Advanced linear algebra

Teo Banica

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

2010 *Mathematics Subject Classification.* 15A18

Key words and phrases. Linear algebra, Matrix theory

Abstract. This is an introduction to advanced linear algebra, with emphasis on geometric aspects, and with some applications included too. We first review basic linear algebra, notably with the spectral theorem in its general form, and with the theory of the resultant and discriminant. Then we discuss the Jordan form and its basic applications to physics, and other advanced decomposition results for the matrices. We then go into positivity topics, involving matrices and bilinear forms, and with a look into curved space-time, and discrete Laplacians. Finally, we discuss the various groups of matrices, with a look at reflection groups, Lie groups, spin matrices, and random matrices.

Preface

This is an introduction to advanced linear algebra, with emphasis on geometric aspects, and with some applications included too. The book is organized itself with emphasis on algebra and symmetry, in 4 parts, having 4 chapters each, with each chapter having 24 pages, and consisting of 4 sections, plus an informal exercise section at the end.

Add to this 8 pages of front matter and 8 pages of back matter, and you have exactly 400 pages, according to the equation $8 + 16 \times 24 + 8 = 400$, that took me years and years to find, after countless organization attempts, with some previous books that I wrote. With this being something nice, making among others the text printer-friendly.

Getting now to the contents of the book, this will be certainly algebraic, but less maniac on algebraic aspects, and often insisting on the underlying geometry. Personally I tend to regard any vector as being something alive and dynamic, and any linear map as being something alive and dynamic too, and any matrix as consisting of course of numbers, which themselves are alive and dynamic objects too. With the "dynamism" of everything coming from the underlying physics, I mean, after all these years spent doing or teaching mathematics, I have yet to meet an interesting vector, of linear map, or matrix, not having something to do with physics. So, this will be for the general philosophy, geometry and physics, in order to understand linear algebra, and vice versa.

As already mentioned, the book is organized in 4 parts, which are as follows:

I - We review here the basic linear algebra, namely vectors, linear maps, matrices, matrix inversion, determinant, eigenvectors, eigenvalues and diagonalization. Then we review more advanced theory, such as the spectral theorem in its various forms, and the theory of the resultant and discriminant. We include as well all sorts of tricks, as for instance the fact that the diagonalizable matrices over $\mathbb{C}$ are dense.

II - Unfortunately not all matrices are diagonalizable, and in this second part, we discuss what can be done when they aren't. We first explain the Jordan form, along with its basic applications to physics, in relation with the dynamical systems. Then we discuss some other advanced decomposition results, and notably the singular value decomposition, and with a look into infinite dimensions and compact operators too.

III - Here we go into positivity and negativity topics, motivated by the positivity and negativity of the matrix eigenvalues, notably for the Hessian matrices. We first discuss the classification of the bilinear forms, in terms of their signature, and do not miss the occasion for talking a bit about curved spacetime, and Lorentz geometry. Then we discuss bistochastic matrices, discrete Fourier analysis, designs and discrete Laplacians.

IV - Finally, for ending in beauty, we discuss what matrices can do when operating together, in groups. Nothing or almost can resist to these frightening formations, and we discuss here the reflection groups, playing the role of Army, the Lie groups, playing the role of Navy, the tricky physics groups, also known as Navy Seals, and the thick squadrons known as random matrices, playing the role of the Air Force, I guess.

In the hope that you will like this book, which comes as a continuation of my basic linear algebra book [7]. Many thanks go to my students, and especially the undergraduate ones, that I often take into SU_2 and SO_3, no matter what the course is about, and who invariably like this stuff. Thanks as well to my colleagues, for countless coffee room discussions about linear algebra, and for some joint research on linear algebra too. Finally, many thanks go to my cats, for their teachings on both linearity and non-linearity.

Cergy, June 2025
Teo Banica

Contents

Part I

Linear algebra

Hey, where did we go
Days when the rains came
Down in the hollow
Playing a new game

CHAPTER 1

Linear maps

1a. Linear maps

As you can see, we live in $\mathbb{R}^3$, and this is where most of the questions in our mathematics take place. However, you also know from calculus, or from physics, that dealing with the mathematics of $\mathbb{R}^3$ is no easy matter. So, for this purpose, doing mathematics in $\mathbb{R}^3$, the best is to regard our space $\mathbb{R}^3$ as being part of a hierarchy of spaces $\mathbb{R}^N$, where you can do mathematics, at varying levels of difficulty, as follows:

(1) First comes $\mathbb{R}$. This has little to no interest in connection with real-life problems, but as you know well from calculus, everything mathematics comes from here, with this meaning sequences, convergence, series, functions, continuity, derivatives, integrals and so on. In this book we will assume the basic theory of $\mathbb{R}$ known, and in case you need from time to time to revise that, go with Rudin [77], or Lax-Terrell [68].

(2) Then comes $\mathbb{R}^2$. This is the entry point to advanced mathematics, because most of the $\mathbb{R}^3$ phenomenona have interesting 2D analogues, quite often capturing the whole point. Sometimes, $\mathbb{R}^2$ can be even your final destination, because many interesting $\mathbb{R}^3$ questions take place in fact in a plane $\mathbb{R}^2 \subset \mathbb{R}^3$. And finally, as another key feature of $\mathbb{R}^2$, we have an isomorphism $\mathbb{R}^2 \simeq \mathbb{C}$, transforming by some kind of magic your 2-variable questions in $\mathbb{R}^2$ into routine 1-variable problems, over the complex numbers $\mathbb{C}$. In this book we will assume $\mathbb{R}^2$, $\mathbb{C}$ reasonably known, and in case you need from time to time a reference here, go with the advanced books of Rudin [78], and Lax-Terrell [69].

(3) Then comes $\mathbb{R}^3$. Here there are no tricks of type $\mathbb{R}^2 \simeq \mathbb{C}$, so we are definitely into several variables, whose functioning we must understand well. But intuition, helped by the $\mathbb{R}^3$ surrounding us, can help a lot. As an interesting feature of $\mathbb{R}^3$, of rather engineering type, and contradicting all the mathematics that you learned, volumes of bodies $V \subset \mathbb{R}^3$ are easier to compute than areas $A \subset \mathbb{R}^2$, simply by plunging them into water, and measuring the water displacement. And isn't this genius. Also, mathematically, on $\mathbb{R}^3$ we have available the vector product $x \times y$, which can be useful for many things.

(4) Then comes $\mathbb{R}^4$. You would say why bothering with it, but the point is that, according to Einstein's relativity theory, our usual $\mathbb{R}^3$ does not really exist, in practice, as strange as this might seem, because the space variables (x, y, z) are in fact connected

to the time variable t. Thus, the correct variable for any physics problem, involving at least a bit or relativity, and there are so many of them, incuding everything having to do with light, electromagnetism, or quantum physics, is in fact (x, y, z, t), which lives in $\mathbb{R}^4$, or rather in a technical, curved version of $\mathbb{R}^4$. Note also that we have $\mathbb{R}^4 \simeq \mathbb{C}^2$.

(5) Then comes $\mathbb{R}^N$. This is actually simpler than both $\mathbb{R}^3$ and $\mathbb{R}^4$, for most matters, and when we said in (3) above, in relation with $\mathbb{R}^3$, that "we are definitely into several variables, whose functioning we must understand well", we meant by this "time to learn several variables, first in $\mathbb{R}^N$, and then in $\mathbb{R}^3$". Also, as another interesting feature of $\mathbb{R}^N$, the vector product $x \times y$ from $\mathbb{R}^3$ has no analogue in $\mathbb{R}^N$, and with this being a good thing, forcing us to rewrite many things that we know from $\mathbb{R}^3$, obtained via $x \times y$, in a more straightforward way in $\mathbb{R}^N$, by using the rock-solid scalar product $< x, y >$.

(6) Finally, we have $\mathbb{R}^\infty$. This is normally reserved for quantum mechanics matters, which live there, in infinite dimensions, and to be more precise in $\mathbb{C}^\infty$, to be fully correct, and in what regards the level of difficulty, with respect to $\mathbb{R}^3, \mathbb{R}^4, \mathbb{R}^N$, this can wildly vary, depending on the type of quantum mechanics questions that you have in mind. That is, for easy questions $\mathbb{R}^\infty$ can be simpler than $\mathbb{R}^N$, for the simple reason that there are less tools available, so less mathematics to be done. However, for difficult questions, $\mathbb{R}^\infty$ can be at the same level of difficulty with $\mathbb{R}^3, \mathbb{R}^4$, or even harder. Finally, forgetting about quantum, knowing a bit about $\mathbb{R}^\infty, \mathbb{C}^\infty$ can be useful for $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3, \ldots$, and this because the real or complex functions on $\mathbb{R}^N$ form spaces which are isomorphic to $\mathbb{R}^\infty, \mathbb{C}^\infty$.

So, this was the general story with mathematics, and more specifically geometry and analysis, motivated by physics questions, the conclusion being as follows:

CONCLUSION 1.1. *Mathematics inside $\mathbb{R}^3$ is a tricky business, best learned:*

(1) *By studying $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3, \mathbb{R}^4, \mathbb{R}^N, \mathbb{R}^\infty$, which are all useful,*
(2) *One at the time, switching dimensions when needed,*
(3) *And by having an eye on $\mathbb{C}, \mathbb{C}^2, \mathbb{C}^N, \mathbb{C}^\infty$ too.*

What about linear algebra, in relation with all this? Well, linear algebra is the key to geometry and analysis, and therefore to physics too, dealing with the most basic phenomena that can appear, the "linear" ones, which are at the core of everything.

You surely know some linear algebra, and I would assume here that you learned this in the rather abstract way that this is taught nowadays, worldwide, meaning a bit of vectors, linear maps and matrices over $\mathbb{R}^2, \mathbb{R}^3$, quickly done, then a lot of abstract study in $\mathbb{R}^N$, or worse, over an arbitrary real vector space, then some sort of incomprehensible things called "determinant" and "diagonalization", and that is pretty much it.

Well, time to review this, at a more advanced level. To start with, we certainly have some business to do with $\mathbb{R}^2, \mathbb{R}^3$, basic things there that you might know or not. Then,

as a main objective, we have to properly understand what the determinant is, and how the diagonalization works. And finally, in view of the few occurrences of $\mathbb{C}$ instead of $\mathbb{R}$ in the above, it is better to talk as well about linear algebra over arbitrary fields F.

In view of this, here will be our plan for the first 2 chapters of the present book:

PLAN 1.2. *We must review the linear algebra that we know, by learning:*

(1) *More geometry in $\mathbb{R}^2, \mathbb{R}^3$, with focus on the linear maps there.*
(2) *The precise and true meaning of the determinant, in $\mathbb{R}^N$.*
(3) *The diagonalization procedure, done geometrically, also in $\mathbb{R}^N$.*
(4) *What happens as well over $\mathbb{C}$, or over an arbitrary field F.*

So this will be our plan, and afterwards in chapters 3-16 we will of course further build on all this, with a number of results that should be new to you, I hope.

Before starting, a few references too. Normally what we will be doing here will be quite self-contained, but quite often coming with very compact proofs, for the linear algebra basics that you are supposed to know. As standard references here, you have Lang [64] if you are more into algebra, and Lax [64] if you are more into analysis. You can check also my basic linear algebra book [7], which is somehow more into geometry.

Getting started for good now, we need a definition for the linear maps, our main objects of study. Leaving the arbitrary fields F for later, here that definition is:

DEFINITION 1.3. *A map $f : \mathbb{R}^N \to \mathbb{R}^M$ is called linear when it satisfies the following equivalent conditions:*

(1) *Algebraic conditions: $f(x + y) = f(x) + f(y)$ and $f(\lambda x) = \lambda f(x)$.*
(2) *f maps lines to lines: $f(tx + (1 - t)y) = tf(x) + (1 - t)f(y)$.*
(3) *Each component of $f(x)$ appears as a linear combination $\sum_i \lambda_i x_i$.*
(4) *$f(x) = Ax$, for some rectangular matrix $A \in M_{M\times N}(\mathbb{R})$.*

To be more precise, these conditions are something very familiar, with each having its own advantages and disadvantages, and the equivalence between them, that you know well too, is not difficult to establish, the idea with all this being as follows:

(1) The algebraic conditions, although a bit abstract, prove to be in practice very useful, and they will be quite often our main workhorses here, in order to understand the linear maps. The name "algebraic conditions" comes from abstract algebra, because we can talk more generally about linear maps between abstract vector spaces $f : V \to W$, and the operations on such vector spaces being the vector sum $x+y$ and the multiplication by scalars λx, being linear simply means "preserving the algebraic structure".

(2) This is something certainly more intuitive, among other justifying the same "linear" for our maps, I mean that linear must certainly come from "line", but go see a line in

the axioms (1), personally I don't see any. In practice, while the equivalence with (1) is something clear, this condition, while intuitive and beautiful, is not very useful in practice. Also, as a technical remark, when saying in the above "maps lines to lines", by line we mean, as above, a dynamic object, with a parameter $t \in \mathbb{R}$ involved. When dropping this convention, and regarding the lines as sets, the equivalence with (1) no longer holds, and we will leave finding a counterexample here as an instructive exercise.

(3) This is also something nice, of old-style flavor, which helps understanding what is going on, and with the equivalence with (1) being clear from definitions. However, as we will see in a moment, this condition is clearly equivalent to (4) too, which is more powerful, and so in practice, our condition is somehow stuck between (1) and (4), which are both more powerful, each in its own way, and so, hard life for this condition.

(4) This is something very powerful, and a true rival to (1), usually surpassing it in power, for nearly all concrete applications. The equivalence comes via (3), because according to that condition we can write each component $f(x)_i$ as a linear combination $\sum_j A_{ij}x_j$, which according to the rules of usual matrix multiplication means $(Ax)_i$. Thus, we have our matrix $A \in M_{M\times N}(\mathbb{R})$ making the formula $f(x) = Ax$ work, as desired.

Before going further, let us record the following result, focusing on the condition (4) in Definition 1.3, and building a bit more on the equivalence with (1):

THEOREM 1.4. *The linear maps $f : \mathbb{R}^N \to \mathbb{R}^M$ are the maps of the form*

$$f(x) = Ax$$

with $A \in M_{M\times N}(\mathbb{R})$, and A can be recaptured via $A_{ij} =< f(e_j), e_i >$.

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

(1) The first assertion follows the above discussion, or even from Definition 1.3 based on the above discussion, if you prefer. In fact, in what follows, $f(x) = Ax$ will be more or less our definition for the linear maps, for most questions that we will investigate.

(2) The second assertion is something quite clear, by thinking a bit on how matrices act on vectors, and how scalar products act too. However, such sort of deep thinking often requires silence around, with not many phones ringing, or folks fighting in the subway, someone watching TV, kids crying and so on, so let me teach you a trick. In case you are working in a noisy environment, nothing beats the matrix units $e_{ij} : e_j \to e_i$, which can be utilized with zero functioning neurons or almost. As an illustration, here is how the

second assertion can be proved, by using them, without pain:

$$\begin{aligned} < f(e_j), e_i > &= < Ae_j, e_i > \\ &= \left\langle \left(\sum_{ij} A_{ij} e_{ij} \right) e_j, e_i \right\rangle \\ &= \left\langle \sum_i A_{ij} e_i, e_i \right\rangle \\ &= A_{ij} \end{aligned}$$

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

In order to understand how the correspondence $f \leftrightarrow A$ works, let us work out some examples. We have here the following statement, which is a must-know:

PROPOSITION 1.5. *The following happen:*

(1) *The rotation of angle $t \in \mathbb{R}$ is given by the following matrix:*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

(2) *The symmetry with respect to the Ox axis rotated by $t/2 \in \mathbb{R}$ is given by:*

$$S_t = \begin{pmatrix} \cos t & \sin t \\ \sin t & -\cos t \end{pmatrix}$$

(3) *The projection on the Ox axis rotated by $t/2 \in \mathbb{R}$ is given by:*

$$P_t = \frac{1}{2} \begin{pmatrix} 1 + \cos t & \sin t \\ \sin t & 1 - \cos t \end{pmatrix}$$

(4) *The projection on the all-one vector $\xi \in \mathbb{R}^N$ is given by:*

$$P = \frac{1}{N} \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$$

(5) *In fact, the projection on $\mathbb{R}x$ is given by $P = ||x||^{-2}(x_i x_j)_{ij}$.*

PROOF. We use the fact, coming from $f(x) = Ax$, or from $A_{ij} =< f(e_j), e_i >$, that the columns of A are the vectors $f(e_1), \ldots, f(e_N)$. With this in hand:

(1) This is clear by drawing a picture, the rows of R_t being the images of e_1, e_2.

(2) This is again clear on the picture, drawn with $t/2$ instead of t, as indicated.

(3) Again, picture drawn with $t/2$, as indicated, plus some easy trigonometry.

(4) This comes from $Px = A(x)\xi$, with $A(x)$ being the average of the entries of x.

(5) Consider a vector $y \in \mathbb{R}^N$. Its projection Py on the space $\mathbb{R}x$ must be a certain multiple of x, and we are led in this way to the following formula:

$$Py = \frac{< y, x >}{< x, x >} x = \frac{1}{||x||^2} < y, x > x$$

With this in hand, we can now compute the entries of P, as follows:

$$\begin{aligned} P_{ij} &= < Pe_j, e_i > \\ &= \frac{1}{||x||^2} < e_j, x >< x, e_i > \\ &= \frac{x_j x_i}{||x||^2} \end{aligned}$$

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

As another piece of general theory now, that you know well, we will need:

DEFINITION 1.6. *A linear map $f : \mathbb{R}^N \to \mathbb{R}^N$ is called diagonalizable if there exist directions $v_1, \ldots, v_N \in \mathbb{R}^N$ such that f multiplies by λ_i in the direction v_i:*

$$f(v_i) = \lambda_i v_i$$

In terms of the writing $f(x) = Ax$, we say that the corresponding matrix $A \in M_N(\mathbb{R})$ is diagonalizable, with eingenvectors v_i and eigenvalues λ_i.

Here by "directions" we mean of course linearly independent directions, and the assumptions above uniquely determine f, which follows to be given by:

$$f\left(\sum_i c_i v_i\right) = \sum_i \lambda_i c_i v_i$$

Obviously, being diagonalizable means to be "good", and being not diagonalizable means to be "bad". In order to understand this, what good and bad mean in linear algebra, let us work out some examples. We have here the following result:

PROPOSITION 1.7. *The following happen:*

(1) *The rotation R_t is not diagonalizable, unless at $t = 0$ where it is the identity, $R_0 = 1$, and at $t = \pi$ where it is minus the identity, $R_\pi = -1$.*
(2) *The symmetry S_t is diagonalizable, with eigenvectors on the symmetry axis, and on its orthogonal, with respective eigenvalues $1, -1$.*
(3) *The projection P_t is diagonalizable, with the eigenvectors exactly as for the symmetry S_t, this time with respective eigenvalues $1, 0$.*
(4) *In fact, any projection is diagonalizable, with eigenvectors on it image, and on the orthogonal of its image, with respective eigenvalues $1, 0$.*

PROOF. All this is self-explanatory and, we insist, with no need for any computation. Of course, if eager for computations, do not worry, we will have some, in this book. □

Still in relation with diagonalization, at the general level, we have:

THEOREM 1.8. *Assuming that a matrix $A \in M_N(\mathbb{R})$ is diagonalizable, with eigenvectors $v_1, \ldots, v_N$ and corresponding eigenvalues $\lambda_1, \ldots, \lambda_N$, we have*

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

with the matrices $P, D \in M_N(\mathbb{R})$ being given by the formulae

$$P = [v_1, \ldots, v_N] \quad , \quad D = diag(\lambda_1, \ldots, \lambda_N)$$

and respectively called passage matrix, and diagonal form of A.

PROOF. We have $Pe_i = v_i$, where $\{e_i\}$ is the standard basis of $\mathbb{R}^N$, and so:

$$APe_i = Av_i = \lambda_i v_i$$

On the other hand, once again by using $Pe_i = v_i$, we have as well:

$$PDe_i = P\lambda_i e_i = \lambda_i Pe_i = \lambda_i v_i$$

Thus we have $AP = PD$, and so $A = PDP^{-1}$, as claimed. □

As an illustration for this, you can have some fun with the various matrices from Proposition 1.5, with in each case the corresponding diagonalization formula $A = PDP^{-1}$ coming without much pain, because Proposition 1.7 tells us what both P, D are, in each case, and the only piece of work remaining is that of figuring out what P^{-1} is.

Summarizing, the diagonalizable matrices are the "good" ones, and their diagonalization is quite often a matter of doing some geometry. Regarding the non-diagonalizable matrices, these actually fall into two classes, "bad" and "evil". The bad ones are those which diagonalize over $\mathbb{C}$, with a main example here being the rotation R_t, and more on this later. As for the evil ones, these are evil, a basic example being as follows:

THEOREM 1.9. *The following matrix is not diagonalizable,*

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

because it has only 1 eigenvector.

PROOF. The above matrix, called J en hommage to Jordan, acts as follows:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

Thus the eigenvector/eigenvalue equation $Jv = \lambda v$ reads:

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}$$

We have then two cases, depending on λ, as follows, which give the result:

(1) For $\lambda \neq 0$ we must have $y = 0$, coming from the second row, and so $x = 0$ as well, coming from the first row, so we have no nontrivial eigenvectors.

(2) As for the case $\lambda = 0$, here we must have $y = 0$, coming from the first row, and so the eigenvectors here are the vectors of the form $\binom{x}{0}$. □

1b. Real geometry

Let us see now what we can do with our linear maps. Perhaps the simplest application, which is something of key importance for both mathematics and physics, is the classification of conics. These are the algebraic curves of degree 2 in the plane:

$$C = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \middle| P(x, y) = 0 \right\}$$

You certainly know from physics that conics appear from gravity, and more specifically, describe the trajectory of one object with respect to another. For instance, in our Solar system, all planets move on ellipses, which are of course conics, around the Sun.

But, are there any other possible trajectories, besides ellipses? And here, with physics and astronomy things are a bit complicated, because with the planets ruled out, we must carefully observe all sorts of small and capricious objects, such as comets, and asteroids. Fortunately math, and linear algebra, come to the rescue, and we first have:

PROPOSITION 1.10. *Up to non-degenerate linear transformations of the plane, which are by definition transformations as follows, assumed to be invertible,*

$$\begin{pmatrix} x \\ y \end{pmatrix} \to A \begin{pmatrix} x \\ y \end{pmatrix}$$

the conics are the circles, parabolas, hyperbolas, along with some degenerate solutions, namely $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$.

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

(1) As a first remark, it looks like we forgot the ellipses, but via linear transformations these become circles, so things fine. As a second remark, all our claimed solutions can appear. Indeed, the circles, parabolas, hyperbolas can appear as follows:

$$x^2 + y^2 = 1 \quad , \quad x^2 = y \quad , \quad xy = 1$$

As for $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$, these can appear too, as follows, and with our polynomial P chosen, whenever possible, to be of degree exactly 2:

$$x^2 = -1 \quad , \quad x^2 + y^2 = 0 \quad , \quad x^2 = 0 \quad , \quad xy = 0 \quad , \quad 0 = 0$$

Observe here that, when dealing with these degenerate cases, assuming $\deg P = 2$ instead of $\deg P \leq 2$ would only rule out $\mathbb{R}^2$ itself, which is not worth it.

(2) Getting now to the proof of our result, classification up to linear transformations, consider an arbitrary conic, written as follows, with $a, b, c, d, e, f \in \mathbb{R}$:

$$ax^2 + by^2 + cxy + dx + ey + f = 0$$

Assume first $a \neq 0$. By making a square out of ax^2, up to a linear transformation in (x, y), we can get rid of the term cxy, and we are left with:

$$ax^2 + by^2 + dx + ey + f = 0$$

In the case $b \neq 0$ we can make two obvious squares, and again up to a linear transformation in (x, y), we are left with an equation as follows:

$$x^2 \pm y^2 = k$$

In the case of positive sign, $x^2 + y^2 = k$, the solutions are the circle, when $k \geq 0$, the point, when $k = 0$, and $\emptyset$, when $k < 0$. As for the case of negative sign, $x^2 - y^2 = k$, which reads $(x-y)(x+y) = k$, here once again by linearity our equation becomes $xy = l$, which is a hyperbola when $l \neq 0$, and two lines when $l = 0$.

(3) In the case $b \neq 0$ the study is similar, with the same solutions, so we are left with the case $a = b = 0$. Here our conic is as follows, with $c, d, e, f \in \mathbb{R}$:

$$cxy + dx + ey + f = 0$$

If $c \neq 0$, by linearity our equation becomes $xy = l$, which produces a hyperbola or two lines, as explained before. As for the remaining case, $c = 0$, here our equation is:

$$dx + ey + f = 0$$

But this is generically the equation of a line, unless we are in the case $d = e = 0$, where our equation is $f = 0$, having as solutions $\emptyset$ when $f \neq 0$, and $\mathbb{R}^2$ when $f = 0$. □

As a continuation of the above, we can now formulate a final result, as follows:

THEOREM 1.11. *The conics, which are the algebraic curves of degree* 2 *in the plane,*

$$C = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \middle| P(x, y) = 0 \right\}$$

with $\deg P \leq 2$*, are up to degeneration the ellipses, parabolas and hyperbolas.*

PROOF. We already have the classification up to linear transformations, and the point is that this classification leads to the classification in general too, by applying linear transformations to the solutions that we found, with the conclusions in the statement. □

Getting back now to physics, our result predicts that there are certain objects in our Solar system moving around the Sun on parabolas, or hyperbolas. And this is indeed true, with certain asteroids doing so, and with the technical remark that these asteroids do not belong in fact to our Solar system, precisely because they are able to escape from the gravitational attraction of the Sun, on parabolic or hyperbolic trajectories.

As another physical remark, there is an interesting discussion to be made here, in relation with degeneracy, because the degenerate conics appearing from mathematics do not exactly coincide with the degenerate conics appearing from physics. For instance in mathematics we have $\emptyset$, the lines, the pairs of lines, and $\mathbb{R}^2$, which cannot appear as gravitational trajectories, while in physics we have the segment, which is the trajectory of a centered free fall, which is obviously not a conic, in a mathematical sense.

This sounds quite interesting, and as a homework for you, reader, we have:

HOMEWORK 1.12. *Fix the foundations of mathematics and physics, as for the mathematical conics to coincide with the physical conics, in the degenerate cases too.*

So long for applications of linear algebra to basic geometry and physics. We can of course use similar methods for many other geometric problems, and we will be back to this, later in this book, when discussing more in detail manifolds and geometry.

Switching topics now, no discussion about matrices and linear algebra would be complete without a word on multivariable calculus, and all the matrices appearing there. In fact, and you might already know this, this is more or less how matrices and linear algebra appeared, in our human mathematics, according to the following scheme:

$$\text{physics} \implies \text{calculus} \implies \text{matrices}$$

But probably too much talking, let us get to work, study the functions of several variables, and see where this study gets us into. And, as a first result here, which is something fundamental, getting us precisely into matrices and linear algebra, we have:

THEOREM 1.13. *The derivative of a function $f : \mathbb{R}^N \to \mathbb{R}^M$, making the formula*

$$f(x+t) \simeq f(x) + f'(x)t$$

work, is the matrix of partial derivatives at x, namely

$$f'(x) = \left(\frac{df_i}{dx_j}(x) \right)_{ij} \in M_{M\times N}(\mathbb{R})$$

acting on the vectors $t \in \mathbb{R}^N$ by usual multiplication.

PROOF. As a first observation, the formula in the statement makes sense indeed, as an equality, or rather approximation, of vectors in $\mathbb{R}^M$, as follows:

$$f\begin{pmatrix} x_1+t_1 \\ \vdots \\ x_N+t_N \end{pmatrix} \simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_1}{dx_1}(x) & \ldots & \frac{df_1}{dx_N}(x) \\ \vdots & & \vdots \\ \frac{df_M}{dx_1}(x) & \ldots & \frac{df_M}{dx_N}(x) \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

In order to prove now this formula, assuming first that we are in the case $M = 1$, the formula here, obtained via a straightforward recurrence, is as follows:

$$\begin{aligned} f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} &\simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \ldots + \frac{df}{dx_N}(x)t_N \\ &= f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \ldots & \frac{df}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \end{aligned}$$

But this gives the result in general too, by writing our function as follows:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_M \end{pmatrix}$$

Indeed, by applying our result above to each f_i, we obtain the desired formula. □

As further good news, for us linear algebraists, we have as well an important square matrix, which appears for the scalar functions, at order 2, as follows:

THEOREM 1.14. *Given a twice differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we have*

$$f(x + t) \simeq f(x) + f'(x)t + \frac{< f''(x)t, t >}{2}$$

with $f'(x) \in M_{1\times N}(\mathbb{R})$ being a row vector, and with $f''(x) \in M_N(\mathbb{R})$, given by

$$f''(x) = \left(\frac{d^2 f}{dx_i dx_j}\right)_{ij}(x)$$

being the Hessian matrix of f, at the point $x \in \mathbb{R}^N$.

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

(1) As a first remark, at $N = 1$ the Hessian matrix is the 1×1 matrix having as entry $f''(x)$, and the formula in the statement is something that we know well, namely:

$$f(x + t) \simeq f(x) + f'(x)t + \frac{f''(x)t^2}{2}$$

(2) In general now, this is in fact something which does not need a whole new proof, because it follows from the one-variable formula above, applied to the restriction of f to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x + t]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$g(r) = f(x + ry)$$

We know from (1) that the Taylor formula for g, at the point $r = 0$, reads:

$$g(r) \simeq g(0) + g'(0)r + \frac{g''(0)r^2}{2}$$

And our claim is that, with $t = ry$, this is precisely the formula in the statement.

(3) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$g'(r) = f'(x + ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned} g''(r) &= (f'(x + ry) \cdot y)' \\ &= \left(\sum_i \frac{df}{dx_i}(x + ry) \cdot y_i \right)' \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x + ry) \cdot \frac{d(x + ry)_j}{dr} \cdot y_i \\ &= \sum_i \sum_j \frac{d^2 f}{dx_i dx_j}(x + ry) \cdot y_i y_j \\ &= < f''(x + ry)y, y > \end{aligned}$$

(4) Time now to conclude. We know that we have $g(r) = f(x + ry)$, and according to our various computations above, we have the following formulae:

$$g(0) = f(x) \quad , \quad g'(0) = f'(x) \quad , \quad g''(0) =< f''(x)y, y >$$

Buit with this data in hand, the usual Taylor formula for our one variable function g, at order 2, at the point $r = 0$, takes the following form, with $t = ry$:

$$\begin{aligned} f(x + ry) &\simeq f(x) + f'(x)ry + \frac{< f''(x)y, y > r^2}{2} \\ &= f(x) + f'(x)t + \frac{< f''(x)t, t >}{2} \end{aligned}$$

Thus, we have obtained the formula in the statement. □

As a conclusion to all this, geometry and physics and analysis are all about matrices and linear algebra, or perhaps vice versa, and with the above being most likely just the tip of the iceberg. Which is good to know, and this will be our philosophy in what follows, develop the theory of matrices and linear algebra, as to get to know about the iceberg.

1c. Complex numbers

Let us discuss now what happens over the complex numbers. You certainly know about these numbers, and hopefully even love them, as any mathematician or physicist should do, with their basic theory being summarized as follows:

THEOREM 1.15. *The complex numbers, $z = a + ib$ with $a, b \in \mathbb{R}$ and with i being a formal number satisying $i^2 = -1$, form a field $\mathbb{C}$. Moreover:*

(1) *We have a field embedding $\mathbb{R} \subset \mathbb{C}$, given by $a \to a + 0 \cdot i$.*
(2) *Additively, we have $\mathbb{C} \simeq \mathbb{R}^2$, with $z = a + ib$ corresponding to (a, b).*
(3) *The length of vectors $r = |z|$, with $z = a + ib$, is given by $r = \sqrt{a^2 + b^2}$.*
(4) *With $z = r(\cos t + i \sin t)$, the products $z = z'z''$ are given by $r = r'r''$, $t = t' + t''$.*
(5) *We have the formula $e^{it} = \cos t + i \sin t$, so we can write $z = re^{it}$.*
(6) *There are N solutions to the equation $z^N = 1$, called N-th roots of unity.*
(7) *Any degree 2 equation with complex coefficients has both roots in $\mathbb{C}$.*

PROOF. We have a field, with $z^{-1} = (a - ib)/(a^2 + b^2)$, and regarding the rest:

(1-3) These assertions are clear. Observe also that we have $r^2 = z\bar{z}$, with $\bar{z} = a - ib$.

(4) We need here the formulae for the sines and cosines of sums, namely:

$$\sin(s + t) = \sin s \cos t + \cos s \sin t$$
$$\cos(s + t) = \cos s \cos t - \sin s \sin t$$

Indeed, with these formulae in hand, we have the following computation, as desired:

$$\begin{aligned} & (\cos s + i \sin s)(\cos t + i \sin t) \\ = \ & (\cos s \cos t + i^2 \sin s \sin t) + i(\sin s \cos t + \cos s \sin t) \\ = \ & (\cos s \cos t - \sin s \sin t) + i(\sin s \cos t + \cos s \sin t) \\ = \ & \cos(s + t) + i \sin(s + t) \end{aligned}$$

(5) In order to prove $e^{it} = \cos t + i \sin t$, consider the following function $f : \mathbb{R} \to \mathbb{C}$:

$$f(t) = \frac{\cos t + i \sin t}{e^{it}}$$

By using $\sin' = \cos$ and $\cos' = -\sin$, coming from the formulae in (4), we have:

$$\begin{aligned} f'(t) &= (e^{-it}(\cos t + i \sin t))' \\ &= -ie^{-it}(\cos t + i \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= e^{-it}(-i \cos t + \sin t) + e^{-it}(-\sin t + i \cos t) \\ &= 0 \end{aligned}$$

We conclude that $f : \mathbb{R} \to \mathbb{C}$ is constant, equal to $f(0) = 1$, as desired.

(6-7) These assertions both follow from (5), with $z = w^k$, with $w = e^{2\pi i/N}$ and $k = 0, 1, \ldots, N - 1$ for (6), and with $\sqrt{re^{it}} = \pm\sqrt{r}e^{it/2}$ needed for (7). □

Many interesting things can be done with the complex numbers. As a first magic result, going well beyond what we can do with the real numbers, we have:

THEOREM 1.16. *The rotation of angle $t \in \mathbb{R}$ in the plane diagonalizes as:*

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

Over the real numbers this is impossible, unless $t = 0, \pi$.

PROOF. The last assertion is something clear, that we already know, coming from the fact that at $t \neq 0, \pi$ our rotation is a "true" rotation, having no eigenvectors in the plane. Regarding the first assertion, the point is that we have the following computation:

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

We have as well a second eigenvector, as follows:

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

Thus our rotation matrix R_t is indeed diagonalizable over $\mathbb{C}$, with the passage matrix and diagonal form being, according to the above formulae, as follows:

$$P = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \quad , \quad D = \begin{pmatrix} e^{-it} & 0 \\ 0 & e^{it} \end{pmatrix}$$

Now by inverting P, we are led to the conclusion in the statement. □

Another thing that we can do with complex numbers is to nicely diagonalize the all-one matrix, that we met in Proposition 1.5. Indeed, over the reals this matrix is certainly diagonalizable, but not in a nice way, due to troubles in finding "canonical" solutions of the following equation, which is the eigenvector equation for $\lambda = 0$:

$$x_1 + \ldots + x_N = 0$$

In the complex setting, however, the roots of unity come to the rescue, via:

PROPOSITION 1.17. *The roots of unity, $\{w^k\}$ with $w = e^{2\pi i/N}$, have the property*

$$\sum_{k=0}^{N-1}(w^k)^s = N\delta_{N|s}$$

for any exponent $s \in \mathbb{N}$, where on the right we have a Kronecker symbol.

PROOF. The numbers in the statement, when written more conveniently as $(w^s)^k$ with $k = 0, \ldots, N-1$, form a certain regular polygon in the plane P_s. Thus, if we denote by C_s the barycenter of this polygon, we have the following formula:

$$\frac{1}{N}\sum_{k=0}^{N-1} w^{ks} = C_s$$

Now observe that in the case $N \not| s$ our polygon P_s is non-degenerate, circling around the unit circle, and having center $C_s = 0$. As for the case $N|s$, here the polygon is degenerate, lying at 1, and having center $C_s = 1$. Thus, we have the following formula:

$$C_s = \delta_{N|s}$$

Thus, we obtain the formula in the statement. Alternatively, the formula in the statement follows of course too by algebrically summing the sum there. □

We have the following definition, inspired by what happens in Proposition 1.17:

DEFINITION 1.18. *The Fourier matrix F_N is the following matrix, with $w = e^{2\pi i/N}$:*

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{pmatrix}$$

That is, $F_N = (w^{ij})_{ij}$, with indices $i, j \in \{0, 1, \ldots, N-1\}$, taken modulo N.

Here the terminology comes from the fact that F_N is the matrix of the Fourier transform over the cyclic group $\mathbb{Z}_N$, and more on this later in this book, when systematically discussing the discrete Fourier transform, in its various versions.

As a first example, at $N = 2$ the root of unity is $w = -1$, and with indices as above, namely $i, j \in \{0, 1\}$, taken modulo 2, our Fourier matrix is as follows:

$$F_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

At $N = 3$ now, the root of unity is $w = e^{2\pi i/3}$, and the Fourier matrix is:

$$F_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & w & w^2 \\ 1 & w^2 & w \end{pmatrix}$$

At $N=4$ now, the root of unit is $w=i$, and the Fourier matrix is:

$$F_4=\begin{pmatrix}1&1&1&1\\1&i&-1&-i\\1&-1&1&-1\\1&-i&-1&i\end{pmatrix}$$

Also, at $N=5$ the root of unity is $w=e^{2\pi i/5}$, and the Fourier matrix is:

$$F_5=\begin{pmatrix}1&1&1&1&1\\1&w&w^2&w^3&w^4\\1&w^2&w^4&w&w^3\\1&w^3&w&w^4&w^2\\1&w^4&w^3&w^2&w\end{pmatrix}$$

You get the point, with how this matrix works. Getting back now to the diagonalization problem for the all-one matrix, this can be solved, in a nice way, as follows:

THEOREM 1.19. *The all-one matrix diagonalizes as follows,*

$$\begin{pmatrix}1&\ldots&\ldots&1\\\vdots&&&\vdots\\\vdots&&&\vdots\\1&\ldots&\ldots&1\end{pmatrix}=\frac{1}{N}F_N\begin{pmatrix}N&&&0\\&0&&\\&&\ddots&\\0&&&0\end{pmatrix}F_N^*$$

with $F_N=(w^{ij})_{ij}$ being the Fourier matrix.

PROOF. We know that the all-one matrix is N times the projection on the all-one vector, so we are left with finding the 0-eigenvectors, which amounts in solving:

$$x_0+\ldots+x_{N-1}=0$$

For this purpose, we can use the root of unity $w=e^{2\pi i/N}$, and more specifically, the following standard formula, coming from Proposition 1.17:

$$\sum_{i=0}^{N-1}w^{ij}=N\delta_{j0}$$

This formula shows that for $j=1,\ldots,N-1$, the vector $v_j=(w^{ij})_i$ is a 0-eigenvector. Moreover, these vectors are pairwise orthogonal, because we have:

$$<v_j,v_k>=\sum_i w^{ij-ik}=N\delta_{jk}$$

Thus, we have our basis $\{v_1,\ldots,v_{N-1}\}$ of 0-eigenvectors, and since the N-eigenvector is $\xi=v_0$, the passage matrix P that we are looking is given by:

$$P=\begin{bmatrix}v_0&v_1&\ldots&v_{N-1}\end{bmatrix}$$

But this is precisely the Fourier matrix, $P = F_N$. In order to finish now, observe that the above computation of $< v_i, v_j >$ shows that $F_N/\sqrt{N}$ is unitary, and so:

$$F_N^{-1} = \frac{1}{N} F_N^*$$

Thus, we are led to the diagonalization formula in the statement. □

Many other things can be done with complex numbers in linear algebra. We will be back to this, on numerous occasions, in the remainder of this book.

1d. Arbitrary fields

As a continuation of the above, which led us from linear algebra over $\mathbb{R}$ to linear algebra over $\mathbb{C}$, let us discuss now linear algebra over arbitrary fields F. We have:

DEFINITION 1.20. *A field is a set F with a sum operation $+$ and a product operation $\times$, subject to the following conditions:*

(1) *$a + b = b + a$, $a + (b + c) = (a + b) + c$, there exists $0 \in F$ such that $a + 0 = 0$, and any $a \in F$ has an inverse $-a \in F$, satisfying $a + (-a) = 0$.*
(2) *$ab = ba$, $a(bc) = (ab)c$, there exists $1 \in F$ such that $a1 = a$, and any $a \neq 0$ has a multiplicative inverse $a^{-1} \in F$, satisfying $aa^{-1} = 1$.*
(3) *The sum and product are compatible via $a(b + c) = ab + ac$.*

In other words, a field satisfies what we can normally expect from "numbers", and as basic examples, we have of course $\mathbb{Q}, \mathbb{R}, \mathbb{C}$. There are many other examples of fields, along the same lines. We can talk for instance about fields like $\mathbb{Q}[\sqrt{2}]$, as follows:

PROPOSITION 1.21. *The following is an intermediate field $\mathbb{Q} \subset F \subset \mathbb{R}$,*

$$\mathbb{Q}[\sqrt{2}] = \left\{ a + b\sqrt{2} \middle| a, b \in \mathbb{Q} \right\}$$

and the same happens for any $\mathbb{Q}[\sqrt{n}]$, with $n \neq m^2$ being not a square.

PROOF. All the field axioms are clearly satisfied, except perhaps for the inversion axiom. But this axiom is satisfied too, due to the following formula:

$$\frac{1}{a + b\sqrt{2}} = \frac{a - b\sqrt{2}}{a^2 - 2b^2}$$

Observe that the denominator is indeed nonzero, due to $a^2 \neq 2b^2$, which follows by reasoning modulo 2. As for the case of $\mathbb{Q}[\sqrt{n}]$ with $n \neq m^2$, this is similar. □

The above result is quite interesting, obviously in relation with arithmetic, and suggests looking into the intermediate fields of numbers, as follows:

$$\mathbb{Q} \subset F \subset \mathbb{C}$$

Getting back now to generalities, the simplest example of field appears to be $\mathbb{Q}$. However, this is not exactly true, because the numbers $0, 1$, whose presence in a field is mandatory, $0, 1 \in F$, can form themselves a field, with structure as follows:

$$1 + 1 = 0$$

To be more precise, according to our field axioms, all operations of type $a * b$ with $a, b = 0, 1$ are uniquely determined, except for $1+1$. You would say that we must normally set $1 + 1 = 2$, with $2 \neq 0$ being a new field element, but the point is that $1 + 1 = 0$ is something natural too, this being the addition modulo 2. And, what we get is a field:

$$\mathbb{F}_2 = \{0, 1\}$$

Let us summarize this finding, along with a bit more, as follows:

PROPOSITION 1.22. *$\mathbb{Q}$ is the simplest field having the property $1 + \ldots + 1 \neq 0$, in the sense that any field F satisfying this condition must contain $\mathbb{Q}$:*

$$\mathbb{Q} \subset F$$

However, in general this fails, for instance for the field $\mathbb{F}_2 = \{0, 1\}$, with addition $1+1 = 0$, and more generally for the field $\mathbb{F}_p$ formed by the integers modulo p, with p prime.

PROOF. Here the first assertion is clear, because $1 + \ldots + 1 \neq 0$ tells us that we have an embedding $\mathbb{N} \subset F$, and then by taking inverses with respect to $+$ and $\times$ we obtain $\mathbb{Q} \subset F$. As for the second assertion, this follows from the above discussion. □

All this is quite intriguing, and is certainly worth some more study. We first have:

THEOREM 1.23. *Given a field F, define its characteristic $p = char(F)$ as being the smallest $p \in \mathbb{N}$ such that the following happens, and as $p = 0$, if this never happens:*

$$\underbrace{1 + \ldots + 1}_{p \text{ times}} = 0$$

Then, assuming $p > 0$, this number p must be prime, we have a field embedding $\mathbb{F}_p \subset F$, and $q = |F|$ must be of the form $q = p^k$, with $k \in \mathbb{N}$. Also, we have the formulae

$$(a + b)^p = a^p + b^p \quad , \quad a^q = a$$

valid for any $a, b \in F$, and the Fermat poynomial $X^q - X$ factorizes as:

$$X^q - X = \prod_{a \in F} (X - a)$$

Also, regardless of p, any finite multiplicative subgroup $G \subset F - \{0\}$ must be cyclic.

PROOF. This is a quite crowded statement, containing a lot of information, but the idea is that all this comes from some elementary arithmetic, as follows:

(1) The various assertions in the beginning, regarding the characteristic $p = char(F)$ and its basic properties, all follow from definitions and from some quick thinking, based on formulae of type $(1 + \ldots + 1)(1 + \ldots + 1) = 1 + \ldots + 1$, inside the field F.

(2) We can also see that the sums $1 + \ldots + 1$, and their quotients, form a minimal subfield $E \subset F$, called prime field. At $p = 0$ we have $E = \mathbb{Q}$. At $p > 0$ we have $E = \mathbb{F}_p$, and $q = |F|$ is given by $q = p^k$, with $k \in \mathbb{N}$ being the dimension of F over E.

(3) The baby Fermat formula $(a + b)^p = a^p + b^p$, which reminds the Fermat little theorem, $a^p = a(p)$ over $\mathbb{Z}$, follows in the same way, namely from the binomial formula, because all the non-trivial binomial coefficients $\binom{p}{s}$ are multiples of p.

(4) As for the Fermat formula $a^q = a$ itself, which implies the assertion about $X^q - X$, this follows from the last assertion, which can be proved via some basic arithmetic inside F, and which for $G = F - \{0\}$ itself, with $|F| = q$, gives $a^{q-1} = 1$, for any $a \neq 0$. □

The above result raises a lot of questions. First, we have the question of understanding the finite fields, $|F| = q < \infty$, with $q = p^k$. And here, we have the following result:

THEOREM 1.24. *Associated to any prime power $q = p^k$ is a certain field $\mathbb{F}_q$, having q elements, obtained as splitting field for the polynomial*

$$P = X^q - X$$

and we obtain in this way all the finite fields.

PROOF. As explained in the above, associated to any finite field F is its characteristic, the smallest number $p \in \mathbb{N}$ such that the following equality happens, inside F:

$$\underbrace{1 + \ldots + 1}_{p \text{ times}} = 0$$

Moreover, it is easy to see that p must be prime, and this leads us into the classification of the finite fields F having characteristic p. But here, at $|F| = p$ we clearly have $\mathbb{F}_p$ as only solution, say because the group $(F, +)$ must be cyclic of order p, and more generally, at $|F| = p^k$ with $k \in \mathbb{N}$, we have the field $\mathbb{F}_q$ in the statement as the only solution. □

Changing topics, in relation with prime numbers and arithmetic, we have as well the following construction, producing a characteristic 0 field, as we like them:

THEOREM 1.25. *Given a prime number p, and a rational number $a/b \in \mathbb{Q}$, we can write $a/b = p^k(c/d)$ with c, d prime to p, and set:*

$$\left|\frac{a}{b}\right| = p^{-k}$$

This function is not exactly a norm, but the following function is a distance on $\mathbb{Q}$,

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

and the completion of $\mathbb{Q}$ with respect to this distance is the field of p-adic numbers $\mathbb{Q}_p$.

PROOF. This is indeed something very standard, the idea being that $|a/b| = p^{-k}$ is not exactly a norm, but satisfies the following conditions, which are even better, and which are elementary to establish, by using some basic arithmetic modulo p:

(1) First axiom: $|x| \geq 0$, with $|x| = 0$ when $x = 0$.

(2) Modified second axiom: $|xy| = |x| \cdot |y|$.

(3) Strong triangle inequality: $|x + y| \leq \max(|x|, |y|)$.

Now with these conditions in hand, it is clear that $d(x, y) = |x-y|$ is indeed a distance on $\mathbb{Q}$. Then, we can perform the completion procedure in the statement, with this being quite similar to the completion procedure which produces $\mathbb{R}$ out of $\mathbb{Q}$. □

Many things can be said about the field of p-adic numbers $\mathbb{Q}_p$, its subring of p-adic integers $\mathbf{Z}_p \subset \mathbb{Q}_p$, and its algebraic completion $\mathbb{Q}_p \subset \bar{\mathbb{Q}}_p$ too, including of course fighting for notations for these objects. We will be back to this later in this book.

Finally, as a further remark here, with our field theory we are not at all away from analysis, quite the opposite. Indeed, while the usual spaces of functions are obviously not fields, analysis remains around the corner, due to the following basic fact:

THEOREM 1.26. *The quotients of complex polynomials, called rational functions, when written in reduced form, as follows, with P, Q prime to each other,*

$$f = \frac{P}{Q}$$

are well-defined and continuous outside the zeroes $P_f \subset \mathbb{C}$ of Q, called poles of f:

$$f : \mathbb{C} - P_f \to \mathbb{C}$$

These functions are stable under summing, making products and taking inverses,

$$\frac{P}{Q} + \frac{R}{S} = \frac{PS + QR}{QS} \quad , \quad \frac{P}{Q} \cdot \frac{R}{S} = \frac{PR}{QS} \quad , \quad \left(\frac{P}{Q}\right)^{-1} = \frac{Q}{P}$$

so they form a field $\mathbb{C}(X)$, called field of rational functions.

PROOF. Almost everything here is clear from definitions, and with the comment that, in what regards the term "pole", this does not come from the Poles who invented this, but rather from the fact that, when trying to draw the graph of f, or rather imagine that graph, which takes place in $2+2 = 4$ real dimensions, we are faced with some sort of tent, which is suspended by infinite poles, which lie, guess where, at the poles of f. □

Again, more on this later in this book, at the appropriate analysis chapter, when discussing rational functions along with complex, holomorphic and harmonic ones.

Now, let us do some naive geometry, say over $\mathbb{F}_q$. However, things are a bit bizarre here, and we have for instance the following result, to start with:

PROPOSITION 1.27. *The circle of radius zero $x^2+y^2=0$ over $\mathbb{F}_p$ is as follows:*

(1) *At $p=2$, this has 2 points.*
(2) *At $p=1(4)$, this has $2p-1$ points.*
(3) *At $p=3(4)$, this has 1 point.*

PROOF. Our circle $x^2+y^2=0$ is formed by the point $(0,0)$, and then of the solutions of $x^2+y^2=0$, with $x,y\neq 0$. But this latter equation is equivalent to $(x/y)^2+1=0$, and so to $(x/y)^2=-1$, so the number of points of our circle is:

$$N=1+(p-1)\#\{r|r^2=-1\}$$

But at $p=2$ this gives $N=1+1\times 1=2$, then at $p=1(4)$ this gives $N=1+(p-1)\times 2=2p-1$, and finally at $p=3(4)$ this gives $N=1+(p-1)\times 0=1$. □

When looking at more general conics, still over finite fields $\mathbb{F}_q$, things do not necessarily improve, and we have some other bizarre results, along the same lines, such as:

THEOREM 1.28. *Any curve over $\mathbb{F}_2$ is a conic. However, this is not the case for $\mathbb{F}_p$ with $p\geq 3$.*

PROOF. This is again something elementary, as follows:

(1) Let us find the conics over $\mathbb{F}_2$. These are given by equations as follows:

$$ax^2+by^2+cxy+dx+ey+f=0$$

Since $x^2=x$ holds in $\mathbb{F}_2$, the first 2 terms dissapear, and we are left with:

$$cxy+dx+ey+f=0$$

– The first case, $c=0$, corresponds to the lines over $\mathbb{F}_2$. But there are 8 such lines, all distinct, given by $r=0$, $x=r$, $y=r$, $x+y=r$, with $r=0,1$.

– The second case, $c\neq 0$, corresponds to the non-degenerate conics over $\mathbb{F}_2$. But there are 8 such conics, all distinct, and distinct as well from the 8 lines found above, given by $xy=r$, $x(y+1)=r$, $(x+1)y=r$, $(x+1)(y+1)=r$, with $r=0,1$.

Summarizing, we have $8+8=16$ conics over $\mathbb{Z}_2$. But since the plane $\mathbb{F}_2\times\mathbb{F}_2$ has $2\times 2=4$ points, there are $2^4=16$ possible curves. Thus, all the curves are conics.

(2) Regarding now $\mathbb{F}_p$ with $p\geq 3$, here the plane $\mathbb{F}_p\times\mathbb{F}_p$ has p^2 points, so there are 2^{p^2} curves. Among these curves, the conics are given by equations as follows:

$$ax^2+by^2+cxy+dx+ey+f=0$$

Thus, we have at most p^6 conics, and since we have $2^{p^2}>p^6$ for any $p\geq 4$, we are done with the case $p\geq 5$. In the remaining case now, $p=3$, the $3^6=729$ possible conics split into the $2^5=243$ ones with $a=0$, and the $2\times 243=486$ ones with $a\neq 0$. But these latter conics appear twice, as we can see by dividing everything by a, and so there

are only $1 \times 243 = 243$ of them. Thus, we have at most $243 + 243 = 486$ conics, and this is smaller than the number of curves of $\mathbb{F}_3 \times \mathbb{F}_3$, which is $2^9 = 512$, as desired. □

Summarizing, better stay away from characteristic p. And this is what we will do below, where our mathematics will be mostly over $\mathbb{R}, \mathbb{C}$, or other characteristic 0 fields.

1e. Exercises

This was an elementary chapter, for the most concerned with things that you are supposed to know, and as exercises, for making sure that you know indeed, we have:

EXERCISE 1.29. *Check the formulae given above of the plane rotations R_t, plane symmetries S_t, and plane projections P_t.*

EXERCISE 1.30. *Diagonalize the plane symmetries S_t and the plane projections P_t, geometrically, without any algebraic computations.*

EXERCISE 1.31. *Prove that the conics are, modulo some degenerate cases, exactly the curves which appear by cutting a two-sided cone with a plane.*

EXERCISE 1.32. *Compute a few first derivatives, and a few Hessian matrices too, for some multivariable functions of your choice.*

EXERCISE 1.33. *Further meditate on the complex numbers, on the various ways of introducing and denoting them, and on the formula $e^{it} = \cos t + i \sin t$.*

EXERCISE 1.34. *Learn more about the Fourier matrices and their properties, and learn as well about the complex Hadamard matrices, which generalize them.*

EXERCISE 1.35. *Learn more about finite fields, about the Fermat polynomial, about splitting fields for polynomials, and about the fields $\mathbb{F}_q$.*

EXERCISE 1.36. *Learn more about the p-adic numbers, including full details on their construction, and what can be done with them.*

As bonus exercise, of genuine linear algebra type, find some other matrices that you can diagonalize geometrically, besides the 2×2 matrices R_t, S_t, P_t discussed above.

CHAPTER 2

Matrix theory

2a. Matrix inversion

We have seen so far that most of the interesting maps $f : \mathbb{R}^N \to \mathbb{R}^N$ that we know, such as the rotations, symmetries and projections, are linear, and can be written in the following form, with $A \in M_N(\mathbb{R})$ being a square matrix:

$$f(v) = Av$$

We develop now more general theory for such linear maps. We will be interested in the question of inverting the linear maps $f : \mathbb{R}^N \to \mathbb{R}^N$. And the point is that this is the same question as inverting the corresponding matrices $A \in M_N(\mathbb{R})$, due to:

THEOREM 2.1. *A linear map $f : \mathbb{R}^N \to \mathbb{R}^N$, written as*

$$f(v) = Av$$

is invertible precisely when A is invertible, and in this case we have $f^{-1}(v) = A^{-1}v$.

PROOF. This is something that we basically know, coming from the fact that, with the notation $f_A(v) = Av$, we have the following formula:

$$f_A f_B = f_{AB}$$

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

In order to study invertibility questions, for matrices or linear maps, let us begin with some examples. In the simplest case, in 2 dimensions, the result is as follows:

THEOREM 2.2. *We have the following inversion formula, for the 2×2 matrices:*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

When $ad - bc = 0$, the matrix is not invertible.

PROOF. We have two assertions to be proved, the idea being as follows:

(1) As a first observation, when $ad - bc = 0$ we must have, for some $\lambda \in \mathbb{R}$:

$$b = \lambda a \quad , \quad d = \lambda c$$

Thus our matrix must be of the following special type:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & \lambda a \\ a & \lambda c \end{pmatrix}$$

But in this case the columns are proportional, and so the linear map associated to the matrix is not invertible, and so the matrix itself is not invertible either.

(2) When $ad - bc \neq 0$, let us look for an inversion formula of the following type:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} * & * \\ * & * \end{pmatrix}$$

We must therefore solve the following equations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} * & * \\ * & * \end{pmatrix} = \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix}$$

But the obvious solution here is as follows:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix}$$

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

In order to deal now with the inversion problem in general, for the arbitrary matrices $A \in M_N(\mathbb{R})$, we will use the same method as the one above, at $N = 2$. Let us write indeed our matrix as follows, with $v_1, \ldots, v_N \in \mathbb{R}^N$ being its column vectors:

$$A = [v_1, \ldots, v_N]$$

We know that, in order for our matrix A to be invertible, its column vectors $v_1, \ldots, v_N$ must be linearly independent. Thus, we are led into the question of understanding when a family of vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ are linearly independent. Now in order to deal with this latter question, let us introduce the following notion:

DEFINITION 2.3. *Associated to any vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ is the volume*

$$\det{}^+(v_1 \ldots v_N) = vol < v_1, \ldots, v_N >$$

of the parallelepiped made by these vectors.

Here the volume is taken in the standard N-dimensional sense. At $N = 1$ this volume is a length, at $N = 2$ this volume is an area, at $N = 3$ this is the usual 3D volume, and so on. In general, the volume of a body $X \subset \mathbb{R}^N$ is by definition the number $vol(X) \in [0, \infty]$ of copies of the unit cube $C \subset \mathbb{R}^N$ which are needed for filling X. Now with this notion in hand, in relation with our inversion problem, we have the following statement:

PROPOSITION 2.4. *The quantity* $\det^+$ *that we constructed, regarded as a function of the corresponding square matrices, formed by column vectors,*

$$\det^+ : M_N(\mathbb{R}) \to \mathbb{R}_+$$

has the property that a matrix $A \in M_N(\mathbb{R})$ *is invertible precisely when* $\det^+(A) > 0$.

PROOF. This follows from the fact that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when its column vectors $v_1, \ldots, v_N \in \mathbb{R}^N$ are linearly independent. But this latter condition is equivalent to the fact that we must have the following strict inequality:

$$vol < v_1, \ldots, v_N >> 0$$

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

Summarizing, all this leads us into the explicit computation of $\det^+$. As a first observation, in 1 dimension we obtain the absolute value of the real numbers:

$$\det^+(a) = |a|$$

In 2 dimensions now, the computation is non-trivial, and we have the following result, making the link with our main result so far, namely Theorem 2.2:

THEOREM 2.5. *In* 2 *dimensions we have the following formula,*

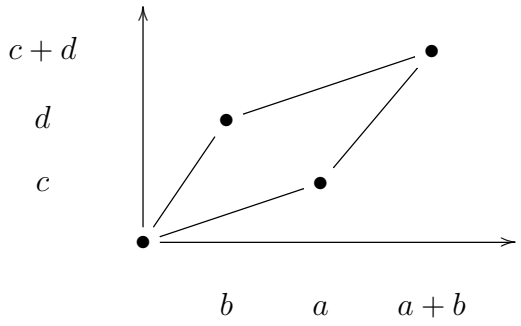
$$\det^+ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = |ad - bc|$$

with $\det^+ : M_2(\mathbb{R}) \to \mathbb{R}_+$ *being the function constructed above.*

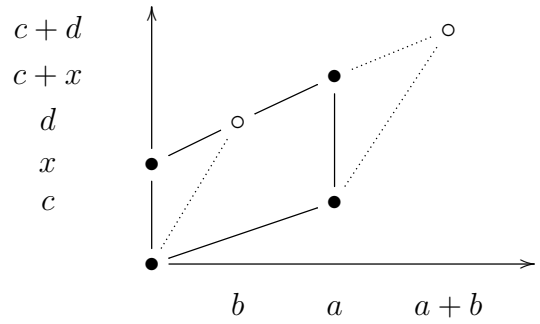
PROOF. We must show that the area of the parallelogram formed by $\binom{a}{c}, \binom{b}{d}$ equals $|ad - bc|$. We can assume $a, b, c, d > 0$ for simplifying, the proof in general being similar. Moreover, by switching if needed the vectors $\binom{a}{c}, \binom{b}{d}$, we can assume that we have:

$$\frac{a}{c} > \frac{b}{d}$$

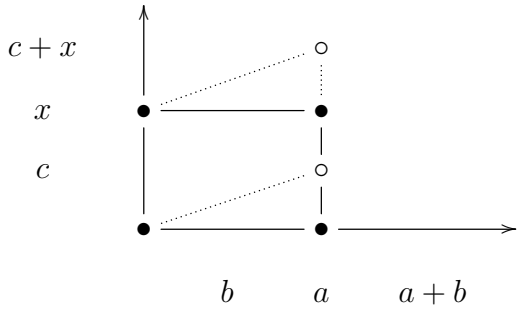
According to these conventions, the picture of our parallelogram is as follows:



Now let us slide the upper side downwards left, until we reach the Oy axis. Our parallelogram, which has not changed its area in this process, becomes:



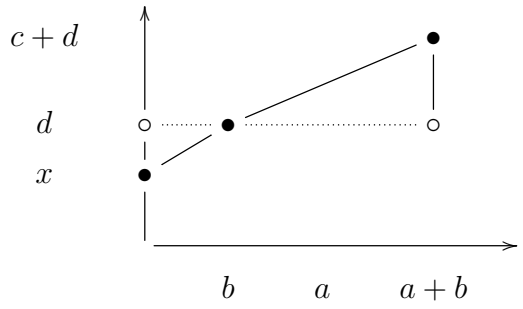
We can further modify this parallelogram, once again by not altering its area, by sliding the right side downwards, until we reach the Ox axis:



Let us compute now the area. Since our two sliding operations have not changed the area of the original parallelogram, this area is given by:

$$A = ax$$

In order to compute the quantity x, observe that in the context of the first move, we have two similar triangles, according to the following picture:



Thus, we are led to the following equation for the number x:

$$\frac{d-x}{b}=\frac{c}{a}$$

By solving this equation, we obtain the following value for x:

$$x=d-\frac{bc}{a}$$

Thus the area of our parallelogram, or rather of the final rectangle obtained from it, which has the same area as the original parallelogram, is given by:

$$A=ax=ad-bc$$

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

All this is very nice, and obviously we have a beginning of theory here. However, when looking carefully, we can see that our theory has a weakness, because:

(1) In 1 dimension the number a, which is the simplest function of a itself, is certainly a better quantity than the number $|a|$.

(2) In 2 dimensions the number $ad-bc$, which is linear in a,b,c,d, is certainly a better quantity than the number $|ad-bc|$.

So, let us upgrade now our theory, by constructing a better function, which takes signed values. In order to do this, we must come up with a way of splitting the systems of vectors $v_1,\ldots,v_N\in\mathbb{R}^N$ into two classes, say positive and negative. And here, the answer is quite clear, because a bit of thinking leads to the following definition:

DEFINITION 2.6. *A system of vectors $v_1,\ldots,v_N\in\mathbb{R}^N$ is called:*

(1) *Oriented, if one can continuously pass from the standard basis to it.*
(2) *Unoriented, otherwise.*

The associated sign is $+$ in the oriented case, and $-$ in the unoriented case.

As a first example, in 1 dimension the basis consists of the single vector $e=1$, which can be continuously deformed into any vector $a>0$. Thus, the sign is the usual one:

$$sgn(a)=\begin{cases}+ & \text{if } a>0\\ - & \text{if } a<0\end{cases}$$

Thus, in connection with our original question, we are definitely on the good track, because when multiplying $|a|$ by this sign we obtain a itself, as desired:

$$a=sgn(a)|a|$$

In 2 dimensions now, the explicit formula of the sign is as follows:

PROPOSITION 2.7. *We have the following formula, valid for any 2 vectors in $\mathbb{R}^2$,*

$$sgn\left[\begin{pmatrix}a\\c\end{pmatrix},\begin{pmatrix}b\\d\end{pmatrix}\right]=sgn(ad-bc)$$

with the sign function on the right being the usual one, in 1 dimension.

PROOF. According to our conventions, the sign of $\binom{a}{c},\binom{b}{d}$ is as follows:

(1) The sign is $+$ when these vectors come in this order with respect to the counterclockwise rotation in the plane, around 0.

(2) The sign is $-$ otherwise, meaning when these vectors come in this order with respect to the clockwise rotation in the plane, around 0.

If we assume now $a,b,c,d>0$ for simplifying, we are left with comparing the angles having the numbers c/a and d/b as tangents, and we obtain in this way:

$$sgn\left[\begin{pmatrix}a\\c\end{pmatrix},\begin{pmatrix}b\\d\end{pmatrix}\right]=\begin{cases}+ & \text{if } \frac{c}{a}<\frac{d}{b}\\ - & \text{if } \frac{c}{a}>\frac{d}{b}\end{cases}$$

But this gives the formula in the statement. The proof in general is similar. □

Once again, in connection with our original question, we are on the good track, because when multiplying $|ad-bc|$ by this sign we obtain $ad-bc$ itself, as desired:

$$ad-bc=sgn(ad-bc)|ad-bc|$$

At the level of the general results now, we have:

PROPOSITION 2.8. *The orientation of a system of vectors changes as follows:*

(1) *If we switch the sign of a vector, the associated sign switches.*
(2) *If we permute two vectors, the associated sign switches as well.*

PROOF. Both these assertions are clear from the definition of the sign, because the two operations in question change the orientation of the system of vectors. □

With the above notion in hand, we can now formulate:

DEFINITION 2.9. *The determinant of $v_1,\ldots,v_N\in\mathbb{R}^N$ is the signed volume*

$$\det(v_1\ldots v_N)=\pm vol<v_1,\ldots,v_N>$$

of the parallelepiped made by these vectors.

In other words, we are upgrading here Definition 2.3, by adding a sign to the quantity $\det^+$ constructed there, as to potentially reach to good additivity properties:

$$\det(v_1\ldots v_N)=\pm\det^+(v_1\ldots v_N)$$

In relation with our original inversion problem for the square matrices, this upgrade does not change what we have so far, and we have the following statement:

THEOREM 2.10. *The quantity* det *that we constructed, regarded as a function of the corresponding square matrices, formed by column vectors,*

$$\det : M_N(\mathbb{R}) \to \mathbb{R}$$

has the property that a matrix $A \in M_N(\mathbb{R})$ *is invertible precisely when* $\det(A) \neq 0$.

PROOF. We know from the above that a matrix $A \in M_N(\mathbb{R})$ is invertible precisely when $\det^+(A) = |\det A|$ is strictly positive, and this gives the result. □

Let us try now to compute the determinant. In 1 dimension we have of course the formula $\det(a) = a$, because the absolute value fits, and so does the sign:

$$\det(a) = sgn(a) \times |a| = a$$

In 2 dimensions now, we have the following result:

THEOREM 2.11. *In* 2 *dimensions we have the following formula,*

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

with $|\,.\,| = \det$ *being the determinant function constructed above.*

PROOF. According to our definition, to the computation in Theorem 2.5, and to the sign formula from Proposition 2.7, the determinant of a 2×2 matrix is given by:

$$\begin{aligned} \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} &= sgn\left[\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix}\right] \times \det{}^+ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ &= sgn\left[\begin{pmatrix} a \\ c \end{pmatrix}, \begin{pmatrix} b \\ d \end{pmatrix}\right] \times |ad - bc| \\ &= sgn(ad - bc) \times |ad - bc| \\ &= ad - bc \end{aligned}$$

Thus, we have obtained the formula in the statement. □

2b. The determinant

In order to discuss now arbitrary dimensions, we will need a number of theoretical results. Here is a first series of formulae, coming straight from the definitions:

THEOREM 2.12. *The determinant has the following properties:*

(1) *When multiplying by scalars, the determinant gets multiplied as well:*

$$\det(\lambda_1 v_1, \ldots, \lambda_N v_N) = \lambda_1 \ldots \lambda_N \det(v_1, \ldots, v_N)$$

(2) *When permuting two columns, the determinant changes the sign:*

$$\det(\ldots, u, \ldots, v, \ldots) = -\det(\ldots, v, \ldots, u, \ldots)$$

(3) *The determinant* $\det(e_1, \ldots, e_N)$ *of the standard basis of* $\mathbb{R}^N$ *is* 1.

PROOF. All this is clear from definitions, as follows:

(1) This follows from definitions, and from Proposition 2.8 (1).

(2) This follows as well from definitions, and from Proposition 2.8 (2).

(3) This is clear from our definition of the determinant. □

As an application of the above result, we have:

THEOREM 2.13. *The determinant of a diagonal matrix is given by:*

$$\begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} = \lambda_1 \ldots \lambda_N$$

That is, we obtain the product of diagonal entries, or of eigenvalues.

PROOF. The above formula is clear by using Theorem 2.12, which gives:

$$\begin{vmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{vmatrix} = \lambda_1 \ldots \lambda_N \begin{vmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{vmatrix} = \lambda_1 \ldots \lambda_N$$

As for the last assertion, this is rather a remark. □

In order to reach to a more advanced theory, let us adopt now the linear map point of view. In this setting, the definition of the determinant reformulates as follows:

THEOREM 2.14. *Given a linear map, written as $f(v) = Av$, its "inflation coefficient", obtained as the signed volume of the image of the unit cube, is given by:*

$$I_f = \det A$$

More generally, I_f is the inflation ratio of any parallelepiped in $\mathbb{R}^N$, via the transformation f. In particular f is invertible precisely when $\det A \neq 0$.

PROOF. The only non-trivial thing in all this is the fact that the inflation coefficient I_f, as defined above, is independent of the choice of the parallelepiped. But this is a generalization of the Thales theorem, which follows from the Thales theorem itself. □

As a first application of the above linear map viewpoint, we have:

THEOREM 2.15. *We have the following formula, valid for any matrices A, B:*

$$\det(AB) = \det A \cdot \det B$$

In particular, we have $\det(AB) = \det(BA)$.

PROOF. The first formula follows from the formula $f_{AB} = f_A f_B$ for the associated linear maps. As for $\det(AB) = \det(BA)$, this is clear from the first formula. □

Getting back now to explicit computations, we have the following key result:

THEOREM 2.16. *The determinant of a diagonalizable matrix*

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

is the product of its eigenvalues, $\det A = \lambda_1 \ldots \lambda_N$.

PROOF. We know that a diagonalizable matrix can be written in the form $A = PDP^{-1}$, with $D = diag(\lambda_1, \ldots, \lambda_N)$. Now by using Theorem 2.15, we obtain:

$$\begin{aligned} \det A &= \det(PDP^{-1}) \\ &= \det(DP^{-1}P) \\ &= \det D \\ &= \lambda_1 \ldots \lambda_N \end{aligned}$$

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

In general now, at the theoretical level, we have the following key result:

THEOREM 2.17. *The determinant has the additivity property*

$$\det(\ldots, u + v, \ldots) = \det(\ldots, u, \ldots) + \det(\ldots, v, \ldots)$$

valid for any choice of the vectors involved.

PROOF. This follows by doing some elementary geometry, in the spirit of the computations in the proof of Theorem 2.5, with several methods being available, as follows:

(1) We can either use the Thales theorem, and then compute the volumes of all the parallelepipeds involved, by using basic algebraic formulae.

(2) Or we can solve the problem in "puzzle" style, the idea being to cut the big parallelepiped, and then recover the small ones, after some manipulations.

(3) We can do as well something hybrid, consisting in deforming the parallelepipeds involved, without changing their volumes, and then cutting and gluing. □

As a basic application of the above result, we have:

THEOREM 2.18. *We have the following results:*

(1) *The determinant of a diagonal matrix is the product of diagonal entries.*
(2) *The same is true for the upper triangular matrices.*
(3) *The same is true for the lower triangular matrices.*

PROOF. All this can be deduced by using our various general formulae, as follows:

(1) This is something that we already know, from Theorem 2.16.

(2) This follows by using our various formulae, then (1), as follows:

$$\begin{vmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{vmatrix} = \begin{vmatrix} \lambda_1 & 0 & & * \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{vmatrix}$$

$$\vdots$$

$$\vdots$$

$$= \begin{vmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{vmatrix}$$

$$= \lambda_1 \ldots \lambda_N$$

(3) This follows as well from our various formulae, then (1), by proceeding this time from right to left, from the last column towards the first column. □

As an important theoretical result now, we have:

THEOREM 2.19. *The determinant of square matrices is the unique map*

$$\det : M_N(\mathbb{R}) \to \mathbb{R}$$

satisfying the conditions found above.

PROOF. Any map $\det' : M_N(\mathbb{R}) \to \mathbb{R}$ satisfying our conditions must indeed coincide with det on the upper triangular matrices, and then on all the matrices. □

Here is now another important theoretical result:

THEOREM 2.20. *The determinant is subject to the row expansion formula*

$$\begin{vmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & \ldots & a_{2N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} & \ldots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N3} & \ldots & a_{NN} \end{vmatrix}$$

$$+ \ldots\ldots + (-1)^{N+1} a_{1N} \begin{vmatrix} a_{21} & \ldots & a_{2,N-1} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{N,N-1} \end{vmatrix}$$

and this method fully computes it, by recurrence.

PROOF. This follows indeed from the fact that the above formula produces a certain function $\det : M_N(\mathbb{R}) \to \mathbb{R}$, which has the properties required by Theorem 2.19. □

We can expand as well over the columns, as follows:

THEOREM 2.21. *The determinant is subject to the column expansion formula*

$$\begin{vmatrix} a_{11} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N1} & \ldots & a_{NN} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & \ldots & a_{2N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & \ldots & a_{1N} \\ a_{32} & \ldots & a_{3N} \\ \vdots & & \vdots \\ a_{N2} & \ldots & a_{NN} \end{vmatrix}$$
$$+ \ldots\ldots + (-1)^{N+1} a_{N1} \begin{vmatrix} a_{12} & \ldots & a_{1N} \\ \vdots & & \vdots \\ a_{N-1,2} & \ldots & a_{N-1,N} \end{vmatrix}$$

and this method fully computes it, by recurrence.

PROOF. This follows indeed by using the same argument as for the rows. □

As a first application of the above methods, we can now prove:

THEOREM 2.22. *The determinant of the* 3×3 *matrices is given by*

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh$$

which can be memorized by using Sarrus' triangle method, "triangles parallel to the diagonal, minus triangles parallel to the antidiagonal".

PROOF. Here is the computation, by using the above results:

$$\begin{aligned} \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} &= a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= aei - afh - bdi + bfg + cdh - ceg \\ &= aei + bfg + cdh - ceg - bdi - afh \end{aligned}$$

Thus, we obtain the formula in the statement. □

Let us discuss now the general formula of the determinant, at arbitrary values $N \in \mathbb{N}$ of the matrix size, generalizing the formulae that we have at $N = 2, 3$. We will need:

DEFINITION 2.23. *A permutation of* $\{1, \ldots, N\}$ *is a bijection, as follows:*

$$\sigma : \{1, \ldots, N\} \to \{1, \ldots, N\}$$

The set of such permutations is denoted S_N.

There are many possible notations for the permutations, the simplest one consisting in writing the numbers $1, \ldots, N$, and below them, their permuted versions:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 5 & 3 \end{pmatrix}$$

Another method, which is better for most purposes, and faster too, remember that time is money, is by denoting permutations as diagrams, going from top to bottom:

$\sigma =$ 

There are many interesting things that can be said about permutations. In what concerns us, we will need the following key result:

THEOREM 2.24. *The permutations have a signature function*

$$\varepsilon : S_N \to \{\pm 1\}$$

which can be defined in the following equivalent ways:

(1) *As $(-1)^c$, where c is the number of inversions.*
(2) *As $(-1)^t$, where t is the number of transpositions.*
(3) *As $(-1)^o$, where o is the number of odd cycles.*
(4) *As $(-1)^x$, where x is the number of crossings.*
(5) *As the sign of the corresponding permuted basis of $\mathbb{R}^N$.*

PROOF. Let us begin with the precise definition of c, t, o, x, as numbers modulo 2:

(1) The idea here is that given any two numbers $i < j$ among $1, \ldots, N$, the permutation can either keep them in the same order, $\sigma(i) < \sigma(j)$, or invert them:

$$\sigma(j) > \sigma(i)$$

Now by making $i < j$ vary over all pairs of numbers in $1, \ldots, N$, we can count the number of inversions, and call it c. This is an integer, $c \in \mathbb{N}$, which is well-defined.

(2) Here the idea, which is something quite intuitive, is that any permutation appears as a product of switches, also called transpositions:

$$i \leftrightarrow j$$

The decomposition as a product of transpositions is not unique, but the number t of the needed transpositions is unique, when considered modulo 2. This follows for instance from the equivalence of (2) with (1,3,4,5), explained below.

(3) Here the point is that any permutation decomposes, in a unique way, as a product of cycles, which are by definition permutations of the following type:

$$i_1 \to i_2 \to i_3 \to \ldots\ldots \to i_k \to i_1$$

Some of these cycles have even length, and some others have odd length. By counting those having odd length, we obtain a well-defined number $o \in \mathbb{N}$.

(4) Here the method is that of drawing the permutation, as we usually do, and by avoiding triple crossings, and then counting the number of crossings. This number x depends on the way we draw the permutations, but modulo 2, we always get the same number. Indeed, this follows from the fact that we can continuously pass from a drawing to each other, and that when doing so, the number of crossings can only jump by ± 2.

Summarizing, we have 4 different definitions for the signature of the permutations, which all make sense, constructed according to (1-4) above. Regarding now the fact that we always obtain the same number, this can be established as follows:

(1)=(2) This is clear, because any transposition inverts once, modulo 2.

(1)=(3) This is clear as well, because the odd cycles invert once, modulo 2.

(1)=(4) This comes from the fact that the crossings correspond to inversions.

(2)=(3) This follows by decomposing the cycles into transpositions.

(2)=(4) This comes from the fact that the crossings correspond to transpositions.

(3)=(4) This follows by drawing a product of cycles, and counting the crossings.

Finally, in what regards the equivalence of all these constructions with (5), here simplest is to use (2). Indeed, we already know that the sign of a system of vectors switches when interchanging two vectors, and so the equivalence between (2,5) is clear. □

Now back to linear algebra, we can formulate a key result, as follows:

THEOREM 2.25. *We have the following formula for the determinant,*

$$\det A = \sum_{\sigma \in S_N} \varepsilon(\sigma) A_{1\sigma(1)} \ldots A_{N\sigma(N)}$$

with the signature function being the one introduced above.

PROOF. This follows by recurrence over $N \in \mathbb{N}$, as follows:

(1) When developing the determinant over the first column, we obtain a signed sum of N determinants of size $(N-1) \times (N-1)$. But each of these determinants can be computed by developing over the first column too, and so on, and we are led to the conclusion that we have a formula as in the statement, with $\varepsilon(\sigma) \in \{-1, 1\}$ being certain coefficients.

(2) But these latter coefficients $\varepsilon(\sigma) \in \{-1, 1\}$ can only be the signatures of the corresponding permutations $\sigma \in S_N$, with this being something that can be viewed again by recurrence, with either of the definitions (1-5) in Theorem 2.24 for the signature. □

The above result is something quite tricky. As a first, basic illustration, in 2 dimensions we recover the usual formula of the determinant, the details being as follows:

$$\begin{aligned}\begin{vmatrix}a & b\\ c & d\end{vmatrix} &= \varepsilon(|\,|)\cdot ad+\varepsilon(\diagup\!\diagdown)\cdot cb\\ &= 1\cdot ad+(-1)\cdot cb\\ &= ad-bc\end{aligned}$$

In 3 dimensions, we recover the Sarrus formula, that we know from Theorem 2.22:

$$\begin{vmatrix}a & b & c\\ d & e & f\\ g & h & i\end{vmatrix}=aei+bfg+cdh-ceg-bdi-afh$$

Observe that the triangles in the Sarrus formula correspond to the permutations of $\{1,2,3\}$, and their signs correspond to the signatures of these permutations:

$$\begin{aligned}\det &= \begin{pmatrix}*&&\\&*&\\&&*\end{pmatrix}+\begin{pmatrix}&*&\\&&*\\*&&\end{pmatrix}+\begin{pmatrix}&&*\\*&&\\&*&\end{pmatrix}\\ &- \begin{pmatrix}&&*\\&*&\\*&&\end{pmatrix}+\begin{pmatrix}&*&\\*&&\\&&*\end{pmatrix}+\begin{pmatrix}*&&\\&&*\\&*&\end{pmatrix}\end{aligned}$$

In 4 dimensions now, by using our technology, we can formulate:

THEOREM 2.26. *The determinant of the 4×4 matrices is given by*

$$\begin{aligned}&\begin{vmatrix}a_1 & a_2 & a_3 & a_4\\ b_1 & b_2 & b_3 & b_4\\ c_1 & c_2 & c_3 & c_4\\ d_1 & d_2 & d_3 & d_4\end{vmatrix}\\ =\ & a_1b_2c_3d_4-a_1b_2c_4d_3-a_1b_3c_2d_4+a_1b_3c_4d_2+a_1b_4c_2d_3-a_1b_4c_3d_2\\ -\ & a_2b_1c_3d_4+a_2b_1c_4d_3+a_2b_3c_1d_4-a_2b_3c_4d_1-a_2b_4c_1d_3+a_2b_4c_3d_1\\ +\ & a_3b_1c_2d_4+a_3b_1c_4d_2-a_3b_2c_1d_4+a_3b_2c_4d_1+a_3b_4c_1d_2-a_3b_4c_2d_1\\ -\ & a_4b_1c_2d_3+a_4b_1c_3d_2-a_4b_2c_1d_3-a_4b_2c_3d_1-a_4b_3c_1d_2+a_4b_3c_2d_1\end{aligned}$$

with the generic term being of the following form, with $\sigma\in S_4$,

$$\pm a_{\sigma(1)}b_{\sigma(2)}c_{\sigma(3)}d_{\sigma(4)}$$

and with the sign being $\varepsilon(\sigma)$, computable by using Theorem 2.24.

PROOF. This follows indeed from Theorem 2.25, with the various permutations appearing in the statement being listed according to the lexicographic order. □

As yet another application, we have the following key result:

THEOREM 2.27. *We have the formula*

$$\det A = \det A^t$$

valid for any square matrix A.

PROOF. This follows from the formula in Theorem 2.25. Indeed, we have:

$$\begin{aligned}
\det A^t &= \sum_{\sigma\in S_N} \varepsilon(\sigma)(A^t)_{1\sigma(1)}\ldots(A^t)_{N\sigma(N)} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{\sigma(1)1}\ldots A_{\sigma(N)N} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{1\sigma^{-1}(1)}\ldots A_{N\sigma^{-1}(N)} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma^{-1})A_{1\sigma^{-1}(1)}\ldots A_{N\sigma^{-1}(N)} \\
&= \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{1\sigma(1)}\ldots A_{N\sigma(N)} \\
&= \det A
\end{aligned}$$

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

There are countless other applications of the formula in Theorem 2.25. Importantly, that formula allows us to deal now with the complex matrices too, as follows:

THEOREM 2.28. *If we define the determinant of a complex matrix $A \in M_N(\mathbb{C})$ to be*

$$\det A = \sum_{\sigma\in S_N} \varepsilon(\sigma)A_{1\sigma(1)}\ldots A_{N\sigma(N)}$$

then this determinant has the same properties as the determinant of the real matrices.

PROOF. This follows by doing some sort of reverse engineering, with respect to what we have been doing in this section, and we reach to the conclusion that det has indeed all the good properties that we are familiar with. Except of course for the properties at the very beginning of this chapter, in relation with volumes, which don't extend well to $\mathbb{C}^N$. Needless to say, all this applies as well to the matrices over an arbitrary field F. □

2c. Some applications

Good news, we have now in our bag all the needed techniques for computing the determinant. So, let us enjoy this knowledge, and do some computations of determinants. As a first result, which is something very classical, and whose proof actually uses some interesting new techniques, going beyond what has been said above, we have:

THEOREM 2.29. *We have the Vandermonde determinant formula*

$$\begin{vmatrix} 1 & 1 & 1 & \ldots\ldots & 1 \\ x_1 & x_2 & x_3 & \ldots\ldots & x_N \\ x_1^2 & x_2^2 & x_3^2 & \ldots\ldots & x_N \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ x_1^{N-1} & x_2^{N-1} & x_3^{N-1} & \ldots\ldots & x_N^{N-1} \end{vmatrix} = \prod_{i<j}(x_i - x_j)$$

valid for any $x_1, \ldots, x_N \in \mathbb{R}$.

PROOF. By expanding over the columns, we see that the determinant in question, say D, is a polynomial in the variables $x_1, \ldots, x_N$, having degree $N-1$ in each variable. Now observe that when setting $x_i = x_j$, for some indices $i \neq j$, our matrix will have two identical columns, and so its determinant D will vanish:

$$x_i = x_j \implies D = 0$$

But this gives us the key to the computation of D. Indeed, D must be divisible by $x_i - x_j$ for any $i \neq j$, and so we must have a formula of the following type:

$$D = c\prod_{i<j}(x_i - x_j)$$

Moreover, since the product on the right is, exactly as D itself, a polynomial in the variables $x_1, \ldots, x_N$, having degree $N-1$ in each variable, we conclude that the quantity c must be a constant, not depending on any of the variables $x_1, \ldots, x_N$:

$$c \in \mathbb{R}$$

In order to finish the computation, it remains to find the value of this constant c. But this can be done for instance by recurrence, and we obtain:

$$c = 1$$

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

Switching topics, an interesting class of matrices, which appear in coding theory end engineering, are the Hadamard matrices, which are the matrices $H \in M_N(\pm 1)$ whose rows are pairwise orthogonal. Here is a basic example, called first Walsh matrix:

$$W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

This matrix is quite trivial, of size 2×2, but by taking tensor powers of it, we have as examples the higher Walsh matrices as well, having size $2^k \times 2^k$, given by:

$$W_{2^k} = W_2^{\otimes k}$$

What happens then in arbitrary size $N \times N$? It is clear that we must have $2|N$, and along the same lines, it is easy to see, by playing around with the first rows, that once your matrix has $N \geq 3$ rows, we must have $4|N$, the precise result being as follows:

PROPOSITION 2.30. *The size of an Hadamard matrix $H \in M_N(\pm 1)$ must satisfy*

$$N \in \{2\} \cup 4\mathbb{N}$$

with this coming from the orthogonality condition between the first 3 rows.

PROOF. By permuting the rows and columns or by multiplying them by -1, as to rearrange the first 3 rows, we can always assume that our matrix looks as follows:

$$H = \begin{pmatrix} \underbrace{1 \ldots\ldots 1}_{x} & \underbrace{1 \ldots\ldots 1}_{y} & \underbrace{1 \ldots\ldots 1}_{z} & \underbrace{1 \ldots\ldots 1}_{t} \\ 1 \ldots\ldots 1 & 1 \ldots\ldots 1 & -1 \ldots -1 & -1 \ldots -1 \\ 1 \ldots\ldots 1 & -1 \ldots -1 & 1 \ldots\ldots 1 & -1 \ldots -1 \end{pmatrix}$$

Now if we denote by x, y, z, t the sizes of the block columns, as indicated, the orthogonality conditions between the first 3 rows give the following system of equations:

$$\begin{aligned} (1 \perp 2) &: x + y = z + t \\ (1 \perp 3) &: x + z = y + t \\ (2 \perp 3) &: x + t = y + z \end{aligned}$$

The numbers x, y, z, t being such that the average of any two equals the average of the other two, and so equals the global average, the solution of our system is $x = y = z = t$. Thus the matrix size $N = x + y + z + t$ must be a multiple of 4, as claimed. □

The above result is something quite interesting, and the point is that a similar analysis with 4 rows or more does not give any further restriction on the possible values of the size $N \in \mathbb{N}$. In fact, we are led in this way to the following famous conjecture:

CONJECTURE 2.31 (Hadamard). *There is an Hadamard matrix of order N,*

$$H \in M_N(\pm 1)$$

for any $N \in 4\mathbb{N}$.

This being said, for engineering purposes, what is concerning is not the Hadamard Conjecture, but rather the fact that Hadamard matrices don't exist at any $N \in \mathbb{N}$. And, a solution to this problem comes from determinants. Following Hadamard, we have:

THEOREM 2.32. *Given a matrix $H \in M_N(\pm 1)$, we have*

$$|\det H| \leq N^{N/2}$$

with equality precisely when H is Hadamard.

PROOF. We use the fact, that we learned in this chapter, that the determinant of a system of N vectors in $\mathbb{R}^N$ is the signed volume of the associated parallelepiped:

$$\det(H_1,\ldots,H_N)=\pm vol<H_1,\ldots,H_N>$$

Now in the case where our vectors have their entries in $\{\pm 1\}$, we therefore have the following inequality, with equality precisely when our vectors are pairwise orthogonal:

$$\begin{aligned}|\det(H_1,\ldots,H_N)| &\leq ||H_1||\times\ldots\times||H_N|| \\ &= (\sqrt{N})^N\end{aligned}$$

Thus, we have obtained the result, straight from the definition of det. □

The above result is quite interesting, and suggests formulating:

DEFINITION 2.33. *A quasi-Hadamard matrix is a square binary matrix*

$$H\in M_N(\pm 1)$$

which maximizes the quantity $|\det H|$.

In practice, the first problem appears at $N=3$, where $|\det H|\leq 4$, and with the quasi-Hadamard matrices here being the following matrix, and those obtained from it by permuting rows and columns, or switching the signs on rows and columns:

$$Q_3=\begin{pmatrix}1&1&1\\1&1&-1\\1&-1&1\end{pmatrix}$$

As a comment, however, all this might look a bit dissapointing, because it is hard to imagine that this matrix Q_3, which does now look as a very interesting matrix, can really play the role of a "generalized Hadamard matrix" at $N=3$. We will come later with more interesting solutions to this latter problem, a first solution being as follows:

$$K_3=\frac{1}{\sqrt{3}}\begin{pmatrix}-1&2&2\\2&-1&2\\2&2&-1\end{pmatrix}$$

To be more precise, this matrix is of course not binary, but it is definitely an interesting matrix, that we will see to be sharing many properties with the Hadamard matrices. Also, we have as well another solution to the $N=3$ problem, which uses complex numbers, and more specifically the number $w=e^{2\pi i/3}$, which is as follows:

$$F_3=\begin{pmatrix}1&1&1\\1&w&w^2\\1&w^2&w\end{pmatrix}$$

We will be back to these questions in chapter 11 below, with a systematic discussion of the Hadamard matrices, and their various generalizations.

As yet another application of the determinants, let us discuss now something quite scary, namely the Gram-Schmidt procedure. We have the following key definition:

DEFINITION 2.34. *The orthogonal polynomials $\{P_k\}_{k\in\mathbb{N}}$ with respect to a real measure $d\mu(x) = f(x)dx$ are the polynomials $P_k \in \mathbb{R}[x]$ of degree k satisfying:*

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

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

As a first observation, the orthogonal polynomials exist indeed for any real measure $d\mu(x) = f(x)dx$, because we can obtain them from the monomials x^k via Gram-Schmidt, as indicated above. However, by using our theory of determinants, we have:

THEOREM 2.35. *The orthogonal polynomials with respect to μ are given by*

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

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

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

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

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

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

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

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

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

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

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

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

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

$$< x, P_2 >= \int_{\mathbb{R}} xP_2(x)d\mu(x) = c_2 \begin{vmatrix} M_0 & M_1 & M_2 \\ M_1 & M_2 & M_3 \\ M_1 & M_2 & M_3 \end{vmatrix} = 0$$

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

2d. Diagonalization

Let us discuss now the diagonalization question for linear maps and matrices. The basic diagonalization theory, formulated in terms of matrices, is as follows:

THEOREM 2.36. *A vector $v \in \mathbb{C}^N$ is called eigenvector of $A \in M_N(\mathbb{C})$, with corresponding eigenvalue λ, when A multiplies by λ in the direction of v:*

$$Av = \lambda v$$

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

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

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

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

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

PROOF. This is something that we know from chapter 1, the idea being as follows:

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

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

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

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

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

THEOREM 2.37. *Let $A \in M_N(\mathbb{C})$, and for any eigenvalue $\lambda \in \mathbb{C}$ define the corresponding eigenspace as being the vector space formed by the corresponding eigenvectors:*

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

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

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

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

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

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

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

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

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

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

On the right we obtain something different, as follows:

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

We conclude from this that the following equality must hold:

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

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

$$\lambda_k c_1 = c_1 \lambda_1$$

$$\vdots$$

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

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

In order to reach now to more advanced results, we can use the following key fact:

THEOREM 2.38. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial:*

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

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

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

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

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

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

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

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

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

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

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

Now recall that we are over $\mathbb{C}$, where the equation $X^2 + 1 = 0$, and in fact any degree 2 equation, has 2 complex roots. We can in fact prove that any polynomial equation, of arbitrary degree $N \in \mathbb{N}$, has exactly N complex solutions, counted with multiplicities:

THEOREM 2.39. *Any polynomial $P \in \mathbb{C}[X]$ decomposes as*

$$P = c(X - a_1) \ldots (X - a_N)$$

with $c \in \mathbb{C}$ and with $a_1, \ldots, a_N \in \mathbb{C}$.

PROOF. The problem is that of proving that our polynomial has at least one root, because afterwards we can proceed by recurrence. We prove this by contradiction. So, assume that P has no roots, and pick a number $z \in \mathbb{C}$ where $|P|$ attains its minimum:

$$|P(z)| = \min_{x \in \mathbb{C}} |P(x)| > 0$$

Since $Q(t) = P(z + t) - P(z)$ is a polynomial which vanishes at $t = 0$, this polynomial must be of the form ct^k + higher terms, with $c \neq 0$, and with $k \geq 1$ being an integer. We obtain from this that, with $t \in \mathbb{C}$ small, we have the following estimate:

$$P(z + t) \simeq P(z) + ct^k$$

Now let us write $t = rw$, with $r > 0$ small, and with $|w| = 1$. Our estimate becomes:

$$P(z + rw) \simeq P(z) + cr^k w^k$$

Now recall that we have assumed $P(z) \neq 0$. We can therefore choose $w \in \mathbb{T}$ such that cw^k points in the opposite direction to that of $P(z)$, and we obtain in this way:

$$\begin{aligned} |P(z + rw)| &\simeq |P(z) + cr^k w^k| \\ &= |P(z)|(1 - |c|r^k) \end{aligned}$$

Now by choosing $r > 0$ small enough, as for the error in the first estimate to be small, and overcame by the negative quantity $-|c|r^k$, we obtain from this:

$$|P(z + rw)| < |P(z)|$$

But this contradicts our definition of $z \in \mathbb{C}$, as a point where $|P|$ attains its minimum. Thus P has a root, and by recurrence it has N roots, as stated. □

Getting back now to linear algebra, we obtain the following result:

THEOREM 2.40. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial*

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

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

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

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

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

The dimensions of these eigenspaces satisfy then the following inequalities,

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

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

PROOF. This follows by combining the above results. By summing the inequalities $\dim(E_\lambda) \leq m_\lambda$ from Theorem 2.38, we obtain an inequality as follows:

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

On the other hand, we know from Theorem 2.37 that our matrix is diagonalizable when we have global equality. Thus, we are led to the conclusion in the statement. □

This was for the main result of linear algebra. There are countless applications of this, and generally speaking, advanced linear algebra consists in building on Theorem 2.40.

In practice now, the use of Theorem 2.40 requires some practice, and skill. In relation with this, let us record as well a useful algorithmic version of the above result:

THEOREM 2.41. *The square matrices $A \in M_N(\mathbb{C})$ can be diagonalized as follows:*

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

PROOF. This is an informal reformulation of Theorem 2.40, with (4) referring to the total number of linearly independent eigenvectors found in (3), and with $A = PDP^{-1}$ in (5) being the usual diagonalization formula, with P, D being as before. □

As a remark here, in step (3) it is always better to start with the eigenvalues having big multiplicity. Indeed, a multiplicity 1 eigenvalue, for instance, can never lead to the end of the computation, via (4), simply because the eigenvectors always exist.

2e. Exercises

This was another basic chapter, all starting level linear algebra material, that you are basically supposed to know, and as exercises on all this, we have:

EXERCISE 2.42. *Have some fun with computing the orientation of various systems of vectors in space, of your choice.*

EXERCISE 2.43. *Forget if needed the definition and theory of the determinant that you learned in school, and relearn it as indicated above, as a signed volume.*

EXERCISE 2.44. *Further meditate on the uniqueness statement for the determinant given above, and learn about some other such uniqueness statements as well.*

EXERCISE 2.45. *Review the theory of permutations and their signatures, and learn also about the alternating group $A_N \subset S_N$.*

EXERCISE 2.46. *Learn more about the real Hadamard matrices, their various properties, and about the Hadamard Conjecture, and its status.*

EXERCISE 2.47. *In relation with Gram-Schmidt, learn about the Legendre, Chebycheff, Jacobi, Laguerre and Hermite polynomials.*

EXERCISE 2.48. *Learn the various standard algebraic and analytic tricks for factorizing polynomials.*

EXERCISE 2.49. *Try to understand what happens to our diagonalization theory in the case of basic non-diagonalizable matrices, such as the basic Jordan block J.*

As bonus exercise, diagonalize some matrices, as many as you can. Normally you cannot call yourself a scientist until you do 3×3 matrices in 15 minutes, or less.

