n=0}^{\infty} f_n(t)v_n(x^j), \\ u^0(x^j) &= \sum_{n=0}^{\infty} u_n^0 v_n(x^j), \end{aligned}$$

where the functions $v_n(x^j)$ satisfy,

$$\begin{aligned} \Delta v_n(x^j) &= \lambda_n v_n(x^j) \\ v_n(x^j)|_L &= 0. \end{aligned}$$

We thus obtain the following infinite system of ordinary equations,

$$\begin{aligned} \dot{u}_n + \lambda_n u_n &= f_n \\ u_n|_{t=0} &= u_n^0. \end{aligned} \tag{14.2}$$

The solution of the homogeneous equation is $u_n^H = u_n^0 e^{-\lambda_n t}$ and using the method of variation of constants we obtain,

$$u_n^I = e^{-\lambda_n t} \int_0^t f_n(\tilde{t}) e^{\lambda_n \tilde{t}} d\tilde{t}. \tag{14.3}$$

The complete solution (respecting the initial condition) is

$$u_n(t) = e^{-\lambda_n t}[u_n^0 + \int_0^t f_n(\tilde{t}) e^{\lambda_n \tilde{t}} d\tilde{t}]. \tag{14.4}$$

Similarly to how we proved that the formal solution for the hyperbolic case was a smooth solution, it can be shown here that for $t > 0$ the solution is, in the spatial variables, twice more differentiable than f and in the temporal variable, once more.

A very important property of this equation is that in general it only admits one solution for $t > 0$, this implies that unlike the equations of mechanics or electromagnetism, this equation privileges a particular temporal direction. Among other things, this indicates that this equation represents or describes phenomena that cannot be derived solely from mechanics and that there must be some kind of thermodynamic information in them.

To see this, let's revisit the example given in the introduction to the Spectral Theorem, 12.2

$$\dot{u} - \frac{d^2 u}{dx^2} = 0 \text{ in } [0,1], \tag{14.5}$$

where we saw that the eigenfunctions $v_n(x)$ with $v_n(0) = v_n(1) = 0$ are $v_n(x) = \sin(\pi n x)$ with eigenvalue $\lambda_n = \pi^2 n^2$.

Taking as initial data $u^0(x) = \sum_{n=1}^{\infty} \frac{(-1)^{\frac{n-1}{2}}}{n^2} \sin(\pi n x)$, which is bounded in $[0, 1]$ since $\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$, we obtain,

$$u(t,x) = \sum_{n=1}^{\infty} \frac{(-1)^{\frac{n-1}{2}} e^{-\pi^2 n^2 t}}{n^2} \sin(\pi n x).$$

But $u(t, 1/2) = \sum_{n=1}^{\infty} \frac{e^{-\pi^2 (2n-1)^2 t}}{(2n-1)^2}$ which is finite $\forall \quad t \geq 0$ and infinite for any $t < 0$ since the nth term tends to infinity with n. On the other hand, given any $u^0(x) = \sum_{n=0}^{\infty} u_n^0 \sin(\pi n x)$, continuous we obtain

$$u(t,x) = \sum_{n=0}^{\infty} u_n^0 e^{-\pi^2 n^2 t} \sin(\pi n x) \tag{14.6}$$

which is infinitely differentiable for all $t > 0$ since given any polynomial $P(n)$ of n, $P(n) e^{-\pi^2 n^2 t} \to 0$ when $n \to \infty$ if $t > 0$.

14.2 Uniqueness and the Maximum Theorem

We now see that the solution obtained in the general case is unique. To do this, we will assume that it is C^1 with respect to time and C^2 with respect to spatial coordinates. To reach this conclusion, we will use the maximum principle. We develop this first for the Dirichlet problem.