CHAPTER 3

Spectral theorems

3a. Self-adjoints

Let us go back to the diagonalization question, discussed in the previous chapter. We have in fact diagonalization results which are far more powerful. We first have:

THEOREM 3.1. *Any matrix $A \in M_N(\mathbb{C})$ which is self-adjoint, $A = A^*$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal and real, then we have:

$$A^* = (UDU^*)^* = UD^*U^* = UDU^* = A$$

In the other sense now, assume that A is self-adjoint, $A = A^*$. Our first claim is that the eigenvalues are real. Indeed, assuming $Av = \lambda v$, we have:

$$\begin{aligned} \lambda < v, v > &= < \lambda v, v > \\ &= < Av, v > \\ &= < v, Av > \\ &= < v, \lambda v > \\ &= \bar{\lambda} < v, v > \end{aligned}$$

Thus we obtain $\lambda \in \mathbb{R}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have then the following computation, using $\lambda, \mu \in \mathbb{R}$:

$$\begin{aligned} \lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, Aw > \\ &= < v, \mu w > \\ &= \mu < v, w > \end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of A span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Av = \lambda v$$

Assuming now that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned} < Aw, v > &= < w, Av > \\ &= < w, \lambda v > \\ &= \lambda < w, v > \\ &= 0 \end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under A. Moreover, since a matrix A is self-adjoint precisely when $< Av, v > \in \mathbb{R}$ for any vector $v \in \mathbb{C}^N$, as one can see by expanding the scalar product, the restriction of A to the subspace $v^\perp$ is self-adjoint. Thus, we can proceed by recurrence, and we obtain the result. □

Observe that, as a consequence of the above result, that you certainly might have heard of, any symmetric matrix $A \in M_N(\mathbb{R})$ is diagonalizable. In fact, we have:

THEOREM 3.2. *Any matrix $A \in M_N(\mathbb{R})$ which is symmetric, $A = A^t$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^t$$

with $U \in O_N$, and with $D \in M_N(\mathbb{R})$ diagonal. The converse holds too.

PROOF. As before, the converse trivially holds, because if we take a matrix of the form $A = UDU^t$, with U orthogonal and D diagonal and real, then we have $A^t = A$. In the other sense now, this follows from Theorem 3.1, and its proof. □

As basic examples of self-adjoint matrices, we have the orthogonal projections:

PROPOSITION 3.3. *The matrices $P \in M_N(\mathbb{C})$ which are projections, $P^2 = P^* = P$, are precisely those which diagonalize as follows,*

$$P = UDU^*$$

with $U \in U_N$, and with $D \in M_N(0, 1)$ being diagonal.

PROOF. Since we have $P^* = P$, by using Theorem 3.1, the eigenvalues must be real. Then, by using $P^2 = P$, assuming that we have $Pv = \lambda v$, we obtain:

$$\begin{aligned} \lambda < v, v > &= < P^2 v, v > \\ &= < Pv, Pv > \\ &= \lambda^2 < v, v > \end{aligned}$$

Thus $\lambda \in \{0, 1\}$, and the diagonalization must be as follows, with $e_i \in \{0, 1\}$:

$$P \sim \begin{pmatrix} e_1 & & \\ & \ddots & \\ & & e_N \end{pmatrix}$$

To be more precise, the number of 1 values is the dimension of the image of P. □

In the real case, the result regarding the projections is as follows:

PROPOSITION 3.4. *The matrices $P \in M_N(\mathbb{R})$ which are projections, $P^2 = P^t = P$, are precisely those which diagonalize as follows,*

$$P = UDU^t$$

with $U \in O_N$, and with $D \in M_N(0, 1)$ being diagonal.

PROOF. This follows indeed from Proposition 3.3, and its proof. □

An important class of self-adjoint matrices, which includes for instance all the projections, are the positive matrices. The theory here is as follows:

THEOREM 3.5. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is positive:*

(1) *$A = B^2$, with $B = B^*$.*
(2) *$A = CC^*$, for some $C \in M_N(\mathbb{C})$.*
(3) *$< Ax, x > \geq 0$, for any vector $x \in \mathbb{C}^N$.*
(4) *$A = A^*$, and the eigenvalues are positive, $\lambda_i \geq 0$.*
(5) *$A = UDU^*$, with $U \in U_N$ and with $D \in M_N(\mathbb{R}_+)$ diagonal.*

PROOF. The idea is that the equivalences in the statement basically follow from some elementary computations, with only Theorem 3.1 needed, at some point:

(1) $\implies$ (2) This is clear, because we can take $C = B$.

(2) $\implies$ (3) This follows from the following computation:

$$< Ax, x > = < CC^*x, x > = < C^*x, C^*x > \geq 0$$

(3) $\implies$ (4) By using the fact that $< Ax, x >$ is real, we have:

$$< Ax, x > = < x, A^*x > = < A^*x, x >$$

Thus we have $A = A^*$, and the remaining assertion, regarding the eigenvalues, follows from the following computation, assuming $Ax = \lambda x$:

$$< Ax, x > = < \lambda x, x > = \lambda < x, x > \geq 0$$

(4) $\implies$ (5) This follows indeed by using Theorem 3.1.

(5) $\implies$ (1) Assuming $A = UDU^*$ with $U \in U_N$, and with $D \in M_N(\mathbb{R}_+)$ diagonal, we can set $B = U\sqrt{D}U^*$. Then B is self-adjoint, and $B^2 = A$, which gives the result. □

Let us record as well the following technical version of the above result:

THEOREM 3.6. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is strictly positive:*

(1) $A = B^2$, *with* $B = B^*$, *invertible.*
(2) $A = CC^*$, *for some* $C \in M_N(\mathbb{C})$ *invertible.*
(3) $< Ax, x >> 0$, *for any nonzero vector* $x \in \mathbb{C}^N$.
(4) $A = A^*$, *and the eigenvalues are strictly positive,* $\lambda_i > 0$.
(5) $A = UDU^*$, *with* $U \in U_N$ *and with* $D \in M_N(\mathbb{R}_+^*)$ *diagonal.*

PROOF. This follows either from Theorem 3.5, by adding the above various extra assumptions, or from the proof of Theorem 3.5, by modifying where needed. □

Let us discuss now some applications of the above. Some very basic examples of symmetric matrices are the adjacency matrices of graphs, and we first have here:

PROPOSITION 3.7. *Given a graph X, with adjacency matrix $d \in M_N(0,1)$, the eigenvalues of d, with eigenvalue λ, can be identified with the functions f satisfying:*

$$\lambda f(i) = \sum_{i-j} f(j)$$

That is, the value of f at each vertex must be the rescaled average, over the neighbors.

PROOF. We have indeed the following computation, valid for any vector f:

$$\begin{aligned}(df)_i &= \sum_j d_{ij} f_j \\ &= \sum_{i-j} d_{ij} f_j + \sum_{i\neq j} d_{ij} f_j \\ &= \sum_{i-j} 1 \cdot f_j + \sum_{i\neq j} 0 \cdot f_j \\ &= \sum_{i-j} f_j\end{aligned}$$

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

The above result is quite interesting, and as an illustration, when assuming that our graph is k-regular, for the particular value $\lambda = k$, the eigenvalue condition reads:

$$f(i) = \frac{1}{k} \sum_{i-j} f(j)$$

Thus, we can see here a relation with harmonic functions. There are many things that can be said here, and we will be back to this later, when talking Laplace operators. Now moving on, we can apply to $d \in M_N(0,1)$ our spectral theorems, and we obtain:

THEOREM 3.8. *The adjacency matrix $d \in M_N(0,1)$ of any graph is diagonalizable, with the diagonalization being of the following type,*

$$d = UDU^t$$

with $U \in O_N$, and with $D \in M_N(\mathbb{R})$ diagonal. Moreover, we have $Tr(D) = 0$.

PROOF. Here the first assertion follows from Theorem 3.2, because d is by definition real and symmetric. As for the last assertion, this deserves some explanations:

(1) Generally speaking, in analogy with the last assertions in Theorem 3.1 and Theorem 3.2, which are something extremely useful, we would like to know under which assumptions on a rotation $U \in O_N$, and on a diagonal matrix $D \in M_N(\mathbb{R})$, the real symmetric matrix $d = UDU^t$ has 0-1 entries, and 0 on the diagonal.

(2) Unfortunately, both these questions are obviously difficult, there is no simple answer to them, and things are like that. So, gone the possibility of a converse. However, as a small consolation, we can make the remark that, with $d = UDU^t$, we have:

$$Tr(d) = Tr(UDU^t) = Tr(D)$$

Thus we have at least $Tr(D) = 0$, as a necessary condition on (U, D), as stated. □

In view of the above difficulties with the bijectivity, it is perhaps wise to formulate as well the graph particular case of Theorem 3.1. The statement here is as follows:

THEOREM 3.9. *The adjacency matrix $d \in M_N(0,1)$ of any graph is diagonalizable, with the diagonalization being of the following type,*

$$d = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{R})$ diagonal. Moreover, we have $Tr(D) = 0$.

PROOF. This follows from Theorem 3.1, via the various remarks from the proof of Proposition 3.7 and Theorem 3.8. But the simplest is to say that the statement itself is just a copy of Theorem 3.8, with $U \in O_N$ replaced by the weaker condition $U \in U_N$. □

As a concrete illustration now for all this, let us look at the simplest graph of them all, namely the simplex, or complete graph. And for this graph, not only the above results successfully apply, but we can also see why Theorem 3.9 is something wise:

THEOREM 3.10. *The adjacency matrix of the simplex diagonalizes as follows,*

$$\begin{pmatrix} 0 & 1 & \ldots & 1 & 1 \\ 1 & 0 & \ldots & 1 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & \ldots & 0 & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N-1 & & & & 0 \\ & -1 & & & \\ & & \ddots & & \\ & & & -1 & \\ 0 & & & & -1 \end{pmatrix} F_N^*$$

with $F_N = (w^{ij})_{ij}$ being the Fourier matrix.

PROOF. The adjacency matrix of the simplex is $d = \mathbb{I}_N - 1_N$, with $\mathbb{I}_N$ being the all-one, or flat matrix. So, let us first attempt to diagonalize the flat matrix:

$$\mathbb{I}_N = \begin{pmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \ldots & \ldots & 1 \end{pmatrix}$$

But here, we already know, since chapter 1, that the simplest way to diagonalize this matrix is as follows, with $F_N = (w^{ij})_{ij}$ being the Fourier matrix:

$$\begin{pmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \ldots & \ldots & 1 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N & & & 0 \\ & 0 & & \\ & & \ddots & \\ 0 & & & 0 \end{pmatrix} F_N^*$$

But this gives the result, by substracting -1 from everything. □

All the above is quite useful, and we will use these results on a regular basis, in what follows. There are also some positivity considerations that can be made, in relation with graphs, and we will be back to this later, when talking Laplacians of graphs.

3b. Rotations, unitaries

Let us discuss now the case of the unitary matrices. We have here:

THEOREM 3.11. *Any matrix $U \in M_N(\mathbb{C})$ which is unitary, $U^* = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because given a matrix of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, we have:

$$\begin{aligned} U^* &= (VDV^*)^* \\ &= VD^*V^* \\ &= VD^{-1}V^{-1} \\ &= (V^*)^{-1}D^{-1}V^{-1} \\ &= (VDV^*)^{-1} \\ &= U^{-1} \end{aligned}$$

Let us prove now the first assertion, stating that the eigenvalues of a unitary matrix $U \in U_N$ belong to $\mathbb{T}$. Indeed, assuming $Uv = \lambda v$, we have:

$$\begin{aligned} < v, v > &= < U^*Uv, v > \\ &= < Uv, Uv > \\ &= < \lambda v, \lambda v > \\ &= |\lambda|^2 < v, v > \end{aligned}$$

Thus $\lambda \in \mathbb{T}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that $Uv = \lambda v$, $Uw = \mu w$. We have then the following computation, using $U^* = U^{-1}$ and $\lambda, \mu \in \mathbb{T}$:

$$\begin{aligned} \lambda < v, w > &= < Uv, w > \\ &= < v, U^*w > \\ &= < v, U^{-1}w > \\ &= < v, \mu^{-1}w > \\ &= \mu < v, w > \end{aligned}$$

Thus $\lambda \neq \mu$ implies $v \perp w$, as claimed. In order now to finish the proof, it remains to prove that the eigenspaces of U span the whole space $\mathbb{C}^N$. For this purpose, we will use a recurrence method. Let us pick an eigenvector of our matrix:

$$Uv = \lambda v$$

Assuming that we have a vector w orthogonal to it, $v \perp w$, we have:

$$\begin{aligned} < Uw, v > &= < w, U^*v > \\ &= < w, U^{-1}v > \\ &= < w, \lambda^{-1}v > \\ &= \lambda < w, v > \\ &= 0 \end{aligned}$$

Thus, if v is an eigenvector, then the vector space $v^\perp$ is invariant under U. Now since U is an isometry, so is its restriction to this space $v^\perp$. Thus this restriction is a unitary, and so we can proceed by recurrence, and we obtain the result. □

Let us record as well the real version of the above result, in a weak form:

PROPOSITION 3.12. *Any matrix $U \in M_N(\mathbb{R})$ which is orthogonal, $U^t = U^{-1}$, is diagonalizable, with the eigenvalues on $\mathbb{T}$. More precisely we have*

$$U = VDV^*$$

with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal.

PROOF. This follows indeed from Theorem 3.11. □

Observe that the above result does not provide us with a complete characterization of the matrices $U \in M_N(\mathbb{R})$ which are orthogonal. To be more precise, the question left is that of understanding when the matrices of type $U = VDV^*$, with $V \in U_N$, and with $D \in M_N(\mathbb{T})$ being diagonal, are real, and this is something non-trivial.

As an illustration for the above, for the simplest unitaries that we know, namely the rotations in the real plane, we have the following result:

THEOREM 3.13. *The rotation of angle $t \in \mathbb{R}$ in the real plane, namely*

$$R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$$

can be diagonalized over the complex numbers, as follows:

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

Over the real numbers this is impossible, unless $t = 0, \pi$.

PROOF. This is indeed something that we know since chapter 1, and we refer to the discussion there for the above assertions, both over $\mathbb{R}$ and $\mathbb{C}$. □

In two complex dimensions now, it is convenient to restrict the attention to the unitaries of determinant 1, which are subject to the following well-known result:

THEOREM 3.14. *We have the following formula,*

$$SU_2 = \left\{ \begin{pmatrix} a & b \\ -\bar{b} & \bar{a} \end{pmatrix} \;\middle|\; |a|^2 + |b|^2 = 1 \right\}$$

which makes SU_2 isomorphic to the unit sphere $S^1_{\mathbb{C}} \subset \mathbb{C}^2$.

PROOF. Consider indeed an arbitrary 2×2 matrix, written as follows:

$$U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Assuming that we have $\det U = 1$, the inverse must be given by:

$$U^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

On the other hand, assuming $U \in U_2$, the inverse must be the adjoint:

$$U^{-1} = \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix}$$

We are therefore led to the following equations, for the matrix entries:

$$d = \bar{a} \quad , \quad c = -\bar{b}$$

Thus our matrix must be of the following special form in the statement. Moreover, since the determinant of this matrix is 1, we must have, as stated:

$$|a|^2+|b|^2=1$$

Thus, we are done with one inclusion. As for the converse, this is clear, the matrices in the statement being unitaries, and of determinant 1, and so being elements of SU_2. Finally, regarding the last assertion, thisis something clear too. □

According to Theorem 3.11 the matrices in Theorem 3.14 are diagonalizable, and we will leave their diagonalization as an instructive exercise. Moving forward now, in 3 dimensions things are more complicated, and in order to discuss this, we will need:

THEOREM 3.15. *We have the following formula,*

$$SU_2=\left\{pc_1+qc_2+rc_3+sc_4\;\middle|\;p^2+q^2+r^2+s^2=1\right\}$$

where c_1,c_2,c_3,c_4 are the following matrices,

$$c_1=\begin{pmatrix}1&0\\0&1\end{pmatrix}\quad,\quad c_2=\begin{pmatrix}i&0\\0&-i\end{pmatrix}\quad,\quad c_3=\begin{pmatrix}0&1\\-1&0\end{pmatrix}\quad,\quad c_4=\begin{pmatrix}0&i\\i&0\end{pmatrix}$$

called Pauli spin matrices.

PROOF. We know from Theorem 3.14 that the group SU_2 can be parametrized by the points of the real sphere $S^3_{\mathbb{R}}\subset\mathbb{R}^4$, in the following way:

$$SU_2=\left\{\begin{pmatrix}p+iq&r+is\\-r+is&p-iq\end{pmatrix}\;\middle|\;p^2+q^2+r^2+s^2=1\right\}$$

But this gives the formula in the statement, with the Pauli matrices c_1,c_2,c_3,c_4 being the coefficients of p,q,r,s, in this parametrization. □

The above result is often the most convenient one, when dealing with SU_2. This is because the Pauli matrices have a number of remarkable properties, as follows:

PROPOSITION 3.16. *The Pauli matrices multiply according to the following formulae,*

$$c_2^2=c_3^2=c_4^2=-1$$
$$c_2c_3=-c_3c_2=c_4$$
$$c_3c_4=-c_4c_3=c_2$$
$$c_4c_2=-c_2c_4=c_3$$

they conjugate according to the following rules,

$$c_1^*=c_1,\ c_2^*=-c_2,\ c_3^*=-c_3,\ c_4^*=-c_4$$

and they form an orthonormal basis of $M_2(\mathbb{C})$, with respect to the scalar product

$$<x,y>=tr(xy^*)$$

with $tr:M_2(\mathbb{C})\to\mathbb{C}$ being the normalized trace of 2×2 matrices, $tr=Tr/2$.

PROOF. The first two assertions, regarding the multiplication and conjugation rules for the Pauli matrices, follow from some elementary computations. As for the last assertion, this follows by using these rules. Indeed, the fact that the Pauli matrices are pairwise orthogonal follows from computations of the following type, for $i \neq j$:

$$< c_i, c_j >= tr(c_i c_j^*) = tr(\pm c_i c_j) = tr(\pm c_k) = 0$$

As for the fact that the Pauli matrices have norm 1, this follows from:

$$< c_i, c_i >= tr(c_i c_i^*) = tr(\pm c_i^2) = tr(c_1) = 1$$

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

Getting now towards SO_3, we first have the following result:

PROPOSITION 3.17. *The adjoint action $SU_2 \curvearrowright M_2(\mathbb{C})$, given by $T_U(A) = UAU^*$, leaves invariant the following real vector subspace of $M_2(\mathbb{C})$,*

$$\mathbb{R}^4 = span(c_1, c_2, c_3, c_4)$$

and we obtain in this way a group morphism $SU_2 \to GL_4(\mathbb{R})$.

PROOF. We have two assertions to be proved, as follows:

(1) We must first prove that, with $E \subset M_2(\mathbb{C})$ being the real vector space in the statement, we have the following implication:

$$U \in SU_2, A \in E \implies UAU^* \in E$$

But this is clear from the multiplication rules for the Pauli matrices, from Proposition 3.16. Indeed, let us write our matrices U, A as follows:

$$U = xc_1 + yc_2 + zc_3 + tc_4$$

$$A = ac_1 + bc_2 + cc_3 + dc_4$$

We know that the coefficients x, y, z, t and a, b, c, d are all real, due to $U \in SU_2$ and $A \in E$. The point now is that when computing UAU^*, by using the various rules from Proposition 3.16, we obtain a matrix of the same type, namely a combination of c_1, c_2, c_3, c_4, with real coefficients. Thus, we have $UAU^* \in E$, as desired.

(2) In order to conclude, let us identify $E \simeq \mathbb{R}^4$, by using the basis c_1, c_2, c_3, c_4. The result found in (1) shows that we have a correspondence as follows:

$$SU_2 \to M_4(\mathbb{R}) \quad , \quad U \to (T_U)_{|E}$$

Now observe that for any $U \in SU_2$ and any $A \in M_2(\mathbb{C})$ we have:

$$T_{U^*} T_U(A) = U^* UAU^* U = A$$

Thus $T_{U^*} = T_U^{-1}$, and so the correspondence that we found can be written as:

$$SU_2 \to GL_4(\mathbb{R}) \quad , \quad U \to (T_U)_{|E}$$

But this a group morphism, due to the following computation:

$$T_UT_V(A)=UVAV^*U^*=T_{UV}(A)$$

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

The above result is quite interesting, and as a continuation of it, we have:

PROPOSITION 3.18. *With respect to the standard basis c_1,c_2,c_3,c_4 of the vector space $\mathbb{R}^4=span(c_1,c_2,c_3,c_4)$, the morphism $T:SU_2\to GL_4(\mathbb{R})$ is given by:*

$$T_U=\begin{pmatrix}1&0&0&0\\0&p^2+q^2-r^2-s^2&2(qr-ps)&2(pr+qs)\\0&2(ps+qr)&p^2+r^2-q^2-s^2&2(rs-pq)\\0&2(qs-pr)&2(pq+rs)&p^2+s^2-q^2-r^2\end{pmatrix}$$

Thus, when looking at T as a group morphism $SU_2\to O_4$, what we have in fact is a group morphism $SU_2\to O_3$, and even $SU_2\to SO_3$.

PROOF. With notations from Proposition 3.17 and its proof, let us first look at the action $L:SU_2\curvearrowright\mathbb{R}^4$ by left multiplication, $L_U(A)=UA$. We have:

$$L_U=\begin{pmatrix}p&-q&-r&-s\\q&p&-s&r\\r&s&p&-q\\s&-r&q&p\end{pmatrix}$$

Similarly, in what regards now the action $R:SU_2\curvearrowright\mathbb{R}^4$ by right multiplication, $R_U(A)=AU^*$, the corresponding matrix is given by:

$$R_U=\begin{pmatrix}p&q&r&s\\-q&p&-s&r\\-r&s&p&-q\\-s&-r&q&p\end{pmatrix}$$

Now by composing, the matrix of the adjoint matrix in the statement is:

$$\begin{aligned}T_U&=R_UL_U\\&=\begin{pmatrix}p&q&r&s\\-q&p&-s&r\\-r&s&p&-q\\-s&-r&q&p\end{pmatrix}\begin{pmatrix}p&-q&-r&-s\\q&p&-s&r\\r&s&p&-q\\s&-r&q&p\end{pmatrix}\\&=\begin{pmatrix}1&0&0&0\\0&p^2+q^2-r^2-s^2&2(qr-ps)&2(pr+qs)\\0&2(ps+qr)&p^2+r^2-q^2-s^2&2(rs-pq)\\0&2(qs-pr)&2(pq+rs)&p^2+s^2-q^2-r^2\end{pmatrix}\end{aligned}$$

Thus, we have the formula in the statement, and this gives the result. □

We can now formulate a famous result, due to Euler-Rodrigues, as follows:

THEOREM 3.19. *We have the Euler-Rodrigues formula*

$$U=\begin{pmatrix}p^2+q^2-r^2-s^2 & 2(qr-ps) & 2(pr+qs)\\ 2(ps+qr) & p^2+r^2-q^2-s^2 & 2(rs-pq)\\ 2(qs-pr) & 2(pq+rs) & p^2+s^2-q^2-r^2\end{pmatrix}$$

with $p^2+q^2+r^2+s^2=1$, for the generic elements of SO_3.

PROOF. We know from the above that we have a group morphism $SU_2\to SO_3$, given by the formula in the statement, and the problem now is that of proving that this is a double cover map, in the sense that it is surjective, and with kernel $\{\pm 1\}$.

(1) Regarding the kernel, this is elementary to compute, as follows:

$$\begin{aligned}\ker(SU_2\to SO_3) &= \left\{U\in SU_2\Big|T_U(A)=A,\forall A\in E\right\}\\ &= \left\{U\in SU_2\Big|UA=AU,\forall A\in E\right\}\\ &= \left\{U\in SU_2\Big|Uc_i=c_iU,\forall i\right\}\\ &= \{\pm 1\}\end{aligned}$$

(2) Thus, we are left with proving that our group morphism $SU_2\to SO_3$ is surjective. As a first computation, let us prove that any rotation $U\in Im(SU_2\to SO_3)$ has an axis. We must look for fixed points of such rotations, and by linearity it is enough to look for fixed points belonging to the sphere $S^2_{\mathbb{R}}\subset\mathbb{R}^3$. Now recall that in our picture for the quotient map $SU_2\to SO_3$, the space $\mathbb{R}^3$ appears as $F=span_{\mathbb{R}}(c_2,c_3,c_4)$, naturally embedded into the space $\mathbb{R}^4$ appearing as $E=span_{\mathbb{R}}(c_1,c_2,c_3,c_4)$. Thus, we must look for fixed points belonging to the sphere $S^3_{\mathbb{R}}\subset\mathbb{R}^4$ whose first coordinate vanishes. But, in our $\mathbb{R}^4=E$ picture, this sphere $S^3_{\mathbb{R}}$ is the group SU_2. Thus, we must look for fixed points $V\in SU_2$ whose first coordinate with respect to c_1,c_2,c_3,c_4 vanishes, which amounts in saying that the diagonal entries of V must be purely imaginary numbers.

(3) Long story short, via our various identifications, we are led into solving the equation $UV=VU$ with $U,V\in SU_2$, and with V having a purely imaginary diagonal. So, with standard notations for SU_2, we must solve the following equation, with $p\in i\mathbb{R}$:

$$\begin{pmatrix}a & b\\ -\bar{b} & \bar{a}\end{pmatrix}\begin{pmatrix}p & q\\ -\bar{q} & \bar{p}\end{pmatrix}=\begin{pmatrix}p & q\\ -\bar{q} & \bar{p}\end{pmatrix}\begin{pmatrix}a & b\\ -\bar{b} & \bar{a}\end{pmatrix}$$

But this is something which is routine. Indeed, by identifying coefficients we obtain the following equations, each appearing twice:

$$b\bar{q}=\bar{b}q\quad,\quad b(p-\bar{p})=(a-\bar{a})q$$

In the case $b = 0$ the only equation which is left is $q = 0$, and reminding that we must have $p \in i\mathbb{R}$, we do have solutions, namely two of them, as follows:

$$V = \pm \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}$$

(4) In the remaining case $b \neq 0$, the first equation reads $b\bar{q} \in \mathbb{R}$, so we must have $q = \lambda b$ with $\lambda \in \mathbb{R}$. Now with this substitution made, the second equation reads $p - \bar{p} = \lambda(a - \bar{a})$, and since we must have $p \in i\mathbb{R}$, this gives $2p = \lambda(a - \bar{a})$. Thus, our equations are:

$$q = \lambda b \quad , \quad p = \lambda \cdot \frac{a - \bar{a}}{2}$$

Getting back now to our problem about finding fixed points, assuming $|a|^2 + |b|^2 = 1$ we must find $\lambda \in \mathbb{R}$ such that the above numbers p, q satisfy $|p|^2 + |q|^2 = 1$. But:

$$\begin{aligned} |p|^2 + |q|^2 &= |\lambda b|^2 + \left| \lambda \cdot \frac{a - \bar{a}}{2} \right|^2 \\ &= \lambda^2(|b|^2 + Im(a)^2) \\ &= \lambda^2(1 - Re(a)^2) \end{aligned}$$

Thus, we have again two solutions to our fixed point problem, given by:

$$\lambda = \pm \frac{1}{\sqrt{1 - Re(a)^2}}$$

Summarizing, we have proved that any rotation $U \in Im(SU_2 \to SO_3)$ has an axis, and with the direction of this axis, corresponding to a pair of opposite points on the sphere $S^2_{\mathbb{R}} \subset \mathbb{R}^3$, being given by the above formulae, via $S^2_{\mathbb{R}} \subset S^3_{\mathbb{R}} = SU_2$.

(5) In order to finish, we must still argue that any rotation $U \in SO_3$ has an axis. But this follows for instance from some topology, by using the induced map $S^2_{\mathbb{R}} \to S^2_{\mathbb{R}}$. Now since $U \in SO_3$ is uniquely determined by its rotation axis, which can be regarded as a point of $S^2_{\mathbb{R}}/\{\pm 1\}$, plus its rotation angle $t \in [0, 2\pi)$, by using $S^2_{\mathbb{R}} \subset S^3_{\mathbb{R}} = SU_2$ we are led to the conclusion that U is uniquely determined by an element of $SU_2/\{\pm 1\}$, and so appears indeed via the Euler-Rodrigues formula, as desired. □

And with the above result, good news, if looking for some difficult exercises, in relation with the spectral theorem for unitaries, try diagonalizing the above matrices. We will be back to this later in this book, when systematically discussing the Lie groups.

3c. Normal matrices

Back to generalities, the self-adjoint matrices and the unitary matrices are particular cases of the general notion of a "normal matrix", and we have here:

THEOREM 3.20. *Any matrix $A \in M_N(\mathbb{C})$ which is normal, $AA^* = A^*A$, is diagonalizable, with the diagonalization being of the following type,*

$$A = UDU^*$$

with $U \in U_N$, and with $D \in M_N(\mathbb{C})$ diagonal. The converse holds too.

PROOF. As a first remark, the converse trivially holds, because if we take a matrix of the form $A = UDU^*$, with U unitary and D diagonal, then we have:

$$\begin{aligned} AA^* &= UDU^* \cdot UD^*U^* \\ &= UDD^*U^* \\ &= UD^*DU^* \\ &= UD^*U^* \cdot UDU^* \\ &= A^*A \end{aligned}$$

In the other sense now, this is something more technical. Our first claim is that a matrix A is normal precisely when the following happens, for any vector v:

$$||Av|| = ||A^*v||$$

Indeed, the above equality can be written as follows:

$$< AA^*v, v >=< A^*Av, v >$$

But this is equivalent to $AA^* = A^*A$, by expanding the scalar products. Our claim now is that A, A^* have the same eigenvectors, with conjugate eigenvalues:

$$Av = \lambda v \implies A^*v = \bar{\lambda}v$$

Indeed, this follows from the following computation, and from the trivial fact that if A is normal, then so is any matrix of type $A - \lambda 1_N$:

$$\begin{aligned} ||(A^* - \bar{\lambda}1_N)v|| &= ||(A - \lambda 1_N)^*v|| \\ &= ||(A - \lambda 1_N)v|| \\ &= 0 \end{aligned}$$

Let us prove now, by using this, that the eigenspaces of A are pairwise orthogonal. Assume that we have two eigenvectors, corresponding to different eigenvalues, $\lambda \neq \mu$:

$$Av = \lambda v \quad , \quad Aw = \mu w$$

We have the following computation, which shows that $\lambda \neq \mu$ implies $v \perp w$:

$$\begin{aligned}\lambda < v, w > &= < \lambda v, w > \\ &= < Av, w > \\ &= < v, A^* w > \\ &= < v, \bar{\mu} w > \\ &= \mu < v, w >\end{aligned}$$

In order to finish, it remains to prove that the eigenspaces of A span the whole $\mathbb{C}^N$. This is something that we have already seen for the self-adjoint matrices, and for unitaries, and we will use here these results, in order to deal with the general normal case. As a first observation, given an arbitrary matrix A, the matrix AA^* is self-adjoint:

$$(AA^*)^* = AA^*$$

Thus, we can diagonalize this matrix AA^*, as follows, with the passage matrix being a unitary, $V \in U_N$, and with the diagonal form being real, $E \in M_N(\mathbb{R})$:

$$AA^* = VEV^*$$

Now observe that, for matrices of type $A = UDU^*$, which are those that we supposed to deal with, we have the following formulae:

$$V = U \quad , \quad E = D\bar{D}$$

In particular, the matrices A and AA^* have the same eigenspaces. So, this will be our idea, proving that the eigenspaces of AA^* are eigenspaces of A. In order to do so, let us pick two eigenvectors v, w of the matrix AA^*, corresponding to different eigenvalues, $\lambda \neq \mu$. The eigenvalue equations are then as follows:

$$AA^* v = \lambda v \quad , \quad AA^* w = \mu w$$

We have the following computation, using the normality condition $AA^* = A^* A$, and the fact that the eigenvalues of AA^*, and in particular μ, are real:

$$\begin{aligned}\lambda < Av, w > &= < \lambda Av, w > \\ &= < A\lambda v, w > \\ &= < AAA^* v, w > \\ &= < AA^* Av, w > \\ &= < Av, AA^* w > \\ &= < Av, \mu w > \\ &= \mu < Av, w >\end{aligned}$$

We conclude that we have $< Av, w >= 0$. But this reformulates as follows:

$$\lambda \neq \mu \implies A(E_\lambda) \perp E_\mu$$

Now since the eigenspaces of AA^* are pairwise orthogonal, and span the whole $\mathbb{C}^N$, we deduce from this that these eigenspaces are invariant under A:

$$A(E_\lambda) \subset E_\lambda$$

But with this result in hand, we can now finish. Indeed, we can decompose the problem, and the matrix A itself, following these eigenspaces of the matrix AA^*, which in practice amounts in saying that we can assume that we only have 1 eigenspace. By rescaling, this is the same as assuming that we have $AA^* = 1$, and so we are now into the unitary case, that we know how to solve, as explained in Theorem 3.11. □

As a first application of our latest spectral theorem, we have the following result:

THEOREM 3.21. *Given a matrix $A \in M_N(\mathbb{C})$, we can construct a matrix $|A|$ as follows, by using the fact that A^*A is diagonalizable, with positive eigenvalues:*

$$|A| = \sqrt{A^*A}$$

This matrix $|A|$ is then positive, and its square is $|A|^2 = A$. In the case $N = 1$, we obtain in this way the usual absolute value of the complex numbers.

PROOF. Consider indeed the matrix A^*A, which is normal. According to Theorem 3.20, we can diagonalize this matrix as follows, with $U \in U_N$, and with D diagonal:

$$A = UDU^*$$

From $A^*A \geq 0$ we obtain $D \geq 0$. But this means that the entries of D are real, and positive. Thus we can extract the square root $\sqrt{D}$, and then set:

$$\sqrt{A^*A} = U\sqrt{D}U^*$$

Thus, we are basically done. Indeed, if we call this latter matrix $|A|$, then we are led to the conclusions in the statement. Finally, the last assertion is clear from definitions. □

As a comment here, it is possible to talk as well about $\sqrt{AA^*}$, which is in general different from $\sqrt{A^*A}$. Note that when A is normal, there is no issue, because we have:

$$AA^* = A^*A \implies \sqrt{AA^*} = \sqrt{A^*A}$$

Now with the above in hand, we can talk about polar decomposition. Let us start with a weak version of the result, regarding the invertible matrices, as follows:

THEOREM 3.22. *Any invertible matrix $A \in M_N(\mathbb{C})$ decomposes as*

$$A = U|A|$$

with $U \in U_N$, and with $|A| = \sqrt{A^*A}$ as above.

PROOF. According to our definition of the modulus, $|A| = \sqrt{A^*A}$, we have:

$$\begin{aligned} < |A|x, |A|y > &= < x, |A|^2 y > \\ &= < x, A^*Ay > \\ &= < Ax, Ay > \end{aligned}$$

Thus we can define a unitary matrix $U \in U_N$ by the following formula:

$$U(|A|x) = Ax$$

But this formula shows that we have $A = U|A|$, as desired. □

Observe that we have uniqueness in the above result, in what regards the choice of the unitary $U \in U_N$, due to the fact that we can write this unitary as follows:

$$U = A(\sqrt{A^*A})^{-1}$$

More generally now, we have the following result, dealing with arbitrary matrices:

THEOREM 3.23. *Any matrix $A \in M_N(\mathbb{C})$ decomposes as*

$$A = U|A|$$

with U being a partial isometry, and with $|A| = \sqrt{A^*A}$ as above.

PROOF. As before, we have the following equality, for any two vectors $x, y \in \mathbb{C}^N$:

$$< |A|x, |A|y > = < Ax, Ay >$$

We conclude that the following linear application is well-defined, and isometric, with the convention that we identify here the matrices with the associated linear maps:

$$U : Im|A| \to Im(A) \quad , \quad |A|x \to Ax$$

Moreover, we can further extend U into a partial isometry $U : \mathbb{C}^N \to \mathbb{C}^N$, by setting $Ux = 0$, for any $x \in \overline{Im|A|}^{\perp}$, and with this convention, the result follows. □

3d. Spectral measures

We would like to discuss now some interesting applications of our various spectral theorems to probability theory. Let us start with something basic, as follows:

DEFINITION 3.24. *Let X be a probability space, that is, a space with a probability measure, and with the corresponding integration denoted E, and called expectation.*

(1) *The random variables are the real functions $f \in L^\infty(X)$.*
(2) *The moments of such a variable are the numbers $M_k(f) = E(f^k)$.*
(3) *The law of such a variable is the measure given by $M_k(f) = \int_{\mathbb{R}} x^k d\mu_f(x)$.*

Here, and in what follows, we use the term "law" for "probability distribution", which means exactly the same thing, and is more convenient. Regarding now the fact that the law μ_f exists indeed, this is true, but not exactly trivial. By linearity, we would like to have a probability measure making hold the following formula, for any $P \in \mathbb{C}[X]$:

$$E(P(f)) = \int_{\mathbb{R}} P(x) d\mu_f(x)$$

By using a standard continuity argument, it is enough to have this formula for the characteristic functions χ_I of the arbitrary measurable sets of real numbers $I \subset \mathbb{R}$:

$$E(\chi_I(f)) = \int_{\mathbb{R}} \chi_I(x) d\mu_f(x)$$

But this latter formula, which reads $P(f \in I) = \mu_f(I)$, can serve as a definition for μ_f, and we are done. Alternatively, assuming some familiarity with measure theory, μ_f is the push-forward of the probability measure on X, via the function $f : X \to \mathbb{R}$.

Let us summarize this discussion in the form of a theorem, as follows:

THEOREM 3.25. *The law μ_f of a random variable f exists indeed, and we have*

$$E(\varphi(f)) = \int_{\mathbb{R}} \varphi(x) d\mu_f(x)$$

for any integrable function $\varphi : \mathbb{R} \to \mathbb{C}$.

PROOF. This follows from the above discussion, and with the precise assumption on $\varphi : \mathbb{R} \to \mathbb{C}$, which is its integrability, in the abstract mathematical sense, being in fact something that we will not really need, in what follows. In fact, for most purposes we will get away with polynomials $\varphi \in \mathbb{C}[X]$, and by linearity this means that we can get away with monomials $\varphi(x) = x^k$, which brings us back to Definition 3.24 (3), as stated. □

Getting now to the case of the matrices $A \in M_N(\mathbb{C})$, here it is quite tricky to figure out what the law of A should mean, based on intuition only. So, in the lack of a bright idea, let us just reproduce Definition 3.24, with a few modifications, as follows:

DEFINITION 3.26. *Let $N \in \mathbb{N}$, and consider the algebra $M_N(\mathbb{C})$ of complex $N \times N$ matrices, with its normalized trace $tr : M_N(\mathbb{C}) \to \mathbb{C}$, given by $tr(A) = Tr(A)/N$.*

(1) *We call random variables the self-adjoint matrices $A \in M_N(\mathbb{C})$.*
(2) *The moments of such a variable are the numbers $M_k(A) = tr(A^k)$.*
(3) *The law of such a variable is the measure given by $M_k(A) = \int_{\mathbb{R}} x^k d\mu_A(x)$.*

Here we have normalized the trace, as to have $tr(1) = 1$, in analogy with the formula $E(1) = 1$ from usual probability. By the way, as a piece of advice here, many confusions appear from messing up tr and Tr, and it is better ot forget about Tr, and always use tr. With the drawback that if you're a physicist, tr might get messed up in quick handwriting

with the reduced Planck constant $\hbar = h/2\pi$. However, shall you ever face this problem, I have an advice here too, namely forgetting about h, and using h instead of $\hbar$.

Another comment is that we assumed in (1) that our matrix is self-adjoint, $A = A^*$, with the adjoint matrix being given, as usual, by the formula $(A^*)_{ij} = \bar{A}_{ji}$. Why this, because for instance at $N = 1$ we would like our matrix, which in the case $N = 1$ is a number, to be real, and so we must assume $A = A^*$. Of course there is still some discussion here, for instance because you might argue that why not assuming instead that the entries of A are real. But let us leave this for later, and in the meantime, just trust me. Or perhaps, let us both trust Heisenberg, who was the first intensive user of complex matrices, and who declared that such matrices must be self-adjoint. More later.

Back to work now, what we have in Definition 3.26 looks quite reasonable, but as before with the usual random variables $f \in L^\infty(X)$, some discussion is needed, in order to understand if the law μ_A exists indeed, and by which mechanism. And, good news here, in the case of the simplest matrices, the real diagonal ones, we have:

THEOREM 3.27. *For any diagonal matrix $A \in M_N(\mathbb{R})$ we have the formula*

$$tr(P(A)) = \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

where $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ are the diagonal entries of A. Thus the measure

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

can be regarded as being the law of A, in the sense of Definition 3.26.

PROOF. Assume indeed that we have a real diagonal matrix, as follows, with the convention that the matrix entries which are missing are by definition 0 entries:

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

The powers of A are then diagonal too, given by the following formula:

$$A^k = \begin{pmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_N^k \end{pmatrix}$$

In fact, given any polynomial $P \in \mathbb{C}[X]$, we have the following formula:

$$P(A) = \begin{pmatrix} P(\lambda_1) & & \\ & \ddots & \\ & & P(\lambda_N) \end{pmatrix}$$

Thus, the first formula in the statement holds indeed. In particular, we conclude that the moments of A are given by the following formula:

$$M_k(A) = tr(A^k) = \frac{1}{N}\sum_i \lambda_i^k$$

On the other hand, with $\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$ as in the statement, we have:

$$\begin{aligned}\int_{\mathbb{R}} x^k d\mu_A(x) &= \frac{1}{N}\sum_i \int_{\mathbb{R}} x^k d\delta_{\lambda_i}(x)\\ &= \frac{1}{N}\sum_i \lambda_i^k\end{aligned}$$

Thus that the law of A exists indeed, and is the measure μ_A, as claimed. □

The point now is that, by using the spectral theorem for self-adjoint matrices, we have the following generalization of Theorem 3.28, dealing with the general case:

THEOREM 3.28. *For a self-adjoint matrix $A \in M_N(\mathbb{C})$ we have the formula*

$$tr(P(A)) = \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

where $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ are the eigenvalues of A. Thus the measure

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

can be regarded as being the law of A, in the sense of Definition 3.26.

PROOF. We already know, from Theorem 3.27, that the result holds indeed for the diagonal matrices. In the general case now, that of an arbitrary self-adjoint matrix, we know from Theorem 3.1 that our matrix is diagonalizable, as follows:

$$A = UDU^*$$

Now observe that the moments of A are given by the following formula:

$$\begin{aligned}tr(A^k) &= tr(UDU^* \cdot UDU^* \ldots UDU^*)\\ &= tr(UD^kU^*)\\ &= tr(D^k)\end{aligned}$$

We conclude from this, by reasoning by linearity, that the matrices A, D have the same law, $\mu_A = \mu_D$, and this gives all the assertions in the statement. □

The above theory is not the end of the story, because we can talk about complex random variables, $f : X \to \mathbb{C}$, and about non-self-adjoint matrices too, $A \neq A^*$. We will see that, with a bit of know-how, we can have some law technology going on, for both.

Let us start with the complex variables $f \in L^\infty(X)$. The main difference with respect to the real case comes from the fact that we have now a pair of variables instead of one, namely $f : X \to \mathbb{C}$ itself, and its conjugate $\bar{f} : X \to \mathbb{C}$. Thus, we are led to:

DEFINITION 3.29. *The moments a complex variable $f \in L^\infty(X)$ are the numbers*

$$M_k(f) = E(f^k)$$

depending on colored integers $k = \circ\bullet\bullet\circ\ldots$, with the conventions

$$f^\emptyset = 1 \quad , \quad f^\circ = f \quad , \quad f^\bullet = \bar{f}$$

and multiplicativity, in order to define the colored powers f^k.

Observe that, since $f, \bar{f}$ commute, we can permute terms, and restrict the attention to exponents of type $k = \ldots\circ\circ\circ\bullet\bullet\bullet\bullet\ldots$, if we want to. However, our various results below will look better without doing this, so we will use Definition 3.29 as stated.

Regarding now the notion of law, this extends too, the result being as follows:

THEOREM 3.30. *Each complex variable $f \in L^\infty(X)$ has a law, which is by definition a complex probability measure μ_f making the following formula hold,*

$$M_k(f) = \int_{\mathbb{C}} z^k d\mu_f(z)$$

for any colored integer k. Moreover, we have in fact the formula

$$E(\varphi(f)) = \int_{\mathbb{C}} \varphi(x) d\mu_f(x)$$

valid for any integrable function $\varphi : \mathbb{C} \to \mathbb{C}$.

PROOF. The first assertion follows exactly as in the real case, and with z^k being defined exactly as f^k, namely by the following formulae, and multiplicativity:

$$z^\emptyset = 1 \quad , \quad z^\circ = z \quad , \quad z^\bullet = \bar{z}$$

As for the second assertion, this basically follows from this by linearity and continuity, by using standard measure theory, again as in the real case. □

Moving ahead towards matrices, all this leads to a mixture of easy and complicated problems. First, Definition 3.29 has the following straightforward analogue:

DEFINITION 3.31. *The moments a matrix $A \in M_N(\mathbb{C})$ are the numbers*

$$M_k(A) = tr(A^k)$$

depending on colored integers $k = \circ\bullet\bullet\circ\ldots$, with the usual conventions

$$A^\emptyset = 1 \quad , \quad A^\circ = A \quad , \quad A^\bullet = A^*$$

and multiplicativity, in order to define the colored powers A^k.

As a first observation about this, unless the matrix is normal, $AA^* = A^*A$, we cannot switch to exponents of type $k = \ldots \circ\circ\circ\bullet\bullet\bullet\bullet\ldots$, as it was theoretically possible for the complex variables $f \in L^\infty(X)$. Here is an explicit counterexample for this:

PROPOSITION 3.32. *The following matrix, which is not normal,*

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

has the property $tr(JJ^*JJ^*) \neq tr(JJJ^*J^*)$.

PROOF. We have the following formulae, which show that J is not normal:

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

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

Let us compute now the quantities in the statement. We first have:

$$tr(JJ^*JJ^*) = tr((JJ^*)^2) = tr \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}$$

On the other hand, we have as well the following formula:

$$tr(JJJ^*J^*) = tr \left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right) = tr \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0$$

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

The above counterexample makes it quite clear that things will be complicated, when attempting to talk about the law of an arbitrary matrix $A \in M_N(\mathbb{C})$. But, there is solution to everything. By being a bit smart, we can formulate things as follows:

DEFINITION 3.33. *The law of a complex matrix $A \in M_N(\mathbb{C})$ is the following functional, on the algebra of polynomials in two noncommuting variables X, X^*:*

$$\mu_A : \mathbb{C} < X, X^* > \to \mathbb{C} \quad , \quad P \to tr(P(A))$$

In the case where we have a complex probability measure $\mu_A \in \mathcal{P}(\mathbb{C})$ such that

$$tr(P(A)) = \int_{\mathbb{C}} P(x)\, d\mu_A(x)$$

we identify this complex measure with the law of A.

As mentioned above, this is something smart, that will take us some time to understand. As a first observation, knowing the law is the same as knowing the moments, because if we write our polynomial as $P = \sum_k c_k X^k$, then we have:

$$tr(P(A)) = tr\left(\sum_k c_k A^k\right) = \sum_k c_k M_k(A)$$

Let us try now to compute some matrix laws, and see what we get. We already did some computations in the real case, and then for the basic 2×2 Jordan block J too, and based on all this, we can formulate the following result, with mixed conclusions:

THEOREM 3.34. *The following happen:*

(1) *If $A = A^*$ then $\mu_A = \frac{1}{N}(\lambda_1 + \ldots + \lambda_N)$, with $\lambda_i \in \mathbb{R}$ being the eigenvalues.*
(2) *If A is diagonal, $\mu_A = \frac{1}{N}(\lambda_1 + \ldots + \lambda_N)$, with $\lambda_i \in \mathbb{C}$ being the eigenvalues.*
(3) *For the basic Jordan block J, the law μ_J is not a complex measure.*
(4) *In fact, assuming $AA^* \neq A^*A$, the law μ_A is not a complex measure.*

PROOF. This follows from the above, with only (4) being new. Assuming $AA^* \neq A^*A$, in order to show that μ_A is not a measure, we can use a positivity trick, as follows:

$$\begin{aligned}
AA^* - A^*A \neq 0 &\implies (AA^* - A^*A)^2 > 0\\
&\implies AA^*AA^* - AA^*A^*A - A^*AAA^* + A^*AA^*A > 0\\
&\implies tr(AA^*AA^* - AA^*A^*A - A^*AAA^* + A^*AA^*A) > 0\\
&\implies tr(AA^*AA^* + A^*AA^*A) > tr(AA^*A^*A + A^*AAA^*)\\
&\implies tr(AA^*AA^*) > tr(AAA^*A^*)
\end{aligned}$$

Thus, we can conclude as in the proof for J, the point being that we cannot obtain both the above numbers by integrating $|z|^2$ with respect to a measure $\mu_A \in \mathcal{P}(\mathbb{C})$. □

Fortunately, by using the spectral theorem for normal matrices, we have:

THEOREM 3.35. *Given a matrix $A \in M_N(\mathbb{C})$ which is normal, $AA^* = A^*A$, we have the following formula, valid for any polynomial $P \in \mathbb{C} < X, X^* >$,*

$$tr(P(A)) = \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

where $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ are the eigenvalues of A. Thus the complex measure

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

is the law of A. In the non-normal case, the law μ_A is not a measure.

PROOF. As before in the diagonal case, since our matrix is normal, $AA^* = A^*A$, knowing its law in the abstract sense of generalized probability is the same as knowing

the restriction of this abstract distribution to the usual polynomials in two variables:

$$\mu_A : \mathbb{C}[X, X^*] \to \mathbb{C} \quad , \quad P \to tr(P(A))$$

In order now to compute this functional, we can write $A = UDU^*$, as in Theorem 3.20, and then change the basis via U, which in practice means that we can simply assume $U = 1$. Thus if we denote by $\lambda_1, \ldots, \lambda_N$ the diagonal entries of D, which are the eigenvalues of A, the law that we are looking for is the following functional:

$$\mu_A : \mathbb{C}[X, X^*] \to \mathbb{C} \quad , \quad P \to \frac{1}{N}(P(\lambda_1) + \ldots + P(\lambda_N))$$

But this functional corresponds to integrating P with respect to the following complex measure, that we agree to still denote by μ_A, and call distribution of A:

$$\mu_A = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$$

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

We will be back to such things later, when discussing the random matrices.

3e. Exercises

This was a quite tricky chapter, and as exercises on this, we have:

EXERCISE 3.36. *Learn more about graphs, and related Laplace operators.*

EXERCISE 3.37. *Work out diagonalization results for various products of graphs.*

EXERCISE 3.38. *Find a geometric proof for the diagonalization of the rotation R_t.*

EXERCISE 3.39. *Learn more about the Pauli spin matrices, and their properties.*

EXERCISE 3.40. *Learn alternative proofs for the spectral theorem for normal matrices.*

EXERCISE 3.41. *Work out the polar decomposition of some matrices, of your choice.*

EXERCISE 3.42. *Learn some probability theory, as to be at ease with the above.*

EXERCISE 3.43. *Find formulae for the colored moments of the Jordan blocks.*

As bonus exercise, learn some operator theory as well, which is related to all this.

CHAPTER 4

Polynomials, roots

4a. Resultant

We have seen in the previous chapters that many linear algebra questions lead us into computing roots of polynomials $P \in \mathbb{C}[X]$. We will investigate here such questions. Let us start with something that we know well, but is always good to remember:

THEOREM 4.1. *The solutions of $ax^2 + bx + c = 0$ with $a, b, c \in \mathbb{C}$ are*

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

with the square root of complex numbers being defined as $\sqrt{re^{it}} = \sqrt{r}e^{it/2}$.

PROOF. We can indeed write our equation in the following way:

$$\begin{aligned} ax^2 + bx + c = 0 &\iff x^2 + \frac{b}{a}x + \frac{c}{a} = 0 \\ &\iff \left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0 \\ &\iff \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2} \\ &\iff x + \frac{b}{2a} = \pm\frac{\sqrt{b^2 - 4ac}}{2a} \end{aligned}$$

Here we have used the fact, mentioned in the statement, that any complex number $z = re^{it}$ has indeed a square root, given by $\sqrt{z} = \sqrt{r}e^{it/2}$, plus in fact a second square root as well, namely $-\sqrt{z}$. Thus, we are led to the conclusion in the statement. □

Moving now to degree 3 and higher, things here are far more complicated, and as a first objective, we would like to understand what the analogue of the discriminant $\Delta = b^2 - 4ac$ is. But even this is something quite tricky, because we would like to have $\Delta = 0$ precisely when $(P, P') \neq 1$, which leads us into the question of deciding, given two polynomials $P, Q \in \mathbb{C}[X]$, if these polynomials have a common root, $(P, Q) \neq 1$, or not.

Fortunately this latter question has a nice answer. We will need:

THEOREM 4.2. *Given a monic polynomial $P \in \mathbb{C}[X]$, factorized as*

$$P = (X - a_1) \ldots (X - a_k)$$

the following happen:

(1) *The coefficients of P are symmetric functions in $a_1, \ldots, a_k$.*
(2) *The symmetric functions in $a_1, \ldots, a_k$ are polynomials in the coefficients of P.*

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

(1) By expanding our polynomial, we have the following formula:

$$P = \sum_{r=0}^{k} (-1)^r \sum_{i_1 < \ldots < i_r} a_{i_1} \ldots a_{i_r} \cdot X^{k-r}$$

Thus the coefficients of P are, up to some signs, the following functions:

$$f_r = \sum_{i_1 < \ldots < i_r} a_{i_1} \ldots a_{i_r}$$

But these are indeed symmetric functions in $a_1, \ldots, a_k$, as claimed.

(2) Conversely now, let us look at the symmetric functions in the roots $a_1, \ldots, a_k$. These appear as linear combinations of the basic symmetric functions, given by:

$$S_r = \sum_i a_i^r$$

Moreover, when allowing polynomials instead of linear combinations, we need in fact only the first k such sums, namely $S_1, \ldots, S_k$. That is, the symmetric functions $\mathcal{F}$ in our variables $a_1, \ldots, a_k$, with integer coefficients, appear as follows:

$$\mathcal{F} = \mathbb{Z}[S_1, \ldots, S_k]$$

(3) The point now is that, alternatively, the symmetric functions in our variables $a_1, \ldots, a_k$ appear as well as linear combinations of the functions f_r that we found in (1), and that when allowing polynomials instead of linear combinations, we need in fact only the first k functions, namely $f_1, \ldots, f_k$. That is, we have as well:

$$\mathcal{F} = \mathbb{Z}[f_1, \ldots, f_k]$$

But this gives the result, because we can pass from $\{S_r\}$ to $\{f_r\}$, and vice versa.

(4) This was for the idea, and in practice now up to you to clarify all the details. In fact, we will also need in what follows the extension of all this to the case where P is no longer assumed to be monic, and with this being, again, exercise for you. □

Getting back now to our original question, namely that of deciding whether two polynomials $P, Q \in \mathbb{C}[X]$ have a common root or not, this has the following nice answer:

THEOREM 4.3. *Given two polynomials $P, Q \in \mathbb{C}[X]$, written as*

$$P = c(X - a_1) \ldots (X - a_k) \quad , \quad Q = d(X - b_1) \ldots (X - b_l)$$

the following quantity, which is called resultant of P, Q,

$$R(P, Q) = c^l d^k \prod_{ij} (a_i - b_j)$$

is a certain polynomial in the coefficients of P, Q, with integer coefficients, and we have $R(P, Q) = 0$ precisely when P, Q have a common root.

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

(1) Given two polynomials $P, Q \in \mathbb{C}[X]$, we can certainly construct the quantity $R(P, Q)$ in the statement, with the role of the normalization factor $c^l d^k$ to become clear later on, and then we have $R(P, Q) = 0$ precisely when P, Q have a common root:

$$R(P, Q) = 0 \iff \exists i, j, a_i = b_j$$

(2) As bad news, however, this quantity $R(P, Q)$, defined in this way, is a priori not very useful in practice, because it depends on the roots a_i, b_j of our polynomials P, Q, that we cannot compute in general. However, and here comes our point, as we will prove below, it turns out that $R(P, Q)$ is in fact a polynomial in the coefficients of P, Q, with integer coefficients, and this is where the power of $R(P, Q)$ comes from.

(3) You might perhaps say, nice, but why not doing things the other way around, that is, formulating our theorem with the explicit formula of $R(P, Q)$, in terms of the coefficients of P, Q, and then proving that we have $R(P, Q) = 0$, via roots and everything. Good point, but this is not exactly obvious, the formula of $R(P, Q)$ in terms of the coefficients of P, Q being something terribly complicated. In short, trust me, let us prove our theorem as stated, and for alternative formulae of $R(P, Q)$, we will see later.

(4) Getting started now, let us expand the formula of $R(P, Q)$, by making all the multiplications there, abstractly, in our head. Everything being symmetric in $a_1, \ldots, a_k$, we obtain in this way certain symmetric functions in these variables, which will be therefore certain polynomials in the coefficients of P. Moreover, due to our normalization factor c^l, these polynomials in the coefficients of P will have integer coefficients.

(5) With this done, let us look now what happens with respect to the remaining variables $b_1, \ldots, b_l$, which are the roots of Q. Once again what we have here are certain symmetric functions in these variables $b_1, \ldots, b_l$, and these symmetric functions must be certain polynomials in the coefficients of Q. Moreover, due to our normalization factor d^k, these polynomials in the coefficients of Q will have integer coefficients.

(6) Thus, we are led to the conclusion in the statement, that $R(P, Q)$ is a polynomial in the coefficients of P, Q, with integer coefficients, and with the remark that the $c^l d^k$ factor is there for these latter coefficients to be indeed integers, instead of rationals. □

All the above might seem a bit complicated, so as an illustration, let us work out an example. Consider the case of a polynomial of degree 2, and a polynomial of degree 1:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

In order to compute the resultant, let us factorize our polynomials:

$$P = a(x - p)(x - q) \quad , \quad Q = d(x - r)$$

The resultant can be then computed as follows, by using the method above:

$$\begin{aligned} R(P, Q) &= ad^2(p - r)(q - r) \\ &= ad^2(pq - (p + q)r + r^2) \\ &= cd^2 + bd^2r + ad^2r^2 \\ &= cd^2 - bde + ae^2 \end{aligned}$$

Finally, observe that $R(P, Q) = 0$ corresponds indeed to the fact that P, Q have a common root. Indeed, the root of Q is $r = -e/d$, and we have:

$$P(r) = \frac{ae^2}{d^2} - \frac{be}{d} + c = \frac{R(P, Q)}{d^2}$$

Regarding now the explicit formula of the resultant $R(P, Q)$, this is something quite complicated, and there are several methods for dealing with this problem. We have:

THEOREM 4.4. *The resultant of two polynomials, written as*

$$P = p_kX^k + \ldots + p_1X + p_0 \quad , \quad Q = q_lX^l + \ldots + q_1X + q_0$$

appears as the determinant of an associated matrix, as follows,

$$R(P, Q) = \begin{vmatrix} p_k & & & q_l & & \\ \vdots & \ddots & & \vdots & \ddots & \\ p_0 & & p_k & q_0 & & q_l \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_0 & & & q_0 \end{vmatrix}$$

with the matrix having size $k + l$, and having 0 coefficients at the blank spaces.

PROOF. This is something clever, due to Sylvester, as follows:

(1) Consider the vector space $\mathbb{C}_k[X]$ formed by the polynomials of degree $< k$:

$$\mathbb{C}_k[X] = \left\{ P \in \mathbb{C}[X] \middle| \deg P < k \right\}$$

This is a vector space of dimension k, having as basis the monomials $1, X, \ldots, X^{k-1}$. Now given polynomials P, Q as in the statement, consider the following linear map:

$$\Phi : \mathbb{C}_l[X] \times \mathbb{C}_k[X] \to \mathbb{C}_{k+l}[X] \quad , \quad (A, B) \to AP + BQ$$

(2) Our first claim is that with respect to the standard bases for all the vector spaces involved, namely those consisting of the monomials $1, X, X^2, \ldots$, the matrix of Φ is the matrix in the statement. But this is something which is clear from definitions.

(3) Our second claim is that $\det \Phi = 0$ happens precisely when P, Q have a common root. Indeed, our polynomials P, Q having a common root means that we can find A, B such that $AP + BQ = 0$, and so that $(A, B) \in \ker \Phi$, which reads $\det \Phi = 0$.

(4) Finally, our claim is that we have $\det \Phi = R(P, Q)$. But this follows from the uniqueness of the resultant, up to a scalar, and with this uniqueness property being elementary to establish, along the lines of the proofs of Theorems 4.2 and 4.3. □

In what follows we will not really need the above formula, so let us just check now that this formula works indeed. Consider our favorite polynomials, as before:

$$P = ax^2 + bx + c \quad , \quad Q = dx + e$$

According to the above result, the resultant should be then, as it should:

$$R(P, Q) = \begin{vmatrix} a & d & 0 \\ b & e & d \\ c & 0 & e \end{vmatrix} = ae^2 - bde + cd^2$$

We will be back to more computations of resultants later.

4b. Discriminant

We can go back now to our original question regarding discriminants, and we have:

THEOREM 4.5. *Given a polynomial $P \in \mathbb{C}[X]$, written as*

$$P(X) = aX^N + bX^{N-1} + cX^{N-2} + \ldots$$

its discriminant, defined as being the following quantity,

$$\Delta(P) = \frac{(-1)^{\binom{N}{2}}}{a} R(P, P')$$

is a polynomial in the coefficients of P, with integer coefficients, and $\Delta(P) = 0$ happens precisely when P has a double root.

PROOF. The fact that the discriminant $\Delta(P)$ is a polynomial in the coefficients of P, with integer coefficients, comes from Theorem 4.3, coupled with the fact that the division by the leading coefficient a is indeed possible, under $\mathbb{Z}$, as being shown by the following

formula, which is written of course a bit informally, coming from Theorem 4.4:

$$R(P,P')=\begin{vmatrix} a & & & Na & & \\ \vdots & \ddots & & \vdots & \ddots & \\ z & & a & y & & Na \\ & \ddots & \vdots & & \ddots & \vdots \\ & & z & & & y \end{vmatrix}$$

Also, the fact that we have $\Delta(P)=0$ precisely when P has a double root is clear from Theorem 4.3. Finally, let us mention that the sign $(-1)^{\binom{N}{2}}$ is there for various reasons, including the compatibility with some well-known formulae, at small values of $N\in\mathbb{N}$, such as $\Delta(P)=b^2-4ac$ in degree 2, that we will discuss in a moment. □

As already mentioned, by using Theorem 4.4, we have an explicit formula for the discriminant, as the determinant of a certain matrix. There is a lot of theory here, and in order to get into this, let us first see what happens in degree 2. Here we have:

$$P=aX^2+bX+c\quad,\quad P'=2aX+b$$

Thus, the resultant is given by the following formula:

$$\begin{aligned} R(P,P') &= ab^2-b(2a)b+c(2a)^2\\ &= 4a^2c-ab^2\\ &= -a(b^2-4ac) \end{aligned}$$

It follows that the discriminant of our polynomial is, as it should:

$$\Delta(P)=b^2-4ac$$

Alternatively, we can use the formula in Theorem 4.4, and we obtain:

$$\begin{aligned} \Delta(P) &= -\frac{1}{a}\begin{vmatrix} a & 2a & \\ b & b & 2a \\ c & & b \end{vmatrix}\\ &= -\begin{vmatrix} 1 & 2 & \\ b & b & 2a \\ c & & b \end{vmatrix}\\ &= -b^2+2(b^2-2ac)\\ &= b^2-4ac \end{aligned}$$

We will be back later to such formulae, in degree 3, and in degree 4 as well, with the comment however, coming in advance, that these formulae are not very beautiful.

At the theoretical level now, we have the following result, which is not trivial:

THEOREM 4.6. *The discriminant of a polynomial P is given by the formula*

$$\Delta(P) = a^{2N-2}\prod_{i<j}(r_i - r_j)^2$$

where a is the leading coefficient, and $r_1, \ldots, r_N$ are the roots.

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

(1) The first thought goes to the formula in Theorem 4.3, so let us see what that formula teaches us, in the case $Q = P'$. Let us write P, P' as follows:

$$P = a(x - r_1)\ldots(x - r_N)$$

$$P' = Na(x - p_1)\ldots(x - p_{N-1})$$

According to Theorem 4.3, the resultant of P, P' is then given by:

$$R(P, P') = a^{N-1}(Na)^N\prod_{ij}(r_i - p_j)$$

And bad news, this is not exactly what we wished for, namely the formula in the statement. That is, we are on the good way, but certainly have to work some more.

(2) Obviously, we must get rid of the roots $p_1, \ldots, p_{N-1}$ of the polynomial P'. In order to do this, let us rewrite the formula that we found in (1) in the following way:

$$\begin{aligned} R(P, P') &= N^N a^{2N-1}\prod_i\left(\prod_j(r_i - p_j)\right) \\ &= N^N a^{2N-1}\prod_i\frac{P'(r_i)}{Na} \\ &= a^{N-1}\prod_i P'(r_i) \end{aligned}$$

(3) In order to compute now P', and more specifically the values $P'(r_i)$ that we are interested in, we can use the Leibnitz rule. So, consider our polynomial:

$$P(x) = a(x - r_1)\ldots(x - r_N)$$

The Leibnitz rule for derivatives tells us that $(fg)' = f'g + fg'$, but then also that $(fgh)' = f'gh + fg'h + fgh'$, and so on. Thus, for our polynomial, we obtain:

$$P'(x) = a\sum_i(x - r_1)\ldots\underbrace{(x - r_i)}_{missing}\ldots(x - r_N)$$

Now when applying this formula to one of the roots r_i, we obtain:

$$P'(r_i) = a(r_i - r_1)\ldots\underbrace{(r_i - r_i)}_{missing}\ldots(r_i - r_N)$$

By making now the product over all indices i, this gives the following formula:

$$\prod_i P'(r_i) = a^N \prod_{i\neq j}(r_i - r_j)$$

(4) Time now to put everything together. By taking the formula in (2), making the normalizations in Theorem 4.5, and then using the formula found in (3), we obtain:

$$\begin{aligned}\Delta(P) &= (-1)^{\binom{N}{2}} a^{N-2} \prod_i P'(r_i)\\ &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i\neq j}(r_i - r_j)\end{aligned}$$

(5) This is already a nice formula, which is very useful in practice, and that we can safely keep as a conclusion, to our computations. However, we can do slightly better, by grouping opposite terms. Indeed, this gives the following formula:

$$\begin{aligned}\Delta(P) &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i\neq j}(r_i - r_j)\\ &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i<j}(r_i - r_j) \cdot \prod_{i>j}(r_i - r_j)\\ &= (-1)^{\binom{N}{2}} a^{2N-2} \prod_{i<j}(r_i - r_j) \cdot (-1)^{\binom{N}{2}} \prod_{i<j}(r_i - r_j)\\ &= a^{2N-2} \prod_{i<j}(r_i - r_j)^2\end{aligned}$$

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

As applications now, the formula in Theorem 4.6 is quite useful for the real polynomials $P \in \mathbb{R}[X]$ in small degree, because it allows to say when the roots are real, or complex, or at least have some partial information about this. For instance, we have:

PROPOSITION 4.7. *Consider a polynomial with real coefficients, $P \in \mathbb{R}[X]$, assumed for simplicity to have nonzero discriminant, $\Delta \neq 0$.*

(1) *In degree 2, the roots are real when $\Delta > 0$, and complex when $\Delta < 0$.*
(2) *In degree 3, all roots are real precisely when $\Delta > 0$.*

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

(1) The first assertion is something that you certainly know well, since ages, formally coming from Theorem 4.1, but let us see how this comes via the formula in Theorem 4.6. In degree $N = 2$, this formula looks as follows, with r_1, r_2 being the roots:

$$\Delta(P) = a^2(r_1 - r_2)^2$$

Thus $\Delta > 0$ amounts in saying that we have $(r_1 - r_2)^2 > 0$. Now since r_1, r_2 are conjugate, and with this being something trivial, meaning no need here for the computations in Theorem 4.1, we conclude that $\Delta > 0$ means that r_1, r_2 are real, as stated.

(2) In degree $N = 3$ now, we know from analysis that P has at least one real root, and the problem is whether the remaining 2 roots are real, or complex conjugate. For this purpose, we can use the formula in Theorem 4.6, which in degree 3 reads:

$$\Delta(P) = a^4(r_1 - r_2)^2(r_1 - r_3)^2(r_2 - r_3)^2$$

We can see that in the case $r_1, r_2, r_3 \in \mathbb{R}$, we have $\Delta(P) > 0$. Conversely now, assume that $r_1 = r$ is the real root, coming from analysis, and that the other roots are $r_2 = z$ and $r_3 = \bar{z}$, with z being a complex number, which is not real. We have then:

$$\begin{aligned}\Delta(P) &= a^4(r - z)^2(r - \bar{z})^2(z - \bar{z})^2\\ &= a^4|r - z|^4(2iIm(z))^2\\ &= -4a^4|r - z|^4 Im(z)^2\\ &< 0\end{aligned}$$

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

4c. Low dimensions

Let us work out now in detail what happens in degree 3, with the explicit computation of the discriminant, in terms of the coefficients. Here the result is as follows:

THEOREM 4.8. *The discriminant of a degree* 3 *polynomial,*

$$P = aX^3 + bX^2 + cX + d$$

is the number $\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$.

PROOF. We have two methods available, based on Theorem 4.3 and Theorem 4.4, and both being instructive, we will try them both. The computations are as follows:

(1) Let us first go the pedestrian way, based on the definition of the resultant, from Theorem 4.3. Consider two polynomials, of degree 3 and degree 2, written as follows:

$$P = aX^3 + bX^2 + cX + d$$

$$Q = eX^2 + fX + g = e(X - s)(X - t)$$

The resultant of these two polynomials is then given by:

$$\begin{aligned}R(P, Q) &= a^2e^3(p - s)(p - t)(q - s)(q - t)(r - s)(r - t)\\ &= a^2 \cdot e(p - s)(p - t) \cdot e(q - s)(q - t) \cdot e(r - s)(r - t)\\ &= a^2Q(p)Q(q)Q(r)\\ &= a^2(ep^2 + fp + g)(eq^2 + fq + g)(er^2 + fr + g)\end{aligned}$$

By expanding, we obtain the following formula for this resultant:

$$\begin{aligned}\frac{R(P,Q)}{a^2} &= e^3p^2q^2r^2+e^2f(p^2q^2r+p^2qr^2+pq^2r^2)\\ &+ e^2g(p^2q^2+p^2r^2+q^2r^2)+ef^2(p^2qr+pq^2r+pqr^2)\\ &+ efg(p^2q+pq^2+p^2r+pr^2+q^2r+qr^2)+f^3pqr\\ &+ eg^2(p^2+q^2+r^2)+f^2g(pq+pr+qr)\\ &+ fg^2(p+q+r)+g^3\end{aligned}$$

Note in passing that we have 27 terms on the right, as we should, and with this kind of check being mandatory, when doing such computations. Next, we have:

$$p+q+r=-\frac{b}{a}\quad,\quad pq+pr+qr=\frac{c}{a}\quad,\quad pqr=-\frac{d}{a}$$

By using these formulae, we can produce some more, as follows:

$$p^2+q^2+r^2=(p+q+r)^2-2(pq+pr+qr)=\frac{b^2}{a^2}-\frac{2c}{a}$$

$$p^2q+pq^2+p^2r+pr^2+q^2r+qr^2=(p+q+r)(pq+pr+qr)-3pqr=-\frac{bc}{a^2}+\frac{3d}{a}$$

$$p^2q^2+p^2r^2+q^2r^2=(pq+pr+qr)^2-2pqr(p+q+r)=\frac{c^2}{a^2}-\frac{2bd}{a^2}$$

By plugging now this data into the formula of $R(P,Q)$, we obtain:

$$\begin{aligned}R(P,Q) &= a^2e^3\cdot\frac{d^2}{a^2}-a^2e^2f\cdot\frac{cd}{a^2}+a^2e^2g\left(\frac{c^2}{a^2}-\frac{2bd}{a^2}\right)+a^2ef^2\cdot\frac{bd}{a^2}\\ &+ a^2efg\left(-\frac{bc}{a^2}+\frac{3d}{a}\right)-a^2f^3\cdot\frac{d}{a}\\ &+ a^2eg^2\left(\frac{b^2}{a^2}-\frac{2c}{a}\right)+a^2f^2g\cdot\frac{c}{a}-a^2fg^2\cdot\frac{b}{a}+a^2g^3\end{aligned}$$

Thus, we have the following formula for the resultant:

$$\begin{aligned}R(P,Q) &= d^2e^3-cde^2f+c^2e^2g-2bde^2g+bdef^2-bcefg+3adefg\\ &- adf^3+b^2eg^2-2aceg^2+acf^2g-abfg^2+a^2g^3\end{aligned}$$

Getting back now to our discriminant problem, with $Q=P'$, which corresponds to $e=3a$, $f=2b$, $g=c$, we obtain the following formula:

$$\begin{aligned}R(P,P') &= 27a^3d^2-18a^2bcd+9a^2c^3-18a^2bcd+12ab^3d-6ab^2c^2+18a^2bcd\\ &- 8ab^3d+3ab^2c^2-6a^2c^3+4ab^2c^2-2ab^2c^2+a^2c^3\end{aligned}$$

By simplifying terms, and dividing by a, we obtain the following formula:

$$-\Delta(P)=27a^2d^2-18abcd+4ac^3+4b^3d-b^2c^2$$

But this gives the formula in the statement, namely:

$$\Delta(P) = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$$

(2) Let us see as well how the computation does, by using Theorem 4.4, which is our most advanced tool, so far. Consider a polynomial of degree 3, and its derivative:

$$P = aX^3 + bX^2 + cX + d$$

$$P' = 3aX^2 + 2bX + c$$

By using now Theorem 4.4 and computing the determinant, we obtain:

$$\begin{aligned}
R(P,P') &= \begin{vmatrix} a & & 3a & & \\ b & a & 2b & 3a & \\ c & b & c & 2b & 3a \\ d & c & & c & 2b \\ & d & & & c \end{vmatrix} \\
&= \begin{vmatrix} a & & & & \\ b & a & -b & 3a & \\ c & b & -2c & 2b & 3a \\ d & c & -3d & c & 2b \\ & d & & & c \end{vmatrix} \\
&= a\begin{vmatrix} a & -b & 3a & \\ b & -2c & 2b & 3a \\ c & -3d & c & 2b \\ d & & & c \end{vmatrix} \\
&= -ad\begin{vmatrix} -b & 3a & \\ -2c & 2b & 3a \\ -3d & c & 2b \end{vmatrix} + ac\begin{vmatrix} a & -b & 3a \\ b & -2c & 2b \\ c & -3d & c \end{vmatrix} \\
&= -ad(-4b^3 - 27a^2d + 12abc + 3abc) \\
&\quad +ac(-2ac^2 - 2b^2c - 9abd + 6ac^2 + b^2c + 6abd) \\
&= a(4b^3d + 27a^2d^2 - 15abcd + 4ac^3 - b^2c^2 - 3abcd) \\
&= a(4b^3d + 27a^2d^2 - 18abcd + 4ac^3 - b^2c^2)
\end{aligned}$$

Now according to Theorem 4.5, the discriminant of our polynomial is given by:

$$\begin{aligned}
\Delta(P) &= -\frac{R(P,P')}{a} \\
&= -4b^3d - 27a^2d^2 + 18abcd - 4ac^3 + b^2c^2 \\
&= b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd
\end{aligned}$$

Thus, we have again obtained the formula in the statement. □

Still talking degree 3 equations, let us try now to solve such an equation $P=0$, with $P=aX^3+bX^2+cX+d$ as above. By linear transformations we can assume $a=1, b=0$, and then it is convenient to write $c=3p, d=2q$. Thus, our equation becomes:

$$x^3+3px+2q=0$$

Regarding such equations, many things can be said, and to start with, we have the following famous result, dealing with real roots, due to Cardano:

THEOREM 4.9. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

the discriminant is $\Delta=-108(p^3+q^2)$. Assuming $p,q\in\mathbb{R}$ and $\Delta<0$, the number

$$x=\sqrt[3]{-q+\sqrt{p^3+q^2}}+\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

is a real solution of our equation.

PROOF. The formula of Δ is clear from definitions, and with $108=4\times 27$. Now with x as in the statement, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}x^3&=\left(\sqrt[3]{-q+\sqrt{p^3+q^2}}+\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot x\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot x\\&=-2q-3px\end{aligned}$$

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

Regarding the other roots, we know from Proposition 4.7 that these are both real when $\Delta<0$, and complex conjugate when $\Delta<0$. Thus, in the context of Theorem 4.9, the other two roots are complex conjugate, the formula for them being as follows:

PROPOSITION 4.10. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

with $p,q\in\mathbb{R}$ and discriminant $\Delta=-108(p^3+q^2)$ negative, $\Delta<0$, the numbers

$$z=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

$$\bar{z}=w^2\sqrt[3]{-q+\sqrt{p^3+q^2}}+w\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

with $w=e^{2\pi i/3}$ are the complex conjugate solutions of our equation.

PROOF. As before, by using $(a+b)^3=a^3+b^3+3ab(a+b)$, we have:

$$\begin{aligned}z^3&=\left(w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}\right)^3\\&=-2q+3\sqrt[3]{-q+\sqrt{p^3+q^2}}\cdot\sqrt[3]{-q-\sqrt{p^3+q^2}}\cdot z\\&=-2q+3\sqrt[3]{q^2-p^3-q^2}\cdot z\\&=-2q-3pz\end{aligned}$$

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

As a conclusion, we have the following statement, unifying the above:

THEOREM 4.11. *For a normalized degree 3 equation, namely*

$$x^3+3px+2q=0$$

the discriminant is $\Delta=-108(p^3+q^2)$. Assuming $p,q\in\mathbb{R}$ and $\Delta<0$, the numbers

$$x=w\sqrt[3]{-q+\sqrt{p^3+q^2}}+w^2\sqrt[3]{-q-\sqrt{p^3+q^2}}$$

with $w=1,e^{2\pi i/3},e^{4\pi i/3}$ are the solutions of our equation.

PROOF. This follows indeed from Theorem 4.9 and Proposition 4.10. Alternatively, we can redo the computation in their proof, which was nearly identical anyway, in the present setting, with x being given by the above formula, by using $w^3=1$. □

As a comment here, the formula in Theorem 4.11 holds of course in the case $\Delta>0$ too, and also when the coefficients are complex numbers, $p,q\in\mathbb{C}$, and this due to the fact that the proof rests on the nearly trivial computation from the proof of Theorem 4.9, or of Proposition 4.10. However, these extensions are quite often not very useful, because when it comes to extract all the above square and cubic roots, for complex numbers, you can well end up with the initial question, the one that you started with.

In higher degree things become quite complicated. In degree 4, to start with, we first have the following result, dealing with the discriminant and its applications:

THEOREM 4.12. *The discriminant of $P=ax^4+bx^3+cx^2+dx+e$ is given by the following formula:*

$$\begin{aligned}\Delta&=256a^3e^3-192a^2bde^2-128a^2c^2e^2+144a^2cd^2e-27a^2d^4\\&\quad+144ab^2ce^2-6ab^2d^2e-80abc^2de+18abcd^3+16ac^4e\\&\quad-4ac^3d^2-27b^4e^2+18b^3cde-4b^3d^3-4b^2c^3e+b^2c^2d^2\end{aligned}$$

In the case $\Delta<0$ we have 2 real roots and 2 complex conjugate roots, and in the case $\Delta>0$ the roots are either all real or all complex.

PROOF. The formula of Δ follows from the definition of the discriminant, from Theorem 4.5, with the resultant computed via Theorem 4.4, as follows:

$$\Delta=\frac{1}{a}\begin{vmatrix}a & & & 4a & & & \\ b & a & & 3b & 4a & & \\ c & b & a & 2c & 3b & 4a & \\ d & c & b & d & 2c & 3b & 4a\\ e & d & c & & d & 2c & 3b\\ & e & d & & & d & 2c\\ & & e & & & & d\end{vmatrix}$$

As for the last assertion, the study here is routine, a bit as in degree 3. □

In practice, as in degree 3, we can do first some manipulations on our polynomials, as to have them in simpler form, and we have the following version of Theorem 4.12:

PROPOSITION 4.13. *The discriminant of $P=x^4+cx^2+dx+e$, normalized degree 4 polynomial, is given by the following formula:*

$$\Delta=16c^4e-4c^3d^2-128c^2e^2+144cd^2e-27d^4+256e^3$$

As before, if $\Delta<0$ we have 2 real roots and 2 complex conjugate roots, and if $\Delta>0$ the roots are either all real or all complex.

PROOF. This is a consequence of Theorem 4.10, with $a=1,b=0$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

$$\Delta=\frac{1}{a}\begin{vmatrix}1 & & & 4 & & & \\ & 1 & & & 4 & & \\ c & & 1 & 2c & & 4 & \\ d & c & & d & 2c & & 4\\ e & d & c & & d & 2c & \\ & e & d & & & d & 2c\\ & & e & & & & d\end{vmatrix}$$

As for the last assertion, this is something that we know, from Theorem 4.12. □

We still have some work to do. Indeed, looking back at what we did in degree 3, the passage there from Theorem 4.8 to Theorem 4.9 was made of two operations, namely "depressing" the equation, that is, getting rid of the next-to-highest term, and then rescaling the coefficients, as for the formula of Δ to become as simple as possible.

In our present setting now, degree 4, with the depressing done as above, in Proposition 4.13, it remains to rescale the coefficients, as for the formula of Δ to become as simple as possible. And here, a bit of formula hunting, in relation with $2,3$ powers, leads to:

THEOREM 4.14. *The discriminant of a normalized degree* 4 *polynomial, written as*

$$P = x^4 + 6px^2 + 4qx + 3r$$

is given by the following formula:

$$\Delta = 256 \times 27 \times \left(9p^4r - 2p^3q^2 - 6p^2r^2 + 6pq^2r - q^4 + r^3\right)$$

In the case $\Delta < 0$ *we have* 2 *real roots and* 2 *complex conjugate roots, and in the case* $\Delta > 0$ *the roots are either all real or all complex.*

PROOF. This follows from Proposition 4.13, with $c = 6p, d = 4q, e = 3r$, but we can deduce this as well directly. Indeed, the formula of Δ follows, quite easily, from:

$$\Delta = \begin{vmatrix} 1 & & & 4 & & & \\ & 1 & & & 4 & & \\ 6p & & 1 & 12p & & 4 & \\ 4q & 6p & & 4q & 12p & & 4 \\ 3r & 4q & 6p & & 4q & 12p & \\ & 3r & 4q & & & 4q & 12p \\ & & 3r & & & & 4q \end{vmatrix}$$

As for the last assertion, this is something that we know from Theorem 4.12. □

Time now to get to the real thing, solving the equation. We have here:

THEOREM 4.15. *The roots of a normalized degree* 4 *equation, written as*

$$x^4 + 6px^2 + 4qx + 3r = 0$$

are as follows, with y *satisfying the equation* $(y^2 - 3r)(y - 3p) = 2q^2$*,*

$$x_1 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} + \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_2 = \frac{1}{\sqrt{2}}\left(-\sqrt{y - 3p} - \sqrt{-y - 3p + \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_3 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} + \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

$$x_4 = \frac{1}{\sqrt{2}}\left(\sqrt{y - 3p} - \sqrt{-y - 3p - \frac{4q}{\sqrt{2y - 6p}}}\right)$$

and with y *being computable via the Cardano formula.*

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

(1) To start with, let us write our equation in the following form:

$$x^4 = -6px^2 - 4qx - 3r$$

The idea will be that of adding a suitable common term, to both sides, as to make square on both sides, as to eventually end with a sort of double quadratic equation. For this purpose, our claim is that what we need is a number y satisfying:

$$(y^2 - 3r)(y - 3p) = 2q^2$$

Indeed, assuming that we have this number y, our equation becomes:

$$\begin{aligned}(x^2 + y)^2 &= x^4 + 2x^2y + y^2\\ &= -6px^2 - 4qx - 3r + 2x^2y + y^2\\ &= (2y - 6p)x^2 - 4qx + y^2 - 3r\\ &= (2y - 6p)x^2 - 4qx + \frac{2q^2}{y - 3p}\\ &= \left(\sqrt{2y - 6p}\cdot x - \frac{2q}{\sqrt{2y - 6p}}\right)^2\end{aligned}$$

(2) Which looks very good, leading us to the following degree 2 equations:

$$x^2 + y + \sqrt{2y - 6p}\cdot x - \frac{2q}{\sqrt{2y - 6p}} = 0$$

$$x^2 + y - \sqrt{2y - 6p}\cdot x + \frac{2q}{\sqrt{2y - 6p}} = 0$$

Now let us write these two degree 2 equations in standard form, as follows:

$$x^2 + \sqrt{2y - 6p}\cdot x + \left(y - \frac{2q}{\sqrt{2y - 6p}}\right) = 0$$

$$x^2 - \sqrt{2y - 6p}\cdot x + \left(y + \frac{2q}{\sqrt{2y - 6p}}\right) = 0$$

(3) Regarding the first equation, the solutions there are as follows:

$$x_1 = \frac{1}{2}\left(-\sqrt{2y - 6p} + \sqrt{-2y - 6p + \frac{8q}{\sqrt{2y - 6p}}}\right)$$

$$x_2 = \frac{1}{2}\left(-\sqrt{2y - 6p} - \sqrt{-2y - 6p + \frac{8q}{\sqrt{2y - 6p}}}\right)$$

As for the second equation, the solutions there are as follows:

$$x_3=\frac{1}{2}\left(\sqrt{2y-6p}+\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

$$x_4=\frac{1}{2}\left(\sqrt{2y-6p}-\sqrt{-2y-6p-\frac{8q}{\sqrt{2y-6p}}}\right)$$

(4) Now by cutting a $\sqrt{2}$ factor from everything, this gives the formulae in the statement. As for the last claim, regarding the nature of y, this comes from Cardano. □

We still have to compute the number y appearing in the above via Cardano, and the result here, adding to what we already have in Theorem 4.15, is as follows:

THEOREM 4.16 (continuation). *The value of y in the previous theorem is*

$$y=t+p+\frac{a}{t}$$

where the number t is given by the formula

$$t=\sqrt[3]{b+\sqrt{b^2-a^3}}$$

with $a=p^2+r$ and $b=2p^2-3pr+q^2$.

PROOF. The legend has it that this is what comes from Cardano, but depressing and normalizing and solving $(y^2-3r)(y-3p)=2q^2$ makes it for too many operations, so the most pragmatic is to simply check this equation. With y as above, we have:

$$\begin{aligned}y^2-3r&=t^2+2pt+(p^2+2a)+\frac{2pa}{t}+\frac{a^2}{t^2}-3r\\&=t^2+2pt+(3p^2-r)+\frac{2pa}{t}+\frac{a^2}{t^2}\end{aligned}$$

With this in hand, we have the following computation:

$$\begin{aligned}(y^2-3r)(y-3p)&=\left(t^2+2pt+(3p^2-r)+\frac{2pa}{t}+\frac{a^2}{t^2}\right)\left(t-2p+\frac{a}{t}\right)\\&=t^3+(a-4p^2+3p^2-r)t+(2pa-6p^3+2pr+2pa)\\&\quad+(3p^2a-ra-4p^2a+a^2)\frac{1}{t}+\frac{a^3}{t^3}\\&=t^3+(a-p^2-r)t+2p(2a-3p^2+r)+a(a-p^2-r)\frac{1}{t}+\frac{a^3}{t^3}\\&=t^3+2p(-p^2+3r)+\frac{a^3}{t^3}\end{aligned}$$

Now by using the formula of t in the statement, this gives:

$$\begin{aligned}(y^2-3r)(y-3p) &= b+\sqrt{b^2-a^3}-4p^2+6pr+\frac{a^3}{b+\sqrt{b^2-a^3}}\\ &= b+\sqrt{b^2-a^3}-4p^2+6pr+b-\sqrt{b^2-a^3}\\ &= 2b-4p^2+6pr\\ &= 2(2p^2-3pr+q^2)-4p^2+6pr\\ &= 2q^2\end{aligned}$$

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

4d. Further results

In degree 5 and more, things become complicated, and the conceptual explanations for what happens here come from the Galois theory of field extensions. We first have:

THEOREM 4.17. *Given a field extension $E\subset F$, we can talk about its Galois group G, as the group of automorphisms of F fixing E. The intermediate fields*

$$E\subset K\subset F$$

are then in correspondence with the subgroups $H\subset G$, with such a field K corresponding to the subgroup H consisting of automorphisms $g\in G$ fixing K.

PROOF. This is something self-explanatory, and follows indeed from some algebra, under suitable assumptions, in order for that algebra to properly apply. □

Getting now towards polynomials and their roots, we have here:

THEOREM 4.18. *Given a field F and a polynomial $P\in F[X]$, we can talk about the abstract splitting field of P, where this polynomial decomposes as:*

$$P(X)=c\prod_i(X-a_i)$$

In particular, any field F has a certain algebraic closure $\bar{F}$, where all the polynomials $P\in F[X]$, and in fact all polynomials $P\in\bar{F}[X]$ too, have roots.

PROOF. This is again something self-explanatory, which follows from Theorem 4.17 and from some extra algebra, under suitable assumptions, in order for that extra algebra to properly apply. Regarding the construction at the end, as main example here we have $\bar{\mathbb{R}}=\mathbb{C}$. However, as an interesting fact, $\bar{\mathbb{Q}}\subset\mathbb{C}$ is a proper subfield. □

Good news, with this in hand, we can now elucidate the structure of finite fields:

THEOREM 4.19. *For any prime power $q = p^k$ there is a unique field $\mathbb{F}_q$ having q elements. At $k = 1$ this is the usual $\mathbb{F}_p$. In general, this is the splitting field of:*

$$P = X^q - X$$

Moreover, we can construct an explicit model for $\mathbb{F}_q$, at $q = p^2$ or higher, as

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

with $Q \in \mathbb{F}_p[X]$ being a suitable irreducible polynomial, of degree k.

PROOF. There are several assertions here, the idea being as follows:

(1) The first assertion, regarding the existence and uniqueness of $\mathbb{F}_q$, follows from Theorem 4.18. Indeed, we know from chapter 1 that given a finite field, $|F| = q$ with $k \in \mathbb{N}$, the Fermat polynomial $P = X^q - X$ factorizes as follows:

$$X^q - X = \prod_{a \in F} (X - a)$$

Thus F must be the splitting field of P, and so is unique. As for the existence, this follows also from Theorem 4.18, telling us that this splitting field always exists.

(2) In what regards now the modeling of $\mathbb{F}_q$, at $q = p$ there is nothing to do, because we have our usual $\mathbb{F}_p$ here. At $q = p^2$ and higher, we know from commutative algebra that we have an isomorphism as follows, whenever $Q \in \mathbb{F}_p[X]$ is taken irreducible:

$$\mathbb{F}_q = \mathbb{F}_p[X]/(Q)$$

(3) Regarding now the best choice of the irreducible polynomial $Q \in \mathbb{F}_p[X]$, providing us with a good model for the finite field $\mathbb{F}_q$, that we can use in practice, this question depends on the value of $q = p^k$, and many things can be said here. All in all, our models are quite similar to $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(4) To be more precise, at the simplest exponent, $q = 4$, to start with, we can use $Q = X^2 + X + 1$, with this being actually the unique possible choice of a degree 2 irreducible polynomial $Q \in \mathbb{F}_2[X]$, and this leads to a model as follows:

$$\mathbb{F}_4 = \left\{ 0, 1, a, a + 1 \,\middle|\, a^2 = a + 1 \right\}$$

To be more precise here, we assume of course that the characteristic of our model is $p = 2$, which reads $x + x = 0$ for any x, and so determines the addition table. As for the multiplication table, this is uniquely determined by the following formulae:

$$a^2 = -a - 1 = a + 1$$

(5) Next, at exponents of type $q = p^2$ with $p \geq 3$ prime, we can use $Q = X^2 - r$, with r being a non-square modulo p, and with $(p - 1)/2$ choices here. We are led to:

$$\mathbb{F}_{p^2} = \left\{ a + b\gamma \,\middle|\, \gamma^2 = r \right\}$$

Here, as before with $\mathbb{F}_4$, our formula is something self-explanatory. Observe the analogy with $\mathbb{C} = \mathbb{R}[i]$, with i being a formal number satisfying $i^2 = -1$.

(6) Finally, at $q = p^k$ with $k \geq 3$ things become more complicated, but the main idea remains the same. We have for instance models for $\mathbb{F}_8$, $\mathbb{F}_{27}$ using $Q = X^3 - X - 1$, and a model for $\mathbb{F}_{16}$ using $Q = X^4 + X + 1$. Many other things can be said here. □

As another application of the above, which motivated Galois, we have:

THEOREM 4.20. *Unlike in degree $N \leq 4$, there is no formula for the roots of polynomials of degree $N = 5$ and higher, with the reason for this, coming from Galois theory, being that S_5 is not solvable. The simplest numeric example is $P = X^5 - X - 1$.*

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

(1) The first assertion, for generic polynomials, is due to Abel-Ruffini, but Galois theory helps in better understanding this, and comes with a number of bonus points too, namely the possibility of formulating a finer result, with Abel-Ruffini's original "generic", which was something algebraic, being now replaced by an analytic "generic", and also with the possibility of dealing with concrete polynomials, such as:

$$P = X^5 - X - 1$$

(2) Regarding now the details of the Galois proof of the Abel-Ruffini theorem, assume that the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, a bit as for the degree 2 equation, or for the degree 3 and 4 equations, via Cardano. Then, algebrically speaking, this gives rise to a tower of fields as folows, with $F_0 = F$, and each F_{i+1} being obtained from F_i by adding a root, $F_{i+1} = F_i(x_i)$, with $x_i^{n_i} \in F_i$:

$$F_0 \subset F_1 \subset \ldots \subset F_k$$

(3) In order for Galois theory to apply well to this situation, we must make all the extensions normal, which amounts in replacing each $F_{i+1} = F_i(x_i)$ by its extension $K_i(x_i)$, with K_i extending F_i by adding a n_i-th root of unity. Thus, with this replacement, we can assume that the tower in (2) in normal, meaning that all Galois groups are cyclic.

(4) Now by Galois theory, at the level of the corresponding Galois groups we obtain a tower of groups as follows as follows, which is a resolution of the last group G_k, the Galois group of P, in the sense of group theory, in the sense that all quotients are cyclic:

$$G_1 \subset G_2 \subset \ldots \subset G_k$$

As a conclusion, Galois theory tells us that if the roots of a polynomial $P \in F[X]$ can be computed by using iterated roots, then its Galois group $G = G_k$ must be solvable.

(5) In the generic case, the conclusion is that Galois theory tells us that, in order for all polynomials of degree 5 to be solvable, via square roots, the group S_5, which appears there as Galois group, must be solvable, in the sense of group theory. But this is wrong, because the alternating subgroup $A_5 \subset S_5$ is simple, and therefore not solvable.

(6) Finally, regarding the polynomial $P = X^5 - X - 1$, some elementary computations here, based on arithmetic over $\mathbb{F}_2, \mathbb{F}_3$, and involving various cycles of length $2, 3, 5$, show that its Galois group is S_5. Thus, we have our counterexample.

(7) To be more precise, our polynomial factorizes over $\mathbb{F}_2$ as follows:

$$X^5 - X - 1 = (X^2 + X + 1)(X^3 + X^2 + 1)$$

We deduce from this the existence of an element $\tau\sigma \in G \subset S_5$, with $\tau \in S_5$ being a transposition, and with $\sigma \in S_5$ being a 3-cycle, disjoint from it. Thus, we have:

$$\tau = (\tau\sigma)^3 \in G$$

(8) On the other hand since $P = X^5 - X - 1$ is irreducible over $\mathbb{F}_5$, we have as well available a certain 5-cycle $\rho \in G$. Now since $< \tau, \rho >= S_5$, we conclude that the Galois group of P is full, $G = S_5$, and by (4) and (5) we have our counterexample.

(9) Finally, as mentioned in (1), all this shows as well that a random polynomial of degree 5 or higher is not solvable by square roots, and with this being an elementary consequence of the main result from (5), via some standard analysis arguments. □

Very nice all this, and we have now answers to most of the questions that we were having. The story is of course not over here, with at least two important follow-ups:

(1) The original Galois theory, complemented by its modern ramifications, is a huge construction, that you can further learn from algebraic number theory books.

(2) On the other hand, regarding polynomials and their roots, there are countless algebraic and analytic tricks, that you can learn from analytic number theory books.

Getting back now to Earth, and to basic linear algebra and mathematics, we can use the resultant and discriminant technology developed above, in relation with our diagonalization questions for the usual matrices, as formulated in chapter 2, as follows:

THEOREM 4.21. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and in this case, the matrix is diagonalizable:*

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

PROOF. This follows from the general theory that we have, as follows:

(1) To start with, the fact that a matrix is diagonalizable when the eigenvalues are different is something elementary, that we know well from chapter 2.

(2) The equivalence (1) $\iff$ (2) is something that we know from chapter 2 too, coming from the basic theory of the characteristic polynomial.

(3) As for the equivalences (2) $\iff$ (3) $\iff$ (4) $\iff$ (5), which are valid for any poynomial P, these follow from the above theory of the resultant and discriminant. □

The above result is quite interesting, and as a continuation to it, we can now formulate a quite tricky and powerful result, having countless potential applications, as follows:

THEOREM 4.22. *The following happen, inside $M_N(\mathbb{C})$:*

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

PROOF. These are quite advanced linear algebra results, which can be proved as follows, with the technology that we have so far:

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

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

(3) This follows from (2), via the standard fact that the matrices having distinct eigenvalues are diagonalizable, that we know from Theorem 4.21. There are of course some other proofs as well, for instance by putting the matrix in Jordan form, and we will discuss this later in this book, after working out the Jordan form. □

As a first observation, the above result is something extremely useful, more or less allowing you in practice to assume that any matrix $A \in M_N(\mathbb{C})$ is diagonalizable. But of course do not try this at home, unless you know what you're doing.

As an application of the above results, and of our methods in general, we can now establish a number of useful and interesting linear algebra results, as follows:

THEOREM 4.23. *The following happen:*

(1) *We have $P_{AB} = P_{BA}$, for any two matrices $A, B \in M_N(\mathbb{C})$.*
(2) *AB, BA have the same eigenvalues, with the same multiplicities.*
(3) *If A has eigenvalues $\lambda_1, \ldots, \lambda_N$, then $f(A)$ has eigenvalues $f(\lambda_1), \ldots, f(\lambda_N)$.*

PROOF. These results, which are quite non-trivial to prove with bare hands, can be all deduced by using the density tricks from Theorem 4.22, as follows:

(1) To start with, it follows from definitions that the characteristic polynomial of a matrix is invariant under conjugation, in the sense that we have:

$$P_C = P_{ACA^{-1}}$$

Now observe that, when assuming that A is invertible, we have:

$$AB = A(BA)A^{-1}$$

Thus, we have the result when A is invertible. By using now Theorem 4.22 (1), we conclude that this formula holds for any matrix A, by continuity.

(2) This is a reformulation of (1) above, via the fact that P encodes the eigenvalues, with multiplicities, which is hard to prove with bare hands. Let us also mention here that such things are well-known to fail for the infinite matrices, a basic counterexample here being provided by the shift $A = S$ and its adjoint $B = S^*$, which are given by:

$$S = \begin{pmatrix} 0 & 0 & 0 & \ldots \\ 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix} \quad , \quad S^* = \begin{pmatrix} 0 & 1 & 0 & \ldots \\ 0 & 0 & 1 & \ldots \\ 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

Indeed, we have the following two product formulae, for these infinite matrices:

$$SS^* = \begin{pmatrix} 0 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix} \quad , \quad S^*S = \begin{pmatrix} 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

Thus SS^* is a projection, having 0 as eigenvalue, while S^*S is the identity, having only 1 eigenvalues. More on this later in this book, when doing infinite dimensions.

(3) This is something more informal, the idea being that this is clear for the diagonal matrices D, then for the diagonalizable matrices PDP^{-1}, and finally for all the matrices, by using Theorem 4.22 (3), provided that f has suitable regularity properties. We will be back to all this later in this book, with full details, when doing spectral theory. □

As a conclusion to all this, there is a nice and fruitful relationship between linear algebra on one hand, and the theory of the resultant and discriminant on the other hand, with applications in both senses. We will be back to this, later in this book.

Let us also mention that many of the above results extend to the case where we are dealing with linear algebra and polynomials over an arbitrary field F. We will be back to this later in this book, when systematically discussing arithmetic aspects.

And good news, that is all. As a matter however of making sure that we have not forgotten anything basic, in this introductory Part I, it is probably wise to ask the cat. And cat, who as usual, is more of a physicist than a mathematician, declares:

CAT 4.24. *Matrices being generically diagonalizable, if you assume that they are diagonalizable, you might fail sometimes, but can certainly catch some mice.*

Thanks cat, it is all about hunting matters, indeed. This being said, we will have a look right next, in chapter 5 below, at the non-diagonalizable case too, following Jordan and others. This is certainly interesting, mathematically speaking, and with a bit of luck, we might even catch something, perhaps not mice, but some form of human food.

4e. Exercises

This was our first truly advanced chapter, and as exercises on this, we have:

EXERCISE 4.25. *Clarify what has been said above, about symmetric functions.*

EXERCISE 4.26. *Clarify as well all the details in relation with the resultant.*

EXERCISE 4.27. *Learn the other formulations of the Cardano formula in degree* 3*.*

EXERCISE 4.28. *Complete the Cardano proof in degree* 4 *by using the degree* 3 *formula.*

EXERCISE 4.29. *Learn the other formulations of the Cardano formula in degree* 4*.*

EXERCISE 4.30. *Learn more about field extensions, and Galois theory.*

EXERCISE 4.31. *Work out all the Galois theory details for* $P = X^5 - X - 1$*.*

EXERCISE 4.32. *Find some other proofs for the density of diagonalizable matrices.*

As bonus exercise, learn some basic algebraic geometry. All good old stuff.

Part II

Advanced results

Oh, my life is changing everyday
In every possible way
And oh, my dreams
It's never quite as it seems

CHAPTER 5

Jordan form

5a. Linear equations

Welcome to advanced linear algebra. We know from Part I that the generic matrices $A \in M_N(\mathbb{C})$ are diagonalizable, but this is of course not the end of the story. The question being, what if the particular matrix $A \in M_N(\mathbb{C})$ appearing on our path, in relation with this or that problem, was chosen, say by the Devil, to be non-diagonalizable?

Note in passing that when doing mathematics over the real numbers, $F = \mathbb{R}$, the generic matrices $A \in M_N(\mathbb{R})$ are no longer diagonalizable. Thus, Devil's sphere of influence goes way beyond the counterexamples that appear in the complex case.

In answer now, work, and more work, and even more work. We will see in the present Part II all sorts of tricks, invented by the mathematicians, in order to deal with the non-diagonalizable matrices. And with these consisting of the Jordan form, which is the main theorem around, plus a myriad of other useful decomposition techniques.

Before that, however, some motivational talk. Here is a good, concrete question, which appears in mathematics, physics, and related disciplines, that we would like to solve:

QUESTION 5.1. *How to solve differential equations?*

To be more precise, this question appears indeed in all sorts of contexts, all across physics and science, and with all this needing no further presentation, I hope.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\vdots$$

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

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

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

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

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

PROOF. This is indeed clear from definitions, because with $y = (x, \dot{x}, \ldots, x^{(k-1)})$, in the context of Definition 5.5, the system there takes the following form, as desired:

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

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

$$\vdots$$

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

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

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

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

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

$$\vdots$$

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

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

$$x' = Ax$$

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

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

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

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

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

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

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

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

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

5b. Matrix exponential

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

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

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

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

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

where $||x|| = \sqrt{\sum |x_i|^2}$ as usual, is a norm on $M_N(\mathbb{C})$. Also, we have

$$||AB|| \leq ||A|| \cdot ||B||$$

for any two matrices $A, B \in M_N(\mathbb{C})$.

PROOF. All this is clear from definitions, the idea being as follows:

(1) Regarding the norm conditions, $||A|| \geq 0$ with equality precisely when $A = 0$ is clear, $||\lambda A|| = |\lambda| \cdot ||A||$ is clear too, and finally $||A + B|| \leq ||A|| + ||B||$ is clear too.

(2) Regarding now the last assertion, $||AB|| \leq ||A|| \cdot ||B||$, this follows from:

$$\begin{aligned} ||AB|| &= \sup_{||x||=1} ||ABx|| \\ &\leq ||A|| \sup_{||x||=1} ||Bx|| \\ &= ||A|| \cdot ||B|| \end{aligned}$$

(3) Finally, as a comment, we already saw in fact such things in Part I, in an indirect form, when talking about density results inside $M_N(\mathbb{C})$. Note also that the space $M_N(\mathbb{C})$ being finite dimensional, all the possible norms on it are equivalent. □

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

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

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

and these exponentials have the following basic properties:

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

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

(1) We have indeed the following computation, using the various properties of the norm, and notably the formula $||AB|| \leq ||A|| \cdot ||B||$, from Theorem 5.8:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Now by making products, we obtain the following formulae:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

THEOREM 5.13. *We have the following formula,*

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

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

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

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

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

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

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

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

And finally, since the diagonalizable matrices are dense, as we know well since chapter 4, we get by continuity our result in general. As simple as that. □

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

5c. The Jordan form

In order to advance, let us go back to the general diagonalization material from chapter 2. We know from there how to diagonalize the real or complex matrices, in case these are indeed diagonalizable. Also, we have seen several general results, known as "spectral theorems", guaranteeing that a matrix is diagonalizable, in chapter 3.

In order to deal with the general case, let us start with the following definition:

DEFINITION 5.14. *Given a matrix $A \in M_N(\mathbb{C})$, and a vector $x \in \mathbb{C}^N$, we set*

$$C_x = span(x, Ax, A^2x, \ldots)$$

and call it cyclic subspace of A, associated to x.

Here the terminology comes from the fact that A acts in a somewhat cyclic way on C_x, or at least on the above spanning vectors, according to the following formula:

$$A(A^ix) = A^{i+1}x$$

In order to have some mathematics going, out of this observation, the first remark is that the cyclic space $C_x \subset \mathbb{C}^N$ is of course finite dimensional. Thus, in the above definition, we can say that C_x appears as follows, with $k \in \mathbb{N}$ being chosen minimal, such that the space on the right coincides indeed with C_x, as constructed above:

$$C_x = span(x, Ax, A^2x, \ldots, A^{k-1}x)$$

In practice, the number $k \in \mathbb{N}$ must be minimal, as to have a formula as follows:

$$A^kx = a_0x + a_1Ax + a_2A^2x + \ldots + a_{k-1}A^{k-1}x$$

And with this, we are now ready to state our first theorem about the arbitrary matrices, going beyond the general diagonalization material from chapter 2, as follows:

THEOREM 5.15. *Given a matrix $A \in M_N(\mathbb{C})$ and a vector $x \in \mathbb{C}^N$, the restriction of A to the cyclic subspace C_x, with respect to the basis $x, Ax, A^2x, \ldots, A^{k-1}x$, is*

$$C = \begin{pmatrix} 0 & 0 & \ldots & 0 & a_0 \\ 1 & 0 & \ldots & 0 & a_1 \\ 0 & 1 & \ldots & 0 & a_2 \\ & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & a_{k-1} \end{pmatrix}$$

where $a_i \in \mathbb{C}$ are such that $A^kx = a_0x + a_1Ax + a_2A^2x + \ldots + a_{k-1}A^{k-1}x$.

PROOF. This follows from the above discussion. Indeed, in what regards the first $k-1$ columns of the restriction, these are indeed those above, with this coming from:

$$A(A^ix) = A^{i+1}x$$

As for the last column, this is again the one above, with this coming from:

$$A(A^{k-1}x) = a_0x + a_1Ax + a_2A^2x + \ldots + a_{k-1}A^{k-1}x$$

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

In order to further advance, our next observation is that, in the context of Theorem 5.15, what only matters is the following polynomial:

$$P(t) = t^k - a_{k-1}t^{k-1} - \ldots - a_2t^2 - a_1t - a_0$$

So, let us reformulate everything in terms of such polynomials. We are led in this way into the following notion, which is something independent of the above discussion:

DEFINITION 5.16. *Given an arbitrary monic polynomial, written as*

$$P(t) = t^k + b_{k-1}t^{k-1} + \ldots + b_2t^2 + b_1t + b_0$$

the following matrix,

$$C_P = \begin{pmatrix} 0 & 0 & \ldots & 0 & -b_0 \\ 1 & 0 & \ldots & 0 & -b_1 \\ 0 & 1 & \ldots & 0 & -b_2 \\ & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & -b_{k-1} \end{pmatrix}$$

is called its companion matrix.

Which looks quite good, so our plan now will be to study such companion matrices, and come back afterwards to Theorem 5.15. In what regards the first task, we have:

THEOREM 5.17. *The companion matrix C_P of a polynomial P,*

$$C_P = \begin{pmatrix} 0 & 0 & \ldots & 0 & -b_0 \\ 1 & 0 & \ldots & 0 & -b_1 \\ 0 & 1 & \ldots & 0 & -b_2 \\ & & \ddots & & \vdots \\ 0 & 0 & \ldots & 1 & -b_{k-1} \end{pmatrix}$$

has the following properties:

(1) *Its characteristic polynomial is P.*
(2) *Its minimal polynomial is P, too.*
(3) *All the eigenspaces are 1-dimensional.*
(4) *C_P is diagonalizable when the roots of P are distinct.*

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

(1) In order to compute the characteristic polynomial, we switch the first two rows, and we eliminate t from the first column. This leads to the following formula:

$$\begin{aligned} \det(t - C_P) &= \begin{vmatrix} t & 0 & 0 & \ldots & 0 & 0 & b_0 \\ -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ & & \ddots & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & t & 0 & b_{k-3} \\ 0 & 0 & 0 & \ldots & -1 & t & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t + b_{k-1} \end{vmatrix} \\ &= -\begin{vmatrix} -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ t & 0 & 0 & \ldots & 0 & 0 & b_0 \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ & & \ddots & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & t & 0 & b_{k-3} \\ 0 & 0 & 0 & \ldots & -1 & t & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t + b_{k-1} \end{vmatrix} \\ &= -\begin{vmatrix} -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ 0 & t^2 & 0 & \ldots & 0 & 0 & b_0 + b_1 t \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ & & \ddots & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & t & 0 & b_{k-3} \\ 0 & 0 & 0 & \ldots & -1 & t & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t + b_{k-1} \end{vmatrix} \end{aligned}$$

Next, we switch the second and third rows, and we eliminate t^2 from the second column, again with the help of the -1 on the diagonal. We obtain in this way:

$$\det(t-C_P) = \begin{vmatrix} -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ 0 & t^2 & 0 & \ldots & 0 & 0 & b_0+b_1t \\ & & \ddots & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & t & 0 & b_{k-3} \\ 0 & 0 & 0 & \ldots & -1 & t & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t+b_{k-1} \end{vmatrix}$$

$$= \begin{vmatrix} -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ 0 & 0 & 0 & \ldots & 0 & 0 & b_0+b_1t+b_2t^2 \\ & & \ddots & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & t & 0 & b_{k-3} \\ 0 & 0 & 0 & \ldots & -1 & t & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t+b_{k-1} \end{vmatrix}$$

And so on by reccurence, and in the end we obtain, as desired:

$$\det(t-C_P) = (-1)^{k-1}\begin{vmatrix} -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ 0 & 0 & -1 & \ldots & 0 & 0 & b_3 \\ & & & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & -1 & 0 & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t+b_{k-1} \\ 0 & 0 & 0 & \ldots & 0 & 0 & P(t) \end{vmatrix}$$

$$= P(t)$$

(2) Regarding now the minimal polynomial, this is clearly P too.

(3) In order to discuss now the eigenspaces, assume $\det(\lambda - C_P) = 0$, which means $P(\lambda) = 0$. We know from the above that we have a row equivalence, as follows:

$$\lambda - C_P \sim \begin{pmatrix} -1 & t & 0 & \ldots & 0 & 0 & b_1 \\ 0 & -1 & t & \ldots & 0 & 0 & b_2 \\ 0 & 0 & -1 & \ldots & 0 & 0 & b_3 \\ & & & \ddots & & \vdots & \\ 0 & 0 & 0 & \ldots & -1 & 0 & b_{k-2} \\ 0 & 0 & 0 & \ldots & 0 & -1 & t+b_{k-1} \\ 0 & 0 & 0 & \ldots & 0 & 0 & P(t) \end{pmatrix}$$

Thus, the eigenspaces are indeed 1-dimensional, as stated.

(4) Finally, the fact that our companion matrix C_P is diagonalizable precisely when the roots of the polynomial P are distinct is clear too, from the above formulae. □

Time now to go back to cyclic subspaces. We are first led to the following result:

THEOREM 5.18 (Cayley-Hamilton). *Any matrix $A \in M_N(\mathbb{C})$ satisfies:*

$$P_A(A) = 0$$

In particular, the minimal polynomial divides the characteristic polynomial.

PROOF. In order to prove this, pick a nonzero vector $x \in \mathbb{C}^N$, construct the associated cyclic subspace C_x, and then pick a complement for C_x, as to have:

$$\mathbb{C}^N = C_x \oplus V$$

With respect to this decomposition, our matrix A becomes block-diagonal:

$$A = \begin{pmatrix} C_P & B \\ 0 & D \end{pmatrix}$$

At the level of the characteristic polynomial, this gives a formula as follows:

$$P_A(t) = P_{C_P}(t)P_D(t) = P(t)P_D(t)$$

We know from Theorem 5.17 that we have $P(C_p) = 0$, and it follows that we have $P_A(A)x = 0$. But since $x \neq 0$ was arbitrary, this gives $P_A(A) = 0$, as desired. □

As a second result now, which truly advances us, in or study, we have:

THEOREM 5.19. *Any matrix $A \in M_N(\mathbb{C})$ can be written, up to a base change, as*

$$A = \begin{pmatrix} C_{P_1} & & \\ & \ddots & \\ & & C_{P_k} \end{pmatrix}$$

with each C_{P_i} being a companion matrix. In this picture we have

$$P_A = P_1 \ldots P_k$$

and A is diagonalizable precisely when each P_i has distinct roots.

PROOF. This follows indeed from what we have in the above, via a straightforward recurrence, and we will leave the details here as an instructive exercise. □

Many other things can be said about Theorem 5.19, which is called cyclic subspace decomposition, and for more on this, we refer to the literature on the subject.

Next, we are led in this way to the Jordan form, which applies too to any matrix:

THEOREM 5.20. *Any matrix $A \in M_N(\mathbb{C})$ can be written, up to a base change, as*

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

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

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

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

PROOF. We can deduce this from the cyclic subspace decomposition, as follows:

(1) As a first ingredient, due to Jordan-Chevalley, we can decompose any matrix $A \in M_N(\mathbb{C})$ as $A = B + C$, with B diagonalizable, $C^N = 0$, and $BC = CB$. Indeed, let us factor the minimal polynomial of our matrix $A \in M_N(\mathbb{C})$, as follows:

$$R_A(t) = (t - \lambda_1)^{m_1} \ldots (t - \lambda_k)^{m_k}$$

Now if we set $L_i = \ker(A - \lambda_i)^{m_i}$, we have a direct sum decomposition, as follows:

$$\mathbb{C}^N = L_1 \oplus \ldots \oplus L_k$$

But with this done, we can define matrices $B, C \in M_N(\mathbb{C})$ block-diagonally, by:

$$B_{|L_i} = \lambda_i 1_{L_i} \quad , \quad C_{|L_i} = A_{|L_i} - \lambda_i 1_{L_i}$$

Now observe that we have indeed $A = B + C$, with B being diagonalizable, and with $BC = CB$. Finally, since $R_A(A) = 0$, we have as well $C^N = 0$, as desired.

(2) Next, let us apply the Jordan-Chevalley decomposition, as performed above, to a companion matrix C_P coming from a polynomial of type $P(t) = (t - \lambda)^k$. We conclude that such a companion matrix must be similar to a Jordan block, as follows:

$$C_P \sim \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

(3) Let us turn now to the proof of the theorem. By using the Jordan-Chevalley decomposition, it is enough to prove the theorem for matrices of type $A = \lambda 1_N + C$, with $C^N = 0$. But here, the result follows from the cyclic subspace decomposition. □

5d. Basic applications

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

To be more precise, let us write the matrix to be exponentiated in Jordan form, as in Theorem 5.20, as follows, with P denoting the passage matrix used there:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

THEOREM 5.21. *For a matrix written in Jordan form, as follows,*

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

the corresponding exponential is given by the following formula,

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

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

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

with s being the size of our Jordan block.

PROOF. This follows indeed from the above discussion. □

And good news, this is all we need to know, being obviously something very powerful, closing any further mathematical discussion about exponentiation. We will be back to this in the next chapter, with a systematic discussion of the differential equations.

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

As yet another application, getting back to our spectral measure considerations from chapter 3, recall from there that any normal matrix has a spectral measure, formed by the Dirac masses at the eigenvalues. In the non-normal case, things can be quite complicated. In particular, we have some interesting computations here for the Jordan blocks.

We can apply other complex functions to our matrices, under suitable assumptions. All this is quite technical, called "functional calculus", and as a basic result here, coming from the Cauchy formula, we can apply any holomorphic function to any matrix.

Passed the holomorphic functions, things become more complicated. In the normal case, we can apply continuous functions, and even measurable ones, to our matrices. Indeed, this follows from our spectral theorems, developed in chapter 3.

We will be back to this in chapter 8 below, with some further results.

5e. Exercises

This was a quite fundamental chapter, at the origin of all possible advaced and modern linear algebra topics, and as exercises on all this, we have:

EXERCISE 5.22. *In relation with equations, have a look at the non-uniform gravitational falls too, first in* 1 *dimension, and then in* 2 *dimensions.*

EXERCISE 5.23. *Also in relation with equations, and with the general theory developed above, learn if needed the implicit function theorem.*

EXERCISE 5.24. *Learn a bit about normed spaces, generalities about them, as needed in the above, in order to talk about the exponential of matrices.*

EXERCISE 5.25. *Try remembering, and then finding matrix analogues, of some other formulae involving* exp*, that you know from calculus.*

EXERCISE 5.26. *Work out all the details of the proof of the Cayley-Hamilton theorem, and find some applications of this theorem too.*

EXERCISE 5.27. *Fill in all the details for the proof of the block decomposition into companion matrices, and work out some applications of this, too.*

EXERCISE 5.28. *Work out all the details for the Jordan decomposition theorem, along the lines explained in the above.*

EXERCISE 5.29. *Learn as well some explicit algorithms for finding the Jordan form, based on the above material, or on some alternative approaches too.*

As bonus exercise for this chapter, and no surprise here, put various matrices of your choice in Jordan form, the more the better, and the bigger the better, too.

CHAPTER 6

Dynamical systems

6a. Differential equations

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

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

$$\vdots$$

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

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

$$\dot{y}_1 = y_2 \quad , \quad \dot{y}_2 = y_3 \quad , \quad \ldots \quad , \quad \dot{y}_{k-1} = y_k \quad , \quad \dot{y}_k = f(t, y)$$

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

$$\dot{z}_1 = 1 \quad , \quad \dot{z}_2 = z_3 \quad , \quad \ldots \quad , \quad \dot{z}_k = z_{k+1} \quad , \quad \dot{z}_{k+1} = f(z)$$

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

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

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

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

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

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

In view of this observation, consider the following function:

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

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

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

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

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

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

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

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

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

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

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

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

$$T_- = \lim_{x \searrow x_1} F(x) \in [-\infty, 0) \quad , \quad T_+ = \lim_{x \nearrow x_2} F(x) \in (0, \infty]$$

Since $\lim_{t \searrow T_-} \varphi(t) = x_1$, the solution φ exists for any $t < 0$ precisely when:

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

Similarly, since $\lim_{t \nearrow T_+} \varphi(t) = x_2$, the solution φ exists for any $t > 0$ when:

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

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

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

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

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

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

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

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

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

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

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

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

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

As a continuation of the above discussion, dealing with the case $f(x_0) \neq 0$, it remains now to discuss the case $f(x_0) = 0$. Here we have the trivial solution $\varphi(t) = x_0$, and we can have as well non-trivial solutions. Assume for instance that we have:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6b. Functional analysis

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

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

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

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

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

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

Based on this observation, consider the following function:

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

In terms of this function, our original equation reads:

$$K(x) = x$$

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

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

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

We will see that the answers to these latter questions are yes, yes, yes. However, this is something quite technical, which will take some time. Let us start with:

DEFINITION 6.7. *A normed space is a complex vector space V with a map*

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

called norm, subject to the following conditions:

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

When V is complete with respect to $d(x, y) = ||x - y||$, we say that V is a Banach space.

In relation with this, observe that the function $d(x, y) = ||x - y||$ is indeed a distance, with the key distance axiom, which is the triangle inequality $d(x, y) \leq d(x, z) + d(y, z)$, coming from our third norm axiom above, namely $||x + y|| \leq ||x|| + ||y||$.

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

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

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

In order to construct further examples, let us start with a basic result, as follows:

THEOREM 6.8 (Jensen). *Given a convex function $f : \mathbb{R} \to \mathbb{R}$, we have the following inequality, for any $x_1, \ldots, x_N \in \mathbb{R}$, and any $\lambda_1, \ldots, \lambda_N > 0$ summing up to 1,*

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

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

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

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

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

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

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

(2) But this means, via a simple argument, by approximating numbers $t\in[0,1]$ by sums of powers 2^{-k}, that for any $t\in[0,1]$ we must have:

$$f(tx+(1-t)y)\leq tf(x)+(1-t)f(y)$$

Alternatively, via yet another simple argument, this time by doing some geometry with triangles, this means that we must have:

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

But then, again alternatively, by combining the above two simple arguments, the following must happen, for any $\lambda_1,\ldots,\lambda_N>0$ summing up to 1:

$$f(\lambda_1x_1+\ldots+\lambda_Nx_N)\leq\lambda_1f(x_1)+\ldots+\lambda_Nx_N$$

(3) Summarizing, all our Jensen inequalities, at $N=2$ and at $N\in\mathbb{N}$ arbitrary, are equivalent. The point now is that, if we look at what the first Jensen inequality, that we took as definition for the convexity, means, this is simply equivalent to:

$$f''(x)\geq 0$$

(4) Thus, we are led to the conclusions in the statement, regarding the convex functions. As for the concave functions, the proof here is similar. Alternatively, we can say that f is concave precisely when $-f$ is convex, and get the results from what we have. □

As a basic application of the Jensen inequality, we have:

PROPOSITION 6.9. *For $p\in(1,\infty)$ we have the following inequality,*

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\leq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

and for $p\in(0,1)$ we have the following reverse inequality,

$$\left|\frac{x_1+\ldots+x_N}{N}\right|^p\geq\frac{|x_1|^p+\ldots+|x_N|^p}{N}$$

with in both cases equality precisely when $|x_1|=\ldots=|x_N|$.

PROOF. This follows indeed from Theorem 6.8, because we have:

$$(x^p)''=p(p-1)x^{p-2}$$

Thus x^p is convex for $p>1$ and concave for $p<1$, which gives the results. □

As another basic application of the Jensen inequality, we have:

THEOREM 6.10 (Young). *We have the following inequality,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

valid for any $a, b \geq 0$, and any exponents $p, q > 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$.

PROOF. We use the logarithm function, which is concave on $(0, \infty)$, due to:

$$(\log x)'' = \left(-\frac{1}{x}\right)' = -\frac{1}{x^2}$$

Thus we can apply the Jensen inequality, and we obtain in this way:

$$\begin{aligned} \log\left(\frac{a^p}{p} + \frac{b^q}{q}\right) &\geq \frac{\log(a^p)}{p} + \frac{\log(b^q)}{q} \\ &= \log(a) + \log(b) \\ &= \log(ab) \end{aligned}$$

Now by exponentiating, we obtain the Young inequality. □

Moving forward now, as a consequence of the Young inequality, we have:

THEOREM 6.11 (Hölder). *Assuming that $p, q \geq 1$ are conjugate, in the sense that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

we have the following inequality, valid for any two vectors $x, y \in \mathbb{C}^N$,

$$\sum_i |x_i y_i| \leq \left(\sum_i |x_i|^p\right)^{1/p} \left(\sum_i |y_i|^q\right)^{1/q}$$

with the convention that an ∞ exponent produces a $\max |x_i|$ quantity.

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

(1) Assume first that we are dealing with finite exponents, $p, q \in (1, \infty)$. By linearity we can assume that x, y are normalized, in the following way:

$$\sum_i |x_i|^p = \sum_i |y_i|^q = 1$$

In this case, we want to prove that the following inequality holds:

$$\sum_i |x_i y_i| \leq 1$$

For this purpose, we use the Young inequality, which gives, for any i:

$$|x_i y_i| \leq \frac{|x_i|^p}{p} + \frac{|y_i|^q}{q}$$

By summing now over $i = 1, \ldots, N$, we obtain from this, as desired:

$$\begin{aligned}\sum_i |x_i y_i| &\leq \sum_i \frac{|x_i|^p}{p} + \sum_i \frac{|y_i|^q}{q}\\ &= \frac{1}{p} + \frac{1}{q}\\ &= 1\end{aligned}$$

(2) In the case $p = 1$ and $q = \infty$, or vice versa, the inequality holds too, trivially, with the convention that an ∞ exponent produces a max quantity, according to:

$$\lim_{p\to\infty} \left(\sum_i |x_i|^p\right)^{1/p} = \max |x_i|$$

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

As a consequence now of the Hölder inequality, we have:

THEOREM 6.12 (Minkowski). *Assuming $p \in [1, \infty]$, we have the inequality*

$$\left(\sum_i |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}$$

for any two vectors $x, y \in \mathbb{C}^N$, with our usual conventions at $p = \infty$.

PROOF. We have indeed the following estimate, using the Hölder inequality, and the conjugate exponent $q \in [1, \infty]$, given by $1/p + 1/q = 1$:

$$\begin{aligned}&\sum_i |x_i + y_i|^p\\ =\ &\sum_i |x_i + y_i| \cdot |x_i + y_i|^{p-1}\\ \leq\ &\sum_i |x_i| \cdot |x_i + y_i|^{p-1} + \sum_i |y_i| \cdot |x_i + y_i|^{p-1}\\ \leq\ &\left(\sum_i |x_i|^p\right)^{1/p} \left(\sum_i |x_i + y_i|^{(p-1)q}\right)^{1/q} + \left(\sum_i |y_i|^p\right)^{1/p} \left(\sum_i |x_i + y_i|^{(p-1)q}\right)^{1/q}\\ =\ &\left[\left(\sum_i |x_i|^p\right)^{1/p} + \left(\sum_i |y_i|^p\right)^{1/p}\right] \left(\sum_i |x_i + y_i|^p\right)^{1-1/p}\end{aligned}$$

Here we have used $(p-1)q = p$ at the end, coming from $1/p + 1/q = 1$. Now by dividing both sides by the last quantity at the end, we obtain the result. □

Good news, done with inequalities, and as a consequence of the above results, and more specifically of the Minkowski inequality obtained above, we can formulate:

THEOREM 6.13. *Given an exponent $p \in [1,\infty]$, the formula*

$$||x||_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on $\mathbb{C}^N$, making it a Banach space.

PROOF. Here the normed space assertion follows from the Minkowski inequality, and the Banach space assertion is trivial, our space being finite dimensional. □

Very nice all this, but you might wonder at this point, what is the relation of all this with functions. In answer, Theorem 6.13 can be reformulated as follows:

THEOREM 6.14. *Given an exponent $p \in [1,\infty]$, the formula*

$$||f||_p = \left(\int |f(x)|^p\right)^{1/p}$$

with usual conventions at $p = \infty$, defines a norm on the space of functions

$$f : \{1, \ldots, N\} \to \mathbb{C}$$

making it a Banach space.

PROOF. This is a just fancy reformulation of Theorem 6.13, by using the fact that the space formed by the functions $f : \{1, \ldots, N\} \to \mathbb{C}$ is canonically isomorphic to $\mathbb{C}^N$. □

In order to further extend the above result, let us start with:

THEOREM 6.15. *Given two functions $f, g : X \to \mathbb{C}$ and an exponent $p \geq 1$, we have*

$$\left(\int_X |f+g|^p\right)^{1/p} \leq \left(\int_X |f|^p\right)^{1/p} + \left(\int_X |g|^p\right)^{1/p}$$

called Minkowski inequality. Also, assuming that $p, q \geq 1$ satisfy $1/p + 1/q = 1$, we have

$$\int_X |fg| \leq \left(\int_X |f|^p\right)^{1/p} \left(\int_X |g|^q\right)^{1/q}$$

called Hölder inequality. These inequalities hold as well for ∞ values of the exponents.

PROOF. This is very standard, exactly as in the case of sequences, as follows:

(1) Let us first prove Hölder, in the case of finite exponents, $p, q \in (1,\infty)$. By linearity we can assume that f, g are normalized, in the following way:

$$\int_X |f|^p = \int_X |g|^q = 1$$

We can use as before the Young inequality, which gives, for any $x \in X$:

$$|f(x)g(x)| \leq \frac{|f(x)|^p}{p} + \frac{|g(x)|^q}{q}$$

By integrating now over $x \in X$, we obtain from this, as desired:

$$\int_X |fg| \leq \int_X \frac{|f(x)|^p}{p} + \int_X \frac{|g(x)|^q}{q} = 1$$

(2) Regarding now Minkowski, again in case $p \in (1, \infty)$, this follows from:

$$\begin{aligned}
& \int_X |f+g|^p \\
= \ & \int_X |f+g| \cdot |f+g|^{p-1} \\
\leq \ & \int_X |f| \cdot |f+g|^{p-1} + \int_X |g| \cdot |f+g|^{p-1} \\
\leq \ & \left(\int_X |f|^p\right)^{1/p} \left(\int_X |f+g|^{(p-1)q}\right)^{1/q} + \left(\int_X |g|^p\right)^{1/p} \left(\int_X |f+g|^{(p-1)q}\right)^{1/q} \\
= \ & \left[\left(\int |f|^p\right)^{1/p} + \left(\int_X |g|^p\right)^{1/p}\right] \left(\int_X |f+g|^p\right)^{1-1/p}
\end{aligned}$$

(3) Finally, in the infinite exponent cases we have similar results, which are trivial this time, with the convention that an ∞ exponent produces an essential supremum. □

We can now extend Theorem 6.14, into something very general, as follows:

THEOREM 6.16. *Given a measured space X, and $p \in [1, \infty]$, the following space, with the convention that functions are identified up to equality almost everywhere,*

$$L^p(X) = \left\{ f : X \to \mathbb{C} \,\middle|\, \int_I |f(x)|^p dx < \infty \right\}$$

is a vector space, and the following quantity

$$||f||_p = \left(\int_X |f(x)|^p\right)^{1/p}$$

is a norm on it, making it a Banach space.

PROOF. This follows indeed from Theorem 6.15, with due attention to the null sets, and this because of the first normed space axiom, namely:

$$||x|| = 0 \implies x = 0$$

To be more precise, in order for this axiom to hold, we must identify the functions up to equality almost everywhere, as indicated in the statement. □

6c. Existence, uniqueness

Getting now towards our ODE business, existence and uniqueness results, as explained before, we would like to use some fixed point technology. So, let us formulate:

DEFINITION 6.17. *Let V be a Banach space, and $K : C \subset V \to C$ be a linear map, with C being closed. We say that K is a contraction if*

$$||K(x) - K(y)|| \leq \theta ||x - y||$$

for some $\theta \in [0, 1)$. Also, we call fixed point of K any $x \in C$ such that $K(x) = x$.

Observe that the fixed point of a contraction, if it exists, is unique, due to our assumption $\theta < 1$. Now with these notions in hand, we have the following result:

THEOREM 6.18. *Any contraction $K : C \subset V \to C$ has a unique fixed point $\bar{x} \in C$, which can be obtained by starting with any point $x \in C$, and iterating K:*

$$\bar{x} = \lim_{n \to \infty} K^n(x)$$

In addition, we have the following estimate,

$$||K^n(x) - \bar{x}|| \leq \frac{\theta^n}{1 - \theta} ||K(x) - x||$$

valid for any $x \in C$, regarding the convergence $K^{(n)}(x) \to \bar{x}$.

PROOF. As explained in the above, the uniqueness of the fixed point is clear, coming from our assumption $\theta < 1$. Regarding now the existence part, and the precise estimate in the statement too, pick $x = x_0 \in C$, and set $x_n = K^n(x_0)$. We have then:

$$\begin{aligned} ||x_{n+1} - x_n|| &\leq \theta ||x_n - x_{n-1}|| \\ &\leq \theta^2 ||x_{n-1} - x_{n-2}|| \\ &\vdots \\ &\leq \theta^n ||x_1 - x_0|| \end{aligned}$$

Now by using the triangle inequality, we obtain from this, for $n > m$:

$$\begin{aligned} ||x_n - x_m|| &\leq \sum_{j=m+1}^{n} ||x_j - x_{j-1}|| \\ &\leq \theta^m \sum_{j=0}^{n-m-1} \theta^j ||x_1 - x_0|| \\ &\leq \frac{\theta^m}{1 - \theta} ||x_1 - x_0|| \end{aligned}$$

Thus the sequence $\{x_n\}$ is Cauchy, and since we are in a Banach space, this sequence converges. Moreover, since $C \subset V$ was chosen closed, the limit belongs to C:

$$x_n \to \bar{x} \in C$$

Now since our map K was assumed to be a contraction, it is continuous, and by continuity we obtain, as desired, that we have indeed a fixed point, due to:

$$||K(\bar{x}) - \bar{x}|| = \lim_{n\to\infty} ||x_{n+1} - x_n|| = 0$$

Finally, in what regards the estimate at the end, in the statement, let us go back to the main estimate obtained before, which was as follows, for any $n > m$:

$$||x_n - x_m|| \leq \frac{\theta^m}{1-\theta}||x_1 - x_0||$$

But this gives, with $m \to \infty$, the estimate in the statement, as desired. □

Now by getting back to our ODE questions, recall from our discussion before that the map which was needing fixed points was as follows:

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

Thus, we are led into the question on whether such a map K is a contraction or not. In order to discuss this, let us introduce the following technical definition:

DEFINITION 6.19. *A map $f \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, is called locally Lipschitz with respect to x, uniformly with respect to t, if for any $V \subset U$ compact we have*

$$\frac{|f(t,x) - f(t,y)|}{||x - y||} \leq L$$

for any $(t, x) \neq (t, y) \in V$, for a certain number $L \in (0, \infty)$.

Observe that in the case $L \leq 1$, our map is a contraction, at any t. Now with this notion in hand, we can formulate, following Cauchy-Lipschitz and Picard-Lindelöf:

THEOREM 6.20. *An equation as follows, with $f \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, has a unique local solution,*

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

provided that f is locally Lipschitz with respect to x, uniformly with respect to t.

PROOF. Consider, as already indicated above, the following map:

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

We assume for simplifying $t_0 = 0$. In order to verify that K is a contraction, for $t > 0$ small, consider the following Banach space, with $T > 0$ to be determined later:

$$V = C(I, \mathbb{R}^N) \quad , \quad I = [0, T]$$

Let also $\delta > 0$, and consider the following closed ball, inside this space V:

$$C = \bar{B}_\delta(x_0)$$

We would like to apply Theorem 6.18, and in order to do so, we need to check two things, namely that we have indeed $K : C \to C$, and that K is a contraction.

(1) Let us first check that we have $K : C \to C$. For this purpose, let us set:

$$W = [0, T] \times C \subset U$$

We have then the following estimate, coming from definitions:

$$\begin{aligned} |K(x)(t) - x_0| &\leq \int_0^t |f(s, x(s))| ds \\ &\leq t \max_{(t,v) \in W} |f(t, x)| \end{aligned}$$

In view of this, consider the number appearing on the right, namely:

$$M = \max_{(t,v) \in W} |f(t, x)|$$

With this notation, we conclude from our estimate above that we have:

$$TM \leq \delta \implies |K(x)(t) - x_0| \leq \delta, \ \forall t \in [0, T]$$

On the other hand, inside the Banach space $C([0, T], \mathbb{R}^N)$, we have:

$$||K(x) - x_0|| = \sup_{t \in [0,T]} |K(x)(t) - x_0|$$

Thus, under the above assumption $TM \leq \delta$, the following happens:

$$||K(x) - x_0|| \leq \delta$$

But this shows that we have $K(x) \in \bar{B}_\delta(x_0) = C$, and so that we have, as desired:

$$K : C \to C$$

(2) With this done, let us turn now to the second check, that of the fact that our linear map K is indeed a contraction. For this purpose, we use the Lipschitz property of f from the statement, or rather from Definition 6.19, namely:

$$\frac{|f(t, x) - f(t, y)|}{||x - y||} \leq L$$

By using this, and integrating, we obtain the following estimate:

$$\begin{aligned}\int_0^t |f(s,x(s))-f(s,y(s))|ds &\leq L\int_0^t |x(s)-y(s)|ds\\ &\leq Lt\sup_{0\leq s\leq t}|x(s)-y(s)|\end{aligned}$$

Thus, in terms of our linear map K, we have the following estimate:

$$||K(x)-K(y)||\leq LT||x-y||$$

But this shows that, with $T\leq 1/L$, we have indeed a contraction, as desired.

(3) Summarizing, we have shown that we have $K:C\to C$, and that this map is a contraction. Thus Theorem 6.18 applies, and gives the result. □

Before getting into further theory, let us discuss a simple application of the above. Consider the following linear equation, that we certainly know how to solve:

$$\dot{x}=x\quad,\quad x(0)=1$$

Observe that $f(t,x)=x$ is indeed Lipschitz as in Definition 6.19, with $L=1$. Regarding now the linear map K, this is given by the following formula:

$$\begin{aligned}K(x)(t) &= x_0+\int_{t_0}^t f(s,x(s))ds\\ &= 1+\int_0^t x(s)ds\end{aligned}$$

By choosing now $y=1$ as starting point, the iteration goes as follows:

$$K(y)=1+\int_0^t 1ds=1+t$$

$$K^2(y)=1+\int_0^t(1+s)ds=1+t+\frac{t^2}{2}$$

$$K^3(y)=1+\int_0^t\left(1+s+\frac{s^2}{2}\right)ds=1+t+\frac{t^2}{2}+\frac{t^3}{6}$$

$$\vdots$$

Thus we obtain in the limit, as we should, the following solution:

$$K^\infty(y)=\sum_{n=0}^\infty\frac{t^n}{n!}=e^t$$

Getting now to technical comments, in relation with Theorem 6.20, many things can be said here, and here are two of them, which are of particular importance:

(1) In the context of Theorem 6.20, it is possible to prove that if $f \in C^k(U, \mathbb{R}^N)$ with $k \geq 1$, then the solution is C^{k+1}. This is indeed elementary, by recurrence on k.

(2) Also in the context of Theorem 6.20, assume that $[t_0, T] \times \mathbb{R}^N \subset U$ is such that:

$$\int_{t_0}^{T} L(t)dt < \infty \quad , \quad L(t) = \sup_{x \neq y \in \mathbb{R}^N} \frac{|f(t,x) - f(t,y)|}{||x - y||}$$

Then, by suitably changing the Banach space norm, and suitably modifying the contraction principle too, it is possible to prove that the solution is defined on $[t_0, T]$.

We refer to the ODE literature for more on the above, which is something quite standard. As a main question now that we would like to solve, we have:

QUESTION 6.21. *How does the solution depend on the initial data?*

Obviously, this question is of key importance, in relation with our general order vs chaos problematics. However, this question is non-trivial, and our tools so far, which are quite abstract, do not provide a direct answer to it. So, we have to work some more.

In order to solve our question, let us begin with a key technical statement, of classical analysis type, not obviously related to equations, due to Gronwall, as follows:

PROPOSITION 6.22. *Assume that a function ψ satisfies the estimate*

$$\psi(t) \leq \alpha(t) + \int_0^t \beta(s)\psi(s)ds$$

for any $t \in [0, T]$, with $\alpha(t) \in \mathbb{R}$, and $\beta(t) > 0$. We have then

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s) \exp\left(\int_s^t \beta(r)dr\right) ds$$

for any $t \in [0, T]$. Moreover, assuming that α is increasing, we have

$$\psi(t) \leq \alpha(t) \exp\left(\int_0^t \beta(s)ds\right)$$

for any $t \in [0, T]$.

PROOF. This is something quite tough, and for the story, it happened to me more than once, when teaching this to our graduate math students in Cergy, for one student to leave the class during or after the proof, in protest, never to be seen again. Well, in the hope that these protesting kids found some friends, spouses and jobs, not quite sure about that, and here is the proof of the result, that I personally find quite cute:

(1) Let us first prove the first assertion, which is the main one. For this purpose, we use a trick. Consider the following function:

$$\phi(t) = \exp\left(-\int_0^t \beta(s)ds\right)$$

We have then the following computation, using the Leibnitz rule for derivatives, and also using at the end our assumption on ψ from the statement:

$$\begin{aligned}
& \frac{d}{dt}\left[\phi(t)\int_0^t \beta(s)\psi(s)ds\right] \\
= & \left[\frac{d}{dt}\,\phi(t)\right]\int_0^t \beta(s)\psi(s)ds + \phi(t)\left[\frac{d}{dt}\int_0^t \beta(s)\psi(s)ds\right] \\
= & -\beta(t)\phi(t)\int_0^t \beta(s)\psi(s)ds + \phi(t)\beta(t)\psi(t) \\
= & \beta(t)\psi(t)\left(\psi(t) - \int_0^t \beta(s)\psi(s)ds\right) \\
\leq & \alpha(t)\beta(t)\phi(t)
\end{aligned}$$

Now by integrating with respect to t, we obtain from this:

$$\phi(t)\int_0^t \beta(s)\psi(s)ds \leq \int_0^t \alpha(s)\beta(s)\phi(s)ds$$

We conclude that we have the following estimate:

$$\int_0^t \beta(s)\psi(s)ds \leq \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds$$

By adding now $\alpha(t)$ to both sides, we obtain the following estimate:

$$\alpha(t) + \int_0^t \beta(s)\psi(s)ds \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds$$

But in this situation, we can use once again our assumption on ψ from the statement, and we obtain the following estimate:

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\frac{\phi(s)}{\phi(t)}ds$$

Now let us look at the fraction on the right. This is given by:

$$\begin{aligned}\frac{\phi(s)}{\phi(t)} &= \frac{\exp\left(-\int_0^s \beta(r)dr\right)}{\exp\left(-\int_0^t \beta(r)dr\right)}\\ &= \exp\left(\int_0^t \beta(r)dr - \int_0^s \beta(r)dr\right)\\ &= \exp\left(\int_s^t \beta(r)dr\right)\end{aligned}$$

We conclude that the estimate that we found above reads:

$$\psi(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds$$

But this is precisely what we wanted to prove, the first estimate in the statement.

(2) With this done, let us turn now to the second assertion in the statement. So, assume that the function α there is increasing. We have then:

$$\begin{aligned}\psi(t) &\leq \alpha(t) + \int_0^t \alpha(s)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds\\ &\leq \alpha(t) + \int_0^t \alpha(t)\beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds\\ &= \alpha(t)\left[1 + \int_0^t \beta(s)\exp\left(\int_s^t \beta(r)dr\right)ds\right]\\ &= \alpha(t)\left[1 + \int_0^t \beta(s)\exp\left(\int_0^t \beta(r)dr - \int_0^s \beta(r)dr\right)ds\right]\\ &= \alpha(t)\left[1 + \exp\left(\int_0^t \beta(r)dr\right)\int_0^t \beta(s)\exp\left(-\int_0^s \beta(r)dr\right)ds\right]\end{aligned}$$

Now recall that we can consider, as in (1), the following function:

$$\phi(t) = \exp\left(-\int_0^t \beta(s)ds\right)$$

The derivative of this function satisfies then the following formula:

$$\phi'(t) = -\beta(t)\phi(t)$$

Thus, we have the following formula, for this derivative:

$$\phi'(s) = -\beta(s)\exp\left(-\int_0^s \beta(r)dr\right)$$

We conclude that the estimate found before reformulates as:

$$\begin{aligned}\psi(t) &\leq \alpha(t)\left[1+\exp\left(\int_0^t\beta(r)dr\right)\int_0^t\beta(s)\exp\left(-\int_0^s\beta(r)dr\right)ds\right]\\ &= \alpha(t)\left[1+\exp\left(\int_0^t\beta(r)dr\right)(-\phi')\Big|_0^t\right]\\ &= \alpha(t)\left[1+\exp\left(\int_0^t\beta(r)dr\right)(1-\phi(t))\right]\end{aligned}$$

In order to finish, consider the following number, depending on t:

$$K=\int_0^t\beta(r)dr$$

In terms of this number, the estimate that we found above reads:

$$\begin{aligned}\psi(t) &\leq \alpha(t)(1+e^K(1-e^{-K}))\\ &= \alpha(t)(1+e^K-1)\\ &= \alpha(t)e^K\end{aligned}$$

Thus, as a conclusion, we have reached to the following estimate:

$$\psi(t)\leq\alpha(t)\exp\left(\int_0^t\beta(s)ds\right)$$

But this is exactly what we wanted to prove, namely second estimate in the statement, and so, eventually, done. So, very good all this, and still with me, I hope. □

As a continuation of the above, we won't leave such beautiful things like this, we would definitely love to spend more time with them, we have:

PROPOSITION 6.23. *Assume that a function ψ satisfies the estimate*

$$\psi(t)\leq\alpha(t)+\int_0^t(\beta\psi(s)+\gamma)ds$$

for any $t\in[0,T]$, with $\alpha\in\mathbb{R}$, $\beta\geq 0$ and $\gamma\in\mathbb{R}$. We have then

$$\psi(t)\leq\alpha\exp(\beta t)+\frac{\gamma}{\beta}(\exp(\beta t)-1)$$

for any $t\in[0,T]$.

PROOF. In order to prove this result, consider the following function:

$$\widetilde{\psi}(t)=\psi(t)+\frac{\gamma}{\beta}$$

In terms of this function $\widetilde{\psi}$, our assumption on ψ in the statement reads:

$$\widetilde{\psi} - \frac{\gamma}{\beta} \leq \alpha + \beta \int_0^t \widetilde{\psi}(s)ds$$

Thus, our modified function $\widetilde{\psi}$ satisfies the following estimate:

$$\widetilde{\psi} \leq \left(\alpha + \frac{\gamma}{\beta}\right) + \beta \int_0^t \widetilde{\psi}(s)ds$$

Thus, we can apply the second assertion in Proposition 6.22, with $\alpha(t) = \alpha + \gamma/\beta$, and $\beta(t) = \beta$, both constant functions, and we obtain in this way:

$$\widetilde{\psi} \leq \left(\alpha + \frac{\gamma}{\beta}\right) \exp(\beta t)$$

But this gives, in terms of the original function ψ, the following estimate:

$$\begin{aligned} \psi(t) &\leq \left(\alpha + \frac{\gamma}{\beta}\right) \exp(\beta t) - \frac{\gamma}{\beta} \\ &= \alpha \exp(\beta t) + \frac{\gamma}{\beta}(\exp(\beta t) - 1) \end{aligned}$$

Thus, we have reached to the conclusion in the statement. □

Now back to the ODE, the above results apply, and we can answer Question 6.21. To be more precise, in the general context of Theorem 6.20, we have the following result:

THEOREM 6.24. *Assume that $f, g \in C(U, \mathbb{R}^N)$, with $U \subset \mathbb{R}^{N+1}$ open, are locally Lipschitz with respect to x, uniformly with respect to t. If x, y are solutions of*

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

$$\dot{y} = g(t, y) \quad , \quad y(t_0) = y_0$$

then we have the following estimate, for any t in the interval of definition of x, y,

$$||x(t) - y(t)|| \leq ||x_0 - y_0|| e^{L|t-t_0|} + \frac{M}{L}\left(e^{L|t-t_0|} - 1\right)$$

with the constant M on the right being given by the following formula,

$$M = \sup_{(t,x) \in U} |f(t, x) - g(t, x)|$$

and with $L > 0$ being a common Lipschitz constant for both f, g.

PROOF. We know from Theorem 6.20 that the above equations have indeed solutions. We can assume for simplifying that we have $t_0 = 0$. Now observe that we have:

$$\begin{aligned} & ||x(t) - y(t)|| \\ \leq\ & ||x_0 - y_0|| + \int_0^t |f(s, x(s)) - g(s, y(s))| ds \\ \leq\ & ||x_0 - y_0|| + \int_0^t \Big(|f(s, x(s)) - f(s, y(s))| + |f(s, y(s)) - g(s, y(s))| \Big) ds \\ \leq\ & ||x_0 - y_0|| + \int_0^t \Big(L||x(s) - y(s)|| + M \Big) ds \end{aligned}$$

In view of this estimate, consider the following function:

$$\psi(t) = ||x(t) - y(t)||$$

In terms of this function, the estimate that we found above reads:

$$\psi(t) \leq ||x_0 - y_0|| + \int_0^t (L\psi(s) + M) ds$$

But this shows that the Gronwall estimate from Proposition 6.23 applies, with the following choices for the constants $\alpha \in \mathbb{R}$, $\beta \geq 0$ and $\gamma \in \mathbb{R}$ appearing there:

$$\alpha = ||x_0 - y_0|| \quad , \quad \beta = L \quad , \quad \gamma = M$$

So, let us apply Proposition 6.23, with these values of α, β, γ. We obtain:

$$\begin{aligned} \psi(t) &\leq \alpha \exp(\beta t) + \frac{\gamma}{\beta}(\exp(\beta t) - 1) \\ &= ||x_0 - y_0|| e^{L|t - t_0|} + \frac{M}{L}\left(e^{L|t-t_0|} - 1\right) \end{aligned}$$

But this is exactly the estimate in the statement, as desired. □

6d. Dynamical systems

We discuss in the remainder of this chapter a very fruitful linearization idea, in the context of the dynamical systems, based on the following principle:

PRINCIPLE 6.25 (Linearization). *In order to deal with an arbitrary, non-linear system*

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

we can write the function f as follows, with $A = f'(x) \in M_N(\mathbb{R})$ being its derivative,

$$f(x) = Ax + o(||x||)$$

and then use, by perturbing, the results regarding the linear system $\dot{x} = Ax$.

Which sounds very good. In practice now, let us first go back to the general theory of the linear equations $\dot{x} = Ax$, as developed in chapter 5. We will need:

NOTATIONS 6.26. *Given $A \in M_N(\mathbb{R})$, we write its characteristic polynomial as*

$$P(z) = \prod_i (z - \alpha_i)^{a_i}$$

so that we have a direct sum decomposition of the ambient space, as follows:

$$\mathbb{C}^N = \bigoplus_i \ker\left[(A - \alpha_i)^{a_i}\right]$$

We also consider the corresponding geometric multiplicities, given by

$$g_i = \dim \ker(A - a_i)$$

and satisfying $g_i \leq a_i$, with equalities when A is diagonalizable.

We refer to chapter 5 for more on all this, theory and applications. Now back to the linear systems, $\dot{x} = Ax$, let us formulate the following key definition:

DEFINITION 6.27. *We say that a linear system $\dot{x} = Ax$ is hyperbolic when*

$$Re(\alpha) \neq 0$$

for any eigenvalue α. In this case, we consider the linear spaces

$$E^{\pm} = \bigoplus_{\pm Re(\alpha_i) < 0} \ker\left[(A - \alpha_i)^{a_i}\right]$$

which are therefore in direct sum position, $\mathbb{C}^N = E^+ \oplus E^-$.

So, studying these hyperbolic linear systems, and then extending our results to the hyperbolic non-linear systems, by using the derivative, will be our job, in what follows.

In what regards the study in the linear case, this is quickly done, by using the Jordan form for the matrix $A \in M_N(\mathbb{R})$, with the result being as follows:

THEOREM 6.28. *For a hyperbolic linear system $\dot{x} = Ax$, the following happen:*

(1) *The spaces $E^{\pm}$ are both invariant by the flow of the system.*
(2) *Any integral curve departing from $E^{\pm}$ converges to 0, with $t \to \pm\infty$.*
(3) *In fact, we have the following explicit estimate for the decay,*

$$|e^{tA} x_{\pm}| \leq C e^{\pm t\alpha} |x_{\pm}|$$

for any $\pm t > 0$ and any $x_{\pm} \in E^{\pm}$, with $\alpha > 0$ subject to

$$\alpha < \min\left\{|Re(\alpha_i)| : \pm Re(\alpha_i) < 0\right\}$$

and with $C > 0$ depending on α.

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

(1) This is something which is obvious.

(2) This is something that we already know, as a consequence of our general results from before, and which follows also from (3), that we will prove next.

(3) We just discuss here the proof of the "+" result, with the proof of the "−" result being similar. We put A in Jordan form, and we consider the following quantity:

$$m = \min\left\{|Re(\alpha_i)| : Re(\alpha_i) < 0\right\}$$

Now let $\alpha < m$ as in the statement, and set $\varepsilon = m - \alpha$. Then, for any eigenvalue satisfying $Re(\alpha_i) < 0$, the entry of maximal absolute value, say M_i, of the corresponding component $e^{tJ_{a_i}}$ of the matrix e^{tA}, can be estimated as follows:

$$\begin{aligned} M_i &= \frac{|t^n e^{a_i t}|}{F} \\ &\leq \frac{|t^n e^{-\varepsilon t}|e^{-\alpha t}}{F} \\ &\leq Ce^{-\alpha t} \end{aligned}$$

To be more precise, here F is a certain factorial, namely $F = (s-1)!$, with s being the size of the Jordan block, and $C > 0$ at the end is a certain constant, depending on this nunber F, and on α. Thus, we are led to the conclusion in the statement. □

Now consider a non-linear equation $\dot{x} = f(x)$, and denote by $\Phi(t, x)$ its flow, describing the solution in time t, with initial data $x(0) = x$. We have:

DEFINITION 6.29. *We associate to the equation $\dot{x} = f(x)$ the following sets,*

$$W^{\pm}(x_0) = \left\{x \,\middle|\, \lim_{t\to\pm\infty} \Phi(t, x) = x_0\right\}$$

gathering the initial data x such that the solution converges to x_0, with $t \to \pm\infty$.

Observe that both the above sets $W^{\pm}(x_0)$ are stable under the flow. In order now to compute these sets, we use our linearization idea. So, let us introduce as well:

DEFINITION 6.30. *We associate to the equation $\dot{x} = f(x)$ the sets*

$$M^{\pm,\alpha} = \left\{x \,\middle|\, \gamma_{\pm}(x) \subset U(x_0),\ \sup_{\pm t\geq 0} e^{\pm\alpha t}|\Phi(t, x) - x_0| < \infty\right\}$$

and then we consider the intersection of these sets, over eigenvalues,

$$M^{\pm}(x_0) = \bigcup_{\alpha>0} M^{\pm,\alpha}$$

which in the linear case, $\dot{x} = Ax$, are the spaces $E^{\pm}$ that we knew from before.

To be more precise here, in the linear case, $\dot{x} = Ax$, the spaces $M^{\pm,\alpha}$ constructed above correspond to the spaces $E^{\pm,\alpha}$ spanned by the eigenvectors of A corresponding to the eingenvalues satisfying $Re(\lambda) \geq \alpha$ and $Re(\lambda) \leq -\alpha$, and so by intersecting, we obtain indeed the spaces $E^{\pm}$ that we knew from before, as claimed in the above.

We can now formulate our main linearization result, as follows:

THEOREM 6.31. *For a hyperbolic point x_0, the following happen:*

(1) $M^{\pm}(x_0)$ *is a* C^1 *manifold.*
(2) $M^{\pm}(x_0)$ *is tangent to* $E^{\pm}$ *at* 0.
(3) $M^{\pm}(x_0) = W^{\pm}(x_0)$.

PROOF. The idea here is that of using the standard direct sum decomposition $\mathbb{R}^N = E^+ \oplus E^-$, in order to decompose everything, and prove the various assertions. For background and details, we refer to any geometry and dynamical systems books. □

6e. Exercises

Tough analytic chapter that we had here, and as exercises, we have:

EXERCISE 6.32. *Learn more about the Bernoulli equations, and their solutions.*

EXERCISE 6.33. *Learn more about the Riccati equations, and their solutions.*

EXERCISE 6.34. *Clarify what has been said above, about convex functions and Jensen.*

EXERCISE 6.35. *Learn more about L^p spaces, and about Banach spaces, in general.*

EXERCISE 6.36. *In case you skipped the proof of Gronwall, go back and read it.*

EXERCISE 6.37. *Learn more about hyperbolic linear systems, and their properties.*

EXERCISE 6.38. *Apply the theory developed in this chapter, to some concrete equations.*

EXERCISE 6.39. *Have a look as well at the non-linear hyperbolic systems.*

As bonus exercise, and no surprise here, purchase an ODE book, and read it.

CHAPTER 7

Singular values

7a. Functional calculus

Getting back to more traditional linear algebra, we have seen in chapter 5 the construction and some interesting applications of the exponential of the matrices:

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

As a first topic for this chapter, which is something very useful to know, we would like to systematically discuss the functional calculus for the complex matrices:

$$A \to f(A)$$

The general principle here, that we already met in chapter 4, is something quite intuitive, and which is very simple to state, as follows:

PRINCIPLE 7.1. *Under suitable regularity assumptions, on both functions and matrices, we can apply complex functions $f : \mathbb{C} \to \mathbb{C}$ to the complex matrices $A \in M_N(\mathbb{C})$,*

$$A \to f(A)$$

and at the level of the corresponding eigenvalues, if $\lambda_1, \ldots, \lambda_N \in \mathbb{C}$ are the eigenvalues of A, then $f(\lambda_1), \ldots, f(\lambda_N) \in \mathbb{C}$ should be the eigenvalues of A.

This is obviously something quite general, and potentially something very useful too, for all sorts of purposes. And, as explained in chapter 4, one good reason for which we can expect this principle to hold comes from the density of the diagonal matrices.

Indeed, assume first that our matrix is diagonalizable, as follows:

$$A = P \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix} P^{-1}$$

In this case, we can in principle define a matrix $f(A) \in M_N(\mathbb{C})$ by the following formula, and with its eigenvalues being indeed those predicted by Principle 7.1:

$$\begin{aligned} f(A) &= f\left[P\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}P^{-1}\right] \\ &= Pf\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}P^{-1} \\ &= P\begin{pmatrix}f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_N)\end{pmatrix}P^{-1} \end{aligned}$$

As for the general case, where our matrix $A \in M_N(\mathbb{C})$ is no longer assumed to be diagonalizable, this should normally follow from what we have above, by using the fact, that we know well from chapter 4, that the diagonalizable matrices are dense.

So, this was for the story, and in addition to this, we have already seen in chapter 5 how this method effectively works for the exponential function, $f(x) = e^x$.

In practice now, things are more complicated that this, because both our computation for the diagonal matrices, and the extension to the general case via density, normally require some regularity assumptions, on both f and A, in order to truly work.

Summarizing, things to do for us, and for dealing with this problem, we will use:

METHOD 7.2. *In order to establish Principle 7.1, we can use:*

(1) *Various direct methods.*
(2) *Density of the diagonalizable matrices.*
(3) *The Jordan form, and other more specialized results.*

Getting started now, let us first discuss in detail the case of the polynomials. As a warm-up target function here, we have $f(x) = x^2$, with the result being as follows:

THEOREM 7.3. *We can apply $f(x) = x^2$ to any matrix $A \in M_N(\mathbb{C})$,*

$$A \to A^2$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $\lambda_1^2, \ldots, \lambda_N^2$ are the eigenvalues of A^2.

PROOF. This does not look difficult to establish, but let us have a detailed discussion about this, which will be quite instructive, by following Method 7.2:

(1) Starting with bare hands, let us see if we can solve the problem at $N = 2$, via a direct computation, without using any trick. So, consider a 2×2 matrix, as follows:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

The square of this matrix is then given by the following formula:

$$A^2 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + bc & ab + bd \\ ac + cd & bc + d^2 \end{pmatrix}$$

Now let us compare the eigenvalues of the matrices A, A^2. Those of the initial matrix A, say $r, s \in \mathbb{C}$, are subject to the following two equations:

$$r + s = Tr(A) = a + d$$

$$rs = \det(A) = ad - bc$$

As for the eigenvalues of the squared matrix A^2, say $R, S \in \mathbb{C}$, these are subject to some similar equations, which are as follows:

$$R + S = Tr(A^2) = a^2 + d^2 + 2bc$$

$$RS = \det(A^2) = \det(A)^2$$

The second equality suggests that we should have $R = r^2, S = s^2$, so let us prove now that it is indeed so. For this purpose, we just have to check that the numbers $R = r^2, S = s^2$ sum up to the quantity computed above, and this is done as follows:

$$\begin{aligned} R + S &= r^2 + s^2 \\ &= (r + s)^2 - 2rs \\ &= (a + d)^2 - 2(ad - bc) \\ &= a^2 + d^2 + 2bc \end{aligned}$$

Summarizing, result proved with bare hands at $N = 2$. However, when thinking a bit, such methods will become quite complicated at $N \geq 3$, so we will stop here, with this.

(2) Still with bare hands, but allowing us some tricks, namely the use of square roots of complex numbers, here is how the result can be established, at any $N \in \mathbb{N}$. Consider our matrix $A \in M_N(\mathbb{C})$, and let us factorize its characteristic polynomial, as follows:

$$\det(A - x) = \prod_i (\lambda_i - x)$$

We have then the following computation, for the characteristic polynomial of A^2:

$$\begin{aligned}\det(A^2-x) &= \det((A-\sqrt{x})(A+\sqrt{x}))\\ &= \det(A-\sqrt{x})\det(A+\sqrt{x})\\ &= \prod_i(\lambda_i-\sqrt{x})\prod_i(\lambda_i+\sqrt{x})\\ &= \prod_i(\lambda_i-\sqrt{x})(\lambda_i+\sqrt{x})\\ &= \prod_i(\lambda_i^2-x)\end{aligned}$$

Thus, claimed proved, eventually, with bare hands, or almost.

(3) Getting now to more advanced tricks, bringing heavy simplifications, we can use here density arguments. Indeed, assume first that our matrix is diagonalizable:

$$A=P\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}P^{-1}$$

In this case, we have the following computation, for the squared matrix:

$$\begin{aligned}A^2 &= P\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}P^{-1}\cdot P\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}P^{-1}\\ &= P\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_N\end{pmatrix}^2P^{-1}\\ &= P\begin{pmatrix}\lambda_1^2 & & \\ & \ddots & \\ & & \lambda_N^2\end{pmatrix}P^{-1}\end{aligned}$$

Thus, claim proved in this case, and the general case follows now by using the fact, that we know well from chapter 4, that the diagonalizable matrices are dense.

(4) Finally, for our discussion to be complete, let us see as well what we get, by using a nuclear bomb, that is, the Jordan form. So, let us write the matrix to be squared in Jordan form, as in chapter 5, as follows, with P denoting the passage matrix:

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

The square of this matrix is then given by the following formula:

$$e^A = P\begin{pmatrix} J_1^2 & & \\ & \ddots & \\ & & J_k^2 \end{pmatrix} P^{-1}$$

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

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

The square of this Jordan block is then given by the following formula:

$$J = \begin{pmatrix} \lambda^2 & 2\lambda & 1 & & & \\ & \lambda^2 & 2\lambda & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \lambda^2 & 2\lambda & 1 \\ & & & & \lambda^2 & 2\lambda \\ & & & & & \lambda^2 \end{pmatrix}$$

Thus, we have the square λ^2 of the eigenvalue λ on the diagonal, and by putting everything together, we are again led to the conclusion in the statement. □

As a next target, we have $f(x) = x^k$, with $k \in \mathbb{N}$. Here the result is as follows:

THEOREM 7.4. *We can apply $f(x) = x^k$ to any matrix $A \in M_N(\mathbb{C})$,*

$$A \to A^k$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $\lambda_1^k, \ldots, \lambda_N^k$ are the eigenvalues of A^k.

PROOF. We must extend the proof of Theorem 7.3, and skipping the discussion there at the totally bare hand level, a simple way is by using the roots of unity. Consider indeed our matrix $A \in M_N(\mathbb{C})$, and let us factorize its characteristic polynomial, as follows:

$$\det(A - x) = \prod_i (\lambda_i - x)$$

We have then the following computation for the matrix A^k, with $w = e^{2\pi i/k}$:

$$\begin{aligned}
\det(A^k - x) &= \det\left(\prod_j \left(A - w^j \sqrt[k]{x}\right)\right) \\
&= \prod_j \det\left(A - w^j \sqrt[k]{x}\right) \\
&= \prod_j \prod_i \left(\lambda_i - w^j \sqrt[k]{x}\right) \\
&= \prod_i \prod_j \left(\lambda_i - w^j \sqrt[k]{x}\right) \\
&= \prod_i (\lambda_i^k - x)
\end{aligned}$$

Thus, we are led to the conclusion in the statement. Alternatively, we can of course use our usual, and elegant, density argument, or the Jordan form. □

With a bit more work, we can have the result for arbitrary polynomials, as follows:

THEOREM 7.5. *We can apply any $f \in \mathbb{C}[X]$ to any matrix $A \in M_N(\mathbb{C})$,*

$$A \to f(A)$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $f(\lambda_1), \ldots, f(\lambda_N)$ are those of $f(A)$.

PROOF. We must extend here the proof of Theorem 7.4, So, consider our matrix $A \in M_N(\mathbb{C})$, and let us factorize its characteristic polynomial, as follows:

$$\det(A - x) = \prod_i (\lambda_i - x)$$

Now fix $x \in \mathbb{C}$, and let us factorize the polynomial $f(z) - x$, as follows:

$$f(z) - x = c \prod_j (z - x_j)$$

We have then the following computation, for the matrix $f(A)$:

$$\begin{aligned}\det(f(A)-x) &= \det\left(c\prod_j(A-x_j)\right)\\ &= c^k\prod_j\det(A-x_j)\\ &= c^k\prod_j\prod_i(\lambda_i-x_j)\\ &= \prod_i\left(c\prod_j(\lambda_i-x_j)\right)\\ &= \prod_i(f(\lambda_i)-x)\end{aligned}$$

Thus, we are led to the conclusion in the statement. Alternatively, we can of course use our usual, and elegant, density argument, or the Jordan form. □

Getting now to more complicated functions, we first have the inverse function:

$$x\to x^{-1}$$

This function can only be applied to the invertible matrices, and we have here:

THEOREM 7.6. *We can apply $x\to x^{-1}$ to any invertible matrix $A\in M_N(\mathbb{C})$,*

$$A\to A^{-1}$$

and if $\lambda_1,\ldots,\lambda_N$ are the eigenvalues of A, then $\lambda_1^{-1},\ldots,\lambda_N^{-1}$ are those of A^{-1}.

PROOF. There are many possible arguments that we can use here, as follows:

(1) To start with, we can say that, given an invertible matrix $A\in M_N(\mathbb{C})$, the matrix $A-\lambda$ is invertible precisely when $A^{-1}-\lambda^{-1}$ is, and this due to:

$$\begin{aligned}A^{-1}-\lambda^{-1} &= A^{-1}(1-\lambda^{-1}A)\\ &= \lambda^{-1}A^{-1}(\lambda-A)\end{aligned}$$

(2) Alternatively, we can say that we have the following formula:

$$\begin{aligned}\det(A^{-1}-x) &= \det(A^{-1}(1-Ax))\\ &= \det(-xA^{-1}(A-x^{-1}))\\ &= (-x)^N\det(A^{-1})\det(A-x^{-1})\\ &= \frac{(-x)^N}{\det A}\cdot\det(A-x^{-1})\end{aligned}$$

Thus, we are once again led to the conclusion in the statement. Alternatively, we can of course use our usual density argument, or the Jordan form. □

Time now to have a break, and look at what we have in the above. In answer, what we have are Theorem 7.5 and Theorem 7.6, which in addition come with quite similar proofs. So, let us unify now these two statements. For this purpose, we will need:

DEFINITION 7.7. *A rational function $f \in \mathbb{C}(X)$ is a quotient of polynomials:*

$$f = \frac{P}{Q}$$

Assuming that P, Q are prime to each other, we can regard f as a usual function,

$$f : \mathbb{C} - X \to \mathbb{C}$$

with X being the set of zeroes of Q, also called poles of f.

We should mention here that the term "poles" comes from the fact that, if you want to imagine the graph of such a rational function f, in two complex dimensions, what you get is some sort of tent, supported by poles of infinite height, situated at the zeroes of Q. For more on all this, and on complex analysis in general, we refer as usual to Rudin [78]. Although a look at an abstract algebra book can be interesting as well.

Now that we have our class of functions, the next step consists in applying them to matrices. Here we cannot expect $f(A)$ to make sense for any f and any A, for instance because A^{-1} is defined only when A is invertible. We are led in this way to:

DEFINITION 7.8. *Given a matrix $A \in M_N(\mathbb{C})$, and a rational function $f = P/Q$ having poles outside the eigenvalues of A, we can construct the following matrix,*

$$f(A) = P(A)Q(A)^{-1}$$

that we can denote as a usual fraction, as follows,

$$f(A) = \frac{P(A)}{Q(A)}$$

due to the fact that $P(A), Q(A)$ commute, so that the order is irrelevant.

To be more precise, $f(A)$ is indeed well-defined, and the fraction notation is justified too. In more formal terms, we can say that we have a morphism of algebras as follows, with $\mathbb{C}(X)^A$ being the rational functions having poles outside the eigenvalues of A:

$$\mathbb{C}(X)^A \to M_N(\mathbb{C}) \quad , \quad f \to f(A)$$

Summarizing, we have now a good class of functions, generalizing both the polynomials and the inverse map $x \to x^{-1}$. We can now unify Theorems 7.5 and 7.6, as follows:

THEOREM 7.9. *Given a matrix $A \in M_N(\mathbb{C})$, we can apply to it any rational function $f \in \mathbb{C}(X)$ having its poles outside the eigenvalues of A,*

$$A \to f(A)$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $f(\lambda_1), \ldots, f(\lambda_N)$ are those of $f(A)$.

PROOF. We pick a scalar $\lambda \in \mathbb{C}$, we write $f = P/Q$, and we set:

$$F = P - \lambda Q$$

By using what we found in Theorem 7.5, for this polynomial $F \in \mathbb{C}[X]$, we have the following equivalence, with $\sigma(.)$ standing by definition for the set of eigenvalues:

$$\begin{aligned}
\lambda \in \sigma(f(A)) &\iff F(A) \notin M_N(\mathbb{C})^{-1}\\
&\iff 0 \in \sigma(F(A))\\
&\iff 0 \in F(\sigma(A))\\
&\iff \exists \mu \in \sigma(A), F(\mu) = 0\\
&\iff \lambda \in f(\sigma(A))
\end{aligned}$$

Thus, the eigenvalue set is the good one, and with a bit more work, say by using characteristic polynomials, as in the proofs of Theorem 7.5 and Theorem 7.6, the multiplicities match too. Thus, we are led to the conclusion in the statement. □

7b. General functions

Getting now to more general functions, our next objective will be that of unifying Theorem 7.9, dealing with the rational functions $f \in \mathbb{C}(X)$, with what we know from chapter 5 regarding the exponential, $f(x) = e^x$. Hang on, this will take some time.

To start with, we can extend what we know about $f(x) = e^x$, as follows:

THEOREM 7.10. *Given a matrix $A \in M_N(\mathbb{C})$, we can apply to it any entire holomorphic function $f : \mathbb{C} \to \mathbb{C}$,*

$$A \to f(A)$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $f(\lambda_1), \ldots, f(\lambda_N)$ are those of $f(A)$.

PROOF. For the existence part, this is something that we already know for $f(x) = e^x$, and the proof in general is similar, by using the Taylor series of f, as follows:

$$f(x) = \sum_k c_k x^k \quad \implies \quad f(A) = \sum_k c_k A^k$$

As for the second assertion, again this is something that we know for $f(x) = e^x$, coming via density of the diagonalizable matrices, or via the Jordan form, and the proof in general is similar, as explained earlier in this chapter, for the polynomials. □

Very nice all this, but as a drawback, we are now again in the dark, with Theorem 7.9 with Theorem 7.10 waiting to be unified. So, still lots of work to be done.

In general, the holomorphic functions are not entire, and the above method won't cover the rational functions $f \in \mathbb{C}(X)^A$ that we want to generalize. Thus, we must use

something else. And the answer here comes from the Cauchy formula:

$$f(a) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-a}\,dz$$

Indeed, given a rational function $f \in \mathbb{C}(X)^A$, the matrix $f(A) \in M_N(\mathbb{C})$, as constructed in Definition 7.8, can be recaptured in an analytic way, as follows:

$$f(A) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-A}\,dz$$

Now given an arbitrary function $f \in Hol(\sigma(A))$, we can define $f(A) \in M_N(\mathbb{C})$ by the exactly same formula, and we obtain in this way the desired correspondence:

$$Hol(\sigma(A)) \to M_N(\mathbb{C}) \quad , \quad f \to f(A)$$

This was for the plan. In practice now, all this needs a bit of care, with many verifications needed, and with the technical remark that a winding number must be added to the above Cauchy formulae, for things to be correct. Let us start with:

DEFINITION 7.11. *If γ is a loop in $\mathbb{C}$ the number of times γ goes around a point $z \in \mathbb{C} - \gamma$ is computed by the following integral, called winding number:*

$$Ind(\gamma, z) = \frac{1}{2\pi i}\int_\gamma \frac{d\xi}{\xi - z}$$

We say that γ turns around z if $Ind(\gamma, z) = 1$, and that it does not turn if $Ind(\gamma, z) = 0$. Otherwise, we say that γ turns around z many times, or in the bad sense, or both.

Let $f : U \to \mathbb{C}$ be an holomorphic function defined on an open subset of $\mathbb{C}$, and γ be a loop in U. If $Ind(\gamma, z) \neq 0$ for $z \in \mathbb{C} - U$ then $f(z)$ is given by the Cauchy formula:

$$Ind(\gamma, z)f(z) = \frac{1}{2\pi i}\int_\gamma \frac{f(\xi)}{\xi - z}\,d\xi$$

Also, if $Ind(\gamma, z) = 0$ for $z \in \mathbb{C} - U$ then the integral of f on γ is zero:

$$\int_\gamma f(\xi)\,d\xi = 0$$

It is convenient to use formal combinations of loops, called cycles:

$$\Sigma = n_1\gamma_1 + \ldots + n_r\gamma_r$$

The winding number for Σ is by definition the corresponding linear combination of winding numbers of its loop components, and the Cauchy formula holds for arbitrary cycles. Now by getting back to our questions regarding matrices, we can formulate:

DEFINITION 7.12. *Let $A \in M_N(\mathbb{C})$, and let $f : U \to \mathbb{C}$ be an holomorphic function defined on an open set containing $\sigma(A)$. we can define a matrix $f(A)$ by the formula*

$$f(A) = \frac{1}{2\pi i}\int_\Sigma \frac{f(\xi)}{\xi - A}\,d\xi$$

where Σ is a cycle in $U - \sigma(A)$ which turns around $\sigma(A)$ and doesn't turn around $\mathbb{C} - U$.

The formula makes sense because Σ is in $U - \sigma(A)$. Also, $f(A)$ is independent of the choice of Σ. Indeed, let Σ_1 and Σ_2 be two cycles. Their difference $\Sigma_1 - \Sigma_2$ is a cycle which doesn't turn around $\sigma(a)$, neither around $\mathbb{C} - U$. The function $z \to f(z)/(z - A)$ being holomorphic $U - \sigma(A) \to M_N(\mathbb{C})$, its integral on $\Sigma_1 - \Sigma_2$ must be zero:

$$\int_{\Sigma_1 - \Sigma_2} \frac{f(\xi)}{\xi - A}\,d\xi = 0$$

Thus $f(A)$ is the same with respect to Σ_1 and to Σ_2, and so Definition 7.12 is fully justified. Now with this definition in hand, we first have the following result:

PROPOSITION 7.13. *We have the formula*

$$f(A)g(A) = (fg)(A)$$

whenever the equality makes sense.

PROOF. Let Σ_1 be a cycle in $U - \sigma(A)$ around $\sigma(A)$ and consider the following set:

$$Int(\Sigma_1) = \left\{z \in \mathbb{C} - \Sigma_1 \middle| Ind(\Sigma_1, z) \neq 0\right\} \cup \Sigma_1$$

This is a compact set, included in U and containing the spectrum of A:

$$\sigma(T) \subset Int(\Sigma_1) \subset U$$

Let Σ_2 be a cycle in $U - Int(\Sigma_1)$ turning around $Int(\Sigma_1)$. Consider two holomorphic functions f, g defined around $\sigma(A)$, so that the statement make sense. We have:

$$\begin{aligned} f(A)g(A) &= \left(\frac{1}{2\pi i}\right)^2 \left(\int_{\Sigma_1} \frac{f(\xi)}{\xi - A}\,d\xi\right)\left(\int_{\Sigma_2} \frac{g(\eta)}{\eta - A}\,d\eta\right) \\ &= \left(\frac{1}{2\pi i}\right)^2 \int_{\Sigma_1}\int_{\Sigma_2} \frac{f(\xi)g(\eta)}{(\xi - A)(\eta - A)}\,d\eta d\xi \end{aligned}$$

In order to integrate, we can use the following identity:

$$\frac{1}{(\xi - A)(\eta - A)} = \frac{1}{(\eta - \xi)(\xi - A)} + \frac{1}{(\xi - \eta)(\eta - A)}$$

Thus our integral, and so our formula for $f(A)g(A)$, splits into two terms. The first term can be computed by integrating first over Σ_2, and we obtain:

$$\frac{1}{2\pi i}\int_{\Sigma_1} \frac{f(\xi)g(\xi)}{\xi - A}\,d\xi = (fg)(A)$$

As for the second term, here we can integrate first over Σ_1, and we get:

$$\frac{1}{2\pi i}\int_{\Sigma_2}\frac{g(\eta)}{\eta-A}\left(\frac{1}{2\pi i}\int_{\Sigma_1}\frac{f(\xi)}{\xi-\eta}d\xi\right)d\eta=0$$

It follows that $f(A)g(A)$ is equal to $(fg)(A)$, as claimed. □

We can now formulate our theorem regarding holomorphic functions, as follows:

THEOREM 7.14. *Given $A\in M_N(\mathbb{C})$, we have a morphism of algebras as follows, where $Hol(\sigma(A))$ is the algebra of functions which are holomorphic around $\sigma(A)$,*

$$Hol(\sigma(A))\to M_N(\mathbb{C})\quad,\quad f\to f(A)$$

which extends the previous rational functional calculus $f\to f(A)$. We have:

$$\sigma(f(A))=f(\sigma(A))$$

Moreover, if $\sigma(A)$ is contained in an open set U and $f_n,f:U\to\mathbb{C}$ are holomorphic functions such that $f_n\to f$ uniformly on compact subsets of U then $f_n(A)\to f(A)$.

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

(1) Consider indeed the algebra $Hol(\sigma(A))$, with the convention that two functions are identified if they coincide on an open set containing $\sigma(A)$. We have then a construction $f\to f(A)$ as in the statement, provided by Definition 7.12 and Proposition 7.13.

(2) Let us prove now that our construction extends the one for rational functions. Since $1,z$ generate $\mathbb{C}(X)$, it is enough to show that $f(z)=1$ implies $f(A)=1$, and that $f(z)=z$ implies $f(A)=A$. For this purpose, we prove that $f(z)=z^n$ implies $f(A)=A^n$ for any n. But this follows by integrating over a circle γ of big radius, as follows:

$$\begin{aligned}f(A)&=\frac{1}{2\pi i}\int_\gamma\frac{\xi^n}{\xi-A}\,d\xi\\&=\frac{1}{2\pi i}\int_\gamma\xi^{n-1}\left(1-\frac{A}{\xi}\right)^{-1}d\xi\\&=\frac{1}{2\pi i}\int_\gamma\xi^{n-1}\left(\sum_{k=0}^{\infty}\xi^{-k}A^k\right)d\xi\\&=\sum_{k=0}^{\infty}\left(\frac{1}{2\pi i}\int_\gamma\xi^{n-k-1}d\xi\right)A^k\\&=A^n\end{aligned}$$

(3) Regarding $\sigma(f(A))=f(\sigma(A))$, it is enough to prove that this equality holds on the point 0, and we can do this by double inclusion, as follows:

"$\supset$". Assume that $f(\sigma(A))$ contains 0, and let $z_0 \in \sigma(A)$ be such that $f(z_0) = 0$. Consider the function $g(z) = f(z)/(z - z_0)$. We have $g(A)(A - z_0) = f(A)$ by using the morphism property. Since $A - z_0$ is not invertible, $f(A)$ is not invertible either.

"$\subset$". Assume now that $f(\sigma(A))$ does not contain 0. With the holomorphic function $g(z) = 1/f(z)$ we get $g(A) = f(A)^{-1}$, so $f(A)$ is invertible, and we are done.

(4) Finally, regarding the last assertion, this is clear from definitions. And with the remark that this can be applied to holomorphic functions written as series:

$$f(z) = \sum_{k=0}^{\infty} c_k (z - z_0)^k$$

Indeed, if this is the expansion of f around z_0, with convergence radius r, and if $\sigma(A)$ is contained in the disc centered at z_0 of radius r, then $f(A)$ is given by:

$$f(A) = \sum_{k=0}^{\infty} c_k (A - z_0)^k$$

Summarizing, we have proved the result, and fully extended Theorem 7.9. □

As a conclusion to all this, you would say, good work that we did, with Theorem 7.14 seemingly being a ultimate result on the subject, that we can be proud of. Well, not exactly. What we found is just the tip of the iceberg, with many more functions, such as the continuous ones, or even the measurable ones, still waiting to be investigated.

Let us start with the following result, based on the spectral theorems from chapter 3:

THEOREM 7.15. *Given a normal matrix $A \in M_N(\mathbb{C})$, we have a morphism of algebras as follows, extending the previous holomorphic functional calculus,*

$$C(\sigma(A)) \to M_N(\mathbb{C}) \quad , \quad f \to f(A)$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $f(\lambda_1), \ldots, f(\lambda_N)$ are those of $f(A)$.

PROOF. The idea here is to "complete" our previous functional calculus results. Indeed, the simplest is to start with the polynomial calculus morphism, namely:

$$\mathbb{C}[X] \to M_N(\mathbb{C}) \quad , \quad P \to P(A)$$

We know from the above that this morphism is continuous, and is in fact isometric, when regarding the polynomials $P \in \mathbb{C}[X]$ as functions on $\sigma(A)$:

$$||P(A)|| = ||P_{|\sigma(A)}||$$

We conclude from this that we have a unique isometric extension, as follows:

$$C(\sigma(A)) \to M_N(\mathbb{C}) \quad , \quad f \to f(A)$$

It remains to prove $\sigma(f(A)) = f(\sigma(A))$, and we can do this by double inclusion:

"$\subset$" Given a continuous function $f \in C(\sigma(A))$, we must prove that we have:

$$\lambda \notin f(\sigma(A)) \implies \lambda \notin \sigma(f(A))$$

For this purpose, consider the following function, which is well-defined:

$$\frac{1}{f-\lambda} \in C(\sigma(A))$$

We can therefore apply this function to A, and we obtain:

$$\left(\frac{1}{f-\lambda}\right) A = \frac{1}{f(A)-\lambda}$$

In particular $f(A)-\lambda$ is invertible, so $\lambda \notin \sigma(f(A))$, as desired.

"$\supset$" Given a continuous function $f \in C(\sigma(A))$, we must prove that we have:

$$\lambda \in f(\sigma(A)) \implies \lambda \in \sigma(f(A))$$

But this is the same as proving that we have:

$$\mu \in \sigma(A) \implies f(\mu) \in \sigma(f(A))$$

For this purpose, we approximate our function by polynomials, $P_n \to f$, and we examine the following convergence, which follows from $P_n \to f$:

$$P_n(A) - P_n(\mu) \to f(A) - f(\mu)$$

We know from polynomial functional calculus that we have:

$$P_n(\mu) \in P_n(\sigma(A)) = \sigma(P_n(T))$$

Thus, the matrices $P_n(A) - P_n(\mu)$ are not invertible. On the other hand, we know that the set formed by the invertible matrices is open, so its complement is closed. Thus the limit $f(A) - f(\mu)$ is not invertible either, and so $f(\mu) \in \sigma(f(A))$, as desired. □

At a more advanced level, we have as well the following result:

THEOREM 7.16. *Given a normal matrix $A \in M_N(\mathbb{C})$, we have a morphism of algebras as follows, with L^∞ standing for the abstract measurable functions*

$$L^\infty(\sigma(A)) \to M_N(\mathbb{C}) \quad , \quad f \to f(A)$$

and if $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of A, then $f(\lambda_1), \ldots, f(\lambda_N)$ are those of $f(A)$.

PROOF. As before, the idea will be that of "completing" what we have. To be more precise, we can use the Riesz theorem and a polarization trick, as follows:

(1) Given a vector $x \in \mathbb{C}^N$, consider the following functional:

$$C(\sigma(A)) \to \mathbb{C} \quad , \quad g \to < g(A)x, x >$$

By the Riesz theorem, this functional must be the integration with respect to a certain measure μ on the space $\sigma(A)$. Thus, we have a formula as follows:

$$< g(A)x, x >= \int_{\sigma(A)} g(z)d\mu(z)$$

Now given an arbitrary Borel function $f \in L^\infty(\sigma(A))$, as in the statement, we can define a number $< f(A)x, x >\in \mathbb{C}$, by using exactly the same formula, namely:

$$< f(A)x, x >= \int_{\sigma(A)} f(z)d\mu(z)$$

Thus, we have managed to define numbers $< f(A)x, x >\in \mathbb{C}$, for all vectors $x \in \mathbb{C}^N$, and in addition we can recover these numbers as follows, with $g_n \in C(\sigma(A))$:

$$< f(A)x, x >= \lim_{g_n \to f} < g_n(A)x, x >$$

(2) In order to define now numbers $< f(A)x, y >\in \mathbb{C}$, for all vectors $x, y \in \mathbb{C}^N$, we can use a polarization trick. Indeed, for any matrix $B \in M_N(\mathbb{C})$ we have:

$$\begin{aligned} < B(x+y), x+y > &= < Bx, x > + < By, y > \\ &\quad + < Bx, y > + < By, x > \end{aligned}$$

By replacing $y \to iy$, we have as well the following formula:

$$\begin{aligned} < B(x+iy), x+iy > &= < Bx, x > + < By, y > \\ &\quad -i < Bx, y > +i < By, x > \end{aligned}$$

By multiplying this latter formula by i, we obtain the following formula:

$$\begin{aligned} i < B(x+iy), x+iy > &= i < Bx, x > +i < By, y > \\ &\quad + < Bx, y > - < By, x > \end{aligned}$$

Now by summing this latter formula with the first one, we obtain:

$$\begin{aligned} < B(x+y), x+y > +i < B(x+iy), x+iy > &= (1+i)[< Bx, x > + < By, y >] \\ &\quad +2 < Bx, y > \end{aligned}$$

(3) But with this, we can now finish. Indeed, by combining (1,2), given a Borel function $f \in L^\infty(\sigma(A))$, we can define numbers $< f(A)x, y >\in \mathbb{C}$ for any $x, y \in \mathbb{C}^N$, and it is routine to check, by using approximation by continuous functions $g_n \to f$ as in (1), that we obtain in this way a matrix $f(A) \in M_N(\mathbb{C})$, having all the desired properties. □

7c. Singular values

Still with me, I hope, after all the above mathematics, and in case you liked it, there will be some more later, in chapter 8, and time now for some applications.

Let us first discuss some basic decomposition results for the matrices $A \in M_N(\mathbb{C})$. We know that any $a \in \mathbb{C}$ can be written as follows, with $b, c \in \mathbb{R}$:

$$a = b + ic$$

Also, we know that both the real and imaginary parts $b, c \in \mathbb{R}$, and more generally any real number $d \in \mathbb{R}$, can be written as follows, with $e, f \geq 0$:

$$d = e - f$$

Here is the matrix-theoretical generalization of these results:

THEOREM 7.17. *Given a matrix $A \in M_N(\mathbb{C})$, the following happen:*

(1) *We can write $A = B + iC$, with $B, C \in M_N(\mathbb{C})$ being self-adjoint.*
(2) *When $A = A^*$, we can write $A = E - F$, with $E, F \in M_N(\mathbb{C})$ being positive.*
(3) *Thus, we can write any A as a linear combination of 4 positive matrices.*

PROOF. All this follows from basic spectral theory, as follows:

(1) This is something which comes from the following decomposition formula:

$$A = \frac{A + A^*}{2} + i \cdot \frac{A - A^*}{2i}$$

(2) This follows from the spectral theorem for self-adjoint matrices, by applying some suitable functions. Indeed, we can use the following decomposition formula on $\mathbb{R}$:

$$1 = \chi_{[0,\infty)} + \chi_{(-\infty,0)}$$

To be more precise, let us multiply by z, and rewrite this formula as follows:

$$z = \chi_{[0,\infty)} z - \chi_{(-\infty,0)}(-z)$$

Now by applying these measurable functions to A, we obtain as formula as follows, with both the matrices $A_+, A_- \in M_N(\mathbb{C})$ being positive, as desired:

$$A = A_+ - A_-$$

(3) This follows indeed by combining the results in (1) and (2) above. □

Going ahead with our decomposition results, another basic thing that we know about complex numbers is that any $a \in \mathbb{C}$ appears as a real multiple of a unitary:

$$a = re^{it}$$

In the case of the arbitrary matrices, finding the correct analogue of this key decomposition result is something quite tricky. As a basic result here, we have:

THEOREM 7.18. *Given a matrix $A \in M_N(\mathbb{C})$, the following happen:*

(1) *When $A = A^*$ and $||A|| \leq 1$, we can write A as an average of 2 unitaries:*

$$A = \frac{U + V}{2}$$

(2) *In the general $A = A^*$ case, we can write A as a rescaled sum of unitaries:*

$$A = \lambda(U + V)$$

(3) *Thus, in general, we can write A as a rescaled sum of 4 unitaries.*

PROOF. This follows from the results that we have, as follows:

(1) Assuming $A = A^*$ and $||A|| \leq 1$, it follows that we have:

$$1 - A^2 \geq 0$$

Our claim is that the decomposition result that we are looking can be taken as follows, with both the components on the right being unitaries:

$$A = \frac{A + i\sqrt{1 - A^2}}{2} + \frac{A - i\sqrt{1 - A^2}}{2}$$

To be more precise, the square root can be extracted in the usual way, and the check of the unitarity of the components goes as follows:

$$\begin{aligned}(A + i\sqrt{1 - A^2})(A - i\sqrt{1 - A^2}) &= A^2 - i^2\left(\sqrt{1 - A^2}\right)^2\\ &= A^2 + (1 - A^2)\\ &= 1\end{aligned}$$

(2) This simply follows by applying (1) to the following matrix:

$$A' = \frac{A}{||A||}$$

(3) Assuming first $||A|| \leq 1$, we know from Theorem 7.17 (1) that we can write A as follows, with B, C being self-adjoint, and satisfying $||B||, ||C|| \leq 1$:

$$A = B + iC$$

Now by applying (1) to both B and C, we obtain a decomposition of A as follows:

$$A = \frac{U + V + W + X}{2}$$

In general, we can apply this to the matrix $A/||A||$, and we obtain the result. □

All this gets us into the multiplicative theory of the complex numbers, that we will attempt to generalize now. As a first construction, that we would like to generalize to the complex matrix setting, we have the construction of the modulus, as follows:

$$|a| = \sqrt{a\bar{a}}$$

The point now is that, as we already know from chapter 3, we can indeed generalize this construction, by using the spectral theorem for the normal matrices, as follows:

THEOREM 7.19. *Given a matrix $A \in M_N(\mathbb{C})$, we can construct a matrix $|A|$ as follows, by using the fact that A^*A is diagonalizable, with positive eigenvalues:*

$$|A| = \sqrt{A^*A}$$

This matrix $|A|$ is then positive, and its square is $|A|^2 = A$. In the case $N = 1$, we obtain in this way the usual absolute value of the complex numbers.

PROOF. This is something that we already know, from chapter 3. Indeed, we can diagonalize our matrix as follows, with $U \in U_N$, and with D being diagonal:

$$A = UDU^*$$

From $A^*A \geq 0$ we obtain $D \geq 0$. But this means that the entries of D are real, and positive. Thus we can extract the square root $\sqrt{D}$, and then set:

$$\sqrt{A^*A} = U\sqrt{D}U^*$$

Thus, we are basically done. Indeed, if we call this latter matrix $|A|$, then we are led to the conclusions in the statement. Finally, the last assertion is clear from definitions. □

As a comment here, it is possible to talk as well about $\sqrt{AA^*}$, which is in general different from $\sqrt{A^*A}$. Note that when A is normal, there is no issue, because we have:

$$AA^* = A^*A \implies \sqrt{AA^*} = \sqrt{A^*A}$$

Regarding now the polar decomposition formula, for the complex matrices, this is again something that we know from chapter 3, the result here being as follows:

THEOREM 7.20. *Given a matrix $A \in M_N(\mathbb{C})$, the following happen:*

(1) *When A is invertible, we have $A = U|A|$, with U being a unitary.*
(2) *In general, we still have $A = U|A|$, with U being a partial isometry.*

PROOF. This is something that we know since chapter 3, but always good to talk about it again. According to our definition of the modulus, $|A| = \sqrt{A^*A}$, we have:

$$\begin{aligned} < |A|x, |A|y > &= < x, |A|^2 y > \\ &= < x, A^*Ay > \\ &= < Ax, Ay > \end{aligned}$$

We conclude that the following linear application is well-defined, and isometric:

$$U : Im|A| \to Im(A) \quad , \quad |A|x \to Ax$$

But now we can further extend this linear isometric map U into a partial isometry $U : \mathbb{C}^N \to \mathbb{C}^N$, in a straightforward way, by setting:

$$Ux = 0 \quad , \quad \forall x \in Im|A|^\perp$$

And the point is that, with this convention, the result follows. □

As a continuation of this, let us discuss now the singular value theorem, which is a key result in linear algebra. The result can be formulated, a bit abstractly, as follows:

THEOREM 7.21. *We can write the action of any matrix $A \in M_N(\mathbb{C})$ in the following form, with $\{e_n\}$, $\{f_n\}$ being orthonormal families, and with $\lambda_n \geq 0$:*

$$A(x) = \sum_n \lambda_n < x, e_n > f_n$$

The numbers λ_n, called singular values of A, are the eigenvalues of the modulus $|A|$. In fact, the polar decomposition of A is given by $A = U|A|$, with

$$|A|(x) = \sum_n \lambda_n < x, e_n > e_n$$

and with U being given by $Ue_n = f_n$, and $U = 0$ on the complement of $span(e_i)$.

PROOF. This basically comes from what we already have, as follows:

(1) Given two orthonormal families $\{e_n\}$, $\{f_n\}$, and a sequence of real numbers $\lambda_n \geq 0$, consider the linear map given by the formula in the statement, namely:

$$A(x) = \sum_n \lambda_n < x, e_n > f_n$$

The adjoint of this linear map is the given by the following formula:

$$A^*(x) = \sum_n \lambda_n < x, f_n > e_n$$

Thus, when composing A^* with A, we obtain the following linear map:

$$A^*A(x) = \sum_n \lambda_n^2 < x, e_n > e_n$$

Now by extracting the square root, we obtain the formula in the statement, namely:

$$|A|(x) = \sum_n \lambda_n < x, e_n > e_n$$

(2) Conversely, consider a matrix $A \in M_N(\mathbb{C})$. Then A^*A is self-adjoint, so we have a formula as follows, with $\{e_n\}$ being a certain orthonormal family, and with $\lambda_n \geq 0$:

$$A^*A(x) = \sum_n \lambda_n^2 < x, e_n > e_n$$

By extracting the square root we obtain the formula of $|A|$ in the statement, namely:

$$|A|(x) = \sum_n \lambda_n < x, e_n > e_n$$

Moreover, by setting $U(e_n) = f_n$, we obtain a second orthonormal family, $\{f_n\}$, such that the following formula holds:

$$A(x) = U|A| = \sum_n \lambda_n < x, e_n > f_n$$

Thus, our matrix $A \in M_N(\mathbb{C})$ appears indeed as in the statement. □

As before with the polar decomposition, there are many possible applications of the singular value theorem. We will be back to this, on several occasions, in what follows.