Theorem 14.1 (of the Maximum) *Let $u \in C^2(S)$ and $\Delta u \geq 0$ in S, then*

$$\max_{\bar{S}} u = \max_{\partial S} u$$

Proof: If $\Delta u > 0$ this simply follows from the fact that if the maximum were at $p \in S$ then $\frac{\partial^2 u}{\partial (x^k)^2}\big|_p \leq 0 \quad \forall \quad k = 1, ..., n$ and therefore, by continuity of the second derivatives, $\Delta u \leq 0$, in an entire region within S, which is a contradiction. For the case $\Delta u \geq 0$ we use the function $v = |x|^2$. Since $\Delta v > 0$ in S then

$$\Delta(u + \varepsilon v) > 0 \quad \text{in } S \ \ \forall \varepsilon > 0 \tag{14.7}$$

and thus

$$\max_{\bar{S}} (u + \varepsilon v) = \max_{\partial S} (u + \varepsilon v)$$

and therefore

$$\max_{\bar{S}} u + \varepsilon \min_{\bar{S}} v \leq \max_{\partial S} u + \varepsilon \max_{\partial S} v$$

and taking $\varepsilon \to 0$ we obtain the desired equality. In the case where $\Delta u = 0$ then it also holds that

$$\min_{\bar{S}} u = \min_{\partial S} u$$

(Simply using that $\min(u) = -\max(-u)$).

This result gives us another proof of the uniqueness of solutions to the Dirichlet problem for the Laplacian.

Theorem 14.2 (Uniqueness of the Dirichlet Problem) *The problem*

$$\begin{aligned} \Delta u &= f \quad in\ S \\ u|_{\partial S} &= u_0, \end{aligned} \tag{14.8}$$

has at most a unique solution in $C^2(S)$.

Proof: Let u and $\tilde{u}$ in $C^2(S)$ be solutions, then $\delta = u - \tilde{u}$ satisfies $\Delta\delta = 0$ and $\delta|\partial S = 0$ but then

$$\max_{\bar{S}} \delta = \max_{\partial S} \delta = 0$$

and

$$\min_{\bar{S}} \delta = \min_{\partial S} \delta = 0$$

therefore we conclude that $\delta = 0$.

Exercise: For which other elliptic equations does this uniqueness proof hold?

Can we obtain a similar result for the heat equation?

Theorem 14.3 (Uniqueness for the Heat Equation) *There exists at most a unique solution* $u \in C^1[0,T] \times C^2(S)$ *to the problem,*

$$\begin{aligned} \dot{u} - \Delta u &= f \quad in\ S \\ u|_{t=0} &= u^0, \\ u|_L &= u^1, \end{aligned} \tag{14.9}$$

u^0 *and* u^1 *continuous.*

The proof of this theorem is a trivial application of the following lemma.

Lemma 14.1 *Let* $u \in C^1[0,T] \times C^2(S)$ *be continuous in* $\bar{\Omega}$ *and satisfying* $u_t - \Delta u \leq 0$. *Then*

$$\max_{\bar{\Omega}} u = \max_{\partial'\Omega} u$$

where $\partial'\Omega = S_0 \cup L$.

Proof: First, consider the case $u_t - \Delta u < 0$ in Ω. Let $0 < \varepsilon < T$ and $\Omega_\varepsilon = \cup_{\tilde{t}\in(0,T-\varepsilon)} S_{\tilde{t}}$. Since u is continuous in $\bar{\Omega}_\varepsilon$ there will exist $p \in \bar{\Omega}_\varepsilon$ such that $u(p) = \max_{\bar{\Omega}_\varepsilon} u$. If $p \in \Omega_\varepsilon$ then there $u_t = 0$ and $\Delta u \leq 0$ which gives us a contradiction. If

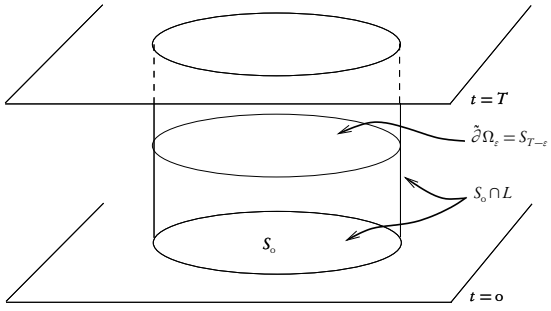

Figure 14.2: Proof of Lemma 14.1

$p \in \tilde{\partial}\Omega_\varepsilon = S_{T-\varepsilon}$ we would have $u_t \geq 0$ and $\Delta u \leq 0$, which also gives us a contradiction. It follows then that $p \in \partial'\Omega_\varepsilon$ and

$$\max_{\bar{\Omega}_\varepsilon} u = \max_{\partial'\Omega_\varepsilon} u \leq \max_{\partial'\Omega} u$$

letting $\varepsilon \to 0$ we obtain the desired result.