As a technical remark now, it is possible to slightly improve a part of the above statement. Consider indeed a linear map of the following form, with $\{e_n\}$, $\{f_n\}$ being orthonormal families as before, and with λ_n being now complex numbers:

$$A(x) = \sum_n \lambda_n < x, e_n > f_n$$

The adjoint of this linear map is the given by the following formula:

$$A^*(x) = \sum_n \bar{\lambda}_n < x, f_n > e_n$$

Thus, when composing A^* with A, we obtain the following linear map:

$$A^*A(x) = \sum_n |\lambda_n|^2 < x, e_n > e_n$$

Now by extracting the square root, we conclude that the polar decomposition of A is given by $A = U|A|$, with the modulus $|A|$ being as follows:

$$|A|(x) = \sum_n |\lambda_n| < x, e_n > e_n$$

As for the partial isometry U, this is given by $Ue_n = w_n f_n$, and $U = 0$ on the complement of $span(e_i)$, where $w_n \in \mathbb{T}$ are such that $\lambda_n = |\lambda_n| w_n$.

As already mentioned in the above, there are many possible applications of the singular value theorem. We will be back to this, on several occasions, in what follows. Also, we will discuss in chapter 8 below a remarkable generalization of the above results, to the

case of certain special infinite matrices, whose associated linear operators have suitable compactness properties, making them quite similar to the usual matrices.

7d. Triangularization

We discuss in the remainder of this chapter a number of more specialized decomposition results for the square matrices, which can stand as a useful complement to the main theoretical results that we have so far, namely Jordan decomposition and singular value decomposition, and which are very useful, in the context of applied linear algebra.

Let us start with something very basic, namely:

THEOREM 7.22. *Any complex matrix can be put in upper triangular form,*

$$A \sim \begin{pmatrix} U_{11} & U_{12} & \ldots & U_{1N} \\ 0 & U_{22} & \ldots & U_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & U_{NN} \end{pmatrix}$$

up to a suitable change of basis.

PROOF. This is indeed something very standard, and with the algorithm for doing so being something very intuitive. We will leave this as an exercise. □

We can of course use the same method for reaching to a lower triangular matrix:

THEOREM 7.23. *Any complex matrix can be put in lower triangular form,*

$$A \sim \begin{pmatrix} L_{11} & 0 & \ldots & 0 \\ L_{21} & L_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \ldots & L_{NN} \end{pmatrix}$$

up to a suitable change of basis.

PROOF. This follows from Theorem 7.22, by transposing the matrix in question, and in practice, is again something very standard. □

Of particular interest in applied linear algebra is the following decomposition result, which is something more specialized:

THEOREM 7.24. *Under suitable assumptions, we can put the matrices in LU form,*

$$A \sim \begin{pmatrix} L_{11} & 0 & \ldots & 0 \\ L_{21} & L_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \ldots & L_{NN} \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} & \ldots & U_{1N} \\ 0 & U_{22} & \ldots & U_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & U_{NN} \end{pmatrix}$$

up to a suitable change of basis.

PROOF. This is again something standard, obtained by further building on the above, and we will leave some thinking and reading here as an instructive exercise. □

As a technical version of the above result, which is again very useful, in relation with various applied linear algebra questions, we have:

THEOREM 7.25. *Under suitable assumptions, we can put the matrices in LDU form,*

$$A\sim\begin{pmatrix}1&0&\ldots&0\\L_{21}&1&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\L_{N1}&L_{N2}&\ldots&1\end{pmatrix}\begin{pmatrix}D_{11}&0&\ldots&0\\0&D_{22}&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&D_{NN}\end{pmatrix}\begin{pmatrix}1&U_{12}&\ldots&U_{1N}\\0&1&\ldots&U_{2N}\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&1\end{pmatrix}$$

up to a suitable change of basis.

PROOF. This is again something standard, that we will leave as an exercise. □

As a continuation of this, we have some further decomposition results of similar type, which hold in general, called LUP and PLU factorizations. We first have:

THEOREM 7.26. *We can put the matrices in LUP form,*

$$A=\begin{pmatrix}L_{11}&0&\ldots&0\\L_{21}&L_{22}&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\L_{N1}&L_{N2}&\ldots&L_{NN}\end{pmatrix}\begin{pmatrix}U_{11}&U_{12}&\ldots&U_{1N}\\0&U_{22}&\ldots&U_{2N}\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&U_{NN}\end{pmatrix}P$$

with P being a suitable permutation matrix.

PROOF. This is again something standard, coming as usual, as an exercise. □

Along the same lines, we have as well the following result:

THEOREM 7.27. *We can put the matrices in PLU form,*

$$A=P\begin{pmatrix}L_{11}&0&\ldots&0\\L_{21}&L_{22}&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\L_{N1}&L_{N2}&\ldots&L_{NN}\end{pmatrix}\begin{pmatrix}U_{11}&U_{12}&\ldots&U_{1N}\\0&U_{22}&\ldots&U_{2N}\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&U_{NN}\end{pmatrix}$$

with P being a suitable permutation matrix.

PROOF. This is again something standard, which follows from Theorem 7.26, by transposing the matrix in question. Of course, many other things can be said here. □

Along the same lines, and at a more specialized level, we can make use of both row and column permutations, and we are led in this way to the following result:

THEOREM 7.28. *We can put the matrices in the following form,*

$$PAQ=\begin{pmatrix}L_{11}&0&\ldots&0\\L_{21}&L_{22}&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\L_{N1}&L_{N2}&\ldots&L_{NN}\end{pmatrix}\begin{pmatrix}U_{11}&U_{12}&\ldots&U_{1N}\\0&U_{22}&\ldots&U_{2N}\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&U_{NN}\end{pmatrix}$$

with P,Q being suitable permutation matrices.

PROOF. This is again something standard, that we will leave as an exercise. □

Switching topics now, but still in relation with the triangular matrices, we have:

THEOREM 7.29. *Any square matrix can be put in QR form,*

$$A=Q\begin{pmatrix}R_{11}&R_{12}&\ldots&R_{1N}\\0&R_{22}&\ldots&R_{2N}\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&R_{NN}\end{pmatrix}$$

with Q being an orthogonal matrix.

PROOF. This is again something standard, coming as before, as an exercise. □

We can of course use the same method for reaching to a lower triangular matrix:

THEOREM 7.30. *Any square matrix can be put in LQ form,*

$$A=\begin{pmatrix}L_{11}&0&\ldots&0\\L_{21}&L_{22}&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\L_{N1}&L_{N2}&\ldots&L_{NN}\end{pmatrix}Q$$

with Q being an orthogonal matrix.

PROOF. This follows indeed from Theorem 7.29, by transposing the matrix in question, and in practice, is again something very standard. □

As a remark here, besides the QR and LQ decompositions discussed in the above, we can talk as well about QL and RQ decompositions, in the obvious way.

There are as well some interesting generalizations of the above, to the case of the rectangular matrices. But more on rectangular matrices later in this book, when really needing them, and we will be back to this notably in chapter 11 below.

Also, there are many other decomposition results for the matrices, which are more specialized, quite often making a clever use of various classes of unitary matrices. For more on all this, we refer to any truly advanced, applied linear algebra book. In what concerns us, we will be back to this later in this book, when discussing the Lie groups.

7e. Exercises

This was a quite advanced chapter, and as exercises on all this, we have:

EXERCISE 7.31. *Further develop the functional calculus with* $f(x) = \sqrt{x}$.

EXERCISE 7.32. *Further develop the functional calculus with* $f(x) = \chi_{[0,\infty)}(x)$.

EXERCISE 7.33. *Work out all the details of the triangularization procedure.*

EXERCISE 7.34. *Work out the details of the* LU *decomposition.*

EXERCISE 7.35. *Work out the details of the* LDU *decomposition.*

EXERCISE 7.36. *Work out the details of the* LUP *decomposition.*

EXERCISE 7.37. *Work out the details of the* QR *decomposition.*

EXERCISE 7.38. *Learn as well some further, more specialized decomposition results.*

As bonus exercise, learn some systematic complex analysis, as much as you can.

CHAPTER 8

Infinite matrices

8a. Infinite matrices

We discuss here some extensions of the above results, and notably of the singular value decomposition theorem, to the case of the infinite matrices. Let us start with:

DEFINITION 8.1. *A Hilbert space is a complex vector space H with a scalar product $< x, y >$, which will be linear at left and antilinear at right,*

$$< \lambda x, y >= \lambda < x, y > \quad , \quad < x, \lambda y >= \bar{\lambda} < x, y >$$

and which is complete with respect to corresponding norm

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

in the sense that any sequence $\{x_n\}$ which is a Cauchy sequence, having the property $||x_n - x_m|| \to 0$ with $n, m \to \infty$, has a limit, $x_n \to x$.

Here our convention for the scalar products, written $< x, y >$ and being linear at left, is one among others, often used by mathematicians. At the level of examples, we have:

THEOREM 8.2. *Given an index set I, which can be finite or not, the space of square-summable vectors having indices in I, namely*

$$l^2(I) = \left\{ (x_i)_{i \in I} \middle| \sum_i |x_i|^2 < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

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

When I is finite, $I = \{1, \ldots, N\}$, we obtain in this way the usual space $H = \mathbb{C}^N$.

PROOF. We have already met such things before, but let us recall all this:

(1) We know that $l^2(I) \subset \mathbb{C}^I$ is the space of vectors satisfying $||x|| < \infty$. We want to prove that $l^2(I)$ is a vector space, that $< x, y >$ is a scalar product on it, that $l^2(I)$ is complete with respect to $||.||$, and finally that for $|I| < \infty$ we have $l^2(I) = \mathbb{C}^{|I|}$.

(2) The last assertion, $l^2(I) = \mathbb{C}^{|I|}$ for $|I| < \infty$, is clear, because in this case the sums are finite, so the condition $||x|| < \infty$ is automatic. So, we know at least one thing.

(3) Regarding the rest, our claim here, which will more or less prove everything, is that for any two vectors $x, y \in l^2(I)$ we have the Cauchy-Schwarz inequality:

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

But this follows from the positivity of the following degree 2 quantity, depending on a real variable $t \in \mathbb{R}$, and on a variable on the unit circle, $w \in \mathbb{T}$:

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

(4) Now with Cauchy-Schwarz proved, everything is straightforward. We first obtain, by raising to the square and expanding, that for any $x, y \in l^2(I)$ we have:

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

Thus $l^2(I)$ is indeed a vector space, the other vector space conditions being trivial.

(5) Also, $< x, y >$ is surely a scalar product on this vector space, because all the conditions for a scalar product are trivially satisfied.

(6) Finally, the fact that our space $l^2(I)$ is indeed complete with respect to its norm $||.||$ follows in the obvious way, the limit of a Cauchy sequence $\{x_n\}$ being the vector $y = (y_i)$ given by $y_i = \lim_{n\to\infty} x_{ni}$, with all the verifications here being trivial. □

Going now a bit more abstract, we have, more generally, the following result:

THEOREM 8.3. *Given an arbitrary space X with a positive measure μ on it, the space of square-summable complex functions on it, namely*

$$L^2(X) = \left\{ f : X \to \mathbb{C} \Big| \int_X |f(x)|^2 \, d\mu(x) < \infty \right\}$$

is a Hilbert space, with scalar product as follows:

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

When $X = I$ is discrete, meaning that the measure μ on it is the counting measure, $\mu(\{x\}) = 1$ for any $x \in X$, we obtain in this way the previous spaces $l^2(I)$.

PROOF. This is something routine, remake of Theorem 8.2, as follows:

(1) The proof of the first, and main assertion is something perfectly similar to the proof of Theorem 8.2, by replacing everywhere the sums by integrals.

(2) With the remark that we forgot to say in the statement that the L^2 functions are by definition taken up to equality almost everywhere, $f = g$ when $||f - g|| = 0$.

(3) As for the last assertion, when μ is the counting measure all our integrals here become usual sums, and so we recover in this way Theorem 8.2. □

As a third and last theorem about Hilbert spaces, that we will need, we have:

THEOREM 8.4. *Any Hilbert space H has an orthonormal basis $\{e_i\}_{i\in I}$, which is by definition a set of vectors whose span is dense in H, and which satisfy*

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

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

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

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

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

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

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

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

(4) Finally, everything at the end is clear from definitions, except perhaps for the fact that $L^2(\mathbb{R})$ is separable. But here we can argue that, since functions can be approximated by polynomials, we have a countable algebraic basis, namely $\{x^n\}_{n\in\mathbb{N}}$, called the Weierstrass basis, that we can orthogonalize afterwards by using Gram-Schmidt. □

Moving ahead, now that we know what our vector spaces are, we can talk about infinite matrices with respect to them. And the situation here is as follows:

THEOREM 8.5. *Given a Hilbert space H, consider the linear operators $T : H \to H$, and for each such operator define its norm by the following formula:*

$$||T|| = \sup_{||x||=1} ||Tx||$$

The operators which are bounded, $||T|| < \infty$, form then a complex algebra $B(H)$, which is complete with respect to $||.||$. When H comes with a basis $\{e_i\}_{i\in I}$, we have

$$B(H) \subset \mathcal{L}(H) \subset M_I(\mathbb{C})$$

where $\mathcal{L}(H)$ is the algebra of all linear operators $T : H \to H$, and $\mathcal{L}(H) \subset M_I(\mathbb{C})$ is the correspondence $T \to M$ obtained via the usual linear algebra formulae, namely:

$$T(x) = Mx \quad , \quad M_{ij} =< Te_j, e_i >$$

In infinite dimensions, none of the above two inclusions is an equality.

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

(1) The fact that we have indeed an algebra, satisfying the product condition in the statement, follows from the following estimates, which are all elementary:

$$||S+T|| \leq ||S|| + ||T|| \quad , \quad ||\lambda T|| = |\lambda| \cdot ||T|| \quad , \quad ||ST|| \leq ||S|| \cdot ||T||$$

(2) Regarding now the completness assertion, if $\{T_n\} \subset B(H)$ is Cauchy then $\{T_n x\}$ is Cauchy for any $x \in H$, so we can define the limit $T = \lim_{n\to\infty} T_n$ by setting:

$$Tx = \lim_{n\to\infty} T_n x$$

Let us first check that the application $x \to Tx$ is linear. We have:

$$\begin{aligned} T(x+y) &= \lim_{n\to\infty} T_n(x+y) \\ &= \lim_{n\to\infty} T_n(x) + T_n(y) \\ &= \lim_{n\to\infty} T_n(x) + \lim_{n\to\infty} T_n(y) \\ &= T(x) + T(y) \end{aligned}$$

Similarly, we have $T(\lambda x) = \lambda T(x)$, and we conclude that $T \in \mathcal{L}(H)$.

(3) With this done, it remains to prove now that we have $T \in B(H)$, and that $T_n \to T$ in norm. For this purpose, observe that we have:

$$\begin{aligned} ||T_n - T_m|| \leq \varepsilon \ , \ \forall n, m \geq N &\implies ||T_n x - T_m x|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n, m \geq N \\ &\implies ||T_n x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \ , \ \forall n \geq N \\ &\implies ||T_N x - Tx|| \leq \varepsilon \ , \ \forall ||x|| = 1 \\ &\implies ||T_N - T|| \leq \varepsilon \end{aligned}$$

But this gives both $T \in B(H)$, and $T_N \to T$ in norm, and we are done.

(4) Regarding the embeddings, the correspondence $T \to M$ in the statement is indeed linear, and its kernel is $\{0\}$, so we have indeed an embedding as follows, as claimed:

$$\mathcal{L}(H) \subset M_I(\mathbb{C})$$

In finite dimensions we have an isomorphism, because any $M \in M_N(\mathbb{C})$ determines an operator $T : \mathbb{C}^N \to \mathbb{C}^N$, given by $< Te_j, e_i >= M_{ij}$. However, in infinite dimensions, we have matrices not producing operators, as for instance the all-one matrix.

(5) As for the examples of linear operators which are not bounded, these are more complicated, coming from logic, and we will not need them in what follows. □

Finally, as a second and last basic result regarding the operators, we will need:

THEOREM 8.6. *Each operator $T \in B(H)$ has an adjoint $T^* \in B(H)$, given by:*

$$< Tx, y >=< x, T^*y >$$

The operation $T \to T^*$ is antilinear, antimultiplicative, involutive, and satisfies:

$$||T|| = ||T^*|| \quad , \quad ||TT^*|| = ||T||^2$$

When H comes with a basis $\{e_i\}_{i\in I}$, the operation $T \to T^*$ corresponds to

$$(M^*)_{ij} = \overline{M}_{ji}$$

at the level of the associated matrices $M \in M_I(\mathbb{C})$.

PROOF. This is standard too, and can be proved in 3 steps, as follows:

(1) The existence of the adjoint operator T^*, given by the formula in the statement, comes from the fact that the function $\varphi(x) =< Tx, y >$ being a linear map $H \to \mathbb{C}$, we must have a formula as follows, for a certain vector $T^*y \in H$:

$$\varphi(x) =< x, T^*y >$$

Moreover, since this vector is unique, T^* is unique too, and we have as well:

$$(S + T)^* = S^* + T^* \quad , \quad (\lambda T)^* = \bar{\lambda} T^* \quad , \quad (ST)^* = T^* S^* \quad , \quad (T^*)^* = T$$

Observe also that we have indeed $T^* \in B(H)$, because:

$$\begin{aligned} ||T|| &= \sup_{||x||=1} \sup_{||y||=1} < Tx, y > \\ &= \sup_{||y||=1} \sup_{||x||=1} < x, T^*y > \\ &= ||T^*|| \end{aligned}$$

(2) Regarding now $||TT^*|| = ||T||^2$, which is a key formula, observe that we have:

$$||TT^*|| \leq ||T|| \cdot ||T^*|| = ||T||^2$$

On the other hand, we have as well the following estimate:

$$\begin{aligned} ||T||^2 &= \sup_{||x||=1} | < Tx, Tx > | \\ &= \sup_{||x||=1} | < x, T^*Tx > | \\ &\leq ||T^*T|| \end{aligned}$$

By replacing $T \to T^*$ we obtain from this $||T||^2 \leq ||TT^*||$, as desired.

(3) Finally, when H comes with a basis, the formula $< Tx, y >=< x, T^*y >$ applied with $x = e_i$, $y = e_j$ translates into the formula $(M^*)_{ij} = \overline{M}_{ji}$, as desired. □

8b. Spectral radius

We discuss now the diagonalization problem for the operators $T \in B(H)$, in analogy with the diagonalization problem for the usual matrices $A \in M_N(\mathbb{C})$. We first have:

DEFINITION 8.7. *The spectrum of an operator $T \in B(H)$ is the set*

$$\sigma(T) = \left\{\lambda \in \mathbb{C}\middle| T - \lambda \notin B(H)^{-1}\right\}$$

where $B(H)^{-1} \subset B(H)$ is the set of invertible operators.

As a basic example, in the finite dimensional case, $H = \mathbb{C}^N$, the spectrum of a usual matrix $A \in M_N(\mathbb{C})$ is the collection of its eigenvalues, taken without multiplicities. We will see many other examples. In general, the spectrum has the following properties:

PROPOSITION 8.8. *The spectrum of $T \in B(H)$ contains the eigenvalue set*

$$\varepsilon(T) = \left\{\lambda \in \mathbb{C}\middle| \ker(T - \lambda) \neq \{0\}\right\}$$

and $\varepsilon(T) \subset \sigma(T)$ is an equality in finite dimensions, but not in infinite dimensions.

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

(1) First of all, the eigenvalue set is indeed the one in the statement, because $Tx = \lambda x$ tells us precisely that $T - \lambda$ must be not injective. The fact that we have $\varepsilon(T) \subset \sigma(T)$ is clear as well, because if $T - \lambda$ is not injective, it is not bijective.

(2) In finite dimensions we have $\varepsilon(T) = \sigma(T)$, because $T - \lambda$ is injective if and only if it is bijective, with the boundedness of the inverse being automatic.

(3) In infinite dimensions we can assume $H = l^2(\mathbb{N})$, and the shift operator $S(e_i) = e_{i+1}$ is injective but not surjective. Thus $0 \in \sigma(T) - \varepsilon(T)$. □

Philosophically, the best way of thinking at this is as follows: the numbers $\lambda \notin \sigma(T)$ are good, because we can invert $T - \lambda$, the numbers $\lambda \in \sigma(T) - \varepsilon(T)$ are bad, because so they are, and the eigenvalues $\lambda \in \varepsilon(T)$ are evil. Welcome to operator theory.

Let us develop now some general theory. Here is a first basic result regarding the spectra, inspired from what happens in finite dimensions, and which shows that things do not necessarily extend without troubles to the infinite dimensional setting:

THEOREM 8.9. *We have the following formula, valid for any operators S, T:*

$$\sigma(ST) \cup \{0\} = \sigma(TS) \cup \{0\}$$

In finite dimensions we have $\sigma(ST) = \sigma(TS)$, but this fails in infinite dimensions.

PROOF. There are several assertions here, the idea being as follows:

(1) Let us first prove the main assertion, stating that $\sigma(ST), \sigma(TS)$ coincide outside 0. We first prove that we have the following implication:

$$1 \notin \sigma(ST) \implies 1 \notin \sigma(TS)$$

Assume indeed that $1 - ST$ is invertible, with inverse denoted R:

$$R = (1 - ST)^{-1}$$

We have then the following formulae, relating our variables R, S, T:

$$RST = STR = R - 1$$

By using $RST = R - 1$, we have the following computation:

$$\begin{aligned}(1 + TRS)(1 - TS) &= 1 + TRS - TS - TRSTS \\ &= 1 + TRS - TS - TRS + TS \\ &= 1\end{aligned}$$

A similar computation, using $STR = R - 1$, shows that we have:

$$(1 - TS)(1 + TRS) = 1$$

Thus $1 - TS$ is invertible, with inverse $1 + TRS$, which proves our claim. Now by multiplying by scalars, we deduce that for any $\lambda \in \mathbb{C} - \{0\}$ we have, as desired:

$$\lambda \notin \sigma(ST) \implies \lambda \notin \sigma(TS)$$

(2) Regarding now the counterexample to the formula $\sigma(ST) = \sigma(TS)$, in general, let us take S to be the shift on $H = L^2(\mathbb{N})$, given by the following formula:

$$S(e_i) = e_{i+1}$$

As for T, we can take it to be the adjoint of S, which is the following operator:

$$S^*(e_i) = \begin{cases} e_{i-1} & \text{if } i > 0 \\ 0 & \text{if } i = 0 \end{cases}$$

Let us compose now these two operators. In one sense, we have:

$$S^*S = 1 \implies 0 \notin \sigma(SS^*)$$

In the other sense, however, the situation is different, as follows:

$$SS^* = Proj(e_0^\perp) \implies 0 \in \sigma(SS^*)$$

Thus, the spectra do not match on 0, and we have our counterexample, as desired. □

Let us develop now some systematic theory for the computation of the spectra, based on what we know about the eigenvalues of the usual complex matrices. As a first result, which is well-known for the usual matrices, and extends well, we have:

THEOREM 8.10. *We have the "rational functional calculus" formula*

$$\sigma(f(T)) = f(\sigma(T))$$

valid for any rational function $f \in \mathbb{C}(X)$ having poles outside $\sigma(T)$.

PROOF. This can be proved in two steps, as follows:

(1) Assume first that our rational function $f \in \mathbb{C}(X)$ is a usual polynomial $P \in \mathbb{C}[X]$. We pick a scalar $\lambda \in \mathbb{C}$, and we decompose the polynomial $P - \lambda$, as follows:

$$P(X) - \lambda = c(X - r_1) \ldots (X - r_n)$$

We have then the following equivalences, which give the result:

$$\begin{aligned}
\lambda \notin \sigma(P(T)) &\iff P(T) - \lambda \in B(H)^{-1} \\
&\iff c(T - r_1) \ldots (T - r_n) \in B(H)^{-1} \\
&\iff T - r_1, \ldots, T - r_n \in B(H)^{-1} \\
&\iff r_1, \ldots, r_n \notin \sigma(T) \\
&\iff \lambda \notin P(\sigma(T))
\end{aligned}$$

(2) In general now, we pick a scalar $\lambda \in \mathbb{C}$, we write $f = P/Q$, and we set $F = P - \lambda Q$. By using what we found in (1), for this polynomial $F \in \mathbb{C}[X]$, we obtain:

$$\begin{aligned}
\lambda \in \sigma(f(T)) &\iff F(T) \notin B(H)^{-1} \\
&\iff 0 \in \sigma(F(T)) \\
&\iff 0 \in F(\sigma(T)) \\
&\iff \exists \mu \in \sigma(T), F(\mu) = 0 \\
&\iff \lambda \in f(\sigma(T))
\end{aligned}$$

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

As a first application of the above methods, we have the following key result:

THEOREM 8.11. *The following happen:*

(1) *For a unitary operator, $U^* = U^{-1}$, we have $\sigma(U) \subset \mathbb{T}$.*
(2) *For a self-adjoint operator, $T = T^*$, we have $\sigma(T) \subset \mathbb{R}$.*

PROOF. This is something quite tricky, based on Theorem 8.10, as follows:

(1) Assuming $U^* = U^{-1}$, we have the following norm computation:

$$||U|| = \sqrt{||UU^*||} = \sqrt{1} = 1$$

Now if we denote by D the unit disk, we obtain from this:

$$\sigma(U) \subset D$$

On the other hand, once again by using $U^* = U^{-1}$, we have as well:

$$||U^{-1}|| = ||U^*|| = ||U|| = 1$$

Thus, as before with D being the unit disk in the complex plane, we have:

$$\sigma(U^{-1}) \subset D$$

Now by using Theorem 8.10, we obtain $\sigma(U) \subset D \cap D^{-1} = \mathbb{T}$, as desired.

(2) Consider the following rational function, depending on a parameter $r \in \mathbb{R}$:

$$f(z) = \frac{z + ir}{z - ir}$$

Then for $r >> 0$ the operator $f(T)$ is well-defined, and we have:

$$\left(\frac{T + ir}{T - ir}\right)^* = \frac{T - ir}{T + ir} = \left(\frac{T + ir}{T - ir}\right)^{-1}$$

Thus $f(T)$ is unitary, and by (1) we have $\sigma(T) \subset f^{-1}(\mathbb{T}) = \mathbb{R}$, as desired. □

In order to formulate our next result, we will need the following notion:

DEFINITION 8.12. *Given an operator $T \in B(H)$, its spectral radius*

$$\rho(T) \in \big[0, ||T||\big]$$

is the radius of the smallest disk centered at 0 containing $\sigma(T)$.

Now with this notion in hand, we have the following key result:

THEOREM 8.13. *The spectral radius of an operator $T \in B(H)$ is given by*

$$\rho(T) = \lim_{n\to\infty} ||T^n||^{1/n}$$

and in this formula, we can replace the limit by an inf.

PROOF. We have several things to be proved, the idea being as follows:

(1) Our first claim is that the numbers $u_n = ||T^n||^{1/n}$ satisfy:

$$(n + m)u_{n+m} \leq nu_n + mu_m$$

Indeed, we have the following estimate, using the Young inequality $ab \leq a^p/p + b^q/q$, with exponents $p = (n + m)/n$ and $q = (n + m)/m$:

$$\begin{aligned} u_{n+m} &= ||T^{n+m}||^{1/(n+m)} \\ &\leq ||T^n||^{1/(n+m)}||T^m||^{1/(n+m)} \\ &\leq ||T^n||^{1/n} \cdot \frac{n}{n+m} + ||T^m||^{1/m} \cdot \frac{m}{n+m} \\ &= \frac{nu_n + mu_m}{n+m} \end{aligned}$$

(2) Our second claim is that the second assertion holds, namely:

$$\lim_{n\to\infty} ||T^n||^{1/n} = \inf_n ||T^n||^{1/n}$$

For this purpose, we just need the inequality found in (1). Indeed, fix $m \geq 1$, let $n \geq 1$, and write $n = lm + r$ with $0 \leq r \leq m-1$. By using twice $u_{ab} \leq u_b$, we get:

$$u_n \leq \frac{1}{n}(lmu_{lm} + ru_r) \leq u_m + \frac{r}{n}\, u_1$$

It follows that we have $\limsup_n u_n \leq u_m$, which proves our claim.

(3) Summarizing, we are left with proving the main formula, which is as follows, and with the remark that we already know that the sequence on the right converges:

$$\rho(T) = \lim_{n\to\infty} ||T^n||^{1/n}$$

In one sense, we can use the polynomial calculus formula $\sigma(T^n) = \sigma(T)^n$. Indeed, this gives the following estimate, valid for any n, as desired:

$$\begin{aligned}
\rho(T) &= \sup_{\lambda\in\sigma(T)} |\lambda| \\
&= \sup_{\rho\in\sigma(T)^n} |\rho|^{1/n} \\
&= \sup_{\rho\in\sigma(T^n)} |\rho|^{1/n} \\
&= \rho(T^n)^{1/n} \\
&\leq ||T^n||^{1/n}
\end{aligned}$$

(4) For the reverse inequality, we fix a number $\rho > \rho(T)$, and we want to prove that we have $\rho \geq \lim_{n\to\infty} ||T^n||^{1/n}$. By using the Cauchy formula, we have:

$$\begin{aligned}
\frac{1}{2\pi i}\int_{|z|=\rho} \frac{z^n}{z-T}\, dz &= \frac{1}{2\pi i}\int_{|z|=\rho} \sum_{k=0}^{\infty} z^{n-k-1}T^k\, dz \\
&= \sum_{k=0}^{\infty} \frac{1}{2\pi i}\left(\int_{|z|=\rho} z^{n-k-1}dz\right) T^k \\
&= \sum_{k=0}^{\infty} \delta_{n,k+1}T^k \\
&= T^{n-1}
\end{aligned}$$

By applying the norm we obtain from this formula:

$$\begin{aligned} ||T^{n-1}|| &\leq \frac{1}{2\pi}\int_{|z|=\rho}\left|\left|\frac{z^n}{z-T}\right|\right| dz \\ &\leq \rho^n\cdot\sup_{|z|=\rho}\left|\left|\frac{1}{z-T}\right|\right| \end{aligned}$$

Since the sup does not depend on n, by taking n-th roots, we obtain in the limit:

$$\rho\geq\lim_{n\to\infty}||T^n||^{1/n}$$

Now recall that ρ was by definition an arbitrary number satisfying $\rho>\rho(T)$. Thus, we have obtained the following estimate, valid for any $T\in B(H)$:

$$\rho(T)\geq\lim_{n\to\infty}||T^n||^{1/n}$$

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

In the case of the normal elements, we have the following finer result:

THEOREM 8.14. *The spectral radius of a normal element,*

$$TT^*=T^*T$$

is equal to its norm.

PROOF. We can proceed in two steps, as follows:

Step 1. In the case $T=T^*$ we have $||T^n||=||T||^n$ for any exponent of the form $n=2^k$, by using the formula $||TT^*||=||T||^2$, and by taking n-th roots we get:

$$\rho(T)\geq||T||$$

Thus, we are done with the self-adjoint case, with the result $\rho(T)=||T||$.

Step 2. In the general normal case $TT^*=T^*T$ we have $T^n(T^n)^*=(TT^*)^n$, and by using this, along with the result from Step 1, applied to TT^*, we obtain:

$$\begin{aligned} \rho(T) &= \lim_{n\to\infty}||T^n||^{1/n} \\ &= \sqrt{\lim_{n\to\infty}||T^n(T^n)^*||^{1/n}} \\ &= \sqrt{\lim_{n\to\infty}||(TT^*)^n||^{1/n}} \\ &= \sqrt{\rho(TT^*)} \\ &= \sqrt{||T||^2} \\ &= ||T|| \end{aligned}$$

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

8c. Normal operators

By using Theorem 8.14 we can say a number of non-trivial things about the normal operators, commonly known as "spectral theorem for normal operators". We first have:

THEOREM 8.15. *Given $T \in B(H)$ normal, we have a morphism of algebras*

$$\mathbb{C}[X] \to B(H) \quad , \quad P \to P(T)$$

having the properties $||P(T)|| = ||P_{|\sigma(T)}||$, and $\sigma(P(T)) = P(\sigma(T))$.

PROOF. This is an improvement of Theorem 8.10 for polynomials, in the normal case, with the extra assertion being the norm estimate. But the element $P(T)$ being normal, we can apply to it the spectral radius formula for normal elements, and we obtain:

$$\begin{aligned} ||P(T)|| &= \rho(P(T)) \\ &= \sup_{\lambda \in \sigma(P(T))} |\lambda| \\ &= \sup_{\lambda \in P(\sigma(T))} |\lambda| \\ &= ||P_{|\sigma(T)}|| \end{aligned}$$

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

At a more advanced level now, we have the following result:

THEOREM 8.16. *Given $T \in B(H)$ normal, we have a morphism of algebras*

$$C(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

which is isometric, $||f(T)|| = ||f||$, and has the property $\sigma(f(T)) = f(\sigma(T))$.

PROOF. The idea here is to "complete" the morphism in Theorem 8.15. Indeed, by Stone-Weierstrass, that morphism has a unique isometric extension, as follows:

$$C(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

It remains to prove $\sigma(f(T)) = f(\sigma(T))$, and we can do this by double inclusion:

"$\subset$" Given a continuous function $f \in C(\sigma(T))$, we must prove that we have:

$$\lambda \notin f(\sigma(T)) \implies \lambda \notin \sigma(f(T))$$

For this purpose, consider the following function, which is well-defined:

$$\frac{1}{f - \lambda} \in C(\sigma(T))$$

We can therefore apply this function to T, and we obtain:

$$\left(\frac{1}{f - \lambda}\right) T = \frac{1}{f(T) - \lambda}$$

In particular $f(T) - \lambda$ is invertible, so $\lambda \notin \sigma(f(T))$, as desired.

"$\supset$" Given a continuous function $f \in C(\sigma(T))$, we must prove that we have:

$$\lambda \in f(\sigma(T)) \implies \lambda \in \sigma(f(T))$$

But this is the same as proving that we have:

$$\mu \in \sigma(T) \implies f(\mu) \in \sigma(f(T))$$

For this purpose, we approximate our function by polynomials, $P_n \to f$, and we examine the following convergence, which follows from $P_n \to f$:

$$P_n(T) - P_n(\mu) \to f(T) - f(\mu)$$

We know from polynomial functional calculus that we have:

$$P_n(\mu) \in P_n(\sigma(T)) = \sigma(P_n(T))$$

Thus, the operators $P_n(T) - P_n(\mu)$ are not invertible. On the other hand, we know that the set formed by the invertible operators is open, so its complement is closed. Thus the limit $f(T) - f(\mu)$ is not invertible either, and so $f(\mu) \in \sigma(f(T))$, as desired. □

Even more generally now, we have the following result:

THEOREM 8.17. *Given $T \in B(H)$ normal, we have a morphism of algebras as follows, with L^∞ standing for abstract measurable functions, or Borel functions,*

$$L^\infty(\sigma(T)) \to B(H) \quad , \quad f \to f(T)$$

which is isometric, $||f(T)|| = ||f||$, and has the property $\sigma(f(T)) = f(\sigma(T))$.

PROOF. As before, the idea will be that of "completing" what we have:

(1) Given a vector $x \in H$, consider the following functional:

$$C(\sigma(T)) \to \mathbb{C} \quad , \quad g \to < g(T)x, x >$$

By the Riesz theorem, this functional must be the integration with respect to a certain measure μ on the space $\sigma(T)$. Thus, we have a formula as follows:

$$< g(T)x, x >= \int_{\sigma(T)} g(z)d\mu(z)$$

Now given an arbitrary Borel function $f \in L^\infty(\sigma(T))$, as in the statement, we can define a number $< f(T)x, x > \in \mathbb{C}$, by using exactly the same formula, namely:

$$< f(T)x, x >= \int_{\sigma(T)} f(z)d\mu(z)$$

Thus, we have managed to define numbers $< f(T)x, x > \in \mathbb{C}$, for all vectors $x \in H$, and in addition we can recover these numbers as follows, with $g_n \in C(\sigma(T))$:

$$< f(T)x, x >= \lim_{g_n \to f} < g_n(T)x, x >$$

(2) In order to define now numbers $< f(T)x, y > \in \mathbb{C}$, for all vectors $x, y \in H$, we can use a polarization trick. Indeed, for any operator $S \in B(H)$ we have:

$$< S(x+y), x+y > = < Sx, x > + < Sy, y > + < Sx, y > + < Sy, x >$$

By replacing $y \to iy$, we have as well the following formula:

$$< S(x+iy), x+iy > = < Sx, x > + < Sy, y > -i < Sx, y > +i < Sy, x >$$

By multiplying this formula by i, and summing with the first one, we obtain:

$$\begin{aligned} < S(x+y), x+y > +i < S(x+iy), x+iy > &= (1+i)[< Sx, x > + < Sy, y >] \\ &+ 2 < Sx, y > \end{aligned}$$

(3) But with this, we can finish. Indeed, by combining (1,2), given a Borel function $f \in L^\infty(\sigma(T))$, we can define numbers $< f(T)x, y > \in \mathbb{C}$ for any $x, y \in H$, and we obtain in this way a certain operator $f(T) \in B(H)$, having all the desired properties. □

Good news, we can now diagonalize the normal operators. Let us start with:

THEOREM 8.18. *Any self-adjoint operator $T \in B(H)$ can be diagonalized,*

$$T = U^* M_f U$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to T, with $f : X \to \mathbb{R}$ being a certain function, once again associated to T, and with

$$M_f(g) = fg$$

being the usual multiplication operator by f, on the Hilbert space $L^2(X)$.

PROOF. The construction of U, f can be done in several steps, as follows:

(1) We first prove the result in the special case where our operator T has a cyclic vector $x \in H$, with this meaning that the following holds:

$$\overline{span\left(T^k x \middle| n \in \mathbb{N}\right)} = H$$

For this purpose, let us go back to the proof of Theorem 8.17. We will use the following formula from there, with μ being the measure on $X = \sigma(T)$ associated to x:

$$< g(T)x, x > = \int_{\sigma(T)} g(z) d\mu(z)$$

Our claim is that we can define a unitary $U : H \to L^2(X)$, first on the dense part spanned by the vectors $T^k x$, by the following formula, and then by continuity:

$$U[g(T)x] = g$$

Indeed, the following computation shows that U is well-defined, and isometric:

$$\begin{aligned} ||g(T)x||^2 &= < g(T)x, g(T)x > \\ &= < g(T)^* g(T)x, x > \\ &= < |g|^2(T)x, x > \\ &= \int_{\sigma(T)} |g(z)|^2 d\mu(z) \\ &= ||g||_2^2 \end{aligned}$$

We can then extend U by continuity into a unitary $U : H \to L^2(X)$, as claimed. Now observe that we have the following formula:

$$\begin{aligned} UTU^* g &= U[Tg(T)x] \\ &= U[(zg)(T)x] \\ &= zg \end{aligned}$$

Thus our result is proved in the present case, with U as above, and with $f(z) = z$.

(2) We discuss now the general case. Our first claim is that H has a decomposition as follows, with each H_i being invariant under T, and admitting a cyclic vector x_i:

$$H = \bigoplus_i H_i$$

Indeed, this is something elementary, the construction being by recurrence in finite dimensions, in the obvious way, and by using the Zorn lemma in general. Now with this decomposition in hand, we can make a direct sum of the diagonalizations obtained in (1), for each of the restrictions $T_{|H_i}$, and we obtain the formula in the statement. □

We have the following technical generalization of the above result:

THEOREM 8.19. *Any family of commuting self-adjoint operators $T_i \in B(H)$ can be jointly diagonalized,*

$$T_i = U^* M_{f_i} U$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to $\{T_i\}$, with $f_i : X \to \mathbb{R}$ being certain functions, once again associated to T_i, and with

$$M_{f_i}(g) = f_i g$$

being the usual multiplication operator by f_i, on the Hilbert space $L^2(X)$.

PROOF. This is similar to the proof of Theorem 8.18, by suitably modifying the measurable calculus formula, and the measure μ itself, as to have this formula working for all the operators T_i. With this modification done, everything extends. □

We can now discuss the case of the arbitrary normal operators, as follows:

THEOREM 8.20. *Any normal operator $T \in B(H)$ can be diagonalized,*

$$T = U^* M_f U$$

with $U : H \to L^2(X)$ being a unitary operator from H to a certain L^2 space associated to T, with $f : X \to \mathbb{C}$ being a certain function, once again associated to T, and with

$$M_f(g) = fg$$

being the usual multiplication operator by f, on the Hilbert space $L^2(X)$.

PROOF. Consider the decomposition of T into its real and imaginary parts:

$$T = \frac{T + T^*}{2} + i \cdot \frac{T - T^*}{2i}$$

We know that the real and imaginary parts are self-adjoint operators. Now since T was assumed to be normal, $TT^* = T^*T$, these real and imaginary parts commute:

$$\left[\frac{T + T^*}{2} \, , \, \frac{T - T^*}{2i} \right] = 0$$

Thus Theorem 8.19 applies to these real and imaginary parts, and gives the result. □

Getting now to applications, the above results are quite powerful, and many things can be said, in analogy with what we know about usual matrices. Let us record here:

THEOREM 8.21. *Any bounded operator $T \in B(H)$ can be decomposed as*

$$T = U|T|$$

with U being a partial isometry, and with $|T| = \sqrt{T^*T}$.

PROOF. The operator T^*T being self-adjoint, and even positive, in the sense that we have $< T^*Tx, x > \geq 0$ for any $x \in H$, we can extract its square root $|T| = \sqrt{T^*T}$, by using the continuous functional calculus. Now observe that we have the following formula:

$$\begin{aligned} < |T|x, |T|y > &= < x, |T|^2 y > \\ &= < x, T^*Ty > \\ &= < Tx, Ty > \end{aligned}$$

We conclude that the following linear application is well-defined, and isometric:

$$U : Im|T| \to Im(T) \quad , \quad |T|x \to Tx$$

Now by continuity we can extend this isometry U into an isometry between certain Hilbert subspaces of H, as follows:

$$U : \overline{Im|T|} \to \overline{Im(T)} \quad , \quad |T|x \to Tx$$

Moreover, we can further extend U into a partial isometry $U : H \to H$, by setting $Ux = 0$, for any $x \in \overline{Im|T|}^{\perp}$, and with this convention, the result follows. □

8d. Compact operators

We restrict now the attention to the compact operators, which share many properties with the usual matrices. Let us start with a basic definition, as follows:

DEFINITION 8.22. *An operator $T \in B(H)$ is said to be of finite rank if its image*

$$Im(T) \subset H$$

is finite dimensional. The set of such operators is denoted $F(H)$.

There are many interesting examples of finite rank operators, the most basic ones being the finite rank projections, on the finite dimensional subspaces $K \subset H$. We have:

PROPOSITION 8.23. *The set of finite rank operators*

$$F(H) \subset B(H)$$

is a two-sided $*$-ideal.

PROOF. It is clear that $F(H)$ is a vector space. Let us prove now that $F(H)$ is stable under $*$. Given $T \in F(H)$, we can regard it as an invertible operator, as follows:

$$T : (\ker T)^\perp \to Im(T)$$

We conclude that we have the following dimension equality:

$$\dim((\ker T)^\perp) = \dim(Im(T))$$

On the other hand, we have equalities as follows, which give the result:

$$\begin{aligned} \dim(Im(T^*)) &= \dim(\overline{Im(T^*)}) \\ &= \dim((\ker T)^\perp) \\ &= \dim(Im(T)) \end{aligned}$$

Regarding now the ideal property, this follows from the following two formulae, valid for any $S, T \in B(H)$, which are once again clear from definitions:

$$\dim(Im(ST)) \leq \dim(Im(T))$$

$$\dim(Im(TS)) \leq \dim(Im(T))$$

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

Let us discuss now the compact operators. These are introduced as follows:

DEFINITION 8.24. *An operator $T \in B(H)$ is said to be compact if the closed set*

$$\overline{T(B_1)} \subset H$$

is compact, where $B_1 \subset H$ is the unit ball. The set of such operators is denoted $K(H)$.

In finite dimensions any operator is compact. In general, as a first observation, any finite rank operator is compact. We have in fact the following result:

PROPOSITION 8.25. *Any finite rank operator is compact,*

$$F(H) \subset K(H)$$

and the finite rank operators are dense inside the compact operators.

PROOF. The first assertion is clear, because if $Im(T)$ is finite dimensional, then the following subset is closed and bounded, and so it is compact:

$$\overline{T(B_1)} \subset Im(T)$$

Regarding the second assertion, let us pick a compact operator $T \in K(H)$, and a number $\varepsilon > 0$. By compactness of T we can find a finite set $S \subset B_1$ such that:

$$T(B_1) \subset \bigcup_{x \in S} B_\varepsilon(Tx)$$

Consider now the orthogonal projection P onto the following finite dimensional space:

$$E = span\left(Tx \middle| x \in S\right)$$

Since the set S is finite, this space E is finite dimensional, and so P is of finite rank, $P \in F(H)$. Now observe that for any norm one $y \in H$ and any $x \in S$ we have:

$$\begin{aligned} ||Ty - Tx||^2 &= ||Ty - PTx||^2 \\ &= ||Ty - PTy + PTy - PTx||^2 \\ &= ||Ty - PTy||^2 + ||PTx - PTy||^2 \end{aligned}$$

Now by picking $x \in S$ such that the ball $B_\varepsilon(Tx)$ covers the point Ty, we conclude from this that we have the following estimate:

$$||Ty - PTy|| \leq ||Ty - Tx|| \leq \varepsilon$$

Thus we have $||T - PT|| \leq \varepsilon$, which gives the density result. □

Quite remarkably, the set of compact operators is closed, and we have:

THEOREM 8.26. *The set of compact operators*

$$K(H) \subset B(H)$$

is a closed two-sided $*$-ideal.

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

(1) It is clear that $K(H)$ is a vector space. In order to prove now that $K(H)$ is closed, assume that $T_n \in K(H)$ converges to $T \in B(H)$. Given $\varepsilon > 0$, pick $N \in \mathbb{N}$ such that:

$$||T - T_N|| \leq \varepsilon$$

By compactness of T_N we can find a finite set $S \subset B_1$ such that:

$$T_N(B_1) \subset \bigcup_{x \in S} B_\varepsilon(T_N x)$$

We conclude that for any $y \in B_1$ there exists $x \in S$ such that:

$$\begin{aligned} ||Ty - Tx|| &\leq ||Ty - T_N y|| + ||T_N y - T_N x|| + ||T_N x - Tx|| \\ &\leq \varepsilon + \varepsilon + \varepsilon \\ &= 3\varepsilon \end{aligned}$$

Thus, we have an inclusion as follows, showing that T is indeed compact:

$$T(B_1) \subset \bigcup_{x \in S} B_{3\varepsilon}(Tx)$$

(2) Regarding now the fact that $K(H)$ is stable under involution, this follows from Proposition 8.23, Proposition 8.25 and (1). Indeed, by using Proposition 8.25, given $T \in K(H)$ we can write it as a limit of finite rank operators, as follows:

$$T = \lim_{n \to \infty} T_n$$

Now by applying the adjoint, we have as well $T^* = \lim_{n \to \infty} T_n^*$. We know from Proposition 8.23 that the operators T_n^* are of finite rank, and so compact by Proposition 8.25, and by using (1) we obtain that T^* is compact too, as desired.

(3) Finally, regarding the ideal property of $K(H)$, this is clear from definitions. □

Here is now a second key result regarding the compact operators:

THEOREM 8.27. *A bounded operator $T \in B(H)$ is compact precisely when*

$$Te_n \to 0$$

for any orthonormal system $\{e_n\} \subset H$.

PROOF. We have two implications to be proved, the idea being as follows:

" $\implies$ " Assume that T is compact. By contradiction, assume $Te_n \not\to 0$. This means that there exists $\varepsilon > 0$ and a subsequence satisfying $||Te_{n_k}|| > \varepsilon$, and by replacing $\{e_n\}$ with this subsequence, we can assume that the following holds, with $\varepsilon > 0$:

$$||Te_n|| > \varepsilon$$

Since T is compact, a certain subsequence $\{Te_{n_k}\}$ must converge. Thus, by replacing once again $\{e_n\}$ with a subsequence, we can assume that we have, with $x \neq 0$:

$$Te_n \to x$$

But this is a contradiction, as desired, because we obtain in this way:

$$\begin{aligned} < x, x > &= \lim_{n \to \infty} < Te_n, x > \\ &= \lim_{n \to \infty} < e_n, T^* x > \\ &= 0 \end{aligned}$$

"$\Longleftarrow$" Assume $Te_n \to 0$, for any orthonormal system $\{e_n\} \subset H$. In order to prove $T \in K(H)$, we will prove that T is in the closure of the space of finite rank operators:

$$T \in \overline{F(H)}$$

We do this by contradiction. So, assume that there exists $\varepsilon > 0$ such that:

$$S \in F(H) \implies ||T - S|| > \varepsilon$$

As a first observation, by using $S = 0$ we obtain $||T|| > \varepsilon$. Thus, we can find a norm one vector $e_1 \in H$ such that the following holds:

$$||Te_1|| > \varepsilon$$

Our claim, which will bring the desired contradiction, is that we can construct by recurrence vectors $e_1, \ldots, e_n$ such that the following holds, for any i:

$$||Te_i|| > \varepsilon$$

Indeed, assume that we have constructed $e_1, \ldots, e_n$. Let $E \subset H$ be the linear space spanned by these vectors, and set $P = Proj(E)$. Since TP has finite rank, our assumption above shows that we have $||T - TP|| > \varepsilon$. Thus, we can find $x \in H$ such that:

$$||(T - TP)x|| > \varepsilon$$

We have then $x \notin E$, and so we can consider the following nonzero vector:

$$y = (1 - P)x$$

With this nonzero vector y constructed, now let us set:

$$e_{n+1} = \frac{y}{||y||}$$

This vector e_{n+1} is then orthogonal to E, has norm one, and satisfies:

$$||Te_{n+1}|| \geq ||y||^{-1}\varepsilon \geq \varepsilon$$

Thus we are done with our construction by recurrence, and this contradicts our assumption that $Te_n \to 0$, for any orthonormal system $\{e_n\} \subset H$, as desired. □

Let us discuss now the spectral theory of the compact operators. We first have:

PROPOSITION 8.28. *Assuming that $T \in B(H)$, with $\dim H = \infty$, is compact and self-adjoint, the following happen:*

(1) *The eigenvalues of T form a sequence $\lambda_n \to 0$.*
(2) *All eigenvalues $\lambda_n \neq 0$ have finite multiplicity.*

PROOF. We prove both the assertions at the same time. For this purpose, we fix a number $\varepsilon > 0$, we consider all the eigenvalues satisfying $|\lambda| \geq \varepsilon$, and for each such eigenvalue we consider the corresponding eigenspace $E_\lambda \subset H$. Let us set:

$$E = span\left(E_\lambda \,\middle|\, |\lambda| \geq \varepsilon\right)$$

Our claim, which will prove both (1) and (2), is that this space E is finite dimensional. In now to prove now this claim, we can proceed as follows:

(1) We know that we have $E \subset Im(T)$. Our claim is that we have:

$$\bar{E} \subset Im(T)$$

Indeed, assume that we have a sequence $g_n \in E$ which converges, $g_n \to g \in \bar{E}$. Let us write $g_n = Tf_n$, with $f_n \in H$. By definition of E, the following condition is satisfied:

$$h \in E \implies ||Th|| \geq \varepsilon ||h||$$

Now since the sequence $\{g_n\}$ is Cauchy we obtain from this that the sequence $\{f_n\}$ is Cauchy as well, and with $f_n \to f$ we have $Tf_n \to Tf$, as desired.

(2) Consider now the projection $P \in B(H)$ onto the above space $\bar{E}$. The composition PT is then as follows, surjective on its target:

$$PT : H \to \bar{E}$$

On the other hand since T is compact so must be PT, and if follows from this that the space $\bar{E}$ is finite dimensional. Thus E itself must be finite dimensional too, and as explained in the beginning of the proof, this gives (1) and (2), as desired. □

In order to construct now eigenvalues, we will need:

PROPOSITION 8.29. *If T is compact and self-adjoint, one of the numbers*

$$||T|| \ , \ -||T||$$

must be an eigenvalue of T.

PROOF. We know from the spectral theory of the self-adjoint operators that the spectral radius $||T||$ of our operator T is attained, and so one of the numbers $||T||, -||T||$ must be in the spectrum. In order to prove now that one of these numbers must actually appear as an eigenvalue, we must use the compactness of T, as follows:

(1) First, we can assume $||T|| = 1$. By functional calculus this implies $||T^3|| = 1$ too, and so we can find a sequence of norm one vectors $x_n \in H$ such that:

$$| < T^3 x_n, x_n > | \to 1$$

By using our assumption $T = T^*$, we can rewrite this formula as follows:

$$| < T^2 x_n, Tx_n > | \to 1$$

Now since T is compact, and $\{x_n\}$ is bounded, we can assume, up to changing the sequence $\{x_n\}$ to one of its subsequences, that the sequence Tx_n converges:

$$Tx_n \to y$$

Thus, the convergence formula found above reformulates as follows, with $y \neq 0$:

$$| < Ty, y > | = 1$$

(2) Our claim now, which will finish the proof, is that this latter formula implies $Ty = \pm y$. Indeed, by using Cauchy-Schwarz and $||T|| = 1$, we have:

$$| < Ty, y > | \leq ||Ty|| \cdot ||y|| \leq 1$$

We know that this must be an equality, so Ty, y must be proportional. But since T is self-adjoint the proportionality factor must be ± 1, and so we obtain, as claimed:

$$Ty = \pm y$$

Thus, we have constructed an eigenvector for $\lambda = \pm 1$, as desired. □

We can further build on the above results in the following way:

PROPOSITION 8.30. *If T is compact and self-adjoint, there is an orthogonal basis of H made of eigenvectors of T.*

PROOF. We use Proposition 8.28. According to the results there, we can arrange the nonzero eigenvalues of T, taken with multiplicities, into a sequence $\lambda_n \to 0$. Let $y_n \in H$ be the corresponding eigenvectors, and consider the following space:

$$E = \overline{span(y_n)}$$

The result follows then from the following observations:

(1) Since we have $T = T^*$, both E and its orthogonal $E^\perp$ are invariant under T.

(2) On the space E, our operator T is by definition diagonal.

(3) On the space $E^\perp$, our claim is that we have $T = 0$. Indeed, assuming that the restriction $S = T_{E^\perp}$ is nonzero, we can apply Proposition 8.29 to this restriction, and we obtain an eigenvalue for S, and so for T, contradicting the maximality of E. □

With the above results in hand, we can now formulate a first theorem, as follows:

THEOREM 8.31. *Assuming that $T \in B(H)$, with $\dim H = \infty$, is compact and self-adjoint, the following happen:*

(1) *The spectrum $\sigma(T) \subset \mathbb{R}$ consists of a sequence $\lambda_n \to 0$.*
(2) *All spectral values $\lambda \in \sigma(T) - \{0\}$ are eigenvalues.*
(3) *All eigenvalues $\lambda \in \sigma(T) - \{0\}$ have finite multiplicity.*
(4) *There is an orthogonal basis of H made of eigenvectors of T.*

PROOF. This follows from the various results established above:

(1) In view of Proposition 8.28 (1), this will follow from (2) below.

(2) Assume that $\lambda \neq 0$ belongs to the spectrum $\sigma(T)$, but is not an eigenvalue. By using Proposition 8.30, let us pick an orthonormal basis $\{e_n\}$ of H consisting of eigenvectors of T, and then consider the following operator:

$$Sx = \sum_n \frac{< x, e_n >}{\lambda_n - \lambda} e_n$$

Then S is an inverse for $T-\lambda$, and so we have $\lambda \notin \sigma(T)$, as desired.

(3) This is something that we know, from Proposition 8.28 (2).

(4) This is something that we know too, from Proposition 8.30. □

Finally, we have the following result, regarding the general case:

THEOREM 8.32. *The compact operators $T \in B(H)$, with $\dim H = \infty$, are the operators of the following form, with $\{e_n\}$, $\{f_n\}$ being orthonormal families, and with $\lambda_n \searrow 0$:*

$$T(x) = \sum_n \lambda_n < x, e_n > f_n$$

The numbers λ_n, called singular values of T, are the eigenvalues of $|T|$. In fact, the polar decomposition of T is given by $T = U|T|$, with

$$|T|(x) = \sum_n \lambda_n < x, e_n > e_n$$

and with U being given by $Ue_n = f_n$, and $U = 0$ on the complement of $span(e_i)$.

PROOF. This basically follows from Theorem 8.31, as follows:

(1) Given two orthonormal families $\{e_n\}$, $\{f_n\}$, and a sequence of real numbers $\lambda_n \searrow 0$, consider the linear operator given by the formula in the statement, namely:

$$T(x) = \sum_n \lambda_n < x, e_n > f_n$$

Our first claim is that T is bounded. Indeed, when assuming $|\lambda_n| \leq \varepsilon$ for any n, which is something that we can do if we want to prove that T is bounded, we have:

$$\begin{aligned} ||T(x)||^2 &= \left|\sum_n \lambda_n < x, e_n > f_n\right|^2 \\ &= \sum_n |\lambda_n|^2 | < x, e_n > |^2 \\ &\leq \varepsilon^2 \sum_n | < x, e_n > |^2 \\ &\leq \varepsilon^2 ||x||^2 \end{aligned}$$

(2) The next observation is that this operator is indeed compact, because it appears as the norm limit, $T_N \to T$, of the following sequence of finite rank operators:

$$T_N = \sum_{n \leq N} \lambda_n < x, e_n > f_n$$

(3) Regarding now the polar decomposition assertion, for the above operator, this follows once again from definitions. Indeed, the adjoint is given by:

$$T^*(x) = \sum_n \lambda_n < x, f_n > e_n$$

Thus, when composing T^* with T, we obtain the following operator:

$$T^*T(x) = \sum_n \lambda_n^2 < x, e_n > e_n$$

Now by extracting the square root, we obtain the formula in the statement, namely:

$$|T|(x) = \sum_n \lambda_n < x, e_n > e_n$$

(4) Conversely now, assume that $T \in B(H)$ is compact. Then T^*T, which is self-adjoint, must be compact as well, and so by Theorem 8.31 we have a formula as follows, with $\{e_n\}$ being a certain orthonormal family, and with $\lambda_n \searrow 0$:

$$T^*T(x) = \sum_n \lambda_n^2 < x, e_n > e_n$$

By extracting the square root we obtain the formula of $|T|$ in the statement, and then by setting $U(e_n) = f_n$ we obtain a second orthonormal family, $\{f_n\}$, such that:

$$T(x) = U|T| = \sum_n \lambda_n < x, e_n > f_n$$

Thus, our compact operator $T \in B(H)$ appears indeed as in the statement. □

8e. Exercises

This was a quite straightforward chapter, and as exercises, we have:

EXERCISE 8.33. *In relation with separability, learn about orthogonal polynomials.*

EXERCISE 8.34. *Clarify the details in the proof of the measurable calculus theorem.*

EXERCISE 8.35. *Jointly diagonalize the families of commuting normal operators.*

EXERCISE 8.36. *Check the positivity details, in the proof of the polar decomposition.*

EXERCISE 8.37. *Learn as well about the strictly positive operators, $T > 0$.*

EXERCISE 8.38. *Work out other decomposition results, for the linear operators.*

EXERCISE 8.39. *Learn about the trace class operators, and their properties.*

EXERCISE 8.40. *Learn as well about Hilbert-Schmidt operators, and their properties.*

As bonus exercise, in addition to this, learn some operator algebras as well.

Part III

Positive matrices

She'll carry on through it all
She's a waterfall
She'll carry on through it all
She's a waterfall

CHAPTER 9

Hessian matrices

9a. Calculus, Jacobian

Welcome to positivity. We have certainly met positive matrices in the above, but that material was quite theoretical, and time now to have a closer look at positivity, which is what makes this world go round, via various subtle mathematical mechanisms.

We will start our discussion with a review of multivariable calculus, where the positivity properties of the various matrices appearing there, such as the Jacobian or the Hessian ones, is of key importance. Let us first recall the basics. We first have:

THEOREM 9.1. *The functions $f:\mathbb{R}^N\to\mathbb{R}^M$ can be locally approximated as*

$$f(x+t)\simeq f(x)+f'(x)t$$

with $f'(x)$ being by definition the matrix of partial derivatives at x,

$$f'(x)=\left(\frac{df_i}{dx_j}(x)\right)_{ij}\in M_{M\times N}(\mathbb{R})$$

acting on the vectors $t\in\mathbb{R}^N$ by usual multiplication.

PROOF. This is obviously something a bit informal, because the precise assumptions needed on f are not mentioned. More on this in a moment, and in the meantime:

(1) First of all, at $N=M=1$ what we have is a usual 1-variable function $f:\mathbb{R}\to\mathbb{R}$, and the formula in the statement is something that we know well, namely:

$$f(x+t)\simeq f(x)+f'(x)t$$

(2) Let us discuss now the case $N=2,M=1$. Here what we have is a function $f:\mathbb{R}^2\to\mathbb{R}$, and by using twice the basic approximation result from (1), we obtain:

$$\begin{aligned}f\begin{pmatrix}x_1+t_1\\x_2+t_2\end{pmatrix}&\simeq f\begin{pmatrix}x_1+t_1\\x_2\end{pmatrix}+\frac{df}{dx_2}(x)t_2\\&\simeq f\begin{pmatrix}x_1\\x_2\end{pmatrix}+\frac{df}{dx_1}(x)t_1+\frac{df}{dx_2}(x)t_2\\&=f\begin{pmatrix}x_1\\x_2\end{pmatrix}+\begin{pmatrix}\frac{df}{dx_1}(x)&\frac{df}{dx_2}(x)\end{pmatrix}\begin{pmatrix}t_1\\t_2\end{pmatrix}\end{aligned}$$

(3) More generally, we can deal in this way with the general case $M = 1$, with the formula here, obtained via a straightforward recurrence, being as follows:

$$\begin{aligned} f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} &\simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \frac{df}{dx_1}(x)t_1 + \ldots + \frac{df}{dx_N}(x)t_N \\ &= f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df}{dx_1}(x) & \ldots & \frac{df}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \end{aligned}$$

(4) But this gives the result in the case where both $N, M \in \mathbb{N}$ are arbitrary too. Indeed, we can apply (3) to each of the components $f_i : \mathbb{R}^N \to \mathbb{R}$, and we get:

$$f_i\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} \simeq f_i\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_i}{dx_1}(x) & \ldots & \frac{df_i}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

But this collection of M formulae tells us precisely that the following happens, as an equality, or rather approximation, of vectors in $\mathbb{R}^M$:

$$f\begin{pmatrix} x_1 + t_1 \\ \vdots \\ x_N + t_N \end{pmatrix} \simeq f\begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} + \begin{pmatrix} \frac{df_1}{dx_1}(x) & \ldots & \frac{df_1}{dx_N}(x) \\ \vdots & & \vdots \\ \frac{df_M}{dx_1}(x) & \ldots & \frac{df_M}{dx_N}(x) \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$$

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

At a more advanced level now, fully rigorous, we have the following result:

THEOREM 9.2. *For a function $f : X \to \mathbb{R}^M$, with $X \subset \mathbb{R}^N$, the following conditions are equivalent, and in this case we say that f is continuously differentiable:*

(1) *f is differentiable, and the map $x \to f'(x)$ is continuous.*
(2) *f has partial derivatives, which are continuous with respect to $x \in X$.*

If these conditions are satisfied, $f'(x)$ is the matrix fomed by the partial derivatives at x.

PROOF. We already know, from Theorem 9.1, that the last assertion holds. Regarding now the proof of the equivalence, this goes as follows:

(1) $\implies$ (2) Assuming that f is differentiable, we know from Theorem 9.1 that $f'(x)$ is the matrix fomed by the partial derivatives at x. Thus, for any $x, y \in X$:

$$\frac{df_i}{dx_j}(x) - \frac{df_i}{dx_j}(y) = f'(x)_{ij} - f'(y)_{ij}$$

By applying now the absolute value, we obtain from this the following estimate:

$$\begin{aligned}\left|\frac{df_i}{dx_j}(x)-\frac{df_i}{dx_j}(y)\right| &= |f'(x)_{ij}-f'(y)_{ij}|\\ &= |(f'(x)-f'(y))_{ij}|\\ &\leq ||f'(x)-f'(y)||\end{aligned}$$

But this gives the result, because if the map $x\to f'(x)$ is assumed to be continuous, then the partial derivatives follow to be continuous with respect to $x\in X$.

(2) $\implies$ (1) This is something more technical. For simplicity, let us assume $M=1$, the proof in general being similar. Given $x\in X$ and $\varepsilon>0$, let us pick $r>0$ such that the ball $B=B_x(r)$ belongs to X, and such that the following happens, over B:

$$\left|\frac{df}{dx_j}(x)-\frac{df}{dx_j}(y)\right|<\frac{\varepsilon}{N}$$

Our claim is that, with this choice made, we have the following estimate, for any $t\in\mathbb{R}^N$ satisfying $||t||<r$, with A being the vector of partial derivatives at x:

$$|f(x+t)-f(x)-At|\leq\varepsilon||t||$$

In order to prove this claim, the idea will be that of suitably applying the mean value theorem, over the N directions of $\mathbb{R}^N$. Indeed, consider the following vectors:

$$t^{(k)}=\begin{pmatrix}t_1\\ \vdots\\ t_k\\ 0\\ \vdots\\ 0\end{pmatrix}$$

In terms of these vectors, we have the following formula:

$$f(x+t)-f(x)=\sum_{j=1}^{N}f(x+t^{(j)})-f(x+t^{(j-1)})$$

Also, the mean value theorem gives a formula as follows, with $s_j\in[0,1]$:

$$f(x+t^{(j)})-f(x+t^{(j-1)})=\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})\cdot t_j$$

But, according to our assumption on $r>0$ from the beginning, the derivative on the right differs from $\frac{df}{dx_j}(x)$ by something which is smaller than ε/N:

$$\left|\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})-\frac{df}{dx_j}(x)\right|<\frac{\varepsilon}{N}$$

Now by putting everything together, we obtain the following estimate:

$$\begin{aligned}
|f(x+t)-f(x)-At| &= \left|\sum_{j=1}^N f(x+t^{(j)})-f(x+t^{(j-1)})-\frac{df}{dx_j}(x)\cdot t_j\right| \\
&\leq \sum_{j=1}^N \left|f(x+t^{(j)})-f(x+t^{(j-1)})-\frac{df}{dx_j}(x)\cdot t_j\right| \\
&= \sum_{j=1}^N \left|\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})\cdot t_j-\frac{df}{dx_j}(x)\cdot t_j\right| \\
&= \sum_{j=1}^N \left|\frac{df}{dx_j}(x+s_jt^{(j)}+(1-s_j)t^{(j-1)})-\frac{df}{dx_j}(x)\right|\cdot|t_j| \\
&\leq \sum_{j=1}^N \frac{\varepsilon}{N}\cdot|t_j| \\
&\leq \varepsilon||t||
\end{aligned}$$

Thus we have proved our claim, and this gives the result. □

Generally speaking, Theorems 9.1 and 9.2 are all you need to know for upgrading from one variable calculus to multivariable calculus. As a standard result here, we have:

THEOREM 9.3. *We have the chain derivative formula*

$$(f\circ g)'(x)=f'(g(x))\cdot g'(x)$$

as an equality of matrices.

PROOF. Consider indeed a composition of functions, as follows:

$$f:\mathbb{R}^N\to\mathbb{R}^M\quad,\quad g:\mathbb{R}^K\to\mathbb{R}^N\quad,\quad f\circ g:\mathbb{R}^K\to\mathbb{R}^M$$

According to Theorem 9.1, the derivatives of these functions are certain linear maps, corresponding to certain rectangular matrices, as follows:

$$f'(g(x))\in M_{M\times N}(\mathbb{R})\quad,\quad g'(x)\in M_{N\times K}(\mathbb{R})\qquad(f\circ g)'(x)\in M_{M\times K}(\mathbb{R})$$

Thus, our formula makes sense indeed. As for proof, this comes from:

$$\begin{aligned}
(f\circ g)(x+t) &= f(g(x+t)) \\
&\simeq f(g(x)+g'(x)t) \\
&\simeq f(g(x))+f'(g(x))g'(x)t
\end{aligned}$$

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

Getting now to integration matters, as a key result here, we have:

THEOREM 9.4. *Given a transformation $\varphi = (\varphi_1, \ldots, \varphi_N)$, we have*

$$\int_E f(x)dx = \int_{\varphi^{-1}(E)} f(\varphi(t))|J_\varphi(t)|dt$$

with the J_φ quantity, called Jacobian, being given by

$$J_\varphi(t) = \det\left[\left(\frac{d\varphi_i}{dx_j}(x)\right)_{ij}\right]$$

and with this generalizing the usual formula from one variable calculus.

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

(1) Observe first that the above formula generalizes indeed the change of variable formula in 1 dimension, the point here being that the absolute value on the derivative appears as to compensate for the lack of explicit bounds for the integral.

(2) As a second observation, we can assume if we want, by linearity, that we are dealing with the constant function $f = 1$. For this function, our formula reads:

$$vol(E) = \int_{\varphi^{-1}(E)} |J_\varphi(t)|dt$$

In terms of $D = \varphi^{-1}(E)$, this amounts in proving that we have:

$$vol(\varphi(D)) = \int_D |J_\varphi(t)|dt$$

And here, as a first remark, our formula is clear for the linear maps φ, by using the definition of the determinant of real matrices, as a signed volume.

(3) However, the extension of this to the case of non-linear maps φ is something non-trivial, so we will not follow this path. In order to prove now the result, as stated, our first claim is that the validity of the theorem is stable under taking compositions of transformations φ. In order to prove this claim, consider a composition, as follows:

$$\varphi : E \to F \quad , \quad \psi : D \to E \quad , \quad \varphi \circ \psi : D \to F$$

Assuming that the theorem holds for φ, ψ, we deduce that we have, as desired:

$$\begin{aligned}
\int_F f(x)dx &= \int_E f(\varphi(s))|J_\varphi(s)|ds \\
&= \int_D f(\varphi \circ \psi(t))|J_\varphi(\psi(t))| \cdot |J_\psi(t)|dt \\
&= \int_D f(\varphi \circ \psi(t))|J_{\varphi\circ\psi}(t)|dt
\end{aligned}$$

(4) Next, as a key ingredient, let us examine the case where we are in $N = 2$ dimensions, and our transformation φ has one of the following special forms:

$$\varphi(x, y) = (\psi(x, y), y) \quad , \quad \varphi(x, y) = (x, \psi(x, y))$$

By symmetry, it is enough to deal with the first case. Here the Jacobian is $d\psi/dx$, and by replacing if needed $\psi \to -\psi$, we can assume that this Jacobian is positive, $d\psi/dx > 0$. Now by assuming as before that $D = \varphi^{-1}(E)$ is a rectangle, $D = [a, b] \times [c, d]$, we can prove our formula by using the change of variables in 1 dimension, as follows:

$$\begin{aligned}
\int_E f(s)ds &= \int_{\varphi(D)} f(x, y)dxdy \\
&= \int_c^d \int_{\psi(a,y)}^{\psi(b,y)} f(x, y)dxdy \\
&= \int_c^d \int_a^b f(\psi(x, y), y)\frac{d\psi}{dx}\, dxdy \\
&= \int_D f(\varphi(t))J_\varphi(t)dt
\end{aligned}$$

(5) But with this, we can now prove the theorem, in $N = 2$ dimensions. Indeed, given a transformation $\varphi = (\varphi_1, \varphi_2)$, consider the following two transformations:

$$\phi(x, y) = (\varphi_1(x, y), y) \quad , \quad \psi(x, y) = (x, \varphi_2 \circ \phi^{-1}(x, y))$$

We have then $\varphi = \psi \circ \phi$, and by using (4) for ψ, ϕ, which are of the special form there, and then (3) for composing, we conclude that the theorem holds indeed for φ, as desired. Thus, theorem proved in $N = 2$ dimensions, and the extension of the above proof to arbitrary N dimensions is straightforward, that we will leave here as an exercise. □

At the level of the main applications, in 2 dimensions, we have:

PROPOSITION 9.5. *We have polar coordinates in 2 dimensions,*

$$\begin{cases} x &= r\cos t \\ y &= r\sin t \end{cases}$$

the corresponding Jacobian being $J = r$.

PROOF. This is indeed elementary, with the Jacobian being as follows:

$$J = \begin{vmatrix} \cos t & -r\sin t \\ \sin t & r\cos t \end{vmatrix} = r$$

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

We can now compute the Gauss integral, which is the best calculus formula ever:

THEOREM 9.6. *We have the following formula,*

$$\int_{\mathbb{R}} e^{-x^2} dx = \sqrt{\pi}$$

called Gauss integral formula.

PROOF. Let I be the above integral. By using polar coordinates, we obtain:

$$\begin{aligned} I^2 &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r dr dt \\ &= 2\pi \int_0^{\infty} \left(-\frac{e^{-r^2}}{2}\right)' dr \\ &= \pi \end{aligned}$$

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

Moving now to 3 dimensions, we have here the following result:

PROPOSITION 9.7. *We have spherical coordinates in* 3 *dimensions,*

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

the corresponding Jacobian being $J(r,s,t) = r^2\sin s$.

PROOF. This is again elementary, with the Jacobian being given by:

$$\begin{aligned} & J(r,s,t) \\ = & \begin{vmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{vmatrix} \\ = & r^2\sin s\sin t \begin{vmatrix} \cos s & -r\sin s \\ \sin s\sin t & r\cos s\sin t \end{vmatrix} + r\sin s\cos t \begin{vmatrix} \cos s & -r\sin s \\ \sin s\cos t & r\cos s\cos t \end{vmatrix} \\ = & r\sin s\sin^2 t \begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} + r\sin s\cos^2 t \begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s(\sin^2 t + \cos^2 t) \begin{vmatrix} \cos s & -r\sin s \\ \sin s & r\cos s \end{vmatrix} \\ = & r\sin s \times 1 \times r \\ = & r^2\sin s \end{aligned}$$

Thus, we have indeed the formula in the statement. □

Let us work out now the general spherical coordinate formula, in arbitrary N dimensions. The formula here, which generalizes those at $N = 2, 3$, is as follows:

THEOREM 9.8. *We have spherical coordinates in N dimensions,*

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

the corresponding Jacobian being given by the following formula,

$$J(r,t) = r^{N-1}\sin^{N-2}t_1\sin^{N-3}t_2\ \ldots\ \sin^2 t_{N-3}\sin t_{N-2}$$

and with this generalizing the known formulae at $N = 2, 3$.

PROOF. As before, the fact that we have spherical coordinates is clear. Regarding now the Jacobian, also as before, by developing over the last column, we have:

$$\begin{aligned} J_N &= r\sin t_1\ldots\sin t_{N-2}\sin t_{N-1}\times\sin t_{N-1}J_{N-1} \\ &+ r\sin t_1\ldots\sin t_{N-2}\cos t_{N-1}\times\cos t_{N-1}J_{N-1} \\ &= r\sin t_1\ldots\sin t_{N-2}(\sin^2 t_{N-1}+\cos^2 t_{N-1})J_{N-1} \\ &= r\sin t_1\ldots\sin t_{N-2}J_{N-1} \end{aligned}$$

Thus, we obtain the formula in the statement, by recurrence. □

As an application, let us compute the volumes of spheres. We will need:

PROPOSITION 9.9. *We have the following formulae,*

$$\int_0^{\pi/2}\cos^p t\,dt = \int_0^{\pi/2}\sin^p t\,dt = \left(\frac{\pi}{2}\right)^{\varepsilon(p)}\frac{p!!}{(p+1)!!}$$

where $\varepsilon(p) = 1$ if p is even, and $\varepsilon(p) = 0$ if p is odd, and where

$$m!! = (m-1)(m-3)(m-5)\ldots$$

with the product ending at 2 if m is odd, and ending at 1 if m is even.

PROOF. Let us first compute the integral on the left in the statement:

$$I_p = \int_0^{\pi/2}\cos^p t\,dt$$

We do this by partial integration. We have the following formula:

$$\begin{aligned} (\cos^p t\sin t)' &= p\cos^{p-1}t(-\sin t)\sin t+\cos^p t\cos t \\ &= p\cos^{p+1}t - p\cos^{p-1}t+\cos^{p+1}t \\ &= (p+1)\cos^{p+1}t - p\cos^{p-1}t \end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain the following formula:

$$(p+1)I_{p+1} = pI_{p-1}$$

Thus we can compute I_p by recurrence, and we obtain:

$$\begin{aligned} I_p &= \frac{p-1}{p}\, I_{p-2} \\ &= \frac{p-1}{p} \cdot \frac{p-3}{p-2}\, I_{p-4} \\ &\vdots \\ &= \frac{p!!}{(p+1)!!}\, I_{1-\varepsilon(p)} \end{aligned}$$

But $I_0 = \frac{\pi}{2}$ and $I_1 = 1$, so we get the result. As for the second formula, this follows from the first one, with $t = \frac{\pi}{2} - s$. Thus, we have proved both formulae in the statement. □

We can now compute the volumes of the spheres, as follows:

THEOREM 9.10. *The volume of the unit sphere in $\mathbb{R}^N$ is given by*

$$V = \left(\frac{\pi}{2}\right)^{[N/2]} \frac{2^N}{(N+1)!!}$$

with our usual convention $N!! = (N-1)(N-3)(N-5)\ldots$

PROOF. Let us denote by S^+ the positive part of the unit sphere S, obtained by cutting this sphere in 2^N parts. At the level of volumes we have $V = 2^N V^+$, with:

$$\begin{aligned} V^+ &= \int_{B^+} 1 \\ &= \int_0^1 \int_0^{\pi/2} \ldots \int_0^{\pi/2} r^{N-1} \sin^{N-2} t_1 \ldots \sin t_{N-2}\, dr dt_1 \ldots dt_{N-1} \\ &= \int_0^1 r^{N-1}\, dr \int_0^{\pi/2} \sin^{N-2} t_1\, dt_1 \ldots \int_0^{\pi/2} \sin t_{N-2} dt_{N-2} \int_0^{\pi/2} 1 dt_{N-1} \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{(N-2)!!}{(N-1)!!} \cdot \frac{(N-3)!!}{(N-2)!!} \cdots \frac{2!!}{3!!} \cdot \frac{1!!}{2!!} \cdot 1 \\ &= \frac{1}{N} \times \left(\frac{\pi}{2}\right)^{[N/2]} \times \frac{1}{(N-1)!!} \\ &= \left(\frac{\pi}{2}\right)^{[N/2]} \frac{1}{(N+1)!!} \end{aligned}$$

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

As main particular cases of the above formula, we have:

THEOREM 9.11. *The volumes of the low-dimensional spheres are as follows:*

(1) *At $N = 1$, the length of the unit interval is $V = 2$.*
(2) *At $N = 2$, the area of the unit disk is $V = \pi$.*
(3) *At $N = 3$, the volume of the unit sphere is $V = \frac{4\pi}{3}$*
(4) *At $N = 4$, the volume of the corresponding unit sphere is $V = \frac{\pi^2}{2}$.*

PROOF. Some of these results are well-known, but we can obtain all of them as particular cases of the general formula in Theorem 9.10, as follows:

(1) At $N = 1$ we obtain $V = 1 \cdot \frac{2}{1} = 2$.

(2) At $N = 2$ we obtain $V = \frac{\pi}{2} \cdot \frac{4}{2} = \pi$.

(3) At $N = 3$ we obtain $V = \frac{\pi}{2} \cdot \frac{8}{3} = \frac{4\pi}{3}$.

(4) At $N = 4$ we obtain $V = \frac{\pi^2}{4} \cdot \frac{16}{8} = \frac{\pi^2}{2}$. □

There are many other applications of the above. Also, it is possible as well to compute arbitrary polynomial integrals over the spheres, by using the same method as in the proof of Theorem 9.10, coupled with a finer version of Proposition 9.9. More about this later.

9b. Higher derivatives

Regarding now the higher derivatives, the situation here is more complicated. As a first result on the subject, regarding the functions $f : \mathbb{R}^2 \to \mathbb{R}$, we have:

THEOREM 9.12. *The double derivatives satisfy the formula*

$$\frac{d^2 f}{dxdy} = \frac{d^2 f}{dydx}$$

called Clairaut formula.

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

(1) Before pulling out a formal proof, as an intuitive justification for our formula, let us consider a product of power functions, $f(z) = x^p y^q$. We have then:

$$\frac{d^2 f}{dxdy} = \frac{d}{dx}\left(\frac{dx^p y^q}{dy}\right) = \frac{d}{dx}\left(qx^p y^{q-1}\right) = pqx^{p-1}y^{q-1}$$

$$\frac{d^2 f}{dydx} = \frac{d}{dy}\left(\frac{dx^p y^q}{dx}\right) = \frac{d}{dy}\left(px^{p-1} y^q\right) = pqx^{p-1}y^{q-1}$$

Next, let us consider a linear combination of power functions, $f(z) = \sum_{pq} c_{pq} x^p y^q$, which can be finite or not. We have then, by using the above computation:

$$\frac{d^2 f}{dxdy} = \frac{d^2 f}{dydx} = \sum_{pq} c_{pq} pq x^{p-1} y^{q-1}$$

Thus, we can see that our commutation formula for derivatives holds indeed, and this due to the fact that the functions in x, and in y, commute. Of course, all this does not prove our formula, in general. But exercise for you, to have this idea fully working.

(2) Getting now to more standard techniques, given a point in the complex plane, $z = a + ib$, consider the following functions, depending on $h, k \in \mathbb{R}$ small:

$$u(h,k) = f(a+h,b+k) - f(a+h,b)$$

$$v(h,k) = f(a+h,b+k) - f(a,b+k)$$

$$w(h,k) = f(a+h,b+k) - f(a+h,b) - f(a,b+k) + f(a,b)$$

By the mean value theorem, for $h, k \neq 0$ we can find $\alpha, \beta \in \mathbb{R}$ such that:

$$\begin{aligned} w(h,k) &= u(h,k) - u(0,k) \\ &= h \cdot \frac{d}{dx} u(\alpha h, k) \\ &= h\left(\frac{d}{dx} f(a+\alpha h, b+k) - \frac{d}{dx} f(a+\alpha h, b)\right) \\ &= hk \cdot \frac{d}{dy} \cdot \frac{d}{dx} f(a+\alpha h, b+\beta k) \end{aligned}$$

Similarly, again for $h, k \neq 0$, we can find $\gamma, \delta \in \mathbb{R}$ such that:

$$\begin{aligned} w(h,k) &= v(h,k) - v(h,0) \\ &= k \cdot \frac{d}{dy} v(h, \delta k) \\ &= k\left(\frac{d}{dy} f(a+h, b+\delta k) - \frac{d}{dy} f(a, b+\delta k)\right) \\ &= hk \cdot \frac{d}{dx} \cdot \frac{d}{dy} f(a+\gamma h, b+\delta k) \end{aligned}$$

Now by dividing everything by $hk \neq 0$, we conclude from this that the following equality holds, with the numbers $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ being found as above:

$$\frac{d}{dy} \cdot \frac{d}{dx} f(a+\alpha h, b+\beta k) = \frac{d}{dx} \cdot \frac{d}{dy} f(a+\gamma h, b+\delta k)$$

But with $h, k \to 0$ we get from this the Clairaut formula, at $z = a+ib$, as desired. □

With the above result in hand, we can now develop the theory of higher derivatives. Let us record here the following key result, happening at order 2, and which does the job, the job in analysis being usually that of finding the minima or maxima:

THEOREM 9.13. *Given a function $f : \mathbb{R}^N \to \mathbb{R}$, construct its Hessian, as being:*

$$f''(x) = \left(\frac{d^2 f}{dx_i dx_j}(x) \right)_{ij}$$

We have then the following approximation of f around a given point $x \in \mathbb{R}^N$,

$$f(x+t) \simeq f(x) + f'(x)t + \frac{< f''(x)t, t >}{2}$$

relating the positivity properties of $f''(x)$ to the local minima and maxima of f.

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

(1) As a first observation, at $N = 1$ the Hessian matrix as constructed above is the 1×1 matrix having as entry the second derivative $f''(x)$, and the formula in the statement is something that we know well, from one-variable calculus, namely:

$$f(x+t) \simeq f(x) + f'(x)t + \frac{f''(x)t^2}{2}$$

(2) In general now, this is in fact something which does not need a new proof, because it follows from the one-variable formula above, applied to the restriction of f to the following segment in $\mathbb{R}^N$, which can be regarded as being a one-variable interval:

$$I = [x, x+t]$$

To be more precise, let $y \in \mathbb{R}^N$, and consider the following function, with $r \in \mathbb{R}$:

$$g(r) = f(x+ry)$$

We know from (1) that the Taylor formula for g, at the point $r = 0$, reads:

$$g(r) \simeq g(0) + g'(0)r + \frac{g''(0)r^2}{2}$$

And our claim is that, with $t = ry$, this is precisely the formula in the statement.

(3) So, let us see if our claim is correct. By using the chain rule, we have the following formula, with on the right, as usual, a row vector multiplied by a column vector:

$$g'(r) = f'(x+ry) \cdot y$$

By using again the chain rule, we can compute the second derivative as well:

$$\begin{aligned} g''(r) &= (f'(x+ry)\cdot y)' \\ &= \left(\sum_i \frac{df}{dx_i}(x+ry)\cdot y_i\right)' \\ &= \sum_i\sum_j \frac{d^2f}{dx_i dx_j}(x+ry)\cdot \frac{d(x+ry)_j}{dr}\cdot y_i \\ &= \sum_i\sum_j \frac{d^2f}{dx_i dx_j}(x+ry)\cdot y_i y_j \\ &= < f''(x+ry)y, y > \end{aligned}$$

(4) Time now to conclude. We know that we have $g(r) = f(x+ry)$, and according to our various computations above, we have the following formulae:

$$g(0) = f(x) \quad , \quad g'(0) = f'(x) \quad , \quad g''(0) =< f''(x)y, y >$$

But with this data in hand, the usual Taylor formula for our one variable function g, at order 2, at the point $r = 0$, takes the following form, with $t = ry$:

$$\begin{aligned} f(x+ry) &\simeq f(x) + f'(x)ry + \frac{< f''(x)y, y > r^2}{2} \\ &= f(x) + f'(x)t + \frac{< f''(x)t, t >}{2} \end{aligned}$$

Thus, we have obtained the formula in the statement.

(5) Finally, the last assertion is standard. Indeed, in order for f to have a local extremum at x, we must have $f'(x) = 0$, so our Taylor formula becomes:

$$f(x+t) \simeq f(x) + \frac{< f''(x)t, t >}{2}$$

Thus, the local extrema of f are related to the positivity properties of f''. □

Next in line, we can talk as well about higher derivatives, as follows:

THEOREM 9.14. *Given a function $f : \mathbb{R}^N \to \mathbb{R}$, we can talk about its higher derivatives, defined recursively via the following formula,*

$$\frac{d^k f}{dx_{i_1}\ldots dx_{i_k}} = \frac{d}{dx_{i_1}}\cdots\frac{d}{dx_{i_k}}(f)$$

provided that all these derivatives exist indeed. Moreover, due to the Clairaut formula,

$$\frac{d^2 f}{dx_i dx_j} = \frac{d^2 f}{dx_j dx_i}$$

the order in which these higher derivatives are computed is irrelevant.

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

(1) First of all, we can talk about the quantities in the statement, with the remark however that at each step of our recursion, the corresponding partial derivative can exist of not. We will say in what follows that our function is k times differentiable if the quantities in the statement exist at any $l \leq k$, and smooth, if this works with $k = \infty$.

(2) Regarding now the second assertion, this is something more tricky. Let us first recall from the above that the second derivatives of a twice differentiable function of two variable $f : \mathbb{R}^2 \to \mathbb{R}$ are subject to the Clairaut formula, namely:

$$\frac{d^2 f}{dxdy} = \frac{d^2 f}{dydx}$$

(3) But this result clearly extends to our function $f : \mathbb{R}^N \to \mathbb{R}$, simply by ignoring the unneeded variables, so we have the Clairaut formula in general, also called Schwarz formula, which is the one in the statement, namely:

$$\frac{d^2 f}{dx_i dx_j} = \frac{d^2 f}{dx_j dx_i}$$

(4) Now observe that this tells us that the order in which the higher derivatives are computed is irrelevant. That is, we can permute the order of our partial derivative computations, and a standard way of doing this is by differentiating first with respect to x_1, as many times as needed, then with respect to x_2, and so on. Thus, the collection of our partial derivatives can be written, in a more convenient form, as follows:

$$\frac{d^k f}{dx_1^{k_1} \ldots dx_N^{k_N}} = \frac{d^{k_1}}{dx_1^{k_1}} \cdots \frac{d^{k_N}}{dx_N^{k_N}}(f)$$

(5) To be more precise, here $k \in \mathbb{N}$ is as usual the global order of our derivatives, the exponents $k_1, \ldots, k_N \in \mathbb{N}$ are subject to the condition $k_1 + \ldots + k_N = k$, and the operations on the right are the familiar one-variable higher derivative operations.

(6) This being said, for certain tricky questions it is more convenient not to order the indices, or rather to order them according to what order best fits your computation, so what we have in the statement is the good formula, and (4-5) are mere remarks. □

Regarding now the Taylor formula in several variables, that we already know to hold at order $k = 1, 2$, at higher order things become more complicated, as follows:

THEOREM 9.15. *Given an order k differentiable function $f : \mathbb{R}^N \to \mathbb{R}$, we have*

$$f(x + t) \simeq f(x) + f'(x)t + \frac{< f''(x)t, t >}{2} + \ldots$$

and this helps in identifying the local extrema, when $f'(x) = 0$ and $f''(x) = 0$.

PROOF. The study here is very similar to that at $k = 2$, from the proof of Theorem 9.13, with everything coming from the usual Taylor formula, applied on:

$$I = [x, x + t]$$

We will leave this as an instructive exercise, for your long Summer nights. □

As a conclusion to this, multivariable calculus leads us into positivity and negativity considerations for the various matrices formed by the partial derivatives, or by the higher partial derivatives. We will be back to this question, later in this chapter.

9c. Laplace operator

Before getting back to linear algebra, let us discuss a bit more the mathematics of the second derivative. We have here the following principle, in need to be explained:

PRINCIPLE 9.16. *The second derivative of a function $f : \mathbb{R}^N \to \mathbb{R}$, making the formula*

$$f(x + t) \simeq f(x) + f'(x)t + \frac{< f''(x)t, t >}{2}$$

work, is its Hessian matrix $f''(x) \in M_N(\mathbb{R})$, given by the following formula:

$$f''(x) = \left(\frac{d^2 f}{dx_i dx_j} \right)_{ij}$$

However, when needing a number, as second derivative, the trace of $f''(x)$, denoted

$$\Delta f = \sum_{i=1}^{N} \frac{d^2 f}{dx_i^2}$$

and called Laplacian of f, usually does the job.

In order to discuss this, and more specifically the last assertion, regarding the Laplacian, which is something new, let us start with the one variable functions. We have the following result about them, which is something a bit heuristic, and good to know:

PROPOSITION 9.17. *Intuitively, the second derivative of a function $f : \mathbb{R} \to \mathbb{R}$,*

$$f''(x) \in \mathbb{R}$$

computes how much different is $f(x)$, compared to the average of $f(y)$, with $y \simeq x$.

PROOF. As already mentioned, this is something a bit heuristic, but which is good to know. Let us write the Taylor formula at order 2, as such, and with $t \to -t$ too:

$$f(x + t) \simeq f(x) + f'(x)t + \frac{f''(x)}{2} t^2$$

$$f(x - t) \simeq f(x) - f'(x)t + \frac{f''(x)}{2} t^2$$

By making the average, we obtain the following formula:

$$\frac{f(x+t)+f(x-t)}{2}\simeq f(x)+\frac{f''(x)}{2}\,t^2$$

Thus, thinking a bit, we are led to the conclusion in the statement, modulo some discussion regarding what the average of $t^2/2$ exactly is. It is of course possible to say more here, but we will not really need all the details, in what follows. □

In several variables now, exactly the same happens, as follows:

PROPOSITION 9.18. *Intuitively, the following quantity, called Laplacian of f,*

$$\Delta f=\sum_{i=1}^{N}\frac{d^2f}{dx_i^2}$$

measures how much different is $f(x)$, compared to the average of $f(y)$, with $y\simeq x$.

PROOF. Again, this is something a bit heuristic, but which is good to know. Let us write, as before, the Taylor formula at order 2, as such, and with $t\to -t$ too:

$$f(x+t)\simeq f(x)+f'(x)t+\frac{<f''(x)t,t>}{2}$$

$$f(x-t)\simeq f(x)-f'(x)t+\frac{<f''(x)t,t>}{2}$$

By making the average, we obtain the following formula:

$$\frac{f(x+t)+f(x-t)}{2}\simeq f(x)+\frac{<f''(x)t,t>}{2}$$

Thus, thinking a bit, we are led to the conclusion in the statement, modulo some discussion about integrating all this, that we will not really need, in what follows. □

With this understood, the problem is now, what can we say about the mathematics of Δ? As a first observation, which is a bit speculative, the Laplace operator appears by applying twice the gradient operator, in a somewhat formal sense, as follows:

$$\begin{aligned}\Delta f&=\sum_{i=1}^{N}\frac{d^2f}{dx_i^2}\\&=\sum_{i=1}^{N}\frac{d}{dx_i}\cdot\frac{df}{dx_i}\\&=\left\langle\begin{pmatrix}\frac{d}{dx_1}\\\vdots\\\frac{d}{dx_N}\end{pmatrix},\begin{pmatrix}\frac{df}{dx_1}\\\vdots\\\frac{df}{dx_N}\end{pmatrix}\right\rangle\\&=<\nabla,\nabla f>\end{aligned}$$

Thus, it is possible to write a formula of type $\Delta = \nabla^2$, with the convention that the square of the gradient ∇ is taken in a scalar product sense, as above. However, this can be a bit confusing, and in what follows, we will not use this notation.

Instead of further thinking at this, and at double derivatives in general, let us formulate a more straightforward question, inspired by linear algebra, as follows:

QUESTION 9.19. *The Laplace operator being linear,*

$$\Delta(af + bg) = a\Delta f + b\Delta g$$

what can we say about it, inspired by usual linear algebra?

In answer now, the space of functions $f : \mathbb{R}^N \to \mathbb{R}$, on which Δ acts, being infinite dimensional, the usual tools from linear algebra do not apply as such, and we must be extremely careful. For instance, we cannot really expect to diagonalize Δ, via some sort of explicit procedure, as we usually do in linear algebra, for the usual matrices.

Thinking some more, there is actually a real bug too with our problem, because at $N = 1$ this problem becomes "what can we say about the second derivatives $f'' : \mathbb{R} \to \mathbb{R}$ of the functions $\varphi : \mathbb{R} \to \mathbb{R}$, inspired by linear algebra", with answer "not much".

And by thinking even more, still at $N = 1$, there is a second bug too, because if a function $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable, nothing will guarantee that its second derivative $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable too. Thus, we have some issues with the domain and range of Δ, regarded as linear operator, and these problems will persist at higher N.

So, shall we trash Question 9.19? Not so quick, because, very remarkably, some magic comes at $N = 2$ and higher in relation with complex analysis, according to:

PRINCIPLE 9.20. *The functions $f : \mathbb{R}^N \to \mathbb{R}$ which are 0-eigenvectors of Δ,*

$$\Delta f = 0$$

called harmonic functions, have the following properties:

(1) *At $N = 1$, nothing spectacular, these are just the linear functions.*
(2) *At $N = 2$, these are, locally, the real parts of holomorphic functions.*
(3) *At $N \geq 3$, these still share many properties with the holomorphic functions.*

In order to understand this principle, which is something quite deep, or at least get introduced to it, let us first look at the case $N = 2$. Here, any function $f : \mathbb{R}^2 \to \mathbb{R}$ can be regarded as function $f : \mathbb{C} \to \mathbb{R}$, depending on the following variable:

$$z = x + iy$$

But, in view of this, it is natural to enlarge the attention to the functions $f : \mathbb{C} \to \mathbb{C}$, and ask which of these functions are harmonic, $\Delta f = 0$. And here, we have the following remarkable result, making the link with complex analysis:

THEOREM 9.21. *Any holomorphic function $f : \mathbb{C} \to \mathbb{C}$, when regarded as function*

$$f : \mathbb{R}^2 \to \mathbb{C}$$

is harmonic. Moreover, the conjugates $\bar{f}$ of holomorphic functions are harmonic too.

PROOF. The first assertion comes from the following computation, with $z = x + iy$:

$$\begin{aligned}
\Delta z^n &= \frac{d^2 z^n}{dx^2} + \frac{d^2 z^n}{dy^2} \\
&= \frac{d(nz^{n-1})}{dx} + \frac{d(inz^{n-1})}{dy} \\
&= n(n-1)z^{n-2} - n(n-1)z^{n-2} \\
&= 0
\end{aligned}$$

As for the second assertion, this follows from $\Delta \bar{f} = \overline{\Delta f}$, which is clear from definitions, and which shows that if φ is harmonic, then so is its conjugate $\bar{\varphi}$. □

Many more things can be said, along these lines, notably a proof of the assertion (2) in Principle 9.20, which is however a quite tough piece of mathematics, and then with a clarification of the assertion (3) too, from that same principle, which again requires some substantial mathematics. For more on all this, you can check for instance Rudin [78].

9d. Positive matrices

Getting back now to our linear algebra business, we have seen that looking for the local extrema of a function, or other properties, leads us into looking at the positivity property of the Hessian, and into the mathematics of the Laplacian as well.

Thus, time to have a closer look at positivity, for the usual matrices, and for the infinite matrices too. The positivity theory, for arbitrary matrices, is as follows:

THEOREM 9.22. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is positive:*

(1) *$A = B^2$, with $B = B^*$.*
(2) *$A = CC^*$, for some $C \in M_N(\mathbb{C})$.*
(3) *$< Ax, x > \geq 0$, for any vector $x \in \mathbb{C}^N$.*
(4) *$A = A^*$, and the eigenvalues are positive, $\lambda_i \geq 0$.*
(5) *$A = UDU^*$, with $U \in U_N$ and with $D \in M_N(\mathbb{R}_+)$ diagonal.*

PROOF. This is something that we know from chapter 3, the idea being that the result follows from some elementary computations, and from the spectral theorem:

(1) $\implies$ (2) This is clear, because we can take $C = B$.

(2) $\implies$ (3) This follows indeed from the following computation:

$$< CC^*x, x >=< C^*x, C^*x >\geq 0$$

(3) $\implies$ (4) By using the fact that the numbers $< Ax, x >$ is real, we have:

$$< Ax, x >=< x, A^*x >=< A^*x, x >$$

Thus we have $A = A^*$, and the remaining assertion, regarding the eigenvalues, follows from the following computation, assuming $Ax = \lambda x$:

$$< Ax, x >=< \lambda x, x >= \lambda < x, x >\geq 0$$

(4) $\implies$ (5) This follows indeed by using the spectral theorem.

(5) $\implies$ (1) Assuming $A = UDU^*$ with $U \in U_N$, and with $D \in M_N(\mathbb{R}_+)$ diagonal, we can set $B = U\sqrt{D}U^*$. Then B is self-adjoint, and $B^2 = UDU^* = A$, as desired. □

Let us record as well the following useful technical version of the above result:

THEOREM 9.23. *For a matrix $A \in M_N(\mathbb{C})$ the following conditions are equivalent, and if they are satisfied, we say that A is strictly positive:*

(1) *$A = B^2$, with $B = B^*$, invertible.*
(2) *$A = CC^*$, for some $C \in M_N(\mathbb{C})$ invertible.*
(3) *$< Ax, x >> 0$, for any nonzero vector $x \in \mathbb{C}^N$.*
(4) *$A = A^*$, and the eigenvalues are strictly positive, $\lambda_i > 0$.*
(5) *$A = UDU^*$, with $U \in U_N$ and with $D \in M_N(\mathbb{R}_+^*)$ diagonal.*

PROOF. This follows either from Theorem 9.22, by adding the above various extra assumptions, or from the proof of Theorem 9.22, by modifying where needed. □

In practice now, there are many interesting examples of positive matrices, in the real life, and various applications of them, quite often coming in relation with the analysis of the multivariable functions. Many interesting things can be said here, and we will be back to this on a regular basis, in the present Part III of this book.

Getting back to generalities, let us see as well what happens in infinite dimensions, as a continuation of our discussion about linear operators from chapter 8. We have here:

THEOREM 9.24. *For an operator $T \in B(H)$, the following are equivalent:*

(1) *$< Tx, x >\geq 0$, for any $x \in H$.*
(2) *T is normal, and $\sigma(T) \subset [0, \infty)$.*
(3) *$T = S^2$, for some $S \in B(H)$ satisfying $S = S^*$.*
(4) *$T = R^*R$, for some $R \in B(H)$.*

If these conditions are satisfied, we call T positive, and write $T \geq 0$.

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

(1) $\implies$ (2) Assuming $< Tx, x > \geq 0$, with $S = T - T^*$ we have:

$$\begin{aligned} < Sx, x > &= < Tx, x > - < T^*x, x > \\ &= < Tx, x > - < x, Tx > \\ &= < Tx, x > - \overline{< Tx, x >} \\ &= 0 \end{aligned}$$

The next step is to use a polarization trick, as follows:

$$\begin{aligned} < Sx, y > &= < S(x+y), x+y > - < Sx, x > - < Sy, y > - < Sy, x > \\ &= - < Sy, x > \\ &= < y, Sx > \\ &= \overline{< Sx, y >} \end{aligned}$$

Thus we must have $< Sx, y > \in \mathbb{R}$, and with $y \to iy$ we obtain $< Sx, y > \in i\mathbb{R}$ too, and so $< Sx, y > = 0$. Thus $S = 0$, which gives $T = T^*$. Now since T is self-adjoint, it is normal as claimed. Moreover, by self-adjointness, we have:

$$\sigma(T) \subset \mathbb{R}$$

In order to prove now that we have indeed $\sigma(T) \subset [0, \infty)$, as claimed, we must invert $T + \lambda$, for any $\lambda > 0$. For this purpose, observe that we have:

$$\begin{aligned} < (T+\lambda)x, x > &= < Tx, x > + < \lambda x, x > \\ &\geq < \lambda x, x > \\ &= \lambda ||x||^2 \end{aligned}$$

But this shows that $T + \lambda$ is injective. In order to prove now the surjectivity, and the boundedness of the inverse, observe first that we have:

$$\begin{aligned} Im(T+\lambda)^\perp &= \ker(T+\lambda)^* \\ &= \ker(T+\lambda) \\ &= \{0\} \end{aligned}$$

Thus $Im(T + \lambda)$ is dense. On the other hand, observe that we have:

$$\begin{aligned} ||(T+\lambda)x||^2 &= < Tx + \lambda x, Tx + \lambda x > \\ &= ||Tx||^2 + 2\lambda < Tx, x > + \lambda^2 ||x||^2 \\ &\geq \lambda^2 ||x||^2 \end{aligned}$$

Thus for any vector in the image $y \in Im(T + \lambda)$ we have:

$$||y|| \geq \lambda \left|\left|(T+\lambda)^{-1} y\right|\right|$$

As a conclusion to what we have so far, $T+\lambda$ is bijective and invertible as a bounded operator from H onto its image, with the following norm bound:

$$||(T+\lambda)^{-1}|| \leq \lambda^{-1}$$

But this shows that $Im(T+\lambda)$ is complete, hence closed, and since we already knew that $Im(T+\lambda)$ is dense, our operator $T+\lambda$ is surjective, and we are done.

(2) $\implies$ (3) Since T is normal, and with spectrum contained in $[0,\infty)$, we can use the continuous functional calculus formula for the normal operators from chapter 8, with the function $f(x)=\sqrt{x}$, as to construct a square root $S=\sqrt{T}$.

(3) $\implies$ (4) This is trivial, because we can set $R=S$.

(4) $\implies$ (1) This is clear, because we have the following computation:

$$< R^*Rx, x >=< Rx, Rx >= ||Rx||^2$$

Thus, we have the equivalences in the statement. □

It is possible to talk as well about strictly positive operators, and we have here:

THEOREM 9.25. *For an operator $T \in B(H)$, the following are equivalent:*

(1) *T is positive and invertible.*
(2) *T is normal, and $\sigma(T) \subset (0,\infty)$.*
(3) *$T=S^2$, for some $S \in B(H)$ invertible, satisfying $S=S^*$.*
(4) *$T=R^*R$, for some $R \in B(H)$ invertible.*

If these conditions are satisfied, we call T strictly positive, and write $T>0$.

PROOF. Our claim is that the above conditions (1-4) are precisely the conditions (1-4) in Theorem 9.24, with the assumption "T is invertible" added. Indeed:

(1) This is clear by definition.

(2) In the context of Theorem 9.24 (2), namely when T is normal, and $\sigma(T) \subset [0,\infty)$, the invertibility of T, which means $0 \notin \sigma(T)$, gives $\sigma(T) \subset (0,\infty)$, as desired.

(3) In the context of Theorem 9.24 (3), namely when $T=S^2$, with $S=S^*$, by using the basic properties of the functional calculus for normal operators, the invertibility of T is equivalent to the invertibility of its square root $S=\sqrt{T}$, as desired.

(4) In the context of Theorem 9.24 (4), namely when $T=RR^*$, the invertibility of T is equivalent to the invertibility of R. This can be either checked directly, or deduced via the equivalence (3) $\iff$ (4) from Theorem 9.24, by using the above argument (3). □

As a subtlety now, we have the following complement to the above result:

THEOREM 9.26. *For a strictly positive operator, $T>0$, we have*

$$< Tx, x >> 0 \quad , \quad \forall x \neq 0$$

but the converse of this fact is not true, unless we are in finite dimensions.

PROOF. We have several things to be proved, the idea being as follows:

(1) Regarding the main assertion, the inequality can be deduced as follows, by using the fact that the operator $S = \sqrt{T}$ is invertible, and in particular injective:

$$\begin{aligned} < Tx, x > &= < S^2 x, x > \\ &= < Sx, S^* x > \\ &= < Sx, Sx > \\ &= ||Sx||^2 \\ &> 0 \end{aligned}$$

(2) In finite dimensions, assuming $< Tx, x >> 0$ for any $x \neq 0$, we know from Theorem 9.24 that we have $T \geq 0$. Thus we have $\sigma(T) \subset [0, \infty)$, and assuming by contradiction $0 \in \sigma(T)$, we obtain that T has $\lambda = 0$ as eigenvalue, and the corresponding eigenvector $x \neq 0$ has the property $< Tx, x >= 0$, contradiction. Thus $T > 0$, as claimed.

(3) Regarding now the counterexample, consider the following operator on $l^2(\mathbb{N})$:

$$T = \begin{pmatrix} 1 & & & \\ & \frac{1}{2} & & \\ & & \frac{1}{3} & \\ & & & \ddots \end{pmatrix}$$

This operator T is well-defined and bounded, and we have $< Tx, x >> 0$ for any $x \neq 0$. However T is not invertible, and so the converse does not hold, as stated. □

9e. Exercises

This was a quite routine analysis chapter, and as exercises on this, we have:

EXERCISE 9.27. *Check the details in the proof of the change of variable formula.*

EXERCISE 9.28. *Learn some other proofs of the change of variable formula.*

EXERCISE 9.29. *Compute the Gauss integral without polar coordinates. Can you?*

EXERCISE 9.30. *Compute the arbitrary polynomial integrals over the spheres.*

EXERCISE 9.31. *Work out the multivariable Taylor formula, at order $k \in \mathbb{N}$.*

EXERCISE 9.32. *Learn more, from physicists, about the Laplace operator.*

EXERCISE 9.33. *Learn about the harmonic functions, and their various properties.*

EXERCISE 9.34. *Make a list, from physics, of interesting positive matrices or operators.*

As bonus exercise, read if needed Rudin [77], [78], or an equivalent text.

CHAPTER 10

Forms, signature

10a. Bilinear forms

It is good time now to talk about geometry. We already know about conics, from chapter 1. However, when getting to $\mathbb{R}^3$, we are right away into a dillema, because the plane curves have two generalizations. First we have the algebraic curves in $\mathbb{R}^3$:

DEFINITION 10.1. *An algebraic curve in $\mathbb{R}^3$ is a curve as follows,*

$$C = \left\{(x, y, z) \in \mathbb{R}^3 \middle| P(x, y, z) = 0,\ Q(x, y, z) = 0\right\}$$

appearing as the joint zeroes of two polynomials P, Q.

These curves look of course like the usual plane curves, and at the level of the phenomena that can appear, these are similar to those in the plane, involving singularities and so on, but also knotting, which is a new phenomenon. However, it is hard to say something with bare hands about knots. We will be back to this, later in this book.

On the other hand, as another natural generalization of the plane curves, and this might sound a bit surprising, we have the surfaces in $\mathbb{R}^3$, constructed as follows:

DEFINITION 10.2. *An algebraic surface in $\mathbb{R}^3$ is a surface as follows,*

$$S = \left\{(x, y, z) \in \mathbb{R}^3 \middle| P(x, y, z) = 0\right\}$$

appearing as the zeroes of a polynomial P.

The point indeed is that, as it was the case with the plane curves, what we have here is something defined by a single equation. And with respect to many questions, having a single equation matters a lot, and this is why surfaces in $\mathbb{R}^3$ are "simpler" than curves in $\mathbb{R}^3$. In fact, believe me, they are even the correct generalization of the curves in $\mathbb{R}^2$.

As an example of what can be done with surfaces, which is very similar to what we did with the conics $C \subset \mathbb{R}^2$ in chapter 1, we have the following result:

THEOREM 10.3. *The degree 2 surfaces $S \subset \mathbb{R}^3$, called quadrics, are the ellipsoid*

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 + \left(\frac{z}{c}\right)^2 = 1$$

which is the only compact one, plus 16 more, which can be explicitly listed.

PROOF. We will be quite brief here, because we intend to rediscuss all this in a moment, with full details, in arbitrary N dimensions, the idea being as follows:

(1) The equations for a quadric $S \subset \mathbb{R}^2$ are best written as follows, with $A \in M_3(\mathbb{R})$ being a matrix, $B \in M_{1\times 3}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$< Au, u > +Bu + C = 0$$

(2) By doing now the linear algebra, and we will come back to this in a moment, with details, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo degeneracy and linear transformations, with signed sums of squares, as follows:

$$\pm x^2 \pm y^2 \pm z^2 = 0, 1$$

(3) Thus the sphere is the only compact quadric, up to linear transformations, and by applying now linear transformations to it, we are led to the ellipsoids in the statement.

(4) As for the other quadrics, there are many of them, a bit similar to the parabolas and hyperbolas in 2 dimensions, and some work here leads to a 16 item list. □

With this done, instead of further insisting on the surfaces $S \subset \mathbb{R}^3$, or getting into their rivals, the curves $C \subset \mathbb{R}^3$, which appear as intersections of such surfaces, $C = S \cap S'$, let us get instead to arbitrary N dimensions, see what the axiomatics looks like there, with the hope that this will clarify our dimensionality dillema, curves vs surfaces.

So, moving to N dimensions, we have here the following definition, to start with:

DEFINITION 10.4. *An algebraic hypersurface in $\mathbb{R}^N$ is a space of the form*

$$S = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P(x_1, \ldots, x_N) = 0, \forall i\right\}$$

appearing as the zeroes of a polynomial $P \in \mathbb{R}[x_1, \ldots, x_N]$.

Again, this is a quite general definition, covering both the plane curves $C \subset \mathbb{R}$ and the surfaces $S \subset \mathbb{R}^2$, which is certainly worth a systematic exploration. But, no hurry with this, for the moment we are here for talking definitons and axiomatics.

In order to have now a full collection of beasts, in all possible dimensions $N \in \mathbb{N}$, and of all possible dimensions $k \in \mathbb{N}$, we must intersect such algebraic hypersurfaces. We are led in this way to the zeroes of families of polynomials, as follows:

DEFINITION 10.5. *An algebraic manifold in $\mathbb{R}^N$ is a space of the form*

$$X = \left\{(x_1, \ldots, x_N) \in \mathbb{R}^N \middle| P_i(x_1, \ldots, x_N) = 0, \forall i\right\}$$

with $P_i \in \mathbb{R}[x_1, \ldots, x_N]$ being a family of polynomials.

As a first observation, as already mentioned, such a manifold appears as an intersection of hypersurfaces S_i, those associated to the various polynomials P_i:

$$X = S_1 \cap \ldots \cap S_r$$

There is actually a bit of a discussion needed here, regarding the parameter $r \in \mathbb{N}$, shall we allow this parameter to be $r = \infty$ too, or not. However, with some abstract commutative algebra helping, the idea is that allowing $r = \infty$ forces in fact $r < \infty$.

As an announcement now, good news, what we have in Definition 10.5 is the good and final notion of algebraic manifold, very general, and with the branch of mathematics studying such manifolds being called algebraic geometry. In what follows we will discuss a bit what can be done with this, as a continuation of our previous work on the plane curves, at the elementary level, by using our accumulated linear algebra knowledge.

Let us first look more in detail at the hypersurfaces. We have here:

THEOREM 10.6. *The degree 2 hypersurfaces $S \subset \mathbb{R}^N$, called quadrics, are up to degeneracy and to linear transformations the hypersurfaces of the following form,*

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

and with the sphere being the only compact one.

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

(1) The equations for a quadric $S \subset \mathbb{R}^N$ are best written as follows, with $A \in M_N(\mathbb{R})$ being a matrix, $B \in M_{1\times N}(\mathbb{R})$ being a row vector, and $C \in \mathbb{R}$ being a constant:

$$< Ax, x > +Bx + C = 0$$

(2) By doing the linear algebra, or by invoking the theorem of Sylvester on quadratic forms, we are left, modulo linear transformations, with signed sums of squares:

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

(3) To be more precise, with linear algebra, by evenly distributing the terms $x_i x_j$ above and below the diagonal, we can assume that our matrix $A \in M_N(\mathbb{R})$ is symmetric. Thus A must be diagonalizable, and by changing the basis of $\mathbb{R}^N$, as to have it diagonal, our equation becomes as follows, with $D \in M_N(\mathbb{R})$ being now diagonal:

$$< Dx, x > +Ex + F = 0$$

(4) But now, by making squares in the obvious way, which amounts in applying yet another linear transformation to our quadric, the equation takes the following form, with $G \in M_N(-1, 0, 1)$ being diagonal, and with $H \in \{0, 1\}$ being a constant:

$$< Gx, x >= H$$

(5) Now barring the degenerate cases, we can further assume $G \in M_N(-1,1)$, and we are led in this way to the equation claimed in (2) above, namely:

$$\pm x_1^2 \pm \ldots \pm x_N^2 = 0, 1$$

(6) In particular we see that, up to some degenerate cases, namely emptyset and point, the only compact quadric, up to linear transformations, is the one given by:

$$x_1^2 + \ldots + x_N^2 = 1$$

(7) But this is the unit sphere, so are led to the conclusions in the statement. □

Many other things can be said, as a continuation of this, and we can only recommend here learning some algebraic geometry, say from Harris [45] or Shafarevich [86].

10b. Smooth manifolds

We already know about the algebraic curves, then surfaces and other algebraic manifolds, generalizing the conics, from the above. A second idea now, in order to generalize the conics, is to look at the smooth manifolds, in the following sense:

DEFINITION 10.7. *A smooth manifold is a space X which is locally isomorphic to $\mathbb{R}^N$. To be more precise, this space X must be covered by charts, bijectively mapping open pieces of it to open pieces of $\mathbb{R}^N$, with the changes of charts being C^∞ functions.*

It is of course possible to talk as well about C^k manifolds, with $k < \infty$, but this is rather technical material, that we will not get into, in this book.

As basic examples of smooth manifolds, we have of course $\mathbb{R}^N$ itself, or any open subset $X \subset \mathbb{R}^N$, with only 1 chart being needed here. Other basic examples include the circle, or curves like ellipses and so on, for obvious reasons. To be more precise, the unit circle can be covered by 2 charts as above, by using polar coordinates, in the obvious way, and then by applying dilations, translations and other such transformations, namely bijections which are smooth, we obtain a whole menagery of circle-looking manifolds.

Here is a more precise statement in this sense, covering the conics:

THEOREM 10.8. *The following are smooth manifolds, in the plane:*

(1) *The circles.*
(2) *The ellipses.*
(3) *The non-degenerate conics.*
(4) *Smooth deformations of these.*

PROOF. All this is quite intuitive, the idea being as follows:

(1) Consider the unit circle, $x^2 + y^2 = 1$. We can write then $x = \cos t$, $y = \sin t$, with $t \in [0, 2\pi)$, and we seem to have here the solution to our problem, just using 1 chart. But this is of course wrong, because $[0, 2\pi)$ is not open, and we have a problem at 0. In

practice we need to use 2 such charts, say with the first one being with $t \in (0, 3\pi/2)$, and the second one being with $t \in (\pi, 5\pi/2)$. As for the fact that the change of charts is indeed smooth, this comes by writing down the formulae, or just thinking a bit, and arguing that this change of chart being actually a translation, it is automatically linear.

(2) This follows from (1), by pulling the circle in both the Ox and Oy directions, and the formulae here, based on those for ellipses from chapter 1, are left to you reader.

(3) We already have the ellipses, and the case of the parabolas and hyperbolas is elementary as well, and in fact simpler than the case of the ellipses. Indeed, a parablola is clearly homeomorphic to $\mathbb{R}$, and a hyperbola, to two copies of $\mathbb{R}$.

(4) This is something which is clear too, depending of course on what exactly we mean by "smooth deformation", and by using a bit of multivariable calculus if needed. □

In higher dimensions, as basic examples, we have the spheres, as shown by:

THEOREM 10.9. *The sphere is a smooth manifold.*

PROOF. There are several proofs for this, all instructive, as follows:

(1) A first idea is to use spherical coordinates, which are as follows:

$$\begin{cases} x_1 &= r\cos t_1 \\ x_2 &= r\sin t_1\cos t_2 \\ \vdots & \\ x_{N-1} &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\cos t_{N-1} \\ x_N &= r\sin t_1\sin t_2\ldots\sin t_{N-2}\sin t_{N-1} \end{cases}$$

Indeed, these produce explicit charts for the sphere.

(2) A second idea, which makes use of less charts, namely 2 charts only, is to use the stereographic projection, which is given by inverse maps as follows:

$$\Phi : \mathbb{R}^N \to S^N_{\mathbb{R}} - \{\infty\} \quad , \quad \Psi : S^N_{\mathbb{R}} - \{\infty\} \to \mathbb{R}^N$$

To be more precise, the formulae of these maps, which are elementary to establish, are as follows, with the convention $\mathbb{R}^{N+1} = \mathbb{R} \times \mathbb{R}^N$, and with the coordinate of $\mathbb{R}$ denoted x_0, and with the coordinates of $\mathbb{R}^N$ denoted $x_1, \ldots, x_N$:

$$\Phi(v) = (1, 0) + \frac{2}{1 + ||v||^2}(-1, v) \quad , \quad \Psi(c, x) = \frac{x}{1 - c}$$

Indeed, we get in this way explicit charts for the sphere.

(3) We have as well cylindrical coordinates, as well as many other types of more specialized coordinates, which can be useful in physics, plus of course, in disciplines like geography, economics and so on. There are many interesting computations that can be

done here, and we will be back to these, on a regular basis in what follows, once we will know more about smooth manifolds, and their properties. □

Other basic examples of smooth manifolds include the projective spaces:

THEOREM 10.10. *The projective space $P_{\mathbb{R}}^{N-1}$ is a smooth manifold, with charts*

$$(x_1, \ldots, x_N) \to \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_N}{x_i} \right)$$

where $x_i \neq 0$. This manifold is compact, and of dimension $N-1$.

PROOF. We know that $P_{\mathbb{R}}^{N-1}$ appears by definition as the space of lines in $\mathbb{R}^N$ passing through the origin, so we have the following formula, with $\sim$ being the proportionality of vectors, given as usual by $x \sim y$ when $x = \lambda y$, for some scalar $\lambda \neq 0$:

$$P_{\mathbb{R}}^{N-1} = \mathbb{R}^N - \{0\} / \sim$$

Alternatively, we can restrict if we want the attention to the vectors on the unit sphere $S_{\mathbb{R}}^{N-1} \subset \mathbb{R}^N$, and this because any line in $\mathbb{R}^N$ passing through the origin will certainly cross this sphere. Moreover, it is clear that our line will cross the sphere in exactly two points $\pm x$, and we conclude that we have the following formula, with $\sim$ being now the proportionality of vectors on the sphere, given by $x \sim y$ when $x = \pm y$:

$$P_{\mathbb{R}}^{N-1} = S_{\mathbb{R}}^{N-1} / \sim$$

With this discussion made, let us get now to what is to be proved. Obviously, once we fix an index $i \in \{1, \ldots, N\}$, the condition $x_i \neq 0$ on the vectors $x \in \mathbb{R}^N - \{0\}$ defines an open subset $U_i \subset P_{\mathbb{R}}^{N-1}$, and the open subsets that we get in this way cover $P_{\mathbb{R}}^{N-1}$:

$$P_{\mathbb{R}}^{N-1} = U_1 \cup \ldots \cup U_N$$

Moreover, the map in the statement is injective $U_i \to \mathbb{R}^{N-1}$, and it is clear too that the changes of charts are C^∞. Thus, we have our smooth manifold, as claimed. □

As a continuation of this, we can talk about Riemannian manifolds, which are those smooth manifolds where we can talk about length, area, volume, integration and so on.

The theory of Riemannian manifolds is far more advanced with respect to what can be done with the arbitrary smooth manifolds, for somewhat obvious reasons. At the level of applications, we will see later in this chapter that, save for a slight modification in the axioms, we can use such manifolds for understanding our surrounding space-time.

As a last topic of discussion, in regards with general differential geometry, let us go back now to the optimization questions from the end of the previous chapter. Thinking well, the functions that we have to minimize or maximize, in the real life, are often defined on a manifold, instead of being defined on the whole $\mathbb{R}^N$. Fortunately, the good old principle $f'(x) = 0$ can be adapted to the manifold case, as follows:

PRINCIPLE 10.11. *In order for a function $f : X \to \mathbb{R}$ defined on a manifold X to have a local extremum at $x \in X$, we must have, as usual*

$$f'(x) = 0$$

but with this taking into account the fact that the equations defining the manifold count as well as "zero", and so must be incorporated into the formula $f'(x) = 0$.

In what follows, we will take this principle as granted. In practice, the idea is that we must have a formula as follows, with g_i being the constraint functions for our manifold X, and with $\lambda_i \in \mathbb{R}$ being certain scalars, called Lagrange multipliers:

$$f'(x) = \sum_i \lambda_i g_i'(x)$$

As a basic illustration for this, our claim is that, by using a suitable manifold, and a suitable function, and Lagrange multipliers, we can prove in this way the Hölder inequality, that we know well of course, but without any computation. Let us start with:

PROPOSITION 10.12. *For any exponent $p > 1$, the following set*

$$S_p = \left\{ x \in \mathbb{R}^N \middle| \sum_i |x_i|^p = 1 \right\}$$

is a submanifold of $\mathbb{R}^N$.

PROOF. We know from the above that the unit sphere in $\mathbb{R}^N$ is a manifold. In our terms, this solves our problem at $p = 2$, because this unit sphere is:

$$S_2 = \left\{ x \in \mathbb{R}^N \middle| \sum_i x_i^2 = 1 \right\}$$

Now observe that we have a bijection $S_p \simeq S_2$, at least on the part where all the coordinates are positive, $x_i > 0$, given by the following function:

$$x_i \to x_i^{2/p}$$

Thus we obtain that S_p is indeed a manifold, as claimed. □

We already know that the manifold S_p constructed above is the unit sphere, in the case $p = 2$. In order to have a better geometric picture of what is going on, in general, observe that S_p can be constructed as well at $p = 1$, as follows:

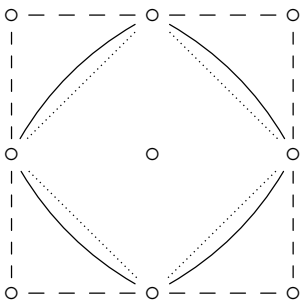
$$S_1 = \left\{ x \in \mathbb{R}^N \middle| \sum_i |x_i| = 1 \right\}$$

However, this is no longer a manifold, as we can see for instance at $N = 2$, where we obtain a square. Now observe that we can talk as well about $p = \infty$, as follows:

$$S_\infty = \left\{ x \in \mathbb{R}^N \Big| \sup_i |x_i| = 1 \right\}$$

This letter set is no longer a manifold either, as we can see for instance at $N = 2$, where we obtain again a square, containing the previous square, the one at $p = 1$.

With these limiting constructions in hand, we can have now a better geometric picture of what is going on, in the general context of Proposition 10.12. Indeed, let us draw, at $N = 2$ for simplifying, our sets S_p at the values $p = 1, 2, \infty$ of the exponent:



We can see that what we have is a small square, at $p = 1$, becoming smooth and inflating towards the circle, in the parameter range $p \in (1, 2]$, and then further inflating, in the parameter range $p \in [2, \infty)$, towards the big square appearing at $p = \infty$.

With these preliminaries in hand, we can formulate our result, as follows:

THEOREM 10.13. *The local extrema over S_p of the function*

$$f(x) = \sum_i x_i y_i$$

can be computed by using Lagrange multipliers, and this gives

$$\left| \sum_i x_i y_i \right| \leq \left(\sum_i |x_i|^p \right)^{1/p} \left(\sum_i |y_i|^q \right)^{1/q}$$

with $1/p + 1/q = 1$, that is, the Hölder inequality, with a purely geometric proof.

PROOF. We can restrict the attention to the case where all the coordinates are positive, $x_i > 0$ and $y_i > 0$. The derivative of the function in the statement is:

$$f'(x) = (y_1, \ldots, y_N)$$

On the other hand, we know that the manifold S_p appears by definition as the set of zeroes of the function $\varphi(x) = \sum_i x_i^p - 1$, having derivative as follows:

$$\varphi'(x) = p(x_1^{p-1}, \ldots, x_N^{p-1})$$

Thus, by using Lagrange multipliers, the critical points of f must satisfy:

$$(y_1, \ldots, y_N) \sim (x_1^{p-1}, \ldots, x_N^{p-1})$$

In other words, the critical points must satisfy $x_i = \lambda y_i^{1/(p-1)}$, for some $\lambda > 0$, and by using now $\sum_i x_i^p = 1$ we can compute the precise value of λ, and we get:

$$\lambda = \left(\sum_i y_i^{p/(p-1)} \right)^{-1/p}$$

Now let us see what this means. Since the critical point is unique, this must be a maximum of our function, and we conclude that for any $x \in S_p$, we have:

$$\sum_i x_i y_i \leq \sum_i \lambda y_i^{1/(p-1)} \cdot y_i = \left(\sum_i y_i^{p/(p-1)} \right)^{1-1/p} = \left(\sum_i y_i^q \right)^{1/q}$$

Thus we have Hölder, and the general case follows from this, by rescaling. □

There are many other possible applications of the Lagrange multipliers, to all sorts of science questions, and for more on all this, including of course a proof of Principle 10.11 too, we refer to any solid differential geometry book, of pure or applied type.

10c. Relativity theory

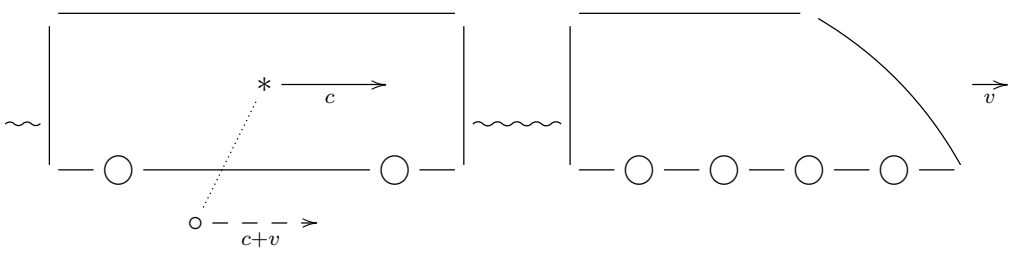
Let us discuss now some applications of the above to basic theoretical physics. Based on experiments by Fizeau, then Michelson-Morley and others, and some physics by Maxwell and Lorentz too, Einstein came upon the following principles:

FACT 10.14 (Einstein principles). *The following happen:*

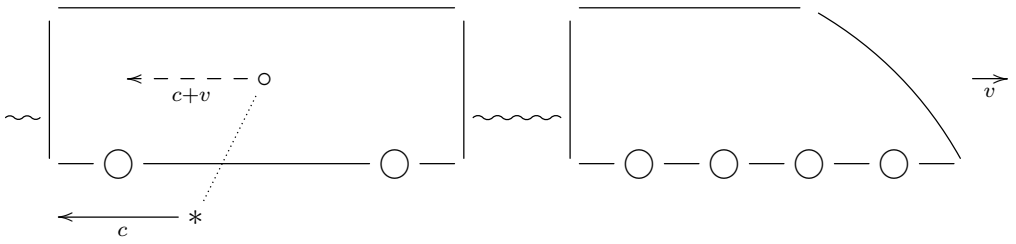
(1) *Light travels in vacuum at a finite speed, $c < \infty$.*
(2) *This speed c is the same for all inertial observers.*
(3) *In non-vacuum, the light speed is lower, $v < c$.*
(4) *Nothing can travel faster than light, $v \not> c$.*

The point now is that, obviously, something is wrong here. Indeed, assuming for instance that we have a train, running in vacuum at speed $v > 0$, and someone on board lights a flashlight $*$ towards the locomotive, then an observer $\circ$ on the ground will see the

light traveling at speed $c+v>c$, which is a contradiction:



Equivalently, with the same train running, in vacuum at speed $v>0$, if the observer on the ground lights a flashlight $*$ towards the back of the train, then viewed from the train, that light will travel at speed $c+v>c$, which is a contradiction again:



Summarizing, Fact 10.14 implies $c+v=c$, so contradicts classical mechanics, which therefore needs a fix. By dividing all speeds by c, as to have $c=1$, and by restricting the attention to the 1D case, to start with, we are led to the following puzzle:

PUZZLE 10.15. *How to define speed addition on the space of* 1D *speeds, which is*

$$I=[-1,1]$$

with our $c=1$ *convention, as to have* $1+c=1$*, as required by physics?*

In view of our basic geometric knowledge, a natural idea here would be that of wrapping $[-1,1]$ into a circle, and then stereographically projecting on $\mathbb{R}$. Indeed, we can then "import" to $[-,1,1]$ the usual addition on $\mathbb{R}$, via the inverse of this map.

So, let us see where all this leads us. First, the formula of our map is as follows:

PROPOSITION 10.16. *The map wrapping* $[-1,1]$ *into the unit circle, and then stereographically projecting on* $\mathbb{R}$ *is given by the formula*

$$\varphi(u)=\tan\left(\frac{\pi u}{2}\right)$$

with the convention that our wrapping is the most straightforward one, making correspond $\pm1\to i$*, with negatives on the left, and positives on the right.*

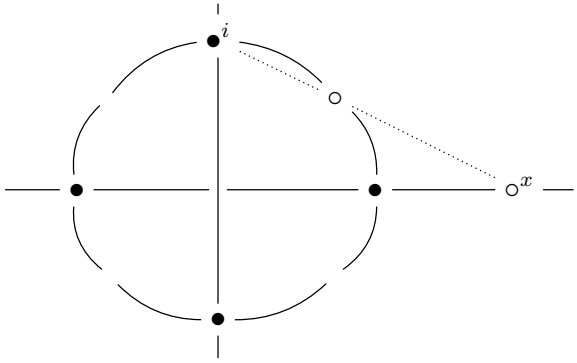
PROOF. Regarding the wrapping, as indicated, this is given by:

$$u \to e^{it} \quad , \quad t = \pi u - \frac{\pi}{2}$$

Indeed, this correspondence wraps $[-1, 1]$ as above, the basic instances of our correspondence being as follows, and with everything being fine modulo 2π:

$$-1 \to \frac{\pi}{2} \quad , \quad -\frac{1}{2} \to -\pi \quad , \quad 0 \to -\frac{\pi}{2} \quad , \quad \frac{1}{2} \to 0 \quad , \quad 1 \to \frac{\pi}{2}$$

Regarding now the stereographic projection, the picture here is as follows:



Thus, by Thales, the formula of the stereographic projection is as follows:

$$\frac{\cos t}{x} = \frac{1 - \sin t}{1} \implies x = \frac{\cos t}{1 - \sin t}$$

Now if we compose our wrapping operation above with the stereographic projection, what we get is, via the above Thales formula, and some trigonometry:

$$\begin{aligned} x &= \frac{\cos t}{1 - \sin t} \\ &= \frac{\cos\left(\pi u - \frac{\pi}{2}\right)}{1 - \sin\left(\pi u - \frac{\pi}{2}\right)} \\ &= \frac{\cos\left(\frac{\pi}{2} - \pi u\right)}{1 + \sin\left(\frac{\pi}{2} - \pi u\right)} \\ &= \frac{\sin(\pi u)}{1 + \cos(\pi u)} \\ &= \frac{2\sin\left(\frac{\pi u}{2}\right)\cos\left(\frac{\pi u}{2}\right)}{2\cos^2\left(\frac{\pi u}{2}\right)} \\ &= \tan\left(\frac{\pi u}{2}\right) \end{aligned}$$

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

The above result is very nice, but when it comes to physics, things do not work, for instance because of the wrong slope of the function $\varphi(u) = \tan\left(\frac{\pi u}{2}\right)$ at the origin, which makes our summing on $[-1, 1]$ not compatible with the Galileo addition, at low speeds.

So, what to do? Obviously, trash Proposition 10.16, and start all over again. Getting back now to Puzzle 10.15, this has in fact a simpler solution, based this time on algebra, and which in addition is the good, physically correct solution, as follows:

THEOREM 10.17. *If we sum the speeds according to the Einstein formula*

$$u +_e v = \frac{u + v}{1 + uv}$$

then the Galileo formula still holds, approximately, for low speeds

$$u +_e v \simeq u + v$$

and if we have $u = 1$ or $v = 1$, the resulting sum is $u +_e v = 1$.

PROOF. All this is self-explanatory, and clear from definitions, and with the Einstein formula of $u +_e v$ itself being just an obvious solution to Puzzle 10.15, provided that, importantly, we know 0 geometry, and rely on very basic algebra only. □

So, very nice, problem solved, at least in 1D. But, shall we give up with geometry, and the stereographic projection? Certainly not, let us try to recycle that material. In order to do this, let us recall that the usual trigonometric functions are given by:

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \quad , \quad \cos x = \frac{e^{ix} + e^{-ix}}{2} \quad , \quad \tan x = \frac{e^{ix} - e^{-ix}}{i(e^{ix} + e^{-ix})}$$

The point now is that, and you might know this from calculus, the above functions have some natural "hyperbolic" or "imaginary" analogues, constructed as follows:

$$\sinh x = \frac{e^{x} - e^{-x}}{2} \quad , \quad \cosh x = \frac{e^{x} + e^{-x}}{2} \quad , \quad \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

But the function on the right, tanh, starts reminding the formula of Einstein addition, from Theorem 10.17. So, we have our idea, and we are led to the following result:

THEOREM 10.18. *The Einstein speed summation in 1D is given by*

$$\tanh x +_e \tanh y = \tanh(x + y)$$

with $\tanh : [-\infty, \infty] \to [-1, 1]$ *being the hyperbolic tangent function.*

PROOF. This follows by putting together our various formulae above, but it is perhaps better, for clarity, to prove this directly. Our claim is that we have:

$$\tanh(x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$

But this can be checked via direct computation, from the definitions, as follows:

$$\begin{aligned}
&\frac{\tanh x+\tanh y}{1+\tanh x\tanh y}\\
=\ &\left(\frac{e^x-e^{-x}}{e^x+e^{-x}}+\frac{e^y-e^{-y}}{e^y+e^{-y}}\right)\Big/\left(1+\frac{e^x-e^{-x}}{e^x+e^{-x}}\cdot\frac{e^y-e^{-y}}{e^y+e^{-y}}\right)\\
=\ &\frac{(e^x-e^{-x})(e^y+e^{-y})+(e^x+e^{-x})(e^y-e^{-y})}{(e^x+e^{-x})(e^y+e^{-y})+(e^x-e^{-x})(e^y+e^{-y})}\\
=\ &\frac{2(e^{x+y}-e^{-x-y})}{2(e^{x+y}+e^{-x-y})}\\
=\ &\tanh(x+y)
\end{aligned}$$

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

Very nice all this, hope you agree. As a conclusion, passing from the Riemann stereographic projection sum to the Einstein summation basically amounts in replacing:

$$\tan\to\tanh$$

Let us formulate as well this finding more philosophically, as follows:

Conclusion 10.19. *The Einstein speed summation in 1D is the imaginary analogue of the summation on* $[-1,1]$ *obtained via Riemann's stereographic projection.*

Getting now to several dimensions, we have an analogue of Puzzle 10.15 here, and after doing the math, we are led to the following conclusion:

Theorem 10.20. *When defining the Einstein speed summation in* 3D *as*

$$u+_e v=\frac{1}{1+<u,v>}\left(u+v+\frac{u\times(u\times v)}{1+\sqrt{1-||u||^2}}\right)$$

in $c=1$ *units, the following happen:*

(1) *When* $u\sim v$*, we recover the previous* 1D *formula.*
(2) *We have* $||u||,||v||<1\implies||u+_e v||<1$*.*
(3) *When* $||u||=1$*, we have* $u+_e v=u$*.*
(4) *When* $||v||=1$*, we have* $||u+_e v||=1$*.*
(5) *However,* $||v||=1$ *does not imply* $u+_e v=v$*.*
(6) *Also, the formula* $u+_e v=v+_e u$ *fails.*

In addition, the above formula is physically correct, agreeing with experiments.

Proof. This is something quite tricky, with the key physics claim at the end being indeed true, and with the idea with the mathematical part being as follows:

(1) This is something which follows from definitions.

(2) In order to simplify notation, let us set $\delta = \sqrt{1 - ||u||^2}$, which is the inverse of the quantity $\gamma = 1/\sqrt{1 - ||u||^2}$. With this convention, we have:

$$\begin{aligned} u +_e v &= \frac{1}{1+ < u, v >} \left(u + v + \frac{< u, v > u - ||u||^2 v}{1 + \delta} \right) \\ &= \frac{(1 + \delta + < u, v >)u + (1 + \delta - ||u||^2)v}{(1+ < u, v >)(1 + \delta)} \end{aligned}$$

Taking now the squared norm and computing gives the following formula:

$$||u +_e v||^2 = \frac{(1 + \delta)^2 ||u + v||^2 + (||u||^2 - 2(1 + \delta))(||u||^2 ||v||^2 - < u, v >^2)}{(1+ < u, v >)^2 (1 + \delta)^2}$$

But this formula can be further processed by using $\delta = \sqrt{1 - ||u||^2}$, and by navigating through the various quantities which appear, we obtain, as a final product:

$$||u +_e v||^2 = \frac{||u + v||^2 - ||u||^2 ||v||^2 + < u, v >^2}{(1+ < u, v >)^2}$$

But this type of formula is exactly what we need, for what we want to do. Indeed, by assuming $||u||, ||v|| < 1$, we have the following estimate:

$$\begin{aligned} ||u +_e v||^2 < 1 &\iff ||u + v||^2 - ||u||^2 ||v||^2 + < u, v >^2 < (1+ < u, v >)^2 \\ &\iff ||u + v||^2 - ||u||^2 ||v||^2 < 1 + 2 < u, v > \\ &\iff ||u||^2 + ||v||^2 - ||u||^2 ||v||^2 < 1 \\ &\iff (1 - ||u||^2)(1 - ||v||^2) > 0 \end{aligned}$$

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

(3) This is something elementary, coming from definitions.

(4) This comes from the squared norm formula established in the proof of (2) above, because when assuming $||v|| = 1$, we obtain:

$$\begin{aligned} ||u +_e v||^2 &= \frac{||u + v||^2 - ||u||^2 + < u, v >^2}{(1+ < u, v >)^2} \\ &= \frac{||u||^2 + 1 + 2 < u, v > - ||u||^2 + < u, v >^2}{(1+ < u, v >)^2} \\ &= \frac{1 + 2 < u, v > + < u, v >^2}{(1+ < u, v >)^2} \\ &= 1 \end{aligned}$$

(5) This is clear, from the obvious lack of symmetry of our formula.

(6) This is again clear, from the obvious lack of symmetry of our formula. □

Time now to draw some concrete conclusions, from the above speed computations. Since speed $v = d/t$ is distance over time, we must fine-tune distance d, or time t, or both. Let us first discuss, following as usual Einstein, what happens to time t. Here the result, which might seem quite surprising, at a first glance, is as follows:

THEOREM 10.21. *Relativistic time is subject to Lorentz dilation*

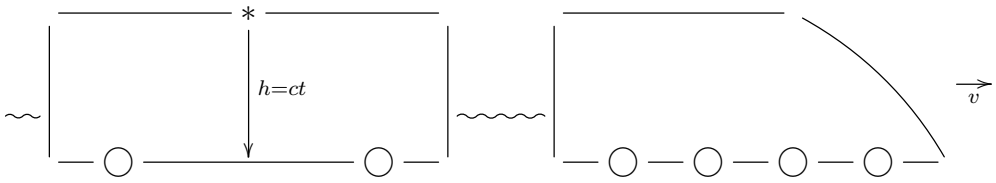
$$t \to \gamma t$$

where the number $\gamma \geq 1$, called Lorentz factor, is given by the formula

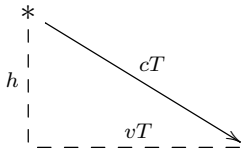
$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

with v being the moving speed, at which time is measured.

PROOF. Assume indeed that we have a train, moving to the right with speed v, through vacuum. In order to compute the height h of the train, the passenger onboard switches on the ceiling light bulb, measures the time t that the light needs to hit the floor, by traveling at speed c, and concludes that the train height is $h = ct$:



On the other hand, an observer on the ground will see here something different, namely a right triangle, with on the vertical the height of the train h, on the horizontal the distance vT that the train has traveled, and on the hypotenuse the distance cT that light has traveled, with T being the duration of the event, according to his watch:



Now by Pythagoras applied to this triangle, we have the following formula:

$$h^2 + (vT)^2 = (cT)^2$$

Thus, the observer on the ground will reach to the following formula for h:

$$h = \sqrt{c^2 - v^2} \cdot T$$

But h must be the same for both observers, so we have the following formula:

$$\sqrt{c^2 - v^2} \cdot T = ct$$

It follows that the two times t and T are indeed not equal, and are related by:

$$T = \frac{ct}{\sqrt{c^2 - v^2}} = \frac{t}{\sqrt{1 - v^2/c^2}} = \gamma t$$

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

Let us discuss now what happens to length. We have here the following result:

THEOREM 10.22. *Relativistic length is subject to Lorentz contraction*

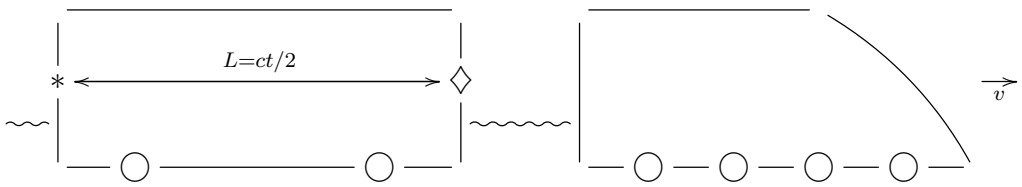
$$L \to L/\gamma$$

where the number $\gamma \geq 1$, called Lorentz factor, is given by the usual formula

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

with v being the moving speed, at which length is measured.

PROOF. As before in the proof of Theorem 10.21, meaning in the same train traveling at speed v, in vacuum, imagine now that the passenger wants to measure the length L of the car. For this purpose he switches on the light bulb, now at the rear of the car, and measures the time t needed for the light to reach the front of the car, and get reflected back by a mirror installed there, according to the following scheme:



He concludes that, as marked above, the length L of the car is given by:

$$L = \frac{ct}{2}$$

Now viewed from the ground, the duration of the event is $T = T_1 + T_2$, where $T_1 > T_2$ are respectively the time needed for the light to travel forward, among others for beating v, and the time for the light to travel back, helped this time by v. More precisely, if l denotes the length of the train car viewed from the ground, the formula of T is:

$$T = T_1 + T_2 = \frac{l}{c - v} + \frac{l}{c + v} = \frac{2lc}{c^2 - v^2}$$

With this data, the formula $T = \gamma t$ of time dilation established before reads:

$$\frac{2lc}{c^2 - v^2} = \gamma t = \frac{2\gamma L}{c}$$

Thus, the two lengths L and l are indeed not equal, and related by:

$$l = \frac{\gamma L(c^2 - v^2)}{c^2} = \gamma L\left(1 - \frac{v^2}{c^2}\right) = \frac{\gamma L}{\gamma^2} = \frac{L}{\gamma}$$

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

10d. Curved spacetime

With the above discussed, time now to get into the real thing, namely happens to our usual $\mathbb{R}^4$. The result here, which is something quite tricky, is as follows:

THEOREM 10.23. *In the context of a relativistic object moving with speed v along the x axis, the frame change is given by the Lorentz transformation*

$$x' = \gamma(x - vt)$$

$$y' = y$$

$$z' = z$$

$$t' = \gamma(t - vx/c^2)$$

with $\gamma = 1/\sqrt{1 - v^2/c^2}$ being as usual the Lorentz factor.

PROOF. We know that, with respect to the non-relativistic formulae, x is subject to the Lorentz dilation by γ, and we obtain as desired:

$$x' = \gamma(x - vt)$$

Regarding y, z, these are obviously unchanged, so done with these too. Finally, for t we must use the reverse Lorentz transformation, given by the following formulae:

$$x = \gamma(x' + vt')$$

$$y = y'$$

$$z = z'$$

By using the formula of x' we can compute t', and we obtain the following formula:

$$\begin{aligned} t' &= \frac{x - \gamma x'}{\gamma v} \\ &= \frac{x - \gamma^2(x - vt)}{\gamma v} \\ &= \frac{\gamma^2 vt + (1 - \gamma^2)x}{\gamma v} \end{aligned}$$

On the other hand, we have the following computation:

$$\gamma^2 = \frac{c^2}{c^2 - v^2} \implies \gamma^2(c^2 - v^2) = c^2 \implies (\gamma^2 - 1)c^2 = \gamma^2 v^2$$

Thus we can finish the computation of t' as follows:

$$\begin{aligned} t' &= \frac{\gamma^2 vt + (1-\gamma^2)x}{\gamma v} \\ &= \frac{\gamma^2 vt - \gamma^2 v^2 x/c^2}{\gamma v} \\ &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned}$$

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

Now since y, z are irrelevant, we can put them at the end, and put the time t first, as to be close to x. By multiplying as well the time equation by c, our system becomes:

$$\begin{aligned} ct' &= \gamma(ct - vx/c) \\ x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \end{aligned}$$

In linear algebra terms, the result is as follows:

THEOREM 10.24. *The Lorentz transformation is given by*

$$\begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix}$$

where $\gamma = 1/\sqrt{1 - v^2/c^2}$ *as usual, and where* $\beta = v/c$.

PROOF. In terms of $\beta = v/c$, replacing v, the system looks as follows:

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ x' &= \gamma(x - \beta ct) \\ y' &= y \\ z' &= z \end{aligned}$$

But this gives the formula in the statement. □

As a nice and basic theoretical application of the Lorentz transform, this brings a new viewpoint on the Einstein speed addition formula, the result being follows:

THEOREM 10.25. *The speed addition formula in* $3D$ *relativity is*

$$u +_e v = \frac{1}{1 + < u, v >}\left(u + v + \frac{u \times (u \times v)}{1 + \sqrt{1 - ||u||^2}}\right)$$

in $c = 1$ *units.*

PROOF. We already know this, but the point is that we can derive this as well from the formula of the Lorentz transform, by computing some derivatives, as follows:

(1) The idea will be that of differentiating x, y, z, t in the formulae for the inverse Lorentz transform, which are as follows:

$$x = \gamma(x' + ut')$$
$$y = y'$$
$$z = z'$$
$$t = \gamma(t' + ux'/c^2)$$

(2) Indeed, by differentiating these equalities, we obtain the following formulae:

$$dx = \gamma(dx' + udt')$$
$$dy = dy'$$
$$dz = dz'$$
$$dt = \gamma(dt' + udx'/c^2)$$

(3) Now by dividing the first three formulae by the fourth one, we obtain:

$$\frac{dx}{dt} = \frac{dx' + udt'}{dt' + udx'/c^2}$$
$$\frac{dy}{dt} = \frac{dy'}{\gamma(dt' + udx'/c^2)}$$
$$\frac{dz}{dt} = \frac{dz'}{\gamma(dt' + udx'/c^2)}$$

(4) We can make these look better by dividing everywhere by dt', and we get:

$$\frac{dx}{dt} = \frac{dx'/dt' + u}{1 + u/c^2 \cdot dx'/dt'}$$
$$\frac{dy}{dt} = \frac{dy'/dt'}{\gamma(1 + u/c^2 \cdot dx'/dt')}$$
$$\frac{dz}{dt} = \frac{dz'/dt'}{\gamma(1 + u/c^2 \cdot dx'/dt')}$$

(5) In terms of speeds now, this means that we have, with $w = u +_e v$:

$$w_x = \frac{v_x + u}{1 + u/c^2 \cdot v_x}$$
$$w_y = \frac{v_y}{\gamma(1 + u/c^2 \cdot v_x)}$$
$$w_z = \frac{v_z}{\gamma(1 + u/c^2 \cdot v_x)}$$

(6) Now in $c = 1$ units, these formulae are as follows, with $w = u +_e v$:

$$w_x = \frac{v_x + u}{1 + uv_x}$$

$$w_y = \frac{v_y}{\gamma(1 + uv_x)}$$

$$w_z = \frac{v_z}{\gamma(1 + uv_x)}$$

(7) In vector notation now, the above formulae show that we have:

$$\begin{aligned} u +_e v &= \frac{1}{1 + uv_x} \begin{pmatrix} v_x + u \\ v_y/\gamma \\ v_z/\gamma \end{pmatrix} \\ &= \frac{1}{1 + < u, v >} \left(u + \begin{pmatrix} v_x \\ v_y/\gamma \\ v_z/\gamma \end{pmatrix} \right) \end{aligned}$$

(8) On the other hand, we have the following computation:

$$\begin{aligned} u \times (u \times v) &= \begin{pmatrix} u \\ 0 \\ 0 \end{pmatrix} \times \left[\begin{pmatrix} u \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \right] \\ &= \begin{pmatrix} u \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ -uv_z \\ uv_y \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ -u^2 v_y \\ -u^2 v_z \end{pmatrix} \end{aligned}$$

(9) We deduce from this that we have the following formula:

$$\begin{aligned} v + \frac{u \times (u \times v)}{1 + \sqrt{1 - u^2}} &= \begin{pmatrix} v_x \\ v_y(1 - u^2/(1 + \sqrt{1 - u^2})) \\ v_z(1 - u^2/(1 + \sqrt{1 - u^2})) \end{pmatrix} \\ &= \begin{pmatrix} v_x \\ v_y\sqrt{1 - u^2} \\ v_z\sqrt{1 - u^2} \end{pmatrix} \\ &= \begin{pmatrix} v_x \\ v_y/\gamma \\ v_z/\gamma \end{pmatrix} \end{aligned}$$

(10) Here we have used the following identity, which is something trivial:

$$1-\frac{u^2}{1+\sqrt{1-u^2}}=\sqrt{1-u^2}$$

(11) The point now is that the formula from (9) shows that the speed addition formula established in (7) can be written in the following way:

$$u+_e v=\frac{1}{1+<u,v>}\left(u+v+\frac{u\times(u\times v)}{1+\sqrt{1-||u||^2}}\right)$$

Summarizing, we have recovered the formula for speed addition in relativity, from before, in our present configuration, with u assumed to be along the Ox axis.

(12) Finally, there is a discussion for passing from the standard configuration investigated above, where the movement is along the Ox axis, to the general configuration, where the movement is arbitary. But this can be done either by decomposing one speed vector with respect to the other, or simply by arguing that everything is rotation invariant. □

Many other things can be said, as a continuation of the above. For the moment, what we have will do, with the conclusion being that, when it comes to high speeds, $v\simeq c$, spacetime is curved. So, let us try now to better understand this phenomenon.

To start with, recall that in the non-relativistic setting two events are separated by space Δx and time Δt, with these two separation variables being independent. In relativistic physics this is no longer true, and the correct analogue of this comes from:

THEOREM 10.26. *The following quantity, called relativistic spacetime separation*

$$\Delta s^2=c^2\Delta t^2-(\Delta x^2+\Delta y^2+\Delta z^2)$$

is invariant under relativistic frame changes.

PROOF. This is something important, and as before with such things, we will take our time, and carefully understand how this result works:

(1) Let us first examine the case of the standard configuration. We must prove that the quantity $K=c^2t^2-x^2-y^2-z^2$ is invariant under Lorentz transformations, in the standard configuration. For this purpose, observe that we have:

$$K=\left\langle\begin{pmatrix}1&0&0&0\\0&-1&0&0\\0&0&-1&0\\0&0&0&-1\end{pmatrix}\begin{pmatrix}ct\\x\\y\\z\end{pmatrix},\begin{pmatrix}ct\\x\\y\\z\end{pmatrix}\right\rangle$$

Now recall that the Lorentz transformation is given in standard configuration by the following formula, where $\gamma = 1/\sqrt{1 - v^2/c^2}$ as usual, and where $\beta = v/c$:

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix}$$

Thus, if we denote by L the matrix of the Lorentz transformation, and by E the matrix found before, we must prove that for any vector ξ we have:

$$< E\xi, \xi >=< EL\xi, L\xi >$$

Since the matrix L is symmetric, we have the following formula:

$$< EL\xi, L\xi >=< LEL\xi, \xi >$$

Thus, we must prove that the following happens:

$$E = LEL$$

But this is the same as proving that we have the following equality:

$$L^{-1}E = EL$$

Moreover, by using the fact that the passage $L \to L^{-1}$ is given by $\beta \to -\beta$, what we eventually want to prove is that:

$$L_{-\beta}E = EL_\beta$$

So, let us prove this. As usual we can restrict the attention to the upper left corner, call that NW corner, and here we have the following computation:

$$\begin{aligned} (L_{-\beta}E)_{NW} &= \begin{pmatrix} \gamma & \beta\gamma \\ \beta\gamma & \gamma \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \\ &= \begin{pmatrix} \gamma & -\beta\gamma \\ \beta\gamma & -\gamma \end{pmatrix} \end{aligned}$$

On the other hand, we have as well the following computation:

$$\begin{aligned} (EL_\beta)_{NW} &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \gamma & -\beta\gamma \\ -\beta\gamma & \gamma \end{pmatrix} \\ &= \begin{pmatrix} \gamma & -\beta\gamma \\ \beta\gamma & -\gamma \end{pmatrix} \end{aligned}$$

The matrices on the right being equal, this gives the result.

(2) Now let us prove the invariance in general. The matrix of the Lorentz transformation being a bit complicated, in this case, as explained in the above, the best is to use for

this the raw formulae of x', t', that we found in the above, namely:

$$x' = x + (\gamma - 1)\frac{< v, x > v}{||v||^2} - \gamma t v$$

$$t' = \gamma\left(t - \frac{< v, x >}{c^2}\right)$$

With these formulae in hand, we have the following computation:

$$\begin{aligned}
& (ct')^2 - ||x'||^2 \\
= \ & \gamma^2\left(ct - \frac{< v, x >}{c}\right)^2 \\
& -||x||^2 - (\gamma - 1)^2\frac{< v, x >^2}{||v||^2} - \gamma^2 t^2||v||^2 \\
& -2(\gamma - 1)\frac{< v, x >^2}{||v||^2} + 2\gamma t < v, x > +2\gamma(\gamma - 1)t < v, x > \\
= \ & \gamma^2 t^2(c^2 - ||v||^2) - ||x||^2 \\
& + < v, x > (-2\gamma^2 t + 2\gamma t + 2\gamma(\gamma - 1)t) \\
& + < v, x >^2 \left(\frac{\gamma^2}{c^2} - \frac{(\gamma - 1)^2}{||v||^2} - \frac{2(\gamma - 1)}{||v||^2}\right) \\
= \ & c^2 t^2 - ||x||^2 + < v, x >^2 \left(\frac{\gamma^2}{c^2} - \frac{\gamma^2 - 1}{||v||^2}\right) \\
= \ & c^2 t^2 - ||x||^2
\end{aligned}$$

Here we have used the following trivial formula, for the coefficient of t^2:

$$\gamma^2(c^2 - ||v||^2) = \frac{c^2 - ||v||^2}{1 - ||v||^2/c^2} = c^2$$

Also, we have used the following formula, for the coefficient of $< v, x >^2$:

$$\begin{aligned}
\frac{\gamma^2}{c^2} - \frac{\gamma^2 - 1}{||v||^2} &= \gamma^2\left(\frac{1}{c^2} - \frac{1}{||v||^2}\right) + \frac{1}{||v||^2} \\
&= \frac{1}{1 - ||v||^2/c^2} \cdot \frac{||v||^2 - c^2}{c^2||v||^2} + \frac{1}{||v||^2} \\
&= \frac{c^2}{c^2 - ||v||^2} \cdot \frac{||v||^2 - c^2}{c^2||v||^2} + \frac{1}{||v||^2} \\
&= -\frac{1}{||v||^2} + \frac{1}{||v||^2} \\
&= 0
\end{aligned}$$

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

In relation with the above, it is possible to do some sort of reverse engineering, by recovering the formula of the Lorentz transform, from the spacetime separation invariance. Thus, we are led in this way to yet another axiomatization of the theory.

Finally, what we have in Theorem 10.26 suggests that curved spacetime is some sort of Riemannian manifold, that is, a smooth manifold whose tangent spaces come with scalar products, but with a $-$ sign appearing in the various formulae, instead of a $+$.

This is indeed true, with such manifolds being called "Lorentz manifolds", and with all this being related to everything algebraic and differential geometry, and linear algebra positivity and negativity too, that we learned in the beginning of this chapter.

Which sounds quite exciting, doesn't it. There are many things that can be learned here, both math and physics, and in the hope that you will go this way, and enjoy it.

10e. Exercises

This was an introduction to modern math and physics, and as exercises, we have:

EXERCISE 10.27. *Have some fun with learning about plane algebraic curves.*

EXERCISE 10.28. *Learn some commutative algebra, notably the Nullstellensatz.*

EXERCISE 10.29. *Learn how to compute lengths of curves, and areas of surfaces.*

EXERCISE 10.30. *Learn the formal definition and basic properties of smooth manifolds.*

EXERCISE 10.31. *Learn also about Riemannian manifolds, as much as you can.*

EXERCISE 10.32. *Still in relation with geometry, learn about knots and links, too.*

EXERCISE 10.33. *Recover the Einstein speed summation in $3D$, by yourself.*

EXERCISE 10.34. *Meditate on the $c \pm v$ speeds used for the Lorentz contraction.*

As bonus exercise, that we warmly recommend, read Einstein's book [26].

CHAPTER 11

Special matrices

11a. Circulant matrices

Back now to more traditional linear algebra, we discuss in this chapter various classes of "special matrices". As a first and central example here, we have the flat matrix:

DEFINITION 11.1. *The flat matrix $\mathbb{I}_N$ is the all-one $N \times N$ matrix:*

$$\mathbb{I}_N = \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$$

Equivalently, $\mathbb{I}_N/N$ is the orthogonal projection on the all-one vector $\xi \in \mathbb{C}^N$.

A first interesting question regarding $\mathbb{I}_N$ concerns its diagonalization. As explained in chapter 1, this is best solved by using complex numbers, in the following way:

THEOREM 11.2. *The flat matrix $\mathbb{I}_N$ diagonalizes as follows,*

$$\begin{pmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \ldots & \ldots & 1 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} F_N^*$$

where $F_N = (w^{ij})_{ij}$ with $w = e^{2\pi i/N}$ is the Fourier matrix.

PROOF. The flat matrix being N times the projection on the all-one vector, we are left with finding the 0-eigenvectors, which amounts in solving the following equation:

$$x_0 + \ldots + x_{N-1} = 0$$

But for this purpose, we use the root of unity $w = e^{2\pi i/N}$, which is subject to:

$$\sum_{i=0}^{N-1} w^{ij} = N\delta_{j0}$$

Indeed, this formula shows that for $j = 1, \ldots, N-1$, the vector $v_j = (w^{ij})_i$ is a 0-eigenvector. Moreover, these vectors are pairwise orthogonal, because we have:

$$< v_j, v_k >= \sum_i w^{ij-ik} = N\delta_{jk}$$

Thus, we have our basis $\{v_1, \ldots, v_{N-1}\}$ of 0-eigenvectors, and since the N-eigenvector is $\xi = v_0$, the passage matrix P that we are looking is given by:

$$P = \begin{bmatrix} v_0 & v_1 & \ldots & v_{N-1} \end{bmatrix}$$

But this is precisely the Fourier matrix, $P = F_N$. In order to finish now, observe that the above computation of $< v_i, v_j >$ shows that $F_N/\sqrt{N}$ is unitary, and so:

$$F_N^{-1} = \frac{1}{N} F_N^*$$

Thus, we are led to the diagonalization formula in the statement. □

Now observe that the flat matrix $\mathbb{I}_N$ is circulant, in the following sense:

DEFINITION 11.3. *A real or complex matrix M is called circulant if*

$$M_{ij} = \xi_{j-i}$$

for a certain vector ξ, with the indices taken modulo N.

The circulant matrices are beautiful mathematical objects, which appear of course in many serious problems as well. As an example, at $N = 4$, we must have:

$$M = \begin{pmatrix} a & b & c & d \\ d & a & b & c \\ c & d & a & b \\ b & c & d & a \end{pmatrix}$$

The point now is that, while certainly gently looking, these matrices can be quite diabolic, when it comes to diagonalization, and other problems. For instance, when M is real, the computations with M are usually quite complicated over the real numbers. Fortunately the complex numbers and the Fourier matrices are there, and we have:

THEOREM 11.4. *For a matrix $M \in M_N(\mathbb{C})$, the following are equivalent:*

(1) *M is circulant, in the sense that we have, for a certain vector $\xi \in \mathbb{C}^N$:*

$$M_{ij} = \xi_{j-i}$$

(2) *M is Fourier-diagonal, in the sense that, for a certain diagonal matrix Q:*

$$M = F_N Q F_N^*$$

In addition, if these conditions hold, then ξ, Q are related by the formula

$$\xi = F_N^* q$$

where $q \in \mathbb{C}^N$ is the column vector formed by the diagonal entries of Q.

PROOF. This follows from some basic computations with roots of unity, as follows:

(1) $\implies$ (2) Assuming $M_{ij} = \xi_{j-i}$, the matrix $Q = F_N^* M F_N$ is indeed diagonal, as shown by the following computation:

$$\begin{aligned} Q_{ij} &= \sum_{kl} w^{-ik} M_{kl} w^{lj} \\ &= \sum_{kl} w^{jl-ik} \xi_{l-k} \\ &= \sum_{kr} w^{j(k+r)-ik} \xi_r \\ &= \sum_r w^{jr} \xi_r \sum_k w^{(j-i)k} \\ &= N\delta_{ij} \sum_r w^{jr} \xi_r \end{aligned}$$

(2) $\implies$ (1) Assuming $Q = diag(q_1, \ldots, q_N)$, the matrix $M = F_N Q F_N^*$ is indeed circulant, as shown by the following computation:

$$M_{ij} = \sum_k w^{ik} Q_{kk} w^{-jk} = \sum_k w^{(i-j)k} q_k$$

To be more precise, in this formula the last term depends only on $j - i$, and so shows that we have $M_{ij} = \xi_{j-i}$, with ξ being the following vector:

$$\xi_i = \sum_k w^{-ik} q_k = (F_N^* q)_i$$

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

As a basic illustration for the above result, for the circulant matrix $M = \mathbb{I}_N$ we recover in this way the diagonalization result from Theorem 11.1, namely:

$$\begin{pmatrix} 1 & \ldots & \ldots & 1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ 1 & \ldots & \ldots & 1 \end{pmatrix} = \frac{1}{N} F_N \begin{pmatrix} N & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} F_N^*$$

There are many other examples, as for instance those coming from the adjacency matrices of the circulant graphs, and with the remark that, up to a copy of 1_N, the flat matrix $\mathbb{I}_N$ is indeed of this type. We will leave some exploration here as an exercise.

The above result is something quite powerful, and very useful, and suggests doing everything in Fourier, when dealing with circulant matrices. And we can use here:

THEOREM 11.5. *The various basic sets of $N \times N$ circulant matrices are as follows, with the convention that associated to any $q \in \mathbb{C}^N$ is the matrix $Q = diag(q_1, \ldots, q_N)$:*

(1) *The set of all circulant matrices is:*

$$M_N(\mathbb{C})^{circ} = \left\{ F_N Q F_N^* \middle| q \in \mathbb{C}^N \right\}$$

(2) *The set of all circulant unitary matrices is:*

$$U_N^{circ} = \left\{ \frac{1}{N} F_N Q F_N^* \middle| q \in \mathbb{T}^N \right\}$$

(3) *The set of all circulant orthogonal matrices is:*

$$O_N^{circ} = \left\{ \frac{1}{N} F_N Q F_N^* \middle| q \in \mathbb{T}^N, \bar{q}_i = q_{-i}, \forall i \right\}$$

In addition, in this picture, the first row vector of $F_N Q F_N^*$ is given by $\xi = F_N^* q$.

PROOF. All this follows from Theorem 11.4, as follows:

(1) This assertion, along with the last one, is Theorem 11.4 itself.

(2) This is clear from (1), and from the fact that the rescaled matrix $F_N/\sqrt{N}$ is unitary, because the eigenvalues of a unitary matrix must be on the unit circle $\mathbb{T}$.

(3) This follows from (2), because the matrix is real when $\xi_i = \bar{\xi}_i$, and in Fourier transform, $\xi = F_N^* q$, this corresponds to the condition $\bar{q}_i = q_{-i}$. □

There are many other interesting things that can be said about the circulant matrices, along the above lines. Importantly, all this can be generalized to the matrices which are $(N_1, \ldots, N_k)$ patterned, with the matrix doing the job here being as follows:

$$F_{N_1,\ldots,N_k} = F_{N_1} \otimes \ldots \otimes F_{N_k}$$

To be more precise here, in order to have some better notations and formalism, consider a finite abelian group G, decomposed as $G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_k}$. We can then talk about the corresponding Fourier matrix F_G, which coincides with the above one:

$$F_G = F_{N_1} \otimes \ldots \otimes F_{N_k}$$

We will discuss more in detail such matrices, which are called generalized Fourier matrices, and are basic examples of complex Hadamard matrices, in a moment. In the meantime, we have the following result, which is standard in discrete Fourier analysis, extending what we previously knew from the above, in the circulant case:

THEOREM 11.6. *For a matrix $A \in M_N(\mathbb{C})$, the following are equivalent,*

(1) *A is G-invariant, $A_{ij} = \xi_{j-i}$, for a certain vector $\xi \in \mathbb{C}^N$,*

(2) *A is Fourier-diagonal, $A = F_G Q F_G^*$, for a certain diagonal matrix Q,*

and if so, $\xi = F_G^* q$, where $q \in \mathbb{C}^N$ is the vector formed by the diagonal entries of Q.

PROOF. This is something that we know from the above in the cyclic case, $G = \mathbb{Z}_N$, and the proof in general is similar, by using matrix indices as follows:

$$i, j \in G$$

To be more precise, in order to get started, with our generalization, let us decompose our finite abelian group G as a product of cyclic groups, as follows:

$$G = \mathbb{Z}_{N_1} \times \ldots \times \mathbb{Z}_{N_s}$$

The corresponding Fourier matrix decomposes then as well, as follows:

$$F_G = F_{N_1} \otimes \ldots \otimes F_{N_s}$$

Now if we set $w_i = e^{2\pi i/N_i}$, this means that we have the following formula:

$$(F_G)_{ij} = w_1^{i_1 j_1} \ldots w_s^{i_s j_s}$$

We can now prove the equivalence in the statement, as follows:

(1) $\implies$ (2) Assuming $A_{ij} = \xi_{j-i}$, the matrix $Q = F_G^* A F_G$ is diagonal, as shown by the following computation, with all indices being group elements:

$$\begin{aligned}
Q_{ij} &= \sum_{kl} \overline{(F_G)}_{ki} A_{kl} (F_G)_{lj} \\
&= \sum_{kl} w_1^{-k_1 i_1} \ldots w_s^{-k_s i_s} \cdot \xi_{l-k} \cdot w_1^{l_1 j_1} \ldots w_s^{l_s j_s} \\
&= \sum_{kl} w_1^{l_1 j_1 - k_1 i_1} \ldots w_s^{l_s j_s - k_s i_s} \xi_{l-k} \\
&= \sum_{kr} w_1^{(k_1+r_1) j_1 - k_1 i_1} \ldots w_s^{(k_s+r_s) j_s - k_s i_s} \xi_r \\
&= \sum_{r} w_1^{r_1 j_1} \ldots w_s^{r_s j_s} \xi_r \sum_{k} w_1^{k_1 (j_1 - i_1)} \ldots w_s^{k_s (j_s - i_s)} \\
&= \sum_{r} w_1^{r_1 j_1} \ldots w_s^{r_s j_s} \xi_r \cdot N_1 \delta_{i_1 j_1} \ldots N_s \delta_{i_s j_s} \\
&= N \delta_{ij} \sum_{r} (F_G)_{jr} \xi_r
\end{aligned}$$

(2) $\implies$ (1) Assuming $Q = diag(q_1, \ldots, q_N)$, the matrix $A = F_G Q F_G^*$ is G-invariant, as shown by the following computation, again with all indices being group elements:

$$\begin{aligned} A_{ij} &= \sum_{kl} (F_G)_{ik} Q_{kk} \overline{(F_G)}_{kj} \\ &= \sum_{k} w_1^{i_1 k_1} \ldots w_s^{i_s k_s} \cdot q_k \cdot w_1^{-j_1 k_1} \ldots w_s^{-j_s k_s} \\ &= \sum_{k} w_1^{(i_1 - j_1) k_1} \ldots w_s^{(i_s - j_s) k_s} q_k \end{aligned}$$

To be more precise, in this formula the last term depends only on $j - i$, and so shows that we have $A_{ij} = \xi_{j-i}$, with ξ being the following vector:

$$\begin{aligned} \xi_i &= \sum_{k} w_1^{-i_1 k_1} \ldots w_s^{-i_s k_s} q_k \\ &= \sum_{k} (F_G^*)_{ik} q_k \\ &= (F_G^* q)_i \end{aligned}$$

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

As before with the circulant matrices, there are many illustrations for the above result, as for instance those coming from the adjacency matrices of the graphs having a transitive action of an abelian group. We will leave some exploration here as an exercise.

11b. Hadamard matrices

Switching topics, but somehow still in relation with the Fourier matrices, taken in a generalized sense, we would like to discuss now the Hadamard matrices. Let us begin with the real Hadamard matrix case, which is very old, and very interesting too:

DEFINITION 11.7. *An Hadamard matrix is a square binary matrix*

$$H \in M_N(\pm 1)$$

whose rows are pairwise orthogonal.

Here is a basic example of such a matrix, called first Walsh matrix:

$$W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

This matrix is quite trivial, of size 2×2, but by taking tensor powers of it, we have as examples the higher Walsh matrices as well, having size $2^k \times 2^k$, given by:

$$W_{2^k} = W_2^{\otimes k}$$

Observe that the first Walsh matrix W_2 is in fact the Fourier matrix of $\mathbb{Z}_2$, and that, more generally, the Walsh matrix W_{2^k} is the Fourier matrix of $\mathbb{Z}_2^k$. We will be back to this, group-theoretical aspects of the Hadamard matrices, later in this chapter.

Getting back to Definition 11.7 as stated, we have seen that we have examples of such matrices at any $N = 2^k$. But, what happens then in arbitrary size $N \times N$? It is clear that we must have $2|N$, and along the same lines, it is easy to see, by playing around with the first rows, that once your matrix has $N \geq 3$ rows, we must have $4|N$:

PROPOSITION 11.8. *The size of an Hadamard matrix $H \in M_N(\pm 1)$ must satisfy*

$$N \in \{2\} \cup 4\mathbb{N}$$

with this coming from the orthogonality condition between the first 3 rows.

PROOF. This is something that we already met, in chapter 2, the idea being that by permuting the rows and columns or by multiplying them by -1, as to rearrange the first 3 rows, we can always assume that our matrix looks as follows:

$$H = \begin{pmatrix} 1 \ldots\ldots 1 & 1 \ldots\ldots 1 & 1 \ldots\ldots 1 & 1 \ldots\ldots 1 \\ 1 \ldots\ldots 1 & 1 \ldots\ldots 1 & -1 \ldots -1 & -1 \ldots -1 \\ 1 \ldots\ldots 1 & -1 \ldots -1 & 1 \ldots\ldots 1 & -1 \ldots -1 \\ \underbrace{\cdots\cdots\cdots}_{x} & \underbrace{\cdots\cdots\cdots}_{y} & \underbrace{\cdots\cdots\cdots}_{z} & \underbrace{\cdots\cdots\cdots}_{t} \end{pmatrix}$$

Now if we denote by x, y, z, t the sizes of the block columns, as indicated, the orthogonality conditions between the first 3 rows give $x = y = z = t$. Thus the matrix size $N = x + y + z + t$ must be a multiple of 4, as claimed. □

The above result is something quite interesting, and the point is that a similar analysis with 4 rows or more does not give any further restriction on the possible values of the size $N \in \mathbb{N}$. In fact, we are led in this way to the following famous conjecture:

CONJECTURE 11.9 (Hadamard). *There is an Hadamard matrix of order N,*

$$H \in M_N(\pm 1)$$

for any $N \in 4\mathbb{N}$.

Normally this is an analytic question, because in practice the number of Hadamard matrices grows exponentially with N, and so in order to prove the conjecture, you just need a modest lower estimate on this number. But, no one knows how to do this, and this despite the Hadamard conjecture being open for more than 100 years.

This being said, what we can do with our algebraic methods is to verify at least the Hadamard conjecture at small values of $N \in 4\mathbb{N}$. And here, with $N = 4, 8$ being solved by the Walsh matrices, we are faced with constructing a matrix at $N = 12$.

In order to solve this question, let $q = p^k$ be an odd prime power, and set:

$$\chi(a) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{if } a = b^2, b \neq 0 \\ -1 & \text{otherwise} \end{cases}$$

Then set $Q_{ab} = \chi(b - a)$, with indices in $\mathbb{F}_q$. With these conventions, the Paley construction of Hadamard matrices, which works well at $N = 12$, is as follows:

THEOREM 11.10. *Given an odd prime power $q = p^k$, construct $Q_{ab} = \chi(b - a)$ as above. We have then constructions of Hadamard matrices, as follows:*

(1) *Paley 1: if $q = 3(4)$ we have a matrix of size $N = q + 1$, as follows:*

$$P_N^1 = 1 + \begin{pmatrix} 0 & 1 & \ldots & 1 \\ -1 & & & \\ \vdots & & Q & \\ -1 & & & \end{pmatrix}$$

(2) *Paley 2: if $q = 1(4)$ we have a matrix of size $N = 2q + 2$, as follows:*

$$P_N^2 = \begin{pmatrix} 0 & 1 & \ldots & 1 \\ 1 & & & \\ \vdots & & Q & \\ 1 & & & \end{pmatrix} \quad : \quad 0 \to \begin{pmatrix} 1 & -1 \\ -1 & -1 \end{pmatrix} \quad , \quad \pm 1 \to \pm \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

These matrices are skew-symmetric ($H + H^t = 2$), respectively symmetric ($H = H^t$).

PROOF. In order to simplify the presentation, we denote by 1 all the identity matrices, of any size, and by $\mathbb{I}$ all the rectangular all-one matrices, of any size as well. It is elementary to check that the matrix $Q_{ab} = \chi(a - b)$ has the following properties:

$$QQ^t = q1 - \mathbb{I} \quad , \quad Q\mathbb{I} = \mathbb{I}Q = 0$$

In addition, we have the following formulae, which are elementary as well, coming from the fact that -1 is a square in $\mathbb{F}_q$ precisely when $q = 1(4)$:

$$q = 1(4) \implies Q = Q^t \quad , \quad q = 3(4) \implies Q = -Q^t$$

With these observations in hand, the proof goes as follows:

(1) With our above conventions for 1 and $\mathbb{I}$, the matrix in the statement is:

$$P_N^1 = \begin{pmatrix} 1 & \mathbb{I} \\ -\mathbb{I} & 1 + Q \end{pmatrix}$$

With this formula in hand, the Hadamard matrix condition follows from:

$$\begin{aligned} P_N^1(P_N^1)^t &= \begin{pmatrix} 1 & \mathbb{I} \\ -\mathbb{I} & 1+Q \end{pmatrix} \begin{pmatrix} 1 & -\mathbb{I} \\ \mathbb{I} & 1-Q \end{pmatrix} \\ &= \begin{pmatrix} N & 0 \\ 0 & \mathbb{I}+1-Q^2 \end{pmatrix} \\ &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix} \end{aligned}$$

(2) If we denote by G, F the 2×2 matrices in the statement, which replace respectively the $0, 1$ entries, then we have the following formula for our matrix:

$$P_N^2 = \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes F + 1 \otimes G$$

With this formula in hand, the Hadamard matrix condition follows from:

$$\begin{aligned} (P_N^2)^2 &= \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix}^2 \otimes F^2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes G^2 + \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes (FG+GF) \\ &= \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} \otimes 2 + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes 2 + \begin{pmatrix} 0 & \mathbb{I} \\ \mathbb{I} & Q \end{pmatrix} \otimes 0 \\ &= \begin{pmatrix} N & 0 \\ 0 & N \end{pmatrix} \end{aligned}$$

Finally, the last assertion is clear, from the above formulae relating Q, Q^t. □

In practice now, by using the Walsh and Paley constructions, the next problem when verifying the Hadamard Conjecture appears at $N = 92$. But here, we have:

THEOREM 11.11. *Assuming that $A, B, C, D \in M_K(\pm 1)$ are circulant, symmetric, pairwise commute and satisfy the condition*

$$A^2 + B^2 + C^2 + D^2 = 4K$$

the following $4K \times 4K$ matrix is Hadamard, called of Williamson type:

$$H = \begin{pmatrix} A & B & C & D \\ -B & A & -D & C \\ -C & D & A & -B \\ -D & -C & B & A \end{pmatrix}$$

Moreover, matrices A, B, C, D as above exist at $K = 23$, where $4K = 92$.

PROOF. Consider the standard quaternion units $1, i, j, k \in M_4(0, 1)$. These matrices are by definition those describing the positions of the A, B, C, D entries in the matrix H from the statement, and so this matrix can be written as follows:

$$H = A \otimes 1 + B \otimes i + C \otimes j + D \otimes k$$

Assuming now that A, B, C, D are symmetric, we have:

$$\begin{aligned} HH^t &= (A\otimes 1+B\otimes i+C\otimes j+D\otimes k)\\ &\quad (A\otimes 1-B\otimes i-C\otimes j-D\otimes k)\\ &= (A^2+B^2+C^2+D^2)\otimes 1-([A,B]-[C,D])\otimes i\\ &\quad -([A,C]-[B,D])\otimes j-([A,D]-[B,C])\otimes k \end{aligned}$$

Now assume that our matrices A, B, C, D pairwise commute, and satisfy the condition in the statement. In this case, it follows from the above formula that we have:

$$HH^t=4K$$

Thus, we obtain indeed an Hadamard matrix, as claimed. However, finding such matrices is in general a difficult task, and this is where Williamson's extra assumption in the statement, that A, B, C, D should be taken circulant, comes from. Finally, regarding the $K = 23$ and $N = 92$ example, this comes via a computer search. □

At higher N things become more technical, and more complicated constructions, along the lines of those of Paley and Williamson, are needed. Quite curiously, as of now, early 21th century, the human knowledge stops at the number of the beast, namely:

$$\mathfrak{N}=666$$

That is, explicit examples of Hadamard matrices have been constructed for all multiples of four $N \leq 664$, but no such matrix is known so far at $N = 668$.

Switching topics now, another well-known open question concerns the circulant case. Given a binary vector $\gamma \in (\pm 1)^N$, one can ask whether the matrix $H \in M_N(\pm 1)$ defined by $H_{ij} = \gamma_{j-i}$ is Hadamard or not. Here is a solution to the problem:

$$K_4=\begin{pmatrix} -1 & 1 & 1 & 1\\ 1 & -1 & 1 & 1\\ 1 & 1 & -1 & 1\\ 1 & 1 & 1 & -1 \end{pmatrix}$$

More generally, any vector $\gamma \in (\pm 1)^4$ satisfying $\sum \gamma_i = \pm 1$ is a solution to the problem. The following conjecture, from the 50s, states that there are no other solutions:

Conjecture 11.12 (Circulant Hadamard Conjecture (CHC)). *The only Hadamard matrices which are circulant are*

$$K_4=\begin{pmatrix} -1 & 1 & 1 & 1\\ 1 & -1 & 1 & 1\\ 1 & 1 & -1 & 1\\ 1 & 1 & 1 & -1 \end{pmatrix}$$

and its conjugates, regardless of the value of $N \in \mathbb{N}$.

The fact that such a simple-looking problem is still open might seem quite surprising. Indeed, if we denote by $S \subset \{1, \ldots, N\}$ the set of positions of the -1 entries of γ, the Hadamard matrix condition is simply, for any $k \neq 0$, taken modulo N:

$$|S \cap (S + k)| = |S| - N/4$$

Thus, the above conjecture simply states that at $N \neq 4$, such a set S cannot exist. This is a well-known problem in combinatorics, raised by Ryser a long time ago.

11c. Complex Hadamard

We have seen that the theory of real Hadamard matrices leads to many difficult questions. As a further twist to the plot, bringing some sort of solution to this, we have:

THEOREM 11.13. *When enlarging the attention to the complex Hadamard matrices, $H \in M_N(\mathbb{T})$ having the rows pairwise orthogonal, the Fourier matrix,*

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & w & w^2 & \ldots & w^{N-1} \\ 1 & w^2 & w^4 & \ldots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \ldots & w^{(N-1)^2} \end{pmatrix}$$

with $w = e^{2\pi i/N}$, provides an example of such a matrix, at any $N \in \mathbb{N}$. Thus, the Hadamard Conjecture problematics dissapears, in the complex setting.

PROOF. We have seen before that the rescaling $U = F_N/\sqrt{N}$ is unitary. Thus the rows of U are pairwise orthogonal, and so follow to be the rows of F_N. □

In view of the above result, let us study more in detail the complex Hadamard matrices. Many examples can be constructed, quite often by using the combinatorics of roots of unity, and as a basic example here, we have the tensor products of Fourier matrices:

$$F_{N_1,\ldots,N_k} = F_{N_1} \otimes \ldots \otimes F_{N_k}$$

Moreover, we have as well deformations, and we are led in this way to:

THEOREM 11.14. *The only Hadamard matrices at $N = 4$ are, up to equivalence*

$$F_4^q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & q & -1 & -q \\ 1 & -q & -1 & q \end{pmatrix}$$

with $q \in \mathbb{T}$, which appear as suitable deformations of $W_4 = F_2 \otimes F_2$.

PROOF. This is something quite self-explanatory, and we will leave working out all this, namely finding the correct meaning of the equivalence relation, and of the deformation notion used, along of course with the proof, as an instructive exercise. □

Moving on, at $N = 5$ the situation is considerably more complicated, with F_5 being the only matrix. The key technical result here, due to Haagerup, is as follows:

PROPOSITION 11.15. *Given an Hadamard matrix $H \in M_5(\mathbb{T})$, chosen dephased,*

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & x & * & * \\ 1 & y & b & * & * \\ 1 & * & * & * & * \\ 1 & * & * & * & * \end{pmatrix}$$

the numbers a, b, x, y must satisfy the equation $(x - y)(x - ab)(y - ab) = 0$.

PROOF. Consider the upper 3-row truncation of H, which looks as follows:

$$H' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & x & p & q \\ 1 & y & b & r & s \end{pmatrix}$$

By using the orthogonality of the rows, we have:

$$(1 + a + x)(1 + \bar{b} + \bar{y})(1 + \bar{a}y + b\bar{x}) = -(p + q)(r + s)(\bar{p}r + \bar{q}s)$$

On the other hand, by using $p, q, r, s \in \mathbb{T}$, we have:

$$\begin{aligned} (p + q)(r + s)(\bar{p}r + \bar{q}s) &= (r + p\bar{q}s + \bar{p}qr + s)(\bar{r} + \bar{s}) \\ &= 1 + p\bar{q}\bar{r}s + \bar{p}q + \bar{r}s + r\bar{s} + p\bar{q} + \bar{p}qr\bar{s} + 1 \\ &= 2Re(1 + p\bar{q} + r\bar{s} + p\bar{q}r\bar{s}) \\ &= 2Re[(1 + p\bar{q})(1 + r\bar{s})] \end{aligned}$$

We conclude that we have the following formula, involving a, b, x, y only:

$$(1 + a + x)(1 + \bar{b} + \bar{y})(1 + \bar{a}y + b\bar{x}) \in \mathbb{R}$$

Now this is a product of type $(1 + \alpha)(1 + \beta)(1 + \gamma)$, with the first summand being 1, and with the last summand, namely $\alpha\beta\gamma$, being real as well, as shown by the above general $p, q, r, s \in \mathbb{T}$ computation. Thus, when expanding, and we are left with:

$$\begin{aligned} &(a + x) + (\bar{b} + \bar{y}) + (\bar{a}y + b\bar{x}) + (a + x)(\bar{b} + \bar{y}) \\ + \ &(a + x)(\bar{a}y + b\bar{x}) + (\bar{b} + \bar{y})(\bar{a}y + b\bar{x}) \in \mathbb{R} \end{aligned}$$

By expanding all the products, our formula looks as follows:

$$\begin{aligned} &a + x + \bar{b} + \bar{y} + \bar{a}y + b\bar{x} + a\bar{b} + a\bar{y} + \bar{b}x + x\bar{y} \\ + \ &1 + ab\bar{x} + \bar{a}xy + b + \bar{a}\bar{b}y + \bar{x} + \bar{a} + b\bar{x}\bar{y} \in \mathbb{R} \end{aligned}$$

By removing from this all terms of type $z + \bar{z}$, we are left with:

$$a\bar{b} + x\bar{y} + ab\bar{x} + \bar{a}\bar{b}y + \bar{a}xy + b\bar{x}\bar{y} \in \mathbb{R}$$

Now by getting back to our Hadamard matrix, all this remains true when transposing it, which amounts in interchanging $x \leftrightarrow y$. Thus, we have as well:

$$a\bar{b} + \bar{x}y + ab\bar{y} + \bar{a}\bar{b}x + \bar{a}xy + b\bar{x}\bar{y} \in \mathbb{R}$$

By substracting now the two equations that we have, we obtain:

$$x\bar{y} - \bar{x}y + ab(\bar{x} - \bar{y}) + \bar{a}\bar{b}(y - x) \in \mathbb{R}$$

Now observe that this number, say Z, is purely imaginary, because $\bar{Z} = -Z$. Thus our equation reads $Z = 0$. On the other hand, we have the following formula:

$$\begin{aligned} abxyZ &= abx^2 - aby^2 + a^2b^2(y - x) + xy(y - x) \\ &= (y - x)(a^2b^2 + xy - ab(x + y)) \\ &= (y - x)(ab - x)(ab - y) \end{aligned}$$

Thus, our equation $Z = 0$ corresponds to the formula in the statement. □

We are led in this way to the following theorem, due to Haagerup:

THEOREM 11.16. *The only Hadamard matrix at $N = 5$ is the Fourier matrix,*

$$F_5 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 & w^4 \\ 1 & w^2 & w^4 & w & w^3 \\ 1 & w^3 & w & w^4 & w^2 \\ 1 & w^4 & w^3 & w^2 & w \end{pmatrix}$$

with $w = e^{2\pi i/5}$, up to the standard equivalence relation for such matrices.

PROOF. Assume that have an Hadamard matrix $H \in M_5(\mathbb{T})$, chosen dephased, and written as in Proposition 11.15, with emphasis on the upper left 2×2 subcorner.

(1) We know from Proposition 11.15, applied to H itself, and to its transpose H^t as well, that the entries a, b, x, y must satisfy the following equations:

$$(a - b)(a - xy)(b - xy) = 0$$

$$(x - y)(x - ab)(y - ab) = 0$$

Our first claim is that, by doing some combinatorics, we can actually obtain from this $a = b$ and $x = y$, up to the equivalence relation for the Hadamard matrices:

$$H \sim \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & x & * & * \\ 1 & x & a & * & * \\ 1 & * & * & * & * \\ 1 & * & * & * & * \end{pmatrix}$$

Indeed, the above two equations lead to 9 possible cases, the first of which is, as desired, $a = b$ and $x = y$. As for the remaining 8 cases, here again things are determined

by 2 parameters, and in practice, we can always permute the first 3 rows and 3 columns, and then dephase our matrix, as for our matrix to take the above special form.

(2) With this result in hand, the combinatorics of the scalar products between the first 3 rows, and between the first 3 columns as well, becomes something which is quite simple to investigate. By doing a routine study here, and then completing it with a study of the lower right 2×2 corner as well, we are led to 2 possible cases, as follows:

$$H \sim \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & b & c & d \\ 1 & b & a & d & c \\ 1 & c & d & a & b \\ 1 & d & c & b & a \end{pmatrix} \quad , \quad H \sim \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & b & c & d \\ 1 & b & a & d & c \\ 1 & c & d & b & a \\ 1 & d & c & a & b \end{pmatrix}$$

(3) Our claim now is that the first case is in fact not possible. Indeed, we must have:

$$\begin{aligned} a+b+c+d &= -1 \\ 2Re(a\bar{b}) + 2Re(c\bar{d}) &= -1 \\ 2Re(a\bar{c}) + 2Re(b\bar{d}) &= -1 \\ 2Re(a\bar{d}) + 2Re(b\bar{c}) &= -1 \end{aligned}$$

Now since $|Re(x)| \leq 1$ for any $x \in \mathbb{T}$, we deduce from the second equation that:

$$Re(a\bar{b}) \leq 1/2$$

In other words, the arc length between a, b satisfies:

$$\theta(a, b) \geq \pi/3$$

The same argument applies to c, d, and to the other pairs of numbers in the last 2 equations. Now since our equations are invariant under permutations of a, b, c, d, we can assume that a, b, c, d are ordered in this way on the unit circle, and by the above, separated by $\geq \pi/3$ arc lengths. But this tells us that we have the following inequalities:

$$\theta(a, c) \geq 2\pi/3 \quad , \quad \theta(b, d) \geq 2\pi/3$$

These two inequalities give the following estimates:

$$Re(a\bar{c}) \leq -1/2 \quad , \quad Re(b\bar{d}) \leq -1/2$$

But these estimates contradict the third equation. Thus, our claim is proved.

(4) Summarizing, we have proved so far that our matrix must be as follows:

$$H \sim \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & a & b & c & d \\ 1 & b & a & d & c \\ 1 & c & d & b & a \\ 1 & d & c & a & b \end{pmatrix}$$

We are now in position of finishing. The orthogonality equations are as follows:

$$\begin{aligned} a+b+c+d &= -1\\ 2Re(a\bar{b})+2Re(c\bar{d}) &= -1\\ a\bar{c}+c\bar{b}+b\bar{d}+d\bar{a} &= -1 \end{aligned}$$

The third equation can be written in the following equivalent form:

$$\begin{aligned} Re[(a+b)(\bar{c}+\bar{d})] &= -1\\ Im[(a-b)(\bar{c}-\bar{d})] &= 0 \end{aligned}$$

By using now $a,b,c,d\in\mathbb{T}$, we obtain from this:

$$\frac{a+b}{a-b}\in i\mathbb{R}\quad,\quad \frac{c+d}{c-d}\in i\mathbb{R}$$

Thus we can find $s,t\in\mathbb{R}$ such that:

$$a+b=is(a-b)\quad,\quad c+d=it(c-d)$$

By plugging in these values, our system of equations simplifies, as follows:

$$\begin{aligned} (a+b)+(c+d) &= -1\\ |a+b|^2+|c+d|^2 &= 3\\ (a+b)(\bar{c}+\bar{d}) &= -1 \end{aligned}$$

Now observe that the last equation implies in particular that we have:

$$|a+b|^2\cdot|c+d|^2=1$$

Thus $|a+b|^2,|c+d|^2$ must be roots of the following polynomial:

$$X^2-3X+1=0$$

But this gives the following equality of sets:

$$\left\{|a+b|\,,\,|c+d|\right\}=\left\{\frac{\sqrt{5}+1}{2}\,,\,\frac{\sqrt{5}-1}{2}\right\}$$

This is good news, because we are now into 5-th roots of unity. To be more precise, we have 2 cases to be considered, the first one being as follows, with $z\in\mathbb{T}$:

$$a+b=\frac{\sqrt{5}+1}{2}\,z\quad,\quad c+d=-\frac{\sqrt{5}-1}{2}\,z$$

From $a+b+c+d=-1$ we obtain $z=-1$, and by using this we obtain $b=\bar{a}$, $d=\bar{c}$. Thus we have the following formulae:

$$Re(a)=\cos(2\pi/5)\quad,\quad Re(c)=\cos(\pi/5)$$

We conclude that we have $H\sim F_5$, as claimed. As for the second case, with a,b and c,d interchanged, this leads to $H\sim F_5$ as well. □

In view of the above, we are led to the question of finding the Hadamard matrices which are "isolated". Let us begin with some notations. We denote by X_p an unspecified neighborhood of a point in a manifold, $p \in X$. Also, for $q \in \mathbb{T}_1$, meaning that $q \in \mathbb{T}$ is close to 1, we define q^r with $r \in \mathbb{R}$ by $(e^{it})^r = e^{itr}$. With these conventions, we have:

PROPOSITION 11.17. *For $H \in X_N$ and $A \in M_N(\mathbb{R})$, the following are equivalent:*

(1) *The following is an Hadamard matrix, for any $q \in \mathbb{T}_1$:*

$$H_{ij}^q = H_{ij} q^{A_{ij}}$$

(2) *The following equations hold, for any $i \neq j$ and any $q \in \mathbb{T}_1$:*

$$\sum_k H_{ik} \bar{H}_{jk} q^{A_{ik} - A_{jk}} = 0$$

(3) *The following equations hold, for any $i \neq j$ and any $\varphi : \mathbb{R} \to \mathbb{C}$:*

$$\sum_k H_{ik} \bar{H}_{jk} \varphi(A_{ik} - A_{jk}) = 0$$

(4) *For any $i \neq j$ and any $r \in \mathbb{R}$, with $E_{ij}^r = \{k | A_{ik} - A_{jk} = r\}$, we have:*

$$\sum_{k \in E_{ij}^r} H_{ik} \bar{H}_{jk} = 0$$

If these conditions are satisfied, we call the matrix H^q an affine deformation of H.