To handle the case $u_t - \Delta u \leq 0$ in Ω, consider $v = u - kt$, $k > 0$. Then $v_t - \Delta v = u_t - \Delta u - k < 0$ and therefore

$$\max_{\bar{\Omega}} u = \max_{\bar{\Omega}}(v + kt) \leq \max_{\bar{\Omega}}(v + kT) \leq \max_{\partial'\Omega} v; +; kT; \leq; \max_{\partial'\Omega} u + kT, \quad (14.10)$$

where in the last inequality we have used that $\max v \leq \max u$ if $k \geq 0$, $t \geq 0$. Taking the limit, $k \to 0$ we obtain the desired result.

Exercise: Prove that if $\frac{\partial u}{\partial t} - \Delta u \geq 0$ then $\min_{\bar{\Omega}} u = \min_{\partial'\Omega} u$.

CHAPTER 15

GROUPS

15.1 Introduction

Groups play a fundamental role in contemporary physics. This is because symmetries are observed in nature, or their correlate, via Noether's theorem, conserved quantities, that is, quantities that do not vary during the interactions of the various fields over their respective time scales. These symmetries are invariances under transformations, such as spatial or temporal translations, rotations, or transformations between different fields, such as those accounting for the conservation of electric charge, baryon number, etc. All these transformations form groups, and determining their structure is a fundamental ingredient in constructing models that approximate reality.

We begin with the abstract definition of a group:

Definition: A group is a set G and an assignment, $\Psi : G \times G \to G$, called *multiplication*, and usually denoted in the same way, $\Psi(g, g') := g\,g'$, satisfying the following conditions:

- Associativity: $g(g'g'') = (g\,g')g''$
- There exists an element in G, usually denoted by e, such that

$$ge = eg = e \;\; \forall\, g \in G$$

- For each element $g \in G$ there exists another element in G, called its inverse, g^{-1}, such that,

$$g\,g^{-1} = g^{-1}g = e$$

Example: Let the set Z be the set of real numbers, including negatives and zero. Let the product be given by addition, $\Psi(n, m) = n + m$. This operation is clearly associative, $\Psi(n, \Psi(m, l)) = \Psi(n, m+l) = n+m+l = \Psi(n+m, l) = \Psi(\Psi(n, m), l)$. In this case, the identity is the number zero, $\Psi(0, n) = \Psi(n, 0) = n$ and the inverse of the element n is the number $-n$.

Example: Let the set $R - \{0\}$ and the multiplication be the usual one. See that this set with this operation is a group.

Example: Now let's see the most important example of a group, as we will see later all groups can be obtained by this construction. Let S be any set, let $P(S)$ be the set of injective and surjective maps from S to S, $\Psi : S \to S$. Let the product be the composition of maps. Note that the composition of maps is clearly associative,

$$\Psi o(\Phi o \chi)(s) = \Psi o \Phi(\chi(s)) = \Psi(\Phi(\chi(s))) = (\Psi o \Phi)(\chi(s)) = (\Psi o \Phi) o \chi(s).$$

On the other hand, the composition of invertible maps gives an invertible map and therefore the multiplication is closed in $P(S)$. Since the considered maps are invertible, their inverses as maps are their inverses as groups and the identity as a map is the identity in the group.

Example: Let $S = \{1, 2\}$ be a set of two elements. $P(S)$ consists of only two elements, the identity, $e(1, 2) = (1, 2)$ and the permutation $g(1, 2) := (2, 1)$. Note that g is its own inverse, $g\,o\,g(12) = g(2, 1) = (1, 2)$.

Abstractly, we can define this group by its multiplication table,

Table 15.1: Multiplication table of $P(1, 2)$

e	g
g	e

Example: Let's see the following case, let $S = \{1, 2, 3\}$. In this case, we have the following maps: $e(1, 2, 3) = (1, 2, 3)$, $f_1(1, 2, 3) = (1, 3, 2)$, $f_2(1, 2, 3) = (3, 2, 1)$, $f_3(1, 2, 3) = (2, 1, 3)$, $g(1, 2, 3) = (3, 1, 2)$, $g'(1, 2, 3) = (2, 3, 1)$. Their multiplication table is as follows:

Table 15.2: Multiplication table of $P(1, 2, 3)$

e	g	g'	f_1	f_2	f_3
g	g'	e	f_3	f_1	f_2
g'	e'	g	f_2	f_3	f_1
f_1	f_2	f_3	e	g	g'
f_2	f_3	f_1	g'	e	g
f_3	f_1	f_2	g	g'	e

Exercise: Consider the group of all linear maps from $\mathbb{R}^2 \mapsto \mathbb{R}^2$ that preserve an equi-

lateral triangle. Construct its multiplication table and compare it with the previous table.

Exercise: Consider the space of all linear maps from $\mathbb{R}^2 \mapsto \mathbb{R}^2$ that preserve a square. Compare its table with that of $P(1,2,3,4)$.

15.2 Isomorphisms

As we have seen, the group of linear transformations in $\mathbb{R}^2$ that preserve an equilateral triangle and the group $P(1,2,3)$ have the same multiplication table. For all practical purposes, these two groups are identical in an abstract sense, the study of one gives us all the information about the other. This is not an exception, in physics the same groups appear as transformations in very diverse situations and spaces, and it is not so easy to establish the relationship between them. Formally, the relationship between them can be defined as follows:

Definition: Two groups, G and H are *isomorphic* if there exists an invertible map between them, $\Psi : G \mapsto H$ such that,

$$\Psi(g)\Psi(g') = \Psi(g\,g').$$

Note that on the left side the product is the product of H, while on the right side the product is that of G. That is, the map Ψ preserves the product between the spaces. If two groups are isomorphic, then they are identical in terms of their intrinsic properties.

Exercise: Find the map that makes $P(1,2)$ and the group of all linear maps from $\mathbb{R}^2 \mapsto \mathbb{R}^2$ that preserve an equilateral triangle isomorphic.

15.3 Subgroups

Definition: A subset H of G is a subgroup of G if it is a group in itself, with the multiplication operation inherited from G. That is, if given two elements, $h,\ h' \in H$, $hh' \in H$, meaning it is closed under the product, and if $h \in H$, then $h^{-1} \in H$. Note that both conditions already ensure that $e \in H$.

Example: In $P(1,2,3)$ the elements e, g, g' form a subgroup of H.

Exercise: Find a non-trivial subgroup (the identity element is always a subgroup) of the group given by the set $\mathbb{R} - \{0\}$ with the usual multiplication.

Example: If H_1 and H_2 are subgroups, then their intersection, $H_1 \cap H_2$, is also a subgroup. Given a subset S of G, the intersection of all groups containing it is a subgroup. It is called the subgroup generated by S.

Exercise: Prove the statements of the previous example.

Exercise: Consider a finite set S and the group of permutations on it, $P(S)$. Let S' be a subset of S and consider all maps that, when restricted to S', are the identity. See that this set is a subgroup.

Exercise: Let A be a subset of G. See that A is a subgroup if and only if $\forall\ a,\ a' \in A$, $a^{-1}a' \in A$.

Exercise: Let $P(S)$, with S finite. Let

$$A := \{\mu \in P(S) | \exists\ s,\ s'\ s''\ \text{distinct, such that}\ \mu(s) = s',\ \mu(s') = s''\ \mu(s'') = s.\}$$

Show that the subgroup generated by A is not the whole $P(S)$.

15.4 The Universal Construction

We saw several examples of groups that appear as transformation groups, that is, as invertible maps from a set to itself. In reality, all groups can be obtained in this way. We cite the following theorem, which we will not prove, but which is very important as it places groups in their proper dimension:

Theorem 15.1 *Given a group G, there exists a set S such that G is isomorphic to a subgroup of $P(S)$.*

Later we will see a way to establish this result.

Example: Let the group whose set consists of the integers and the multiplication be addition. See that this group is isomorphic to the group of translations on Z: $\Psi_n : Z \mapsto Z$, given by $\Psi_n(m) = n + m$.

Example: Let the group whose set is $R - \{0\}$ with multiplication given by the product. This group can be obtained as the set of invertible maps of the real line onto itself given by, $\Psi_x : I\!R \mapsto I\!R$ given by $\Psi_x(y) := xy$. These maps are called dilations.

15.5 Linear Groups

Linear groups are those groups that appear as subgroups of the group of linear and invertible maps from $I\!R^n \mapsto I\!R^n$ or their complex versions, the maps from $C^n \mapsto C^n$.

The largest of them is the complete set, called the *general linear group*, and denoted $GL(n, I\!R^n/C^n)$, that is, the set of invertible $n \times n$ square matrices, with real or complex elements.