PROOF. These equivalences are all elementary, and can be proved as follows:

(1) $\iff$ (2) Indeed, the scalar products between the rows of H^q are:

$$\begin{aligned} < H_i^q, H_j^q > &= \sum_k H_{ik} q^{A_{ik}} \bar{H}_{jk} \bar{q}^{A_{jk}} \\ &= \sum_k H_{ik} \bar{H}_{jk} q^{A_{ik} - A_{jk}} \end{aligned}$$

(2) $\implies$ (4) This follows from the following formula, and from the fact that the power functions $\{q^r | r \in \mathbb{R}\}$ over the unit circle $\mathbb{T}$ are linearly independent:

$$\sum_k H_{ik} \bar{H}_{jk} q^{A_{ik} - A_{jk}} = \sum_{r \in \mathbb{R}} q^r \sum_{k \in E_{ij}^r} H_{ik} \bar{H}_{jk}$$

(4) $\implies$ (3) This follows from the following formula:

$$\sum_k H_{ik} \bar{H}_{jk} \varphi(A_{ik} - A_{jk}) = \sum_{r \in \mathbb{R}} \varphi(r) \sum_{k \in E_{ij}^r} H_{ik} \bar{H}_{jk}$$

(3) $\implies$ (2) This simply follows by taking $\varphi(r) = q^r$. □

In order to understand the above deformations, which are "affine" in a certain sense, as suggested at the end of the statement, it is convenient to enlarge the attention to all types of deformations. We keep using the neighborhood notation X_p introduced above, and we consider functions of type $f : X_p \to Y_q$, which by definition satisfy $f(p) = q$. We have the following definition, further clarifying the terminology in Proposition 11.17:

DEFINITION 11.18. *Let $H \in M_N(\mathbb{C})$ be a complex Hadamard matrix.*

(1) *A deformation of H is a smooth function $f : \mathbb{T}_1 \to (X_N)_H$.*
(2) *The deformation is called "affine" if $f_{ij}(q) = H_{ij}q^{A_{ij}}$, with $A \in M_N(\mathbb{R})$.*
(3) *We call "trivial" the deformations of type $f_{ij}(q) = H_{ij}q^{a_i+b_j}$, with $a, b \in \mathbb{R}^N$.*

Here the adjective "affine", which is used in the same context as in Proposition 11.17, comes from the formula $f_{ij}(e^{it}) = H_{ij}e^{iA_{ij}t}$, because the function $t \to A_{ij}t$ which produces the exponent is indeed affine. As for the adjective "trivial", this comes from the fact that the affine deformations of type $f(q) = (H_{ij}q^{a_i+b_j})_{ij}$ are obtained from H by multiplying the rows and columns by certain numbers in $\mathbb{T}$, so are automatically Hadamard.

Getting now to the examples, and skipping some elementary theory and constructions, we have an interesting construction due to McNulty and Weigert, that we would like to explain now. This construction is based on the following simple fact:

THEOREM 11.19. *Assuming that $K \in M_N(\mathbb{C})$ is Hadamard, so is the matrix*

$$H_{ia,jb} = \frac{1}{\sqrt{Q}}K_{ij}(L_i^*R_j)_{ab}$$

provided that $\{L_1, \ldots, L_N\} \subset \sqrt{Q}U_Q$ and $\{R_1, \ldots, R_N\} \subset \sqrt{Q}U_Q$ are such that

$$\frac{1}{\sqrt{Q}}L_i^*R_j \in \sqrt{Q}U_Q$$

with $i, j = 1, \ldots, N$, are complex Hadamard.

PROOF. The check of the unitarity of the matrix in the statement can be done as follows, by using our various assumptions on the various matrices involved:

$$\begin{aligned} < H_{ia}, H_{kc} > &= \frac{1}{Q}\sum_{jb} K_{ij}(L_i^*R_j)_{ab}\bar{K}_{kj}\overline{(L_k^*R_j)}_{cb} \\ &= \sum_j K_{ij}\bar{K}_{kj}(L_i^*L_k)_{ac} \\ &= N\delta_{ik}(L_i^*L_k)_{ac} \\ &= NQ\delta_{ik}\delta_{ac} \end{aligned}$$

The entries of our matrix being in addition on the unit circle, we are done. □

As a concrete input for the above construction, we can use:

PROPOSITION 11.20. *For $q \geq 3$ prime, the matrices $\{F_q, DF_q, \ldots, D^{q-1}F_q\}$, where F_q is the Fourier matrix, and where*

$$D = diag\left(1, 1, w, w^3, w^6, w^{10}, \ldots, w^{\frac{q^2-1}{8}}, \ldots, w^{10}, w^6, w^3, w\right)$$

with $w = e^{2\pi i/q}$, are such that $\frac{1}{\sqrt{q}}E_i^*E_j$ is complex Hadamard, for any $i \neq j$.

PROOF. With by definition $0, 1, \ldots, q-1$ as indices for our matrices, as usual in a Fourier analysis context, the formula of the above matrix D is:

$$D_c = w^{0+1+\ldots+(c-1)} = w^{\frac{c(c-1)}{2}}$$

Since we have $\frac{1}{\sqrt{q}}E_i^*E_j \in \sqrt{q}U_q$, we just need to check that these matrices have entries belonging to $\mathbb{T}$, for any $i \neq j$. With $k = j - i$, these entries are given by:

$$\frac{1}{\sqrt{q}}(E_i^*E_j)_{ab} = \frac{1}{\sqrt{q}}(F_q^*D^kF_q)_{ab} = \frac{1}{\sqrt{q}}\sum_c w^{c(b-a)}D_c^k$$

Now observe that with $s = b - a$, we have the following formula:

$$\begin{aligned}
\left|\sum_c w^{cs}D_c^k\right|^2 &= \sum_{cd} w^{cs-ds}w^{\frac{c(c-1)}{2}\cdot k-\frac{d(d-1)}{2}\cdot k} \\
&= \sum_{cd} w^{(c-d)\left(\frac{c+d-1}{2}\cdot k+s\right)} \\
&= \sum_{de} w^{e\left(\frac{2d+e-1}{2}\cdot k+s\right)} \\
&= \sum_e \left(w^{\frac{e(e-1)}{2}\cdot k+es}\sum_d w^{edk}\right) \\
&= \sum_e w^{\frac{e(e-1)}{2}\cdot k+es}\cdot q\delta_{e0} \\
&= q
\end{aligned}$$

Thus the entries are on the unit circle, and we are done. □

Next, we have the following result, making use of Gauss sums:

PROPOSITION 11.21. *The matrices $G_k = \frac{1}{\sqrt{q}}F_q^*D^kF_q$, with $D = diag\left(w^{\frac{c(c-1)}{2}}\right)$, and with $k \neq 0$ are circulant, their first row vectors V^k being given by*

$$V_i^k = \delta_q\left(\frac{k/2}{q}\right)w^{\frac{q^2-1}{8}\cdot k}\cdot w^{-\frac{\frac{i}{k}(\frac{i}{k}-1)}{2}}$$

where $\delta_q = 1$ if $q = 1(4)$ and $\delta_q = i$ if $q = 3(4)$, and with all inverses being taken in $\mathbb{Z}_q$.

PROOF. The above matrices G_k are indeed circulant, their first vectors being:

$$V_i^k = \frac{1}{\sqrt{q}} \sum_c w^{\frac{c(c-1)}{2} \cdot k + ic}$$

But this is a Gauss sum, and by computing the square, we obtain:

$$(V_i^k)^2 = \delta_q^2 \cdot w^{\frac{q^2-1}{4} \cdot k} \cdot w^{-\frac{i}{k}(\frac{i}{k}-1)}$$

By extracting now the square root, we obtain a formula as follows:

$$V_i^k = \pm\delta_q \cdot w^{\frac{q^2-1}{8} \cdot k} \cdot w^{-\frac{\frac{i}{k}(\frac{i}{k}-1)}{2}}$$

And with Gauss computing for us the sign, this leads to the above formula. □

Let us combine now all the above results. We obtain the following statement:

THEOREM 11.22. *Let $q \geq 3$ be prime, consider subsets $S, T \subset \{0, 1, \ldots, q-1\}$ satisfying the conditions $|S| = |T|$ and $S \cap T = \emptyset$, and write:*

$$S = \{s_1, \ldots, s_N\} \quad , \quad T = \{t_1, \ldots, t_N\}$$

Then, with the matrix V being as above, the following matrix,

$$H_{ia,jb} = K_{ij} V_{b-a}^{t_j - s_i}$$

is complex Hadamard, provided that the matrix $K \in M_N(\mathbb{C})$ is complex Hadamard.

PROOF. This follows indeed by using the general construction in Theorem 11.19, with input coming from Proposition 11.20 and Proposition 11.21. □

The above construction covers many interesting examples of Hadamard matrices, known to be isolated, such as the Tao matrix, which is as follows, with $w = e^{2\pi i/3}$:

$$T_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & w & w & w^2 & w^2 \\ 1 & w & 1 & w^2 & w^2 & w \\ 1 & w & w^2 & 1 & w & w^2 \\ 1 & w^2 & w^2 & w & 1 & w \\ 1 & w^2 & w & w^2 & w & 1 \end{pmatrix}$$

For more on Hadamard matrices, real and complex, there are many texts available.

11d. Bistochastic matrices

Switching again topics, but with discrete Fourier analysis being as usual around the corner, a very basic definition, that you might already know, is as follows:

DEFINITION 11.23. *A square matrix $M \in M_N(\mathbb{C})$ is called bistochastic if each row and each column sum up to the same number:*

$$\begin{matrix} M_{11} & \ldots & M_{1N} & \to & \lambda \\ \vdots & & \vdots & & \\ M_{N1} & \ldots & M_{NN} & \to & \lambda \\ \downarrow & & \downarrow & & \\ \lambda & & \lambda & & \end{matrix}$$

If this happens only for the rows or columns, M is called row or column stochastic.

As the name indicates, the above matrices are useful in statistics, with the case of the matrices having entries in $[0,1]$, summing up to $\lambda = 1$, being the important one. As a basic example of a bistochastic matrix, we have the flat matrix, which is as follows:

$$\mathbb{I}_N = \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix}$$

Observe that the rescaling $P_N = \mathbb{I}_N/N$ has the property mentioned above, namely entries in $[0,1]$, summing up to $\lambda = 1$. In fact, this matrix $P_N = \mathbb{I}_N/N$ is a very familiar object in linear algebra, being the projection on the all-one vector, namely:

$$\xi = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Getting back now to the general case, the various notions of stochasticity in Definition 11.23 are closely related to this vector ξ, due to the following simple fact:

PROPOSITION 11.24. *Let $M \in M_N(\mathbb{C})$ be a square matrix.*

(1) *M is row stochastic, with sums λ, when $M\xi = \lambda\xi$.*
(2) *M is column stochastic, with sums λ, when $M^t\xi = \lambda\xi$.*
(3) *M is bistochastic, with sums λ, when $M\xi = M^t\xi = \lambda\xi$.*

PROOF. The first assertion is clear from definitions, because when multiplying a matrix by ξ, we obtain the vector formed by the row sums:

$$\begin{pmatrix} M_{11} & \ldots & M_{1N} \\ \vdots & & \vdots \\ M_{N1} & \ldots & M_{NN} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} M_{11} + \ldots + M_{1N} \\ \vdots \\ M_{N1} + \ldots + M_{NN} \end{pmatrix}$$

As for the second, and then third assertion, these are both clear from this. □

As an observation here, we can reformulate if we want the above statement in a purely matrix-theoretic form, by using the flat matrix $\mathbb{I}_N$, as follows:

PROPOSITION 11.25. *Let $M \in M_N(\mathbb{C})$ be a square matrix.*

(1) *M is row stochastic, with sums λ, when $M\mathbb{I}_N = \lambda\mathbb{I}_N$.*
(2) *M is column stochastic, with sums λ, when $\mathbb{I}_N M = \lambda\mathbb{I}_N$.*
(3) *M is bistochastic, with sums λ, when $M\mathbb{I}_N = \mathbb{I}_N M = \lambda\mathbb{I}_N$.*

PROOF. This follows from Proposition 11.24, and from the fact that both the rows and columns of the flat matrix $\mathbb{I}_N$ are copies of the all-one vector, namely:

$$\xi = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

Alternatively, we have the following formula, with $S_1, \ldots, S_N$ being the row sums of our matrix M, which gives the assertion (1):

$$\begin{pmatrix} M_{11} & \ldots & M_{1N} \\ \vdots & & \vdots \\ M_{N1} & \ldots & M_{NN} \end{pmatrix} \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} = \begin{pmatrix} S_1 & \ldots & S_1 \\ \vdots & & \vdots \\ S_N & \ldots & S_N \end{pmatrix}$$

As for the second, and then third assertion, these are both clear from this. □

In what follows, we will be mainly interested in the bistochastic matrices which are unitary, $M \in U_N$. As the simplest example here, which is a familiar object in quantum physics, we have the following matrix $K_N \in O_N$, obtained by suitably modifying the flat matrix $\mathbb{I}_N$, as to make the rows pairwise orthogonal, and of norm one:

$$K_N = \frac{1}{N} \begin{pmatrix} 2-N & 2 & \ldots & 2 \\ 2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 2 \\ 2 & \ldots & 2 & 2-N \end{pmatrix}$$

As a first result now regarding the unitary bistochastic matrices, we have:

THEOREM 11.26. *For a unitary matrix $U \in U_N$, the following are equivalent:*

(1) *H is bistochastic, with sums λ.*
(2) *H is row stochastic, with sums λ, and $|\lambda| = 1$.*
(3) *H is column stochastic, with sums λ, and $|\lambda| = 1$.*

PROOF. By using a symmetry argument we just need to prove (1) $\iff$ (2), and both the implications are elementary, as follows:

(1) $\implies$ (2) If we denote by $U_1, \ldots, U_N \in \mathbb{C}^N$ the rows of U, we have indeed:

$$\begin{aligned} 1 &= < U_1, U_1 > \\ &= \sum_i < U_1, U_i > \\ &= \sum_i \sum_j U_{1j} \bar{U}_{ij} \\ &= \sum_j U_{1j} \sum_i \bar{U}_{ij} \\ &= \sum_j U_{1j} \cdot \bar{\lambda} \\ &= \lambda \cdot \bar{\lambda} \\ &= |\lambda|^2 \end{aligned}$$

(2) $\implies$ (1) Consider the all-one vector $\xi = (1)_i \in \mathbb{C}^N$. The fact that U is row-stochastic with sums λ reads:

$$\begin{aligned} \sum_j U_{ij} = \lambda, \forall i \quad &\iff \quad \sum_j U_{ij}\xi_j = \lambda\xi_i, \forall i \\ &\iff \quad U\xi = \lambda\xi \end{aligned}$$

Also, the fact that U is column-stochastic with sums λ reads:

$$\begin{aligned} \sum_i U_{ij} = \lambda, \forall j \quad &\iff \quad \sum_j U_{ij}\xi_i = \lambda\xi_j, \forall j \\ &\iff \quad U^t\xi = \lambda\xi \end{aligned}$$

We must prove that the first condition implies the second one, provided that the row sum λ satisfies $|\lambda| = 1$. But this follows from the following computation:

$$\begin{aligned} U\xi = \lambda\xi \quad &\implies \quad U^*U\xi = \lambda U^*\xi \\ &\implies \quad \xi = \lambda U^*\xi \\ &\implies \quad \xi = \bar{\lambda} U^t\xi \\ &\implies \quad U^t\xi = \lambda\xi \end{aligned}$$

Thus, we have proved both the implications, and we are done. □

From now on all our bistochastic matrices will be assumed to be normalized, with the sum on all rows and columns being equal to $\lambda = 1$. We have the following result:

THEOREM 11.27. *We have groups as follows:*

(1) $B_N \subset O_N$, *consisting of the orthogonal matrices which are bistochastic.*
(2) $C_N \subset U_N$, *consisting of the unitary matrices which are bistochastic.*

PROOF. We know from Theorem 11.26 that the sets of bistochastic matrices B_N, C_N in the statement appear as follows, with ξ being the all-one vector:

$$B_N = \left\{U \in O_N \middle| U\xi = \xi\right\}$$

$$C_N = \left\{U \in U_N \middle| U\xi = \xi\right\}$$

It is then clear that both B_N, C_N are stable under the multiplication, contain the unit, and are stable by inversion. Thus, we have indeed groups, as stated. □

Following Idel-Wolf and related work, we would like to discuss now a non-trivial result involving the complex bistochastic group C_N. We will need some geometric preliminaries. The complex projective space appears by definition as follows:

$$P_{\mathbb{C}}^{N-1} = (\mathbb{C}^N - \{0\})/ < x = \lambda y >$$

Inside this projective space, we have the Clifford torus, constructed as follows:

$$\mathbb{T}^{N-1} = \left\{(z_1, \ldots, z_N) \in P_{\mathbb{C}}^{N-1} \middle| |z_1| = \ldots = |z_N|\right\}$$

With these conventions, we have the following result:

PROPOSITION 11.28. *For a unitary matrix $U \in U_N$, the following are equivalent:*

(1) *There exist $L, R \in U_N$ diagonal such that $U' = LUR$ is bistochastic.*
(2) *The standard torus $\mathbb{T}^N \subset \mathbb{C}^N$ satisfies $\mathbb{T}^N \cap U\mathbb{T}^N \neq \emptyset$.*
(3) *The Clifford torus $\mathbb{T}^{N-1} \subset P_{\mathbb{C}}^{N-1}$ satisfies $\mathbb{T}^{N-1} \cap U\mathbb{T}^{N-1} \neq \emptyset$.*

PROOF. These equivalences are all elementary, as follows:

(1) $\implies$ (2) Assuming that $U' = LUR$ is bistochastic, which in terms of the all-1 vector ξ means $U'\xi = \xi$, if we set $f = R\xi \in \mathbb{T}^N$ we have:

$$Uf = \bar{L}U'\bar{R}f = \bar{L}U'\xi = \bar{L}\xi \in \mathbb{T}^N$$

Thus we have $Uf \in \mathbb{T}^N \cap U\mathbb{T}^N$, which gives the conclusion.

(2) $\implies$ (1) Given $g \in \mathbb{T}^N \cap U\mathbb{T}^N$, we can define R, L as follows:

$$R = \begin{pmatrix} g_1 & & \\ & \ddots & \\ & & g_N \end{pmatrix} \quad , \quad \bar{L} = \begin{pmatrix} (Ug)_1 & & \\ & \ddots & \\ & & (Ug)_N \end{pmatrix}$$

With these values for L, R, we have then the following formulae:

$$R\xi = g \quad , \quad \bar{L}\xi = Ug$$

Thus the matrix $U' = LUR$ is bistochastic, because:

$$U'\xi = LUR\xi = LUg = \xi$$

(2) $\implies$ (3) This is clear, because $\mathbb{T}^{N-1} \subset P_{\mathbb{C}}^{N-1}$ appears as the projective image of $\mathbb{T}^N \subset \mathbb{C}^N$, and so $\mathbb{T}^{N-1} \cap U\mathbb{T}^{N-1}$ appears as the projective image of $\mathbb{T}^N \cap U\mathbb{T}^N$.

(3) $\implies$ (2) We have indeed the following equivalence:

$$\mathbb{T}^{N-1} \cap U\mathbb{T}^{N-1} \neq \emptyset \iff \exists \lambda \neq 0, \lambda\mathbb{T}^N \cap U\mathbb{T}^N \neq \emptyset$$

But $U \in U_N$ implies $|\lambda| = 1$, and this gives the result. □

The point now is that the condition (3) above is something familiar in symplectic geometry, and known to hold for any $U \in U_N$. We are led in this way to:

THEOREM 11.29. *Any unitary matrix $U \in U_N$ can be put in bistochastic form,*

$$U' = LUR$$

with $L, R \in U_N$ being both diagonal, via a certain non-explicit method.

PROOF. As already mentioned, the condition $\mathbb{T}^{N-1} \cap U\mathbb{T}^{N-1} \neq \emptyset$ in Proposition 11.28 (3) is something quite natural in symplectic geometry. To be more precise:

(1) The Clifford torus $\mathbb{T}^{N-1} \subset P_{\mathbb{C}}^{N-1}$ is a Lagrangian submanifold, and the map $\mathbb{T}^{N-1} \to U\mathbb{T}^{N-1}$ is a Hamiltonian isotopy. For more on this, see Arnold [2].

(2) A non-trivial result of Biran-Entov-Polterovich and Cho states that the Clifford torus $\mathbb{T}^{N-1}$ cannot be displaced from itself via a Hamiltonian isotopy.

(3) Thus, the above-mentioned result tells us that $\mathbb{T}^{N-1} \cap U\mathbb{T}^{N-1} \neq \emptyset$ holds indeed, for any $U \in U_N$. We therefore obtain the result, via Proposition 11.28. □

11e. Exercises

This was a truly exciting linear algebra chapter, and as exercises, we have:

EXERCISE 11.30. *Learn about circulant graphs, and their adjacency matrices.*

EXERCISE 11.31. *Learn about vertex-transitive graphs, and their adjacency matrices.*

EXERCISE 11.32. *Try proving the HC or CHC, without insisting too much, tough.*

EXERCISE 11.33. *Learn about quasi-Hadamard and almost Hadamard matrices.*

EXERCISE 11.34. *Classify, with full details, the Hadamard matrices at $N = 3, 4$.*

EXERCISE 11.35. *Learn about the various known Hadamard matrices at $N = 6$.*

EXERCISE 11.36. *Learn about the Sinkhorn algorithm, and normal form.*

EXERCISE 11.37. *Learn more about the Idel-Wolf theorem, and its ingredients.*

As bonus exercise, learn more about design theory, and Hadamard matrices.

CHAPTER 12

Graph theory

12a. Graphs, Laplacian

We discuss here some further applications of linear algebra to discrete mathematics, and more specifically, to the finite graphs. As usual with discrete mathematics, as we got used to in this book, discrete Fourier analysis will be somewhere around the corner.

Let us start with some generalities. The finite graphs X are described by their adjacency matrices $d \in M_N(0,1)$, with N being the number of vertices, and with this being something very useful. Consider for instance the following basic question:

QUESTION 12.1. *Given a graph X, with a distinguished vertex $*$:*

(1) *What is the number L_k of length k loops on X, based at $*$?*
(2) *Equivalently, what is the measure μ having L_k as moments?*

To be more precise, we are mainly interested in the first question, counting loops on graphs, with this being notoriously related to many applied mathematics questions, of discrete type. As for the second question, this is a useful reformulation of it.

In practice now, many things can be said, starting with the following basic result, featuring the adjacency matrix $d \in M_N(0,1)$ and its diagonalization, which is something quite elementary, and which can be very helpful for explicit computations:

THEOREM 12.2. *Given a graph X, with adjacency matrix $d \in M_N(0,1)$, we have:*

$$L_k = (d^k)_{**}$$

When writing $d = UDU^t$ with $U \in O_N$ and $D = diag(\lambda_1, \ldots, \lambda_N)$ with $\lambda_i \in \mathbb{R}$, we have

$$L_k = \sum_i U_{*i}^2 \lambda_i^k$$

and the real probability measure μ having these numbers as moments is given by

$$\mu = \sum_i U_{*i}^2 \delta_{\lambda_i}$$

with the delta symbols standing as usual for Dirac masses.

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

(1) According to the usual rule of matrix multiplication, the formula for the powers of the adjacency matrix $d \in M_N(0,1)$ is as follows:

$$\begin{aligned}(d^k)_{i_0i_k} &= \sum_{i_1,\ldots,i_{k-1}} d_{i_0i_1}d_{i_1i_2}\ldots d_{i_{k-1}i_k}\\ &= \sum_{i_1,\ldots,i_{k-1}} \delta_{i_0-i_1}\delta_{i_1-i_2}\ldots\delta_{i_{k-1}-i_k}\\ &= \sum_{i_1,\ldots,i_{k-1}} \delta_{i_0-i_1-\ldots-i_{k-1}-i_k}\\ &= \#\Big\{i_0-i_1-\ldots-i_{k-1}-i_k\Big\}\end{aligned}$$

In particular, with $i_0 = i_k = *$, we obtain the following formula, as claimed:

$$(d^k)_{**} = \#\Big\{*-i_1-\ldots-i_{k-1}-*\Big\} = L_k$$

(2) Now since $d \in M_N(0,1)$ is symmetric, this matrix is diagonalizable, with the diagonalization being as follows, with $U \in O_N$, and $D = diag(\lambda_1,\ldots,\lambda_N)$ with $\lambda_i \in \mathbb{R}$:

$$d = UDU^t$$

By using this formula, we obtain the second formula in the statement:

$$\begin{aligned}L_k &= (d^k)_{**}\\ &= (UD^kU^t)_{**}\\ &= \sum_i U_{*i}\lambda_i^k(U^t)_{i*}\\ &= \sum_i U_{*i}^2\lambda_i^k\end{aligned}$$

(3) Finally, the last assertion is clear from this, because the moments of the measure in the statement, $\mu = \sum_i U_{*i}^2\delta_{\lambda_i}$, are the following numbers:

$$\begin{aligned}M_k &= \int_{\mathbb{R}} x^k d\mu(x)\\ &= \sum_i U_{*i}^2\lambda_i^k\\ &= L_k\end{aligned}$$

Observe also that μ is indeed of mass 1, because all rows of $U \in O_N$ must be of norm 1, and so $\sum_i U_{*i}^2 = 1$. Thus, we are led to the conclusions in the statement. □

Summarizing, foundations laid for counting loops via linear algebra, and I will leave it to you, to have some fun with all this, for some simple graphs of your choice.

Moving forward, in relation with this, but at a more conceptual level, we can formulate the following definition, making an interesting link with calculus and geometry:

DEFINITION 12.3. *We call Laplacian of a graph X the matrix*

$$L = v - d$$

with v being the diagonal valence matrix, and d being the adjacency matrix.

This definition is inspired by differential geometry, or just by multivariable calculus, and more specifically by the well-known Laplace operator there, given by:

$$\Delta f = \sum_{i=1}^{N} \frac{d^2 f}{dx_i^2}$$

More on this in a moment, but as a word regarding terminology, we have:

WARNING 12.4. *The graph Laplacian above is in fact the negative Laplacian,*

$$L = -\Delta$$

with our preference for it, negative, coming from the fact that it is positive, $L \geq 0$.

Which sounds like a bad joke, but this is how things are, and more on this a moment. In practice now, the graph Laplacian is given by the following formula:

$$L_{ij} = \begin{cases} v_i & \text{if } i = j \\ -1 & \text{if } i - j \\ 0 & \text{otherwise} \end{cases}$$

Alternatively, we have the following formula, for the entries of the Laplacian:

$$L_{ij} = \delta_{ij} v_i - \delta_{i-j}$$

With these formulae in hand, we can formulate, as our first result on the subject:

PROPOSITION 12.5. *A function on a graph is harmonic, $Lf = 0$, precisely when*

$$f_i = \frac{1}{v_i} \sum_{i-j} f_j$$

that is, when the value at each vertex is the average over the neighbors.

PROOF. We have indeed the following computation, for any function f:

$$\begin{aligned}(Lf)_i &= \sum_j L_{ij}f_j\\ &= \sum_j(\delta_{ij}v_i-\delta_{i-j})f_j\\ &= v_if_i-\sum_{i-j}f_j\end{aligned}$$

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

Summarizing, we have some good reasons for calling L the Laplacian, because the solutions of $Lf=0$ satisfy what we would expect from a harmonic function, namely having the "average over the neighborhood" property. With the remark however that the harmonic functions on graphs are something trivial, due to the following fact:

PROPOSITION 12.6. *A function on a graph X is harmonic in the above sense precisely when it is constant over the connected components of X.*

PROOF. This is clear from the equation that we found in Proposition 12.5, namely:

$$f_i=\frac{1}{v_i}\sum_{i-j}f_j$$

Indeed, based on this, we can say for instance that f cannot have variations over a connected component, and so must be constant on these components, as stated. □

At a more advanced level now, let us try to understand the relation with the usual Laplacian from analysis Δ, which is given by the following formula:

$$\Delta f=\sum_{i=1}^N\frac{d^2f}{dx_i^2}$$

In one dimension, $N=1$, the Laplacian is simply the second derivative, $\Delta f=f''$. Now let us recall that the first derivative of a one-variable function is given by:

$$f'(x)=\lim_{t\to0}\frac{f(x+t)-f(x)}{t}$$

We deduce from this, or from the Taylor formula at order 2, to be fully correct, that the second derivative of a one-variable function is given by the following formula:

$$\begin{aligned}f''(x) &= \lim_{t\to0}\frac{f'(x+t)-f'(x)}{t}\\ &= \lim_{t\to0}\frac{f(x+t)-2f(x)+f(x-t)}{t^2}\end{aligned}$$

Now since $\mathbb{R}$ can be thought of as appearing as the continuum limit, $t \to 0$, of the graphs $t\mathbb{Z} \simeq \mathbb{Z}$, this suggests defining the Laplacian of $\mathbb{Z}$ by the following formula:

$$\Delta f(x) = \frac{f(x+1) - 2f(x) + f(x-1)}{1^2}$$

But this is exactly what we have in Definition 12.3, up to a sign switch, the graph Laplacian of $\mathbb{Z}$, as constructed there, being given by the following formula:

$$Lf(x) = 2f(x) - f(x+1) - f(x-1)$$

Summarizing, we have reached to the formula in Warning 12.4, namely:

$$L = -\Delta$$

In arbitrary dimensions now, everything generalizes well, and we have:

THEOREM 12.7. *The Laplacian of graphs is compatible with the usual Laplacian,*

$$\Delta f = \sum_{i=1}^{N} \frac{d^2 f}{dx_i^2}$$

via the following formula, showing that our L is in fact the negative Laplacian,

$$L = -\Delta$$

via regarding $\mathbb{R}^N$ as the continuum limit, $t \to 0$, of the graphs $t\mathbb{Z}^N \simeq \mathbb{Z}^N$.

PROOF. This is something that we know at $N = 1$, and the proof in general is similar. Indeed, at $N = 2$, to start with, the formula that we need is as follows:

$$\begin{aligned}
\Delta f(x,y) &= \frac{d^2 f}{dx^2} + \frac{d^2 f}{dy^2} \\
&= \lim_{t \to 0} \frac{f(x+t,y) - 2f(x,y) + f(x-t,y)}{t^2} \\
&+ \lim_{t \to 0} \frac{f(x,y+t) - 2f(x,y) + f(x,y-t)}{t^2} \\
&= \lim_{t \to 0} \frac{f(x+t,y) + f(x-t,y) + f(x,y+t) + f(x,y-t) - 4f(x,y)}{t^2}
\end{aligned}$$

Now since $\mathbb{R}^2$ can be thought of as appearing as the continuum limit, $t \to 0$, of the graphs $t\mathbb{Z}^2 \simeq \mathbb{Z}^2$, this suggests defining the Laplacian of $\mathbb{Z}^2$ as follows:

$$\Delta f(x) = \frac{f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y)}{1^2}$$

But this is exactly what we have in Definition 12.3, up to a sign switch, the graph Laplacian of $\mathbb{Z}^2$, as constructed there, being given by the following formula:

$$Lf(x) = 4f(x,y) - f(x+1,y) - f(x-1,y) - f(x,y+1) - f(x,y-1)$$

At higher $N \in \mathbb{N}$ the proof is similar, and we will leave this as an exercise. □

Now back to our general graph questions, and to Definition 12.3 as it is, the Laplacian of graphs as constructed there has the following basic properties:

THEOREM 12.8. *The graph Laplacian $L = v - d$ has the following properties:*

(1) *It is symmetric, $L = L^t$.*
(2) *It is positive definite, $L \geq 0$.*
(3) *It is bistochastic, with row and column sums 0.*
(4) *It has 0 as eigenvalue, with the other eigenvalues being positive.*
(5) *The multiplicity of 0 is the number of connected components.*
(6) *In the connected case, the eigenvalues are $0 < \lambda_1 \leq \ldots \leq \lambda_{N-1}$.*

PROOF. All this is straightforward, the idea being as follows:

(1) This is clear from $L = v - d$, both v, d being symmetric.

(2) This follows from the following computation, for any function f on the graph:

$$\begin{aligned} < Lf, f > &= \sum_{ij} L_{ij} f_i f_j \\ &= \sum_{ij} (\delta_{ij} v_i - \delta_{i-j}) f_i f_j \\ &= \sum_{i} v_i f_i f_j - \sum_{i-j} f_i f_j \\ &= \sum_{i \sim j} f_i^2 - \sum_{i-j} f_i f_j \\ &= \frac{1}{2} \sum_{i-j} (f_i - f_j)^2 \\ &\geq 0 \end{aligned}$$

(3) This is again clear from $L = v - d$, and from the definition of v, d.

(4) Here the first assertion comes from (3), and the second one, from (2).

(5) Given an arbitrary graph, we can label its vertices inceasingly, over the connected components, and this makes the adjacency matrix d, so the Laplacian L as well, block diagonal. Thus, we are left with proving that for a connected graph, the multiplicity of 0 is precisely 1. But this follows from the formula from the proof of (2), namely:

$$< Lf, f >= \frac{1}{2} \sum_{i-j} (f_i - f_j)^2$$

Indeed, this formula shows in particular that we have the following equivalence:

$$Lf = 0 \iff f_i = f_j, \forall i - j$$

Now since our graph was assumed to be connected, as per the above beginning of proof, the condition on the right means that f must be constant. Thus, the 0-eigenspace of the Laplacian follows to be 1-dimensional, spanned by the all-1 vector, as desired.

(6) This follows indeed from (4) and (5), and with the remark that in fact we already proved this, in the proof of (5), with the formulae there being very useful in practice. □

12b. Kirchoff formula

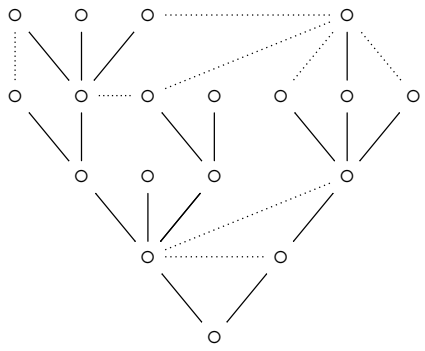
As a main application of our Laplacian technology, which will lead us deep into graph theory, let us discuss now the Kirchoff formula. This is based on the following facts:

PROPOSITION 12.9. *The following happen:*

(1) *Any connected graph has a spanning tree, meaning a tree subgraph, making use of all vertices.*
(2) *For the complete graph K_N, with vertices labeled $1, \ldots, N$, the spanning trees are exactly the trees with vertices labeled $1, \ldots, N$.*

PROOF. Both the assertions are trivial, the idea being as follows:

(1) The fact that any connected graph has indeed a spanning tree is something which is very intuitive, clear on pictures, and we will leave the formal proof, which is not difficult, as an exercise. As an illustration for this, here is a picture of a quite random graph, which, after removal of some of the edges, the dotted ones, becomes indeed a tree:



(2) As for the second assertion, this is something which is clear too, and again we will leave the formal proof, which is not difficult at all, as an exercise. □

In view of the above, the following interesting question appears:

QUESTION 12.10. *Given a connected graph X, with vertices labeled $1, \ldots, N$, how to count its spanning trees? And, for the complete graph K_N, do we really get N^{N-2} such trees, in agreement with the well-known Cayley formula, by using this method?*

Getting to work now, following Kirchoff, the idea will be that of connecting the spanning trees of a connected graph X to the combinatorics of the Laplacian of X, which is by definition given by the following formula, with v being the valence matrix:

$$L = v - d$$

In order to get started, we will just need the fact, which is something trivial, that L is bistochastic, with zero row and column sums. Indeed, this makes the link with the following basic linear algebra fact, that we can use afterwards, for our Laplacian L:

PROPOSITION 12.11. *For a matrix $L \in M_N(\mathbb{R})$ which is bistochastic, with zero row and column sums, the signed minors*

$$T_{ij} = (-1)^{i+j} \det(L^{ij})$$

do not depend on the choice of the indices i, j.

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

(1) Before anything, let us do a quick check at $N = 2$. Here the bistochastic matrices, with zero row and column sums, are as follows, with $a \in \mathbb{R}$:

$$L = \begin{pmatrix} a & -a \\ -a & a \end{pmatrix}$$

But this gives the result, with the number in question being $T_{ij} = a$.

(2) Let us do as well a quick check at $N = 3$. Here the bistochastic matrices, with zero row and column sums, are as follows, with $a, b, c, d \in \mathbb{R}$:

$$L = \begin{pmatrix} a & b & -a-b \\ c & d & -c-d \\ -a-c & -b-d & a+b+c+d \end{pmatrix}$$

But this gives again the result, with the number in question being $T_{ij} = ad - bc$.

(3) In the general case now, where $N \in \mathbb{N}$ is arbitrary, the bistochastic matrices with zero row and column sums are as follows, with $A \in M_n(\mathbb{R})$ with $n = N - 1$ being an arbitary matrix, and with $R_1, \ldots, R_n$ and $C_1, \ldots, C_n$ being the row and column sums of this matrix, and $S = \sum R_i = \sum C_i$ being the total sum of this matrix:

$$L = \begin{pmatrix} A_{11} & \ldots & A_{1n} & -R_1 \\ \vdots & & \vdots & \vdots \\ A_{n1} & \ldots & A_{nn} & -R_n \\ -C_1 & \ldots & -C_n & S \end{pmatrix}$$

We want to prove that the signed minors of L coincide, and by using the symmetries of the problem, it is enough to prove that the following equality holds:

$$L^{n+1,n} = -L^{n+1,n+1}$$

But, what we have on the right is $-\det A$, and what we have on the left is:

$$\begin{aligned}
L^{n+1,n} &= \begin{vmatrix} A_{11} & \ldots & A_{1,n-1} & -R_1 \\ \vdots & & \vdots & \vdots \\ A_{n1} & \ldots & A_{n,n-1} & -R_n \end{vmatrix} \\
&= \begin{vmatrix} A_{11} & \ldots & A_{1,n-1} & A_{11}+\ldots+A_{1,n-1}-R_1 \\ \vdots & & \vdots & \vdots \\ A_{n1} & \ldots & A_{n,n-1} & A_{n1}+\ldots+A_{n,n-1}-R_n \end{vmatrix} \\
&= \begin{vmatrix} A_{11} & \ldots & A_{1,n-1} & -A_{1n} \\ \vdots & & \vdots & \vdots \\ A_{n1} & \ldots & A_{n,n-1} & -A_{nn} \end{vmatrix} \\
&= -\det A
\end{aligned}$$

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

We can now formulate, following Kirchoff, the following key result:

THEOREM 12.12. *Given a connected graph X, with vertices labeled $1,\ldots,N$, the number of spanning trees inside X, meaning tree subgraphs using all vertices, is*

$$T_X = (-1)^{i+j}\det(L^{ij})$$

with $L = v - d$ being the Laplacian, with this being independent on the chosen minor.

PROOF. This is something non-trivial, the idea being as follows:

(1) We know from Proposition 12.11 that the signed minors of L coincide. In other words, we have a common formula as follows, with $T \in \mathbb{Z}$ being a certain number:

$$(-1)^{i+j}\det(L^{ij}) = T$$

Our claim, which will prove the result, is that the number of spanning trees T_X is precisely this common number T. That is, with $i = j = 1$, our claim is that we have:

$$T_X = \det(L^{11})$$

(2) In order to prove our claim, which is non-trivial, we use a trick. We orient all the edges $e = (ij)$ of our graph as to have $i < j$, and we define the ordered incidence matrix of our graph, which is a rectangular matrix, with the vertices i as row indices, and the oriented edges $e = (ij)$ as column indices, by the following formula:

$$E_{ie} = \begin{cases} 1 & \text{if } e = (ij) \\ -1 & \text{if } e = (ji) \\ 0 & \text{otherwise} \end{cases}$$

The point is that, in terms of this matrix, the Laplacian decomposes as follows:

$$L = EE^t$$

(3) Indeed, let us compute the matrix on the right. We have, by definition:

$$(EE^t)_{ij} = \sum_e E_{ie}E_{je}$$

Let us first compute the contributions of type 1×1, to the above sum. These come from the edges e having the property $E_{ie} = E_{je} = 1$. But $E_{ie} = 1$ means $e = (ik)$ with $i < k$, and $E_{je} = 1$ means $e = (jl)$ with $j < l$. Thus, our condition $E_{ie} = E_{je} = 1$ means $i = j$, and $e = (ik)$ with $i < k$, so the contributions of type 1×1 are given by:

$$C_{1\times 1} = \delta_{ij}\#\left\{k\middle| i < k,\ i - k\right\}$$

Similarly, the contributions of type $(-1) \times (-1)$ to our sum come from the equations $E_{ie} = E_{je} = -1$, which read $i = j$ and $e = (ki)$ with $k < i$, so are given by:

$$C_{(-1)\times(-1)} = \delta_{ij}\#\left\{k\middle| k < i,\ i - k\right\}$$

Now observe that by summing, the total 1 contributions to our sum, be them of type 1×1 or $(-1) \times (-1)$, are given by the following formula, v being the valence function:

$$\begin{aligned} C_1 &= C_{1\times 1} + C_{(-1)\times(-1)} \\ &= \delta_{ij}\#\left\{k\middle| i < k,\ i - k\right\} + \delta_{ij}\#\left\{k\middle| k < i,\ i - k\right\} \\ &= \delta_{ij}\#\left\{k\middle| i - k\right\} \\ &= \delta_{ij}v_i \end{aligned}$$

(4) It remains to compute the total -1 contributions to our sum. But here, we first have the contributions of type $1 \times (-1)$, coming from the equations $E_{ie} = 1, E_{je} = -1$. Now since $E_{ie} = 1$ means $e = (ik)$ with $i < k$, and $E_{je} = -1$ means $e = (lj)$ with $l < j$, our equations $E_{ie} = 1, E_{je} = -1$ amount in saying that $e = (ij)$ with $i < j$. We conclude that the contributions of type $(-1) \times 1$ to our sum are given by:

$$C_{1\times(-1)} = \delta_{i-j}\delta_{i<j}$$

Similarly, the contributions of type $(-1) \times 1$ to our sum come from the equations $E_{ie} = -1, E_{je} = 1$, which read $e = (ij)$ with $i < j$, so these are given by:

$$C_{(-1)\times 1} = \delta_{i-j}\delta_{i>j}$$

Now by summing, the total -1 contributions to our sum, be them of type $1 \times (-1)$ or $(-1) \times 1$, are given by the following formula, d being the adjacency matrix:

$$\begin{aligned} C_{-1} &= C_{1\times(-1)} + C_{(-1)\times 1} \\ &= \delta_{i-j}\delta_{i<j} + \delta_{i-j}\delta_{i>j} \\ &= \delta_{i-j} \\ &= d_{ij} \end{aligned}$$

(5) But with this, we can now finish the proof of our claim in (2), as follows:

$$\begin{aligned}(EE^t)_{ij} &= \sum_e E_{ie}E_{je}\\ &= C_1 - C_{-1}\\ &= \delta_{ij}v_i - d_{ij}\\ &= (v-d)_{ij}\\ &= L_{ij}\end{aligned}$$

Thus, we have $EE^t = L$, and claim proved. Note in passing that our formula $EE^t = L$ gives another proof of the well-known property $L \geq 0$ of the Laplacian.

(6) Getting now towards minors, if we denote by F the submatrix of E obtained by deleting the first row, the one coming from the vertex 1, we have, for any $i, j > 1$:

$$\begin{aligned}(FF^t)_{ij} &= \sum_e F_{ie}F_{je}\\ &= \sum_e E_{ie}E_{je}\\ &= (EE^t)_{ij}\\ &= L_{ij}\\ &= (L^{11})_{ij}\end{aligned}$$

We conclude that we have the following equality of matrices:

$$L^{11} = FF^t$$

(7) The point now is that, in order to compute the determinant of this latter matrix, we can use the Cauchy-Binet formula. To be more precise, the Cauchy-Binet formula says that given rectangular matrices A, B, of respective sizes $M \times N$ and $N \times M$, we have the following formula, with A_S, B_S being both $M \times M$ matrices, obtained from A, B by cutting, in the obvious way, with respect to the set of indices S:

$$\det(AB) = \sum_{|S|=M} \det(A_S)\det(B_S)$$

Observe that this formula tells us at $M > N$ that we have $\det(AB) = 0$, as it should be, and at $M = N$ that we have $\det(AB) = \det A \det B$, again as it should be. At $M < N$, which is the interesting case, the Cauchy-Binet formula holds indeed, with the proof being a bit similar to that of the formula $\det(AB) = \det A \det B$ for the square matrices, which itself is not exactly a trivial business. We will leave clarifying all this as an exercise.

(8) Now back to our questions, in the context of our formula $L^{11} = FF^t$ from (6), we can apply Cauchy-Binet to the matrices $A = F$ and $B = F^t$, having respective sizes $(N-1) \times N$ and $N \times (N-1)$. We are led in this way to the following formula,

with S ranging over the subsets of the edge set having size $N-1$, and with F_S being the corresponding square submatrix of E, having size $(N-1)\times(N-1)$, obtained by restricting the attention to the columns indexed by the subset S:

$$\begin{aligned}\det(L^{11}) &= \det(FF^t)\\ &= \sum_S \det(F_S)\det(F_S^t)\\ &= \sum_S \det(F_S)^2\end{aligned}$$

(9) Now comes the combinatorics. The sets S appearing in the above computation specify in fact $N-1$ edges of our graph, and so specify a certain subgraph X_S. But, in this picture, our claim is that we have the following formula:

$$\det(F_S)=\begin{cases}\pm1 & \text{if } X_S \text{ is a spanning tree}\\ 0 & \text{otherwise}\end{cases}$$

Indeed, since the subgraph X_S has N vertices and $N-1$ edges, it must be either a spanning tree, or have a cycle, and the study here goes as follows:

– In the case where X_S is a spanning tree, we pick a leaf of this tree, in theory I mean, by leaving it there, on the tree. The corresponding row of F_S consists then of a ±1 entry, at the neighbor of the leaf, and of 0 entries elsewhere. Thus, by developing $\det(F_S)$ over that row, we are led to a recurrence, which gives $\det(F_S)=\pm1$, as claimed above.

– In the other case, where X_S has a cycle, the sum of the columns of F_S indexed by the vertices belonging to this cycle must be 0. We conclude that in this case we have $\det(F_S)=0$, again as claimed above, and this finishes the proof of our claim.

(10) By putting now everything together, we obtain the following formula:

$$\det(L^{11})=T_X$$

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

As a basic application of the Kirchoff formula, let us apply it to the complete graph K_N. We are led in this way to another proof of the Cayley formula, as follows:

THEOREM 12.13. *The number of spanning trees of the complete graph K_N is*

$$T_{K_N}=N^{N-2}$$

in agreement with the Cayley formula.

PROOF. This is something which is clear from the Kirchoff formula, but let us prove this slowly, as an illustration for the various computations above:

(1) At $N = 2$ the Laplacian of the segment K_2 is given by the following fomula:

$$L = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Thus the common cofactor is 1, which equals the number of spanning trees, $2^0 = 1$.

(2) At $N = 3$ the Laplacian of the triangle K_3 is given by the following fomula:

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

Thus the common cofactor is 3, which equals the number of spanning trees, $3^1 = 3$.

(3) At $N = 4$ the Laplacian of the tetrahedron K_4 is given by the following fomula:

$$L = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}$$

Here the cofactor is $27 - 11 = 16$, which is the number of spanning trees, $4^2 = 16$.

(4) In general, for the complete graph K_N, the Laplacian is as follows:

$$\Delta = \begin{pmatrix} N-1 & -1 & \ldots & -1 & -1 \\ -1 & N-1 & \ldots & -1 & -1 \\ \vdots & \vdots & & \vdots & \vdots \\ -1 & -1 & \ldots & N-1 & -1 \\ -1 & -1 & \ldots & -1 & N-1 \end{pmatrix}$$

Thus, the common cofactor is N^{N-2}, in agreement with the Cayley formula. □

Finally, let us mention that in what regards the counting of trees having N vertices, this time without labeled vertices, things here are far more complicated, and there is no formula available, for the number of such trees. We refer here to the literature.

12c. Into the waves

Time now for some exciting physics, getting straight to the point, waves and heat. We first have the following result, regarding the waves, coming with a graph proof:

THEOREM 12.14. *The wave equation in $\mathbb{R}^N$ is*

$$\ddot{\varphi} = v^2 \Delta \varphi$$

where $v > 0$ is the propagation speed.

PROOF. Before everything, the equation in the statement is what comes out of experiments. However, allowing us a bit of imagination, and trust in this imagination, we can mathematically "prove" this equation, by discretizing, as follows:

(1) Let us first consider the 1D case. In order to understand the propagation of waves, we will model $\mathbb{R}$ as a network of balls, with springs between them, as follows:

$$\cdots \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \bullet \times\!\!\times\!\!\times \cdots$$

Now let us send an impulse, and see how the balls will be moving. For this purpose, we zoom on one ball. The situation here is as follows, l being the spring length:

$$\cdots\cdots \bullet_{\varphi(x-l)} \times\!\!\times\!\!\times \bullet_{\varphi(x)} \times\!\!\times\!\!\times \bullet_{\varphi(x+l)} \cdots\cdots$$

We have two forces acting at x. First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have two springs at x, this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l \\ &= k(\varphi(x+l) - \varphi(x)) - k(\varphi(x) - \varphi(x-l)) \\ &= k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$m \cdot \ddot{\varphi}(x) = k(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

(2) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $B >> 0$ balls, having a total mass M, and spanning a total distance L. Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, which is of course lower than k:

$$m = \frac{M}{B} \quad , \quad k = KB \quad , \quad l = \frac{L}{B}$$

With these changes, our equation of motion found in (1) reads:

$$\ddot{\varphi}(x) = \frac{KB^2}{M}(\varphi(x+l) - 2\varphi(x) + \varphi(x-l))$$

Now observe that this equation can be written, more conveniently, as follows:

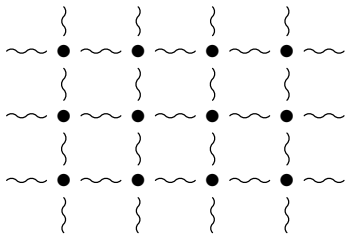
$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{\varphi(x+l) - 2\varphi(x) + \varphi(x-l)}{l^2}$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

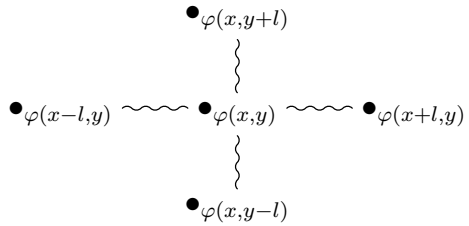
$$\ddot{\varphi}(x) = \frac{KL^2}{M} \cdot \frac{d^2\varphi}{dx^2}(x)$$

We are therefore led to the wave equation in the statement, which is $\ddot{\varphi} = v^2\varphi''$ in our present $N = 1$ dimensional case, the propagation speed being $v = \sqrt{K/M} \cdot L$.

(3) In 2 dimensions now, the same argument carries on. Indeed, we can use here a lattice model as follows, with all the edges standing for small springs:



As before in one dimension, we send an impulse, and we zoom on one ball. The situation here is as follows, with l being the spring length:



We have two forces acting at (x, y). First is the Newton motion force, mass times acceleration, which is as follows, with m being the mass of each ball:

$$F_n = m \cdot \ddot{\varphi}(x, y)$$

And second is the Hooke force, displacement of the spring, times spring constant. Since we have four springs at (x, y), this is as follows, k being the spring constant:

$$\begin{aligned} F_h &= F_h^r - F_h^l + F_h^u - F_h^d \\ &= k(\varphi(x + l, y) - \varphi(x, y)) - k(\varphi(x, y) - \varphi(x - l, y)) \\ &+ k(\varphi(x, y + l) - \varphi(x, y)) - k(\varphi(x, y) - \varphi(x, y - l)) \\ &= k(\varphi(x + l, y) - 2\varphi(x, y) + \varphi(x - l, y)) \\ &+ k(\varphi(x, y + l) - 2\varphi(x, y) + \varphi(x, y - l)) \end{aligned}$$

We conclude that the equation of motion, in our model, is as follows:

$$\begin{aligned} m \cdot \ddot{\varphi}(x, y) &= k(\varphi(x + l, y) - 2\varphi(x, y) + \varphi(x - l, y)) \\ &+ k(\varphi(x, y + l) - 2\varphi(x, y) + \varphi(x, y - l)) \end{aligned}$$

(4) Now let us take the limit of our model, as to reach to continuum. For this purpose we will assume that our system consists of $B^2 >> 0$ balls, having a total mass M, and

spanning a total area L^2. Thus, our previous infinitesimal parameters are as follows, with K being the spring constant of the total system, taken to be equal to k:

$$m = \frac{M}{B^2} \quad , \quad k = K \quad , \quad l = \frac{L}{B}$$

With these changes, our equation of motion found in (3) reads:

$$\begin{aligned} \ddot{\varphi}(x,y) &= \frac{KB^2}{M}(\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)) \\ &+ \frac{KB^2}{M}(\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)) \end{aligned}$$

Now observe that this equation can be written, more conveniently, as follows:

$$\begin{aligned} \ddot{\varphi}(x,y) &= \frac{KL^2}{M} \times \frac{\varphi(x+l,y) - 2\varphi(x,y) + \varphi(x-l,y)}{l^2} \\ &+ \frac{KL^2}{M} \times \frac{\varphi(x,y+l) - 2\varphi(x,y) + \varphi(x,y-l)}{l^2} \end{aligned}$$

With $N \to \infty$, and therefore $l \to 0$, we obtain in this way:

$$\ddot{\varphi}(x,y) = \frac{KL^2}{M}\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y)$$

Thus, we are led in this way to the following wave equation in two dimensions, with $v = \sqrt{K/M} \cdot L$ being the propagation speed of our wave:

$$\ddot{\varphi}(x,y) = v^2\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y)$$

But we recognize at right the Laplace operator, and we are done. As before in 1D, there is of course some discussion to be made here, arguing that our spring model in (3) is indeed the correct one. But do not worry, experiments confirm our findings.

(5) In 3 dimensions now, which is the case of the main interest, corresponding to our real-life world, the same argument carries over, and the wave equation is as follows:

$$\ddot{\varphi}(x,y,z) = v^2\left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2} + \frac{d^2\varphi}{dz^2}\right)(x,y,z)$$

(6) Finally, the same argument, namely a lattice model, carries on in arbitrary N dimensions, and the wave equation here is as follows:

$$\ddot{\varphi}(x_1,\ldots,x_N) = v^2\sum_{i=1}^{N}\frac{d^2\varphi}{dx_i^2}(x_1,\ldots,x_N)$$

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

Moving on, with a bit of work, we can in fact fully solve the 1D wave equation. In order to explain this, we will need a standard calculus result, as follows:

PROPOSITION 12.15. *The derivative of a function of type*

$$\varphi(x)=\int_{g(x)}^{h(x)}f(s)ds$$

is given by the formula $\varphi'(x)=f(h(x))h'(x)-f(g(x))g'(x)$.

PROOF. Consider a primitive of the function that we integrate, $F'=f$. We have:

$$\begin{aligned}\varphi(x)&=\int_{g(x)}^{h(x)}f(s)ds\\&=\int_{g(x)}^{h(x)}F'(s)ds\\&=F(h(x))-F(g(x))\end{aligned}$$

By using now the chain rule for derivatives, we obtain from this:

$$\begin{aligned}\varphi'(x)&=F'(h(x))h'(x)-F'(g(x))g'(x)\\&=f(h(x))h'(x)-f(g(x))g'(x)\end{aligned}$$

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

Now back to the 1D waves, the general result here, due to d'Alembert, along with a little more, in relation with our lattice models above, is as follows:

THEOREM 12.16. *The solution of the 1D wave equation with initial value conditions* $\varphi(x,0)=f(x)$ *and* $\dot{\varphi}(x,0)=g(x)$ *is given by the d'Alembert formula, namely:*

$$\varphi(x,t)=\frac{f(x-vt)+f(x+vt)}{2}+\frac{1}{2v}\int_{x-vt}^{x+vt}g(s)ds$$

In the context of our lattice model discretizations, what happens is more or less that the above d'Alembert integral gets computed via Riemann sums.

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

(1) Let us first check that the d'Alembert solution is indeed a solution of the wave equation $\ddot{\varphi}=v^2\varphi''$. The first time derivative is computed as follows:

$$\dot{\varphi}(x,t)=\frac{-vf'(x-vt)+vf'(x+vt)}{2}+\frac{1}{2v}(vg(x+vt)+vg(x-vt))$$

The second time derivative is computed as follows:

$$\ddot{\varphi}(x,t)=\frac{v^2f''(x-vt)+v^2f(x+vt)}{2}+\frac{vg'(x+vt)-vg'(x-vt)}{2}$$

Regarding now space derivatives, the first one is computed as follows:

$$\varphi'(x,t)=\frac{f'(x-vt)+f'(x+vt)}{2}+\frac{1}{2v}(g'(x+vt)-g'(x-vt))$$

As for the second space derivative, this is computed as follows:

$$\varphi''(x,t)=\frac{f''(x-vt)+f''(x+vt)}{2}+\frac{g''(x+vt)-g''(x-vt)}{2v}$$

Thus we have indeed $\ddot{\varphi}=v^2\varphi''$. As for the initial conditions, $\varphi(x,0)=f(x)$ is clear from our definition of φ, and $\dot{\varphi}(x,0)=g(x)$ is clear from our above formula of $\dot{\varphi}$.

(2) Conversely now, we must show that our solution is unique, but instead of going here into abstract arguments, we will simply solve our equation, which among others will doublecheck out computations in (1). Let us make the following change of variables:

$$\xi=x-vt\quad,\quad\eta=x+vt$$

With this change of variables, which is quite tricky, mixing space and time variables, our wave equation $\ddot{\varphi}=v^2\varphi''$ reformulates in a very simple way, as follows:

$$\frac{d^2\varphi}{d\xi d\eta}=0$$

But this latter equation tells us that our new ξ,η variables get separated, and we conclude from this that the solution must be of the following special form:

$$\varphi(x,t)=F(\xi)+G(\eta)=F(x-vt)+G(x+vt)$$

Now by taking into account the intial conditions $\varphi(x,0)=f(x)$ and $\dot{\varphi}(x,0)=g(x)$, and then integrating, we are led to the d'Alembert formula in the statement.

(3) In regards now with our discretization questions, by using a 1D lattice model with balls and springs as before, what happens to all the above is more or less that the above d'Alembert integral gets computed via Riemann sums, in our model, as stated. □

In $N\geq 2$ dimensions things are considerably more complicated, and for more on all this, we refer to any reasonably advanced theoretical physics or PDE book.

12d. Into the heat

Let us discuss now the heat equation, again by using a lattice model. The general equation here is quite similar to the one for the waves, as follows:

THEOREM 12.17. *Heat diffusion in $\mathbb{R}^N$ is described by the heat equation*

$$\dot{\varphi}=\alpha\Delta\varphi$$

where $\alpha>0$ is the thermal diffusivity of the medium, and Δ is the Laplace operator.

PROOF. The study here is quite similar to the study of waves, as follows:

(1) To start with, as an intuitive explanation for the equation, since the second derivative φ'' in one dimension, or the quantity $\Delta\varphi$ in general, computes the average value of a function φ around a point, minus the value of φ at that point, the heat equation as formulated above tells us that the rate of change $\dot{\varphi}$ of the temperature of the material at

any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material.

(2) The heat equation as formulated above is of course something approximative, and several improvements can be made to it, first by incorporating a term accounting for heat radiation, and then doing several fine-tunings, depending on the material involved. But more on this later, for the moment let us focus on the heat equation above.

(3) In relation with our modelling questions, we can recover this equation a bit as we did for the wave equation before, by using a basic lattice model. Indeed, let us first assume, for simplifying, that we are in the one-dimensional case, $N = 1$. Here our model looks as follows, with distance $l > 0$ between neighbors:

$$\text{---} \circ_{x-l} \overset{l}{\text{---}} \circ_x \overset{l}{\text{---}} \circ_{x+l} \text{---}$$

In order to model heat diffusion, we have to implement the intuitive mechanism explained above, namely "the rate of change of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha > 0$, to the average difference of temperature between that given point and the surrounding material".

(4) In practice, this leads to a condition as follows, expressing the change of the temperature φ, over a small period of time $\delta > 0$:

$$\varphi(x, t + \delta) = \varphi(x, t) + \frac{\alpha\delta}{l^2} \sum_{x \sim y} [\varphi(y, t) - \varphi(x, t)]$$

To be more precise, we have made several assumptions here, as follows:

– General heat diffusion assumption: the change of temperature at any given point x is proportional to the average over neighbors, $y \sim x$, of the differences $\varphi(y, t) - \varphi(x, t)$ between the temperatures at x, and at these neighbors y.

– Infinitesimal time and length conditions: in our model, the change of temperature at a given point x is proportional to small period of time involved, $\delta > 0$, and is inverse proportional to the square of the distance between neighbors, l^2.

(5) Regarding these latter assumptions, the one regarding the proportionality with the time elapsed $\delta > 0$ is something quite natural, physically speaking, and mathematically speaking too, because we can rewrite our equation as follows, making it clear that we have here an equation regarding the rate of change of temperature at x:

$$\frac{\varphi(x, t + \delta) - \varphi(x, t)}{\delta} = \frac{\alpha}{l^2} \sum_{x \sim y} [\varphi(y, t) - \varphi(x, t)]$$

As for the second assumption that we made above, namely inverse proportionality with l^2, this can be justified on physical grounds too, but again, perhaps the best is to do the math, which will show right away where this proportionality comes from.

(6) So, let us do the math. In the context of our 1D model the neighbors of x are the points $x \pm l$, and so the equation that we wrote above takes the following form:

$$\frac{\varphi(x,t+\delta)-\varphi(x,t)}{\delta}=\frac{\alpha}{l^2}\Big[(\varphi(x+l,t)-\varphi(x,t))+(\varphi(x-l,t)-\varphi(x,t))\Big]$$

Now observe that we can write this equation as follows:

$$\frac{\varphi(x,t+\delta)-\varphi(x,t)}{\delta}=\alpha\cdot\frac{\varphi(x+l,t)-2\varphi(x,t)+\varphi(x-l,t)}{l^2}$$

(7) As it was the case with the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l\to 0$, we obtain the following equation:

$$\frac{\varphi(x,t+\delta)-\varphi(x,t)}{\delta}=\alpha\cdot\varphi''(x,t)$$

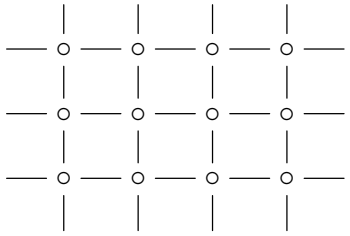
Now with $t\to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,t)=\alpha\cdot\varphi''(x,t)$$

Summarizing, we are done with the 1D case, with our proof being quite similar to the one for the wave equation, from the above.

(8) In practice now, there are of course still a few details to be discussed, in relation with all this, for instance at the end, in relation with the precise order of the limiting operations $l\to 0$ and $\delta\to 0$ to be performed, but these remain minor aspects, because our equation makes it clear, right from the beginning, that time and space are separated, and so that there is no serious issue with all this. And so, fully done with 1D.

(9) With this done, let us discuss now 2 dimensions. Here, as before for the waves, we can use a lattice model as follows, with all lengths being $l>0$, for simplifying:



(10) We have to implement now the physical heat diffusion mechanism, namely "the rate of change of the temperature of the material at any given point must be proportional, with proportionality factor $\alpha>0$, to the average difference of temperature between that given point and the surrounding material". In practice, this leads to a condition as follows,

expressing the change of the temperature φ, over a small period of time $\delta > 0$:

$$\varphi(x,y,t+\delta) = \varphi(x,y,t) + \frac{\alpha\delta}{l^2} \sum_{(x,y)\sim(u,v)} [\varphi(u,v,t) - \varphi(x,y,t)]$$

In fact, we can rewrite our equation as follows, making it clear that we have here an equation regarding the rate of change of temperature at x:

$$\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} = \frac{\alpha}{l^2} \sum_{(x,y)\sim(u,v)} [\varphi(u,v,t) - \varphi(x,y,t)]$$

(11) So, let us do the math. In the context of our 2D model the neighbors of x are the points $(x \pm l, y \pm l)$, so the equation above takes the following form:

$$\begin{aligned} & \frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} \\ = \ & \frac{\alpha}{l^2}\Big[(\varphi(x+l,y,t) - \varphi(x,y,t)) + (\varphi(x-l,y,t) - \varphi(x,y,t))\Big] \\ + \ & \frac{\alpha}{l^2}\Big[(\varphi(x,y+l,t) - \varphi(x,y,t)) + (\varphi(x,y-l,t) - \varphi(x,y,t))\Big] \end{aligned}$$

Now observe that we can write this equation as follows:

$$\begin{aligned} \frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} &= \alpha \cdot \frac{\varphi(x+l,y,t) - 2\varphi(x,y,t) + \varphi(x-l,y,t)}{l^2} \\ &+ \alpha \cdot \frac{\varphi(x,y+l,t) - 2\varphi(x,y,t) + \varphi(x,y-l,t)}{l^2} \end{aligned}$$

(12) As it was the case when modelling the wave equation before, we recognize on the right the usual approximation of the second derivative, coming from calculus. Thus, when taking the continuous limit of our model, $l \to 0$, we obtain the following equation:

$$\frac{\varphi(x,y,t+\delta) - \varphi(x,y,t)}{\delta} = \alpha \left(\frac{d^2\varphi}{dx^2} + \frac{d^2\varphi}{dy^2}\right)(x,y,t)$$

Now with $t \to 0$, we are led in this way to the heat equation, namely:

$$\dot{\varphi}(x,y,t) = \alpha \cdot \Delta\varphi(x,y,t)$$

Finally, in arbitrary N dimensions the same argument carries over, namely a straightforward lattice model, and gives the heat equation, as formulated in the statement. □

Observe that we can use if we want different lenghts $l > 0$ on the vertical and on the horizontal, because these will simplify anyway due to proportionality. Also, for some further mathematical fun, we can build our model on a cylinder, or a torus.

Also, as mentioned before, our heat equation above is something approximative, and several improvements can be made to it, first by incorporating a term accounting for heat

radiation, and also by doing several fine-tunings, depending on the material involved. Some of these improvements can be implemented in the lattice model setting.

Regarding now the mathematics of the heat equation, many things can be said. As a first result here, often used by mathematicians, as to assume $\alpha = 1$, we have:

PROPOSITION 12.18. *Up to a time rescaling, we can assume $\alpha = 1$, as to deal with*

$$\dot{\varphi} = \Delta\varphi$$

called normalized heat equation.

PROOF. This is clear physically speaking, because according to our model, changing the parameter $\alpha > 0$ will result in accelerating or slowing the heat diffusion, in time $t > 0$. Mathematically, this follows via a change of variables, for the time variable t. □

Regarding now the resolution of the heat equation, we have here:

THEOREM 12.19. *The heat equation, normalized as $\dot{\varphi} = \Delta\varphi$, and with initial condition $\varphi(x, 0) = f(x)$, has as solution the function*

$$\varphi(x, t) = (K_t * f)(x)$$

where the function $K_t : \mathbb{R}^N \to \mathbb{R}$, called heat kernel, is given by

$$K_t(x) = (4\pi t)^{-N/2} e^{-||x||^2/4t}$$

with $||x||$ being the usual norm of vectors $x \in \mathbb{R}^N$.

PROOF. According to the definition of the convolution operation $*$, we have to check that the following function satisfies $\dot{\varphi} = \Delta\varphi$, with initial condition $\varphi(x, 0) = f(x)$:

$$\varphi(x, t) = (4\pi t)^{-N/2} \int_{\mathbb{R}^N} e^{-||x-y||^2/4t} f(y) dy$$

But both checks are elementary, coming from definitions. □

Regarding now discretization questions, things here are quite tricky. The idea is to use the Central Limit Theorem (CLT) from probability theory, which is as follows:

THEOREM 12.20 (CLT). *Given random variables $f_1, f_2, f_3, \ldots \in L^\infty(X)$ which are i.i.d., centered, and with variance $t > 0$, we have, with $n \to \infty$, in moments,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} f_i \sim g_t$$

where g_t is the Gaussian law of parameter t, having as density $\frac{1}{\sqrt{2\pi t}} e^{-z^2/2t} dz$.

PROOF. To start with, thanks to the Gauss formula from chapter 9, we can talk indeed about the Gaussian law g_t of parameter $t > 0$, having as density $\frac{1}{\sqrt{2\pi t}} e^{-z^2/2t} dz$. Also, the Fourier transform of this Gaussian law g_t can be computed as follows:

$$\begin{aligned} F_{g_t}(x) &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-z^2/2t+ixz} dz \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-(z/\sqrt{2t}-\sqrt{t/2}\, iz)^2 - tx^2/2} dz \\ &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} e^{-y^2 - tx^2/2} \sqrt{2t}\, dy \\ &= \frac{1}{\sqrt{\pi}} e^{-tx^2/2} \int_{\mathbb{R}} e^{-y^2} dy \\ &= e^{-tx^2/2} \end{aligned}$$

Getting now to what we want to prove, observe first that in terms of moments, the Fourier transform $F_f(x) = E(e^{ixf})$ is given by the following formula:

$$\begin{aligned} F_f(x) &= E\left(\sum_{k=0}^{\infty} \frac{(ixf)^k}{k!} \right) \\ &= \sum_{k=0}^{\infty} \frac{(ix)^k E(f^k)}{k!} \\ &= \sum_{k=0}^{\infty} \frac{i^k M_k(f)}{k!}\, x^k \end{aligned}$$

Thus, the Fourier transform of the variable in the statement is:

$$\begin{aligned} F(x) &= \left[F_f\left(\frac{x}{\sqrt{n}} \right) \right]^n \\ &= \left[1 - \frac{tx^2}{2n} + O(n^{-2}) \right]^n \\ &\simeq \left[1 - \frac{tx^2}{2n} \right]^n \\ &\simeq e^{-tx^2/2} \end{aligned}$$

But this latter function being the Fourier transform of g_t, we obtain the result. □

With the above result in hand, complemented by its higher dimensional analogues, which follow from it, we can talk afterwards about discretizing the heat kernel. We will leave some exploration and reading here as an instructive exercise.

12e. Exercises

We had an exciting science chapter here, and as exercises on this, we have:

EXERCISE 12.21. *Count loops on graphs, as many as you can.*

EXERCISE 12.22. *Diagonalize adjacency matrices, as many as you can.*

EXERCISE 12.23. *Learn some other proofs of the Cayley formula for graphs.*

EXERCISE 12.24. *Clarify everything in relation with the Cauchy-Binet formula.*

EXERCISE 12.25. *Learn more, from physicists, about the various types of waves.*

EXERCISE 12.26. *Learn also, from mathematicians, how to deal with waves at* $N \geq 2$.

EXERCISE 12.27. *Learn more, from physicists, about the versions of the heat equation.*

EXERCISE 12.28. *Learn also, from mathematicians, how to deal with the heat equation.*

As bonus exercise, now that you know about graphs, read some design theory.

Part IV

Matrix groups

But here I am
Next to you
The sky is more blue
In Malibu

CHAPTER 13

Finite groups

13a. Finite groups

We discuss in this final Part IV the various groups that the matrices can form. Let us start with some basic group theory. As a beginning for everything, we have:

DEFINITION 13.1. *A group is a set G endowed with a multiplication operation*

$$(g,h)\to gh$$

which must satisfy the following conditions:

(1) *Associativity: we have, $(gh)k=g(hk)$, for any $g,h,k\in G$.*
(2) *Unit: there is an element $1\in G$ such that $g1=1g=g$, for any $g\in G$.*
(3) *Inverses: for any $g\in G$ there is $g^{-1}\in G$ such that $gg^{-1}=g^{-1}g=1$.*

The multiplication law is not necessarily commutative. In the case where it is, in the sense that $gh=hg$, for any $g,h\in G$, we call G abelian, en hommage to Abel, and we usually denote its multiplication, unit and inverse operation as follows:

$$(g,h)\to g+h\quad,\quad 0\in G\quad,\quad g\to -g$$

However, this is not a general rule, and rather the converse is true, in the sense that if a group is denoted as above, this means that the group must be abelian.

Let us work out now some examples, in the finite group case. The simplest finite group is the cyclic group $\mathbb{Z}_N$, which is something very familiar, constructed as follows:

DEFINITION 13.2. *The cyclic group $\mathbb{Z}_N$ is defined as follows:*

(1) *As the additive group of remainders modulo N.*
(2) *As the multiplicative group of the N-th roots of unity.*

Observe that the above two constructions are indeed equivalent, because if we set $w=e^{2\pi i/N}$, any remainder modulo N defines a N-th root of unity, according to:

$$k\to w^k$$

We obtain in this way all the N-roots of unity, so our correspondence is bijective. Moreover, our correspondence transforms the sum of remainders modulo N into the multiplication of the N-th roots of unity, due to the following formula:

$$w^kw^l=w^{k+l}$$

Thus, the groups constructed in Definition 13.2 (1) and (2) are indeed isomorphic, via $k \to w^k$, and we agree to denote by $\mathbb{Z}_N$ the corresponding group. Observe that this group $\mathbb{Z}_N$ is abelian. We will be back to the abelian groups later, on several occasions.

As a second basic example of a finite group, this time not abelian, we have the symmetric group S_N. This is again something very familiar, appearing as follows:

DEFINITION 13.3. *A permutation of $\{1, \ldots, N\}$ is a bijection, as follows:*

$$\sigma : \{1, \ldots, N\} \to \{1, \ldots, N\}$$

The set of such permutations is denoted S_N.

There are many possible notations for the permutations, the basic one consisting in writing the numbers $1, \ldots, N$, and below them, their permuted versions:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 5 & 3 \end{pmatrix}$$

Another method, which is faster, and that I personally prefer, remember that time is money, is by denoting the permutations as diagrams, acting from top to bottom:

$\sigma =$ 

Here are some basic properties of the permutations, that you surely know about:

THEOREM 13.4. *The permutations have the following properties:*

(1) *There are $N!$ of them.*
(2) *They from a group.*

PROOF. Indeed, in order to construct a permutation $\sigma \in S_N$, we have:

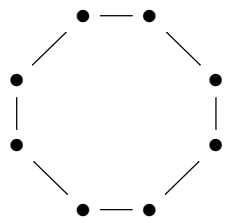
– N choices for the value of $\sigma(N)$.
– $(N-1)$ choices for the value of $\sigma(N-1)$.
– $(N-2)$ choices for the value of $\sigma(N-2)$.
$\vdots$
– and so on, up to 1 choice for the value of $\sigma(1)$.

Thus, we have $N!$ choices, as claimed. As for the second assertion, this is clear. □

The symmetric groups S_N are key objects of group theory, and they have many interesting properties. We will be back to them on many occasions, in what follows.

As a third and last basic example of a finite group, for our purposes here, which is something more advanced, we have the dihedral group D_N, which appears as follows:

DEFINITION 13.5. *The dihedral group D_N is the symmetry group of*



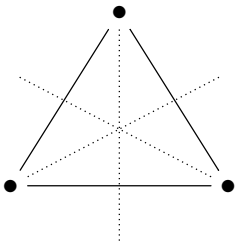
that is, of the regular polygon having N vertices.

In order to understand how this works, here are the basic examples of regular N-gons, at small values of the parameter $N \in \mathbb{N}$, along with their symmetry groups:

$\underline{N = 2}$. Here the N-gon is just a segment, and its symmetries are obviously the identity id, plus the symmetry τ with respect to the middle of the segment:

Thus we have $D_2 = \{id, \tau\}$, which in group theory terms means $D_2 = \mathbb{Z}_2$.

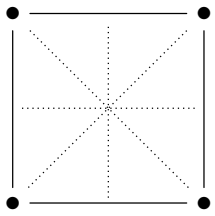
$\underline{N = 3}$. Here the N-gon is an equilateral triangle, and we have 6 symmetries, the rotations of angles $0°$, $120°$, $240°$, and the symmetries with respect to the altitudes:



Alternatively, we can say that the symmetries are all the $3! = 6$ possible permutations of the vertices, and so that in group theory terms, we have $D_3 = S_3$.

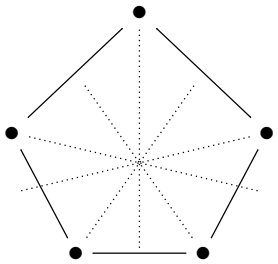
$\underline{N = 4}$. Here the N-gon is a square, and as symmetries we have 4 rotations, of angles $0°, 90°, 180°, 270°$, as well as 4 symmetries, with respect to the 4 symmetry axes, which

are the 2 diagonals, and the 2 segments joining the midpoints of opposite sides:

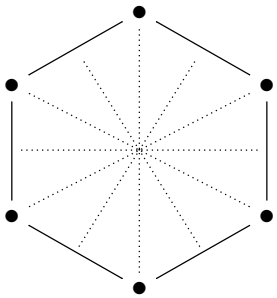


Thus, we obtain as symmetry group some sort of product between $\mathbb{Z}_4$ and $\mathbb{Z}_2$. Observe however that this product is not the usual one, our group being not abelian.

$\underline{N=5}$. Here the N-gon is a regular pentagon, and as symmetries we have 5 rotations, of angles $0°, 72°, 144°, 216°, 288°$, as well as 5 symmetries, with respect to the 5 symmetry axes, which join the vertices to the midpoints of the opposite sides:



$\underline{N=6}$. Here the N-gon is a regular hexagon, and we have 6 rotations, of angles $0°, 60°, 120°, 180°, 240°, 300°$, and 6 symmetries, with respect to the 6 symmetry axes, which are the 3 diagonals, and the 3 segments joining the midpoints of opposite sides:



$\underline{N=7}$. Here the N-gon is a regular heptagon, and we have 7 rotations, of angles kt with $k = 0, 1, \ldots, 6$ and $t = 360°/7$, as well as 7 symmetries, with respect to the 7 symmetry axes, which join the vertices to the midpoints of the opposite sides.

We can see from the above that the various dihedral groups D_N have many common features, and that there are some differences as well, basically coming from the parity of $N \in \mathbb{N}$. In general, we first have the following result, regarding the general case:

PROPOSITION 13.6. *The dihedral group D_N has $2N$ elements, as follows:*

(1) *We have N rotations $R_1, \ldots, R_N$, with R_k being the rotation of angle $2k\pi/N$. When labeling the vertices of the N-gon $1, \ldots, N$, we have $R_k(i) = k + i$.*

(2) *We have N symmetries $S_1, \ldots, S_N$, with S_k being the symmetry with respect to the Ox axis rotated by $k\pi/N$. The symmetry formula is $S_k(i) = k - i$.*

PROOF. This is clear from definitions. To be more precise, D_N consists of:

(1) The N rotations of the N-gon, of angles $2k\pi/N$ with $k = 1, \ldots, N$. But these are exactly the rotations $R_1, \ldots, R_N$ from the statement.

(2) The N symmetries of the N-gon, with respect to the N medians when N is odd, and with the $N/2$ diagonals plus the $N/2$ lines connecting the midpoints of opposite edges, when N is even. But these are the symmetries $S_1, \ldots, S_N$ from the statement. □

With the above description of D_N in hand, we can forget if we want about geometry and the regular N-gon, and talk about D_N abstractly, as follows:

THEOREM 13.7. *The dihedral group D_N is the group having $2N$ elements, $R_1, \ldots, R_N$ and $S_1, \ldots, S_N$, called rotations and symmetries, which multiply as follows,*

$$R_k R_l = R_{k+l} \quad , \quad R_k S_l = S_{k+l}$$
$$S_k R_l = S_{k-l} \quad , \quad S_k S_l = R_{k-l}$$

with all the indices being taken modulo N.

PROOF. With notations from Proposition 13.6, the various compositions between rotations and symmetries can be computed as follows:

$$R_k R_l \ : \ i \to l + i \to k + l + i$$
$$R_k S_l \ : \ i \to l - i \to k + l - i$$
$$S_k R_l \ : \ i \to l + i \to k - l - i$$
$$S_k S_l \ : \ i \to l - i \to k - l + i$$

But these are exactly the formulae for $R_{k+l}, S_{k+l}, S_{k-l}, R_{k-l}$, as stated. Now since a group is uniquely determined by its multiplication rules, this gives the result. □

Observe now that D_N has the same cardinality as $E_N = \mathbb{Z}_N \times \mathbb{Z}_2$. We obviously don't have $D_N \simeq E_N$, because D_N is not abelian, while E_N is. So, our next goal will be that of proving that D_N appears by "twisting" E_N. In order to do this, let us start with:

PROPOSITION 13.8. *The group $E_N = \mathbb{Z}_N \times \mathbb{Z}_2$ is the group having $2N$ elements, $r_1, \ldots, r_N$ and $s_1, \ldots, s_N$, which multiply according to the following rules,*

$$r_k r_l = r_{k+l} \quad , \quad r_k s_l = s_{k+l}$$
$$s_k r_l = s_{k+l} \quad , \quad s_k s_l = r_{k+l}$$

with all the indices being taken modulo N.

PROOF. With the notation $\mathbb{Z}_2 = \{1, \tau\}$, the elements of the product group $E_N = \mathbb{Z}_N \times \mathbb{Z}_2$ can be labeled $r_1, \ldots, r_N$ and $s_1, \ldots, s_N$, as follows:

$$r_k = (k, 1) \quad , \quad s_k = (k, \tau)$$

These elements multiply then according to the formulae in the statement. Now since a group is uniquely determined by its multiplication rules, this gives the result. □

Let us compare now Theorem 13.7 and Proposition 13.8. In order to formally obtain D_N from E_N, we must twist some of the multiplication rules of E_N, namely:

$$s_k r_l = s_{k+l} \to s_{k-l}$$

$$s_k s_l = r_{k+l} \to r_{k-l}$$

Informally, this amounts in following the rule "τ switches the sign of what comes afterwards", and we are led in this way to the following definition:

DEFINITION 13.9. *Given two groups A, G, with an action $A \curvearrowright G$, the crossed product*

$$P = G \rtimes A$$

is the set $G \times A$, with multiplication $(g, a)(h, b) = (gh^a, ab)$.

It is routine to check that P is indeed a group. Observe that when the action is trivial, $h^a = h$ for any $a \in A$ and $h \in H$, we obtain the usual product $G \times A$.

Now with this technology in hand, by getting back to the dihedral group D_N, we can improve Theorem 13.7, into a final result on the subject, as follows:

THEOREM 13.10. *We have a crossed product decomposition as follows,*

$$D_N = \mathbb{Z}_N \rtimes \mathbb{Z}_2$$

with $\mathbb{Z}_2 = \{1, \tau\}$ acting on $\mathbb{Z}_N$ via switching signs, $k^\tau = -k$.

PROOF. We have an action $\mathbb{Z}_2 \curvearrowright \mathbb{Z}_N$ given by the formula in the statement, namely $k^\tau = -k$, so we can consider the corresponding crossed product group:

$$P_N = \mathbb{Z}_N \rtimes \mathbb{Z}_2$$

In order to understand the structure of P_N, we follow Proposition 13.8. The elements of P_N can indeed be labeled $\rho_1, \ldots, \rho_N$ and $\sigma_1, \ldots, \sigma_N$, as follows:

$$\rho_k = (k, 1) \quad , \quad \sigma_k = (k, \tau)$$

Now when computing the products of such elements, we basically obtain the formulae in Proposition 13.8, perturbed as in Definition 13.9. To be more precise, we have:

$$\rho_k \rho_l = \rho_{k+l} \quad , \quad \rho_k \sigma_l = \sigma_{k+l}$$

$$\sigma_k \rho_l = \sigma_{k+l} \quad , \quad \sigma_k \sigma_l = \rho_{k+l}$$

But these are exactly the multiplication formulae for D_N, from Theorem 13.7. Thus, we have an isomorphism $D_N \simeq P_N$ given by $R_k \to \rho_k$ and $S_k \to \sigma_k$, as desired. □

13b. Symmetric groups

We have seen some basic group theory, but you might wonder where is the linear algebra, in relation with this. In order to get to this, following Cayley, we first have:

THEOREM 13.11. *Given a finite group G, we have an embedding as follows,*

$$G \subset S_N \quad , \quad g \to (h \to gh)$$

with $N = |G|$. Thus, any finite group is a permutation group.

PROOF. Given a group element $g \in G$, we can associate to it the following map:

$$\sigma_g : G \to G \quad , \quad h \to gh$$

Since $gh = gh'$ implies $h = h'$, this map is bijective, and so is a permutation of G, viewed as a set. Thus, with $N = |G|$, we can view this map as a usual permutation, $\sigma_G \in S_N$. Summarizing, we have constructed so far a map as follows:

$$G \to S_N \quad , \quad g \to \sigma_g$$

Our first claim is that this is a group morphism. Indeed, this follows from:

$$\sigma_g\sigma_h(k) = \sigma_g(hk) = ghk = \sigma_{gh}(k)$$

It remains to prove that this group morphism is injective. But this follows from:

$$\begin{aligned} g \neq h &\implies \sigma_g(1) \neq \sigma_h(1) \\ &\implies \sigma_g \neq \sigma_h \end{aligned}$$

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

Observe that in the above statement the embedding $G \subset S_N$ that we constructed depends on a particular writing $G = \{g_1, \ldots, g_N\}$, which is needed in order to identify the permutations of G with the elements of the symmetric group S_N. This is not very good, in practice, and as an illustration, for the basic examples of groups that we know, the Cayley theorem provides us with embeddings as follows:

$$\mathbb{Z}_N \subset S_N \quad , \quad D_N \subset S_{2N} \quad , \quad S_N \subset S_{N!}$$

And here the first embedding is the good one, the second one is not the best possible one, but can be useful, and the third embedding is useless. Thus, as a conclusion, the Cayley theorem remains something quite theoretical. We will be back to this later on, with a systematic study of the "representation" problem for the finite groups.

Getting back now to our main series of finite groups, $\mathbb{Z}_N \subset D_N \subset S_N$, these are of course permutation groups, according to the above. However, and perhaps even more interestingly, these are as well subgroups of the orthogonal group O_N:

$$\mathbb{Z}_N \subset D_N \subset S_N \subset O_N$$

In order to explain this, we first have the following key result:

THEOREM 13.12. *We have a group embedding as follows, obtained by regarding S_N as the permutation group of the N coordinate axes of $\mathbb{R}^N$,*

$$S_N \subset O_N$$

which makes $\sigma \in S_N$ correspond to the matrix having 1 on row i and column $\sigma(i)$, for any i, and having 0 entries elsewhere.

PROOF. The first assertion is clear, because the permutations of the N coordinate axes of $\mathbb{R}^N$ are isometries. Regarding now the explicit formula, we have by definition:

$$\sigma(e_j) = e_{\sigma(j)}$$

Thus, the permutation matrix corresponding to σ is given by:

$$\sigma_{ij} = \begin{cases} 1 & \text{if } \sigma(j) = i \\ 0 & \text{otherwise} \end{cases}$$

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

We can combine the above result with the Cayley theorem, and we obtain:

THEOREM 13.13. *Given a finite group G, we have an embedding as follows,*

$$G \subset O_N \quad , \quad g \to (e_h \to e_{gh})$$

with $N = |G|$. Thus, any finite group is an orthogonal matrix group.

PROOF. The Cayley theorem gives an embedding as follows:

$$G \subset S_N \quad , \quad g \to (h \to gh)$$

On the other hand, Theorem 13.12 provides us with an embedding as follows:

$$S_N \subset O_N \quad , \quad \sigma \to (e_i \to e_{\sigma(i)})$$

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

The same remarks as for the Cayley theorem apply. First, the embedding $G \subset O_N$ that we constructed depends on a particular writing $G = \{g_1, \ldots, g_N\}$. And also, for the basic examples of groups that we know, the embeddings that we obtain are as follows:

$$\mathbb{Z}_N \subset O_N \quad , \quad D_N \subset O_{2N} \quad , \quad S_N \subset O_{N!}$$

Summarizing, all this is not very good, and in order to advance, it is better to forget about the Cayley theorem, and build on Theorem 13.12 instead. We have here:

THEOREM 13.14. *We have the following finite groups of matrices:*

(1) *$\mathbb{Z}_N \subset O_N$, the cyclic permutation matrices.*
(2) *$D_N \subset O_N$, the dihedral permutation matrices.*
(3) *$S_N \subset O_N$, the permutation matrices.*

PROOF. This is something self-explanatory, the idea being that Theorem 13.12 provides us with embeddings as follows, given by the permutation matrices:

$$\mathbb{Z}_N \subset D_N \subset S_N \subset O_N$$

In practice now, the groups in the statement appear as follows:

(1) The cyclic permutation matrices are by definition the matrices as follows, with 0 entries elsewhere, and form a group, which is isomorphic to the cyclic group $\mathbb{Z}_N$:

$$U = \begin{pmatrix} & & 1 & & \\ & & & 1 & \\ & & & & \ddots \\ & & & & & 1 \\ 1 & & & & \\ & \ddots & & & \\ & & 1 & & \end{pmatrix}$$

(2) The dihedral matrices are the above cyclic permutation matrices, plus some suitable symmetry permutation matrices, and form a group which is isomorphic to D_N.

(3) The permutation matrices, which by Theorem 13.12 form a group which is isomorphic to S_N, are the $0-1$ matrices having exactly one 1 on each row and column. □

All the above is very nice, and as an informal conclusion to this, let us record:

CONCLUSION 13.15. *In what regards the theory of finite groups:*

(1) *The central object is the symmetric group, $S_N \subset O_N$.*
(2) *All basic finite groups appear as subgroups $G \subset O_N$.*
(3) *Thus, the objects of study are the finite subgroups $G \subset U_N$.*

To be more precise here, we have declared in (1) the symmetric group S_N to be the most important group, and we have also suggested to look at it as a subgroup $S_N \subset O_N$, in view of the above results. Regarding (2), this is actually something which theoretically happens for any finite group, as explained above, and with the basic examples truly appearing as $G \subset O_N$, by construction. As for (3), this is a version of (2) that we have not talked about yet, but that we can expect to happen, because it is not hard to imagine that certain more complicated groups naturally appear as subgroups $G \subset U_N$.

Very nice all this, and good to know, but the problem is now, what to do with our finite groups, $S_N \subset O_N$, or more generally $G \subset O_N$, or even more generally $G \subset U_N$?

And here, you must agree with me that the answer is not very clear. A look at the group theory literature suggests doing a myriad technical things, which are all useful of course, but with none being really elementary, and inviting. So, time to ask the cat:

CAT 13.16. *We cats are interested indeed in the subgroups $G \subset U_N$, which can be finite or not, and what we do with them is to compute the law of $g \to Tr(g)$.*

Thanks cat, this sounds quite interesting, and related indeed to linear algebra, as I was wishing for, in view of the purposes of the present book. So, let us study this problem, the linear algebra of the finite group elements, viewed as matrices, $g \in G \in U_N$.

As a first observation, before talking about the trace, as cat suggests, we can have a look at the determinant, with the result here, which is good to know, being as follows:

PROPOSITION 13.17. *The determinant of the permutation matrices*

$$\det : S_N \subset O_N \to \mathbb{R}$$

is the signature, $\varepsilon : S_N \to \mathbb{Z}_2$. Thus, we have $S_N \cap SO_N = A_N$, inside O_N.

PROOF. The first assertion comes from the Sarrus type formula for the determinant from chapter 2, in terms of permutations and their signatures, which in the case of the permutation matrices simply gives $\det = \varepsilon$. As for the second assertion, this follows from this, by taking the preimage via the application $\det = \varepsilon$ of the trivial group $\{1\}$. □

Getting now to what cat says, study of the trace, again in the simplest case, that of $S_N \subset O_N$, things here become truly interesting, with the result being as follows:

THEOREM 13.18. *The trace of permutation matrices, regarded as function*

$$\chi : S_N \to \mathbb{N} \quad , \quad \chi(\sigma) = Tr(\sigma)$$

counts the number of fixed points, according to the following formula:

$$\chi(\sigma) = \left\{ i \in \{1, \ldots, N\} \middle| \sigma(i) = i \right\}$$

Moreover, the variable $\chi : S_N \to \mathbb{N}$ follows the Poisson law of parameter 1,

$$p_1 = \frac{1}{e} \sum_{k \in \mathbb{N}} \frac{\delta_k}{k!}$$

in the $N \to \infty$ limit.

PROOF. We have several things going on here, the idea being as follows:

(1) Consider indeed the trace of permutation matrices, $\chi(\sigma) = Tr(\sigma)$. The permutation matrices being given by $\sigma_{ij} = \delta_{i\sigma(j)}$, we have the following formula, as claimed:

$$\chi(\sigma) = \sum_i \delta_{i\sigma(i)} = \# \left\{ i \in \{1, \ldots, N\} \middle| \sigma(i) = i \right\}$$

(2) In order to prove now the second assertion, consider the following subsets of S_N:

$$S_N^i = \left\{ \sigma \in S_N \middle| \sigma(i) = i \right\}$$

The set of permutations having no fixed points, called derangements, is then:

$$X_N = \left(\bigcup_i S_N^i \right)^c$$

Now the inclusion-exclusion principle tells us that we have:

$$\begin{aligned}
|X_N| &= \left| \left(\bigcup_i S_N^i \right)^c \right| \\
&= |S_N| - \sum_i |S_N^i| + \sum_{i<j} |S_N^i \cap S_N^j| - \ldots + (-1)^N \sum_{i_1<\ldots<i_N} |S_N^{i_1} \cap \ldots \cap S_N^{i_N}| \\
&= N! - N(N-1)! + \binom{N}{2}(N-2)! - \ldots + (-1)^N \binom{N}{N}(N-N)! \\
&= \sum_{k=0}^{N} (-1)^k \binom{N}{k} (N-k)! \\
&= \sum_{k=0}^{N} (-1)^k \frac{N!}{k!}
\end{aligned}$$

We conclude that the first probability that we are interested in, that for a random permutation $\sigma \in S_N$ to have no fixed points, is given by the following formula:

$$P(\chi = 0) = \frac{|X_N|}{N!} = \sum_{k=0}^{N} \frac{(-1)^k}{k!} \simeq \frac{1}{e}$$

(3) Thus, we are on the good way for establishing the second assertion, and in order now to fully do that, we just have to fine-tune our computation. To be more precise, we would like to prove the following formula, for any $r \in \mathbb{N}$, in the $N \to \infty$ limit:

$$P(\chi = r) \simeq \frac{1}{r!e}$$

(4) We already know, from the above, that this formula holds at $r = 0$. In the general case now, we have to count the permutations $\sigma \in S_N$ having exactly r points. Now since having such a permutation amounts in choosing r points among $1, \ldots, N$, and then

permuting the $N-r$ points left, without fixed points allowed, we have:

$$\begin{aligned}\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\} &= \binom{N}{r}\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}\\ &= \frac{N!}{r!(N-r)!}\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}\\ &= N!\times\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}\end{aligned}$$

By dividing everything by $N!$, we obtain from this the following formula:

$$\frac{\#\left\{\sigma\in S_N\Big|\chi(\sigma)=r\right\}}{N!}=\frac{1}{r!}\times\frac{\#\left\{\sigma\in S_{N-r}\Big|\chi(\sigma)=0\right\}}{(N-r)!}$$

Now by using the computation at $r=0$, that we already have, from (2), it follows that with $N\to\infty$ we have the following estimate:

$$P(\chi=r)\simeq\frac{1}{r!}\cdot P(\chi=0)\simeq\frac{1}{r!}\cdot\frac{1}{e}$$

Thus, we obtain as limiting measure the Poisson law of parameter 1, as stated. □

The above result looks quite exciting, and as further good news, that is not the end of what can be said, because we have also the following result, which is more general:

THEOREM 13.19. *The truncated trace of permutation matrices, defined as*

$$\chi_t:S_N\to\mathbb{R}\quad,\quad\chi(\sigma)=\sum_{i=1}^{[tN]}\sigma_{ii}$$

counts the truncated number of fixed points, according to the following formula:

$$\chi_t(\sigma)=\left\{i\in\{1,\ldots,[tN]\}\Big|\sigma(i)=i\right\}$$

Moreover, the variable $\chi_t:S_N\to\mathbb{R}$ follows the Poisson law of parameter t,

$$p_t=\frac{1}{e^t}\sum_{k\in\mathbb{N}}\frac{t^k\delta_k}{k!}$$

in the $N\to\infty$ limit.

PROOF. We have several things going on here, the idea being as follows:

(1) Let us construct indeed the truncated trace, as in the statement. As before in the case $t=1$, we have the following computation, coming from definitions:

$$\chi_t(\sigma)=\sum_{i=1}^{[tN]}\delta_{i\sigma(i)}=\#\left\{i\in\{1,\ldots,[tN]\}\Big|\sigma(i)=i\right\}$$

(2) Also as before at $t=1$, we obtain by inclusion-exclusion that:

$$\begin{aligned}P(\chi_t=0) &= \frac{1}{N!}\sum_{k=0}^{[tN]}(-1)^k\sum_{i_1<\ldots<i_k<[tN]}|S_N^{i_1}\cap\ldots\cap S_N^{i_k}|\\ &= \frac{1}{N!}\sum_{k=0}^{[tN]}(-1)^k\binom{[tN]}{k}(N-k)!\\ &= \sum_{k=0}^{[tN]}\frac{(-1)^k}{k!}\cdot\frac{[tN]!(N-k)!}{N!([tN]-k)!}\end{aligned}$$

Now with $N\to\infty$, we obtain from this the following estimate:

$$P(\chi_t=0)\simeq\sum_{k=0}^{[tN]}\frac{(-1)^k}{k!}\cdot t^k\simeq e^{-t}$$

(3) More generally now, by counting the permutations $\sigma\in S_N$ having exactly r fixed points among $1,\ldots,[tN]$, again as in the previous proof at $t=1$, we obtain:

$$P(\chi_t=r)\simeq\frac{t^r}{r!e^t}$$

Thus, we obtain in the limit a Poisson law of parameter t, as stated. □

13c. Reflection groups

The above results regarding the symmetric group S_N are quite exciting, and it is tempting to keep going this way, with similar results for $\mathbb{Z}_N$ and D_N. We first have:

PROPOSITION 13.20. *The main character of $\mathbb{Z}_N\subset O_N$ is given by:*

$$\chi(g)=\begin{cases}0 & \text{if } g\neq 1\\ N & \text{if } g=1\end{cases}$$

Thus, at the probabilistic level, we have the following formula,

$$law(\chi)=\left(1-\frac{1}{N}\right)\delta_0+\frac{1}{N}\delta_N$$

telling us that the main character χ follows a Bernoulli law.