A subgroup of it is the set of all matrices with determinant of modulus one. Note that this set is really a subgroup since,

$$\det(A)\det(B) = \det(AB).$$

A smaller subgroup of this is formed by the subset of those with determinant one. These are called the *special linear group*, $SL(n, \mathbb{R}^n/C^n)$.

These spaces have more subgroups, but they do not manifest naturally unless we include more structure. For example, if we introduce a preferred basis, then the set of matrices that are upper (lower) triangular with all diagonal elements non-zero (invertible) are subgroups. More relevant are those that appear when we introduce a scalar product. In that case, we see that the unitary (orthogonal) matrices form subgroups, in fact, if U and U' are unitary, that is,

$$\langle Ux, Uy\rangle = \langle U'x, U'y\rangle = \langle x, y\rangle,$$

their product is also unitary,

$$\langle UU'x, UU'y\rangle = \langle U(U'x), U(U'y)\rangle = \langle U'x, U'y\rangle = \langle x, y\rangle.$$

These groups are denoted as $U(n)$ when we consider complex (unitary) matrices and $O(n)$ for the real case. Those with determinant one are denoted by $SU(n)$ and $SO(n)$.

The group $SO(3)$ is the group of rotations in space. The $O(n)$ incorporates, in addition to rotations, reflections perpendicular to arbitrary planes.

Exercise: Find and describe all these previous groups for $n = 1$ and $n = 2$.

Exercise: See that $U(1)$ has $P(1, 2, 3)$ as a subgroup. $U(1)$ appears in physics as a symmetry group associated with the conservation of electric charge.

Exercise: What other subgroups does $U(1)$ have?

Exercise: See that $U(1)$ is isomorphic to $SO(2)$.

Exercise: See that $SU(2)$ covers $SO(3)$ twice. Hint, use the Pauli matrices. $SU(2)$ appears in physics as one of the main groups of the standard model of particles, as does $SU(3)$.

Exercise:

15.5.1 The group $SO(3)$.

The group $SO(3)$ consists of the orthonormal 3×3 matrices with determinant one. That is, the rotations in $\mathbb{R}^3$. Let's see what properties this set has as a variety. One way to describe this set is as follows: To determine a rotation, we need to give the invariant axis of it, that is, a direction in $\mathbb{R}^3$, which we take as a unit vector in $\mathbb{R}^3$, $\hat{n}$ and the angle of it. To the latter we can assign a range given by $[-\pi, \pi]$ taking into account that the rotation in the direction $\hat{n}$ by an angle π is equal to the rotation by the angle $-\pi$. Each of these rotations can be expressed by the matrix,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

Where we have chosen the coordinate axes so that the z axis coincides with the direction of the rotation. We see thus that the rotation by π is the same rotation as by $-\pi$. We can join the direction of rotation with the angle of it and form a vector so that its direction coincides with the direction of rotation and its magnitude with the angle of it. We see then that $SO(3)$ consists of all points in a ball of radius π, each diameter of it consists of all rotations about a fixed axis. But this set has a peculiarity, we must identify each point of the outer sphere, of diameter π, with its antipode. Thus we achieve a variety that has no boundary. If we draw a curve that intends to leave it, when it reaches the diameter π it appears at the opposite point of that sphere entering the interior of the ball through that point. In particular, if we start such a curve at a point p inside the ball, go towards the diameter π and continue it until we reach the starting point p again, we will have a closed curve that cannot be deformed to the identity curve! [1] That is, as a variety, $SO(3)$ is not simply connected. The following example also shows that it is not a torus.

Exercise: Show graphically that a curve that traverses the previous path twice is deformable to the identity.

This property of $SO(3)$ has important consequences in physics. It is what allows the mathematical existence of spinors, that is, the model that describes all fermionic fields, in particular the electron. This is because spinors, when rotated by an angle of 2π, that is, along one of the axes of $SO(3)$, change sign (they know about the non-deformable curves). While if we rotate them by an angle of 4π they return to their original state.

15.6 Cosets

Given a group G and a subgroup H of it, we can construct two quotient spaces. One of them is obtained by introducing the following equivalence relation:

We will say that $g_1 \approx g_2$ if there exists $g \in G$ such that $g_1 = g h_1$ and $g_2 = g h_2$. With $h_1, h_2 \in H$.