PROOF. The first formula is clear, because the cyclic permutation matrices have 0 on the diagonal, and so 0 as trace, unless the matrix is the identity, having trace N. As for the second formula, this is a probabilistic reformulation of the first one. □

For the dihedral group now, which is the next one in our hierarchy, the computation is more interesting, but the final answer is no longer uniform in N, as follows:

PROPOSITION 13.21. *For the dihedral group $D_N \subset S_N$ we have*

$$law(\chi) = \begin{cases} \left(\frac{3}{4} - \frac{1}{2N}\right)\delta_0 + \frac{1}{4}\delta_2 + \frac{1}{2N}\delta_N & (N \text{ even}) \\ \\ \left(\frac{1}{2} - \frac{1}{2N}\right)\delta_0 + \frac{1}{2}\delta_1 + \frac{1}{2N}\delta_N & (N \text{ odd}) \end{cases}$$

with this law having no asymptotics, with $N \to \infty$.

PROOF. This follows indeed from the fact that the dihedral group D_N consists of:

(1) N symmetries, having each 1 fixed point when N is odd, and having 0 or 2 fixed points, distributed $50 - 50$, when N is even.

(2) N rotations, each having 0 fixed points, except for the identity, which is technically a rotation too, and which has N fixed points.

Thus, we are led to the formula in the statement, and to the final conclusion too. □

All the above does not look very good, so I am afraid that we are a bit stuck with our program, and that I will have to ask again the cat. And cat answers:

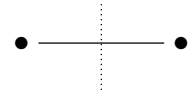
CAT 13.22. *There are many groups $G \subset U_N$, some good, some bad. Try doing your computations for groups which really look good, geometrically speaking.*

Thanks cat, so forgetting now about $\mathbb{Z}_N$, D_N, and other abstract groups and mathematics that we might know, let us think deeply, and try to find a group $G \subset U_N$ which is really interesting, and that we would really like to know about. And here, we have:

DEFINITION 13.23. *The hyperoctahedral group $H_N \subset O_N$ is the group of symmetries of the unit cube in $\mathbb{R}^N$, viewed as a graph, or equivalently, as a metric space.*

The hyperoctahedral group is a quite interesting group, whose definition, as a symmetry group, reminds that of the dihedral group D_N. So, let us start our study in the same way as we did for D_N, with a discussion at small values of $N \in \mathbb{N}$:

$\underline{N = 1}$. Here the 1-cube is the segment, whose symmetries are the identity id, plus the symmetry τ with respect to the middle of the segment:

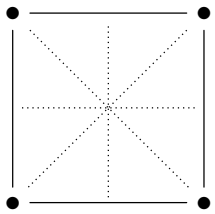


Thus, we obtain the group with 2 elements, which is a very familiar object:

$$H_1 = D_2 = S_2 = \mathbb{Z}_2$$

$\underline{N = 2}$. Here the 2-cube is the square, whose symmetries are the 4 rotations, of angles $0°, 90°, 180°, 270°$, and the 4 symmetries with respect to the 4 symmetry axes, which are

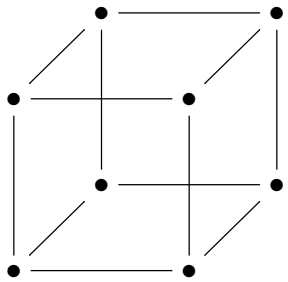
the 2 diagonals, and the 2 segments joining the midpoints of opposite sides:



Thus, we obtain a group with 8 elements, which again is a very familiar object:

$$H_2 = D_4 = \mathbb{Z}_4 \rtimes \mathbb{Z}_2$$

$\underline{N = 3}$. Here the 3-cube is the usual cube in $\mathbb{R}^3$, pictured as follows:



However, in relation with the symmetries, the situation now is considerably more complicated, because, thinking well, this cube has no less than 48 symmetries.

All this looks quite complicated, but fortunately, we have the following result:

PROPOSITION 13.24. *We have the cardinality formula*

$$|H_N| = 2^N N!$$

coming from the fact that H_N is the symmetry group of the coordinate axes of $\mathbb{R}^N$.

PROOF. This follows from some geometric thinking, as follows:

(1) Consider the standard cube in $\mathbb{R}^N$, centered at 0, and having as vertices the points having coordinates ± 1. With this picture in hand, it is clear that the symmetries of the cube coincide with the symmetries of the N coordinate axes of $\mathbb{R}^N$.

(2) In order to count now these latter symmetries, a bit as we did for the dihedral group, observe first that we have $N!$ permutations of these N coordinate axes.

(3) But each of these permutations of the coordinate axes $\sigma \in S_N$ can be further "decorated" by a sign vector $e \in \{\pm 1\}^N$, consisting of the possible ± 1 flips which can be applied to each coordinate axis, at the arrival.

(4) And the point is that, obviously, we obtain in this way all the elements of H_N. Thus, we have the following formula, for the cardinality of H_N:

$$|H_N| = |S_N| \cdot |\mathbb{Z}_2^N| = N! \cdot 2^N$$

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

As in the dihedral group case, it is possible to go beyond this, with a crossed product decomposition, of quite special type, called wreath product decomposition:

THEOREM 13.25. *We have a wreath product decomposition as follows,*

$$H_N = \mathbb{Z}_2 \wr S_N$$

which means by definition that we have a crossed product decomposition

$$H_N = \mathbb{Z}_2^N \rtimes S_N$$

with the permutations $\sigma \in S_N$ acting on the elements $e \in \mathbb{Z}_2^N$ as follows:

$$\sigma(e_1, \ldots, e_k) = (e_{\sigma(1)}, \ldots, e_{\sigma(k)})$$

In particular we have, as found before, the cardinality formula $|H_N| = 2^N N!$.

PROOF. As explained in the proof of Proposition 13.24, the elements of H_N can be identified with the pairs $g = (e, \sigma)$ consisting of a permutation $\sigma \in S_N$, and a sign vector $e \in \mathbb{Z}_2^N$, so that at the level of the cardinalities, we have the following formula:

$$|H_N| = |\mathbb{Z}_2^N \times S_N|$$

To be more precise, given an element $g \in H_N$, the element $\sigma \in S_N$ is the corresponding permutation of the N coordinate axes, regarded as unoriented lines in $\mathbb{R}^N$, and $e \in \mathbb{Z}_2^N$ is the vector collecting the possible flips of these coordinate axes, at the arrival. Now observe that the product formula for two such pairs $g = (e, \sigma)$ is as follows, with the permutations $\sigma \in S_N$ acting on the elements $f \in \mathbb{Z}_2^N$ as in the statement:

$$(e, \sigma)(f, \tau) = (ef^\sigma, \sigma\tau)$$

Thus, we are precisely in the framework of the crossed products, as constructed in chapter 1, and we conclude that we have a crossed product decomposition, as follows:

$$H_N = \mathbb{Z}_2^N \rtimes S_N$$

Thus, we are led to the conclusion in the statement, with the formula $H_N = \mathbb{Z}_2 \wr S_N$ being just a shorthand for the decomposition $H_N = \mathbb{Z}_2^N \rtimes S_N$ that we found. □

Regarding now the trace laws, we can compute them by using the same method as for the symmetric group S_N, namely inclusion-exclusion, and we have:

THEOREM 13.26. *For the hyperoctahedral group $H_N \subset O_N$, the law of the variable*

$$\chi_t = \sum_{i=1}^{[tN]} g_{ii}$$

becomes in the $N \to \infty$ limit the measure

$$b_t = e^{-t} \sum_{k=-\infty}^{\infty} \delta_k \sum_{p=0}^{\infty} \frac{(t/2)^{|k|+2p}}{(|k|+p)!p!}$$

called Bessel law of parameter $t > 0$.

PROOF. We regard H_N as being the symmetry group of the graph $I_N = \{I^1, \ldots, I^N\}$ formed by N segments. The diagonal coefficients are then given by:

$$g_{ii}(g) = \begin{cases} 0 & \text{if } g \text{ moves } I^i \\ 1 & \text{if } g \text{ fixes } I^i \\ -1 & \text{if } g \text{ returns } I^i \end{cases}$$

We denote by $\uparrow g, \downarrow g$ the number of segments among $\{I^1, \ldots, I^s\}$ which are fixed, respectively returned by an element $g \in H_N$. With this notation, we have:

$$g_{11} + \ldots + g_{ss} = \uparrow g - \downarrow g$$

Let us denote by P_N probabilities computed over the group H_N. The density of the law of $g_{11} + \ldots + g_{ss}$ at a point $k \geq 0$ is then given by the following formula:

$$\begin{aligned} D(k) &= P_N(\uparrow g - \downarrow g = k) \\ &= \sum_{p=0}^{\infty} P_N(\uparrow g = k + p, \downarrow g = p) \end{aligned}$$

Assume first that we have $t = 1$. We have here the following computation:

$$\begin{aligned} \lim_{N\to\infty} D(k) &= \lim_{N\to\infty} \sum_{p=0}^{\infty} (1/2)^{k+2p} \binom{k+2p}{k+p} P_N(\uparrow g + \downarrow g = k + 2p) \\ &= \sum_{p=0}^{\infty} (1/2)^{k+2p} \binom{k+2p}{k+p} \frac{1}{e(k+2p)!} \\ &= \frac{1}{e} \sum_{p=0}^{\infty} \frac{(1/2)^{k+2p}}{(k+p)!p!} \end{aligned}$$

As for the general case $0 < t \leq 1$, here the result follows by performing some modifications in the above computation. The asymptotic density is computed as follows:

$$\begin{aligned}\lim_{N\to\infty} D(k) &= \lim_{N\to\infty}\sum_{p=0}^{\infty}(1/2)^{k+2p}\binom{k+2p}{k+p}P_N(\uparrow g+\downarrow g=k+2p)\\ &= \sum_{p=0}^{\infty}(1/2)^{k+2p}\binom{k+2p}{k+p}\frac{t^{k+2p}}{e^t(k+2p)!}\\ &= e^{-t}\sum_{p=0}^{\infty}\frac{(t/2)^{k+2p}}{(k+p)!p!}\end{aligned}$$

Together with $D(-k)=D(k)$, this gives the formula in the statement. □

In the above result the terminology comes from the fact, up to $t\to t/2$, the density of the law is the following function, called Bessel function of the first kind:

$$f_k(t)=\sum_{p=0}^{\infty}\frac{t^{|k|+2p}}{(|k|+p)!p!}$$

Let us further study now these Bessel laws. A key result regarding the Poisson laws is the semigroup formula $p_s * p_t = p_{s+t}$, and in analogy with this, we have:

THEOREM 13.27. *The Bessel laws b_t have the property*

$$b_s * b_t = b_{s+t}$$

so they form a truncated one-parameter semigroup with respect to convolution.

PROOF. We use the formula that we found in Theorem 13.26, written as:

$$b_t = e^{-t}\sum_{k=-\infty}^{\infty}\delta_k\, f_k(t/2)$$

The Fourier transform of this measure is then given by:

$$Fb_t(y) = e^{-t}\sum_{k=-\infty}^{\infty}e^{ky}\, f_k(t/2)$$

We compute now the derivative with respect to t, as follows:

$$Fb_t(y)' = -Fb_t(y)+\frac{e^{-t}}{2}\sum_{k=-\infty}^{\infty}e^{ky}\, f_k'(t/2)$$

On the other hand, the derivative of f_k with $k \geq 1$ is given by:

$$\begin{aligned}
f_k'(t) &= \sum_{p=0}^{\infty} \frac{(k+2p)t^{k+2p-1}}{(k+p)!p!} \\
&= \sum_{p=0}^{\infty} \frac{(k+p)t^{k+2p-1}}{(k+p)!p!} + \sum_{p=0}^{\infty} \frac{p\,t^{k+2p-1}}{(k+p)!p!} \\
&= \sum_{p=0}^{\infty} \frac{t^{k+2p-1}}{(k+p-1)!p!} + \sum_{p=1}^{\infty} \frac{t^{k+2p-1}}{(k+p)!(p-1)!} \\
&= \sum_{p=0}^{\infty} \frac{t^{(k-1)+2p}}{((k-1)+p)!p!} + \sum_{p=1}^{\infty} \frac{t^{(k+1)+2(p-1)}}{((k+1)+(p-1))!(p-1)!} \\
&= f_{k-1}(t) + f_{k+1}(t)
\end{aligned}$$

This computation works in fact for any k, so we get:

$$\begin{aligned}
Fb_t(y)' &= -Fb_t(y) + \frac{e^{-t}}{2} \sum_{k=-\infty}^{\infty} e^{ky}(f_{k-1}(t/2) + f_{k+1}(t/2)) \\
&= -Fb_t(y) + \frac{e^{-t}}{2} \sum_{k=-\infty}^{\infty} e^{(k+1)y} f_k(t/2) + e^{(k-1)y} f_k(t/2) \\
&= -Fb_t(y) + \frac{e^y + e^{-y}}{2}\, Fb_t(y) \\
&= \left(\frac{e^y + e^{-y}}{2} - 1\right) Fb_t(y)
\end{aligned}$$

Thus the log of the Fourier transform is linear in t, and we get the assertion. □

13d. Complex reflections

In order to further discuss all this, we will need a number of probabilistic preliminaries. We recall that, conceptually speaking, the Poisson laws are the laws appearing via the Poisson Limit Theorem (PLT). In order to generalize this construction, as to cover for Bessel laws, we have the following notion, extending the Poisson limit theory:

DEFINITION 13.28. *Associated to any compactly supported positive measure ν on $\mathbb{C}$ is the probability measure*

$$p_\nu = \lim_{n\to\infty} \left(\left(1 - \frac{c}{n}\right)\delta_0 + \frac{1}{n}\nu\right)^{*n}$$

where $c = mass(\nu)$, called compound Poisson law.

In what follows we will be interested in the case where ν is discrete, as is for instance the case for the measure $\nu = t\delta_1$ with $t > 0$, which produces via the above procedure the Poisson laws. To be more precise, we will be mainly interested in the case where ν is a multiple of the uniform measure on the s-th roots of unity. More on this later.

The following result allows us to detect the compound Poisson laws:

PROPOSITION 13.29. *For $\nu = \sum_{i=1}^s c_i\delta_{z_i}$ with $c_i > 0$ and $z_i \in \mathbb{C}$ we have*

$$F_{p_\nu}(y) = \exp\left(\sum_{i=1}^s c_i(e^{iyz_i} - 1)\right)$$

where F denotes as usual the Fourier transform.

PROOF. Let μ_n be the measure appearing in Definition 13.28, namely:

$$\mu_n = \left(1 - \frac{c}{n}\right)\delta_0 + \frac{1}{n}\nu$$

We have the following computation, in the context of Definition 13.28:

$$F_{\mu_n}(y) = \left(1 - \frac{c}{n}\right) + \frac{1}{n}\sum_{i=1}^s c_i e^{iyz_i} \quad \implies \quad F_{\mu_n^{*n}}(y) = \left(\left(1 - \frac{c}{n}\right) + \frac{1}{n}\sum_{i=1}^s c_i e^{iyz_i}\right)^n$$

$$\implies \quad F_{p_\nu}(y) = \exp\left(\sum_{i=1}^s c_i(e^{iyz_i} - 1)\right)$$

Thus, we have obtained the formula in the statement. □

We have as well the following result, providing an alternative to Definition 13.28, and which will be our formulation of the Compound Poisson Limit Theorem (CPLT):

THEOREM 13.30. *For $\nu = \sum_{i=1}^s c_i\delta_{z_i}$ with $c_i > 0$ and $z_i \in \mathbb{C}$ we have*

$$p_\nu = \text{law}\left(\sum_{i=1}^s z_i\alpha_i\right)$$

where the variables α_i are Poisson (c_i), independent.

PROOF. Let α be the sum of Poisson variables in the statement. We have:

$$F_{\alpha_i}(y) = \exp(c_i(e^{iy} - 1) \quad \implies \quad F_{z_i\alpha_i}(y) = \exp(c_i(e^{iyz_i} - 1))$$

$$\implies \quad F_\alpha(y) = \exp\left(\sum_{i=1}^s c_i(e^{iyz_i} - 1)\right)$$

Thus we have the same formula as in Proposition 13.29, as desired. □

Getting back now to the Bessel laws, we have the following result:

THEOREM 13.31. *The Bessel laws b_t are compound Poisson laws, given by*

$$b_t = p_{t\varepsilon}$$

where $\varepsilon = \frac{1}{2}(\delta_{-1} + \delta_1)$ is the uniform measure on $\mathbb{Z}_2$.

PROOF. This follows indeed by comparing the formula of the Fourier transform of b_t, from the proof of Theorem 13.27, with the formula in Proposition 13.29. □

Our next task will be that of unifying and generalizing the results that we have for S_N, H_N. For this purpose, consider the following family of groups:

DEFINITION 13.32. *The complex reflection group $H_N^s = \mathbb{Z}_s \wr S_N$ is given by*

$$H_N^s = M_N(\mathbb{Z}_s \cup \{0\}) \cap U_N$$

with the convention $\mathbb{Z}_\infty = \mathbb{T}$, at $s = \infty$.

Here the fact that we have indeed $H_N^s = \mathbb{Z}_s \wr S_N$ follows as in Theorem 13.25. Observe that at $s = 1, 2$ we obtain the symmetric and hyperoctahedral groups:

$$H_N^1 = S_N \quad , \quad H_N^2 = H_N$$

Another important particular case is $s = \infty$, where we obtain a compact group which is actually not finite, but is of key importance, that we will denote as follows:

$$H_N^\infty = K_N$$

In order to do now the character computations for H_N^s, in general, we need a number of further probabilistic preliminaries. Let us start with the following definition:

DEFINITION 13.33. *The Bessel law of level $s \in \mathbb{N} \cup \{\infty\}$ and parameter $t > 0$ is*

$$b_t^s = p_{t\varepsilon_s}$$

with ε_s being the uniform measure on the s-th roots of unity.

Observe that at $s = 1, 2$ we obtain the Poisson and real Bessel laws:

$$b_t^1 = p_t \quad , \quad b_t^2 = b_t$$

Another important particular case is $s = \infty$, where we obtain a measure which is actually not discrete, that we will denote as follows:

$$b_t^\infty = B_t$$

As a basic result on these laws, generalizing those before about p_t, b_t, we have:

THEOREM 13.34. *The generalized Bessel laws b_t^s have the property*

$$b_t^s * b_{t'}^s = b_{t+t'}^s$$

so they form a truncated one-parameter semigroup with respect to convolution.

PROOF. This follows indeed from the Fourier transform formula from Proposition 13.29, because for the Bessel laws, the log of this Fourier transform is linear in t. □

We can go back now to the reflection groups, and we have the following result:

THEOREM 13.35. *For the group $H_N^s = \mathbb{Z}_s \wr S_N$ we have, with $N \to \infty$,*

$$\chi_t \sim b_t^s$$

where $b_t^s = p_{t\varepsilon_s}$, with ε_s being the uniform measure on the s-th roots of unity.

PROOF. In the case $t = 1$, by arguing as before at $s = 2$, since the limit probability for a random permutation to have exactly k fixed points is $e^{-1}/k!$, we obtain:

$$\lim_{N\to\infty} law(\chi_1) = e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!}\, \varepsilon_s^{*k}$$

On the other hand, we get from the definition of the Bessel law b_1^s, as desired:

$$\begin{aligned} b_1^s &= \lim_{N\to\infty} \left(\left(1 - \frac{1}{N}\right) \delta_0 + \frac{1}{N}\, \varepsilon_s \right)^{*N} \\ &= \lim_{N\to\infty} \sum_{k=0}^{N} \binom{N}{k} \left(1 - \frac{1}{N}\right)^{N-k} \frac{1}{N^k}\, \varepsilon_s^{*k} \\ &= e^{-1} \sum_{k=0}^{\infty} \frac{1}{k!}\, \varepsilon_s^{*k} \end{aligned}$$

When $t > 0$ is arbitrary, we can use the same method, with some modifications where needed, again by arguing as before at $s = 2$, and we obtain the result. □

13e. Exercises

We had an interesting group theory chapter here, and as exercises, we have:

EXERCISE 13.36. *Learn about the decomposition of permutations, as products of cycles.*

EXERCISE 13.37. *Review if needed the basic theory of the Poisson laws.*

EXERCISE 13.38. *Fill in all the details for the computation of $law(\chi_t)$, over S_N.*

EXERCISE 13.39. *Find an alternative proof for the S_N result, by integrating over S_N.*

EXERCISE 13.40. *Learn about the Bessel functions of the first kind, appearing above.*

EXERCISE 13.41. *Learn more about the compound Poisson laws, and their properties.*

EXERCISE 13.42. *Learn about the complex reflection groups, and their classification.*

EXERCISE 13.43. *Learn about the finite abelian groups too, and their classification.*

As bonus exercise, and no surprise here, read some systematic finite group theory.

CHAPTER 14

Compact groups

14a. Lie groups

We have seen the basic theory of the finite subgroups $G \subset U_N$. In this chapter we go for the real thing, namely basic theory of the continuous subgroups $G \subset U_N$. Indeed, these groups are the most important ones, for both mathematics and physics, and historically they came first, from the work of Felix Klein, Sophus Lie and others.

However, getting into this subject, continuous subgroups $G \subset U_N$, is something quite unclear, right from the beginning, because we have the following dillema:

DILLEMA 14.1. *Shall our continuous subgroups $G \subset U_N$ be:*

(1) *Continuous, or smooth.*
(2) *Compact, or locally compact.*

Which does not look like something easy to solve, all possible combinations here seem to lead to substantial difficulties, and I am afraid that, following Klein, Lie and the others, who asked their respective cats about this, I will have to ask the cat too.

Unfortunately cat is gone, so in the lack of some good advice here, we will solve this problem Gordian knot style, by starting with the strongest possible axioms:

DEFINITION 14.2. *A compact Lie group is a compact group G which is at the same time a smooth manifold, with the group operations being smooth.*

Here we are making reference to the material from chapter 10, where we discussed what a smooth manifold is. Also, we have temporarily ditched the assumption $G \subset U_N$, for keeping things simple. We can always add this condition later, coming as an assumption, or why not as a theorem, if we are lucky, say as in the case of the finite groups.

Now inspired by the material from chapter 10, we are led to the following questions, which should normally provide the key to the study of the compact Lie groups:

QUESTIONS 14.3. *Given a compact Lie group G, as above:*

(1) *What can we say about the tangent space at the unit, $\mathfrak{g} = T_1G$?*
(2) *Can $\mathfrak{g}$ be axiomatized? Do we have a correspondence $\mathfrak{g} \leftrightarrow G$?*
(3) *What about $G \subset U_N$, how does this translate in terms of $\mathfrak{g}$?*

In answer now, inspired by what happens in the simplest cases, $G = O_N, U_N$, and by some differential geometry too, and skipping some details here, we have:

DEFINITION 14.4. *A Lie algebra is a vector space $\mathfrak{g}$ with an operation $(x, y) \to [x, y]$, called Lie bracket, subject to the following conditions:*

(1) $[x + y, z] = [x, z] + [y, z]$, $[x, y + z] = [x, y] + [x, z]$.
(2) $[\lambda x, y] = [x, \lambda y] = \lambda[x, y]$.
(3) $[x, x] = 0$.
(4) $[[x, y], z] + [[y, z], x] + [[z, x], y] = 0$.

As a basic example here, consider a usual, associative algebra A. We can define then the Lie bracket on it as being the usual commutator, namely:

$$[x, y] = xy - yx$$

The above axioms (1,2,3) are then clearly satisfied, and in what regards axiom (4), called Jacobi identity, this is satisfied too, the verification being as follows:

$$\begin{aligned} & [[x, y], z] + [[y, z], x] + [[z, x], y] \\ = \ & [xy - yx, z] + [yz - zy, x] + [zx - xz, y] \\ = \ & xyz - yxz - zxy + zyx + yzx - zyx - xyz + xzy + zxy - xzy - yzx + yxz \\ = \ & 0 \end{aligned}$$

We will see in a moment that up to a certain abstract operation $\mathfrak{g} \to U\mathfrak{g}$, which is something very straightforward, called enveloping Lie algebra construction, any Lie algebra appears in this way, with its Lie bracket being formally a commutator:

$$[x, y] = xy - yx$$

In relation now with groups, we have the following fundamental result, making the connection with the theory of Lie groups, denoted as usual by G:

THEOREM 14.5. *Given a Lie group G, that is, a group which is a smooth manifold, with the group operations being smooth, the tangent space at the identity*

$$\mathfrak{g} = T_1(G)$$

is a Lie algebra, with its Lie bracket being basically a usual commutator.

PROOF. This is something non-trivial, the idea being as follows:

(1) Let us first have a look at the orthogonal and unitary groups O_N, N_N. These are both Lie groups, and the corresponding Lie algebras $\mathfrak{o}_N, \mathfrak{u}_N$ can be computed by differentiating the equations defining O_N, U_N, with the conclusion being as follows:

$$\mathfrak{o}_N = \left\{ A \in M_N(\mathbb{R}) \middle| A^t = -A \right\}$$

$$\mathfrak{u}_N = \left\{ B \in M_N(\mathbb{C}) \middle| B^* = -B \right\}$$

This was for the correspondences $O_N \to \mathfrak{o}_N$ and $U_N \to \mathfrak{u}_N$. In the other sense, the correspondences $\mathfrak{o}_N \to O_N$ and $\mathfrak{u}_N \to U_N$ appear by exponentiation, the result here stating that, around 1, the orthogonal matrices can be written as $U = e^A$, with $A \in \mathfrak{o}_N$, and the unitary matrices can be written as $U = e^B$, with $B \in \mathfrak{u}_N$.

(2) Getting now to the Lie bracket, the first observation is that both $\mathfrak{o}_N, \mathfrak{u}_N$ are stable under the usual commutator of the $N \times N$ matrices. Indeed, assuming that $A, B \in M_N(\mathbb{R})$ satisfy $A^t = -A$, $B^t = -B$, their commutator satisfies $[A, B] \in M_N(\mathbb{R})$, and:

$$\begin{aligned}
[A, B]^t &= (AB - BA)^t \\
&= B^t A^t - A^t B^t \\
&= BA - AB \\
&= -[A, B]
\end{aligned}$$

Similarly, assuming that $A, B \in M_N(\mathbb{C})$ satisfy $A^* = -A$, $B^* = -B$, their commutator $[A, B] \in M_N(\mathbb{C})$ satisfies the condition $[A, B]^* = -[A, B]$.

(3) Thus, both tangent spaces $\mathfrak{o}_N, \mathfrak{u}_N$ are Lie algebras, with the Lie bracket being the usual commutator of the $N \times N$ matrices. It remains now to see what happens to the Lie bracket when exponentiating, and the formula here is as follows:

$$e^{[A,B]} = e^{AB-BA}$$

But the term on the right can be understood in terms of the differential geometry of O_N, U_N, and the situation is similar when dealing with an arbitrary Lie group G. □

With this understood, let us go back to the arbitrary Lie algebras, as axiomatized in Definition 14.4. We have the following key result, announced after Definition 14.4:

THEOREM 14.6. *Given a Lie algebra $\mathfrak{g}$, define its enveloping Lie algebra $U\mathfrak{g}$ as being the quotient of the tensor algebra of $\mathfrak{g}$, namely*

$$T(\mathfrak{g}) = \bigoplus_{k=0}^{\infty} \mathfrak{g}^{\otimes k}$$

by the following associative algebra ideal, with x, y ranging over the elements of $\mathfrak{g}$:

$$I = < x \otimes y - y \otimes x - [x, y] >$$

Then $U\mathfrak{g}$ is an associative algebra, so it is a Lie algebra too, with bracket

$$[x, y] = xy - yx$$

and the standard embedding $\mathfrak{g} \subset U\mathfrak{g}$ is a Lie algebra embedding.

PROOF. This is something which is quite self-explanatory, and in what regards the examples, illustrations, and other things that can be said, for instance in relation with the Lie groups, we will leave some further reading here as an instructive exercise. □

Importantly, the above enveloping Lie algebra construction makes as well a link with Hopf algebra theory, and with quantum groups, via the following result:

THEOREM 14.7. *Given a Lie algebra $\mathfrak{g}$, its enveloping Lie algebra $U\mathfrak{g}$ is a cocommutative Hopf algebra, with comultiplication, counit and antipode given by*

$$\Delta : U\mathfrak{g} \to U(\mathfrak{g} \oplus \mathfrak{g}) = U\mathfrak{g} \otimes U\mathfrak{g} \quad , \quad x \to x + x$$

$$\varepsilon : U\mathfrak{g} \to \mathbb{C} \quad , \quad x \to 1$$

$$S : U\mathfrak{g} \to U\mathfrak{g}^{opp} = (U\mathfrak{g})^{opp} \quad , \quad x \to -x$$

via various standard identifications, for the various associative algebras involved.

PROOF. Again, this is something self-explanatory, provided that you already know about Hopf algebras and quantum groups, and in what regards the examples, illustrations, and other things that can be said, we will leave some reading here as an exercise. □

So long for the Lie algebra basics, quicky explained in a few pages. The continuation of the story, bringing answers to Questions 14.3, is more complicated, as follows:

ANSWERS 14.8. *The following happen, under suitable assumptions:*

(1) *We have indeed a correspondence $\mathfrak{g} \leftrightarrow G$, appearing by exponentiation.*
(2) *The Lie algebras $\mathfrak{g}$ can be classified, and the compact Lie groups G too.*
(3) *In practice, we have regular cases ABCD, and exceptional cases EFG.*
(4) *The regular compact Lie groups, of type ABCD, are O_N, U_N, Sp_N.*

In relation with the last assertion, which is what matters the most, O_N, U_N are the usual orthogonal and unitary groups, and $Sp_N \subset U_N$ with $N \in 2\mathbb{N}$ is the symplectic group, appearing as a certain modification of the usual orthogonal group $O_N \subset U_N$.

Observe that, as a consequence, any compact Lie group G is a unitary group, $G \subset U_N$. Finally, many further things can be said in relation with Theorem 14.7, notably with a deformation procedure for O_N, U_N, Sp_N, into certain "quantum groups", invented by Drinfeld and Jimbo. For more on all this, we refer to any good Lie algebra book.

14b. Peter-Weyl

You might wonder at this point why not getting into details in relation with the above, which is certainly first-class mathematics, and that could easily cover the remainder of the present chapter, and perhaps even the remainder of the whole present book.

Well, the problem comes from the cat, who came back from his daily hunting adventures, and jumped back scared when seeing what I was typing. Here is what he says:

CAT 14.9. *All this is too complicated, Lie algebras are advanced science. For an introduction, and some nice applications, go with representations of compact groups.*

Okay cat, I was actually sort of expecting this, a word from you, and this since I fell into Dillema 14.1. So, forgetting now about Definition 14.2, which most likely would lead us into a Nobel Prize, or a Fields Medal, but are we here for such things, let us discuss instead the representations of compact groups. We will need the following notions:

DEFINITION 14.10. *A unitary representation of a compact group G is a continuous group morphism into a unitary group*

$$u : G \to U_N \quad , \quad g \to u_g$$

which can be faithful or not. The character of such a representation is the function

$$\chi : G \to \mathbb{C} \quad , \quad g \to Tr(u_g)$$

where Tr is the usual, unnormalized trace of the $N \times N$ matrices.

At the level of examples, most of the compact groups that we met so far, finite or continuous, naturally appear as closed subgroups $G \subset U_N$. In this case, the embedding $G \subset U_N$ is of course a representation, called fundamental representation.

In general now, let us first discuss the various operations on the representations. We have here the following elementary result, coming from definitions:

PROPOSITION 14.11. *The representations of a compact group G are subject to:*

(1) *Making sums. Given representations u, v, of dimensions N, M, their sum is the $N + M$-dimensional representation $u + v = diag(u, v)$.*
(2) *Making products. Given representations u, v, of dimensions N, M, their product is the NM-dimensional representation $(u \otimes v)_{ia,jb} = u_{ij}v_{ab}$.*
(3) *Taking conjugates. Given a N-dimensional representation u, its conjugate is the N-dimensional representation $(\bar{u})_{ij} = \bar{u}_{ij}$.*
(4) *Spinning by unitaries. Given a N-dimensional representation u, and a unitary $V \in U_N$, we can spin u by this unitary, $u \to VuV^*$.*

PROOF. The fact that the operations in the statement are indeed well-defined, among morphisms from G to unitary groups, is indeed clear from definitions. □

In relation now with characters, we have the following result:

PROPOSITION 14.12. *We have the following formulae, regarding characters*

$$\chi_{u+v} = \chi_u + \chi_v \quad , \quad \chi_{u\otimes v} = \chi_u\chi_v \quad , \quad \chi_{\bar{u}} = \bar{\chi}_u \quad , \quad \chi_{VuV^*} = \chi_u$$

in relation with the basic operations for the representations.

PROOF. All these assertions are elementary, by using the following formulae:

$$Tr(diag(U, V)) = Tr(U) + Tr(V) \quad , \quad Tr(U \otimes V) = Tr(U)Tr(V)$$

$$Tr(\bar{U}) = \overline{Tr(U)} \quad , \quad Tr(VUV^*) = Tr(U)$$

Thus, we are led to the character formulae in the statement. □

Assume now that we are given a closed subgroup $G \subset U_N$. By using the above operations, we can construct a whole family of representations of G, as follows:

DEFINITION 14.13. *Given a closed subgroup $G \subset U_N$, its Peter-Weyl representations are the tensor products between the fundamental representation and its conjugate:*

$$u : G \subset U_N \quad , \quad \bar{u} : G \subset U_N$$

We denote these tensor products $u^{\otimes k}$, with $k = \circ\bullet\bullet\circ\ldots$ being a colored integer, with the colored tensor powers being defined according to the rules

$$u^{\otimes\circ} = u \quad , \quad u^{\otimes\bullet} = \bar{u} \quad , \quad u^{\otimes kl} = u^{\otimes k} \otimes u^{\otimes l}$$

and with the convention that $u^{\otimes\emptyset}$ is the trivial representation $1 : G \to U_1$.

Here are a few examples of such representations, namely those coming from the colored integers of length 2, which will often appear in what follows:

$$u^{\otimes\circ\circ} = u \otimes u \quad , \quad u^{\otimes\circ\bullet} = u \otimes \bar{u}$$

$$u^{\otimes\bullet\circ} = \bar{u} \otimes u \quad , \quad u^{\otimes\bullet\bullet} = \bar{u} \otimes \bar{u}$$

In order to advance, we must develop some general theory. Let us start with:

DEFINITION 14.14. *Given a compact group G, and two of its representations,*

$$u : G \to U_N \quad , \quad v : G \to U_M$$

we define the space of intertwiners between these representations as being

$$Hom(u, v) = \left\{ T \in M_{M\times N}(\mathbb{C}) \middle| Tu_g = v_g T, \forall g \in G \right\}$$

and we use the following conventions:

(1) *We use the notations $Fix(u) = Hom(1, u)$, and $End(u) = Hom(u, u)$.*
(2) *We write $u \sim v$ when $Hom(u, v)$ contains an invertible element.*
(3) *We say that u is irreducible, and write $u \in Irr(G)$, when $End(u) = \mathbb{C}1$.*

The terminology here is standard, with Fix, Hom, End standing for fixed points, homomorphisms and endomorphisms. We will see later that irreducible means indecomposable, in a suitable sense. Here are now a few basic results, regarding these spaces:

THEOREM 14.15. *The spaces of intertwiners have the following properties:*

(1) $T \in Hom(u, v), S \in Hom(v, w) \implies ST \in Hom(u, w)$.
(2) $S \in Hom(u, v), T \in Hom(w, z) \implies S \otimes T \in Hom(u \otimes w, v \otimes z)$.
(3) $T \in Hom(u, v) \implies T^* \in Hom(v, u)$.

In abstract terms, we say that the Hom spaces form a tensor $*$-category.

PROOF. All the formulae in the statement are clear from definitions, via elementary computations. As for the last assertion, this is something coming from (1,2,3). We will be back to tensor categories later on, with more details on this latter fact. □

In order to advance, we will need the following standard linear algebra fact:

PROPOSITION 14.16. *Let $A \subset M_N(\mathbb{C})$ be a $*$-algebra.*

(1) *We have $1 = p_1 + \ldots + p_k$, with $p_i \in A$ being central minimal projections.*
(2) *Each of the spaces $A_i = p_i A p_i$ is a non-unital $*$-subalgebra of A.*
(3) *We have a non-unital $*$-algebra sum decomposition $A = A_1 \oplus \ldots \oplus A_k$.*
(4) *We have unital $*$-algebra isomorphisms $A_i \simeq M_{n_i}(\mathbb{C})$, with $n_i = rank(p_i)$.*
(5) *Thus, we have a $*$-algebra isomorphism $A \simeq M_{n_1}(\mathbb{C}) \oplus \ldots \oplus M_{n_k}(\mathbb{C})$.*

PROOF. Consider indeed an arbitrary $*$-algebra of the $N \times N$ matrices, $A \subset M_N(\mathbb{C})$. Let us first look at the center of this algebra, $Z(A) = A \cap A'$. It is elementary to prove that this center, as an algebra, is of the following form:

$$Z(A) \simeq \mathbb{C}^k$$

Consider now the standard basis $e_1, \ldots, e_k \in \mathbb{C}^k$, and let $p_1, \ldots, p_k \in Z(A)$ be the images of these vectors via the above identification. In other words, these elements $p_1, \ldots, p_k \in A$ are central minimal projections, summing up to 1:

$$p_1 + \ldots + p_k = 1$$

The idea is then that this partition of the unity eventually leads to the block decomposition of A, as in the statement, and we leave the details here as an exercise. □

We can now formulate our first Peter-Weyl type theorem, as follows:

THEOREM 14.17 (PW1). *Let $u : G \to U_N$ be a representation, consider the algebra $A = End(u)$, and write its unit $1 = p_1 + \ldots + p_k$ as above. We have then*

$$u = v_1 + \ldots + v_k$$

with each v_i being an irreducible representation, obtained by restricting u to $Im(p_i)$.

PROOF. This follows indeed from Theorem 14.15 and Proposition 14.16:

(1) We first associate to our representation $u : G \to U_N$ the corresponding action map on $\mathbb{C}^N$. If a linear subspace $V \subset \mathbb{C}^N$ is invariant, the restriction of the action map to V is an action map too, which must come from a subrepresentation $v \subset u$.

(2) Consider now a projection $p \in End(u)$. From $pu = up$ we obtain that the linear space $V = Im(p)$ is invariant under u, and so this space must come from a subrepresentation $v \subset u$. It is routine to check that the operation $p \to v$ maps subprojections to subrepresentations, and minimal projections to irreducible representations.

(3) With these preliminaries in hand, let us decompose the algebra $End(u)$ as above, by using the decomposition $1 = p_1 + \ldots + p_k$ into central minimal projections. If we denote by $v_i \subset u$ the subrepresentation coming from the vector space $V_i = Im(p_i)$, then we obtain in this way a decomposition $u = v_1 + \ldots + v_k$, as in the statement. □

Here is now our second Peter-Weyl theorem, complementing Theorem 14.17:

THEOREM 14.18 (PW2). *Given a closed subgroup $G \subset_u U_N$, any of its irreducible smooth representations*

$$v : G \to U_M$$

appears inside a tensor product of the fundamental representation u and its adjoint $\bar{u}$.

PROOF. This basically follows from Theorem 14.17, by reasoning as follows:

(1) Given $v : G \to U_M$, consider its space of coefficients $C_v \subset C(G)$. The operation $w \to C_w$ is then functorial, mapping subrepresentations into linear subspaces.

(2) A closed subgroup $G \subset_u U_N$ is a Lie group, and a representation $v : G \to U_M$ is smooth when we have an inclusion $C_v \subset < C_u >$. This is indeed well-known.

(3) By definition of the Peter-Weyl representations, as arbitrary tensor products between the fundamental representation u and its conjugate $\bar{u}$, we have:

$$< C_u >= \sum_k C_{u^{\otimes k}}$$

(4) Now by putting together the above observations (2,3) we conclude that we must have an inclusion as follows, for certain exponents $k_1, \ldots, k_p \in \mathbb{N}$:

$$C_v \subset C_{u^{\otimes k_1} \oplus \ldots \oplus u^{\otimes k_p}}$$

(5) By using now (1), we deduce that we have an inclusion $v \subset u^{\otimes k_1} \oplus \ldots \oplus u^{\otimes k_p}$, and by applying Theorem 14.17, this leads to the conclusion in the statement. □

In order to further advance, we will need the following standard fact:

THEOREM 14.19. *Any compact group G has a unique Haar integration, which can be constructed by starting with any faithful positive unital form $\varphi \in C(G)^*$, and setting:*

$$\int_G = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \varphi^{*k}$$

Moreover, for any representation v we have the formula

$$\left(id \otimes \int_G \right) v = P$$

where P is the orthogonal projection onto $Fix(v) = \left\{ \xi \in \mathbb{C}^n \middle| v\xi = \xi \right\}$.

PROOF. This is something very standard, and we will leave the proof here, first in the case where G is finite, which is easier, and then in general, and an instructive analysis exercise. Of course, in case you are stuck at some point, do not hesitate to look it up. □

We will need as well an algebraic ingredient for our study, as follows:

PROPOSITION 14.20. *We have a Frobenius type isomorphism*

$$Hom(v,w) \simeq Fix(v \otimes \bar{w})$$

valid for any two representations v, w.

PROOF. According to the definitions, we have the following equivalences:

$$\begin{aligned} T \in Hom(v,w) &\iff Tv = wT \\ &\iff \sum_j T_{aj} v_{ji} = \sum_b w_{ab} T_{bi}, \forall a, i \end{aligned}$$

On the other hand, we have as well the following equivalences:

$$\begin{aligned} T \in Fix(v \otimes \bar{w}) &\iff (v \otimes \bar{w})T = \xi \\ &\iff \sum_{jb} v_{ij} w_{ab}^* T_{bj} = T_{ai} \forall a, i \end{aligned}$$

But with this in hand, both inclusions follow from the unitarity of v, w. □

Good news, we can now formulate our third Peter-Weyl theorem, as follows:

THEOREM 14.21 (PW3). *The dense subalgebra $\mathcal{C}(G) \subset C(G)$ generated by the coefficients of the fundamental representation decomposes as a direct sum*

$$\mathcal{C}(G) = \bigoplus_{v \in Irr(G)} M_{\dim(v)}(\mathbb{C})$$

with the summands being pairwise orthogonal with respect to $< f, g >= \int_G f\bar{g}$.

PROOF. By combining the previous two Peter-Weyl results, we deduce that we have a linear space decomposition as follows:

$$\mathcal{C}(G) = \sum_{v \in Irr(G)} C_v = \sum_{v \in Irr(G)} M_{\dim(v)}(\mathbb{C})$$

Thus, in order to conclude, it is enough to prove that for any two irreducible representations $v, w \in Irr(G)$, the corresponding spaces of coefficients are orthogonal:

$$v \not\sim w \implies C_v \perp C_w$$

But this follows by Frobenius duality, by integrating. Let us set indeed:

$$P_{ia,jb} = \int_G v_{ij} \bar{w}_{ab}$$

Then P is the orthogonal projection onto the following vector space:

$$Fix(v \otimes \bar{w}) \simeq Hom(v,w) = \{0\}$$

Thus we have $P = 0$, and this gives the result. □

Finally, we have the following result, completing the Peter-Weyl theory:

THEOREM 14.22 (PW4). *The characters of irreducible representations belong to*

$$\mathcal{C}(G)_{central} = \left\{ f \in \mathcal{C}(G) \middle| f(gh) = f(hg), \forall g, h \in G \right\}$$

called algebra of central functions on G, and form an orthonormal basis of it.

PROOF. Observe first that $\mathcal{C}(G)_{central}$ is indeed an algebra, which contains all the characters. Conversely, consider a function $f \in \mathcal{C}(G)$, written as follows:

$$f = \sum_{v \in Irr(G)} f_v$$

The condition $f \in \mathcal{C}(G)_{central}$ states then that for any $v \in Irr(G)$, we must have:

$$f_v \in \mathcal{C}(G)_{central}$$

But this means that f_v must be a scalar multiple of χ_v, so the characters form a basis of $\mathcal{C}(G)_{central}$, as stated. Also, the fact that we have an orthogonal basis follows from Theorem 14.21. As for the fact that the characters have norm 1, this follows from:

$$\int_G \chi_v \bar{\chi}_v = \sum_{ij} \int_G v_{ii} \bar{v}_{jj} = \sum_i \frac{1}{M} = 1$$

Here we have used the fact, coming from Frobenius duality, that the various integrals $\int_G v_{ij} \bar{v}_{kl}$ form altogether the orthogonal projection onto the following vector space:

$$Fix(v \otimes \bar{v}) \simeq End(v) = \mathbb{C}1$$

Thus, the proof of our theorem is now complete. □

14c. Brauer algebras

Getting now to more juicy material, the Lie algebra theory as developed before is not the only way of "linearizing" the Lie groups. As a rival principle, we have:

PRINCIPLE 14.23. *Any finite, or even general compact group G appears as the symmetry group of its corresponding Tannakian category C_G,*

$$G = G(C_G)$$

and by suitably delinearizing C_G, say via a Brauer theorem of type $C_G = span(D_G)$, we can view G as symmetry group of a certain combinatorial object D_G.

Excited about this? Does not look easy, all this material, with both Tannaka and Brauer being quite scary names, in the context of algebra. But, believe me, all this is worth learning, and it is good to have in your bag some cutting-edge technology regarding the groups, such as the results of Tannaka and Brauer. So, we will go for this.

Getting started now, we first have a categorical definition, as follows:

DEFINITION 14.24. *A tensor category over $H = \mathbb{C}^N$ is a collection $C = (C_{kl})$ of linear spaces $C_{kl} \subset \mathcal{L}(H^{\otimes k}, H^{\otimes l})$ satisfying the following conditions:*

(1) *$S, T \in C$ implies $S \otimes T \in C$.*
(2) *If $S, T \in C$ are composable, then $ST \in C$.*
(3) *$T \in C$ implies $T^* \in C$.*
(4) *Each C_{kk} contains the identity operator.*
(5) *$C_{\emptyset k}$ with $k = \circ\bullet, \bullet\circ$ contain the operator $R : 1 \to \sum_i e_i \otimes e_i$.*
(6) *$C_{kl,lk}$ with $k, l = \circ, \bullet$ contain the flip operator $\Sigma : a \otimes b \to b \otimes a$.*

Here, as usual, the tensor powers $H^{\otimes k}$, which are Hilbert spaces depending on a colored integer $k = \circ \bullet \bullet \circ \ldots$, are defined by the following formulae, and multiplicativity:

$$H^{\otimes\emptyset} = \mathbb{C} \quad , \quad H^{\otimes\circ} = H \quad , \quad H^{\otimes\bullet} = \bar{H} \simeq H$$

We have already met such categories, when dealing with the Tannakian categories of the closed subgroups $G \subset U_N$, and our knowledge can be summarized as follows:

PROPOSITION 14.25. *Given a closed subgroup $G \subset U_N$, its Tannakian category*

$$C_{kl} = \left\{ T \in \mathcal{L}(H^{\otimes k}, H^{\otimes l}) \middle| Tg^{\otimes k} = g^{\otimes l} T, \forall g \in G \right\}$$

is a tensor category over $H = \mathbb{C}^N$. Conversely, given a tensor category C over $\mathbb{C}^N$,

$$G = \left\{ g \in U_N \middle| Tg^{\otimes k} = g^{\otimes l} T, \forall k, l, \forall T \in C_{kl} \right\}$$

is a closed subgroup of U_N.

PROOF. This is something that we basically know, the idea being as follows:

(1) Regarding the first assertion, we have to check here the axioms (1-6) in Definition 14.24. The axioms (1-4) being all clear from definitions, let us establish (5). But this follows from the fact that each element $g \in G$ is a unitary, which can be reformulated as follows, with $R : 1 \to \sum_i e_i \otimes e_i$ being the map in Definition 14.24:

$$R \in Hom(1, g \otimes \bar{g}) \quad , \quad R \in Hom(1, \bar{g} \otimes g)$$

Regarding now the condition in Definition 14.24 (6), this comes from the fact that the matrix coefficients $g \to g_{ij}$ and their conjugates $g \to \bar{g}_{ij}$ commute with each other.

(2) Regarding the second assertion, we have to check that the subset $G \subset U_N$ constructed in the statement is a closed subgroup. But this is clear from definitions. □

Summarizing, we have so far precise axioms for the tensor categories $C = (C_{kl})$, given in Definition 14.24, as well as correspondences as follows:

$$G \to C_G \quad , \quad C \to G_C$$

We will prove in what follows that these correspondences are inverse to each other. In order to get started, we first have the following technical result:

PROPOSITION 14.26. *Consider the following conditions:*

(1) $C = C_{G_C}$, *for any tensor category* C.

(2) $G = G_{C_G}$, *for any closed subgroup* $G \subset U_N$.

We have then (1) $\implies$ (2). *Also,* $C \subset C_{G_C}$ *is automatic.*

PROOF. Given $G \subset U_N$, we have $G \subset G_{C_G}$. On the other hand, by using (1) we have $C_G = C_{G_{C_G}}$. Thus, we have an inclusion of closed subgroups of U_N, which becomes an isomorphism at the level of the associated Tannakian categories, so $G = G_{C_G}$. Finally, the fact that we have an inclusion $C \subset C_{G_C}$ is clear from definitions. □

The point now is that it is possible to prove that we have $C_{G_C} \subset C$, by doing some abstract algebra, and we are led in this way to the following conclusion:

THEOREM 14.27. *The Tannakian duality constructions*

$$C \to G_C \quad , \quad G \to C_G$$

are inverse to each other.

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

(1) According to Proposition 14.26, we must prove $C_{G_C} \subset C$. For this purpose, given a tensor category $C = (C_{kl})$ over a Hilbert space H, consider the following $*$-algebra:

$$E_C = \bigoplus_{k,l} C_{kl} \subset \bigoplus_{k,l} B(H^{\otimes k}, H^{\otimes l}) \subset B\left(\bigoplus_k H^{\otimes k}\right)$$

Consider also, inside this $*$-algebra, the following $*$-subalgebra:

$$E_C^{(s)} = \bigoplus_{|k|,|l| \leq s} C_{kl} \subset \bigoplus_{|k|,|l| \leq s} B(H^{\otimes k}, H^{\otimes l}) = B\left(\bigoplus_{|k| \leq s} H^{\otimes k}\right)$$

(2) It is then routine to check that we have equivalences as follows:

$$\begin{aligned} C_{G_C} \subset C &\iff E_{C_{G_C}} \subset E_C \\ &\iff E_{C_{G_C}}^{(s)} \subset E_C^{(s)}, \forall s \\ &\iff E_{C_{G_C}}^{(s)'} \supset E_C^{(s)'}, \forall s \end{aligned}$$

(3) Summarizing, we would like to prove that we have inclusions $E_C^{(s)'} \subset E_{C_{G_C}}^{(s)'}$. But this can be done by doing some abstract algebra, and we refer here to the standard literature on the subject. For more on all this, you have as well my book [7]. □

With this piece of general theory in hand, let us go back to Principle 14.23, and develop the second idea there, namely delinearization and Brauer theorems. We have:

DEFINITION 14.28. *A category of crossing partitions is a collection $D = \bigsqcup_{k,l} D(k,l)$ of subsets $D(k,l) \subset P(k,l)$, having the following properties:*

(1) *Stability under the horizontal concatenation, $(\pi, \sigma) \to [\pi\sigma]$.*
(2) *Stability under vertical concatenation $(\pi, \sigma) \to [\frac{\sigma}{\pi}]$, with matching middle symbols.*
(3) *Stability under the upside-down turning $*$, with switching of colors, $\circ \leftrightarrow \bullet$.*
(4) *Each set $P(k,k)$ contains the identity partition $||\ldots||$.*
(5) *The sets $P(\emptyset, \circ\bullet)$ and $P(\emptyset, \bullet\circ)$ both contain the semicircle $\cap$.*
(6) *The sets $P(k, \bar{k})$ with $|k| = 2$ contain the crossing partition $\diagup\!\!\diagdown$.*

Observe the similarity with Definition 14.24, and more on this in a moment. In order now to construct a Tannakian category out of such a category, we will need:

PROPOSITION 14.29. *Each partition $\pi \in P(k,l)$ produces a linear map*

$$T_\pi : (\mathbb{C}^N)^{\otimes k} \to (\mathbb{C}^N)^{\otimes l}$$

given by the following formula, with $e_1, \ldots, e_N$ being the standard basis of $\mathbb{C}^N$,

$$T_\pi(e_{i_1} \otimes \ldots \otimes e_{i_k}) = \sum_{j_1 \ldots j_l} \delta_\pi \begin{pmatrix} i_1 & \ldots & i_k \\ j_1 & \ldots & j_l \end{pmatrix} e_{j_1} \otimes \ldots \otimes e_{j_l}$$

and with the Kronecker type symbols $\delta_\pi \in \{0, 1\}$ depending on whether the indices fit or not. The assignement $\pi \to T_\pi$ is categorical, in the sense that we have

$$T_\pi \otimes T_\sigma = T_{[\pi\sigma]} \quad , \quad T_\pi T_\sigma = N^{c(\pi,\sigma)} T_{[\frac{\sigma}{\pi}]} \quad , \quad T_\pi^* = T_{\pi^*}$$

where $c(\pi, \sigma)$ are certain integers, coming from the erased components in the middle.

PROOF. This is something elementary, the computations being as follows:

(1) The concatenation axiom follows from the following computation:

$$\begin{aligned}
& (T_\pi \otimes T_\sigma)(e_{i_1} \otimes \ldots \otimes e_{i_p} \otimes e_{k_1} \otimes \ldots \otimes e_{k_r}) \\
= & \sum_{j_1 \ldots j_q} \sum_{l_1 \ldots l_s} \delta_\pi \begin{pmatrix} i_1 & \ldots & i_p \\ j_1 & \ldots & j_q \end{pmatrix} \delta_\sigma \begin{pmatrix} k_1 & \ldots & k_r \\ l_1 & \ldots & l_s \end{pmatrix} e_{j_1} \otimes \ldots \otimes e_{j_q} \otimes e_{l_1} \otimes \ldots \otimes e_{l_s} \\
= & \sum_{j_1 \ldots j_q} \sum_{l_1 \ldots l_s} \delta_{[\pi\sigma]} \begin{pmatrix} i_1 & \ldots & i_p & k_1 & \ldots & k_r \\ j_1 & \ldots & j_q & l_1 & \ldots & l_s \end{pmatrix} e_{j_1} \otimes \ldots \otimes e_{j_q} \otimes e_{l_1} \otimes \ldots \otimes e_{l_s} \\
= & T_{[\pi\sigma]}(e_{i_1} \otimes \ldots \otimes e_{i_p} \otimes e_{k_1} \otimes \ldots \otimes e_{k_r})
\end{aligned}$$

(2) The composition axiom follows from the following computation:

$$\begin{aligned}
& T_\pi T_\sigma(e_{i_1}\otimes\ldots\otimes e_{i_p})\\
= & \sum_{j_1\ldots j_q}\delta_\sigma\begin{pmatrix}i_1&\ldots&i_p\\ j_1&\ldots&j_q\end{pmatrix}\sum_{k_1\ldots k_r}\delta_\pi\begin{pmatrix}j_1&\ldots&j_q\\ k_1&\ldots&k_r\end{pmatrix}e_{k_1}\otimes\ldots\otimes e_{k_r}\\
= & \sum_{k_1\ldots k_r}N^{c(\pi,\sigma)}\delta_{[\sigma]}\begin{pmatrix}i_1&\ldots&i_p\\ k_1&\ldots&k_r\end{pmatrix}e_{k_1}\otimes\ldots\otimes e_{k_r}\\
= & N^{c(\pi,\sigma)}T_{[\sigma]}(e_{i_1}\otimes\ldots\otimes e_{i_p})
\end{aligned}$$

(3) Finally, the involution axiom follows from the following computation:

$$\begin{aligned}
& T_\pi^*(e_{j_1}\otimes\ldots\otimes e_{j_q})\\
= & \sum_{i_1\ldots i_p}<T_\pi^*(e_{j_1}\otimes\ldots\otimes e_{j_q}),e_{i_1}\otimes\ldots\otimes e_{i_p}>e_{i_1}\otimes\ldots\otimes e_{i_p}\\
= & \sum_{i_1\ldots i_p}\delta_\pi\begin{pmatrix}i_1&\ldots&i_p\\ j_1&\ldots&j_q\end{pmatrix}e_{i_1}\otimes\ldots\otimes e_{i_p}\\
= & T_{\pi^*}(e_{j_1}\otimes\ldots\otimes e_{j_q})
\end{aligned}$$

Summarizing, our correspondence is indeed categorical. □

We can now formulate a key theoretical result, as follows:

THEOREM 14.30. *Any category of crossing partitions $D\subset P$ produces a series of compact groups $G=(G_N)$, with $G_N\subset U_N$ for any $N\in\mathbb{N}$, via the formula*

$$C_{kl}=span\left(T_\pi\Big|\pi\in D(k,l)\right)$$

for any k,l, and Tannakian duality. We call such groups easy.

PROOF. Indeed, once we fix an integer $N\in\mathbb{N}$, the various axioms in Definition 14.28 show, via Proposition 14.29, that the following spaces form a Tannakian category:

$$span\left(T_\pi\Big|\pi\in D(k,l)\right)$$

Thus, Tannakian duality applies, and provides us with a closed subgroup $G_N\subset U_N$ such that the following equalities are satisfied, for any colored integers k,l:

$$C_{kl}=span\left(T_\pi\Big|\pi\in D(k,l)\right)$$

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

And with this, good news, done with the general theory. At the level of basic examples now, we have the following key theorem of Brauer, with the convention that a pairing is matching when it pairs $\circ-\circ$ or $\bullet-\bullet$ on the vertical, and $\circ-\bullet$ on the horizontal:

THEOREM 14.31. *We have the following results:*

(1) *U_N is easy, coming from the category of all matching pairings $\mathcal{P}_2$.*

(2) *O_N is easy too, coming from the category of all pairings P_2.*

PROOF. This can be deduced from Tannakian duality, the idea being as follows:

(1) The unitary group U_N being defined via the relations $u^* = u^{-1}$, $u^t = \bar{u}^{-1}$, the associated Tannakian category is $C = span(T_\pi | \pi \in D)$, with:

$$D =< \cap_{\circ\bullet}\ ,\ \cap_{\bullet\circ} >= \mathcal{P}_2$$

(2) The orthogonal group $O_N \subset U_N$ being defined by imposing the relations $u_{ij} = \bar{u}_{ij}$, the associated Tannakian category is $C = span(T_\pi | \pi \in D)$, with:

$$D =< \mathcal{P}_2, \updownarrow, \bullet\updownarrow\circ >= P_2$$

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

Moving now towards finite groups, we first have the following result:

THEOREM 14.32. *The symmetric group S_N, regarded as group of unitary matrices,*

$$S_N \subset O_N \subset U_N$$

via the permutation matrices, is easy, coming from the category of all partitions P.

PROOF. Consider indeed the group S_N, regarded as a group of unitary matrices, with each permutation $\sigma \in S_N$ corresponding to the associated permutation matrix:

$$\sigma(e_i) = e_{\sigma(i)}$$

Consider as well the easy group $G \subset O_N$ coming from the category of all partitions P. Since P is generated by the one-block "fork" partition $Y \in P(2,1)$, we have:

$$C(G) = C(O_N) \Big/ \Big\langle T_Y \in Hom(u^{\otimes 2}, u) \Big\rangle$$

Now observe that we have the following formula:

$$(T_Y u^{\otimes 2})_{i,jk} = \sum_{lm} (T_Y)_{i,lm} (u^{\otimes 2})_{lm,jk} = u_{ij} u_{ik}$$

On the other hand, we have as well the following formula:

$$(uT_Y)_{i,jk} = \sum_l u_{il} (T_Y)_{l,jk} = \delta_{jk} u_{ij}$$

Thus, the relation defining $G \subset O_N$ reformulates as follows:

$$T_Y \in Hom(u^{\otimes 2}, u) \iff u_{ij} u_{ik} = \delta_{jk} u_{ij}, \forall i, j, k$$

In other words, the elements u_{ij} must be projections, which must be pairwise orthogonal on the rows of $u = (u_{ij})$. We conclude that $G \subset O_N$ is the subgroup of matrices $g \in O_N$ having the property $g_{ij} \in \{0, 1\}$. Thus we have $G = S_N$, as desired. □

The hyperoctahedral group H_N is easy as well, the result here being as follows:

THEOREM 14.33. *The hyperoctahedral group H_N, regarded as group of matrices,*

$$S_N \subset H_N \subset O_N$$

is easy, coming from the category of partitions with even blocks P_{even}.

PROOF. This follows as usual from Tannakian duality. To be more precise, consider the following one-block partition, which, as the name indicates, looks like a H letter:

$$H \in P(2,2)$$

The linear map associated to this partition is then given by:

$$T_H(e_i \otimes e_j) = \delta_{ij} e_i \otimes e_i$$

By using this formula, we have the following computation:

$$\begin{aligned}(T_H \otimes id)u^{\otimes 2}(e_a \otimes e_b) &= (T_H \otimes id)\left(\sum_{ijkl} e_{ij} \otimes e_{kl} \otimes u_{ij}u_{kl}\right)(e_a \otimes e_b)\\ &= (T_H \otimes id)\left(\sum_{ik} e_i \otimes e_k \otimes u_{ia}u_{kb}\right)\\ &= \sum_i e_i \otimes e_i \otimes u_{ia}u_{ib}\end{aligned}$$

On the other hand, we have as well the following computation:

$$\begin{aligned}u^{\otimes 2}(T_H \otimes id)(e_a \otimes e_b) &= \delta_{ab}\left(\sum_{ijkl} e_{ij} \otimes e_{kl} \otimes u_{ij}u_{kl}\right)(e_a \otimes e_a)\\ &= \delta_{ab}\sum_{ij} e_i \otimes e_k \otimes u_{ia}u_{ka}\end{aligned}$$

We conclude from this that we have the following equivalence:

$$T_H \in End(u^{\otimes 2}) \iff \delta_{ik}u_{ia}u_{ib} = \delta_{ab}u_{ia}u_{ka}, \forall i,k,a,b$$

We deduce from this that the corresponding closed subgroup $G \subset O_N$ consists of the matrices $g \in O_N$ which are permutation-like, with ± 1 nonzero entries. Thus, the corresponding group is $G = H_N$, and as a conclusion to all this, we have:

$$C(H_N) = C(O_N)\Big/\Big\langle T_H \in End(u^{\otimes 2})\Big\rangle$$

But this means that the hyperoctahedral group H_N is easy, coming from the category of partitions $D =< H >= P_{even}$. Thus, we are led to the conclusion in the statement. □

More generally now, we have in fact the following result, regarding the series of complex reflection groups H_N^s, which covers both the groups S_N, H_N:

THEOREM 14.34. *The complex reflection group $H_N^s = \mathbb{Z}_s \wr S_N$ is easy, the corresponding category P^s consisting of the partitions satisfying the condition*

$$\#\circ = \#\bullet\,(s)$$

as a weighted sum, in each block. In particular, we have the following results:

(1) *S_N is easy, coming from the category P.*
(2) *$H_N = \mathbb{Z}_2 \wr S_N$ is easy, coming from the category P_{even}.*
(3) *$K_N = \mathbb{T} \wr S_N$ is easy, coming from the category $\mathcal{P}_{even}$.*

PROOF. This is something that we already know at $s = 1, 2$, from Theorems 14.32 and 14.33. In general, the proof is similar, based on Tannakian duality. To be more precise, in what regards the main assertion, the idea here is that the one-block partition $\pi \in P(s)$, which generates the category of partitions P^s in the statement, implements the relations producing the subgroup $H_N^s \subset S_N$. As for the last assertions, these are all elementary:

(1) At $s = 1$ we know that we have $H_N^1 = S_N$. Regarding now the corresponding category, here the condition $\#\circ = \#\bullet\,(1)$ is automatic, and so $P^1 = P$.

(2) At $s = 2$ we know that we have $H_N^2 = H_N$. Regarding now the corresponding category, here the condition $\#\circ = \#\bullet\,(2)$ reformulates as follows:

$$\#\circ + \#\bullet = 0(2)$$

Thus each block must have even size, and we obtain, as claimed, $P^2 = P_{even}$.

(3) At $s = \infty$ we know that we have $H_N^\infty = K_N$. Regarding now the corresponding category, here the condition $\#\circ = \#\bullet\,(\infty)$ reads:

$$\#\circ = \#\bullet$$

But this is the condition defining $\mathcal{P}_{even}$, and so $P^\infty = \mathcal{P}_{even}$, as claimed. □

Summarizing, we have many examples. In fact, our list of easy groups has currently become quite big, and here is a selection of the main results that we have so far:

THEOREM 14.35. *We have a diagram of compact groups as follows,*

$$\begin{array}{ccc} K_N & \longrightarrow & U_N \\ \uparrow & & \uparrow \\ H_N & \longrightarrow & O_N \end{array}$$

where $H_N = \mathbb{Z}_2 \wr S_N$ and $K_N = \mathbb{T} \wr S_N$, and all these groups are easy.

PROOF. This follows from the above results. To be more precise, we know that the above groups are all easy, the corresponding categories of partitions being as follows:

$$\begin{array}{ccc} \mathcal{P}_{even} & \longleftarrow & \mathcal{P}_2 \\ \downarrow & & \downarrow \\ P_{even} & \longleftarrow & P_2 \end{array}$$

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

14d. Haar integration

In order to investigate linear independence questions for the vectors $\xi_\pi = T_\pi$, we will use the Gram matrix of these vectors. Let us begin with some standard definitions:

DEFINITION 14.36. *Let $P(k)$ be the set of partitions of $\{1, \ldots, k\}$, and let $\pi, \nu \in P(k)$.*

(1) *We write $\pi \leq \nu$ if each block of π is contained in a block of ν.*
(2) *We let $\pi \vee \nu \in P(k)$ be the partition obtained by superposing π, ν.*

As an illustration here, at $k = 2$ we have $P(2) = \{||, \sqcap\}$, and the order is:

$$|| \leq \sqcap$$

At $k = 3$ we have $P(3) = \{|||, \sqcap|, |\sqcap, \sqcap\!\sqcap\}$, and the order relation is as follows:

$$||| \leq \sqcap|, |\sqcap, |\sqcap \leq \sqcap\!\sqcap$$

Observe also that we have $\pi, \nu \leq \pi \vee \nu$. In fact, $\pi \vee \nu$ is the smallest partition with this property, called supremum of π, ν. Now back to the easy groups, we have:

PROPOSITION 14.37. *The Gram matrix $G_{kN}(\pi, \nu) = < \xi_\pi, \xi_\nu >$ is given by*

$$G_{kN}(\pi, \nu) = N^{|\pi \vee \nu|}$$

where $|.|$ is the number of blocks.

PROOF. According to our formula of the vectors ξ_π, we have:

$$\begin{aligned} < \xi_\pi, \xi_\nu > &= \sum_{i_1 \ldots i_k} \delta_\pi(i_1, \ldots, i_k) \delta_\nu(i_1, \ldots, i_k) \\ &= \sum_{i_1 \ldots i_k} \delta_{\pi \vee \nu}(i_1, \ldots, i_k) \\ &= N^{|\pi \vee \nu|} \end{aligned}$$

Thus, we have obtained the formula in the statement. □

In order to study the Gram matrix, and more specifically to compute its determinant, we will need several standard facts about the partitions. We first have:

DEFINITION 14.38. *The Möbius function of any lattice, and so of P, is given by*

$$\mu(\pi,\nu)=\begin{cases}1 & \text{if } \pi=\nu\\ -\sum_{\pi\leq\tau<\nu}\mu(\pi,\tau) & \text{if } \pi<\nu\\ 0 & \text{if } \pi\not\leq\nu\end{cases}$$

with the construction being performed by recurrence.

As an illustration here, let us go back to the set of 2-point partitions, $P(2)=\{||,\sqcap\}$. Here we have by definition:

$$\mu(||,||)=\mu(\sqcap,\sqcap)=1$$

Also, we know that we have $||<\sqcap$, with no intermediate partition in between, and so the above recurrence procedure gives the following formular:

$$\mu(||,\sqcap)=-\mu(||,||)=-1$$

Finally, we have $\sqcap\not\leq||$, which gives the following formula:

$$\mu(\sqcap,||)=0$$

The interest in the Möbius function comes from the Möbius inversion formula:

$$f(\nu)=\sum_{\pi\leq\nu}g(\pi)\implies g(\nu)=\sum_{\pi\leq\nu}\mu(\pi,\nu)f(\pi)$$

In linear algebra terms, the statement and proof of this formula are as follows:

THEOREM 14.39. *The inverse of the adjacency matrix of P, given by*

$$A_{\pi\nu}=\begin{cases}1 & \text{if } \pi\leq\nu\\ 0 & \text{if } \pi\not\leq\nu\end{cases}$$

is the Möbius matrix of P, given by $M_{\pi\nu}=\mu(\pi,\nu)$.

PROOF. This is well-known, coming for instance from the fact that A is upper triangular. Thus, when inverting, we are led into the recurrence from Definition 14.38. □

As an illustration here, for $P(2)$ the formula $M=A^{-1}$ appears as follows:

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

Now back to our Gram matrix considerations, we have the following result:

PROPOSITION 14.40. *The Gram matrix is given by $G_{kN}=AL$, where*

$$L(\pi,\nu)=\begin{cases}N(N-1)\ldots(N-|\pi|+1) & \text{if } \nu\leq\pi\\ 0 & \text{otherwise}\end{cases}$$

and where $A=M^{-1}$ is the adjacency matrix of $P(k)$.

PROOF. We have the following computation:

$$\begin{aligned} N^{|\pi\vee\nu|} &= \#\left\{i_1,\ldots,i_k\in\{1,\ldots,N\}\Big|\ker i\geq\pi\vee\nu\right\}\\ &= \sum_{\tau\geq\pi\vee\nu}\#\left\{i_1,\ldots,i_k\in\{1,\ldots,N\}\Big|\ker i=\tau\right\}\\ &= \sum_{\tau\geq\pi\vee\nu}N(N-1)\ldots(N-|\tau|+1)\end{aligned}$$

According to Proposition 14.37 and to the definition of A, L, this formula reads:

$$(G_{kN})_{\pi\nu}=\sum_{\tau\geq\pi}L_{\tau\nu}=\sum_{\tau}A_{\pi\tau}L_{\tau\nu}=(AL)_{\pi\nu}$$

Thus, we obtain the formula in the statement. □

With the above result in hand, we can now investigate the linear independence properties of the vectors ξ_π. To be more precise, we have the following result:

THEOREM 14.41. *The determinant of the Gram matrix G_{kN} is given by*

$$\det(G_{kN})=\prod_{\pi\in P(k)}\frac{N!}{(N-|\pi|)!}$$

and in particular, for $N\geq k$, the vectors $\{\xi_\pi|\pi\in P(k)\}$ are linearly independent.

PROOF. According to the formula in Proposition 14.40, we have:

$$\det(G_{kN})=\det(A)\det(L)$$

Now if we order $P(k)$ as usual, with respect to the number of blocks, and then lexicographically, we see that A is upper triangular, and that L is lower triangular. But this shows that $\det(A)=1$, and in what concerns $\det(L)$, this can be computed as well by making the product on the diagonal, and we obtain the number in the statement. □

Now back to the laws of characters, we can formulate:

PROPOSITION 14.42. *For an easy group $G=(G_N)$, coming from a category of partitions $D=(D(k,l))$, the asymptotic moments of the main character are given by*

$$\lim_{N\to\infty}\int_{G_N}\chi^k=\#D(k)$$

where $D(k)=D(\emptyset,k)$, with the limiting sequence on the left consisting of certain integers, and being stationary at least starting from the k-th term.

PROOF. This follows indeed from the Peter-Weyl theory, by using the linear independence result for the vectors ξ_π coming from Theorem 14.41. □

With these preliminaries in hand, we can now state and prove:

THEOREM 14.43. *In the $N \to \infty$ limit, the laws of the main character for the main easy groups, real and complex, and discrete and continuous, are as follows,*

$$\begin{array}{ccc} K_N & \to & U_N \\ \uparrow & & \uparrow \\ H_N & \to & O_N \end{array} \quad : \quad \begin{array}{ccc} B_1 & \to & G_1 \\ \uparrow & & \uparrow \\ b_1 & \to & g_1 \end{array}$$

with these laws, namely the real and complex Gaussian and Bessel laws, being the main limiting laws in real and complex, and discrete and continuous probability.

PROOF. This follows from the above results. To be more precise, we know that the above groups are all easy, the corresponding categories of partitions being as follows:

$$\begin{array}{ccc} \mathcal{P}_{even} & \leftarrow & \mathcal{P}_2 \\ \downarrow & & \downarrow \\ P_{even} & \leftarrow & P_2 \end{array}$$

Thus, we can use Proposition 14.42, are we are led into counting partitions, and then recovering the measures via their moments, and this leads to the result. □

Our aim now is to go beyond what we have, with results regarding the truncated characters. Let us start with a general formula coming from Peter-Weyl, namely:

THEOREM 14.44. *The Haar integration over a closed subgroup $G \subset_u U_N$ is given on the dense subalgebra of smooth functions by the Weingarten type formula*

$$\int_G g_{i_1j_1}^{e_1} \ldots g_{i_kj_k}^{e_k}\, dg = \sum_{\pi,\nu \in D(k)} \delta_\pi(i)\delta_\sigma(j) W_k(\pi,\nu)$$

valid for any colored integer $k = e_1 \ldots e_k$ and any multi-indices i, j, where $D(k)$ is a linear basis of $Fix(u^{\otimes k})$, the associated generalized Kronecker symbols are given by

$$\delta_\pi(i) = < \pi, e_{i_1} \otimes \ldots \otimes e_{i_k} >$$

and $W_k = G_k^{-1}$ is the inverse of the Gram matrix, $G_k(\pi,\nu) = < \pi, \nu >$.

PROOF. This is something very standard, coming from the fact that the above integrals form altogether the orthogonal projection P^k onto the following space:

$$Fix(u^{\otimes k}) = span(D(k))$$

Consider now the following linear map, with $D(k) = \{\xi_k\}$ being as in the statement:

$$E(x) = \sum_{\pi \in D(k)} < x, \xi_\pi > \xi_\pi$$

By a standard linear algebra computation, it follows that we have $P = WE$, where W is the inverse of the restriction of E to the following space:

$$K = span\left(T_\pi \middle| \pi \in D(k)\right)$$

But this restriction is the linear map given by the matrix G_k, and so W is the linear map given by the inverse matrix $W_k = G_k^{-1}$, and this gives the result. □

In the easy case, we have the following more concrete result:

THEOREM 14.45. *For an easy group $G \subset U_N$, coming from a category of partitions $D = (D(k,l))$, we have the Weingarten formula*

$$\int_G g_{i_1 j_1}^{e_1} \ldots g_{i_k j_k}^{e_k} \, dg = \sum_{\pi,\nu \in D(k)} \delta_\pi(i) \delta_\nu(j) W_{kN}(\pi, \nu)$$

for any $k = e_1 \ldots e_k$ and any i, j, where $D(k) = D(\emptyset, k)$, δ are usual Kronecker type symbols, checking whether the indices match, and $W_{kN} = G_{kN}^{-1}$, with

$$G_{kN}(\pi, \nu) = N^{|\pi \vee \nu|}$$

where $|.|$ is the number of blocks.

PROOF. We use the abstract Weingarten formula, from Theorem 14.44. Indeed, the Kronecker type symbols there are then the usual ones, as shown by:

$$\begin{aligned} \delta_{\xi_\pi}(i) &= < \xi_\pi, e_{i_1} \otimes \ldots \otimes e_{i_k} > \\ &= \left\langle \sum_j \delta_\pi(j_1, \ldots, j_k) e_{j_1} \otimes \ldots \otimes e_{j_k}, e_{i_1} \otimes \ldots \otimes e_{i_k} \right\rangle \\ &= \delta_\pi(i_1, \ldots, i_k) \end{aligned}$$

The Gram matrix being as well the correct one, we obtain the result. □

As an application of this, let us discuss the computation of the laws of characters. First, we have the following formula, in the general easy group setting:

PROPOSITION 14.46. *The moments of truncated characters are given by*

$$\int_G (u_{11} + \ldots + u_{ss})^k = Tr(W_{kN} G_{ks})$$

where G_{kN} and $W_{kN} = G_{kN}^{-1}$ are the associated Gram and Weingarten matrices.

PROOF. We have indeed the following computation:

$$\begin{aligned}\int_G(u_{11}+\ldots+u_{ss})^k &= \sum_{i_1=1}^{s}\ldots\sum_{i_k=1}^{s}\int_G u_{i_1i_1}\ldots g_{i_ki_k}\\ &= \sum_{\pi,\sigma\in D(k)}W_{kN}(\pi,\sigma)\sum_{i_1=1}^{s}\ldots\sum_{i_k=1}^{s}\delta_\pi(i)\delta_\sigma(i)\\ &= \sum_{\pi,\sigma\in D(k)}W_{kN}(\pi,\sigma)G_{ks}(\sigma,\pi)\\ &= Tr(W_{kN}G_{ks})\end{aligned}$$

Thus, we have obtained the formula in the statement. □

The idea now is to impose a natural uniformity condition, as follows:

DEFINITION 14.47. *An easy group $G=(G_N)$, coming from a category of partitions $D\subset P$, is called uniform if it satisfies the following equivalent conditions:*

(1) *$G_{N-1}=G_N\cap U_{N-1}$, via the embedding $U_{N-1}\subset U_N$ given by $u\to diag(u,1)$.*
(2) *$G_{N-1}=G_N\cap U_{N-1}$, via the N possible diagonal embeddings $U_{N-1}\subset U_N$.*
(3) *D is stable under the operation which consists in removing blocks.*

Here the equivalence between the above three conditions is something standard, obtained by doing some combinatorics. Also, we have many examples of such groups, the idea here being that the most familiar easy groups $G=(G_N)$ that we know, as for instance the various real and complex rotation and reflection groups, are indeed uniform.

In what follows we will be mostly interested in the condition (3) above, which makes the link with our computations for truncated characters, and simplifies them. To be more precise, by imposing the uniformity condition we obtain the following result:

THEOREM 14.48. *For a uniform easy group $G=(G_N)$, we have the formula*

$$\lim_{N\to\infty}\int_{G_N}\chi_t^k=\sum_{\pi\in D(k)}t^{|\pi|}$$

with $D\subset P$ being the associated category of partitions.

PROOF. We use the general moment formula from Proposition 14.46, namely:

$$\int_G(u_{11}+\ldots+u_{ss})^k=Tr(W_{kN}G_{ks})$$

By setting $s=[tN]$, with $t>0$ being a given parameter, this formula becomes:

$$\int_{G_N}\chi_t^k=Tr(W_{kN}G_{k[tN]})$$

The point now is that in the uniform case the Gram and Weingarten matrices are asymptotically diagonal, and this leads to the formula in the statement. □

We can now formulate some general character results, as follows:

THEOREM 14.49. *With $N \to \infty$, the laws of truncated characters are as follows:*

(1) *For O_N we obtain the Gaussian law g_t.*
(2) *For U_N we obtain the complex Gaussian law G_t.*
(3) *For S_N we obtain the Poisson law p_t.*
(4) *For H_N we obtain the Bessel law b_t.*
(5) *For H_N^s we obtain the generalized Bessel law b_t^s.*
(6) *For K_N we obtain the complex Bessel law B_t.*

Also, for B_N, C_N and for Sp_N we obtain modified normal laws.

PROOF. We use the formula that we found in Theorem 14.48, namely:

$$\lim_{N\to\infty}\int_{G_N}\chi_t^k=\sum_{\pi\in D(k)}t^{|\pi|}$$

By doing now some combinatorics, for instance in relation with the cumulants, this gives the results. We refer here to [7] and various related papers. □

14e. Exercises

This was a tough, modern linear algebra chapter, and as exercises, we have:

EXERCISE 14.50. *Read more about the basic theory of Lie groups and algebras.*

EXERCISE 14.51. *Read as well about the classification of Lie groups and algebras.*

EXERCISE 14.52. *Work out all the details, for the $*$-algebras $A \subset M_N(\mathbb{C})$.*

EXERCISE 14.53. *Work out all the details for the existence of the Haar measure.*

EXERCISE 14.54. *Look up and learn the full proof of Tannakian duality.*

EXERCISE 14.55. *Find some other interesting examples of easy groups.*

EXERCISE 14.56. *Work out the asymptotic law results for characters.*

EXERCISE 14.57. *Work out the asymptotic law results for truncated characters.*

As bonus exercise, try unifying the theories of Lie algebras, and Brauer algebras.

CHAPTER 15

Spin matrices

15a. Quantum physics

Good news, done with mathematics, we have learned enough interesting things, and in the remainder of this book we will get into physics, and mathematical physics. Everything will remain of course linear algebra oriented, and we will be talking about things which are related to the most important matrix groups of them all, namely SU_2 and SO_3.

You have surely heard about the atomic theory, in its golden age, happening around 1900-1920. At that time, Bohr was able to put everything together, and formulate a nice conjecture regarding the functioning of hydrogen ${}_1\mathrm{H}$, and of heavier atoms as well. However, this turned difficult to prove, among others because the electrons are "slippery" particles, having no clear positions and speeds. Nevertheless, Heisenberg came with a clever way of solving this question, by encoding positions and speeds by certain linear operators, instead of numbers, and managed to prove the Bohr conjecture.

Before explaining what Heisenberg was saying, let us hear as well the point of view of Schrödinger, which came a few years later. His idea was to forget about exact things, and try to investigate the hydrogen atom statistically. Let us start with:

QUESTION 15.1. *In the context of the hydrogen atom, assuming that the proton is fixed, what is the probability density $\varphi_t(x)$ of the position of the electron e, at time t,*

$$P_t(e \in V) = \int_V \varphi_t(x)dx$$

as function of an intial probability density $\varphi_0(x)$? Moreover, can the corresponding equation be solved, and will this prove the Bohr claims for hydrogen, statistically?

In order to get familiar with this question, let us first look at examples coming from classical mechanics. In the context of a particle whose position at time t is given by $x_0 + \gamma(t)$, the evolution of the probability density will be given by:

$$\varphi_t(x) = \varphi_0(x) + \gamma(t)$$

However, such examples are somewhat trivial, of course not in relation with the computation of γ, usually a difficult question, but in relation with our questions, and do not apply to the electron. The point indeed is that, in what regards the electron, we have:

FACT 15.2. *In respect with various simple interference experiments:*

(1) *The electron is definitely not a particle in the usual sense.*
(2) *But in most situations it behaves exactly like a wave.*
(3) *But in other situations it behaves like a particle.*

Getting back now to the Schrödinger question, all this suggests to use, as for the waves, an amplitude function $\psi_t(x) \in \mathbb{C}$, related to the density $\varphi_t(x) > 0$ by the formula $\varphi_t(x) = |\psi_t(x)|^2$. Not that a big deal, you would say, because the two are related by simple formulae as follows, with $\theta_t(x)$ being an arbitrary phase function:

$$\varphi_t(x) = |\psi_t(x)|^2 \quad , \quad \psi_t(x) = e^{i\theta_t(x)}\sqrt{\varphi_t(x)}$$

However, such manipulations can be crucial, raising for instance the possibility that the amplitude function satisfies some simple equation, while the density itself, maybe not. And this is what happens indeed. Schrödinger was led in this way to:

CLAIM 15.3 (Schrödinger). *In the context of the hydrogen atom, the amplitude function of the electron $\psi = \psi_t(x)$ is subject to the Schrödinger equation*

$$ih\dot{\psi} = -\frac{h^2}{2m}\Delta\psi + V\psi$$

m being the mass, $h = h_0/2\pi$ the reduced Planck constant, and V the Coulomb potential of the proton. The same holds for movements of the electron under any potential V.

Observe the similarity with the wave equation $\ddot{\varphi} = v^2\Delta\varphi$, and with the heat equation $\dot{\varphi} = \alpha\Delta\varphi$ too. Many things can be said here. Following now Heisenberg and Schrödinger, and then especially Dirac, who did the axiomatization work, we have:

DEFINITION 15.4. *In quantum mechanics the states of the system are vectors of a Hilbert space H, and the observables of the system are linear operators*

$$T : H \to H$$

which can be densely defined, and are taken self-adjoint, $T = T^*$. The average value of such an observable T, evaluated on a state $\xi \in H$, is given by:

$$< T >=< T\xi, \xi >$$

In the context of the Schrödinger mechanics of the hydrogen atom, the Hilbert space is the space $H = L^2(\mathbb{R}^3)$ where the wave function ψ lives, and we have

$$< T >= \int_{\mathbb{R}^3} T(\psi) \cdot \bar{\psi}\, dx$$

which is called "sandwiching" formula, with the operators

$$x \quad , \quad -\frac{ih}{m}\nabla \quad , \quad -ih\nabla \quad , \quad -\frac{h^2\Delta}{2m} \quad , \quad -\frac{h^2\Delta}{2m} + V$$

representing the position, speed, momentum, kinetic energy, and total energy.

In other words, we are doing here two things. First, we are declaring by axiom that various "sandwiching" formulae found before by Heisenberg hold true. And second, we are raising the possibility for other quantum mechanical systems, more complicated, to be described as well by the mathematics of operators on a certain Hilbert space H.

With this discussed, let us develop now quantum mechanics, in a mathematical way, by studying the Schrödinger equation from Claim 15.3. We first have:

PROPOSITION 15.5. *We have the following formula,*

$$\dot{\varphi} = \frac{ih}{2m}\left(\Delta\psi\cdot\bar{\psi} - \Delta\bar{\psi}\cdot\psi\right)$$

for the time derivative of the probability density function $\varphi = |\psi|^2$.

PROOF. According to the Leibnitz product rule, we have the following formula:

$$\dot{\varphi} = \frac{d}{dt}|\psi|^2 = \frac{d}{dt}(\psi\bar{\psi}) = \dot{\psi}\bar{\psi} + \psi\dot{\bar{\psi}}$$

On the other hand, the Schrödinger equation and its conjugate read:

$$\dot{\psi} = \frac{ih}{2m}\left(\Delta\psi - \frac{2m}{h^2}V\psi\right) \quad , \quad \dot{\bar{\psi}} = -\frac{ih}{2m}\left(\Delta\bar{\psi} - \frac{2m}{h^2}V\bar{\psi}\right)$$

By plugging this data, we obtain the following formula:

$$\dot{\varphi} = \frac{ih}{2m}\left[\left(\Delta\psi - \frac{2m}{h^2}V\psi\right)\bar{\psi} - \left(\Delta\bar{\psi} - \frac{2m}{h^2}V\bar{\psi}\right)\psi\right]$$

But this gives, after simplifying, the formula in the statement. □

As an important application of Proposition 15.5, we have:

THEOREM 15.6. *The Schrödinger equation conserves probability amplitudes,*

$$\int_{\mathbb{R}^3}|\psi_0|^2 = 1 \implies \int_{\mathbb{R}^3}|\psi_t|^2 = 1$$

in agreement with the basic probabilistic requirement, $P = 1$ overall.

PROOF. According to the formula in Proposition 15.5, we have:

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^3}|\psi|^2\,dx &= \int_{\mathbb{R}^3}\frac{d}{dt}|\psi|^2\,dx\\ &= \int_{\mathbb{R}^3}\dot{\varphi}\,dx\\ &= \frac{ih}{2m}\int_{\mathbb{R}^3}\left(\Delta\psi\cdot\bar{\psi} - \Delta\bar{\psi}\cdot\psi\right)dx\end{aligned}$$

Now by remembering the definition of the Laplace operator, we have:

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^3}|\psi|^2\,dx &= \frac{ih}{2m}\int_{\mathbb{R}^3}\sum_i\left(\frac{d^2\psi}{dx_i^2}\cdot\bar{\psi}-\frac{d^2\bar{\psi}}{dx_i^2}\cdot\psi\right)dx\\ &= \frac{ih}{2m}\sum_i\int_{\mathbb{R}^3}\frac{d}{dx_i}\left(\frac{d\psi}{dx_i}\cdot\bar{\psi}-\frac{d\bar{\psi}}{dx_i}\cdot\psi\right)dx\\ &= \frac{ih}{2m}\sum_i\int_{\mathbb{R}^2}\left[\frac{d\psi}{dx}\cdot\bar{\psi}-\frac{d\bar{\psi}}{dx}\cdot\psi\right]_{-\infty}^{\infty}\frac{dx}{dx_i}\\ &= 0\end{aligned}$$

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

Moving now towards hydrogen, we have here the following result:

THEOREM 15.7. *In the case of time-independent potentials V, including the Coulomb potential of the proton, the separated solutions of the Schrödinger equation*

$$\psi_t(x)=w_t\phi(x)$$

are given by the following formulae, with E being a certain constant,

$$w=e^{-iEt/h}w_0\quad,\quad E\phi=-\frac{h^2}{2m}\Delta\phi+V\phi$$

with the equation for ϕ being called time-independent Schrödinger equation.

PROOF. By dividing by ψ, the Schrödinger equation becomes:

$$ih\cdot\frac{\dot{w}}{w}=-\frac{h^2}{2m}\cdot\frac{\Delta\phi}{\phi}+V$$

Now since the left-hand side depends only on time, and the right-hand side depends only on space, both quantities must equal a constant E, and this gives the result. □

Moving ahead with theory, we can further build on Theorem 15.7, as follows:

THEOREM 15.8. *In the case of time-independent potentials V, the Schrödinger equation and its time-independent version have the following properties:*

(1) *For solutions of type $\psi=w_t\phi(x)$, the density $\varphi=|\psi|$ is time-independent, and more generally, all quantities of type $<T>$ are time-independent.*
(2) *The time-independent Schrödinger equation can be written as $\widehat{H}\phi=E\phi$, with $H=T+V$ being the total energy, of Hamiltonian.*
(3) *For solutions of type $\psi=w_t\phi(x)$ we have $<H^k>=E^k$ for any k. In particular we have $<H>=E$, and the variance is $<H^2>-<H>^2=0$.*

PROOF. All the formulae are clear indeed from the fact that, when using the sandwiching formula for computing averages, the phases will cancel:

$$< T >= \int_{\mathbb{R}^3} \bar{\psi} \cdot T \cdot \psi \, dx = \int_{\mathbb{R}^3} \bar{\phi} \cdot T \cdot \phi \, dx$$

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

We have as well the following key result, mathematical this time:

THEOREM 15.9. *The solutions of the Schrödinger equation with time-independent potential V appear as linear combinations of separated solutions*

$$\psi = \sum_n c_n e^{-iE_n t/h} \phi_h$$

with the absolute values of the coefficients being given by

$$< H >= \sum_n |c_n|^2 E_n$$

$|c_n|$ being the probability for a measurement to return the energy value E_n.

PROOF. This is something standard, which follows from Fourier analysis, which allows the decomposition of ψ as in the statement, and that we will not really need, in what follows next. As before, for a physical discussion here, we refer to Griffiths [41]. □

In order to solve now the hydrogen atom, by using the Schrödinger equation, the idea will be that of reformulating this equation in spherical coordinates. We have:

THEOREM 15.10. *The time-independent Schrödinger equation in spherical coordinates separates, for solutions of type $\phi = \rho(r)\alpha(s,t)$, into two equations, as follows,*

$$\frac{d}{dr}\left(r^2 \cdot \frac{d\rho}{dr}\right) - \frac{2mr^2}{h^2}(V - E)\rho = K\rho$$

$$\sin s \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\alpha}{ds}\right) + \frac{d^2\alpha}{dt^2} = -K \sin^2 s \cdot \alpha$$

with K being a constant, called radial equation, and angular equation.

PROOF. We use the following well-known formula for the Laplace operator in spherical coordinates, whose proof can be found in any advanced calculus book:

$$\Delta = \frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d}{dr}\right) + \frac{1}{r^2 \sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d}{ds}\right) + \frac{1}{r^2 \sin^2 s} \cdot \frac{d^2}{dt^2}$$

By using this formula, the time-independent Schrödinger equation reformulates as:

$$(V - E)\phi = \frac{h^2}{2m}\left[\frac{1}{r^2} \cdot \frac{d}{dr}\left(r^2 \cdot \frac{d\phi}{dr}\right) + \frac{1}{r^2 \sin s} \cdot \frac{d}{ds}\left(\sin s \cdot \frac{d\phi}{ds}\right) + \frac{1}{r^2 \sin^2 s} \cdot \frac{d^2\phi}{dt^2}\right]$$

Let us look now for separable solutions for this latter equation, consisting of a radial part and an angular part, as in the statement, namely:

$$\phi(r,s,t)=\rho(r)\alpha(s,t)$$

By plugging this function into our equation, we obtain:

$$(V-E)\rho\alpha=\frac{h^2}{2m}\left[\frac{\alpha}{r^2}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)+\frac{\rho}{r^2\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{\rho}{r^2\sin^2 s}\cdot\frac{d^2\alpha}{dt^2}\right]$$

By multiplying everything by $2mr^2/(h^2\rho\alpha)$, and then moving the radial terms to the left, and the angular terms to the right, this latter equation can be written as follows:

$$\frac{2mr^2}{h^2}(V-E)-\frac{1}{\rho}\cdot\frac{d}{dr}\left(r^2\cdot\frac{d\rho}{dr}\right)=\frac{1}{\alpha\sin^2 s}\left[\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}\right]$$

Since this latter equation is now separated between radial and angular variables, both sides must be equal to a certain constant $-K$, and this gives the result. □

Let us first study the angular equation. The result here is as follows:

THEOREM 15.11. *The separated solutions $\alpha=\sigma(s)\theta(t)$ of the angular equation,*

$$\sin s\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d\alpha}{ds}\right)+\frac{d^2\alpha}{dt^2}=-K\sin^2 s\cdot\alpha$$

are given by the following formulae, where $l\in\mathbb{N}$ is such that $K=l(l+1)$,

$$\sigma(s)=P_l^m(\cos s)\quad,\quad\theta(t)=e^{imt}$$

and where $m\in\mathbb{Z}$ is a constant, and with P_l^m being the Legendre function,

$$P_l^m(x)=(-1)^m(1-x^2)^{m/2}\left(\frac{d}{dx}\right)^m P_l(x)$$

where P_l are the Legendre polynomials, given by the following formula:

$$P_l(x)=\frac{1}{2^l l!}\left(\frac{d}{dx}\right)^l(x^2-1)^l$$

These solutions $\alpha=\sigma(s)\theta(t)$ are called spherical harmonics.

PROOF. This follows indeed from a routine study, and with the comment that everything is taken up to linear combinations. We will normalize the wave function later. □

In order to finish our study, it remains to solve the radial equation, for the Coulomb potential V of the proton. As a first manipulation on the radial equation, we have:

PROPOSITION 15.12. *The radial equation, written with $K = l(l+1)$,*

$$(r^2\rho')' - \frac{2mr^2}{h^2}(V - E)\rho = l(l+1)\rho$$

takes with $\rho = u/r$ the following form, called modified radial equation,

$$Eu = -\frac{h^2}{2m} \cdot u'' + \left(V + \frac{h^2 l(l+1)}{2mr^2}\right) u$$

which is a time-independent 1D Schrödinger equation.

PROOF. With $\rho = u/r$ as in the statement, we have:

$$\rho = \frac{u}{r} \quad , \quad \rho' = \frac{u'r - u}{r^2} \quad , \quad (r^2\rho')' = u''r$$

By plugging this data into the radial equation, this becomes:

$$u''r - \frac{2mr}{h^2}(V - E)u = \frac{l(l+1)}{r} \cdot u$$

By multiplying everything by $h^2/(2mr)$, this latter equation becomes:

$$\frac{h^2}{2m} \cdot u'' - (V - E)u = \frac{h^2 l(l+1)}{2mr^2} \cdot u$$

But this gives the formula in the statement. As for the interpretation, as time-independent 1D Schrödinger equation, this is clear as well, and with the comment here that the term added to the potential V is some sort of centrifugal term. □

It remains to solve the above equation, for the Coulomb potential of the proton. And we have here the following result, which proves the original claims by Bohr:

THEOREM 15.13 (Schrödinger). *In the case of the hydrogen atom, where V is the Coulomb potential of the proton, the modified radial equation, which reads*

$$Eu = -\frac{h^2}{2m} \cdot u'' + \left(-\frac{Ke^2}{r} + \frac{h^2 l(l+1)}{2mr^2}\right) u$$

leads to the Bohr formula for allowed energies,

$$E_n = -\frac{m}{2}\left(\frac{Ke^2}{h}\right)^2 \cdot \frac{1}{n^2}$$

with $n \in \mathbb{N}$, the binding energy being

$$E_1 \simeq -2.177 \times 10^{-18}$$

with means $E_1 \simeq -13.591$ eV.

PROOF. This is again something non-trivial, the idea being as follows:

(1) By dividing our modified radial equation by E, this becomes:

$$-\frac{h^2}{2mE}\cdot u''=\left(1+\frac{Ke^2}{Er}-\frac{h^2l(l+1)}{2mEr^2}\right)u$$

In terms of $\alpha=\sqrt{-2mE}/h$, this equation takes the following form:

$$\frac{u''}{\alpha^2}=\left(1+\frac{Ke^2}{Er}+\frac{l(l+1)}{(\alpha r)^2}\right)u$$

In terms of the new variable $p=\alpha r$, this latter equation reads:

$$u''=\left(1+\frac{\alpha Ke^2}{Ep}+\frac{l(l+1)}{p^2}\right)u$$

Now let us introduce a new constant S for our problem, as follows:

$$S=-\frac{\alpha Ke^2}{E}$$

In terms of this new constant, our equation reads:

$$u''=\left(1-\frac{S}{p}+\frac{l(l+1)}{p^2}\right)u$$

(2) The idea will be that of looking for a solution written as a power series, but before that, we must "peel off" the asymptotic behavior. Which is something that can be done, of course, heuristically. With $p\to\infty$ we are led to $u''=u$, and ignoring the solution $u=e^p$ which blows up, our approximate asymptotic solution is:

$$u\sim e^{-p}$$

Similarly, with $p\to 0$ we are led to $u''=l(l+1)u/p^2$, and ignoring the solution $u=p^{-l}$ which blows up, our approximate asymptotic solution is:

$$u\sim p^{l+1}$$

(3) The above heuristic considerations suggest writing our function u as follows:

$$u=p^{l+1}e^{-p}v$$

So, let us do this. In terms of v, we have the following formula:

$$u'=p^le^{-p}\left[(l+1-p)v+pv'\right]$$

Differentiating a second time gives the following formula:

$$u''=p^le^{-p}\left[\left(\frac{l(l+1)}{p}-2l-2+p\right)v+2(l+1-p)v'+pv''\right]$$

Thus the radial equation, as modified in (1) above, reads:

$$pv''+2(l+1-p)v'+(S-2(l+1))v=0$$

(4) We will be looking for a solution v appearing as a power series:

$$v = \sum_{j=0}^{\infty} c_j p^j$$

But our equation leads to the following recurrence formula for the coefficients:

$$c_{j+1} = \frac{2(j+l+1) - S}{(j+1)(j+2l+2)} \cdot c_j$$

(5) We are in principle done, but we still must check that, with this choice for the coefficients c_j, our solution v, or rather our solution u, does not blow up. And the whole point is here. Indeed, at $j >> 0$ our recurrence formula reads, approximately:

$$c_{j+1} \simeq \frac{2c_j}{j}$$

But, surprisingly, this leads to $v \simeq c_0 e^{2p}$, and so to $u \simeq c_0 p^{l+1} e^p$, which blows up.

(6) As a conclusion, the only possibility for u not to blow up is that where the series defining v terminates at some point. Thus, we must have for a certain index j:

$$2(j+l+1) = S$$

In other words, we must have, for a certain integer $n > l$:

$$S = 2n$$

(7) We are almost there. Recall from (1) above that S was defined as follows:

$$S = -\frac{\alpha K e^2}{E} \quad : \quad \alpha = \frac{\sqrt{-2mE}}{h}$$

Thus, we have the following formula for the square of S:

$$S^2 = \frac{\alpha^2 K^2 e^4}{E^2} = -\frac{2mE}{h^2} \cdot \frac{K^2 e^4}{E^2} = -\frac{2mK^2 e^4}{h^2 E}$$

Now by using the formula $S = 2n$ from (6), the energy E must be of the form:

$$E = -\frac{2mK^2 e^4}{h^2 S^2} = -\frac{mK^2 e^4}{2h^2 n^2}$$

Calling this energy E_n, depending on $n \in \mathbb{N}$, we have, as claimed:

$$E_n = -\frac{m}{2}\left(\frac{Ke^2}{h}\right)^2 \cdot \frac{1}{n^2}$$

(8) Thus, we proved the Bohr formula. Regarding numerics, the data is as follows:

$$K = 8.988 \times 10^9 \quad , \quad e = 1.602 \times 10^{-19}$$

$$h = 1.055 \times 10^{-34} \quad , \quad m = 9.109 \times 10^{-31}$$

But this gives the formula of E_1 in the statement. □

In order to further advance, let us formulate our conclusions so far as follows:

THEOREM 15.14. *The wave functions of the hydrogen atom are the following functions, labelled by three quantum numbers, n, l, m,*

$$\phi_{nlm}(r, s, t) = \rho_{nl}(r)\alpha_l^m(s, t)$$

where $\rho_{nl}(r) = p^{l+1}e^{-p}v(p)/r$ with $p = \alpha r$ as before, with the coefficients of v subject to

$$c_{j+1} = \frac{2(j + l + 1 - n)}{(j + 1)(j + 2l + 2)} \cdot c_j$$

and $\alpha_l^m(s, t)$ being the spherical harmonics found before.

PROOF. This follows indeed by putting together all the results obtained so far, and with the remark that everything is up to the normalization of the wave function. □

In order to improve our results, we will need the following standard fact:

PROPOSITION 15.15. *The polynomials $v(p)$ are given by the formula*

$$v(p) = L_{n-l-1}^{2l+1}(p)$$

where the polynomials on the right, called associated Laguerre polynomials, are given by

$$L_q^p(x) = (-1)^p \left(\frac{d}{dx}\right)^p L_{p+q}(x)$$

with L_{p+q} being the Laguerre polynomials, given by the following formula,

$$L_q(x) = \frac{e^x}{q!} \left(\frac{d}{dx}\right)^q (e^{-x}x^q)$$

called Rodrigues formula for the Laguerre polynomials.

PROOF. The story here is very similar to that of the Legendre polynomials. Consider the Hilbert space $H = L^2[0, \infty)$, with the following scalar product on it:

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

The orthogonal basis obtained by applying Gram-Schmidt to the Weierstrass basis $\{x^q\}$ is then formed by the Laguerre polynomials $\{L_q\}$, and this gives the results. □

With the above result in hand, we can now improve our main results, as follows:

THEOREM 15.16. *The wave functions of the hydrogen atom are given by*

$$\phi_{nlm}(r, s, t) = \sqrt{\left(\frac{2}{na}\right)^3 \frac{(n - l - 1)!}{2n(n + l)!}} e^{-r/na} \left(\frac{2r}{na}\right)^l L_{n-l-1}^{2l+1}\left(\frac{2r}{na}\right) \alpha_l^m(s, t)$$

with $\alpha_l^m(s, t)$ being the spherical harmonics found before.

PROOF. This follows indeed by putting together what we have, and then doing some remaining work, concerning the normalization of the wave function. □

15b. Angular momentum

What is next? All sorts of corrections to the solution of the hydrogen atom discussed above, coming from relativity theory, electron spin, and other phenomena. Indeed, and with experiments confirming this, what we have above is just the tip of the iceberg.

However, all this is non-trivial and long business, and as a main objective for us, we would like to talk about electron spin. Following Uhlenbeck, Goudsmit, Pauli and others, we will first talk angular momentum, and then we will axiomatize spin as being the quantity which naturally "complements" the angular momentum. We first have:

PROPOSITION 15.17. *The components of the position operator $x = (x_1, x_2, x_3)$ and momentum operator $p = -ih\nabla$ satisfy the following relations,*

$$[x_i, x_j] = [p_i, p_j] = 0$$

$$[x_i, p_j] = ih\delta_{ij}$$

where $[a, b] = ab - ba$, called canonical commutation relations.

PROOF. All the above formulae are elementary, as follows:

(1) The components of the position operator $x = (x_1, x_2, x_3)$ obviously commute with each other, $x_i x_j = x_j x_i$, which makes their commutators vanish, $[x_i, x_j] = 0$.

(2) Regarding the momentum operator $p = -ih\nabla$, its components are as follows:

$$p_1 = -ih \cdot \frac{d}{dx_1} \quad , \quad p_2 = -ih \cdot \frac{d}{dx_2} \quad , \quad p_3 = -ih \cdot \frac{d}{dx_3}$$

Since partial derivatives commute with each other, we obtain $[p_i, p_j] = 0$.

(3) It remains to prove the last formula, and we have here:

$$\begin{aligned}
[x_i, p_j]f &= (x_i p_j - p_j x_i)f \\
&= -ih\left(x_i \cdot \frac{df}{dx_j} - \frac{d}{dx_j}(x_i f)\right) \\
&= -ih\left(x_i \cdot \frac{df}{dx_j} - \frac{dx_i}{dx_j} \cdot f - x_i \cdot \frac{df}{dx_j}\right) \\
&= ih \cdot \frac{dx_i}{dx_j} \cdot f \\
&= ih\delta_{ij} \cdot f
\end{aligned}$$

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

The above might look a bit complicated, and the simplest way to remember it is that "everyhing commutes", that is, $ab = ba$, except for the coordinates and momenta coordinates taken in the same direction, which are subject to the following rule:

$$x_i p_i = p_i x_i + ih$$

Getting now to angular momentum, it is convenient to change notation, with (x, y, z) instead of (x_1, x_2, x_3). We have the following result, to start with:

THEOREM 15.18. *The components of the angular momentum operator*

$$L = x \times (-ih\nabla)$$

satisfy the following equations,

$$[L_x, L_y] = ihL_z \quad , \quad [L_y, L_z] = ihL_x \quad , \quad [L_z, L_x] = ihL_y$$

called commutation relations for the angular momentum.

PROOF. With the more familiar notation $p = -ih\nabla$ for the momentum, or rather for the associated operator, the components of the angular momentum operator are:

$$L_x = yp_z - zp_y \quad , \quad L_y = zp_x - xp_z \quad , \quad L_z = xp_y - yp_x$$

Let us prove the first commutation relation. We have:

$$\begin{aligned} [L_x, L_y] &= [yp_z - zp_y, zp_x - xp_z] \\ &= [yp_z, zp_x] - [yp_z, xp_z] - [zp_y, zp_x] + [zp_y, xp_z] \end{aligned}$$

By heavily using the commutation relations from Proposition 15.17, we have:

$$[yp_z, zp_x] = yp_z zp_x - zp_x yp_z = y(zp_z - ih)p_x - zyp_x p_z = -ihyp_x$$

$$[yp_z, xp_z] = yp_z xp_z - xp_z yp_z = 0$$

$$[zp_y, zp_x] = zp_y zp_x - zp_x zp_y = 0$$

$$[zp_y, xp_z] = zp_y xp_z - xp_z zp_y = zxp_y p_z - x(zp_z - ih)p_y = ihxp_y$$

We conclude that the commutator that we were computing is given by the following formula, which is precisely the one in the statement:

$$\begin{aligned} [L_x, L_y] &= -ihyp_x + ihxp_y \\ &= ih(xp_y - yp_x) \\ &= ihL_z \end{aligned}$$

The proof of the other two commutation relations is similar, or can be simply obtained by invoking the cyclic invariance $x \to y \to z \to x$ of our problem, which cyclic invariance is not broken by the vector product $\times$ used, and so can indeed be invoked. □

As an interesting consequence of Theorem 15.18, we have:

PROPOSITION 15.19. *The following operator, called square of angular momentum*

$$L^2 = L_x^2 + L_y^2 + L_z^2$$

commutes with all 3 *operators* L_x, L_y, L_z.

PROOF. We have the following computation, to start with:

$$\begin{aligned}[L^2, L_x] &= (L_x^2 + L_y^2 + L_z^2)L_x - L_x(L_x^2 + L_y^2 + L_z^2) \\ &= L_y^2 L_x + L_z^2 L_x - L_x L_y^2 - L_x L_z^2 \\ &= [L_y^2, L_x] + [L_z^2, L_x]\end{aligned}$$

The first commutator can be computed with a trick, as follows:

$$\begin{aligned}[L_y^2, L_x] &= L_y L_y L_x - L_x L_y L_y \\ &= L_y L_y L_x - L_y L_x L_y + L_y L_x L_y - L_x L_y L_y \\ &= L_y[L_y, L_x] + [L_y, L_x]L_y \\ &= L_y(-ihL_z) + (-ihL_z)L_y \\ &= -ih(L_y L_z + L_z L_y)\end{aligned}$$

The second commutator can be computed with the same trick, as follows:

$$\begin{aligned}[L_z^2, L_x] &= L_z L_z L_x - L_x L_z L_z \\ &= L_z L_z L_x - L_z L_x L_z + L_z L_x L_z - L_x L_z L_z \\ &= L_z[L_z, L_x] + [L_z, L_x]L_z \\ &= L_z(ihL_y) + (ihL_y)L_z \\ &= ih(L_z L_y + L_y L_z)\end{aligned}$$

Now by summing we obtain the following commutation relation, as desired:

$$[L^2, L_x] = 0$$

The proof of the other two commutation relations is similar, or we can simply invoke here the cyclic symmetry argument from the end of the proof of Theorem 15.18. □

Let us discuss now the diagonalization of L_x, L_y, L_z. Since these operators do not commute, we cannot hope for a joint diagonalization. Thus, we must choose one of them, and for reasons that will become clear later, when writing things in spherical coordinates, we will choose L_x. In view of Proposition 15.19, this operator L_x does commute with L^2, and so we can hope for a joint diagonalization of L^2, L_x. And, so is what happens:

THEOREM 15.20. *The operators* L^2, L_x *diagonalize as*

$$L^2 f_l^m = h^2 l(l+1) f_l^m$$

$$L_x f_l^m = h m f_l^m$$

where $l \in \mathbb{N}/2$ *and* $m = -l, -l+1, \ldots, l-1, l$.

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

(1) For reasons that will become clear later on, let us introduce two operators as follows, called raising and lowering operators:

$$L_+ = L_y + iL_z \quad , \quad L_- = L_y - iL_z$$

We will often deal with these operators at the same time, using the following notation:

$$L_\pm = L_y \pm iL_z$$

(2) In order to get started, we first have the following computation:

$$\begin{aligned}
[L_x, L_\pm] &= [L_x, L_y] \pm i[L_x, L_z] \\
&= ihL_z \pm i(-ihL_y) \\
&= h(iL_z \pm L_y) \\
&= \pm h(\pm iL_z + L_y) \\
&= \pm hL_\pm
\end{aligned}$$

(3) Our claim now is that the conditions $L^2 f = \lambda f$, $L_x f = \mu f$ imply:

$$L^2(L_\pm f) = \lambda(L_\pm f)$$

$$L_x(L_\pm f) = (\mu \pm h)(L_\pm f)$$

Indeed, the first formula follows from the following computation:

$$\begin{aligned}
L^2(L_\pm f) &= L_\pm(L^2 f) \\
&= L_\pm(\lambda f) \\
&= \lambda(L_\pm f)
\end{aligned}$$

As for the second formula, this follows from the following computation:

$$\begin{aligned}
L_x(L_\pm f) &= L_x L_\pm f \\
&= (L_x L_\pm - L_\pm L_x)f + L_\pm L_x f \\
&= \pm hL_\pm f + L_\pm(\mu f) \\
&= (\mu \pm h)(L_\pm f)
\end{aligned}$$

(4) Now in view of the formulae found in (3), the raising and lowering operators act on the joint eigenfunctions of L^2, L_x, by leaving the L^2 eigenvalue unchanged, and by raising and lowering the eigenvalue of L_x. But both this raising process and lowering process for the eigenvalue of L_x cannot go on forever, because of the following estimate:

$$\begin{aligned}
\lambda &= < L^2 > \\
&= < L_x^2 > + < L_y^2 > + < L_z^2 > \\
&= \mu^2 + < L_y^2 > + < L_z^2 > \\
&\geq \mu^2
\end{aligned}$$

(5) In order to see exactly how the raising and lowering processes terminate, we will need some more computations. We first have the following computation:

$$\begin{aligned} L_\pm L_\mp &= (L_y \pm iL_z)(L_y \mp iL_z) \\ &= L_y^2 + L_z^2 \mp i(L_yL_z - L_zL_y) \\ &= L_y^2 + L_z^2 \mp i(ihL_x) \\ &= L_y^2 + L_z^2 \pm hL_x \\ &= L^2 - L_x^2 \pm hL_x \end{aligned}$$

We conclude from this that we have the following formula:

$$L^2 = L_\pm L_\mp + L_x^2 \mp hL_x$$

Now assuming $L_xf = hlf$, at termination of the raising process, we have:

$$\begin{aligned} L^2(f) &= (L_-L_+ + L_x^2 + hL_x)f \\ &= (0 + h^2l^2 + h^2l)f \\ &= h^2l(l+1)f \end{aligned}$$

Similarly, assuming $L_xf = hl'f$, at termination of the lowering process, we have:

$$\begin{aligned} L^2(f) &= (L_+ - L_- + L_x^2 - hL_x)f \\ &= (0 + h^2l'^2 - h^2l')f \\ &= h^2l'(l'-1)f \end{aligned}$$

Thus $l(l+1) = l'(l'-1)$, and since $l' = l+1$ is impossible, due to raising vs lowering, we must have $l' = -l$, and this leads to the conclusion in the statement. □

Moving ahead now, the above is all we need. The idea in what follows will be that of writing everything in spherical coordinates, and then finding the eigenfunctions.

And, what happens is that we have here the following remarkable result:

THEOREM 15.21. *In spherical coordinates r, s, t we have*

$$L_x = -\frac{ih}{dt}$$

$$L_y = ih\left(\frac{\sin t}{ds} + \frac{\cos s\cos t}{\sin s}\cdot\frac{1}{dt}\right)$$

$$L_z = -ih\left(\frac{\cos t}{ds} - \frac{\cos s\sin t}{\sin s}\cdot\frac{1}{dt}\right)$$

and the spherical harmonics are joint eigenfunctions of L^2, L_x.

PROOF. We recall that, according to our usual, N-dimensional looking conventions, the spherical coordinates that we use are as follows, with $r \in [0,\infty)$ being the radius, $s \in [0,\pi]$ being the polar angle, and $t \in [0,2\pi]$ being the azimuthal angle:

$$\begin{cases} x &= r\cos s \\ y &= r\sin s\cos t \\ z &= r\sin s\sin t \end{cases}$$

(1) We know that we have $L = -ihx \times \nabla$, so let us first compute ∇ in spherical coordinates. We have here, according to the chain rule for derivatives:

$$\begin{aligned} \nabla &= \begin{pmatrix} dr/dx & ds/dx & dt/dx \\ dr/dy & ds/dy & dt/dy \\ dr/dz & ds/dz & dt/dz \end{pmatrix} \begin{pmatrix} d/dr \\ d/ds \\ d/dt \end{pmatrix} \\ &= \begin{pmatrix} dx/dr & dy/dr & dz/dr \\ dx/ds & dy/ds & dz/ds \\ dx/dt & dy/dt & dz/dt \end{pmatrix}^{-1} \begin{pmatrix} d/dr \\ d/ds \\ d/dt \end{pmatrix} \end{aligned}$$

(2) On the other hand, it is routine to check that we have:

$$\begin{pmatrix} dx/dr & dx/ds & dx/dt \\ dy/dr & dy/ds & dy/dt \\ dz/dr & dz/ds & dz/dt \end{pmatrix} = \begin{pmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{pmatrix}$$

It is also routine to see that this latter matrix, say A, satisfies:

$$A^tA = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\sin^2 s \end{pmatrix}$$

Now if we call D the diagonal matrix on the right, we conclude that the matrix, say B, appearing in the above formula of ∇ is given by:

$$\begin{aligned} B &= (A^t)^{-1} \\ &= AD^{-1} \\ &= \begin{pmatrix} \cos s & -r\sin s & 0 \\ \sin s\cos t & r\cos s\cos t & -r\sin s\sin t \\ \sin s\sin t & r\cos s\sin t & r\sin s\cos t \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/r^2 & 0 \\ 0 & 0 & 1/(r^2\sin^2 s) \end{pmatrix} \\ &= \begin{pmatrix} \cos s & -\sin s/r & 0 \\ \sin s\cos t & \cos s\cos t/r & -\sin t/(r\sin s) \\ \sin s\sin t & \cos s\sin t/r & \cos t/(r\sin s) \end{pmatrix} \end{aligned}$$

(3) Thus, the angular momentum operator that we are looking for, $L = -ihx \times \nabla$, written more conveniently as $L = -ihx/r \times r\nabla$, is given by:

$$L = -ih \begin{pmatrix} \cos s \\ \sin s \cos t \\ \sin s \sin t \end{pmatrix} \times \begin{pmatrix} r\cos s & -\sin s & 0 \\ r\sin s\cos t & \cos s\cos t & -\sin t/\sin s \\ r\sin s\sin t & \cos s\sin t & \cos t/\sin s \end{pmatrix} \begin{pmatrix} d/dr \\ d/ds \\ d/dt \end{pmatrix}$$

And computing now the vector product gives the formula for L in the statement.

(4) Now with our explicit formula for L in hand, we next find that the raising and lowering operators are given by:

$$L_\pm = \pm he^{\pm it}\left(\frac{d}{ds} \pm i\,\frac{\cos s}{\sin s}\cdot\frac{1}{dt}\right)$$

Next, we find that these two operators satisfy the following formula:

$$L_+L_- = -h^2\left(\frac{d^2}{ds^2} + \frac{\cos s}{\sin s}\cdot\frac{d}{ds} + \frac{\cos^2 s}{\sin^2 s}\cdot\frac{d^2}{dt^2} + i\frac{d}{dt}\right)$$

And finally, by using this latter formula, we find that L^2 is given by:

$$L^2 = -h^2\left(\frac{1}{\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d}{ds}\right) + \frac{1}{\sin^2 s}\cdot\frac{d^2}{dt^2}\right)$$

(5) With all these formulae in hand, we can now finish. The eigenfunction equation for the above operator L^2, with eigenvalue $h^2l(l+1)$, is as follows:

$$-h^2\left(\frac{1}{\sin s}\cdot\frac{d}{ds}\left(\sin s\cdot\frac{d}{ds}\right) + \frac{1}{\sin^2 s}\cdot\frac{d^2}{dt^2}\right)f = h^2l(l+1)f$$

But this is precisely the angular equation found before. As for the eigenfunction equation for the operator L_x, with eigenvalue hm, this is as follows:

$$-\frac{ih}{dt}\,f = hmf$$

But this is equivalent to the azimuthal equation, and this gives the result. □

15c. Pauli matrices

In order to talk now about spin, we will regard, a bit as in the classical mechanics case, the spin and the angular momentum as being similar quantities. Thus, in analogy with the basic equations for the angular momentum, we should have:

DEFINITION 15.22. *The components of the spin operator are subject to*

$$[S_x, S_y] = ihS_z$$
$$[S_y, S_z] = ihS_x$$
$$[S_z, S_x] = ihS_y$$

called commutation relations for the spin operator.

The point now is that, with the above relations in hand, which are identical to the commutation relations for the angular momentum, all the general results from the previous section, based on that commutation relations, extend to our present setting, simply by changing L into S everywhere. And in particular, we are led in this way to:

THEOREM 15.23. *We have the following diagonalization formulae*

$$S^2 f_s^m = h^2 s(s+1) f_s^m$$

$$S_x f_s^m = h m f_s^m$$

$$S_\pm f_s^m = h\sqrt{s(s+1) - m(m\pm 1)}\, f_s^{m\pm 1}$$

involving the operators $S^2 = S_x^2 + S_y^2 + S_z^2$, S_x *and* $S_\pm = S_y \pm iS_z$.

PROOF. Here the first two formulae are something that we already know, from the previous section, with L, j being replaced by S, s. As for the last formula, this is something that we did not need, in the L, j context, but that we will need now. We want to compute the constants $C_{s,\pm}^m$ making work the raising and lowering formula, namely:

$$S_\pm f_s^m = C_{s,\pm}^m f_s^{m\pm 1}$$

But this can be done by using $S^2 = S_\pm S_\mp + S_x^2 \mp hS_x$ and $S_\pm^* = S_\mp$, and we get:

$$C_{s,+}^m = h\sqrt{s(s+1) - m(m+1)}$$

$$C_{s,-}^m = h\sqrt{s(s+1) - m(m-1)}$$

Thus, we are led to the last formula in the statement, and we are done. □

In practice now, let us look for the simplest realization of spin. We are led, for fixed particles, to a quantum mechanics over $H = \mathbb{C}^2$, with spin up and down being represented by the basis vectors.It remains to see the equations in Theorem 15.23 reformulate, in this $H = \mathbb{C}^2$ setting. But here, not many choices, and we are led in this way to:

DEFINITION 15.24. *In the quantum mechanics of the spin, over* $H = \mathbb{C}^2$, *with*

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad , \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

being spin up and down, the spin is subject to the following equations, for $f = e_1, e_2$,

$$S^2 f = h^2 s(s+1) f$$

$$S_x f = h m_f f$$

$$S_\pm f = h\sqrt{s(s+1) - m_f(m_f \pm 1)}\, \check{f}$$

with parameters $s = 1/2$, $m_{e_1} = 1/2$, $m_{e_2} = -1/2$, *and with* $\{e_1, e_2\} = \{f, \check{f}\}$.

Here all the choices, and notably $s = 1/2$, are very natural in view of Theorem 15.23, because these are the choices providing a "minimal" realization of the equations in Theorem 15.23, in the smallest possible number of dimensions, namely $N = 2$. However, all this comes with a shade of mystery, or at least is not rock-solid enough as to be called theorem, and it is probably safer to use the term "definition", as we did above.

The point now is that the above questions can be solved, the result being:

THEOREM 15.25. *In the above $H = \mathbb{C}^2$ context, of the mechanics of a single, fixed electron, the components of the normalized spin $\sigma = 2S/h$ are as follows,*

$$\sigma_x = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad , \quad \sigma_y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad , \quad \sigma_z = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

called Pauli matrices. In the general, dynamic context, where we already have a Hilbert space H for the wave function, spin can be introduced by using the space

$$H' = H \otimes \mathbb{C}^2$$

and using the above Pauli matrices for it, acting on the $\mathbb{C}^2$ part.

PROOF. As a first observation, we recognize in the above the Pauli matrices from chapter 3, which appeared there mathematically, in relation with SU_2, slightly modified. In what follows we will use the above new convention for the Pauli matrices. Regarding now the proof, the equations in Definition 15.24, written in full detail, are as follows:

$$S^2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{3h^2}{4} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad , \quad S^2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{3h^2}{4} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$S_x \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{h}{2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad , \quad S_x \begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\frac{h}{2} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$S_+ \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad , \quad S_+ \begin{pmatrix} 0 \\ 1 \end{pmatrix} = h \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$S_- \begin{pmatrix} 1 \\ 0 \end{pmatrix} = h \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad , \quad S_- \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Thus, we have the following formulae, for the various matrices involved:

$$S^2 = \frac{3h^2}{4} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \quad S_x = \frac{h}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$S_+ = h \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad , \quad S_- = h \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

In relation with what we want to prove, we have obtained the formula of S_x. Regarding now the formulae of S_y, S_z, these follow by solving the following system:

$$S_+ = S_y + iS_z \quad , \quad S_- = S_y - iS_z$$

To be more precise, the computation for S_y goes as follows:

$$S_y = \frac{S_+ + S_-}{2} = \frac{h}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

As for the computation for S_z, this goes as follows:

$$S_z = \frac{S_+ - S_-}{2i} = \frac{h}{2i}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \frac{h}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

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

As a first consequence of the above, looking quite good, we have:

FACT 15.26. *Electrons have spin* $1/2$.

This comes indeed from the formula $s = 1/2$ in Definition 15.24, and some further speculations are certainly possible. For instance the Pauli matrices all square up to one, $\sigma_i^2 = 1$, so we can say that "it takes $720°$ instead of the usual $360°$ to turn an electron back in place", leading to the conclusion that the electron spin is $360/720 = 1/2$.

15d. Dirac matrices

Getting now to more particle physics, as a continuation of the above, things are quite tricky, and as a guiding principle here, we know that the electron cannot live without the photon. Indeed, in the context of the basic physics of atoms, electrons can jump between energy levels, emitting or absorbing photons, and with this being known to happen even in the absence of external stimuli. Thus, the true "brother" of the electron is not the proton or the neutron, but rather the photon. And so, the minimal extension of quantum mechanics, that we should try to build now, should deal with electrons and photons.

In view of this, let us first look into the photon, try to understand how to make it fit into our theory, and leave the electron for later. As a starting point, we have:

FACT 15.27. *The master equation for free electromagnetic radiation, that is, for free photons, is the wave equation at speed* $v = c$*, namely:*

$$\ddot{\varphi} = c^2\Delta\varphi$$

This equation can be reformulated in the more symmetric form

$$\left(\frac{1}{c^2}\cdot\frac{d^2}{dt^2} - \Delta\right)\varphi = 0$$

with the operator on the left being called the d'Alembertian.

In relation now with the electron, there is an obvious similarity here with the free Schrödinger equation, without potential V, which reads:

$$\left(i\frac{d}{dt}+\frac{h}{2m}\,\Delta\right)\psi=0$$

This similarity suggests looking for a relativistic version of the Schrödinger equation, which is compatible with the wave equation at $v=c$. And coming up with such an equation is not very complicated, the straightforward answer being as follows:

DEFINITION 15.28. *The following abstract mathematical equation,*

$$\left(-\frac{1}{c^2}\cdot\frac{d^2}{dt^2}+\Delta\right)\psi=\frac{m^2c^2}{h^2}\,\psi$$

on a function $\psi=\psi_t(x)$, is called the Klein-Gordon equation.

To be more precise, what we have here is some sort of a speculatory equation, formally obtained from the Schrödinger equation, via a few simple manipulations, as to make it relativistic. And with the relation with photons being something very simple, the thing being that at zero mass, $m=0$, we obtain precisely the wave equation at $v=c$.

All this is very nice, looks like we have a beginning of theory here, both making the electrons relativistic, and unifying them with photons. And isn't this too beautiful to be true. Going ahead now with physics, the following question appears:

QUESTION 15.29. *What does the Klein-Gordon equation really describe?*

And here, unfortunately, bad news all the way. A closer look at the Klein-Gordon equation reveals all sorts of bugs, making it unusable for anything reasonable. And with the main bug, which is enough for disqualifying it, being that, unlike the Schrödinger equation which preserves probability amplitudes $|\psi|^2$, the Klein-Gordon equation does not have this property. Thus, even before trying to understand what the Klein-Gordon equation really describes, we are left with the conclusion that this equation cannot really describe anything reasonable, due to the formal nature of the function ψ involved.

So, this was for the story of the Klein-Gordon equation. Actually this equation was first discovered by Schrödinger himself, in the context of his original work on the Schrödinger equation. But noticing the above bugs with it, Schrödinger dismissed it right away, and then downgraded his objectives, looking for something non-relativistic instead, and then found the Schrödinger equation, leading to the story that we know.

This being said, the Klein-Gordon equation found later a number of interesting applications, the continuation of the story being as follows:

(1) Dirac found a clever way of extracting the "square root" of the Klein-Gordon equation. And this square root equation, called Dirac equation, turned out to be the correct one, making exactly what the Klein-Gordon equation was supposed to do.

(2) Technically speaking, the Klein-Gordon equation is very useful for investigating the Dirac equation, because the components of the solutions of the Dirac equation satisfy the Klein-Gordon equation. More on this later, when discussing the Dirac equation.

(3) Finally, the Klein-Gordon equation was later recognized to describe well the spin 0 particles. But with these particles being something specialized, including the unstable and sowewhat fringe "pions", and the Higgs boson, which is something complicated.

We will briefly discuss all this, in what follows. Getting to work now, following Dirac, the idea is that of extracting the square root of the Klein-Gordon operator, as follows:

PROPOSITION 15.30. *We can extract the square root of the Klein-Gordon operator, via a formula as follows,*

$$-\frac{1}{c^2}\cdot\frac{d^2}{dt^2}+\Delta=\left(\frac{i}{c}\cdot\frac{Pd}{dt}+\frac{Qd}{dx}+\frac{Rd}{dy}+\frac{Sd}{dz}\right)^2$$

by using matrices P,Q,R,S which anticommute, $AB=-BA$, and whose squares equal one, $A^2=1$.

PROOF. We have the following computation, valid for any matrices P,Q,R,S, with the notation $\{A,B\}=AB+BA$:

$$\begin{aligned}\left(\frac{i}{c}\cdot\frac{Pd}{dt}+\frac{Qd}{dx}+\frac{Rd}{dy}+\frac{Sd}{dz}\right)^2 &= -\frac{1}{c^2}\cdot\frac{P^2d^2}{dt^2}+\frac{Q^2d^2}{dx^2}+\frac{R^2d^2}{dy^2}+\frac{S^2d^2}{dz^2}\\ &+\frac{i}{c}\left(\frac{\{P,Q\}d^2}{dtdx}+\frac{\{P,R\}d^2}{dtdy}+\frac{\{P,S\}d^2}{dtdz}\right)\\ &+\frac{\{Q,R\}d^2}{dxdy}+\frac{\{Q,S\}d^2}{dxdz}+\frac{\{R,S\}d^2}{dydz}\end{aligned}$$

Thus, in order to obtain in this way the Klein-Gordon operator, the conditions in the statement must be satisfied. □

As a technical comment here, normally when extracting a square root, we should look for a self-adjoint operator. In view of this, observe that we have:

$$\left(\frac{i}{c}\cdot\frac{Pd}{dt}+\frac{Qd}{dx}+\frac{Rd}{dy}+\frac{Sd}{dz}\right)^*=-\frac{i}{c}\cdot\frac{P^*d}{dt}+\frac{Q^*d}{dx}+\frac{R^*d}{dy}+\frac{S^*d}{dz}$$

Thus, we should normally add the conditions $P^*=-P$ and $Q^*=Q$, $R^*=R$, $S^*=S$ to those above. But, the thing is that due to some subtle reasons, the natural square root of the Klein-Gordon operator is not self-adjoint. More on this later.

Looking for matrices P, Q, R, S as above is not exactly trivial, and the simplest solutions appear in $M_4(\mathbb{C})$, in connection with the Pauli matrices, as follows:

PROPOSITION 15.31. *The simplest matrices P, Q, R, S as above appear as*

$$P = \gamma_0 \quad , \quad Q = i\gamma_1 \quad , \quad R = i\gamma_2 \quad , \quad S = i\gamma_3$$

with $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ being the Dirac matrices, given by

$$\gamma_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad , \quad \gamma_i = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix}$$

where $\sigma_1, \sigma_2, \sigma_3$ are the Pauli spin matrices, given by

$$\sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad , \quad \sigma_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad , \quad \sigma_3 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

that we met before, in the context of the electron spin.

PROOF. We have $\gamma_0^2 = 1$, and by using $\sigma_i^2 = 1$ for any $i = 1, 2, 3$, we have as well the following formula, which shows that we have $(i\gamma_i)^2 = 1$, as needed:

$$\gamma_i^2 = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix} \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

As in what regards the commutators, we first have, for any $i = 1, 2, 3$, the following equalities, which show that γ_0 anticommutes indeed with γ_i:

$$\gamma_0\gamma_i = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix} = \begin{pmatrix} 0 & \sigma_i \\ \sigma_i & 0 \end{pmatrix}$$

$$\gamma_i\gamma_0 = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -\sigma_i \\ -\sigma_i & 0 \end{pmatrix}$$

Regarding now the remaining commutators, observe here that we have:

$$\gamma_i\gamma_j = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix} \begin{pmatrix} 0 & \sigma_j \\ -\sigma_j & 0 \end{pmatrix} = \begin{pmatrix} -\sigma_i\sigma_j & 0 \\ 0 & -\sigma_i\sigma_j \end{pmatrix}$$

Now since the Pauli matrices anticommute, we obtain $\gamma_i\gamma_j = -\gamma_j\gamma_i$, as desired. □

We can now put everything together, and we obtain:

THEOREM 15.32. *The following operator, called Dirac operator,*

$$D = i\left(\frac{\gamma_0 d}{cdt} + \frac{\gamma_1 d}{dx} + \frac{\gamma_2 d}{dy} + \frac{\gamma_3 d}{dz}\right)$$

has the property that its square is the Klein-Gordon operator.

PROOF. With notations from Proposition 15.30 and Proposition 15.31, and by making the choices in Proposition 15.26, we have:

$$\begin{aligned}\frac{i}{c}\cdot\frac{Pd}{dt}+\frac{Qd}{dx}+\frac{Rd}{dy}+\frac{Sd}{dz} &= \frac{i}{c}\cdot\frac{\gamma_0 d}{dt}+\frac{i\gamma_1 d}{dx}+\frac{i\gamma_2 d}{dy}+\frac{i\gamma_3 d}{dz}\\ &= i\left(\frac{\gamma_0 d}{cdt}+\frac{\gamma_1 d}{dx}+\frac{\gamma_2 d}{dy}+\frac{\gamma_3 d}{dz}\right)\end{aligned}$$

Thus, we have here a square root of the Klein-Gordon operator, as desired. □

We can now extract the square root of the Klein-Gordon equation, as follows:

THEOREM 15.33. *We have the following equation, called Dirac equation,*

$$ih\left(\frac{\gamma_0 d}{cdt}+\frac{\gamma_1 d}{dx}+\frac{\gamma_2 d}{dy}+\frac{\gamma_3 d}{dz}\right)\psi=mc\psi$$

obtained by extracting the square root of the Klein-Gordon equation.

PROOF. This is more of a definition, based on the above, but we have called it Theorem, in view of its importance, and for finishing this chapter in beauty, with a theorem. □

In practice now, as usual with such theoretical physics equations, extreme caution is recommended, at least to start with. However, with a bit of patience, the Dirac equation can be systematically studied, and the good news is that, passed a few difficulties, this is indeed a true, magic equation, basically solving the problems mentioned in the beginning of this section. For more on this, you can check any particle physics book.

15e. Exercises

Tough physics chapter that we had here, and as useful exercises, we have:

EXERCISE 15.34. *Learn if needed the basics of electrostatics.*

EXERCISE 15.35. *Learn as well, if needed, the basics of electrodynamics.*

EXERCISE 15.36. *Read about radiation, light, optics and spectroscopy.*

EXERCISE 15.37. *Learn about spectral lines, and Lyman, Balmer, Paschen.*

EXERCISE 15.38. *Read about the Ritz-Rydberg combination principle.*

EXERCISE 15.39. *Read also about Max Planck, quanta, and Bohr's claims.*

EXERCISE 15.40. *Read about the Heisenberg matrix mechanics.*

EXERCISE 15.41. *With all the above learned, read again the present chapter.*

As bonus exercise, for more, read some quantum electrodynamics. All good stuff.

CHAPTER 16

Random matrices

16a. Random matrices

With quantum physics discussed you would say, end of the story, with math and physics and everything, and in particular, end of the present book. However, no advanced linear algebra book would be complete without a word on the random matrices, which are a natural, far-reaching generalization of the usual matrices. These matrices, which appear in a wide array of questions in mathematics and physics, and even in more bizarre contexts, such as economics and finance, are something very simple, as follows:

DEFINITION 16.1. *A random matrix is a matrix with random variables as entries,*

$$Z \in M_N(L^\infty(X))$$

with X being a probability space.

Regarding now the mathematics of such matrices, we will be mainly interested in computing their law. Recall indeed from chapter 3 that any scalar matrix $A \in M_N(\mathbb{C})$ has a certain abstract law, which in the normal case, $AA^* = A^*A$, is a usual probability measure on $\mathbb{C}$, and in the self-adjoint case, $A = A^*$, is a usual probability measure on $\mathbb{R}$. We will see in a moment that the same happens for the random matrices, axiomatized as above, and in view of this, the following question makes sense:

QUESTION 16.2. *What are the laws of the various basic types of random matrices, as for instance those having i.i.d. entries, subject to simple constraints? Also, what happens to these laws in the $N >> 0$ regime, do we have some interesting asymptotics?*

So, this was for the definition and main question regarding the random matrices, and with the motivations for all this coming from a remarkable mixture of first class mathematics and physics, featuring all sorts of interesting questions in probability theory, operator algebras, statistical mechanics, quantum mechanics, and many more.

Excited about this? So am I, and before starting, a piece of advertisement too:

ANSWER 16.3. *The above random matrix questions can all be solved, with very interesting answers, involving among others our favorite groups, SU_2 and SO_3.*

Getting to work now, as a first task, mentioned above, we must extend the spectral measure material from chapter 3, from the case of the matrix algebras $M_N(\mathbb{C})$ to the case

of the random matrix algebras $M_N(L^\infty(X))$. Now since these latter algebras are infinite dimensional, the best is to start our study with a discussion of the operator algebras, coming as a continuation of the operator theory from chapter 8. We first have:

DEFINITION 16.4. *An abstract operator algebra, or C^*-algebra, is a complex algebra A having a norm $||.||$ and an involution $*$, subject to the following conditions:*

(1) *A is closed with respect to the norm.*
(2) *We have $||aa^*|| = ||a||^2$, for any $a \in A$.*

In other words, what we did here is to axiomatize the abstract properties of the operator algebras $A \subset B(H)$, coming from the various general results about linear operators from chapter 8, without any reference to the ambient Hilbert space H.

As basic examples now, we have the usual matrix algebras $M_N(\mathbb{C})$, with the norm and the involution being the usual matrix norm and involution, given by:

$$||A|| = \sup_{||x||=1} ||Ax|| \quad , \quad (A^*)_{ij} = \overline{A}_{ji}$$

Some other basic examples are the algebras $L^\infty(X)$ of essentially bounded functions $f : X \to \mathbb{C}$ on a measured space X, with the usual norm and involution, namely:

$$||f|| = \sup_{x \in X} |f(x)| \quad , \quad f^*(x) = \overline{f(x)}$$

We can put these two basic classes of examples together, as follows:

PROPOSITION 16.5. *The random matrix algebras $A = M_N(L^\infty(X))$ are C^*-algebras, with their usual norm and involution, given by:*

$$||Z|| = \sup_{x \in X} ||Z_x|| \quad , \quad (Z^*)_{ij} = \overline{Z}_{ij}$$

These algebras generalize both the algebras $M_N(\mathbb{C})$, and the algebras $L^\infty(X)$.

PROOF. The fact that the C^*-algebra axioms are satisfied is clear from definitions. As for the last assertion, this follows by taking $X = \{.\}$ and $N = 1$, respectively. □

We can in fact say more about the above algebras, as follows:

THEOREM 16.6. *Any algebra of type $L^\infty(X)$ is an operator algebra, as follows:*

$$L^\infty(X) \subset B(L^2(X)) \quad , \quad f \to (g \to fg)$$

More generally, any random matrix algebra is an operator algebra, as follows,

$$M_N(L^\infty(X)) \subset B\left(\mathbb{C}^N \otimes L^2(X)\right)$$

with the embedding being the above one, tensored with the identity.

PROOF. We have two assertions to be proved, the idea being as follows:

(1) Given $f \in L^\infty(X)$, consider the following operator, acting on $H = L^2(X)$:

$$T_f(g) = fg$$

Observe that T_f is indeed well-defined, and bounded as well, because:

$$||fg||_2 = \sqrt{\int_X |f(x)|^2|g(x)|^2 d\mu(x)} \leq ||f||_\infty ||g||_2$$

The application $f \to T_f$ being linear, involutive, continuous, and injective as well, we obtain in this way a C^*-algebra embedding $L^\infty(X) \subset B(H)$, as desired.

(2) Regarding the second assertion, this is best viewed in the following way:

$$\begin{aligned} M_N(L^\infty(X)) &= M_N(\mathbb{C}) \otimes L^\infty(X) \\ &\subset M_N(\mathbb{C}) \otimes B(L^2(X)) \\ &= B\left(\mathbb{C}^N \otimes L^2(X)\right) \end{aligned}$$

Here we have used (1), and some standard tensor product identifications. □

Our purpose in what follows is to develop the spectral theory of the C^*-algebras, and in particular that of the random matrix algebras $A = M_N(L^\infty(X))$ that we are interested in, one of our objectives being that of talking about spectral measures, in the normal case, in analogy with what we know about the usual matrices. Let us start with:

THEOREM 16.7. *Given an element $a \in A$ of a C^*-algebra, define its spectrum as:*

$$\sigma(a) = \left\{\lambda \in \mathbb{C}\middle| a - \lambda \notin A^{-1}\right\}$$

The following spectral theory results hold, exactly as in the $A = B(H)$ case:

(1) *We have $\sigma(ab) \cup \{0\} = \sigma(ba) \cup \{0\}$.*
(2) *We have $\sigma(f(a)) = f(\sigma(a))$, for any $f \in \mathbb{C}(X)$ having poles outside $\sigma(a)$.*
(3) *The spectrum $\sigma(a)$ is compact, non-empty, and contained in $D_0(||a||)$.*
(4) *The spectra of unitaries ($u^* = u^{-1}$) and self-adjoints ($a = a^*$) are on $\mathbb{T}, \mathbb{R}$.*
(5) *The spectral radius of normal elements ($aa^* = a^*a$) is given by $\rho(a) = ||a||$.*

In addition, assuming $a \in A \subset B$, the spectra of a with respect to A and to B coincide.

PROOF. Here the assertions (1-5), which are of course formulated a bit informally, are well-known for the full operator algebra $A = B(H)$, and the proof in general is similar:

(1) Assuming that $1 - ab$ is invertible, with inverse c, we have $abc = cab = c - 1$, and it follows that $1 - ba$ is invertible too, with inverse $1 + bca$. Thus $\sigma(ab), \sigma(ba)$ agree on $1 \in \mathbb{C}$, and by linearity, it follows that $\sigma(ab), \sigma(ba)$ agree on any point $\lambda \in \mathbb{C}^*$.

(2) The formula $\sigma(f(a)) = f(\sigma(a))$ is clear for polynomials, $f \in \mathbb{C}[X]$, by factorizing $f - \lambda$, with $\lambda \in \mathbb{C}$. Then, the extension to the rational functions is straightforward, because $P(a)/Q(a) - \lambda$ is invertible precisely when $P(a) - \lambda Q(a)$ is.

(3) By using $1/(1-b) = 1 + b + b^2 + \ldots$ for $||b|| < 1$ we obtain that $a - \lambda$ is invertible for $|\lambda| > ||a||$, and so $\sigma(a) \subset D_0(||a||)$. It is also clear that $\sigma(a)$ is closed, so what we have is a compact set. Finally, assuming $\sigma(a) = \emptyset$ the function $f(\lambda) = \varphi((a-\lambda)^{-1})$ is well-defined, for any $\varphi \in A^*$, and by Liouville we get $f = 0$, contradiction.

(4) Assuming $u^* = u^{-1}$ we have $||u|| = 1$, and so $\sigma(u) \subset D_0(1)$. But with $f(z) = z^{-1}$ we obtain via (2) that we have as well $\sigma(u) \subset f(D_0(1))$, and this gives $\sigma(u) \subset \mathbb{T}$. As for the result regarding the self-adjoints, this can be obtained from the result for the unitaries, by using (2) with functions of type $f(z) = (z+it)/(z-it)$, with $t \in \mathbb{R}$.

(5) It is routine to check, by integrating quantities of type $z^n/(z-a)$ over circles centered at the origin, and estimating, that the spectral radius is given by $\rho(a) = \lim ||a^n||^{1/n}$. But in the self-adjoint case, $a = a^*$, this gives $\rho(a) = ||a||$, by using exponents of type $n = 2^k$, and then the extension to the general normal case is straightforward.

(6) Regarding now the last assertion, the inclusion $\sigma_B(a) \subset \sigma_A(a)$ is clear. For the converse, assume $a - \lambda \in B^{-1}$, and set $b = (a-\lambda)^*(a-\lambda)$. We have then:

$$\sigma_A(b) - \sigma_B(b) = \left\{\mu \in \mathbb{C} - \sigma_B(b) \middle| (b-\mu)^{-1} \in B - A\right\}$$

Thus this difference in an open subset of $\mathbb{C}$. On the other hand b being self-adjoint, its two spectra are both real, and so is their difference. Thus the two spectra of b are equal, and in particular b is invertible in A, and so $a - \lambda \in A^{-1}$, as desired. □

We can now a prove a key result about operator algebras, as follows:

THEOREM 16.8 (Gelfand). *If X is a compact space, the algebra $C(X)$ of continuous functions on it $f : X \to \mathbb{C}$ is a C^*-algebra, with usual norm and involution, namely:*

$$||f|| = \sup_{x \in X} |f(x)| \quad , \quad f^*(x) = \overline{f(x)}$$

Conversely, any commutative C^*-algebra is of this form, $A = C(X)$, with

$$X = \left\{\chi : A \to \mathbb{C} \ , \ \text{normed algebra character}\right\}$$

with topology making continuous the evaluation maps $ev_a : \chi \to \chi(a)$.

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

(1) The first assertion is clear from definitions. Observe that we have indeed:

$$||ff^*|| = \sup_{x \in X} |f(x)|^2 = ||f||^2$$

Observe also that the algebra $C(X)$ is commutative, because $fg = gf$.

(2) Conversely, given a commutative C^*-algebra A, let us define X as in the statement. Then X is compact, and $a \to ev_a$ is a morphism of algebras, as follows:

$$ev : A \to C(X)$$

(3) We first prove that ev is involutive. We use the following formula, which is similar to the $z = Re(z) + iIm(z)$ decomposition formula for usual complex numbers:

$$a = \frac{a + a^*}{2} + i \cdot \frac{a - a^*}{2i}$$

Thus it is enough to prove $ev_{a^*} = ev_a^*$ for the self-adjoint elements a. But this is the same as proving that $a = a^*$ implies that ev_a is a real function, which is in turn true, by Theorem 16.7, because $ev_a(\chi) = \chi(a)$ is an element of $\sigma(a)$, contained in $\mathbb{R}$.

(4) Since A is commutative, each element is normal, so ev is isometric:

$$||ev_a|| = \rho(a) = ||a||$$

It remains to prove that ev is surjective. But this follows from the Stone-Weierstrass theorem, because $ev(A)$ is a closed subalgebra of $C(X)$, which separates the points. □

As a main consequence of the Gelfand theorem, we have:

THEOREM 16.9. *For any normal element $a \in A$ we have an identification as follows:*

$$< a >= C(\sigma(a))$$

In addition, given a function $f \in C(\sigma(a))$, we can apply it to a, and we have

$$\sigma(f(a)) = f(\sigma(a))$$

which generalizes the previous rational calculus formula, in the normal case.

PROOF. Since a is normal, the C^*-algebra $< a >$ that is generates is commutative, so if we denote by X the space of the characters $\chi :< a >\to \mathbb{C}$, we have:

$$< a >= C(X)$$

Now since the map $X \to \sigma(a)$ given by evaluation at a is bijective, we obtain:

$$< a >= C(\sigma(a))$$

Thus, we are dealing here with usual functions, and this gives all the assertions. □

In order to get now towards noncommutative probability, we first have to develop the theory of positive elements, and linear forms. First, we have the following result:

PROPOSITION 16.10. *For an element $a \in A$, the following are equivalent:*

(1) *a is positive, in the sense that $\sigma(a) \subset [0, \infty)$.*
(2) *$a = b^2$, for some $b \in A$ satisfying $b = b^*$.*
(3) *$a = cc^*$, for some $c \in A$.*

PROOF. This is something very standard, as follows:

(1) $\implies$ (2) Observe first that $\sigma(a) \subset \mathbb{R}$ implies $a = a^*$. Thus the algebra $< a >$ is commutative, and by using Theorem 16.9, we can set $b = \sqrt{a}$.

(2) $\implies$ (3) This is trivial, because we can simply set $c = b$.

(2) $\implies$ (1) This is clear too, because we have:

$$\sigma(a) = \sigma(b^2) = \sigma(b)^2 \subset \mathbb{R}^2 = [0, \infty)$$

(3) $\implies$ (1) We can proceed here by contradiction. Indeed, by multiplying c by a suitable element of the algebra $< cc^* >$, we are led to the existence of an element $d \neq 0$ satisfying $-dd^* \geq 0$. By writing now $d = x + iy$ with $x = x^*, y = y^*$ we have:

$$dd^* + d^*d = 2(x^2 + y^2) \geq 0$$

Thus $d^*d \geq 0$, which is easily seen to contradict the condition $-dd^* \geq 0$. □

We can talk as well about positive linear forms, as follows:

DEFINITION 16.11. *Consider a linear map $\varphi : A \to \mathbb{C}$.*

(1) *φ is called positive when $a \geq 0 \implies \varphi(a) \geq 0$.*
(2) *φ is called faithful and positive when $a \geq 0, a \neq 0 \implies \varphi(a) > 0$.*

In the commutative case, $A = C(X)$, the positive linear forms appear as follows, with μ being positive, and strictly positive if we want φ to be faithful and positive:

$$\varphi(f) = \int_X f(x) d\mu(x)$$

In general, the positive linear forms can be thought of as being integration functionals with respect to some underlying "positive measures". Based on this, we can formulate:

DEFINITION 16.12. *Let A be a C^*-algebra, given with a positive trace $tr : A \to \mathbb{C}$.*

(1) *The elements $a \in A$ are called random variables.*
(2) *The moments of such a variable are the numbers $M_k(a) = tr(a^k)$.*
(3) *The law of such a variable is the functional $\mu_a : P \to tr(P(a))$.*

Here the exponent $k = \circ \bullet \bullet \circ \ldots$ is by definition a colored integer, and the powers a^k are defined by the following formulae, and multiplicativity:

$$a^{\emptyset} = 1 \quad , \quad a^{\circ} = a \quad , \quad a^{\bullet} = a^*$$

As for the polynomial P, this is a noncommuting $*$-polynomial in one variable:

$$P \in \mathbb{C} < X, X^* >$$

Observe that the law is uniquely determined by the moments, because we have:

$$P(X) = \sum_k \lambda_k X^k \implies \mu_a(P) = \sum_k \lambda_k M_k(a)$$

At the level of the general theory, we have the following key result, extending the various results that we have, regarding the self-adjoint and normal matrices:

THEOREM 16.13. *Let A be a C^*-algebra, with a trace tr, and consider an element $a \in A$ which is normal, in the sense that $aa^* = a^*a$.*

(1) *μ_a is a complex probability measure, satisfying $supp(\mu_a) \subset \sigma(a)$.*
(2) *In the self-adjoint case, $a = a^*$, this measure μ_a is real.*
(3) *Assuming that tr is faithful, we have $supp(\mu_a) = \sigma(a)$.*

PROOF. In the normal case, $aa^* = a^*a$, the Gelfand theorem, or rather the subsequent continuous functional calculus theorem, tells us that we have:

$$< a >= C(\sigma(a))$$

Thus the functional $f(a) \to tr(f(a))$ can be regarded as an integration functional on the algebra $C(\sigma(a))$, and by the Riesz theorem, this latter functional must come from a probability measure μ on the spectrum $\sigma(a)$, in the sense that we must have:

$$tr(f(a)) = \int_{\sigma(a)} f(z)d\mu(z)$$

We are therefore led to the conclusions in the statement, with the uniqueness assertion coming from the fact that the elements a^k, taken as usual with respect to colored integer exponents, $k = \circ \bullet \bullet \circ \ldots$, generate the whole C^*-algebra $C(\sigma(a))$. □

As a first concrete application now, by getting back to the random matrices, and to the various questions raised in the beginning of this chapter, we have:

THEOREM 16.14. *Given a random matrix $Z \in M_N(L^\infty(X))$ which is normal,*

$$ZZ^* = Z^*Z$$

its law, which is by definition the following abstract functional,

$$\mu : \mathbb{C} < X, X^* > \to \mathbb{C} \quad , \quad P \to \frac{1}{N} \int_X tr(P(Z))$$

when restricted to the usual polynomials in two variables,

$$\mu : \mathbb{C}[X, X^*] \to \mathbb{C} \quad , \quad P \to \frac{1}{N} \int_X tr(P(Z))$$

must come from a probability measure on the spectrum $\sigma(Z) \subset \mathbb{C}$, as follows:

$$\mu(P) = \int_{\sigma(T)} P(x)d\mu(x)$$

We agree to use the symbol μ for all these notions.

PROOF. This follows indeed from what we know from Theorem 16.13, applied to the normal element $a = Z$, belonging to the C^*-algebra $A = M_N(L^\infty(X))$. □

At the level of the basic examples, the situation is as follows:

THEOREM 16.15. *The basic random matrices $Z \in M_N(L^\infty(X))$ are as follows:*

(1) *In the case $N = 1$ the random matrix is a usual random variable, $f \in L^\infty(X)$, automatically normal, and its law as defined above is the usual law.*
(2) *In the case $X = \{.\}$ the random matrix is a usual scalar matrix, $A \in M_N(\mathbb{C})$, and in the diagonalizable case, the law is $\mu = \frac{1}{N}(\delta_{\lambda_1} + \ldots + \delta_{\lambda_N})$.*

PROOF. This is clear indeed, with the first assertion coming from definitions, and the second assertion coming by diagonalizing the matrix, as explained in chapter 3. □

At a more advanced level now, the main problem regarding the random matrices is that of computing the law of various classes of such matrices, coming in series:

QUESTION 16.16. *What is the law of random matrices coming in series*

$$Z_N \in M_N(L^\infty(X))$$

in the $N >> 0$ regime?

The general strategy here, coming from physicists, is that of computing first the asymptotic law μ^0, in the $N \to \infty$ limit, and then looking for the higher order terms as well, as to finally reach to a series in N^{-1} giving the law of Z_N, as follows:

$$\mu_N = \mu^0 + N^{-1}\mu^1 + N^{-2}\mu^2 + \ldots$$

As a basic example here, of particular interest are the matrices having i.i.d. complex normal entries, under the constraint $Z = Z^*$. Here the asymptotic law μ^0 is the Wigner semicircle law on $[-2, 2]$. We will discuss this in a moment, after some preliminaries.

16b. Gaussian matrices

We recall that a random matrix algebra is an algebra of type $A = M_N(L^\infty(X))$, and that we are interested in the computation of the laws of the operators $Z \in A$, called random matrices. Regarding the precise classes of random matrices that we are interested in, first we have the complex Gaussian matrices, which are constructed as follows:

DEFINITION 16.17. *A complex Gaussian matrix is a random matrix of type*

$$Z \in M_N(L^\infty(X))$$

which has i.i.d. complex normal entries.

We will see that the above matrices have an interesting, and "central" combinatorics, among all kinds of random matrices, with the study of the other random matrices being usually obtained as a modification of the study of the Gaussian matrices.

As a somewhat surprising remark, using real normal variables in Definition 16.17, instead of the complex ones appearing there, leads nowhere. The correct real versions of the Gaussian matrices are the Wigner random matrices, constructed as follows:

DEFINITION 16.18. *A Wigner matrix is a random matrix of type*

$$Z \in M_N(L^\infty(X))$$

which has i.i.d. complex normal entries, up to the constraint $Z = Z^*$.

In other words, a Wigner matrix must be as follows, with the diagonal entries being real normal variables, $a_i \sim g_t$, for some $t > 0$, the upper diagonal entries being complex normal variables, $b_{ij} \sim G_t$, the lower diagonal entries being the conjugates of the upper diagonal entries, as indicated, and with all the variables a_i, b_{ij} being independent:

$$Z = \begin{pmatrix} a_1 & b_{12} & \ldots & \ldots & b_{1N} \\ \bar{b}_{12} & a_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{N-1} & b_{N-1,N} \\ \bar{b}_{1N} & \ldots & \ldots & \bar{b}_{N-1,N} & a_N \end{pmatrix}$$

As a comment here, for many concrete applications the Wigner matrices are in fact the central objects in random matrix theory, and in particular, they are often more important than the Gaussian matrices. In fact, these are the random matrices which were first considered and investigated, a long time ago, by Wigner himself.

Finally, we will be interested as well in the complex Wishart matrices, which are the positive versions of the above random matrices, constructed as follows:

DEFINITION 16.19. *A complex Wishart matrix is a random matrix of type*

$$Z = YY^* \in M_N(L^\infty(X))$$

with Y being a complex Gaussian matrix.

As before with the Gaussian and Wigner matrices, there are many possible comments that can be made here, of technical or historical nature. First, using in the above real Gaussian variables instead of complex ones leads to a less interesting combinatorics. Also, these matrices were introduced and studied by Marchenko-Pastur not long after Wigner, and so historically came second. Finally, in what regards their combinatorics and applications, these matrices quite often come first, before both the Gaussian and the Wigner ones, with all this being of course a matter of knowledge and taste.

Summarizing, we have three main types of random matrices, which can be somehow designated as "complex", "real" and "positive", and that we will study in what follows. Let us also mention that there are many other interesting classes of random matrices, usually appearing as modifications of the above. More on these later.

Getting to work, let us fix some definitions, for the various types of normal variables appearing above. In the real case, the definition that we will need is as follows:

DEFINITION 16.20. *The normal law of parameter* 1 *is the following measure:*

$$g_1 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

More generally, the normal law of parameter $t > 0$ *is the following measure:*

$$g_t = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t} dx$$

These are also called Gaussian distributions, with "g" standing for Gauss.

The above laws are usually denoted $\mathcal{N}(0,1)$ and $\mathcal{N}(0,t)$, but since we will be doing here all kinds of probability, we will use simplified notations for all our measures. Let us also mention that the normal laws traditionally have 2 parameters, the mean and variance, but here we will not need the mean, all our theory using centered laws. Finally, observe that the above laws have indeed mass 1, as they should, due to the Gauss formula:

$$\int_{\mathbb{R}} e^{-x^2/2t} dx = \int_{\mathbb{R}} e^{-y^2} \sqrt{2t}\, dy = \sqrt{2\pi t}$$

Many things can be said about the normal laws, notably with the following result, that will play a key role in what follows, in our various moment computations:

THEOREM 16.21. *The moments of the normal law are the numbers*

$$M_k(g_t) = \sum_{\pi \in P_2(k)} t^{|\pi|}$$

where $P_2(k)$ *is the set of pairings of* $\{1, \ldots, k\}$*, and* $|.|$ *is the number of blocks.*

PROOF. The moments of the normal law are subject to the following formula:

$$\begin{aligned}
M_k(g_t) &= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} x^k e^{-x^2/2t} dx \\
&= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} (tx^{k-1}) \left(-e^{-x^2/2t} \right)' dx \\
&= \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} t(k-1) x^{k-2} e^{-x^2/2t} dx \\
&= t(k-1) \times \frac{1}{\sqrt{2\pi t}} \int_{\mathbb{R}} x^{k-2} e^{-x^2/2t} dx \\
&= t(k-1) M_{k-2}(g_t)
\end{aligned}$$

On the other hand, let us count the pairings of $\{1, \ldots, k\}$. In order to have such a pairing, we must pair 1 with one of the numbers $2, \ldots, k$, and then use a pairing of the remaining $k-2$ numbers. Thus, we have the following recurrence formula:

$$|P_2(k)| = (k-1)|P_2(k-2)|$$

Thus we obtain the result at $t = 1$, and the general case $t > 0$ follows too. □

In the complex case now, the definition that we will need is as follows:

DEFINITION 16.22. *The complex Gaussian law of parameter $t > 0$ is*

$$G_t = law\left(\frac{1}{\sqrt{2}}(a + ib)\right)$$

where a, b are independent, each following the law g_t.

As before with the real Gaussian laws, many things can be said here, notably with the following result, that will play a key role in what follows, for our computations:

THEOREM 16.23. *The moments of the complex normal law are the numbers*

$$M_k(G_t) = \sum_{\pi\in\mathcal{P}_2(k)} t^{|\pi|}$$

where $\mathcal{P}_2(k)$ are the matching pairings of $\{1, \ldots, k\}$, and $|.|$ is the number of blocks.

PROOF. We can assume that we are in the case $t = 1$, and here a straightforward computation shows that the moments are given by the following formula:

$$M_k = \begin{cases} (|k|/2)! & (k \text{ uniform}) \\ 0 & (k \text{ not uniform}) \end{cases}$$

On the other hand, the numbers $|\mathcal{P}_2(k)|$ are given by the same formula. Indeed, in order to have a matching pairing of k, our exponent $k = \circ\bullet\bullet\circ\ldots$ must be uniform, consisting of p copies of $\circ$ and p copies of $\bullet$, with $p = |k|/2$. But then the matching pairings of k correspond to the permutations of the $\bullet$ symbols, as to be matched with $\circ$ symbols, and so we have $p!$ such pairings. Thus, we are led to the result. □

In practice, we also need to know how to compute joint moments of independent normal variables. We have here the following result, to be heavily used later on:

THEOREM 16.24. *Given independent variables X_i, each following the complex normal law G_t, with $t > 0$ being a fixed parameter, we have the Wick formula*

$$E\left(X_{i_1}^{k_1}\ldots X_{i_s}^{k_s}\right) = t^{s/2}\#\left\{\pi\in\mathcal{P}_2(k)\Big|\pi\leq\ker i\right\}$$

where $k = k_1\ldots k_s$ and $i = i_1\ldots i_s$, for the joint moments of these variables.

PROOF. This is something well-known, and the basis for all possible computations with complex normal variables, which can be proved in two steps, as follows:

(1) Let us first discuss the case where we have a single variable X, which amounts in taking $X_i = X$ for any i in the formula in the statement. What we have to compute here are the moments of X, with respect to colored integer exponents $k = \circ\bullet\bullet\circ\ldots$, and the formula in the statement is equivalent to the one from Theorem 16.23, namely:

$$E(X^k) = t^{|k|/2}|\mathcal{P}_2(k)|$$

(2) In general now, the point is that we obtain the formula in the statement. Indeed, when expanding the product $X_{i_1}^{k_1}\ldots X_{i_s}^{k_s}$ and rearranging the terms, we are left with doing a number of computations as in (1), and then making the product of the expectations that we found. But this amounts precisely in counting the partitions in the statement, with the condition $\pi\leq\ker i$ there standing precisely for the fact that we are doing the various type (1) computations independently, and then making the product. □

Now by getting back to the Gaussian matrices, we have the following result, with $\mathcal{NC}_2(k)=\mathcal{P}_2(k)\cap NC(k)$ standing for the noncrossing pairings of a colored integer k:

THEOREM 16.25. *Given a sequence of Gaussian random matrices*

$$Z_N\in M_N(L^\infty(X))$$

having independent G_t variables as entries, for some fixed $t>0$, we have

$$M_k\left(\frac{Z_N}{\sqrt{N}}\right)\simeq t^{|k|/2}|\mathcal{NC}_2(k)|$$

for any colored integer $k=\circ\bullet\bullet\circ\ldots$, in the $N\to\infty$ limit.

PROOF. This is something standard, which can be done as follows:

(1) We fix $N\in\mathbb{N}$, and we let $Z=Z_N$. Let us first compute the trace of Z^k. With $k=k_1\ldots k_s$, and with the convention $(ij)^\circ=ij,(ij)^\bullet=ji$, we have:

$$\begin{aligned}Tr(Z^k)&=Tr(Z^{k_1}\ldots Z^{k_s})\\&=\sum_{i_1=1}^N\ldots\sum_{i_s=1}^N(Z^{k_1})_{i_1i_2}(Z^{k_2})_{i_2i_3}\ldots(Z^{k_s})_{i_si_1}\\&=\sum_{i_1=1}^N\ldots\sum_{i_s=1}^N(Z_{(i_1i_2)^{k_1}})^{k_1}(Z_{(i_2i_3)^{k_2}})^{k_2}\ldots(Z_{(i_si_1)^{k_s}})^{k_s}\end{aligned}$$

(2) Next, we rescale our variable Z by a $\sqrt{N}$ factor, as in the statement, and we also replace the usual trace by its normalized version, $tr=Tr/N$. Our formula becomes:

$$tr\left(\left(\frac{Z}{\sqrt{N}}\right)^k\right)=\frac{1}{N^{s/2+1}}\sum_{i_1=1}^N\ldots\sum_{i_s=1}^N(Z_{(i_1i_2)^{k_1}})^{k_1}(Z_{(i_2i_3)^{k_2}})^{k_2}\ldots(Z_{(i_si_1)^{k_s}})^{k_s}$$

Thus, the moment that we are interested in is given by:

$$M_k\left(\frac{Z}{\sqrt{N}}\right)=\frac{1}{N^{s/2+1}}\sum_{i_1=1}^N\ldots\sum_{i_s=1}^N\int_X(Z_{(i_1i_2)^{k_1}})^{k_1}(Z_{(i_2i_3)^{k_2}})^{k_2}\ldots(Z_{(i_si_1)^{k_s}})^{k_s}$$

(3) Let us apply now the Wick formula, from Theorem 16.24. We conclude that the moment that we are interested in is given by the following formula:

$$\begin{aligned}
&M_k\left(\frac{Z}{\sqrt{N}}\right)\\
=\ &\frac{t^{s/2}}{N^{s/2+1}}\sum_{i_1=1}^{N}\cdots\sum_{i_s=1}^{N}\#\left\{\pi\in\mathcal{P}_2(k)\Big|\pi\leq\ker\left((i_1i_2)^{k_1},(i_2i_3)^{k_2},\ldots,(i_si_1)^{k_s}\right)\right\}\\
=\ &t^{s/2}\sum_{\pi\in\mathcal{P}_2(k)}\frac{1}{N^{s/2+1}}\#\left\{i\in\{1,\ldots,N\}^s\Big|\pi\leq\ker\left((i_1i_2)^{k_1},(i_2i_3)^{k_2},\ldots,(i_si_1)^{k_s}\right)\right\}
\end{aligned}$$

(4) Our claim now is that in the $N\to\infty$ limit the combinatorics of the above sum simplifies, with only the noncrossing partitions contributing to the sum, and with each of them contributing precisely with a 1 factor, so that we will have, as desired:

$$\begin{aligned}
M_k\left(\frac{Z}{\sqrt{N}}\right)&=t^{s/2}\sum_{\pi\in\mathcal{P}_2(k)}\left(\delta_{\pi\in NC_2(k)}+O(N^{-1})\right)\\
&\simeq t^{s/2}\sum_{\pi\in\mathcal{P}_2(k)}\delta_{\pi\in NC_2(k)}\\
&=t^{s/2}|\mathcal{NC}_2(k)|
\end{aligned}$$

(5) In order to prove this, the first observation is that when k is not uniform, in the sense that it contains a different number of $\circ$, $\bullet$ symbols, we have $\mathcal{P}_2(k)=\emptyset$, and so:

$$M_k\left(\frac{Z}{\sqrt{N}}\right)=t^{s/2}|\mathcal{NC}_2(k)|=0$$

(6) Thus, we are left with the case where k is uniform. Let us examine first the case where k consists of an alternating sequence of $\circ$ and $\bullet$ symbols, as follows:

$$k=\underbrace{\circ\bullet\circ\bullet\ldots\ldots\circ\bullet}_{2p}$$

In this case it is convenient to relabel our multi-index $i=(i_1,\ldots,i_s)$, with $s=2p$, in the form $(j_1,l_1,j_2,l_2,\ldots,j_p,l_p)$. With this done, our moment formula becomes:

$$M_k\left(\frac{Z}{\sqrt{N}}\right)=t^p\sum_{\pi\in\mathcal{P}_2(k)}\frac{1}{N^{p+1}}\#\left\{j,l\in\{1,\ldots,N\}^p\Big|\pi\leq\ker\left(j_1l_1,j_2l_1,j_2l_2,\ldots,j_1l_p\right)\right\}$$

Now observe that, with k being as above, we have an identification $\mathcal{P}_2(k)\simeq S_p$, obtained in the obvious way. With this done too, our moment formula becomes:

$$M_k\left(\frac{Z}{\sqrt{N}}\right)=t^p\sum_{\pi\in S_p}\frac{1}{N^{p+1}}\#\left\{j,l\in\{1,\ldots,N\}^p\Big|j_r=j_{\pi(r)+1},l_r=l_{\pi(r)},\forall r\right\}$$

(7) We are now ready to do our asymptotic study, and prove the claim in (4). Let indeed $\gamma \in S_p$ be the full cycle, which is by definition the following permutation:

$$\gamma = (1\,2\,\ldots\,p)$$

In terms of γ, the conditions $j_r = j_{\pi(r)+1}$ and $l_r = l_{\pi(r)}$ found above read:

$$\gamma\pi \leq \ker j \quad , \quad \pi \leq \ker l$$

Counting the number of free parameters in our moment formula, we obtain:

$$M_k\left(\frac{Z}{\sqrt{N}}\right) = \frac{t^p}{N^{p+1}} \sum_{\pi\in S_p} N^{|\pi|+|\gamma\pi|} = t^p \sum_{\pi\in S_p} N^{|\pi|+|\gamma\pi|-p-1}$$

(8) The point now is that the last exponent is well-known to be ≤ 0, with equality precisely when the permutation $\pi \in S_p$ is geodesic, which in practice means that π must come from a noncrossing partition. Thus we obtain, in the $N \to \infty$ limit, as desired:

$$M_k\left(\frac{Z}{\sqrt{N}}\right) \simeq t^p |\mathcal{NC}_2(k)|$$

This finishes the proof in the case of the exponents k which are alternating, and the case where k is an arbitrary uniform exponent is similar, by permuting everything. □

As a conclusion to this, we have obtained as asymptotic law for the Gaussian matrices a certain mysterious distribution, having as moments some numbers which are similar to the moments of the usual normal laws, but with the "underlying matching pairings being now replaced by underlying matching noncrossing pairings". More on this later.

16c. Wigner and Wishart

Regarding now the Wigner matrices, we have here the following result, coming as a consequence of Theorem 16.25, via some simple algebraic manipulations:

THEOREM 16.26. *Given a sequence of Wigner random matrices*

$$Z_N \in M_N(L^\infty(X))$$

having independent G_t variables as entries, with $t > 0$, up to $Z_N = Z_N^*$, we have

$$M_k\left(\frac{Z_N}{\sqrt{N}}\right) \simeq t^{k/2} |NC_2(k)|$$

for any integer $k \in \mathbb{N}$, in the $N \to \infty$ limit.

PROOF. This can be deduced from a direct computation based on the Wick formula, similar to that from the proof of Theorem 16.25, but the best is to deduce this result from Theorem 16.25 itself. Indeed, we know from there that for Gaussian matrices $Y_N \in$

$M_N(L^\infty(X))$ we have the following formula, valid for any colored integer $K = \circ\bullet\bullet\circ\ldots$, in the $N\to\infty$ limit, with $\mathcal{NC}_2$ standing for noncrossing matching pairings:

$$M_K\left(\frac{Y_N}{\sqrt{N}}\right)\simeq t^{|K|/2}|\mathcal{NC}_2(K)|$$

By doing some combinatorics, we deduce from this that we have the following formula for the moments of the matrices $Re(Y_N)$, with respect to usual exponents, $k\in\mathbb{N}$:

$$\begin{aligned}M_k\left(\frac{Re(Y_N)}{\sqrt{N}}\right)&=2^{-k}\cdot M_k\left(\frac{Y_N}{\sqrt{N}}+\frac{Y_N^*}{\sqrt{N}}\right)\\&=2^{-k}\sum_{|K|=k}M_K\left(\frac{Y_N}{\sqrt{N}}\right)\\&\simeq 2^{-k}\sum_{|K|=k}t^{k/2}|\mathcal{NC}_2(K)|\\&=2^{-k}\cdot t^{k/2}\cdot 2^{k/2}|\mathcal{NC}_2(k)|\\&=2^{-k/2}\cdot t^{k/2}|NC_2(k)|\end{aligned}$$

Now since the matrices $Z_N=\sqrt{2}Re(Y_N)$ are of Wigner type, this gives the result. □

Summarizing, all this brings us into counting noncrossing pairings. So, let us start with some preliminaries here. We first have the following well-known result:

THEOREM 16.27. *The Catalan numbers, which are by definition given by*

$$C_k=|NC_2(2k)|$$

satisfy the following recurrence formula, with initial data $C_0=C_1=1$,

$$C_{k+1}=\sum_{a+b=k}C_aC_b$$

their generating series $f(z)=\sum_{k\geq 0}C_kz^k$ satisfies the equation

$$zf^2-f+1=0$$

and is given by the following explicit formula,

$$f(z)=\frac{1-\sqrt{1-4z}}{2z}$$

and we have the following explicit formula for these numbers:

$$C_k=\frac{1}{k+1}\binom{2k}{k}$$

Numerically, these numbers are $1,1,2,5,14,42,132,429,1430,4862,16796,\ldots$

PROOF. We must count the noncrossing pairings of $\{1,\ldots,2k\}$. Now observe that such a pairing appears by pairing 1 to an odd number, $2a+1$, and then inserting a noncrossing pairing of $\{2,\ldots,2a\}$, and a noncrossing pairing of $\{2a+2,\ldots,2l\}$. We conclude that we have the following recurrence formula for the Catalan numbers:

$$C_k=\sum_{a+b=k-1}C_aC_b$$

In terms of the generating series $f(z)=\sum_{k\geq 0}C_kz^k$, this recurrence formula reads:

$$\begin{aligned}zf^2&=\sum_{a,b\geq 0}C_aC_bz^{a+b+1}\\&=\sum_{k\geq 1}\sum_{a+b=k-1}C_aC_bz^k\\&=\sum_{k\geq 1}C_kz^k\\&=f-1\end{aligned}$$

Thus f satisfies $zf^2-f+1=0$, and by solving this equation, and choosing the solution which is bounded at $z=0$, we obtain the following formula:

$$f(z)=\frac{1-\sqrt{1-4z}}{2z}$$

In order to finish, we use the generalized binomial formula, which gives:

$$\sqrt{1+t}=1-2\sum_{k=1}^{\infty}\frac{1}{k}\binom{2k-2}{k-1}\left(\frac{-t}{4}\right)^k$$

Now back to our series f, we obtain the following formula for it:

$$\begin{aligned}f(z)&=\frac{1-\sqrt{1-4z}}{2z}\\&=\sum_{k=1}^{\infty}\frac{1}{k}\binom{2k-2}{k-1}z^{k-1}\\&=\sum_{k=0}^{\infty}\frac{1}{k+1}\binom{2k}{k}z^k\end{aligned}$$

It follows that the Catalan numbers are given by:

$$C_k=\frac{1}{k+1}\binom{2k}{k}$$

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

In order to recapture now the Wigner measure from its moments, we can use:

PROPOSITION 16.28. *The Catalan numbers are the even moments of*

$$\gamma_1 = \frac{1}{2\pi}\sqrt{4-x^2}dx$$

called standard semicircle law. As for the odd moments of γ_1, these all vanish.

PROOF. The even moments of γ_1 can be computed with the change of variable $x = 2\cos t$, and some trigonometric know-how, and we are led to the following formula:

$$\begin{aligned} M_{2k} &= \frac{1}{\pi}\int_0^2 \sqrt{4-x^2}x^{2k}dx \\ &= \frac{4^{k+1}}{\pi}\int_0^{\pi/2} \cos^{2k}t\sin^2 t\,dt \\ &= C_k \end{aligned}$$

As for the odd moments, these all vanish, because the density of γ_1 is an even function. Thus, we are led to the conclusion in the statement. □

More generally, we have the following result, involving a parameter $t > 0$:

PROPOSITION 16.29. *Given $t > 0$, the real measure having as even moments the numbers $M_{2k} = t^k C_k$ and having all odd moments 0 is the measure*

$$\gamma_t = \frac{1}{2\pi t}\sqrt{4t-x^2}dx$$

called Wigner semicircle law on $[-2\sqrt{t}, 2\sqrt{t}]$.

PROOF. This follows indeed from Proposition 16.28, via a change of variables. □

Now by putting everything together, we obtain the Wigner theorem, as follows:

THEOREM 16.30. *Given a sequence of Wigner random matrices*

$$Z_N \in M_N(L^\infty(X))$$

which by definition have i.i.d. complex normal entries, up to $Z_N = Z_N^*$, we have

$$Z_N \sim \gamma_t$$

in the $N \to \infty$ limit, where $\gamma_t = \frac{1}{2\pi t}\sqrt{4t-x^2}dx$ is the Wigner semicircle law.

PROOF. This follows indeed from all the above, and more specifically, by combining Theorem 16.26, Theorem 16.27 and Proposition 16.29. □

Let us discuss now the Wishart matrices, which are the positive analogues of the Wigner matrices. Quite surprisingly, the computation here leads to the Catalan numbers, but not in the same way as for the Wigner matrices, the result being as follows:

THEOREM 16.31. *Given a sequence of complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times N$ complex Gaussian of parameter $t > 0$, we have

$$M_k\left(\frac{W_N}{N}\right) \simeq t^k C_k$$

for any exponent $k \in \mathbb{N}$, in the $N \to \infty$ limit.

PROOF. There are several possible proofs for this result, as follows:

(1) A first method is by using the formula that we have in Theorem 16.25, for the Gaussian matrices Y_N. Indeed, we know from there that we have the following formula, valid for any colored integer $K = \circ\bullet\bullet\circ\ldots$, in the $N \to \infty$ limit:

$$M_K\left(\frac{Y_N}{\sqrt{N}}\right) \simeq t^{|K|/2}|\mathcal{NC}_2(K)|$$

With $K = \circ\bullet\circ\bullet\ldots$, alternating word of length $2k$, with $k \in \mathbb{N}$, this gives:

$$M_k\left(\frac{Y_N Y_N^*}{N}\right) \simeq t^k|\mathcal{NC}_2(K)|$$

Thus, in terms of the Wishart matrix $W_N = Y_N Y_N^*$ we have, for any $k \in \mathbb{N}$:

$$M_k\left(\frac{W_N}{N}\right) \simeq t^k|\mathcal{NC}_2(K)|$$

The point now is that, by doing some combinatorics, we have:

$$|\mathcal{NC}_2(K)| = |NC_2(2k)| = C_k$$

Thus, we are led to the formula in the statement.

(2) A second method, that we will explain now as well, is by proving the result directly, starting from definitions. The matrix entries of our matrix $W = W_N$ are given by:

$$W_{ij} = \sum_{r=1}^{N} Y_{ir}\bar{Y}_{jr}$$

Thus, the normalized traces of powers of W are given by the following formula:

$$\begin{aligned} tr(W^k) &= \frac{1}{N}\sum_{i_1=1}^{N}\ldots\sum_{i_k=1}^{N} W_{i_1i_2}W_{i_2i_3}\ldots W_{i_ki_1} \\ &= \frac{1}{N}\sum_{i_1=1}^{N}\ldots\sum_{i_k=1}^{N}\sum_{r_1=1}^{N}\ldots\sum_{r_k=1}^{N} Y_{i_1r_1}\bar{Y}_{i_2r_1}Y_{i_2r_2}\bar{Y}_{i_3r_2}\ldots Y_{i_kr_k}\bar{Y}_{i_1r_k} \end{aligned}$$

By rescaling now W by a $1/N$ factor, as in the statement, we obtain:

$$tr\left(\left(\frac{W}{N}\right)^k\right)=\frac{1}{N^{k+1}}\sum_{i_1=1}^{N}\ldots\sum_{i_k=1}^{N}\sum_{r_1=1}^{N}\ldots\sum_{r_k=1}^{N}Y_{i_1r_1}\bar{Y}_{i_2r_1}Y_{i_2r_2}\bar{Y}_{i_3r_2}\ldots Y_{i_kr_k}\bar{Y}_{i_1r_k}$$

By using now the Wick rule, we obtain the following formula for the moments, with $K=\circ\bullet\circ\bullet\ldots$, alternating word of lenght $2k$, and with $I=(i_1r_1,i_2r_1,\ldots,i_kr_k,i_1r_k)$:

$$\begin{aligned}M_k\left(\frac{W}{N}\right)&=\frac{t^k}{N^{k+1}}\sum_{i_1=1}^{N}\ldots\sum_{i_k=1}^{N}\sum_{r_1=1}^{N}\ldots\sum_{r_k=1}^{N}\#\left\{\pi\in\mathcal{P}_2(K)\Big|\pi\leq\ker(I)\right\}\\&=\frac{t^k}{N^{k+1}}\sum_{\pi\in\mathcal{P}_2(K)}\#\left\{i,r\in\{1,\ldots,N\}^k\Big|\pi\leq\ker(I)\right\}\end{aligned}$$

In order to compute this quantity, we use the standard bijection $\mathcal{P}_2(K)\simeq S_k$. By identifying the pairings $\pi\in\mathcal{P}_2(K)$ with their counterparts $\pi\in S_k$, we obtain:

$$M_k\left(\frac{W}{N}\right)=\frac{t^k}{N^{k+1}}\sum_{\pi\in S_k}\#\left\{i,r\in\{1,\ldots,N\}^k\Big|i_s=i_{\pi(s)+1},r_s=r_{\pi(s)},\forall s\right\}$$

Now let $\gamma\in S_k$ be the full cycle, which is by definition the following permutation:

$$\gamma=(1\,2\,\ldots\,k)$$

The general factor in the product computed above is then 1 precisely when following two conditions are simultaneously satisfied:

$$\gamma\pi\leq\ker i\quad,\quad\pi\leq\ker r$$

Counting the number of free parameters in our moment formula, we obtain:

$$M_k\left(\frac{W}{N}\right)=t^k\sum_{\pi\in S_k}N^{|\pi|+|\gamma\pi|-k-1}$$

The point now is that the last exponent is well-known to be ≤ 0, with equality precisely when the permutation $\pi\in S_k$ is geodesic, which in practice means that π must come from a noncrossing partition. Thus we obtain, in the $N\to\infty$ limit:

$$M_k\left(\frac{W}{N}\right)\simeq t^kC_k$$

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

As a consequence of the above result, we have a new look on the Catalan numbers, which is more adapted to our present Wishart matrix considerations, as follows:

PROPOSITION 16.32. *The Catalan numbers $C_k = |NC_2(2k)|$ appear as well as*

$$C_k = |NC(k)|$$

where $NC(k)$ is the set of all noncrossing partitions of $\{1, \ldots, k\}$.

PROOF. This follows indeed from the proof of Theorem 16.31. Observe that we obtain as well a formula in terms of matching pairings of alternating colored integers. □

The direct explanation for the above formula, relating noncrossing partitions and pairings, comes form the following result, which is very useful, and good to know:

PROPOSITION 16.33. *We have a bijection between noncrossing partitions and pairings*

$$NC(k) \simeq NC_2(2k)$$

which is constructed as follows:

(1) *The application $NC(k) \to NC_2(2k)$ is the "fattening" one, obtained by doubling all the legs, and doubling all the strings as well.*
(2) *Its inverse $NC_2(2k) \to NC(k)$ is the "shrinking" application, obtained by collapsing pairs of consecutive neighbors.*

PROOF. The fact that the two operations in the statement are indeed inverse to each other is clear, by computing the corresponding two compositions, with the remark that the construction of the fattening operation requires the partitions to be noncrossing. □

Getting back now to probability, we are led to the question of finding the law having the Catalan numbers as moments, in the above way. The result here is as follows:

PROPOSITION 16.34. *The real measure having the Catalan numbers as moments is*

$$\pi_1 = \frac{1}{2\pi}\sqrt{4x^{-1} - 1}\, dx$$

called Marchenko-Pastur law of parameter 1.

PROOF. The moments of the law π_1 in the statement can be computed with the change of variable $x = 4\cos^2 t$, and some trigonometric know-how, as follows:

$$\begin{aligned} M_k &= \frac{1}{2\pi}\int_0^4 \sqrt{4x^{-1} - 1}\, x^k dx \\ &= \frac{4^{k+1}}{\pi}\int_0^{\pi/2} \cos^{2k} t \sin^2 t\, dt \\ &= C_k \end{aligned}$$

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

Now back to the Wishart matrices, we are led to the following result:

THEOREM 16.35. *Given a sequence of complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times N$ complex Gaussian of parameter $t > 0$, we have

$$\frac{W_N}{tN} \sim \frac{1}{2\pi}\sqrt{4x^{-1} - 1}\, dx$$

with $N \to \infty$, with the limiting measure being the Marchenko-Pastur law π_1.

PROOF. This follows indeed from Theorem 16.31 and Proposition 16.34. □

As a comment now, while the above result is definitely something interesting at $t = 1$, at general $t > 0$ this looks more like a "fake" generalization of the $t = 1$ result, because the law π_1 stays the same, modulo a trivial rescaling. The reasons behind this phenomenon are quite subtle, and skipping some discussion, the point is that Theorem 16.35 is indeed something "fake" at general $t > 0$, and the correct generalization of the $t = 1$ computation, involving more general classes of complex Wishart matrices, is as follows:

THEOREM 16.36. *Given a sequence of general complex Wishart matrices*

$$W_N = Y_N Y_N^* \in M_N(L^\infty(X))$$

with Y_N being $N \times M$ complex Gaussian of parameter 1, we have

$$\frac{W_N}{N} \sim \max(1 - t, 0)\delta_0 + \frac{\sqrt{4t - (x - 1 - t)^2}}{2\pi x}\, dx$$

with $M = tN \to \infty$, with the limiting measure being the Marchenko-Pastur law π_t.

PROOF. This follows once again by using the moment method, the limiting moments in the $M = tN \to \infty$ regime being as follows, after doing the combinatorics:

$$M_k\left(\frac{W_N}{N}\right) \simeq \sum_{\pi \in NC(k)} t^{|\pi|}$$

But these numbers are the moments of the Marchenko-Pastur law π_t, which in addition has the density given by the formula in the statement, and this gives the result. □

16d. Back to groups

Many things can be said about the limiting laws found by Wigner and Marchenko-Pastur, and in what concerns us, we will present an interpretation of them related to our favorite matrix groups, SU_2 and SO_3. In order to discuss this, we will need the following formula, coming as a continuation of some previous formulae from chapter 9:

THEOREM 16.37. *We have the following formula,*

$$\int_0^{\pi/2} \cos^r t \sin^s t\, dt = \left(\frac{\pi}{2}\right)^{\varepsilon(r)\varepsilon(s)} \frac{r!!s!!}{(r+s+1)!!}$$

where $\varepsilon(r) = 1$ if r is even, and $\varepsilon(r) = 0$ if r is odd.

PROOF. Let us call I_{rs} the integral in the statement. In order to do the partial integration, observe that we have the following formula:

$$\begin{aligned}(\cos^r t \sin^s t)' &= r\cos^{r-1} t(-\sin t)\sin^s t + \cos^r t \cdot s\sin^{s-1} t\cos t\\ &= -r\cos^{r-1} t\sin^{s+1} t + s\cos^{r+1} t\sin^{s-1} t\end{aligned}$$

By integrating between 0 and $\pi/2$, we obtain, for $r, s > 0$:

$$rI_{r-1,s+1} = sI_{r+1,s-1}$$

Thus, we can compute I_{rs} by recurrence, and when s is even we obtain:

$$\begin{aligned}I_{rs} &= \frac{r!!s!!}{(r+s)!!} I_{r+s}\\ &= \frac{r!!s!!}{(r+s)!!}\left(\frac{\pi}{2}\right)^{\varepsilon(r+s)} \frac{(r+s)!!}{(r+s+1)!!}\\ &= \left(\frac{\pi}{2}\right)^{\varepsilon(r)\varepsilon(s)} \frac{r!!s!!}{(r+s+1)!!}\end{aligned}$$

Observe that this gives the result for r even as well, by symmetry. In the remaining case now, where both the exponents r, s are odd, we can use once again the equality $rI_{r-1,s+1} = sI_{r+1,s-1}$ found above, and the recurrence gives the following formula:

$$\begin{aligned}I_{rs} &= \frac{r!!s!!}{(r+s-1)!!} I_{r+s-1,1}\\ &= \frac{r!!s!!}{(r+s-1)!!} \cdot \frac{1}{r+s}\\ &= \left(\frac{\pi}{2}\right)^{\varepsilon(r)\varepsilon(s)} \frac{r!!s!!}{(r+s+1)!!}\end{aligned}$$

Thus, we obtain indeed the formula in the statement. □

As an application of the above trigonometric formula, we have:

THEOREM 16.38. *The moments of the hyperspherical variables are*

$$\int_{S_{\mathbb{R}}^{N-1}} x_i^p dx = \frac{(N-1)!!p!!}{(N+p-1)!!}$$

and the rescaled variables $y_i = \sqrt{N}x_i$ become normal and independent with $N \to \infty$.

PROOF. By using spherical coordinates, with input from Theorem 16.37, we are led to the folllowing formula, for the joint moments of the hyperspherical variables:

$$\int_{S_{\mathbb{R}}^{N-1}} x_1^{k_1} \ldots x_N^{k_N}\, dx = \frac{(N-1)!!k_1!!\ldots k_N!!}{(N+\Sigma k_i-1)!!}$$

But this gives the formula in the statement, and the last assertion too. □

Getting back now to our random matrix problematics, here is how the quite mysterious semicircle Wigner law γ_1 appears, geometrically, in relation with the group SU_2:

THEOREM 16.39. *The main character of SU_2 follows the following law,*

$$\gamma_1 = \frac{1}{2\pi}\sqrt{4-x^2}dx$$

which is the Wigner law of parameter 1.

PROOF. We know that SU_2 is the group of unitary rotations $U \in U_2$ of determinant one, and as explained in chapter 3, by solving the equation $U^* = U^{-1}$, we are led to:

$$SU_2 = \left\{ \begin{pmatrix} a+ib & c+id \\ -c+id & a-ib \end{pmatrix} \;\middle|\; a^2+b^2+c^2+d^2=1 \right\}$$

In this picture, the main character is given by the following formula:

$$\chi \begin{pmatrix} a+ib & c+id \\ -c+id & a-ib \end{pmatrix} = 2a$$

We are therefore left with computing the law of the following variable:

$$a \in C(S_{\mathbb{R}}^3)$$

But this is something very familiar, namely a hyperspherical variable at $N=4$, so we can use here Theorem 16.38. We obtain the following moment formula:

$$\begin{aligned} \int_{S_{\mathbb{R}}^3} a^{2k} &= \frac{3!!(2k)!!}{(2k+3)!!} \\ &= 2\cdot\frac{(2k)!}{2^k k!2^{k+1}(k+1)!} \\ &= \frac{1}{4^k}\cdot\frac{1}{k+1}\binom{2k}{k} \\ &= \frac{C_k}{4^k} \end{aligned}$$

Thus the variable $2a \in C(S_{\mathbb{R}}^3)$ follows the Wigner semicircle law γ_1, as claimed. □

Quite nice all this, and things are not over here. We have as well a similar result regarding the Marchenko-Pastur law π_1, involving this time the group SO_3, as follows:

THEOREM 16.40. *The main character of SO_3, modified by adding 1 to it, given in standard Euler-Rodrigues coordinates by*

$$\chi = 4a^2$$

follows a squared semicircle law, or Marchenko-Pastur law π_1.

PROOF. This follows by using the quotient map $SU_2 \to SO_3$, and the result for SU_2. Indeed, by using the Euler-Rodrigues formula, in the context of Theorem 16.39 and its proof, the main character of SO_3, modified by adding 1 to it, is given by:

$$\chi = (3a^2 - b^2 - c^2 - d^2) + 1 = 4a^2$$

Now recall from the proof of Theorem 16.39 that we have:

$$2a \sim \gamma_1$$

On the other hand, a quick comparison between the moment formulae for the Wigner and Marchenko-Pastur laws, which are very similar, shows that we have:

$$f \sim \gamma_1 \implies f^2 \sim \pi_1$$

Thus, with $f = 2p$, we obtain the result in the statement. □

And with this, done with random matrices. Of course, all the above was just the basic theory, and for more, have a look at any random matrix or free probability book.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter. However, there is a lot of further advanced linear algebra to be discovered, quite often in relation with operators and infinite dimensions. In the hope that you will go this way.

Bibliography

[1] V.I. Arnold, Ordinary differential equations, Springer (1973).

[2] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] M.F. Atiyah, K-theory, CRC Press (1964).

[5] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[6] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).

[7] T. Banica, Linear algebra and group theory (2024).

[8] T. Banica, Geometry and topology (2025).

[9] T. Banica, Principles of operator algebras (2024).

[10] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[11] I. Bengtsson and K. Życzkowski, Geometry of quantum states, Cambridge Univ. Press (2006).

[12] N. Berline, E. Getzler and M. Vergne, Heat kernels and Dirac operators, Springer (2004).

[13] B. Blackadar, K-theory for operator algebras, Cambridge Univ. Press (1986).

[14] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[15] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[16] A. Connes, Noncommutative geometry, Academic Press (1994).

[17] A. Connes and M. Marcolli, Noncommutative geometry, quantum fields and motives, AMS (2008).

[18] H.S.M. Coxeter, Regular polytopes, Dover (1948).

[19] W. de Launey and D. Flannery, Algebraic design theory, AMS (2011).

[20] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[21] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).

[22] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).

[23] S. Dodelson, Modern cosmology, Academic Press (2003).

[24] S.K. Donaldson, Riemann surfaces, Oxford Univ. Press (2004).

[25] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[26] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[27] L.C. Evans, Partial differential equations, AMS (1998).

[28] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[29] E. Fermi, Thermodynamics, Dover (1937).

[30] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[31] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[32] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[33] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[34] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[35] A.P. French, Special relativity, Taylor and Francis (1968).

[36] W. Fulton, Algebraic topology, Springer (1995).

[37] W. Fulton and J. Harris, Representation theory, Springer (1991).

[38] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[39] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[40] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[41] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[42] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[43] P. Griffiths and J. Harris, Principles of algebraic geometry, Wiley (1994).

[44] G.H. Hardy and E.M. Wright, An introduction to the theory of numbers, Oxford Univ. Press (1938).

[45] J. Harris, Algebraic geometry, Springer (1992).

[46] R. Hartshorne, Algebraic geometry, Springer (1977).

[47] M.P. Hobson, G.P. Efstathiou and A.N. Lasenby, General relativity, Cambridge Univ. Press (2006).

[48] L. Hörmander, The analysis of linear partial differential operators, Springer (1983).

[49] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[50] K. Huang, Introduction to statistical physics, CRC Press (2001).

[51] K. Huang, Fundamental forces of nature, World Scientific (2007).

[52] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).

[53] J.E. Humphreys, Linear algebraic groups, Springer (1975).

[54] K. Ireland and M. Rosen, A classical introduction to modern number theory, Springer (1982).

[55] N. Jacobson, Basic algebra, Dover (1974).

[56] V.F.R. Jones, Subfactors and knots, AMS (1991).

[57] V.F.R. Jones and V.S Sunder, Introduction to subfactors, Cambridge Univ. Press (1997).

[58] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[59] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[60] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[61] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[62] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[63] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[64] S. Lang, Algebra, Addison-Wesley (1993).

[65] S. Lang, Abelian varieties, Dover (1959).

[66] P. Lax, Linear algebra and its applications, Wiley (2007).

[67] P. Lax, Functional analysis, Wiley (2002).

[68] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[69] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[70] J.M. Lee, Introduction to topological manifolds, Springer (2011).

[71] J.M. Lee, Introduction to smooth manifolds, Springer (2012).

[72] J.M. Lee, Introduction to Riemannian manifolds, Springer (2019).

[73] M.L. Mehta, Random matrices, Elsevier (2004).

[74] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[75] P. Petersen, Linear algebra, Springer (2012).

[76] P. Petersen, Riemannian geometry, Springer (1998).

[77] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[78] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[79] W. Rudin, Fourier analysis on groups, Dover (1972).

[80] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[81] B. Ryden and B.M. Peterson, Foundations of astrophysics, Cambridge Univ. Press (2010).

[82] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[83] B. Schutz, A first course in general relativity, Cambridge Univ. Press (2009).

[84] J.P. Serre, A course in arithmetic, Springer (1973).

[85] J.P. Serre, Linear representations of finite groups, Springer (1977).

[86] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[87] J.H. Silverman, The arithmetic of elliptic curves, Springer (1986).

[88] J.H. Silverman and J.T. Tate, Rational points on elliptic curves, Springer (2015).

[89] B. Singh, Basic commutative algebra, World Scientific (2011).

[90] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[91] C.H. Taubes, Differential geometry, Oxford Univ. Press (2011).

[92] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[93] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).

[94] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[95] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[96] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[97] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[98] H. Weyl, The theory of groups and quantum mechanics, Princeton Univ. Press (1931).

[99] H. Weyl, The classical groups: their invariants and representations, Princeton Univ. Press (1939).

[100] H. Weyl, Space, time, matter, Princeton Univ. Press (1918).

Index