The first two conditions that define an equivalence relation are trivially satisfied. Let's see the third one:

We need to see that if $g_1 \approx g_2$ and $g_2 \approx g_3$, then $g_1 \approx g_3$. The first two conditions give us, $g_1 = g h_1$ and $g_2 = g h_2$ for some $g \in G$ and $g_2 = \tilde{g} \tilde{h}_2$ and $g_3 = \tilde{g} \tilde{h}_3$ for some $\tilde{g} \in G$. Using the two relations for g_2 we see that $\tilde{g} = g h_2 \tilde{h}_2^{-1}$. Therefore, $g_3 = \tilde{g} \tilde{h}_3 = g h_2 \tilde{h}_2^{-1} \tilde{h}_3$. This proves the statement since $h_2 \tilde{h}^{-1}{}_2 \tilde{h}_3 \in H$.

[1] We will say that a closed curve $\gamma(t)$ with $\gamma(0) = \gamma(1) = p$ in a given variety, M, can be deformed to the identity curve $i(t) \equiv p$, if there is a map $\phi(t,s)$, $[0,1] \times [0,1] \mapsto M$ such that $\phi(t,0) = \gamma(t)$ and $\phi(t,1) = i(t) = p$.

Thus, we define $L = G/H_L$, the set of equivalence classes with respect to this equivalence relation. The subscript L refers to making equivalences by multiplying the elements of H on the left. The other quotient space is obtained by multiplication on the right.

How are the elements of L constituted? Let $g \in G$, then we can consider the set,

$$gH := gh,\ h \in H.$$

Note that $g \in H$ since $e \in H$ and that all elements of gH are equivalent to each other and there is no element outside of gH that is equivalent to g. Therefore, these are the elements of L. Note that since these are equivalence classes, each element of G is in one and only one of these equivalence classes.

Exercise: Verify for the case of $P(1,2,3)$ that the right cosets corresponding to taking $P(1,2,3)$ as its own subgroup are the columns of the multiplication table. What do the rows correspond to? This indicates that each row and column is made up of distinct elements.

Let A and B be two elements of L. Taking any two elements of G in each of these equivalence classes, g_A and g_B, we have $A = g_A H$ and $B = g_B H$. Thus, $B = g_B H = g_B g_A^{-1} A$ and we see that given any two elements of L, there exists an invertible map that sends all elements of one to the other. That is, all equivalence classes are isomorphic to each other. Note that this has the following implication. Let G be a group with p elements and H a subgroup of it. The quotient space will have n elements, equivalence classes, each with q elements, the number of elements of H. Since the equivalence relations divide the set G into disjoint classes, we must have $p = nq$. That is, the number of elements of a subgroup divides the number of elements of the group! In particular, if the number of elements of a group is prime, we can immediately conclude that it will only have trivial subgroups, the identity and the complete group.

Example: Let $P(1,2,3) = e, g, g', f_1, f_2, f_3$, this group has the following subgroups, $H_0 := e, g, g'$ and $H_i := e, f_i$. Since H_0 has 3 elements, $(P(1,2,3)/H_0)L$ will have two elements, $L1 = H_0$ and $L_2 = f_1, f_2, f3$.

Exercise: See that $f_i L_1 = L_2$.

Exercise: How many elements does $(P(1,2,3)/H_i)_L$ have? Find them.

15.6.1 Homogeneous Spaces

The set $L = G/H_L$ is not generally a group, but it has interesting properties and these sets often appear in applications, usually under the name of Homogeneous Spaces.

We have already seen that G acts on L, given an element of L, A, the action of $g \in G$ on A is the element of L given by gA. We will denote this map as $\Psi_g : L \mapsto L$. As we have seen, given two elements of L, ;;A and B, there always exists an element $g \in G$ such that $\Psi_g(A) = B$, that is, it acts transitively. The inverse of this map is $\Psi_{g^{-1}}$. This tells us that all points of L have the same structure, the group moves them all around. Hence their name of homogeneous spaces.

Since the previous maps have inverses, they belong to $P(L)$, the space of permutations of L. Thus, we have a map from G to $P(L)$, which, given an element $g \in G$, assigns the map Ψ_g in $P(L)$. We have that,

$$\Psi_g o \Psi_{g'}(A) = \Psi_g(g'A) = g(g'A) = (g\,g')A = \Psi_{g\,g'}(A),$$

that is, a homomorphism between groups, indicating that G is a subgroup of $P(L)$. If we take $L = G/e = G$, we have the group as a homogeneous space and G as a subgroup of $P(G)$. This gives us a proof of the universal construction theorem mentioned earlier.

Example: Let $G = \mathbb{R} - 0$ with the usual multiplication operation among rationals. Let $H = -1, 1$. In this case $G/H = \mathbb{R}^+$.

15.7 Normal Subgroups

A particular class of subgroups are the so-called normal groups. These are the groups where the right and left cosets coincide, that is, $N \in G$ is a normal subgroup if it is a subgroup and also,

$$gN = Ng \qquad \forall g \in G$$

the important property that the cosets generated by normal subgroups have is that they have a closed multiplication operation that turns the set of them into groups. In fact, due to the relation that defines them,

$$gN\ g'N = g\,g'N\,N = g\,g'N$$

the group thus defined is not, in general, a subgroup of G.

Exercise: Show that the intersection of normal subgroups is normal.

In the same way that subgroups can be generated from subsets of G by taking the intersection of all subgroups containing the generating subset. Since the intersection of normal subgroups is a normal subgroup, the intersection of all subgroups containing the generating subset will generate a normal subgroup.

A particularly interesting case is when the subset is the set of elements of G of the form,

$$C := g\,\tilde{g}\,g^{-1}\,\tilde{g}$$

The normal subgroup generated from this set is called the commutator of G. This contains all the elements that do not commute in the sense that if we take the quotient, G/N we obtain an abelian group.

Exercise: Prove that a group is abelian if and only if its commutator subgroup is the identity.

Exercise: Let $P(S)$ be the group of permutations on S. See that the transformations that leave all elements of S invariant, except for a finite number of elements, form a normal subgroup.

Exercise: Show that a subgroup giving rise to only two cosets is necessarily normal.

Exercise: Let Z be the set of elements of G such that if $z \in Z$ then $gz = zg; \forall g \in G$. Show that Z is a normal subgroup.

Bibliography notes: These notes follow very closely a chapter in [1]

BIBLIOGRAPHY

[1] *Mathematical Physics*, Geroch, R., The Univ. of Chicago Press.

[2] *Topología General*, Kelley, J. L., EUDEBA.

[3] *Ordinary Differential Equations*, Arnold, V. I., The MIT Press, 1987.

[4] *Mathematical Methods of Classical Mechanics*, Arnold, V. I., Springer, 1980.

[5] *Differential Equations: Geometric Theory*, Lefschetz, S, Wiley, 1962.

[6] *General Relativity*, Wald, R., The University of Chicago Press, 1993.

[7] *Differential Equations and their applications, an introduction to applied mathematics*, Braum, M. New York, Springer, 1983.

[8] *Ecuaciones diferenciales ordinarias y teoría de control.* Roxin, E. Buenos Aires, EUDEBA, 1968.

[9] *Methods of Modern Mathematical Physics. I. Functional Analysis*, Reed, M., and Simon, B., Academic Press, 1972.

[10] *Real Analyis*, Lang, S., Addison-Wesley, 1983.

[11] *Introduction to Linear Algebra*, Lang, S., Undergraduate Texts in Mathematics, Springer, second edition, 1986.

[12] *Linear Algebra*, Lax, P., Pure and applied mathematics, Wiley, 1996.

[13] *Functional Analysis*, Yosida, K., Springer-Verlag 1980.

[14] *Introductory Rela Analysis*, A.N. Kolmogorov, and S.V. Fomin, Dover, 1970.

[15] *Basic Linear Partial Differential Equations*, Treves, F., Academic Press, 1975.

[16] *Partial Differential Equations*, John, F., Springer-Verlag 1982.

[17] *Introduction to Partial Differential Equations*, Folland, G. B., Mathematical Notes, Princeton Univ. Press 1976.

[18] *Partial Differential Equations I: Basic Theory* M. Taylor, Aplied Mathematical Sciences 115, Springer-Verlag, 1996.

[19] *Ecuaciones de la Física Matemática*, Godunov, S.K. Moscú, MIR, 1978.

[20] *Problemas de Ecuaciones Diferenciales Ordinarias* Kiseliov A. , Krasnov M., and Makarenko G., Moscú, MIR, Moscú, 1968.

[21] *Modern Differential Geometry for Physicists* Isham C., World Scientific Lecture Notes in Physics - Vol. 61 World Scientific, Second edition, 1999.